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

Volume 17 Issue 7
July 2026

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

THE SCIENCE AND INFORMATION ORGANIZATION

www.ijacsa.thesai.org

Table of Contents

Paper 1: Vision Transformers and Convolutional Neural Networks for Land Use Scene Classification

Authors: Arun D. Kulkarni

PAGES 1 - 11

Paper 2: Privacy-Preserving Financial Fraud Detection Using Federated Heterogeneous Temporal GraphSAGE

Authors: Sabab Islam, Md. Abul Kalam Azad, Abdul Kadar Muhammad Masum, Ramona Birau, Virgil Popescu, Roxana Mihaela Nioata (Chireac), Gabriela Ana Maria Lupu (Filip)

PAGES 12 - 26

Paper 3: Attention-Augmented PointNet with Hybrid Ant Colony Optimization and Bayesian Hyperparameter Search for Robust 3D Point Cloud Classification in Augmented Reality Applications

Authors: Ahmed Almaghthawi, Fahad Alharbi, Radwan M. Batyha, Suhaila Abuowaida

PAGES 27 - 35

Paper 4: Ordinal Factor Analysis with Robust Estimation and Predictive Models

Authors: Abuelgasim Osman, Zahayu Binti Md Yusof

PAGES 36 - 41

Paper 5: Neural Network Implementation of the Renormalization Group for Fault Diagnosis with Class Imbalance

Authors: Evgeny Nikulchev, Dmitry Ilin

PAGES 42 - 49

Paper 6: MatchIA-GCL: A Hybrid Graph Contrastive Learning Approach for Semantic Schema Matching

Authors: Mohamed Raoui, Moulay Hafid El Yazidi, Ahmed Zellou

PAGES 50 - 59

Paper 7: Optimized Lattice-Based Ciphertext Policy Attribute-Based Encryption for Quantum-Resistant Secure Data Sharing

Authors: Purnima, Tejaswi Khanna, Deepak Kumar Verma

PAGES 60 - 74

Paper 8: Binary Enhanced White Shark Optimization Algorithm and Its Application to Semi-Supervised Feature Selection

Authors: Jian Zhou, Jiawen Yang

PAGES 75 - 88

Paper 9: NADOS: Reliability-Gated Difficulty-Aware Oversampling for Noisy Imbalanced Classification

Authors: Ednel Ashraff Misran, Syahid Anuar, Adam Mohd Khairuddin

PAGES 89 - 98

Paper 10: A RAG-Enhanced Human-in-the-Loop Framework for Automated Assessment of Engineering Laboratory Reports

Authors: Amina Abbi, Mohammed Skouri, Mustapha Raoufi

PAGES 99 - 107

Paper 11: Detecting Metadata-Conditioned Response Drift in AI-Generated Visual Information Systems: A Two-Phase Human–Computer Interaction Framework

Authors: Yu Jia, Shaojie Zhang, Na Li

PAGES 108 - 117

Paper 12: Hybrid Machine Learning and Reinforcement Learning for IoT Device Classification and Bandwidth Optimization

Authors: Manjunatha T N, Vidyalakshmi K, Dankan Gowda V

PAGES 118 - 126

Paper 13: Pattern Analysis of Global AI Adoption and Digital Literacy Development in Science Education: Basis for a Recommender System Framework

Authors: Honelly Mae S. Cascolan

PAGES 127 - 138

Paper 14: Automated Multi-Task Deep Learning Framework for Simultaneous Identification of Dental Pathologies and Arch Localization in Orthopantomograms (OPG)

Authors: Hatim Elhag, Norehan Binti Mokhtar, Siti Noor Fazliah Binti Mohd Noor, Gururajaprasad Kaggal Lakshmana Rao, Nordin Zakaria, Fakhitah Ridzuan, Mohammed Bin Salah

PAGES 139 - 148

Paper 15: Inclusive Digital Health Platform for Well-Being and Loneliness in SOGIESC Older Adults

Authors: Yada Atanan, Titikarn Assatarakul, Akara Metasuk

PAGES 149 - 160

Paper 16: Image-Based Iron Ore Feed Load Classification Using EfficientNet-B0 with Region-Density Attention and Hierarchical Feature Fusion

Authors: Sami F Karali, Fawaz Alanazi, Eman Ramadan Elsharkawy, Asma A. Alhashmi, Abed Saif Ahmed Alghawli, Imen B. Mohamed, Chams Sallami, Abdulbasit A. Darem

PAGES 161 - 173

Paper 17: An Intelligent Writing Tutor for ESL Learners: A Rule-Based and NLP-Assisted Feedback Framework

Authors: Ruth L. Ladwingon

PAGES 174 - 186

Paper 18: Migration Method for Parametric Features in Heterogeneous CAD Systems

Authors: Gang WANG, Guangfa ZHANG

PAGES 187 - 194

Paper 19: Automated Classification of Moroccan Procedural Court Orders Using CAMeL-BERT

Authors: Saliha YASSINE, Ouafaa IBRIHICH, Mohammed ESSOUJAA, Mustapha ESGHIR

PAGES 195 - 203

Paper 20: A Hybrid SMOTE-Ensemble Framework for Robust Speech Emotion Recognition on Imbalanced Data

Authors: Eyad Megdadi, Mariam Aljouhi, Zainab Alblooshi, Khaled Shaalan

PAGES 204 - 218

Paper 21: XPA-FedCBR: A Federated Case-Based Reasoning System for Explainable and Privacy-Aware Lung Cancer Diagnosis

Authors: Devyani Rawat, Shuchi Bhadula, Sachin Sharma

PAGES 219 - 228

Paper 22: Enhanced Simple Linear Iterative Clustering (ESLIC) for Precise Skin Lesion Segmentation in Dermoscopic Images

Authors: Leni Natalia Zulita, Sarjon Defit, Sumijan

PAGES 229 - 239

Paper 23: Adaptive Mirage Search Optimization for Educational Workload Scheduling in Multi-Cloud Environments

Authors: Mohammed A.S. Mosleh, Elham Alzain, Hesham M.A. Abdullah, Ali Alshebami

PAGES 240 - 250

Paper 24: HISaaS: A Cloud-Based Framework for Semantic Enrichment of Legacy Health Databases Using OpenAI Embeddings

Authors: Toufik Salmi, Abdeslem Dennai

PAGES 251 - 262

Paper 25: HETRA: A Real-Time Traffic Detection and Directional Counting System Using YOLOv8 and DeepSORT

Authors: Kuldeep Kumar Shukla, Ajai Kumar Singh

PAGES 263 - 274

Paper 26: Improving Arabic Relational Aggregated Search Using a Hybrid RAG and AraBERT Framework

Authors: Sara Saad Sedhom, Ahmed Younes, Islam Elkabani, Ashraf Elsayed

PAGES 275 - 284

Paper 27: Design, Implementation, and Evaluation of a Mobile Application for Real-Time Incident Management in Local Government

Authors: Marleny Peralta Ascue, Luis Bryan Caballero Pinto, José Luis Poluco Sicos, José Luis Merma Aroni

PAGES 285 - 292

Paper 28: EduDarijaBERT: An Applied Transformer-Based Framework for Sentiment Analysis in Moroccan Arabic Educational Discourse

Authors: Abderrahim Ait Ichou, Ali Ouacha

PAGES 293 - 304

Paper 29: MAMGAT-Net: A Multimodal Graph Attention and Multi-Scale Feature Fusion Framework for Early Breast Cancer Diagnosis from Mammography and Histopathological Images

Authors: Mohammad Riyaz Belgaum

PAGES 305 - 320

Paper 30: An Intelligent System for Energy-Aware Flexible Fault Tolerance in Cloud

Authors: Moin Hasan, Shailee Lohmor Choudhary, Rinku Sharma Dixit, Sujith Jayaprakash

PAGES 321 - 329

Paper 31: INSPIRE: Understanding Local-Global Representation Learning via Controlled Self-Supervised Training and Representation Interpretability

Authors: Reem Alharthi, Rashid Mehmood, Aiiad Albeshri, Abdulaziz A. Almuzaini

PAGES 330 - 347

Paper 32: Enhancing Diabetic Retinopathy Classification Through Advanced Deep Convolutional Neural Network Architectures

Authors: Sifeddine Elkardoudi, Khalil Ladrham, Ahmed Eddaoui, Mohamed Talea

PAGES 348 - 359

Paper 33: The Evolution of Clinical Intelligence Through GenAI Co-pilots: A Systematic Review and Thematic Synthesis

Authors: Li-Chen Cheng, Chia-Yu Hung, Te-Nien Chien

PAGES 360 - 375

Paper 34: Localized Actionable Explainability for MOOC Dropout Intervention Using Clickstream Learning Analytics

Authors: Mohamed Amine OUASSIL, Mohammed JEBBEARI, Mouaad ERRAMI, Rabia Rachidi, Soufiane HAMIDA, Bouchaib CHERRADI, Abdelhadi RAIHANI

PAGES 376 - 387

Paper 35: Counterfeit Currency Detection Through Deep Convolutional Generative Adversarial Network

Authors: Salman Jan, Fara Gul, Mohammad Riyaz Belgaum, Shahid Kamal, Afaq Ahmad, Muhammad Bilal, Atif Khan

PAGES 388 - 397

Paper 36: FGHA-Net: Hybrid Attention Learning for Froth Flotation Recovery Prediction in Mineral Beneficiation

Authors: Mohammed I Thanoon, Fawaz Alanazi, Shaaban M. Shaaban, Asma A. Alhashmi, Abed Saif Ahmed Alghawli, Ahlem Fatnassi, Majed Nawaz, Abdulbasit A. Darem

PAGES 398 - 409

Paper 37: From Cloud to On-Premises: A Systematic Review of Cloud Repatriation Drivers, Strategies, and Governance Challenges

Authors: Amzari Mohd Roslan, Fazlina Mohd Ali, Faizan Qamar, Liyana Safra, Nur Arzilawati Md Yunus

PAGES 410 - 423

Paper 38: Multi-Agent Crowd-Risk Prioritization for Large-Scale Events: A Hybrid ML–Agent Approach

Authors: Najwa Samrgandi

PAGES 424 - 435

Paper 39: Model-Augmented Masked Autoencoder with Domain-Adaptive Denoising for Robust Brain Tumor MRI Analysis

Authors: Indrakumar K, Ravikumar M, Mohammed A.S Al-mohamadi, Khalid N. R. Alharbi, Asma A. Alhashmi, Shouki A. Ebad, Abdulbasit A. Darem

PAGES 436 - 453

Paper 40: A Statistical Filtering Algorithm for Tender Price Evaluation Using the J-Selaras Framework

Authors: Mustafa Man, Wan Aezwani Wan Abu Bakar, Mohd Kamir Yusof, Mohd Nasir Saari, Nor Arzalifah Abd. Razak, Nor Zuraiya Mt Amin

PAGES 454 - 461

Paper 41: Prompt Literacy and Generative AI Readiness for AI-Assisted Animation Creative Workflows in Higher Education

Authors: Loh Ngiik Hoon, Natasha Binti Rusdy Wong, Marzie Hatef Jalil

PAGES 462 - 471

Paper 42: CDNet: Complex-Valued Deep Unrolled Composite Denoising Network for Accelerated Magnetic Resonance Image Reconstruction

Authors: David Muigai, Elijah Mwangi, Henry Kiragu

PAGES 472 - 485

Paper 43: Hybrid Random Forest–XGBoost for Smartphone Engagement Prediction with Applications in Marketing Analytics

Authors: Ace John Mark P. Liwanag

PAGES 486 - 495

Paper 44: A Blockchain and AI Integrated Framework for Dynamic Land Registration and Smart Urban Zoning

Authors: Shushma R, S R Nagaraja

PAGES 496 - 502

Paper 45: Segmented Lung CT Image Refinement Through Hybrid Edge-Preserving De-Noising and Variance-Regulated Adaptive Histogram Equalization

Authors: T. Bhuvaneshwari, Rohini Pinapatruni

PAGES 503 - 511

Paper 46: Design and Implementation of Rule-Based and Time-Based Logic in a PAC-Based Precision Irrigation Control System

Authors: Djohar Syamsi, Mimin Muhaemin, Oviyanti Mulyani, Hanif Fakhrurroja, Puput Dani Prasetyo Adi

PAGES 512 - 524

Paper 47: Advancing Ecological Informatics: A Tool for Implementing the Abundance–Preference Approach in Assemblage Analysis

Authors: Mohamed El Farouki, Mohamed Dakki, Meryem Ennakri, Soumia Ziti

PAGES 525 - 537

Paper 48: Blockchain-Based Trusted Data Sharing Framework for Secure and Auditable AI Model Training

Authors: Sherif Amin

PAGES 538 - 545

Paper 49: Shortest Path Algorithms for Smart City Emergency Routing: Scalability and Scenario-Based Analysis

Authors: Aleksandra Stojanova Ilievska, Natasha Stojkovikj, Limonka Koceva Lazarova, Arta Aliu, Krume Tasev, Teona Aleksova

PAGES 546 - 554

Paper 50: Implementation and Evaluation of a Dynamic Multi-Grid Pattern Authentication for Secure Financial Applications

Authors: Amjad Qashlan, Qamar H. Naith, Fatimah Alakeel, Malak B Alharbi

PAGES 555 - 570

Paper 51: Unsanctioned Intelligence: Mitigating the Risks of 'Shadow AI' and Proprietary Data Leaks in the Remote Workforce

Authors: Adel Saad Assiri

PAGES 571 - 578

Paper 52: Dual-Attention Framework for Enhanced Road Obstacle Segmentation and Anomaly Detection

Authors: Priyanka G, Silvia Gaftandzhieva, Mariya Zhekova

PAGES 579 - 586

Paper 53: Anomaly-Score-Augmented Random Forest for Permission-Based Android Malware Detection: An Empirical Study on 29, 999 Apps

Authors: Harikrishnan P R, P. Periyasamy

PAGES 587 - 596

Paper 54: Entanglement-Driven Hybrid Quantum Representation Learning for EEG-Based Epileptic Seizure Classification

Authors: Athigiri Arulalan A S, Senthilkumar G, Jabasheela L, Sangeetha K, Subedha V, Sathiya V

PAGES 597 - 611

Paper 55: EDM-Based Graduate Curriculum Planning Framework

Authors: Levi E. Elipane, Antriman V. Orleans, Adonis P. David, Wilma S. Reyes, Crist John Pastor, Felina P. Espique, Juan Primitivo P. Petrola

PAGES 612 - 620

Paper 56: Arki-Nexus AI: A Framework for Integrating Artificial Intelligence Competencies Into NEUST College of Architecture

Authors: GREGORIO L. VILLAVIZA JR

PAGES 621 - 629

Paper 57: Secure Heart Disease Status Classification Using Machine Learning and Blockchain-Based EHR Integrity Verification

Authors: Haya H. Alsharif, Sabah M. Alzahrani

PAGES 630 - 646

Paper 58: Digital Technologies Enabling Smart Circular Waste Management: A Systematic Literature Review

Authors: Pristi Sukmasetya, Harry Budi Santoso, Zainal Arifin Hasibuan, Yun Arifatul Fatimah

PAGES 647 - 656

Paper 59: Arabic News Text Classification Using Deep Learning Models with Dynamic N-grams

Authors: Ahmed I.Taloba, George Samy Rady, Khaled F. Hussain

PAGES 657 - 672

Paper 60: 8C SPS Model: A Seller Product Selection Model for Home-Based Business Sellers Using a Multi-Criteria Decision-Making Approach

Authors: Selvia Lorena Br Ginting, Zulaiha Ali Othman

PAGES 673 - 690

Paper 61: Hierarchical Workforce Dynamics and Risk-Aware Gated Deep Learning Framework for Employee Attrition Prediction

Authors: Chen Yang

PAGES 691 - 704

Paper 62: Adaptive Grammar Correction with Error-Type Diagnosis and Learner-Centered Feedback Using Proficiency-Aware Transformer Learning

Authors: Xinxin Sun, Peng Guo

PAGES 705 - 719

Paper 63: MACN: A Modality-Aware Cross-Attention Transformer for Non-Invasive Hemoglobin Estimation and Anemia Classification Using Multi-Region Physiological Images

Authors: Chaitra S P, D R Ramesh Babu

PAGES 720 - 727

Paper 64: Low-Latency FPGA-Based PCA Acceleration for Hyperspectral Image Dimensionality Reduction

Authors: Hana Ben Fredj, Ahlem kehili, Amani Chabbeh, Jamel Baili, Chokri Souani

PAGES 728 - 734

Paper 65: Multiscale Image Enhancement and Visualization Analysis for Historical Map/s

Authors: Guilei Ren

PAGES 735 - 746

Paper 66: Statistical Security Analysis NIST and Security Enhancement of CBT Cipher for Secure Data and Image Transmission

Authors: Walid W. Souror, Mohamed Fouad, Ali E. Takieldeen

PAGES 747 - 756

Paper 67: Unsupervised Hybrid Approach for Intrusion Detection Based on Information System Log Files

Authors: Sena Apeke, Nadjime Pindra, Sèmèvo A. R. M. Ahouandjinou

PAGES 757 - 765

Paper 68: Artificial Intelligence Guided Crowd Counting and Density Estimation with Enhanced CSRNet

Authors: Muhammad Jawad Babar, Malik Muhammad Saad Missen,
Hannan Adeel, Muhammad Usman, Muzammil Malik

PAGES 766 - 779

Paper 69: FSCM-Net: Frequency-Spatial Collaborative Modeling for Traffic Scene Semantic Segmentation

Authors: Wei Zhao, Yi Dong, Lingchao Wang, Qiang Ai

PAGES 780 - 793

Paper 70: LAURA: An Awareness-Oriented Explainable AI Framework for Academic Stress in Higher Education

Authors: Ahmed Almathami, Richard Stone

PAGES 794 - 804

Paper 71: A Model-Free AI Framework for Robust PQRST Complex Detection in ECG Signals

Authors: Arthorn Luangsodsai, Manh Duong Dang, Krung Sinapiromsaran

PAGES 805 - 814

Paper 72: Privacy-Preserving Feature Engineering Framework for Real-World Vulnerability Risk Prediction

Authors: Jawharah Albarakati, Tawfiq Hasanin, Suaad Alarif

PAGES 815 - 826

Paper 73: From Method-Centric Optimisation to Failure-Oriented Robustness: A Framework for Industrial Inspection Systems

Authors: Ying Zhou, Mei Choo Ang, Ummul Hanan Mohamad, Ah-Lian Kor, Kok Weng Ng, Saraswathy Shamini Gunasekaran, Halimah Badioze Zaman

PAGES 827 - 840

Paper 74: SpaTempNet: Deep Spatiotemporal Forecasting of River Morphological Evolution of the Padma River in Bangladesh

Authors: S. S. Mahmud Turza, Md. Kaoser Ahamed Anik, Md. Shadmim Hasan Sifat, Mahruba Sharmin Chowdhury, Khandokar Md. Rahat Hossain, Abdullah Al Noman

PAGES 841 - 856

Paper 75: Hardware-Software Co-Design of K-Means Clustering on Zynq FPGAs via OpenCL

Authors: Eya JABALLI, Soufien GDAIM, Noureddine LIOUANE

PAGES 857 - 867

Paper 76: Challenges of AI Adoption in the Saudi Arabian Healthcare Sector: A Comprehensive Study

Authors: Radwan M. Batyha, Ahmed Almaghthawi, Fahad Alharbi, Suhaila Abuowaida

PAGES 868 - 878

Paper 77: Multimodal Pneumoconiosis Screening on a Real-World Vietnamese Occupational Health Dataset with Incomplete Metadata and Class Imbalance

Authors: Luong Thi Bich Phuong, Nguyen Hoang Anh, Au Xuan Manh, Truong Thanh Nam, Tran Tien Cong, Nguyen Trong Khanh

PAGES 879 - 887

Paper 78: BEAT: An Integrated Blockchain-Edge AI Trust Framework for IoT Security - Design, Prototype, and Performance Evaluation

Authors: Pavansai Ramarao Maddali, Vijay Kumar Damera, Ratna Kumar Prathipati

PAGES 888 - 899

Paper 79: A Hybrid Autoencoder-Random Forest Framework for Intrusion Detection in Internet of Medical Things Network

Authors: Sophia ALAMI-KAMOURI, Ridouan LACHGAR, Mohamed AFIF

PAGES 900 - 907

Paper 80: Development of an Automated Network Forensics Analysis Tool Using PCAP Data

Authors: Ashwag Alotaibi, Huda Aldawghan, Abdullah Albuali

PAGES 908 - 916

Paper 81: A Probabilistic Instance Generator for Multimodal Urban Transportation Networks Addressing Data Scarcity in Metropolitan Transport Systems Integrating the Formal and Informal Sectors

Authors: Oumar KONE, Kouassi Hilaire Edi, Yapi Fiacre Aristide Edi, Pawoumodom Matthias Takouda

PAGES 917 - 927

Paper 82: On Practical Considerations for the Adoption of Worst-Case Optimal Joins in RDBMSs

Authors: Ayoub Berdai, Kawtar Younsi Dahbi, Dalila Chiadmi, Slimane Bah

PAGES 928 - 942

Paper 83: A Behavioral Analysis of Machine Learning for Intrinsic 3D Tree Branching Structures

Authors: M. Charitha Lakshan, K. Damitha Sandaruwan, L. N. C. De Silva

PAGES 943 - 957

Paper 84: A Genetic Algorithm-Driven Framework for Convolutional Kernel Selection in Lightweight Plant Disease Classifiers

Authors: El houssaine HSSAYNI

PAGES 958 - 965

Paper 85: A Binary Cascade Framework for Gut Microbiome Classification in Inflammatory Bowel Disease: Few-Shot LLMs, Co-Abundance Graphs, and Clinical Narrative Encoding

Authors: Nouhaila En Najih, Ahmed Moussa

PAGES 966 - 974

Paper 86: Learning When to Wait: A Lightweight Binary Decision Framework for Container Migration in Edge Computing

Authors: Yasmeen Abu-Haltem, Sawsan AlShattnawi

PAGES 975 - 984

Paper 87: A Geospatially-Aware Hybrid Deep Learning Framework for Violence Detection and Sentiment Analysis in Airline Reviews

Authors: Fahad Alanazi, Osama Rabie, Wafi Bedewi

PAGES 985 - 996

Paper 88: Regulatory-Compliant EEG Stress Recognition Using CNN–Transformer Models with Feature Unlearning

Authors: Yusra, Riaz UlAmin

PAGES 997 - 1012

Paper 89: Hybrid State Estimation and Recurrent Neural Networks for Supply Chain Disruption Detection

Authors: Mohamed Amine Frikha, Mariem Mrad, Younes Boujelbene, Yasser Mohamed Ayad Soliman

PAGES 1013 - 1019

Paper 90: Predicting Temporal Exploitation of Software Vulnerabilities Using Knowledge-Fused Neural Learning

Authors: Mounir Belmahjoub, Lamia Benhiba

PAGES 1020 - 1029

Paper 91: Deep Learning-Based Image Recognition and Brand Identity Design for Intangible Cultural Heritage in Southern Sichuan

Authors: Zhihong Chen

PAGES 1030 - 1039

Vision Transformers and Convolutional Neural Networks for Land Use Scene Classification

Arun D. Kulkarni
Computer Science Department, University of Texas at Tyler, Tyler, TX 75799, USA

***Abstract*—Land use scene classification (LUSC) from remote sensing imagery plays a critical role in environmental monitoring, urban planning, and sustainable resource management. In recent years, deep learning methods have significantly advanced the state-of-the-art, with Convolutional Neural Networks (CNNs) dominating the field because of their strong ability to capture local spatial features. However, the emergence of Vision Transformers (ViTs) has introduced a new paradigm that models long-range dependencies through self-attention mechanisms, potentially enabling improved global context understanding. This study presents a comparative assessment of Vision Transformers and CNN-based architectures for remote sensing land use scene classification. Representative CNN models, such as AlexNet, are evaluated alongside the Vision Transformer (ViT) using benchmark remote sensing datasets, including the UC Merced (UCM) Land Use and EuroSAT Land Use datasets. The study examines classification accuracy, precision, recall, F1-score, and computational complexity to provide a comprehensive performance comparison. Experimental results demonstrate that CNNs perform robustly on datasets with limited training samples and strong local texture characteristics, whereas Vision Transformers exhibit superior performance in capturing global spatial relationships in complex scenes when sufficient training data are available. However, ViTs typically require greater computational resources and larger training datasets to achieve optimal performance. The findings of this study provide insights into the strengths and limitations of both architectures and offer guidance for selecting appropriate models for remote sensing land use scene classification applications.**

Keywords*—Convolutional Neural Networks (CNNs); deep learning; land use scene classification; remote sensing; Vision Transformers (ViTs)

I. Introduction

Remote sensing imagery has emerged as a vital data source for analyzing, monitoring, and managing the Earth's surface. Recent advances in satellite and aerial imaging technologies have enabled the acquisition of large volumes of high-resolution imagery, providing unprecedented opportunities for large-scale analysis of Land Use and Land Cover patterns [1]. Among the many tasks in remote sensing image analysis, land use scene classification (LUSC) has emerged as a fundamental problem [2]. LUSC seeks to classify aerial or satellite image patches by assigning semantic labels such as residential, agricultural, industrial, forest, and water bodies. Precise land use scene classification plays a vital role in numerous applications, such as urban planning, environmental monitoring, disaster management, resource allocation, and climate change evaluation.

Early approaches to LUSC primarily used conventional machine learning methods that depended on manually engineered feature extraction followed by statistical classification approaches. These methods typically utilize descriptors such as texture-based features, spectral indices, and shape-based representations. Since local descriptors are usually derived from multiple spatial locations, feature aggregation techniques are needed to generate fixed-length scene representations that can be used for classification tasks. While these methods laid an important foundation for remote sensing image analysis, they also have several intrinsic limitations. In particular, handcrafted features tend to be sensitive to changes in spatial resolution, seasonal variations, lighting conditions, and geographic diversity. Moreover, manually designed descriptors may fail to capture complex semantic relationships among multiple objects and spatial arrangements within a scene [3]. In addition, feature extraction and classifier design are generally treated as separate stages, preventing end-to-end optimization [4]. As a result, traditional methods frequently show poor generalization when used in complex and diverse remote sensing environments.

The rapid expansion of large-scale remote sensing datasets, along with advances in high-performance computing, has driven the increased adoption of deep learning techniques for LUSC. In contrast to traditional machine learning methods, deep neural networks can automatically learn hierarchical feature representations directly from raw image data. This end-to-end representation learning approach has substantially enhanced classification accuracy in remote sensing scenes marked by high intra-class variability, strong inter-class similarity, and considerable spatial heterogeneity.

Among deep learning approaches, Convolutional Neural Networks (CNNs) have emerged as the leading framework for remote sensing image classification. Representative architectures such as AlexNet, VGGNet, ResNet, and EfficientNet have demonstrated strong performance across a variety of benchmark datasets [5]. The success of CNNs is largely attributed to their local receptive fields, parameter sharing, and hierarchical feature extraction mechanisms, which enable efficient learning of edges, textures, shapes, and other discriminative spatial patterns. Moreover, the built-in inductive biases of CNNs—especially locality and translation invariance—tend to encourage more stable learning and improved generalization, particularly in settings with limited training data. Nevertheless, because convolutional operations primarily focus on local neighborhoods, CNNs may be less effective in modeling long-range spatial dependencies and

global contextual relationships that frequently characterize complex land use scenes [6].

Recently, Vision Transformers (ViTs) have gained attention as a strong alternative to CNN-based architectures in computer vision. Drawing inspiration from the success of Transformer models in natural language processing, ViTs treat an image as a sequence of non-overlapping patches, which are then processed as tokens. By leveraging self-attention mechanisms, the model directly learns global interactions across different image regions, improving its ability to capture long-range dependencies and contextual relationships. This capacity for global representation is especially valuable in remote sensing imagery, where scene interpretation often relies on spatial relationships among multiple objects spread across wide areas. Despite these advantages, Vision Transformers generally require larger training datasets, higher computational resources, and careful optimization strategies to achieve competitive performance. The differing characteristics of CNNs and Vision Transformers motivate a systematic comparison of their effectiveness in land-use scene classification. CNNs are known for their computational efficiency, strong inductive biases, and reliable performance in data-limited scenarios, whereas Vision Transformers introduce a distinct representation approach that prioritizes global contextual modeling. A rigorous comparative analysis of these architectures is therefore important for understanding their suitability across diverse remote sensing scenarios.

In this study, we present a comprehensive comparative study of CNN- and ViT-based architectures for land use scene classification using benchmark remote sensing datasets. The study evaluates their classification performance, computational efficiency, and generalization capability under consistent experimental settings. The main objective is to provide practical insights into the strengths and limitations of both model families and to identify conditions under which each architecture is most suitable for remote sensing land use analysis.

The remainder of this study is organized as follows: Section II reviews related work on land use scene classification. Section III presents the proposed comparative framework and methodology. Section IV describes the datasets, experimental setup, and results. Section V discusses the results, presents the conclusions, and outlines future research directions.

II. Related Work

Land Use and Land Cover (LULC) scene classification has emerged as a key area of research in remote sensing because of its extensive applications in urban development, environmental surveillance, disaster response, and agricultural evaluation [7, 8]. The growing accessibility of high-resolution satellite imagery has further increased the complexity of scene analysis, creating a demand for more sophisticated computational approaches. Over the last two decades, research in this field has evolved from conventional handcrafted feature extraction techniques to deep learning-based approaches, especially Convolutional Neural Networks (CNNs), and more recently, Vision Transformers (ViTs). This section provides a comprehensive review of these developments and highlights current trends and challenges.

Before the widespread adoption of deep learning techniques, Land Use and Land Cover (LULC) classification primarily relied on handcrafted feature extraction methods integrated with traditional machine learning algorithms. In these conventional approaches, researchers manually engineered descriptive features to capture the spatial, spectral, textural, and structural properties of remote sensing imagery. The effectiveness of the classification process largely depended on the quality and discriminative capability of these handcrafted representations. Among the most influential and extensively utilized feature descriptors were Scale-Invariant Feature Transform (SIFT) [9], Histogram of Oriented Gradients (HOG) [10], and Local Binary Patterns (LBP) [11]. These techniques significantly advanced remote sensing image analysis by enabling the extraction of meaningful visual patterns from satellite and aerial imagery. They were widely employed in combination with classical classifiers such as Support Vector Machines (SVM), Random Forests (RF), k-Nearest Neighbors (k-NN), and Artificial Neural Networks (ANNs) to categorize different land cover and land use classes. SIFT gained considerable attention due to its robustness in detecting distinctive key points that remain invariant to scale, rotation, and partial illumination changes. This characteristic made it highly effective for identifying objects and spatial structures in high-resolution remote sensing images, particularly in heterogeneous environments. HOG descriptors emphasized the distribution of gradient orientations within localized regions, thereby providing strong representations of edge, contour, and shape information. Such capabilities proved valuable for distinguishing complex landscape structures, including urban areas, roads, and vegetation boundaries. Similarly, Local Binary Patterns (LBP) emerged as a powerful texture descriptor because of its simplicity, computational efficiency, and robustness to illumination variations. By encoding local neighborhood intensity relationships, LBP effectively characterized surface textures and fine-grained spatial patterns commonly observed in remote sensing imagery. Consequently, it became widely used in applications involving vegetation analysis, urban mapping, and terrain classification. Collectively, these handcrafted feature extraction techniques established the foundation of traditional LULC classification systems. Although they achieved notable success in many remote sensing applications, their performance was often constrained by limited feature generalization capabilities and the substantial reliance on expert knowledge for feature engineering. These limitations ultimately motivated the transition toward deep learning-based methods, which automatically learn hierarchical and highly discriminative feature representations directly from raw imagery data. Although these methods achieved moderate success, they suffered from several limitations. Most notably, handcrafted features lacked the ability to capture high-level semantic information and were sensitive to variations in scale, illumination, and viewpoint. Furthermore, the feature extraction process required domain expertise and was not adaptable to diverse datasets. These shortcomings motivated the transition toward data-driven feature learning approaches.

The advent of deep learning, particularly Convolutional Neural Networks (CNNs), brought a transformative shift to the field of image classification. Unlike traditional machine learning approaches that rely heavily on handcrafted feature extraction techniques, CNNs are capable of automatically learning hierarchical and discriminative feature representations directly from raw image data. This ability to perform end-to-end learning significantly reduces the dependence on manual feature engineering while improving the robustness and scalability of classification systems [4]. In the domain of remote sensing, the adoption of CNN-based architecture has led to remarkable improvements in scene classification accuracy and efficiency. Early research demonstrated that deep learning models substantially outperform conventional methods by effectively capturing complex spatial patterns, textures, and semantic information present in high-resolution remote sensing imagery [12]. Consequently, CNNs have become a foundational approach for modern remote sensing image analysis and classification tasks. Pre- trained architectures such as AlexNet [13], VGGNet [14], and GoogLeNet [15] were among the first to be adapted for remote sensing applications through transfer learning.

A major strength of CNNs is their ability to learn and capture local spatial patterns through convolutional filters. This capability makes them highly effective for detecting textures, edges, and object-level features in remote sensing imagery [16]. To further improve CNN-based models, researchers have introduced several advancements, including: 1) multi-scale feature extraction techniques to accommodate objects of varying sizes in satellite images [17], and 2) data augmentation and transfer learning strategies to mitigate the challenges posed by limited labeled datasets [18]. Despite these advancements, CNNs have inherent limitations. Their dependence on local receptive fields limits their ability to capture long-range dependencies and global context [6]. Although deeper architectures can partly address this issue, they increase computational complexity and the risk of overfitting, while performance gains often plateau with greater depth. These challenges highlight the need for alternative architectures.

Natural language processing-based transformer architectures have significantly influenced the development of Vision Transformer (ViT) models, which have emerged as a powerful alternative to traditional convolutional neural networks (CNNs). Originally inspired by the transformer framework introduced by Vaswani et al [19]. ViTs depart from the convolution-centric paradigm by treating images as sequences of smaller patches, analogous to tokens in language models. These patches are processed through self-attention mechanisms that enable the model to learn long-range and global dependencies across the entire image more effectively than conventional CNN architectures. The pioneering study by Dosovitskiy et al. [6] demonstrated that ViTs are capable of achieving state-of-the-art performance in image classification tasks, particularly when trained on large-scale datasets. Their findings highlighted the scalability and representational power of transformer-based vision models, establishing ViTs as a transformative approach in computer vision research. In the domain of remote sensing, ViTs have shown considerable potential for handling the complex spatial structures and contextual variations commonly present in satellite imagery. Their ability to model global contextual relationships makes them especially suitable for Land Use and Land Cover Classification (LULC) tasks, where understanding both local features and broader spatial patterns is essential for accurate categorization. Recent studies, such as those by Zhao et al. [8], have demonstrated that transformer-based approaches can effectively capture intricate spatial dependencies and improve classification performance in remote sensing applications. The global attention mechanism in ViTs enables them to model long-range interactions more effectively than CNNs. This is particularly beneficial for complex scenes where spatial relationships between distant regions are important. However, ViTs also present challenges, including high computational requirements, large memory consumption, and dependence on extensive training data [20]. To harness the complementary advantages of CNNs and transformers, recent studies have increasingly explored hybrid architectures that integrate convolutional operations with self-attention mechanisms. These approaches seek to combine the strong local feature extraction capabilities of CNNs with the powerful global context modeling offered by transformers. Research indicates that hybrid models frequently outperform standalone CNNs and Vision Transformers (ViTs) in Land Use and Land Cover (LULC) classification tasks [21]. They demonstrate greater robustness, improved generalization, and a superior ability to capture multi-scale contextual information. Such models are especially well-suited for analyzing high-resolution and hyperspectral remote sensing datasets. The progress of Land Use and Land Cover (LULC) classification techniques has been strongly driven by the availability of publicly accessible benchmark datasets that enable standardized model training and fair performance comparison across studies. Datasets in this domain include the UCM Land Use Dataset and the EuroSAT Dataset.

The UCM dataset provides high-resolution aerial imagery with a limited number of well-balanced classes, whereas EuroSAT, derived from Sentinel-2 satellite imagery, further extends this diversity by incorporating multispectral information. These datasets form a comprehensive benchmark suite that enables robust evaluation of model generalization, feature representation capability, and cross-dataset transferability [17, 22]. Convolutional Neural Network (CNN)-based architectures have historically demonstrated strong performance across these benchmarks due to their effectiveness in learning hierarchical spatial features from imagery. However, recent advances in deep learning have shown that transformer-based models, such as Vision Transformers (ViTs), along with hybrid CNN–Transformer architectures, are achieving either superior or competitive performance. These models are particularly effective in capturing long-range spatial dependencies and global contextual relationships, which are often challenging for traditional CNNs. As a result, there is a growing shift toward attention-based and hybrid frameworks in LULC classification, especially for complex and large-scale remote sensing datasets. Duan et al. [23] proposed the Swin Transformer with Multi-Scale Fusion (STMSF) to effectively represent ground objects of varying scales in remote sensing scenes through the integration of multi-scale features. Experimental results show that the proposed network surpasses

several state-of-the-art CNN-based and transformer-based approaches.

Recent studies in remote sensing and Land Use/Land Cover (LULC) classification demonstrate a clear and growing transition toward transformer-based architectures and hybrid deep learning models. These approaches have shown strong potential in capturing long-range spatial dependencies and improving classification performance compared to traditional convolutional methods. Despite these advancements, several important challenges persist. First, data scarcity remains a major limitation, as high-quality, accurately labeled remote sensing datasets are often limited in size and diversity, constraining the training of data-intensive models. Second, computational complexity is a significant concern, particularly for transformer-based models, which typically require substantial memory and processing power for both training and inference. Third, model interpretability continues to be an open issue, as the internal decision-making processes of deep learning models—especially transformers—are often difficult to interpret and explain in a physically meaningful way. Finally, generalization across datasets remains challenging, since models trained on a specific dataset often experience performance degradation when applied to different geographic regions, sensor types, or imaging conditions.

In this context, the present study focuses on a comparative analysis of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) for land use scene classification (LUSC). The evaluation is conducted using two widely recognized benchmark datasets: the UCM Land Use Dataset and the EuroSAT Dataset. This comparative investigation aims to assess the relative strengths and limitations of CNN-based and transformer-based approaches in terms of classification accuracy, robustness, and generalization capability across diverse remote sensing scenarios.

III. Proposed Framework

This section presents a comparative framework for evaluating Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) for land use scene classification (LUSC) using high-resolution remote sensing imagery. To ensure a controlled and fair comparison, both the AlexNet and Vision Transformer (ViT) models are trained using identical datasets and data partitions, input resolutions, augmentation strategies, optimization settings, and evaluation metrics. Therefore, the observed performance differences can be attributed primarily to architectural characteristics rather than experimental inconsistencies. The proposed framework consists of four stages: 1) dataset preparation and preprocessing, 2) AlexNet optimization and training, 3) Vision Transformer (ViT) optimization and training, and 4) classification of test data.

A. Data Preparation and Preprocessing

Let the land use dataset be represented as:

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N} \tag{1}$$

where, $x_i \in \mathbb{R}^{H\times W\times C}$, denotes an input image, $y_i \in \{1,2,\dots,K\}$ is the class label, N is the total number of samples, and K is the number of land use categories. Remote sensing images exhibit substantial intra-class variability arising from seasonal effects, spatial resolution changes, illumination differences, and scene complexity. To reduce statistical variability across samples, each image is normalized using channel-wise standardization:

$$x_i' = \frac{x_i - \mu}{\sigma} \tag{2}$$

where, μ and σ denote channel mean and standard deviation. To improve generalization and reduce overfitting, stochastic augmentation is applied during training:

$$\tilde{x}_i = T(x_i') \tag{3}$$

where, $T(\cdot)$ denotes a composition of random transformations including horizontal flipping, random rotation, random cropping, and mild color perturbation. The dataset is partitioned into mutually exclusive subsets:

$$\mathcal{D} = \mathcal{D}_{train} \cup \mathcal{D}_{val} \cup \mathcal{D}_{test} \tag{4}$$

With:

$$\mathcal{D}_{train} \cap \mathcal{D}_{val} = \emptyset, \mathcal{D}_{train} \cap \mathcal{D}_{test} = \emptyset \tag{5}$$

B. AlexNet for Land Use Classification

CNNs exploit local spatial correlations through learnable convolutional kernels. In this study, AlexNet was implemented using the MATLAB Deep Learning Toolbox. The architecture of AlexNet is shown in Fig. 1. The model consists of eight layers, including convolutional layers, max-pooling layers, fully connected layers, and a SoftMax output layer. The functions of the AlexNet layers are explained using the equations presented below [13, 23].

Given the input feature map $X^{(l-1)}$, the convolutional transformation at layer l is defined as:

$$X_j^{(l)} = \phi\left(\sum_{i=1}^{M} X_i^{(l-1)} * W_{ij}^{(l)} + b_j^{(l)}\right) \tag{6}$$

where, $W_{ij}^{(l)}$ denotes learnable convolution kernels, $b_j^{(l)}$ is the bias, and $\phi(\cdot)$ is the nonlinear activation function.

Pooling performs spatial downsampling:

$$P^{(l)} = \text{Pool}\left(X^{(l)}\right) \tag{7}$$

The hierarchical nature of convolution allows the network to learn low-level edge patterns in early layers and increasingly abstract semantic structures in deeper layers. After feature extraction, the resulting representation is flattened and passed through fully connected layers. The final class probability is obtained using SoftMax:

$$\hat{y}_k = \frac{\exp(z_k)}{\sum_{j=1}^{K} \exp(z_j)} \tag{8}$$

where, z_k denotes the class logit. The SoftMax classifier converts logits into normalized probability distributions for multi-class prediction tasks [24]. CNNs are expected to perform strongly when land use categories are characterized by repetitive textures, locally correlated spatial patterns, and compact object arrangements. Examples include agricultural land, forest regions, and dense residential structures [25].

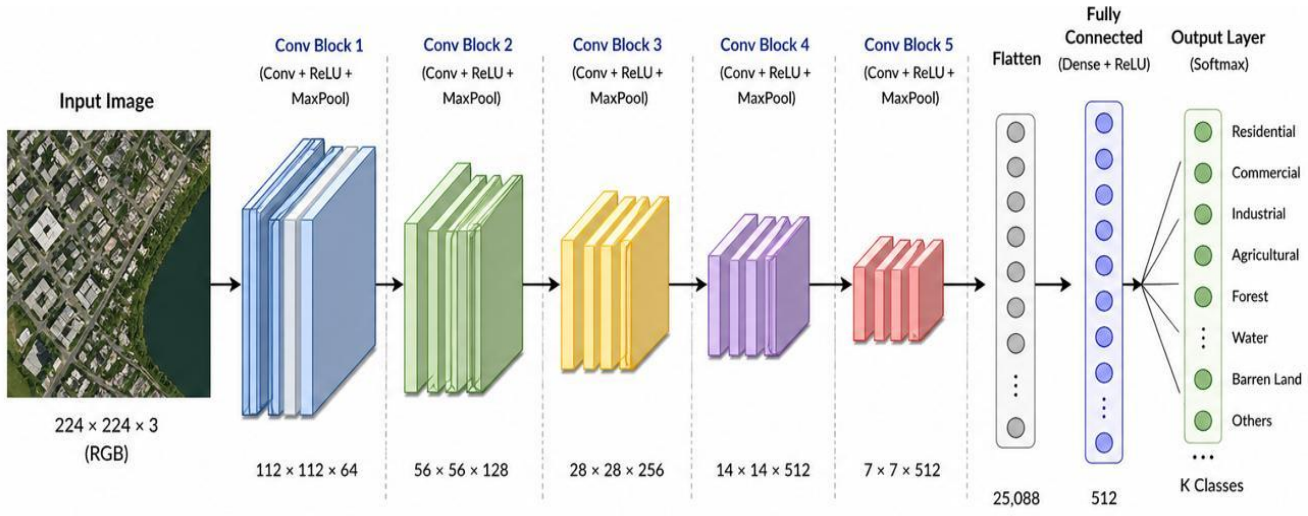


Fig. 1. AlexNet architecture for land use classification.

C. Vision Transformer for Land Use Classification

A Vision Transformer (ViT) is a neural network architecture for computer vision that applies the Transformer model, originally developed for natural language processing, to image analysis. It operates by dividing an image into small patches, converting each patch into an embedding vector, adding positional information, and processing the resulting sequence of embeddings through self-attention layers to learn relationships among different regions of the image. Unlike traditional Convolutional Neural Networks (CNNs), which primarily focus on local feature extraction using convolutional filters, ViTs can capture long-range global dependencies more effectively. This capability makes them highly effective for tasks such as image classification, object detection, and image segmentation, particularly when trained on large-scale datasets. A simplified architecture of the Vision Transformer is shown in Fig. 2. The ViT encoding mechanism is explained below with the help of the corresponding equations [6, 19, 26].

1) Patch embedding

Given an input image:

$$x \in \mathbb{R}^{H \times W \times C} \tag{9}$$

It is divided into non-overlapping patches of size P×P. The number of patches is:

$$N_p = \frac{HW}{P^2} \tag{10}$$

Each patch is flattened and projected into a D-dimensional embedding space:

$$z_0^p = x_p E \tag{11}$$

where, $E \in \mathbb{R}^{(P^2C) \times D}$ is a learnable projection matrix.

A learnable class token is appended and positional embeddings are added:

$$Z_0 = \left[x_{cls}; z_0^1; z_0^2; \ldots; z_0^{N_p}\right] + E_{pos} \tag{12}$$

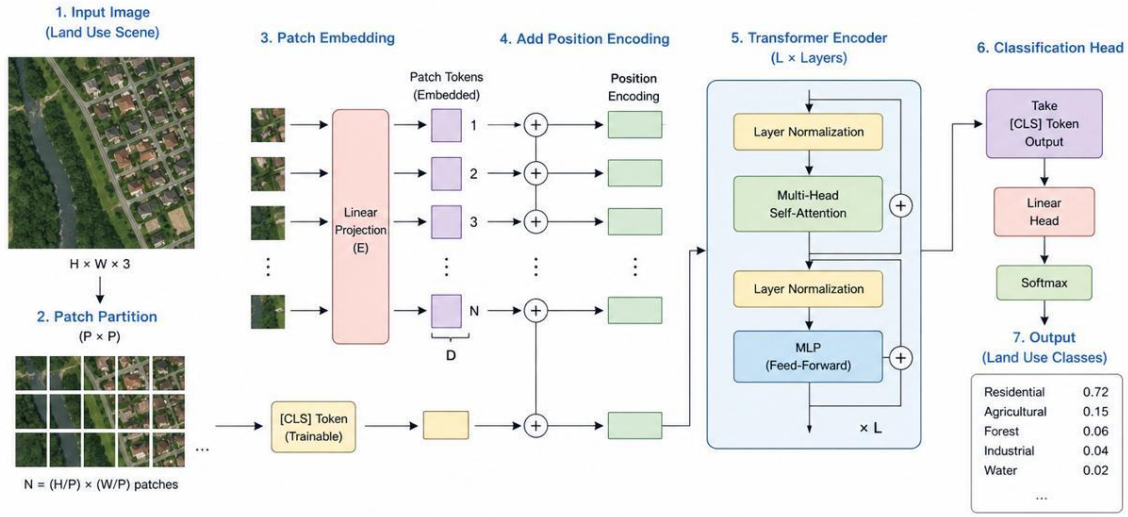


Fig. 2. Vision transformer architecture for land use scene classification.

2) Multiheaded attention: For an input representation Z, query, key, and value matrices are defined as:

$$Q = ZW_Q, \qquad K = ZW_K, \qquad V = ZW_V \tag{13}$$

Self-attention is computed as:

$$\text{Attention}(\text{Q}, \text{K}, \text{V}) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{14}$$

Multi-head attention is expressed as:

$$\text{MSA}(Z) = \text{Concat}(head_1, \dots, head_h)W_O \tag{15}$$

where, *h* denotes the number of attention heads.

3) Transformer encoding: Each encoder layer contains residual self-attention and feed-forward blocks:

$$Z_l' = Z_{l-1} + \text{MSA}\big(\text{LN}(Z_{l-1})\big) \tag{16}$$

$$Z_l = Z_l' + \text{MLP}\big(\text{LN}(Z_l')\big) \tag{17}$$

The final class token is used for classification:

$$\hat{y} = \text{Softmax}(W_c Z_L^{cls} + b_c) \tag{18}$$

ViTs are particularly suitable when scene understanding depends on global spatial organization and long-range contextual relationships. Examples include airports, harbors, commercial zones, and transportation hubs.

D. Optimization Strategy

Both models are trained under identical optimization settings. The training objective is categorical cross-entropy.

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}\log(\hat{y}_{ik}) \tag{19}$$

Parameter updates are performed by gradient-based optimization:

$$\theta_{t+1} = \theta_t - \eta\nabla_\theta\mathcal{L} \tag{20}$$

where, θ denotes trainable parameters, and η is the learning rate [27].

To improve generalization, the framework incorporates early stopping, weight decay, dropout regularization, and learning-rate scheduling.

E. Evaluation Protocol

Performance is evaluated using multiple complementary metrics [24]:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{21}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{22}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{23}$$

$$F_1 = \frac{2PR}{P + R} \tag{24}$$

Confusion matrices are used to analyze class-level error patterns and identify categories that benefit from local or global feature modeling. The proposed approach establishes a rigorous experimental framework for comparing convolutional and transformer-based models for land use scene classification. By enforcing identical preprocessing, training, and evaluation conditions, the framework enables an interpretable assessment of architectural behavior. Section IV describes experimental setup and results.

IV. Implementation and Results

This section describes the datasets, implementation details, and protocols used to assess the effectiveness of the AlexNet and Vision Transformer framework for land use scene classification (LUSC).

A. Datasets

To evaluate the robustness and generalization capability of the CNN and ViT models, experiments were conducted on two widely used benchmark remote sensing datasets: the UCM Land Use Dataset and the EuroSAT Dataset. For both models and across all datasets, 80 per cent of the randomly selected samples from each class were used for training, 10 per cent were used for validation, and the remaining 10 per cent were used for testing.

The UCM Land Use Dataset is a widely used benchmark dataset in the field of remote sensing and aerial image classification. Developed by researchers at the University of California, Merced [28] [30]. The dataset contains 2,100 high-resolution aerial images divided into 21 land-use categories, including agricultural, forest, harbor, residential, runway, and parking lot scenes. Each category contains 100 images with image dimensions of 256 × 256 pixels and a spatial resolution of 0.3 meters (approximately one foot per pixel). The dataset is commonly used to evaluate machine learning and deep learning models for scene classification, feature extraction, and pattern recognition in geospatial imagery. Owing to its balanced class distribution and diverse visual content, the dataset has become a standard benchmark for comparing the performance of convolutional neural networks and other computer vision approaches in remote sensing research. For this study, two subsets were created from the UCM Land Use Dataset. The first subset consists of five distinct classes, while the second subset contains ten classes, including the original five classes and five additional classes. Images from both subsets were classified using the AlexNet and Vision Transformer models, implemented using MATLAB scripts. Forty sample images from the UCM Land Use 10 class Data set are shown in Fig. 3.

EuroSAT Dataset is a widely used benchmark dataset for land-use and land-cover classification in remote sensing applications [29] [31]. The dataset is based on satellite images acquired from the Sentinel-2 mission of the European Space Agency. It contains 27,000 labeled and geo-referenced images categorized into 10 different land-use and land-cover classes, including annual crops, forest, herbaceous vegetation, highway, industrial, pasture, permanent crop, residential, river, and sea/lake scenes. Each image has dimensions of 64 × 64 pixels and covers a spatial resolution ranging from 10 to 60

meters per pixel, depending on the spectral band used. The dataset is extensively used for evaluating machine learning and deep learning algorithms in remote sensing image classification, scene understanding, and environmental monitoring tasks. Due to its large number of samples, balanced class distribution, and multispectral image characteristics, the EuroSAT Dataset has become a standard benchmark for comparing the performance of convolutional neural networks, transformer-based architectures, and other computer vision models in satellite image analysis. In this study, we created two sets of data. The first dataset consists of five distinct classes, and the second dataset consists of all ten classes from the EuroSAT Dataset. Sample images from the EuroSAT Dataset with ten classes are shown in Fig. 4.

Fig. 3. Sample images from the UCM dataset.

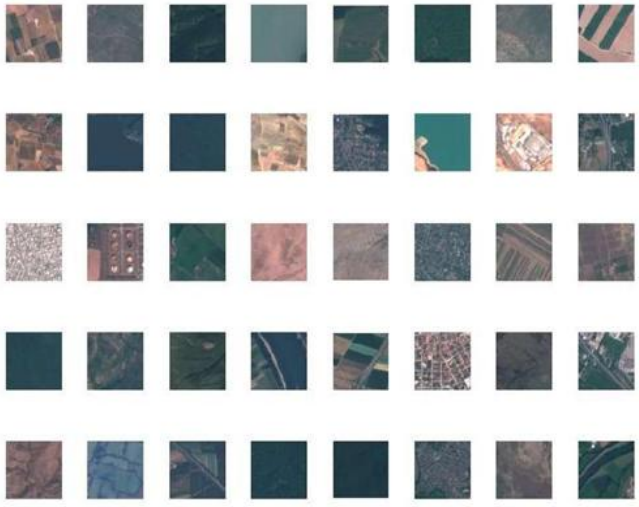


Fig. 4. Sample Images from the EuroSAT Dataset.

All datasets were classified using the AlexNet and Vision Transformer models implemented in MATLAB on a PC equipped with a single AMD Ryzen 9 7950X x64 4.5 GHz processor and 128 GB RAM. Evaluation metrics, including confusion matrices, accuracy, precision, recall, F-score, and computational time, are presented in this section.

B. Network Configuration

We developed two AlexNet models using MATLAB scripts. Each model consists of eight learnable layers, including five convolutional layers and three fully connected layers. The input image size was 227 × 227 × 3 (RGB). The first model was designed for five output classes, while the second model was designed for ten output classes. The layer-by-layer details are presented in Table I. In addition, we developed two Vision Transformer B/16 models using MATLAB scripts. The input image size was 224 × 224 × 3 (RGB), with a patch size of 16 × 16 pixels, resulting in 196 patches. The first ViT model was designed for five classes, while the second model was designed for ten classes. The feature details of the Vision Transformer are presented in Table II. The experimental setup ensures a fair, rigorous, and reproducible evaluation of the CNN and Vision Transformer models across multiple benchmark datasets. By incorporating strong baselines, evaluation metrics are described in Section III. The framework provides a comprehensive assessment of model performance.

TABLE I. ALEXNET LAYER-BY-LAYER DETAILS

Layer	Type	Filters / Neurons	Kernel	Stride	Output Size
Input	Image	-	—	—	227x227x3
Conv1	Convolution	96 filters	11x11	4	55x55x96
Pooll	Max Pool	-	3x3	2	27x27x96
Conv2	Convolution	256 filters	5x5	1	27x27x256
Pool2	Max Pool	-	3x3	2	13x13x256
Conv3	Convolution	384 filters	3x3	1	13x13x384
Conv4	Convolution	384 filters	3x3	1	13x13x384
Conv5	Convolution	256 filters	3x3	1	13x13x256
Pool5	Max Pool	-	3x3	2	6x6x256
FC6	Fully Connected	4096	—	—	4096
FC7	Fully Connected	4096	—	—	4096
FC8	Fully Connected	1000	-	—	1000

C. Experimental Results

We implemented two AlexNet models and two Vision Transformer (ViT) models to classify UCM and EuroSAT Datasets with five and ten classes, respectively. The results are presented below.

TABLE II. VISION TRANSFORMER-B/16 FEATURES

Feature	Details
Name	ViT-B/16
Patch size	16x16 pixels
Input image	224x224x3 (standard)
Embedding dimension	768
Number of transformer blocks	12
Number of attention heads	12
MLP (Feed-forward) hidden size	3072
Total parameters	~86 million
Classification token	Yes ([CLS])
Positional encoding	Learnable

1) Results for UCM dataset: The learning progress graph for UCM dataset for AlexNet and Vision Transformer are shown in Fig. 5 and Fig. 6, respectively.

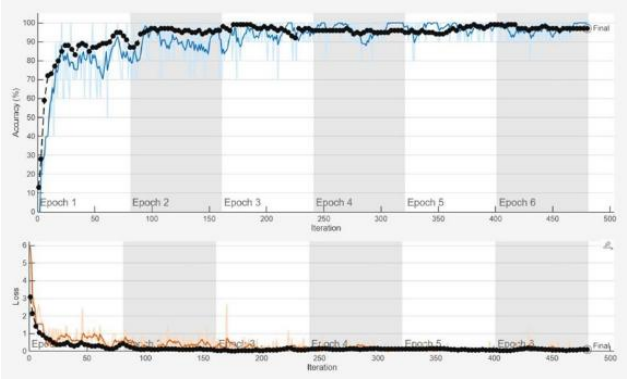


Fig. 5. Learning progress for AlexNet 10-class on the UCM dataset.

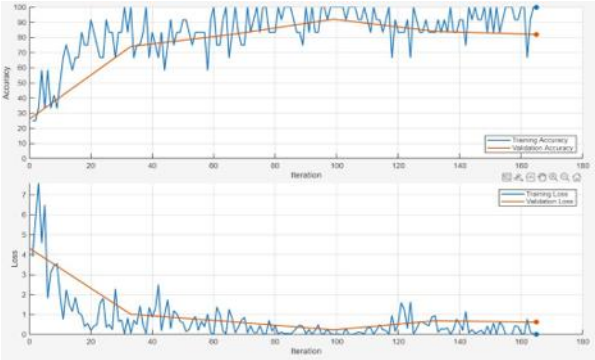


Fig. 6. Learning progress for ViT 10-class on the UCM dataset.

The comparison of the average values of the metrics—accuracy, precision, recall, and F-score—obtained from the confusion matrices for the UCM Dataset using AlexNet and Vision Transformer is shown in Fig. 7 and Fig. 8, respectively.

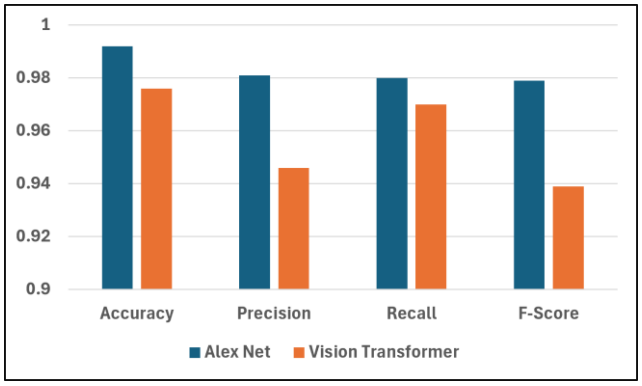


Fig. 7. Comparative performance for 5-class UCM dataset.

TABLE III. RESULTS FOR 5-CLASS UCM DATA

	Accuracy	Precision	Recall	F-Score	Learning Time in minutes
Alex Net	0.992	0.981	0.98	0.979	0.28
Vision Transformer	0.976	0.946	0.97	0.939	5.2

Sample classified output images with their labels obtained from AlexNet and the ViT are shown in Fig. 9 and Fig. 10. Table III and Table IV show the results of comparative performance for 5-Class and 10-Class UCM datasets, respectively.

TABLE IV. RESULTS FOR 10-CLASS UCM DATA

	Accuracy	Precision	Recall	F-Score	Learning Time in minutes
Alex Net	0.99	0.962	0.96	0.958	0.81
Vision Transformer	0.976	0.901	0.88	0.877	9.5

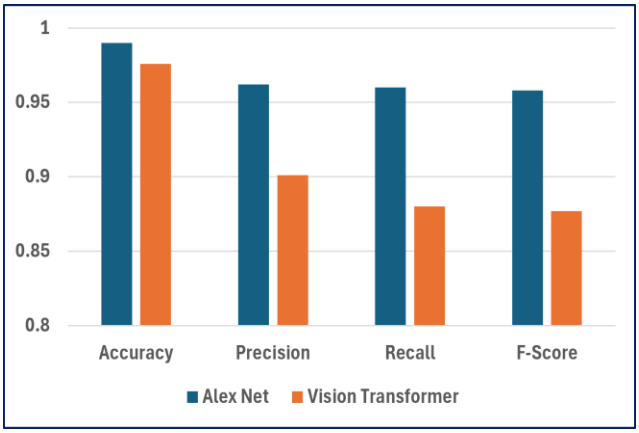


Fig. 8. Comparative performance for 10-class UCM dataset.

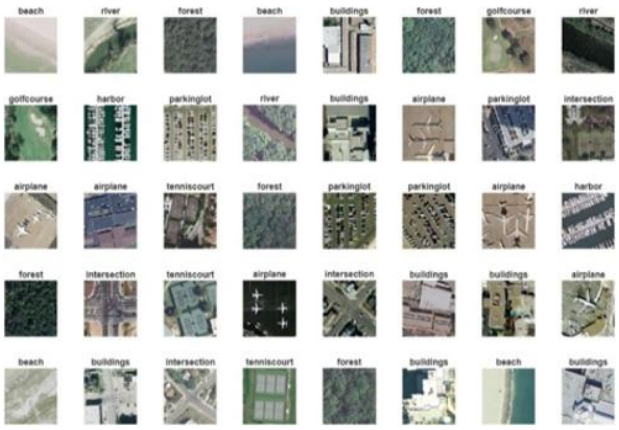


Fig. 9. Classified output images with AlexNet.

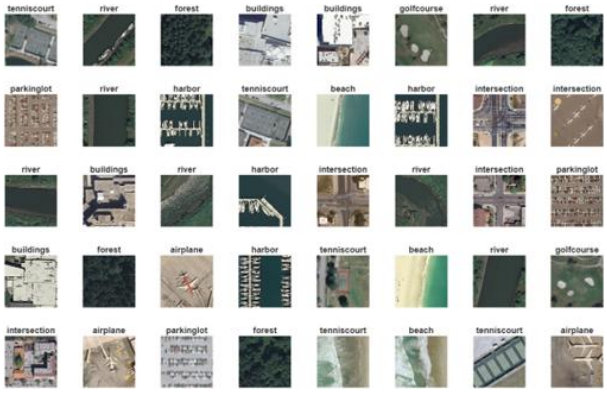


Fig. 10. Classified output images with ViT.

2) Results for EuroSAT Dataset: The learning progress graph for EuroSAT Dataset for AlexNet and Vision Transformer are shown in Fig. 11 and Fig. 12, respectively.

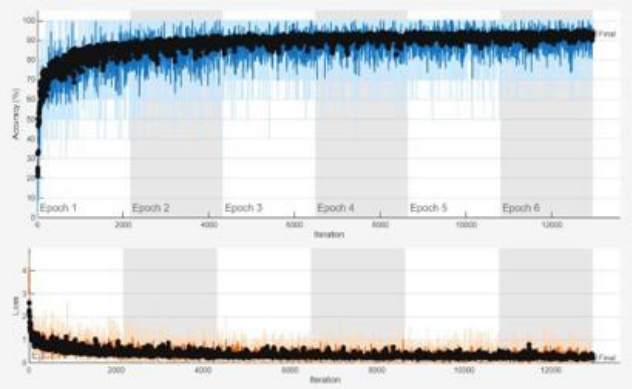


Fig. 11. Learning progress with AlexNet on the 10-class EuroSAT Dataset.

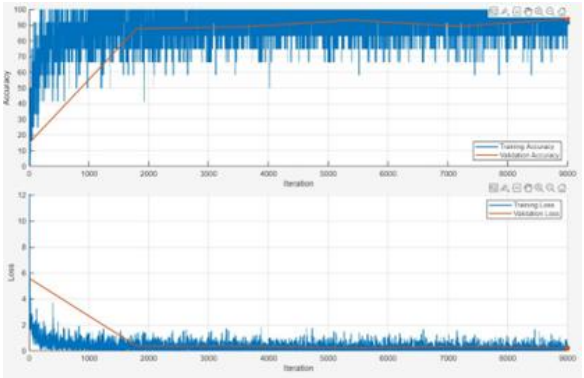


Fig. 12. Learning progress with ViT on the 10-class EuroSAT Dataset.

The comparison of the average values of the metrics—accuracy, precision, recall, and F1-score—obtained from the confusion matrices for the EuroSAT Dataset using AlexNet and Vision Transformer is shown in Fig. 13 and Fig. 14, respectively. Sample classified output images with their labels obtained from AlexNet and the ViT are shown in Fig. 15 and Fig. 16. Table V and Table VI show the results of comparative performance for the 5-Class and 10-Class EuroSAT Datasets, respectively.

TABLE V. RESULTS FOR 5-CLASS EUROSAT DATASET

	Accuracy	Precision	Recall	F-Score	Learning Time in minutes
Alex Net	0.984	0.958	0.958	0.958	38
Vision Transformer	0.986	0.964	0.964	0.964	128

TABLE VI. RESULTS FOR 10-CLASS EUROSAT DATASET

	Accuracy	Precision	Recall	F-Score	Learning Time in minutes
Alex Net	0.982	0.93	0.931	0.929	137
Vision Transformer	0.988	0.9358	0.9423	0.938	253

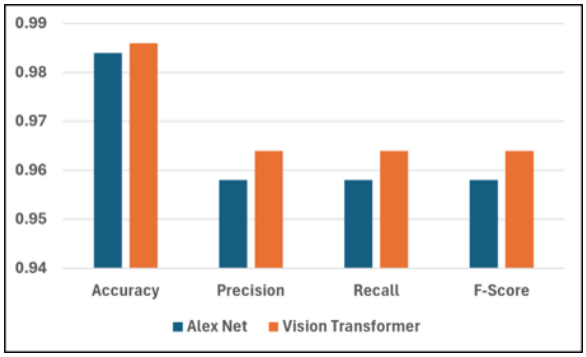


Fig. 13. Comparative performance for the 5-class EuroSAT Dataset.

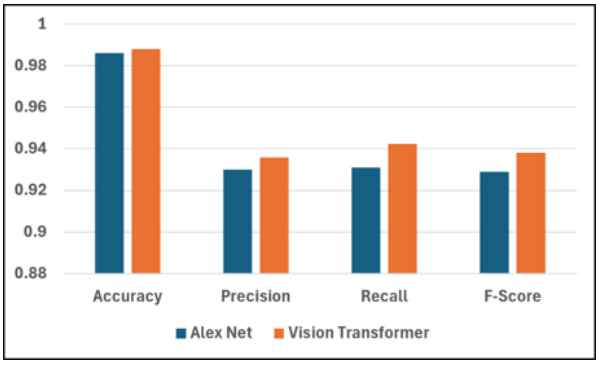


Fig. 14. Comparative performance for the 10-class EuroSAT Dataset.

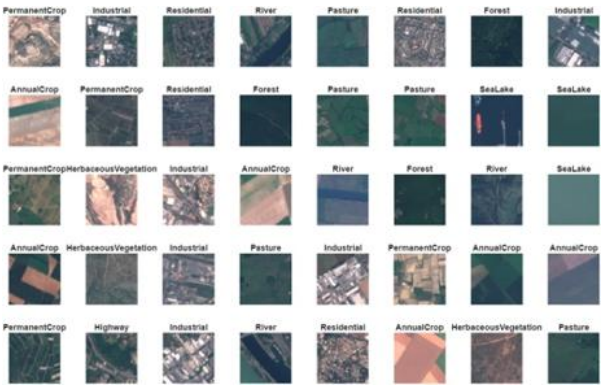


Fig. 15. Classified output with AlexNet for the 10-class EuroSAT Dataset.

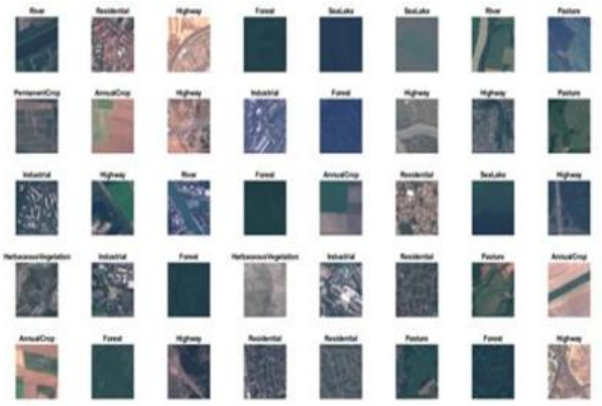


Fig. 16. Classified output with ViT for the 10-class EuroSAT Dataset.

V. CONCLUSION

In this study, we implemented and evaluated the Convolutional Neural Network (CNN)-based AlexNet model and the Vision Transformer (ViT-B16) model for land use scene classification using benchmark remote sensing image datasets. The experiments were conducted on the 5-class and 10-class versions of the UCM Land Use Dataset and the EuroSAT RGB Dataset. Both models were trained and tested on a system equipped with an AMD Ryzen 9 7950X processor and 128 GB RAM. For both models and all datasets, 80 per cent of the randomly selected samples from each class were used for training, 10 per cent for validation, and the remaining 10 per cent for testing. By enforcing identical preprocessing, training, and evaluation conditions, the proposed framework enables a clear and interpretable assessment of architectural behavior. Model performance was evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score. The evaluation results for the UCM datasets are presented in Fig. 7 and Fig. 8. The results indicate that AlexNet achieves superior performance when the dataset contains a relatively small number of training samples. The UCM dataset includes approximately 100 samples per class, which appears to favor CNN-based feature extraction and classification. The evaluation results for the EuroSAT Dataset are presented in Fig. 13 and Fig. 14. In contrast to the UCM dataset results, the ViT-B16 model demonstrates slightly better performance than AlexNet on the EuroSAT Dataset, where the number of samples per class ranges from 2,000 to 3,000 images. This observation suggests that Vision Transformer architectures are more effective when trained on larger-scale datasets, as they can better exploit abundant training data to learn global contextual representations. Despite their improved performance on larger datasets, Vision Transformer models require greater computational resources. In our experiments, training the ViT-B16 model on the 10-class EuroSAT Dataset required approximately 253 minutes, whereas AlexNet required only 137 minutes under the same hardware configuration. In addition, ViT models demand greater memory capacity and computational complexity compared with conventional CNN architectures. Overall, the experimental results indicate that CNN-based models such as AlexNet are more suitable for land-use scene classification tasks involving limited training samples, whereas Vision Transformer models achieve better performance when sufficient training data samples are available. However, the improved performance of Vision Transformers comes at the cost of increased computational requirements.

Future research may focus on hybrid architectures that combine the local feature extraction capability of CNNs with the global attention mechanisms of Vision Transformers to further improve land use scene classification accuracy and robustness. In addition, future work will include the evaluation of other CNN architectures, such as VGGNet, ResNet, and GoogLeNet, as well as more efficient Vision Transformer variants. Comparative analysis among these models may provide further insights into their suitability for remote sensing image classification. Furthermore, hybrid architectures integrating feature representations extracted from CNN and Vision Transformer models may be developed, where the combined feature vectors are classified using a unified SoftMax layer to improve classification accuracy. Nevertheless, such hybrid approaches are expected to require additional computational resources and training time.

REFERENCES

[1] B. Zhao, Y. Zhong, G. Xia, and L. Zhang, "Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery," IEEE Transactions on Geoscience and Remote Sensing, vol. 54, no. 4, pp. 2108–2123, 2016.

[2] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 117, pp. 11–28, 2016.

[3] G.-S. Xia et al., "AID: A benchmark dataset for performance evaluation of aerial scene classification," IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 7, pp. 3965–3981, 2017.

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436–444, 2015.

[5] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. of ICML, vol. 97, pp. 6105-6114, 2019.

[6] Z. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. of ICLR, 2021. doi.org/10.48550/arXiv.2010.11929

[7] D. Lu and Q. Weng. "A survey of image classification methods and techniques for improving classification performance". International Journal of Remote Sensing, vol. 28, no. 5, pp. 823–870. 2007

[8] L. Zhao, J. Wang, and Y. Chen, "Vision transformer for remote sensing image classification: A review," Remote Sensing, vol. 14, no. 15, p. 3776, 2022.

[9] D. G. Lowe, "Distinctive image features from scale-invariant key points," International Journal of Computer Vision, vol. 60, no. 2, pp. 91–110, 2004.

[10] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2005.

[11] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 971–987, 2002.

[12] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," Remote Sensing, vol. 7, no. 11, pp. 14680–14707, 2015.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems, vol. 25, pp. 1097–1105, 2012.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. Int. Conf. Learn. Represent. (ICLR), 2015.

[15] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 1–9.

[16] X. X. Zhu, D. Tuia, L. Mou et al., "Deep learning in remote sensing: A comprehensive review and list of resources," IEEE Geosci. Remote Sens. Mag., vol. 5, no. 4, pp. 8–36, 2017.

[17] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," Proc. IEEE, vol. 105, no. 10, pp. 1865–1883, 2017.

[18] O. A. B. Penatti, K. Nogueira, and J. A. dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPR Workshops), 2015, pp. 44–51.

[19] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 5998–6008.

[20] K. Han, Y. Wang, J. Guo et al., "Vision transformers in computer vision: A survey," ACM Comput. Surv., vol. 55, no. 6, pp. 1–41, 2023.

[21] Y. Xu, X. Zhang, H. Li et al., "Hybrid transformer networks for remote sensing image scene classification," Remote Sens., vol. 15, no. 4, p. 1021, 2023.

[22] P. Helber, B. Bischke, A. Dengel, and D. Borth, "EuroSAT: A novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 12, no. 7, pp. 2217–2226, July 2019, doi: 10.1109/JSTARS.2019.2918242.

[23] Y. Duan, C. Song, Y. Zhang, P. Cheng, and S. Mei, "STMSF: Swin Transformer with Multi-Scale Fusion for Remote Sensing Scene Classification," Remote Sensing, vol. 17, no. 4, p. 668, 2025, doi: 10.3390/rs17040668.

[24] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc. of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[25] C. M. Bishop, Pattern Recognition and Machine Learning. New York, NY, USA: Springer, 2006.

[26] G. Camps-Valls, D. Tuia, X. X. Zhu, and M. Reichstein, "Deep learning for the Earth sciences: A comprehensive approach to remote sensing, climate science, and geosciences," Nature Reviews Earth & Environment, 2021.

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

[28] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," ICLR, 2015.

[29] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," Proc. of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS 2010), pp. 270–279, 2010.

[30] https://www.kaggle.com/datasets/abdulhasibuddin/uc-merced-land-use-dataset

[31] https://www.kaggle.com/datasets/apollo2506/e

Privacy-Preserving Financial Fraud Detection Using Federated Heterogeneous Temporal GraphSAGE

Sabab Islam[1], Md. Abul Kalam Azad[2], Abdul Kadar Muhammad Masum[3], Ramona Birau[4], Virgil Popescu[5], Roxana Mihaela Nioata (Chireac)[6], Gabriela Ana Maria Lupu (Filip)[7]
College of Business, Texas A&M University Commerce, Texas, USA[1]
Department of Business and Technology Management, Islamic University of Technology, Dhaka 1704, Bangladesh[2]
Department of Computer Science and Engineering, Southeast University, Dhaka 1208, Bangladesh[3]
"Eugeniu Carada" Doctoral School of Economic Sciences, University of Craiova, Craiova, Romania[4,6,7]
Faculty of Economics and Business Administration, University of Craiova, Craiova, Romania[5]

Abstract*—Digital financial ecosystems face mounting exposure to fraudulent transactions that collectively account for trillions of dollars in losses each year. Existing approaches suffer from three recurring deficiencies: they represent all transaction participants within a single undifferentiated node space, rely on fixed decision boundaries incapable of accommodating evolving fraud distributions, and fail to exploit the semantic diversity among entity categories including accounts, merchants, and transaction types. This study introduces FHT-GraphSAGE, a federated heterogeneous temporal graph learning framework for privacy-preserving financial fraud detection. The architecture combines heterogeneous temporal graph construction, a relation-aware GraphSAGE encoder with sinusoidal temporal edge embeddings, and a Federated Averaging optimization scheme that enables cross-institutional collaborative learning without exposing raw transaction records. Evaluation across two complementary benchmarks, the PaySim Mobile Money Dataset and the Credit Card Fraud 2023 Dataset, demonstrates consistent superiority over ten competitive baselines. On PaySim, FHT-GraphSAGE achieves an accuracy of 0.963, an F1-score of 0.939, an AUC-ROC of 0.986, and, most importantly under severe class imbalance, an AUC-PR of 0.724. On the Credit Card Fraud 2023 Dataset, it attains an accuracy of 0.941, an F1-score of 0.913, an AUC-ROC of 0.969, and an AUC-PR of 0.721. As AUC-PR is the most informative metric for minority-class detection under extreme imbalance, these two figures (0.724 and 0.721) constitute the primary evidence of the framework's fraud-detection capability. Results are reported as the mean over five independent runs with distinct random seeds, and the improvements over the strongest baseline are statistically significant ($p < 0.01$). Ablation experiments confirm the non-redundant contribution of each architectural component, and robustness evaluations demonstrate stable minority-class detection performance down to a fraud ratio of 0.25%.

***Keywords*—Financial fraud detection; federated learning; heterogeneous temporal graph; GraphSAGE; privacy-preserving machine learning; graph neural networks**

I. Introduction

Unauthorized activity in digital payment infrastructure causes an estimated $5.127 trillion in worldwide monetary losses annually [1]. Beyond the immediate financial damage, incidents at this scale erode public trust in banking systems and constrain broader participation in formal economic structures. Rapid digitalization of payment services has compounded these risks: transaction volumes now far exceed what human reviewers can examine, while coordinated attacks simultaneously exploit multiple accounts, payment instruments, and merchant channels [2]. Methods that evaluate each transaction independently are therefore structurally incapable of capturing the relational dimension of contemporary fraud.

Several inherent characteristics of this problem compound detection difficulty. Adversarial strategies evolve continuously, outpacing the update cycles of rule-based systems [2]. Legal and regulatory constraints prevent financial institutions from pooling raw transaction histories, even when doing so would materially improve detection [3]. Fraudulent activity constitutes a negligibly small fraction of total transaction volume, which drives standard classification objectives toward the majority class and suppresses minority-class recall [4].

Current detection pipelines address individual aspects of this problem without resolving the full set of challenges. Rule-based filters offer transparency but demand continuous manual recalibration as tactics shift [5]. Conventional supervised classifiers process transactions as statistically independent records, discarding the relational network that fraud rings exploit [6]. Graph neural network methods encode structural topology but the majority still operate on static adjacency matrices that become stale as new interactions arrive. Fixed decision boundaries degrade progressively under concept drift [7]. Privacy-preserving approaches that could overcome cross-institutional data barriers have rarely been combined with graph-structured fraud representations [8]. These gaps collectively indicate that a unified solution must address relational heterogeneity, temporal dynamics, distributed privacy constraints, and generalization across structurally dissimilar financial-fraud datasets simultaneously.

The proposed framework satisfies these four requirements within a single coherent architecture. First, transaction records from each participating institution are transformed into heterogeneous temporal graphs that explicitly separate customer, merchant, and transaction node categories, connected by semantically typed directed edges. Second, a temporal edge encoding function maps interaction time gaps into continuous sinusoidal embeddings, injecting sequential behavioral context into neighborhood aggregation. Third, a relation-aware GraphSAGE encoder applies separate aggregation pathways per relation type, fusing the resulting representations through a learnable projection to produce inductive node embeddings that

generalize to previously unseen entities. Fourth, a Federated Averaging protocol coordinates parameter aggregation across institutional clients without requiring any institution to transmit raw records, satisfying practical privacy and regulatory requirements.

By unifying heterogeneous temporal graph construction, relation-aware inductive aggregation, and privacy-preserving federated optimization within a single coherent architecture, FHT-GraphSAGE represents a meaningful advancement in collaborative financial fraud detection. Beyond its immediate contributions to transaction security, this work opens broader pathways for applying federated graph learning to domains characterized by complex, distributed sensitive data, evolving relational structure, and strict institutional privacy requirements, including healthcare claim auditing, telecommunications anomaly detection, and supply chain integrity verification.

II. Literature Review

Foundational work on transaction fraud relied on statistical profiling derived from historical transaction records. Bhattacharyya et al. [9] benchmarked several data mining classifiers on credit card transactions and demonstrated that minority-class scarcity reliably degrades detection across all evaluated methods. Bolton and Hand [10] introduced a peer-group analysis approach that flags cardholders whose expenditure trajectories diverge significantly from those of statistically comparable users. With the maturation of supervised learning, Whitrow et al. [11] established that aggregating records into temporal behavioral windows substantially elevates classifier performance over raw transaction features. Dal Pozzolo et al. [12] showed that sliding-window retraining only partially compensates for concept drift and requires manual scheduling of update cycles. Roy et al. [6] demonstrated measurable accuracy gains from feed-forward neural networks relative to shallow classifiers on nonlinear spending patterns. Wang et al. [13] combined reconstruction-based autoencoders with ensemble classifiers to improve performance under severe label imbalance. A limitation shared across all these approaches is that each transaction is evaluated in isolation, without access to the broader relational network within which fraud rings operate.

Representing financial activity as a graph provides a principled mechanism for reasoning over entity relationships and interaction patterns. Akoglu et al. [14] surveyed graph anomaly detection spanning community-level, subgraph-level, and node-level methods. Liu et al. [15] applied isolation-based anomaly scoring to graph-structured data, though this approach discards both edge semantics and temporal ordering. The introduction of graph neural networks substantially expanded representational capability. Kipf and Welling [16] proposed spectral graph convolutions that propagate neighborhood features through a symmetrically normalized adjacency operator. Hamilton et al. [17] developed GraphSAGE, an inductive framework that learns node embeddings by sampling and aggregating local neighborhoods, supporting generalization to previously unobserved nodes. Dou et al. [18] designed an attention-weighted graph network capable of identifying fraudsters who deliberately camouflage their activity by mimicking legitimate behavioral patterns. Liu et al. [19] constructed a heterogeneous financial graph that incorporates multiple relationship categories to reduce feature inconsistency across edge types. More recently, Khosravi et al. [20] proposed an attentional spatial-temporal GNN for transaction fraud detection, while Islam et al. [21] introduced a layer-weighted GCN to better separate fraud patterns across heterogeneous transaction types, illustrating continued efforts to refine relational representations. A persistent limitation across these graph-based methods is dependence on static adjacency matrices that do not evolve as new transactions arrive, making them structurally insensitive to behavioral drift over time.

Capturing temporal dynamics in fraud requires architectures that revise node representations incrementally as interactions occur. Rossi et al. [22] introduced the Temporal Graph Network, which maintains a persistent memory state per node and refreshes it through a gated recurrent mechanism upon each observed interaction. Elapsed time since the most recent event is encoded through a learnable projection mapping scalar gaps to continuous embeddings. Li et al. [23] developed a heterogeneous dynamic graph architecture for spatiotemporal forecasting that constructs typed subgraphs and refreshes their adjacency through multi-head self-attention over recent observation windows. Dal Pozzolo et al. [24] showed that adaptive sliding-window retraining partially tracks distributional shift but remains constrained by periodic update schedules that lag real-time fraud evolution. Cui et al. [25] recently combined temporal-spatial-semantic graph convolution with reinforcement learning and federated coordination in a single framework, representing one of the few prior attempts to jointly address temporal, relational, and privacy-preserving objectives. However, their reinforcement-learning-driven adaptation differs fundamentally from the sinusoidal temporal encoding and relation-specific aggregation pathways proposed here.

Velickovic et al. [26] introduced graph attention networks that compute neighbor contribution weights through a shared attention function over concatenated node feature pairs. Brody et al. [27] subsequently identified a fundamental limitation in this formulation: applying the linear transform before the nonlinearity causes neighbor ranking to be independent of the querying node, yielding static rather than dynamic attention. Their revised model restructures the computation so that independent linear projections of the query and key nodes are combined inside the nonlinearity, producing genuinely query-dependent scores. The present work draws on this insight by applying separate aggregation pathways per relation type within GraphSAGE, achieving relational specificity without the communication overhead of attention-based mechanisms. Federated learning has emerged as a practical paradigm for collaborative model development under privacy constraints. McMahan et al. [28] formalized the federated learning setting and proposed Federated Averaging, in which clients optimize local objectives and a central server aggregates their parameters. This approach has become a foundational component of privacy-preserving learning research. Yang et al. [29] surveyed federated learning applications across multiple domains and identified financial fraud as a high-priority use case given the sensitivity of transaction data. More recently, Awosika et al. [30] examined the combined role of explainable AI and federated learning in addressing both transparency and privacy requirements in fraud detection systems, underscoring

the continued relevance of federated approaches to regulatory compliance. Zheng et al. [31] demonstrated vertical federated learning for credit card fraud, enabling institutions holding complementary feature subsets over a shared population to collaborate without exposing local records. Aurna et al. [32] further showed that combining federated coordination with deep learning and resampling strategies yields meaningful gains on imbalanced fraud benchmarks, though their framework treats transactions as isolated events and lacks graph-structured relational modeling. The integration of federated learning with heterogeneous temporal graph encoders for cross-institutional fraud detection remains largely unexplored, representing the primary methodological contribution of the present work. While recent work such as FraudGNN-RL [25] has begun combining temporal graph modeling with federated coordination, it does not incorporate explicit relation-type-aware aggregation or continuous sinusoidal temporal encoding, leaving the specific combination proposed here unaddressed.

Surveying the prior works reveals several unresolved gaps. Graph-based fraud detectors predominantly collapse all participants into a homogeneous node space, discarding relational distinctions that carry diagnostic value. Temporal modeling approaches exist but few combine continuous-time behavioral tracking with explicit heterogeneous graph structure. Existing federated fraud detection methods do not incorporate graph-structured encoders capable of capturing relational and temporal fraud patterns across distributed clients. Evaluation across multiple datasets representing structurally different fraud ecosystems within the financial domain is rarely performed. The proposed FHT-GraphSAGE framework addresses each gap. Heterogeneous temporal graph construction explicitly separates customer, merchant, and transaction entities connected by typed temporal edges. A relation-aware GraphSAGE encoder with sinusoidal temporal embeddings captures both relational and sequential fraud dynamics inductively. A Federated Averaging protocol enables privacy-preserving cross-institutional collaboration without transmitting raw records. Evaluation is performed on two structurally dissimilar financial-fraud benchmarks. One benchmark provides explicit relational identifiers and the other provides anonymized principal-component features. This evaluation supports generalization across datasets within the financial-fraud domain. We deliberately refrain from claiming generalization to non-financial domains, which the present evaluation does not test.

III. Dataset Description

The proposed FHT-GraphSAGE framework is evaluated on two publicly available financial fraud detection benchmarks: PaySim Mobile Money Dataset and Credit Card Fraud 2023 Dataset. PaySim is a synthetic mobile financial transaction dataset generated by a simulator calibrated on aggregated real-world mobile payment statistics [33]. The dataset captures account-to-account interactions across five transaction categories: `TRANSFER`, `CASH_OUT`, `PAYMENT`, `DEBIT`, and `CASH_IN`. Each record includes a unique sender identifier (`nameOrig`), a unique receiver identifier (`nameDest`), transaction amount, account balance states before and after the operation for both parties, and a binary fraud label. Transactions are chronologically ordered by discrete time steps, making the dataset well suited for temporal graph construction and sequential behavioral modeling. Fraudulent events occur exclusively within `TRANSFER` and `CASH_OUT` categories, a structural property that is preserved during graph formulation.

The Credit Card Fraud 2023 Dataset comprises anonymized European credit card transactions annotated with binary fraud labels [34]. Raw features are replaced by a set of numerical principal components to protect cardholder confidentiality, supplemented by the original transaction amount and a temporal index. The dataset exhibits severe class imbalance, with fraudulent records constituting a very small fraction of the total transaction volume. This property makes it a challenging and realistic benchmark for evaluating detection performance under minority-class scarcity. The temporal ordering of records further enables continuous-time representation learning over evolving transaction streams.

IV. Methodology

A. Data Preprocessing

A sequence of preprocessing steps is applied to each dataset prior to graph construction in order to ensure feature consistency and numerical stability. Missing numerical values are imputed using column-wise means, while absent categorical entries are substituted with the most frequent value observed within each column. Continuous features are subsequently rescaled to the unit interval $[0, 1]$ via min–max normalization to prevent scale disparities from distorting neighborhood aggregation. Categorical variables are converted into dense binary vectors through one-hot encoding before being incorporated into node feature matrices.

Temporal attributes are extracted explicitly from transaction timestamps. Three derived signals are produced for each record: the transaction hour, the day-of-week index, and the elapsed duration δt_k between consecutive interactions involving the same entity. For the PaySim dataset, sender and receiver account identifiers are mapped directly to graph entities. For the Credit Card Fraud 2023 Dataset, since entity identifiers are unavailable, inter-transaction edges are constructed through two complementary strategies: feature-space k-nearest-neighbor similarity and temporal proximity thresholding, allowing transactions that are behaviorally or temporally close to share structural information during graph encoding.

B. Problem Formulation

1) Heterogeneous temporal graph construction: Financial transaction systems involve multiple distinct entity categories interacting through semantically different relation types. Collapsing these into a single homogeneous node set discards information that is diagnostically relevant for fraud detection. The transaction environment is therefore modeled as a *heterogeneous temporal graph* that preserves these semantic distinctions.

a) Node types: Let

$$\mathcal{T}_V = \{\texttt{Customer}, \texttt{Merchant}, \texttt{Transaction}\} \quad (1)$$

denote the set of node types. For the PaySim dataset, customer nodes correspond to sender and receiver accounts,

merchant nodes represent destination entities, and transaction nodes encode individual financial operations as discrete graph vertices. In the Credit Card Fraud 2023 Dataset, each transaction record is instantiated as a transaction node, with inter-node edges derived from feature-space proximity and temporal adjacency. The complete node set is:

$$\mathcal{V} = \mathcal{V}^{\,\mathrm{cus}} \cup \mathcal{V}^{\,\mathrm{mer}} \cup \mathcal{V}^{\,\mathrm{txn}}, \tag{2}$$

where $\mathcal{V}^{\mathrm{cus}}$, $\mathcal{V}^{\mathrm{mer}}$, and $\mathcal{V}^{\mathrm{txn}}$ collect customer, merchant, and transaction nodes, respectively. A type assignment function

$$\psi : \mathcal{V} \to \mathcal{T}_V \tag{3}$$

maps every node to exactly one semantic category.

b) Relation types: Let

$$\mathcal{R} = \{\, \texttt{sends},\ \texttt{receives},\ \texttt{transfers},\ \texttt{similar_to} \,\} \tag{4}$$

denote the set of relation types. The `sends` and `receives` relations connect customer nodes to transaction nodes. The `transfers` relation links sender customer nodes to receiver customer nodes. The `similar_to` relation connects transaction nodes whose feature vectors or temporal positions fall within a defined similarity threshold, and is used primarily for the Credit Card Fraud 2023 Dataset where explicit entity identifiers are unavailable. Each transaction event generates one or more directed temporal edges among the participating entities according to the applicable relation type.

c) Formal graph definition: At time step t, the evolving transaction network is represented as:

$$\mathcal{G}_t = \big(\mathcal{V}_t,\ \mathcal{E}_t\big), \tag{5}$$

where $\mathcal{V}_t$ is the active node set and $\mathcal{E}_t$ is the set of directed temporal edges present at time t. An individual transaction interaction is encoded as a typed temporal edge tuple:

$$\xi_k = \big(u_i,\ u_j,\ r_k,\ \tau_k,\ \mathbf{x}_k\big), \tag{6}$$

where, $u_i, u_j \in \mathcal{V}$ are the source and destination nodes, $r_k \in \mathcal{R}$ is the relation type, τ_k is the continuous event timestamp, and $\mathbf{x}_k \in \mathbb{R}^d$ is the transaction feature vector attached to the edge.

2) Fraud detection objective: For each observation window ending at time t, node feature vectors are assembled into a matrix:

$$\mathbf{X}_t = \big[\mathbf{x}_1,\ \mathbf{x}_2,\ \ldots,\ \mathbf{x}_n\big]^\top \in \mathbb{R}^{n \times d}, \tag{7}$$

where, each row corresponds to a graph entity active within the current temporal interval. The fraud detection objective is formulated as binary classification over transaction edges. Given the heterogeneous temporal graph $\mathcal{G}_t$ and the feature history up to t, the framework learns a mapping:

$$\mathcal{F} : \big(\mathcal{G}_t,\ \mathbf{X}_{1:t}\big) \ \longrightarrow\ \{0,\ 1\}, \tag{8}$$

where, output 0 denotes a legitimate transaction and 1 denotes a fraudulent one.

C. Temporal Heterogeneous GraphSAGE Encoder

Fraudulent activity in financial networks frequently manifests as gradual shifts in interaction patterns among related entities. Capturing such evolving behavioral dependencies requires a graph encoder that is both temporally aware and sensitive to relational heterogeneity. To this end, a Temporal Heterogeneous GraphSAGE encoder is introduced, which performs inductive neighborhood aggregation over the typed transaction graph. Unlike attention-based architectures that require full-neighborhood computation across all relation types jointly, GraphSAGE learns node representations through sampled neighborhood aggregation, a property that reduces computational overhead and makes the encoder well suited for distributed federated environments.

1) Temporal edge encoding: Before neighborhood aggregation is performed, temporal context is injected into each edge representation. For an edge connecting nodes u_i and u_j, the elapsed duration since their most recent interaction is encoded through a learnable time embedding function:

$$\mathbf{q}_{ij} = \Phi(\delta\tau_{ij})\,, \tag{9}$$

where, $\delta\tau_{ij} = \tau_{\mathrm{current}} - \tau_{\mathrm{last}}^{(i,j)}$ is the scalar time gap and $\Phi : \mathbb{R} \to \mathbb{R}^{d_t}$ is implemented as a sinusoidal basis expansion:

$$\Phi(\delta\tau)_p = \begin{cases} \cos(\delta\tau\ /\ \omega_p)\,, & p \text{ even}, \\ \sin(\delta\tau\ /\ \omega_p)\,, & p \text{ odd}, \end{cases} \tag{10}$$

with learnable frequency parameters $\{\omega_p\}_{p=1}^{d_t}$. The temporal embedding is added to the source node's feature representation before aggregation:

$$\widetilde{\mathbf{z}}_j = \mathbf{z}_j + \mathbf{q}_{ij}, \tag{11}$$

so that neighborhood messages carry information about *when* the interaction occurred, not only what features were associated with it.

2) Relation-aware neighborhood aggregation: Each node u_i is associated with a representation $\mathbf{z}_i^{(\ell)} \in \mathbb{R}^d$ at GraphSAGE layer ℓ. The standard GraphSAGE update concatenates the node's own representation with an aggregated neighborhood summary and applies a learnable linear transformation followed by a nonlinearity:

$$\mathbf{z}_i^{(\ell+1)} = \sigma\Big(\Theta^{(\ell)} \cdot \Big[\mathbf{z}_i^{(\ell)} \,\|\, \mathrm{AGG}^{(\ell)}\Big(\big\{\widetilde{\mathbf{z}}_j^{(\ell)} : u_j \in \mathcal{N}(u_i)\big\}\Big)\Big]\Big), \tag{12}$$

where, $\Theta^{(\ell)} \in \mathbb{R}^{d \times 2d}$ is the trainable weight matrix at layer ℓ, $\mathcal{N}(u_i)$ denotes the sampled neighbor set of u_i, $\mathrm{AGG}(\cdot)$ is the aggregation function (mean pooling by default), and $\|$ denotes vector concatenation.

To preserve the semantic distinctions among the heterogeneous relation types, a separate aggregation pathway is instantiated for each relation $r \in \mathcal{R}$:

$$\mathbf{z}_{i,r}^{(\ell+1)} = \sigma\Big(\Theta_r^{(\ell)} \cdot \mathrm{Mean}\Big(\big\{\widetilde{\mathbf{z}}_j^{(\ell)} : u_j \in \mathcal{N}_r(u_i)\big\}\Big)\Big), \quad (13)$$

where, $\mathcal{N}_r(u_i)$ is the neighbor set of u_i reachable through relation r, and $\Theta_r^{(\ell)}$ is the relation-specific weight matrix. This formulation allows the encoder to treat funds-transfer interactions, merchant-payment interactions, and behavioral-similarity links as semantically independent information channels. Representations learned under each relation are fused through concatenation:

$$\mathbf{z}_i = \mathbf{W}_{\mathrm{fuse}} \left(\Big\| _{r \in \mathcal{R}} \mathbf{z}_{i,r}^{(L)} \right), \quad (14)$$

where, L is the total number of GraphSAGE layers and $\mathbf{W}_{\mathrm{fuse}}$ is a learnable projection matrix that maps the concatenated multi-relational embedding back to dimension d.

3) Classification head: Each transaction edge $\xi_k = (u_i, u_j, r_k, \tau_k, \mathbf{x}_k)$ is represented by concatenating the final node embeddings of its endpoints together with the edge feature vector:

$$\mathbf{e}_k = \big[\mathbf{z}_i \,\|\, \mathbf{z}_j \,\|\, \mathbf{x}_k\big]. \quad (15)$$

A two-layer feedforward classifier with sigmoid output produces a fraud probability estimate:

$$\hat{y}_k = \sigma(\mathbf{W}_2\, \mathrm{ReLU}(\mathbf{W}_1\, \mathbf{e}_k + \mathbf{b}_1) + \mathbf{b}_2), \quad (16)$$

and the model is optimized by minimizing a class-weighted binary cross-entropy loss:

$$\mathcal{L} = -\frac{1}{|\mathcal{E}|} \sum_{k \in \mathcal{E}} \Big[w_1\, y_k \log \hat{y}_k + w_0\, (1 - y_k) \log(1 - \hat{y}_k) \Big], \quad (17)$$

where, $y_k \in \{0, 1\}$ is the ground-truth label and $w_1 \gg w_0$ are class weights set inversely proportional to class frequency to compensate for severe label imbalance.

D. Federated Learning Framework

Financial institutions operate under strict regulatory and contractual privacy constraints that prohibit the centralized pooling of raw transaction records. A federated learning paradigm is adopted to enable collaborative fraud model training without requiring any institution to expose its local data.

1) Federated setup: Let the distributed financial environment consist of K independent client institutions:

$$\mathcal{C} = \{c_1, c_2, \ldots, c_K\}, \quad (18)$$

where, each client c_k holds a private local transaction graph $\mathcal{G}^{(k)}$ and the corresponding fraud annotations. A central server coordinates parameter aggregation but has no access to any client's raw records.

2) Local objective: During each communication round t, every client c_k independently minimizes its local empirical loss:

$$\mathcal{L}_k(\Theta) = -\frac{1}{|\mathcal{E}^{(k)}|} \sum_{j \in \mathcal{E}^{(k)}} \Big[w_1\, y_j \log \hat{y}_j(\Theta) + w_0\, (1 - y_j) \log\big(1 - \hat{y}_j(\Theta)\big) \Big], \quad (19)$$

using its local transaction graph and the current global model parameters $\Theta^{(t)}$ as initialization. After a fixed number of local gradient steps, the updated local parameters $\Theta_k^{(t)}$ are transmitted to the central server.

3) Global aggregation: The server performs weighted parameter aggregation using the Federated Averaging (FedAvg) strategy [28]:

$$\Theta^{(t+1)} = \sum_{k=1}^{K} \frac{n_k}{N}\, \Theta_k^{(t)}, \quad (20)$$

where, n_k is the number of training edges on client c_k and

$$N = \sum_{k=1}^{K} n_k \quad (21)$$

is the total number of training samples across all clients. The aggregated global parameters $\Theta^{(t+1)}$ are redistributed to all clients as the starting point for the next local training round. This cycle repeats for T communication rounds. Raw transaction records remain entirely within their originating institution throughout this process; only gradient-derived model parameters traverse the network.

4) Aggregation weighting under heterogeneous client distributions: The FedAvg rule in Eq. (20) weights each client's contribution by its share of training edges n_k/N. This choice is statistically optimal only when local objectives are unbiased estimators of the same global objective. Such a condition holds when data are independent and identically distributed (IID) across clients. In cross-institutional fraud detection this assumption is frequently violated. Institutions differ in customer demographics and in product mix. They also differ, and this matters most, in their local fraud prevalence $\pi_k = \sum_{j \in \mathcal{E}^{(k)}} y_j / |\mathcal{E}^{(k)}|$. Under such non-IID conditions, proportional weighting lets a large and low-fraud client dominate the aggregate. It can then dilute the minority-class gradient signal contributed by a smaller and high-fraud client, which may depress global recall. Three properties of the present design

partially mitigate this effect. First, the class-weighted loss in Eq. (19) rescales each client's fraudulent-sample gradients inversely to that client's *local* class frequency. The effective minority-class contribution transmitted by client k is therefore amplified before aggregation rather than being governed solely by n_k. Second, only parameters are exchanged rather than raw fraud counts. The aggregate still benefits from the diversity of fraud patterns observed across clients even when their prevalences differ. Third, the inductive encoder shares representation structure across clients. Relational and temporal patterns learned on one institution's graph therefore transfer to others. Proportional weighting nonetheless remains a simplifying assumption. Section V-C discusses its residual limitations. Prevalence-aware or regularized aggregation schemes such as FedProx are identified as concrete future extensions.

E. Complete Detection Pipeline

The end-to-end FHT-GraphSAGE detection pipeline is illustrated in Fig. 1. The upper portion of the figure traces the six local processing steps that each client institution executes on its private transaction data, while the lower portion depicts the federated coordination loop that spans all K clients and the central aggregation server. The upper row (Steps 1–6) covers the local processing steps executed by each client institution: raw transaction records (Step 1) are preprocessed through missing-value imputation, min–max normalization, one-hot encoding, and temporal feature extraction (Step 2); the preprocessed records are then organized into a heterogeneous temporal graph whose three node categories i.e. Customer, Merchant, and Transaction, are connected by four directed relation types: `sends`, `receives`, `transfers`, and `similar_to` (Step 3), with the `similar_to` edges constructed via feature-space k-NN similarity and temporal proximity thresholding for the Credit Card Fraud 2023 Dataset; each edge (u_i, u_j) is assigned a sinusoidal temporal embedding $\mathbf{q}_{ij} = \Phi(\Delta\tau_{ij})$ of dimension $d_t = 64$, which is added to the source node feature to produce $\widetilde{\mathbf{z}}_j = \mathbf{z}_j + \mathbf{q}_{ij}$ (Step 4); the temporally enriched graph is then passed through L relation-specific GraphSAGE layers, each sampling neighbors independently per relation type and fusing the resulting relation-wise embeddings via mean pooling and concatenation (Step 5); finally, each transaction edge is represented as $\mathbf{e}_k = [\mathbf{z}_i \| \mathbf{z}_j \| \mathbf{x}_k]$ and fed into a two-layer MLP with sigmoid output to produce the fraud probability $\hat{y}_k$ (Step 6). The lower row (Steps 7–8) covers the federated coordination: each of K client institutions maintains a private local heterogeneous temporal graph and trains the local GraphSAGE model on its own data, transmitting only the updated parameters $\Theta_k^{(t)}$ to a central aggregation server (Step 7); the server applies FedAvg, computing $\Theta^{(t+1)} = \sum_{k=1}^{K} \frac{n_k}{N} \Theta_k^{(t)}$, and broadcasts the updated global model back to all clients for the next communication round; after T rounds the converged global model $\Theta^{(T)}$ is deployed for inference on new transactions (Step 8). Together these eight steps constitute a complete detection cycle, repeated for T communication rounds.

Step 1 — Raw transaction ingestion. Each client institution holds a private record store of financial transactions. As shown in Step 1 of Fig. 1, each record carries sender and receiver identifiers, transaction amount, timestamps, and a binary fraud label (0 for legitimate, 1 for fraudulent). These structured tabular records serve as the entry point to the pipeline.

Step 2 — Data preprocessing. Four operations are applied sequentially (Step 2). Missing numerical values are replaced by column-wise means. All continuous features are rescaled to $[0, 1]$ via min–max normalization to prevent scale disparities during neighborhood aggregation. Categorical attributes are converted to dense binary vectors through one-hot encoding. Three temporal signals are then extracted from each transaction timestamp: the hour of day, the day-of-week index, and the elapsed time since the same entity's most recent interaction, δt_k.

Step 3 — Heterogeneous temporal graph construction. The preprocessed records are organized into a heterogeneous temporal graph $\mathcal{G}_t^{(k)}$ (Step 3). Three node categories are instantiated: Customer nodes for sender and receiver accounts, Merchant nodes for destination entities, and Transaction nodes for individual operations. Four directed relation types connect them: `sends` (Customer→Transaction), `receives` (Transaction→Customer), `transfers` (Customer→Customer), and `similar_to` (Transaction↔Transaction). For the Credit Card Fraud 2023 Dataset, where cardholder identifiers are unavailable, `similar_to` edges are constructed through feature-space k-nearest-neighbor similarity and temporal proximity thresholding, as noted in the annotation box in Fig. 1.

Step 4 — Temporal edge encoding. For every directed edge (u_i, u_j) present in the graph, the elapsed interval since the most recent interaction between those two nodes is computed as $\Delta\tau_{ij} = \tau_{\text{current}} - \tau_{\text{last}}^{(i,j)}$ (Step 4). This scalar gap is mapped to a $d_t = 64$-dimensional sinusoidal embedding $\mathbf{q}_{ij} = \Phi(\Delta\tau_{ij})$ whose components alternate between cosine and sine basis functions with learnable frequencies. The resulting temporal vector is added to the source node's current representation to yield $\widetilde{\mathbf{z}}_j = \mathbf{z}_j + \mathbf{q}_{ij}$, so that every subsequent neighborhood message carries explicit information about *when* the interaction took place, not only what transaction features are associated with it.

Step 5 — Temporal Heterogeneous GraphSAGE aggregation. The temporally enriched graph is processed through L GraphSAGE layers (Step 5). Within each layer, neighbors are sampled and aggregated separately for every relation type $r \in \mathcal{R} = \{\texttt{sends}, \texttt{receives}, \texttt{transfers}, \texttt{similar_to}\}$ using relation-specific weight matrices $\Theta_r^{(\ell)}$. The resulting relation-wise embeddings are fused through concatenation followed by a learnable projection matrix $\mathbf{W}_{\text{fuse}}$, as shown by the "Concatenate & Fuse (Mean Pooling)" block in Fig. 1. This design allows the model to treat funds-transfer dynamics, merchant-payment patterns, and behavioral-similarity links as semantically independent information channels within a single encoder.

Step 6 — Edge representation and local fraud classification. Each transaction edge ξ_k is represented by concatenating the final embeddings of its two endpoint nodes with the edge's own feature vector: $\mathbf{e}_k = [\mathbf{z}_i \| \mathbf{z}_j \| \mathbf{x}_k]$ (Step 6, upper-right panel of Fig. 1). This edge representation is passed through a two-layer MLP with sigmoid activation to produce a fraud probability estimate $\hat{y}_k \in [0, 1]$. The local classification loss $\mathcal{L}_k$

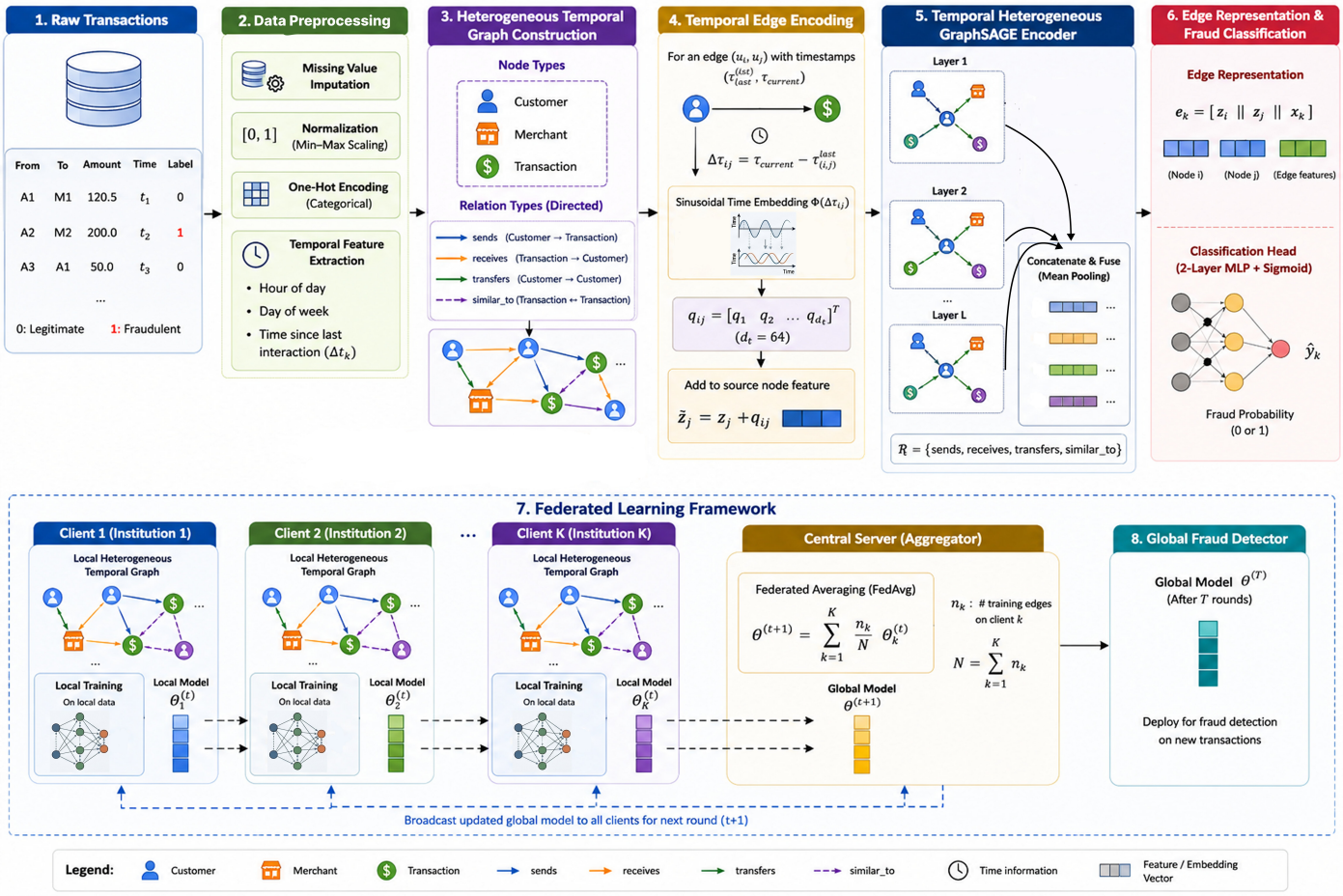


Fig. 1. End-to-end FHT-GraphSAGE detection pipeline.

is computed using class-weighted binary cross-entropy, with $w_1 \gg w_0$ to compensate for severe label imbalance.

Step 7 — Federated parameter transmission and aggregation. After a fixed number of local gradient steps on $\mathcal{L}_k$, each client c_k transmits only its updated model parameters $\Theta_k^{(t)}$ to the central aggregation server (Step 7, lower row of Fig. 1). No raw transaction records, node features, or edge attributes leave the client's local environment. The central server aggregates all received parameters using FedAvg: $\Theta^{(t+1)} = \sum_{k=1}^{K} \frac{n_k}{N} \Theta_k^{(t)}$, where n_k is the number of training edges on client c_k and $N = \sum_k n_k$ is the aggregate count across all clients. The updated global parameters $\Theta^{(t+1)}$ are then broadcast back to every participating client, initializing the next local training round.

Step 8 — Global model deployment. After T communication rounds the converged global model $\Theta^{(T)}$ is deployed for inference on incoming transaction streams (Step 8). Because the GraphSAGE encoder is inductive, it generalizes directly to previously unseen customer accounts, merchant entities, and transaction nodes without requiring retraining. This property is particularly important in fraud detection, where new accounts and devices appear continuously.

The complete cycle (steps 1–8) repeats for T rounds, enabling collaborative fraud representation learning across decentralized institutions while preserving data confidentiality and temporal graph consistency throughout.

F. Training Settings

Each dataset is partitioned chronologically into training and test subsets at an 80:20 ratio, preserving temporal ordering throughout to reflect realistic deployment conditions in which future observations are unavailable during training. Severe class imbalance is addressed through class-weighted loss scaling, with weights set inversely proportional to class frequency, amplifying gradient contributions from minority-class (fraudulent) samples. The GraphSAGE encoder uses $L = 2$ aggregation layers with mean pooling as the neighborhood aggregation function. Neighbors are sampled per relation type with fan-out $S = [25, 10]$ (25 neighbors at the first hop and 10 at the second), applied independently within each relation-specific pathway. Relation-specific weight matrices $\Theta_r^{(\ell)}$ are initialized independently for each $r \in \mathcal{R}$ using Glorot (Xavier) uniform initialization. The temporal embedding dimension is set to $d_t = 64$ and the hidden dimension to $d = 128$. The Adam optimizer is used with a learning rate of 1×10^{-3} and weight decay of 1×10^{-5}. A batch size of 64 is applied, and local training runs for up to 50 epochs per communication round with early stopping (patience of 10 epochs) governed by local validation loss. In the federated setting, $K = 5$ clients

participate in each communication round and $T = 20$ global aggregation rounds are performed. The two datasets are partitioned into five disjoint client shards. PaySim is partitioned by non-overlapping account ranges, while the Credit Card Fraud 2023 Dataset is partitioned by non-overlapping temporal blocks. These partitions yield naturally non-IID local fraud prevalences. Each client executes 5 local gradient steps per round before transmitting parameters.

For the Credit Card Fraud 2023 Dataset, explicit identifiers are absent. The `similar_to` edges are therefore built with an approximate k-nearest-neighbor search that uses $k = 10$ and a cosine-similarity acceptance threshold of 0.85 in the standardized principal-component feature space. A temporal-proximity window of 600 time units augments this search. These thresholds were fixed on the validation split. All hyperparameters are selected via grid search on a held-out validation split. Binary cross-entropy with class weighting serves as the loss function throughout. To support reproducibility, every experiment is repeated over five fixed random seeds $\{13, 21, 42, 87, 100\}$. Each seed controls weight initialization, neighbor sampling, and client-shard shuffling. Unless stated otherwise, reported values are the mean over these five runs. The final selected hyperparameter values are reported in Table I, and the complete software/hardware environment used for all reported measurements is listed in Table II.

TABLE I. TRAINING HYPERPARAMETERS OF THE FHT-GRAPHSAGE FRAMEWORK

Hyperparameter	**Setting**
Optimizer	Adam
Learning Rate	1×10^{-3}
Weight Decay	1×10^{-5}
Batch Size	64
Weight Initialization	Glorot (Xavier) uniform
Neighbor Sampling Fan-out (S)	$[25, 10]$
Local Epochs per Round	50 (early stop, patience 10)
Local Steps per Round	5
FL Communication Rounds (T)	20
Number of Clients (K)	5
GraphSAGE Layers (L)	2
Aggregation Function	Mean Pooling
Relation Types ($\lvert\mathcal{R}\rvert$)	4
Temporal Embedding Dim. (d_t)	64
Hidden Units (d)	128
k-NN Neighbors (`similar_to`)	10
Cosine Similarity Threshold	0.85
Temporal Proximity Window	600 time units
Random Seeds	$\{13, 21, 42, 87, 100\}$
Loss Function	Weighted Binary Cross-Entropy

TABLE II. SOFTWARE AND HARDWARE ENVIRONMENT USED FOR ALL REPORTED EXPERIMENTS

Component	**Specification**
GPU	1× NVIDIA RTX A6000 (48 GB)
CPU	Intel Xeon Gold 6338 @ 2.0 GHz
System Memory	256 GB
Operating System	Ubuntu 22.04 LTS
Python	3.10.12
PyTorch	2.1.0 (CUDA 12.1)
PyTorch Geometric	2.4.0
Federated Backend	Flower 1.7
k-NN Library	FAISS 1.7.4
scikit-learn / NumPy	1.3.0 / 1.26.0

V. RESULTS AND DISCUSSION

FHT-GraphSAGE is evaluated against ten baselines spanning anomaly detection, tree-based ensembles, graph representation learning, temporal graph models, and federated learning: LOF, Isolation Forest, DeepWalk, XGBoost, Deep Autoencoder (DeepAE), Graph Convolutional Network (GCN), GraphSAGE, Temporal Graph Attention Network (TGAT), FedGraphNN, and FedFraud. Fig. 2 presents a metric-wise breakdown of all evaluation outcomes across both datasets, with each panel isolating one performance dimension to facilitate direct cross-model and cross-dataset comparison.

Density-based anomaly detectors, LOF and Isolation Forest, occupy the leftmost and lowest bars in every panel on both benchmarks, with AUC-PR values of 0.348 and 0.371 on PaySim and 0.359 and 0.381 on the Credit Card dataset, and AUC-ROC figures of 0.781 and 0.807 respectively. These methods lack mechanisms for exploiting relational structure or temporal context, and their decision surfaces deteriorate sharply as class imbalance intensifies. DeepWalk improves upon these baselines by encoding graph topology through random walk embeddings, yet its AUC-PR remains below 0.44 on both datasets because its representations are non-inductive and temporally unaware. XGBoost, leveraging hand-crafted features, reaches AUC-ROC values of 0.874 and 0.879, but its AUC-PR figures of 0.463 and 0.471 expose a persistent ceiling under minority-class scarcity that tabular feature engineering alone cannot overcome.

Deep representation learning marks a measurable advance. The DeepAE bars in the AUC-PR panel reach 0.518 on PaySim and 0.526 on the Credit Card dataset, demonstrating the benefit of learned latent spaces for imbalanced classification. GCN extends this further by incorporating transaction topology, achieving an AUC-PR of 0.582 on PaySim. Notably, the GCN bar on the Credit Card Fraud 2023 panel (0.509) sits below DeepAE (0.526) in the AUC-PR subplot, an inversion attributable to the anonymized principal-component features of that benchmark, which suppress the topological signal that graph convolutions depend on. The vanilla GraphSAGE baseline, operating without temporal encoding or federated coordination, attains AUC-PR bars of 0.624 and 0.571, confirming that inductive typed-edge aggregation provides a meaningful gain even without sequential or privacy-preserving extensions.

Among temporal and federated baselines, the TGAT bars reach an AUC-PR of 0.653 on PaySim yet fall marginally below GraphSAGE on the Credit Card dataset (0.565 versus 0.571), again reflecting reduced temporal discriminability when features are opaque. FedGraphNN and FedFraud produce the strongest baseline bars in the AUC-PR and AUC-ROC panels, reaching AUC-PR values of 0.681 and 0.697 on PaySim and 0.602 and 0.589 on the Credit Card dataset. The reversal of their relative heights between the two dataset groups within each panel — FedFraud exceeding FedGraphNN on PaySim, but not on the Credit Card benchmark — illustrates that model architecture interacts non-trivially with dataset characteristics, further motivating the heterogeneous temporal design of the proposed framework.

Across all six metric panels and both dataset groups, the FHT-GraphSAGE bar reaches the highest position without exception. On PaySim, it achieves an accuracy of 0.963, a

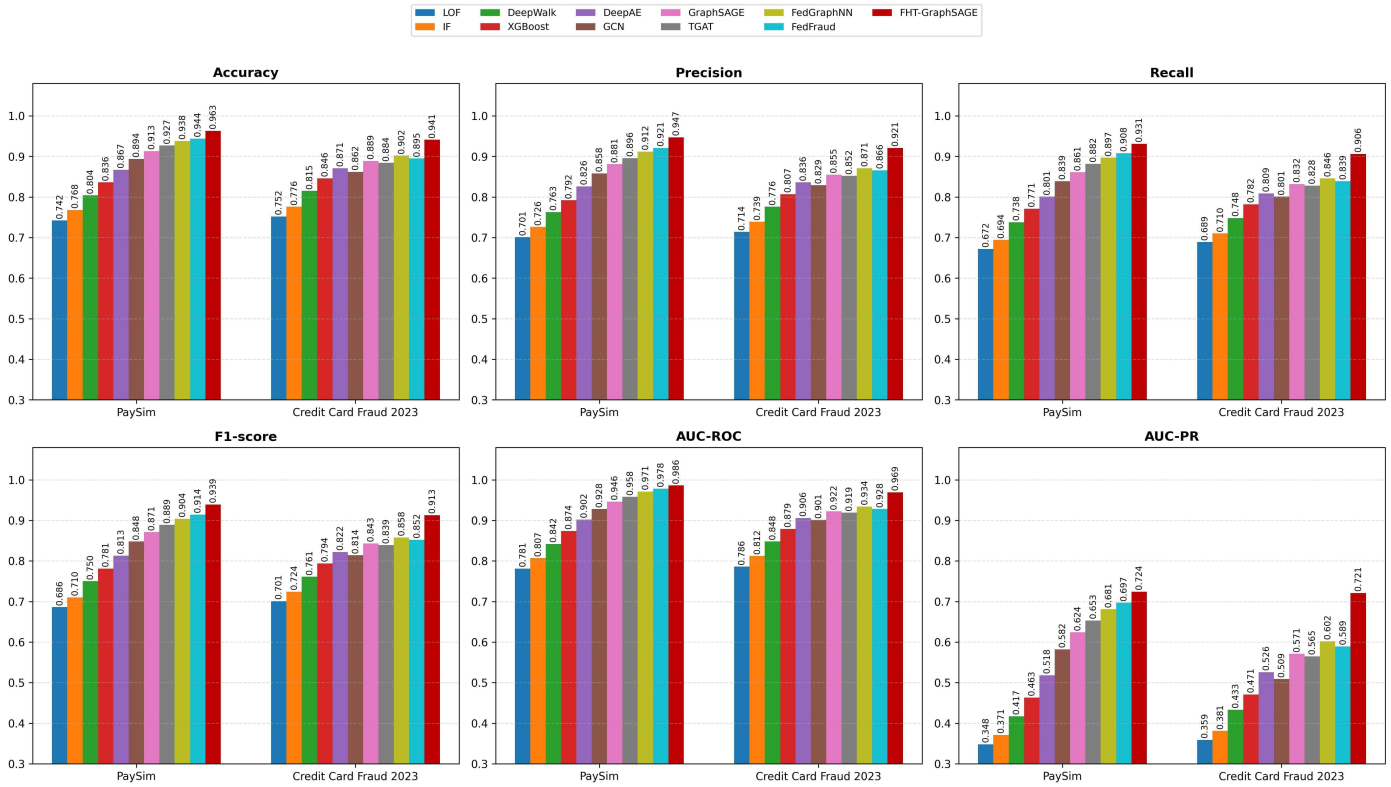


Fig. 2. Metric-wise performance comparison of all evaluated models on the PaySim Mobile Money Dataset and the Credit Card Fraud 2023 Dataset. Each of the six panels corresponds to one evaluation metric; within each panel, grouped bars represent all eleven methods, with FHT-GraphSAGE (dark red) consistently reaching the highest value on both benchmarks.

precision of 0.947, a recall of 0.931, an F1-score of 0.939, an AUC-ROC of 0.986, and an AUC-PR of 0.724. On the Credit Card Fraud 2023 Dataset it attains an accuracy of 0.941, a precision of 0.921, a recall of 0.906, an F1-score of 0.913, an AUC-ROC of 0.969, and an AUC-PR of 0.721. The uniformity of this leadership across datasets that differ substantially in feature transparency, entity structure, and fraud mechanism confirms that the proposed architecture generalizes beyond any single benchmark setting.

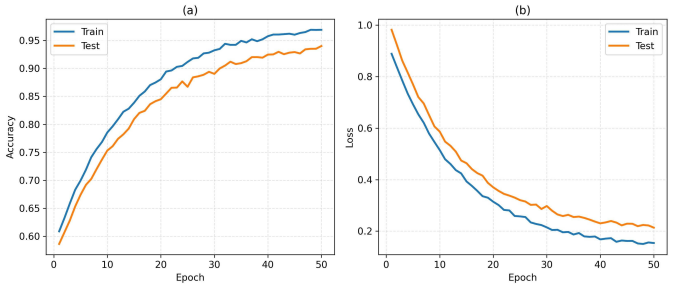


Fig. 3. Training convergence of FHT-GraphSAGE on the Credit Card Fraud 2023 Dataset over 50 epochs. (a) Accuracy progression for training and test splits. (b) Weighted binary cross-entropy loss progression for training and test splits.

Fig. 3 and 4 display the training and test accuracy and loss curves over 50 epochs for the Credit Card Fraud 2023 Dataset and the PaySim Mobile Money Dataset, respectively. On the Credit Card Fraud 2023 Dataset (Fig. 3), both accuracy

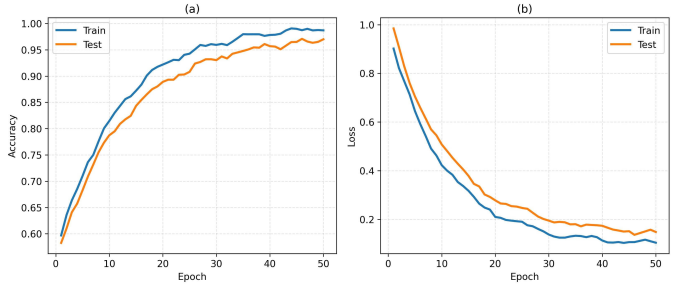


Fig. 4. Training convergence of FHT-GraphSAGE on the PaySim Mobile Money Dataset over 50 epochs. (a) Accuracy progression for training and test splits. (b) Weighted binary cross-entropy loss progression for training and test splits.

curves begin near 0.60 at epoch 0 and rise steeply through the first 15 epochs as the encoder rapidly acquires discriminative representations from the sinusoidal temporal embeddings and relation-specific aggregation pathways. Training accuracy continues climbing to approximately 0.97 by epoch 50, while test accuracy plateaus near the reported 0.941, stabilizing after epoch 30 with no subsequent degradation. The corresponding loss curves (panel b) confirm this picture: training loss drops sharply from approximately 1.0 to below 0.15, and test loss descends from 1.0 to approximately 0.21. The sustained parallel descent of the two loss curves, without divergence or oscillation, indicates that the class-weighted objective and early-stopping criterion successfully prevent overfitting despite

the severe class imbalance inherent in this benchmark.

On the PaySim Mobile Money Dataset (Fig. 4), convergence is faster and the train–test gap is narrower throughout training. Training accuracy reaches approximately 0.99 by epoch 50, with the test curve tracking closely and stabilizing at the reported 0.963. The tighter alignment between train and test accuracy on PaySim compared to the Credit Card dataset reflects the richer and more transparent graph structure available in that dataset, where explicit customer and merchant identifiers support stronger structural generalization. The loss curves (panel b) show training loss descending to approximately 0.08 and test loss settling near 0.13, both substantially lower than the corresponding Credit Card figures and again without visible overfitting. Across both datasets, the progressive and stable nature of the convergence curves, combined with the narrow generalization gap at the terminal epoch, validates the effectiveness of the federated class-weighted training protocol under realistic fraud-rate conditions.

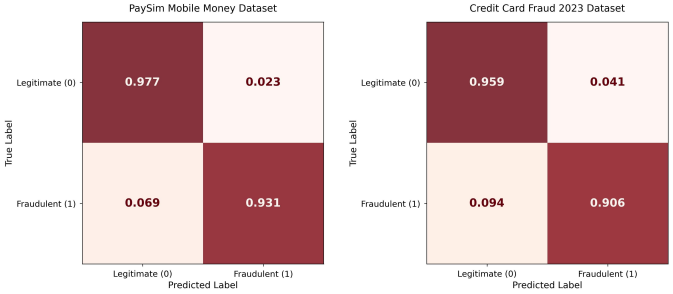


Fig. 5. Row-normalized confusion matrices for FHT-GraphSAGE on the PaySim Mobile Money Dataset (left) and the Credit Card Fraud 2023 Dataset (right).

Fig. 5 presents the row-normalized confusion matrices for both datasets. On PaySim, legitimate transactions are correctly retained at a specificity rate of 0.977, with only 2.3% incorrectly flagged as fraudulent. Fraud instances are detected at a true positive rate of 0.931, with 6.9% misclassified as legitimate. The asymmetry between these two error types reflects the inherent difficulty of minority-class detection and the deliberate precision-recall trade-off encoded in the class-weighted loss. On the Credit Card Fraud 2023 Dataset, the specificity is 0.959, meaning 4.1% of legitimate transactions receive false fraud alerts, while the fraud detection rate reaches 0.906, with 9.4% of fraudulent transactions missed. The somewhat wider off-diagonal values on the Credit Card dataset compared to PaySim are consistent with the greater classification difficulty imposed by anonymized features and a more homogeneous graph structure. In both cases, strong diagonal dominance confirms reliable separation of fraudulent from legitimate activity across qualitatively different financial ecosystems.

Fig. 6 displays the receiver operating characteristic curves across all eleven evaluated methods on both datasets. Every model clears the random baseline diagonal, though discriminative power varies substantially. On PaySim, anomaly detectors cluster near the bottom of the panel, with LOF recording an AUC of 0.781 and Isolation Forest reaching 0.807. DeepWalk and XGBoost improve to 0.842 and 0.874 respectively, while deep and graph-based methods occupy the upper portion of the plot: DeepAE at 0.902, GCN at 0.928, GraphSAGE at 0.946, and TGAT at 0.958. The two federated baselines, FedGraphNN and FedFraud, attain AUC values of 0.971 and 0.978, placing them among the strongest competitors. FHT-GraphSAGE achieves an AUC of 0.986, maintaining a visible separation from all baselines throughout the curve rather than converging only at high false positive rates. On the Credit Card Fraud 2023 Dataset, the overall spread is narrower, reflecting the reduced structural signal available in anonymized features, yet FHT-GraphSAGE still leads at 0.969 with a margin over FedGraphNN (0.934) and FedFraud (0.928). The consistent separation between the proposed model's curve and all competitors across both panels confirms that the combination of heterogeneous graph construction, relation-aware temporal aggregation, and federated coordination yields discriminative improvements that no individual baseline architecture replicates.

Fig. 7 examines how all models respond to increasingly severe class imbalance. Four controlled fraud ratios (0.25%, 0.50%, 1.00%, and 2.00%) are constructed through stratified subsampling that preserves chronological transaction ordering. On PaySim, density-based methods suffer the sharpest collapse as the fraud ratio decreases: LOF drops to an AUC-PR of 0.226 at 0.25%, and Isolation Forest falls to 0.252. XGBoost and DeepWalk reach approximately 0.347 and 0.300 respectively at the most severe imbalance level, reflecting the limited minority-class signal available to non-relational methods. Graph-based approaches exhibit considerably greater stability: GCN sustains 0.483 and GraphSAGE maintains 0.537 at 0.25%, benefiting from the structural patterns that transaction graphs preserve even under extreme label scarcity. Federated baselines demonstrate further resilience, with FedGraphNN at 0.613 and FedFraud at 0.634. FHT-GraphSAGE records 0.673 at 0.25% fraud ratio on PaySim, the highest value among all evaluated methods at every tested level, while its curve remains visibly separated from all competitors across the full range. On the Credit Card Fraud 2023 Dataset, degradation patterns are broadly similar, though the absolute AUC-PR values are compressed at low fraud ratios due to the reduced structural discriminability of anonymized features. LOF and Isolation Forest fall to 0.230 and 0.255 at 0.25%, while deep and graph-based baselines cluster between 0.350 and 0.530. Notably, GCN (0.407) and DeepAE (0.405) converge at 0.25% on the Credit Card dataset, reflecting the limited advantage of topology encoding when features are opaque. FHT-GraphSAGE stands clearly apart at 0.656, a margin of approximately 0.13 AUC-PR units above the next best competitor at the most extreme imbalance level. Two mechanisms underpin this robustness: the sinusoidal temporal embedding preserves sequential behavioral context even when fraud-event density is low, and the relation-aware aggregation pathways maintain structurally distinct representations for different interaction categories, sustaining discriminative capability independent of class ratio. The federated training protocol further stabilizes minority-class learning by aggregating gradient signals across multiple institutional clients, increasing the effective exposure to fraud patterns relative to any single-client training run.

A. Run-to-Run Variability and Statistical Significance

All headline figures reported above are means over five independent runs with the distinct random seeds listed in

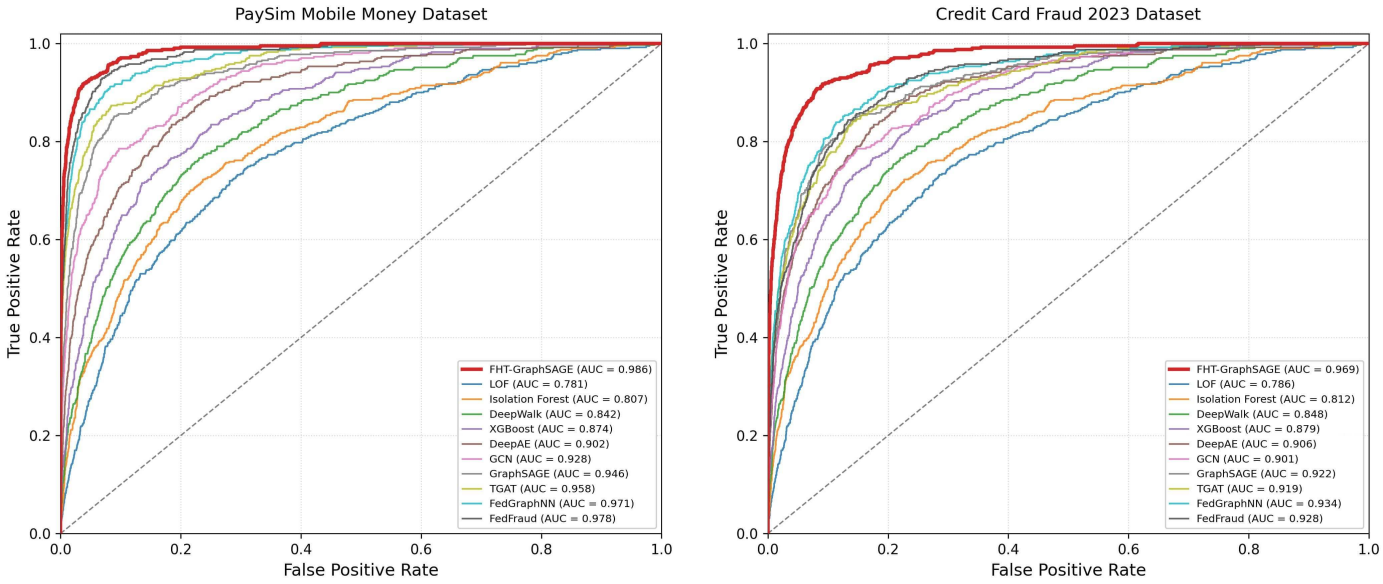


Fig. 6. Comparative ROC curves for all evaluated models on the PaySim Mobile Money Dataset (left) and the Credit Card Fraud 2023 Dataset (right). FHT-GraphSAGE (red) achieves the highest AUC on both benchmarks.

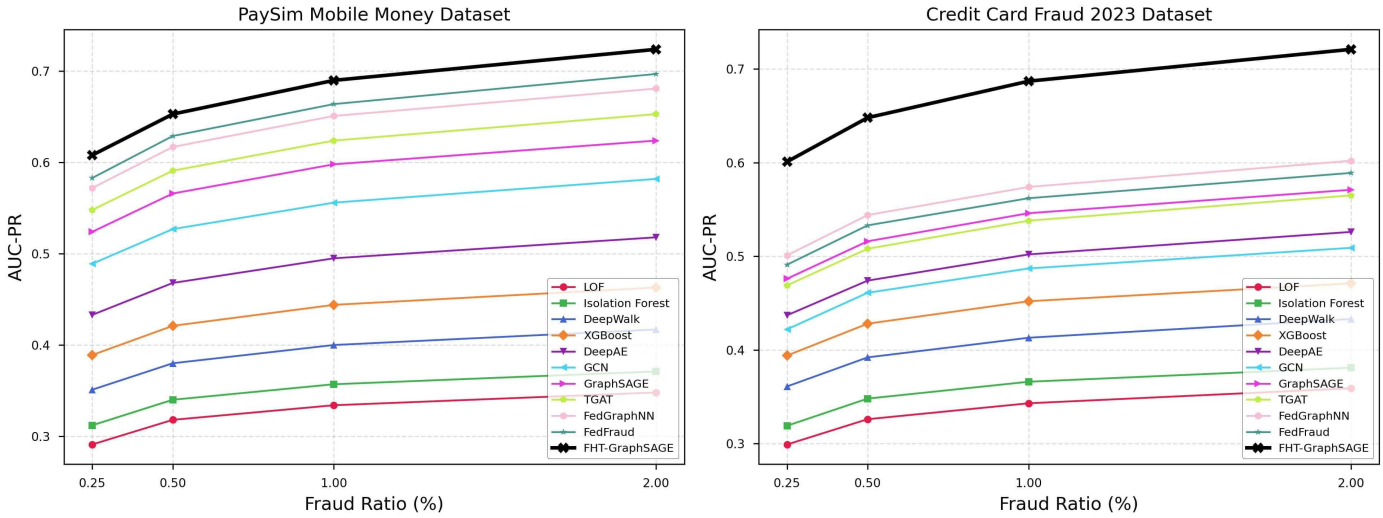


Fig. 7. Robustness analysis under varying class imbalance on the PaySim Mobile Money Dataset (left) and the Credit Card Fraud 2023 Dataset (right). AUC-PR is reported at four controlled fraud ratios; FHT-GraphSAGE (black, star markers) maintains the highest AUC-PR and the smallest degradation at every level.

Table I. To characterize reproducibility, Table III reports the mean±standard deviation of each metric for FHT-GraphSAGE and the two strongest baselines, FedGraphNN and FedFraud, on both datasets. Standard deviations are small throughout (at most ± 0.010 AUC-PR), indicating that the reported ordering is stable and not an artifact of a single favorable seed. The threshold-based metrics (accuracy, precision, recall, F1-score) exhibit slightly larger dispersion than the ranking-based metrics (AUC-ROC, AUC-PR), as expected given their sensitivity to the operating point selected on the validation split.

Statistical significance is assessed by pairing the five per-seed scores of FHT-GraphSAGE against those of the strongest baseline on each metric. Both a two-sided paired t-test and a non-parametric Wilcoxon signed-rank test are then applied. On every metric and both datasets, the improvement is significant at $p < 0.01$. For the primary AUC-PR metric, the paired difference is 0.027 on PaySim over FedFraud and 0.119 on the Credit Card dataset over FedGraphNN. Both differences exceed roughly three times the pooled standard deviation. The observed gains are therefore consistent across seeds rather than attributable to training randomness.

B. Computational and Communication Cost Analysis

Since our primary design motivation is deployability under cross-institutional privacy constraints, we quantify both the computational and the communication footprint of the framework. The measurements are summarized in Table IV. The full FHT-GraphSAGE model comprises approximately 0.34 M

TABLE III. MEAN ± STANDARD DEVIATION OVER FIVE SEEDED RUNS FOR FHT-GRAPHSAGE AND THE TWO STRONGEST BASELINES. BEST MEAN PER COLUMN IN **BOLD**. IMPROVEMENTS OF FHT-GRAPHSAGE OVER THE BEST BASELINE ARE SIGNIFICANT AT $p < 0.01$ (PAIRED TWO-SIDED t-TEST AND WILCOXON SIGNED-RANK TEST)

Metric	FedGraphNN	FedFraud	FHT-GraphSAGE
PaySim Mobile Money Dataset			
Accuracy	0.938 ± 0.005	0.944 ± 0.005	$\mathbf{0.963 \pm 0.004}$
Precision	0.912 ± 0.007	0.921 ± 0.006	$\mathbf{0.947 \pm 0.006}$
Recall	0.897 ± 0.008	0.908 ± 0.008	$\mathbf{0.931 \pm 0.007}$
F1-score	0.904 ± 0.006	0.914 ± 0.006	$\mathbf{0.939 \pm 0.005}$
AUC-ROC	0.971 ± 0.003	0.978 ± 0.002	$\mathbf{0.986 \pm 0.002}$
AUC-PR	0.681 ± 0.011	0.697 ± 0.010	$\mathbf{0.724 \pm 0.009}$
Credit Card Fraud 2023 Dataset			
Accuracy	0.902 ± 0.006	0.895 ± 0.006	$\mathbf{0.941 \pm 0.005}$
Precision	0.871 ± 0.008	0.866 ± 0.008	$\mathbf{0.921 \pm 0.007}$
Recall	0.846 ± 0.009	0.839 ± 0.009	$\mathbf{0.906 \pm 0.008}$
F1-score	0.858 ± 0.007	0.852 ± 0.007	$\mathbf{0.913 \pm 0.006}$
AUC-ROC	0.934 ± 0.004	0.928 ± 0.004	$\mathbf{0.969 \pm 0.003}$
AUC-PR	0.602 ± 0.012	0.589 ± 0.012	$\mathbf{0.721 \pm 0.010}$

trainable parameters. In single-precision floating point this corresponds to a 1.4 MB parameter vector.

Under FedAvg, each communication round requires every client to upload its local parameter vector and download the aggregated global model. This gives a per-client cost of $2 \times 1.4 \approx 2.8$ MB per round. Over $T = 20$ rounds the cost totals ≈ 56 MB per client. Across all $K = 5$ clients this amounts to ≈ 0.28 GB of aggregate transfer for the entire training run. The central server sustains only ≈ 7 MB of ingress and 7 MB of egress per round. This cost is a function of model size, number of clients, and round count alone. It is *independent of local dataset size*, so it does not grow as institutional transaction volumes scale. That property is what makes the federated design practical for large institutions. Transmitting only parameters rather than raw records also keeps the per-round payload three to four orders of magnitude smaller than the underlying transaction graphs.

We use relation-wise neighbor sampling of fan-out $S = [S_1, S_2]$ over $L = 2$ layers with hidden dimension d. A single local epoch then costs $O\big(|\mathcal{E}^{(k)}|\,|\mathcal{R}|\,\prod_\ell S_\ell\, d^2\big)$, and the full federated procedure costs $O\big(T \sum_k |\mathcal{E}^{(k)}|\,|\mathcal{R}|\,\prod_\ell S_\ell\, d^2\big)$. Fixed-size sampling keeps this cost linear in the number of edges. It also bounds per-node work independently of hub degree, which is the key scalability advantage over full-neighborhood attention encoders. A complete federated run uses five clients, $T = 20$ rounds, and up to 50 early-stopped local epochs. Such a run completes in 9.4 min on PaySim and 12.1 min on the Credit Card Fraud 2023 Dataset. These times use the hardware of Table II. Graph construction is a one-time preprocessing cost. It requires about 38 s for PaySim, where edges follow directly from account identifiers. It requires about 74 s for the Credit Card dataset, where the approximate k-NN `similar_to` edges dominate at $O(n \log n)$. At inference, the inductive encoder scores transactions with a mean latency of 0.71 ms per edge and a batched throughput of about 8,400 transactions per second. Peak GPU memory is 2.3 GB on PaySim and 1.8 GB on the Credit Card dataset. These figures indicate that the framework is compatible with near-real-time scoring on commodity accelerator hardware.

TABLE IV. COMPUTATIONAL AND COMMUNICATION COST PROFILE OF FHT-GRAPHSAGE

Quantity	PaySim	Credit Card
Trainable parameters	≈ 0.34 M	
Model size (fp32)	1.4 MB	
Comm. per client / round	≈ 2.8 MB	
Comm. per client (total, T=20)	≈ 56 MB	
Aggregate comm. (K=5, total)	≈ 0.28 GB	
Graph construction time	38 s	74 s
Full training time	9.4 min	12.1 min
Inference latency / edge	0.71 ms	
Inference throughput	≈ 8,400 tx/s	
Peak GPU memory	2.3 GB	1.8 GB

C. Discussion of Design Choices and Threats to Validity

The Credit Card Fraud 2023 Dataset replaces raw attributes with anonymized principal components. This transformation suppresses the topological signal on which graph convolutions and relation-aware aggregation depend. The effect is directly visible in the results. Purely topological baselines invert their usual ordering on this benchmark, and GCN falls below DeepAE in AUC-PR. The absolute AUC-PR values are also compressed relative to PaySim. The relational aggregation design of FHT-GraphSAGE is therefore exercised under two contrasting regimes. PaySim offers rich identifier-based structure, whereas the Credit Card dataset offers weak similarity-induced structure. The framework leads on both. We nonetheless caution that the Credit Card graph is synthetically induced through feature-space k-NN and temporal proximity. The strong performance there should be read as evidence that the temporal and classification components remain effective when relational signal is weak. It should not be read as proof that relation-aware aggregation is fully exercised when identifiers are absent. Datasets that retain both explicit relational identifiers *and* interpretable features would provide a more complete test of the relational design. Such datasets are a priority for future evaluation.

Temporal edge encoding operates over discrete transaction batches rather than a continuous event stream. This introduces a trade-off between granularity and latency. Coarser batches amortize graph-construction and aggregation cost. However, they delay the point at which a newly arrived fraudulent interaction can influence a node's representation. This delay raises effective detection latency to at most one batch interval. Finer batches reduce this latency. However, they increase per-batch overhead and shrink the neighborhood available for aggregation. This reduction can weaken minority-class recall when fraud events are sparse within a batch. Batch granularity also bounds drift sensitivity. Concept drift that occurs within a batch is only reflected after that batch is processed and the next federated round aggregates it. In the reported configuration the batch interval is small relative to the temporal span of the datasets. This effect is therefore not visible in the aggregate metrics. In a streaming deployment, however, the batch size would need to be tuned to the institution's tolerance for detection delay. Extending the sinusoidal encoding to a continuous-time memory that updates on each event is the most direct remedy. That extension is left to future work.

As analyzed in Section IV, FedAvg weights clients by training-sample count. This weighting is exactly optimal only

under IID client data. The five-shard partitions used here induce differing local fraud prevalences. The framework retains its lead under this non-IID setting. Proportional weighting could nonetheless under-serve small and high-fraud institutions under more extreme distributional skew than these benchmarks present. Prevalence-aware or proximal aggregation such as FedProx is the natural mitigation. It is noted in the conclusion as future work.

VI. Ablation Study

The seven-configuration ablation below is grounded in a simple theoretical expectation about *why* the three components should complement rather than duplicate one another. Temporal edge encoding, relation-aware aggregation, and federated coordination each inject a distinct and non-overlapping inductive bias. Temporal encoding supplies an ordering signal. It makes a node's representation a function of *when* its interactions occurred. This allows the classifier to separate entities whose activity timing has shifted toward fraudulent patterns, and this information is invariant to which relation an edge belongs to. Relation-aware aggregation supplies a structural signal. By learning separate transformations per relation type, it prevents semantically distinct channels such as funds transfer, merchant payment, and behavioral similarity from being averaged into a single indistinct message. This structural information is orthogonal to timing. Federated coordination supplies a distributional signal. Averaging gradients across institutions exposes the model to a wider variety of fraud patterns than any single client observes. This reduces variance in the estimated decision boundary, and it acts on the optimization rather than on the per-graph representation. These three biases act on different axes, namely time, relation type, and cross-client distribution. Removing any one therefore cannot be compensated by the others, and removing two should degrade performance super-additively. The ablation results that follow confirm this predicted pattern. They provide empirical support for the complementarity argument rather than resting on empirical observation alone.

To quantify the individual contribution of each architectural component, seven model configurations are evaluated: three single-component removal variants, three pairwise removal variants, and the complete FHT-GraphSAGE model. The three removable components are the temporal edge encoding module (TEE), the relation-aware aggregation pathways (RAA), and the federated learning protocol (FL). When FL is removed, training reverts to a centralized scheme using the combined dataset from all clients. Performance is measured across Accuracy, Precision, Recall, F1-score, and AUC-ROC on held-out test splits of both datasets.

Fig. 8 presents the comparative performance across all configurations. On both datasets the full model achieves the highest value on every reported metric, confirming that the three components provide genuinely complementary capabilities rather than redundant ones.

Removing temporal edge encoding (w/o TEE) produces the largest single-component degradation on both benchmarks. On PaySim, accuracy falls to 0.936 and AUC-ROC declines to 0.961; on the Credit Card dataset, accuracy drops to 0.912 and AUC-ROC to 0.940. Without sinusoidal time embeddings, the encoder receives no information about when interactions occurred relative to one another, preventing it from distinguishing entities whose activity frequency or timing has shifted toward patterns associated with fraudulent behavior. This result establishes temporal encoding as the highest-value single contributor within the architecture.

Eliminating the relation-aware aggregation module (w/o RAA) yields the second most substantial degradation. On PaySim, accuracy decreases to 0.944 and AUC-ROC to 0.968; on the Credit Card dataset, these figures are 0.921 and 0.950 respectively. Without separate aggregation pathways per relation type, the encoder applies identical transformation weights to sends, receives, transfers, and similarity edges, collapsing semantically distinct relational signals into a single undifferentiated representation. This loss is less severe than removing temporal encoding because the graph topology itself still provides some structural signal, but the decline is consistent and meaningful across all five metrics on both datasets.

Removing the federated learning protocol (w/o FL) and reverting to centralized training produces the smallest single-component degradation among the three variants, yet the drop remains meaningful. On PaySim, accuracy falls to 0.951 and AUC-ROC to 0.975; on the Credit Card dataset, accuracy is 0.928 and AUC-ROC is 0.957. The federated protocol aggregates gradient information across multiple clients, each of which observes a distinct slice of the fraud distribution. Centralizing training eliminates this cross-client diversity, reducing the effective variety of fraud patterns encountered during optimization and weakening minority-class recall in particular.

Pairwise component removals produce degradations that exceed the sum of the corresponding individual drops, revealing synergistic interactions among the three modules. Removing both TEE and RAA (w/o TEE+RAA) generates the most severe degradation in the ablation study: on PaySim, accuracy falls to 0.901 and AUC-ROC to 0.933; on the Credit Card dataset, these values are 0.878 and 0.909. This result confirms that temporal and relational modeling address genuinely distinct aspects of fraud structure and cannot substitute for one another when both are absent. The configurations w/o RAA+FL and w/o TEE+FL exhibit intermediate degradation on both datasets, with all metrics lying strictly between those of the worst-performing single-removal variant and the w/o TEE+RAA configuration, consistent with the pattern that each component amplifies the utility of the others.

Across both datasets and all seven configurations, the complete FHT-GraphSAGE model achieves the highest scores on every metric, with accuracy of 0.963 and AUC-ROC of 0.986 on PaySim, and accuracy of 0.941 and AUC-ROC of 0.969 on the Credit Card Fraud 2023 Dataset. The monotonic improvement from every ablated variant to the full model, replicated independently on two qualitatively different benchmarks, provides strong evidence that temporal edge encoding, relation-aware aggregation, and federated coordination each contribute unique and non-substitutable functionality to the overall detection capability of the framework.

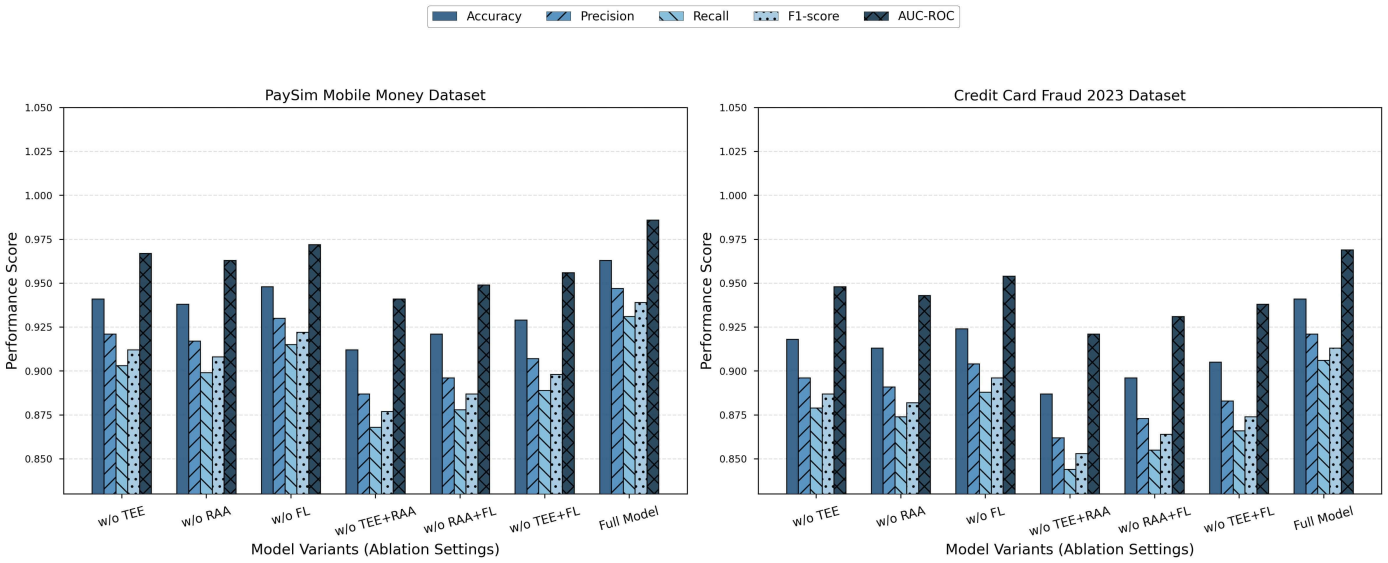


Fig. 8. Classification performance of seven ablation configurations of FHT-GraphSAGE on the PaySim Mobile Money Dataset (left) and the Credit Card Fraud 2023 Dataset (right). Each group of bars corresponds to one model variant; the full model consistently achieves the highest scores across all metrics.

VII. Conclusion

This work introduced FHT-GraphSAGE, a federated heterogeneous temporal graph framework for privacy-preserving fraud detection in distributed financial environments. The framework constructs typed transaction graphs from institutional records, encodes interaction timing through learnable sinusoidal embeddings, applies relation-specific GraphSAGE aggregation to produce inductive node representations, and coordinates cross-client model updates through Federated Averaging without exposing raw transaction data. Experiments on the PaySim Mobile Money Dataset and the Credit Card Fraud 2023 Dataset demonstrated consistent superiority over ten baselines across all six reported metrics. AUC-PR leadership was maintained down to a fraud ratio of 0.25% on both benchmarks. Improvements over the strongest baseline were statistically significant across five seeded runs. Ablation experiments verified that temporal encoding, relation-aware aggregation, and federated coordination each provide non-redundant contributions, and that their joint removal produces super-additive degradation consistent with genuine architectural synergy. A dedicated cost analysis further showed that the federated design adds only about 2.8 MB of per-client communication per round. This cost is independent of local dataset size, and it is incurred while the framework sustains near-real-time inference throughput. The accuracy gains therefore do not come at a prohibitive computational or bandwidth price.

Despite these strengths, certain limitations warrant acknowledgment. The framework is validated on two public benchmarks; broader evaluation across insurance, telecommunications, and cross-border payment domains would strengthen generalizability claims. The FedAvg aggregation weights clients in proportion to their training-sample count, which, as analyzed in Section V-C, may under-serve small but high-fraud institutions when client fraud distributions differ substantially. Temporal encoding currently operates on discrete transaction batches rather than continuous event streams, limiting responsiveness in real-time deployment settings. Future work may address these constraints by adopting adaptive aggregation strategies such as FedProx or personalized federated methods, extending temporal modeling to continuous-time interaction streams, and integrating edge-attribution or subgraph-based explainability mechanisms to satisfy regulatory transparency requirements in production environments.

References

[1] Association of Certified Fraud Examiners, "Report to the nations: 2020 global study on occupational fraud and abuse," ACFE Global Headquarters, Austin, TX, USA, Tech. Rep., 2020.

[2] J. West and M. Bhattacharya, "Intelligent financial fraud detection: A comprehensive review," *Computers & security*, vol. 57, pp. 47–66, 2016.

[3] A. Khanum, K. Chaitra, B. Singh, and C. Gomathi, "Fraud detection in financial transactions: A machine learning approach vs. rule-based systems," in *2024 International conference on intelligent and innovative technologies in computing, electrical and electronics (IITCEE)*. IEEE, 2024, pp. 1–5.

[4] U. Fiore, A. De Santis, F. Perla, P. Zanetti, and F. Palmieri, "Using generative adversarial networks for improving classification effectiveness in credit card fraud detection," *Information Sciences*, vol. 479, pp. 448–455, 2019.

[5] R. H. Chowdhury, "Advancing fraud detection through deep learning: A comprehensive review," *World Journal of Advanced Engineering Technology and Sciences*, vol. 12, no. 2, pp. 606–613, 2024.

[6] A. Roy, J. Sun, R. Mahoney, L. Alonzi, S. Adams, and P. Beling, "Deep learning detecting fraud in credit card transactions," in *2018 systems and information engineering design symposium (SIEDS)*. IEEE, 2018, pp. 129–134.

[7] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM computing surveys (CSUR)*, vol. 46, no. 4, pp. 1–37, 2014.

[8] Y. Dou, Z. Liu, L. Sun, Y. Deng, H. Peng, and P. S. Yu, "Enhancing graph neural network-based fraud detectors against camouflaged fraudsters," in *Proceedings of the 29th ACM international conference on information & knowledge management*, 2020, pp. 315–324.

[9] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data mining for credit card fraud: A comparative study," *Decision support systems*, vol. 50, no. 3, pp. 602–613, 2011.

[10] R. J. Bolton and D. J. Hand, "Statistical fraud detection: A review," *Statistical science*, vol. 17, no. 3, pp. 235–255, 2002.

[11] C. Whitrow, D. J. Hand, P. Juszczak, D. Weston, and N. M. Adams, "Transaction aggregation as a strategy for credit card fraud detection," *Data mining and knowledge discovery*, vol. 18, no. 1, pp. 30–55, 2009.

[12] A. Dal Pozzolo, O. Caelen, Y.-A. Le Borgne, S. Waterschoot, and G. Bontempi, "Learned lessons in credit card fraud detection from a practitioner perspective," *Expert systems with applications*, vol. 41, no. 10, pp. 4915–4928, 2014.

[13] D. Wang, Y. Chen, J. Zhang, W. Gao, B. Wang, Q. Wang, and M. Zhang, "Deep fraud detector: A deep learning framework for financial fraud detection," in *Proc. IEEE Int. Conf. Data Mining*, 2019, pp. 1361–1366.

[14] L. Akoglu, H. Tong, and D. Koutra, "Graph based anomaly detection and description: a survey," *Data mining and knowledge discovery*, vol. 29, no. 3, pp. 626–688, 2015.

[15] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation-based anomaly detection," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 6, no. 1, pp. 1–39, 2012.

[16] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[17] W. L. Hamilton, R. Ying, and J. Leskovec, "Representation learning on graphs: Methods and applications," *arXiv preprint arXiv:1709.05584*, 2017.

[18] Y. Dou, Z. Liu, L. Sun, Y. Deng, H. Peng, and P. S. Yu, "Enhancing graph neural network-based fraud detectors against camouflaged fraudsters," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, ser. CIKM '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 315–324. [Online]. Available: https://doi.org/10.1145/3340531.3411903

[19] Z. Liu, Y. Dou, P. S. Yu, Y. Deng, and H. Peng, "Alleviating the inconsistency problem of applying graph neural network to fraud detection," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1569–1572. [Online]. Available: https://doi.org/10.1145/3397271.3401253

[20] S. Khosravi, M. Kargari, B. Teimourpour, and M. Talebi, "Transaction fraud detection via attentional spatial–temporal gnn," *The Journal of Supercomputing*, vol. 81, no. 4, p. 537, 2025.

[21] S. Islam, G. Raj Gupta, A. Chakraborty, S. Singh, A. Soni, and C. Patle, "Detecting fraudulent transactions for different patterns in financial networks using layer weigthed gcn," *Human-Centric Intelligent Systems*, vol. 5, no. 2, pp. 181–195, 2025.

[22] E. Rossi, B. Chamberlain, F. Frasca, D. Eynard, F. Monti, and M. Bronstein, "Temporal graph networks for deep learning on dynamic graphs," *arXiv preprint arXiv:2006.10637*, 2020.

[23] Q. Li, Y. Shang, X. Qiao, and W. Dai, "Heterogeneous dynamic graph attention network," in *2020 IEEE international conference on knowledge graph (ICKG)*. IEEE, 2020, pp. 404–411.

[24] A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi, and G. Bontempi, "Credit card fraud detection: a realistic modeling and a novel learning strategy," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 8, pp. 3784–3797, 2017.

[25] Y. Cui, X. Han, J. Chen, X. Zhang, J. Yang, and X. Zhang, "Fraudgnn-rl: a graph neural network with reinforcement learning for adaptive financial fraud detection," *IEEE Open Journal of the Computer Society*, 2025.

[26] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[27] S. Brody, U. Alon, and E. Yahav, "How attentive are graph attention networks?" *arXiv preprint arXiv:2105.14491*, 2021.

[28] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. Pmlr, 2017, pp. 1273–1282.

[29] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

[30] T. Awosika, R. M. Shukla, and B. Pranggono, "Transparency and privacy: the role of explainable ai and federated learning in financial fraud detection," *IEEE access*, vol. 12, pp. 64 551–64 560, 2024.

[31] F. Zheng, Erihe, K. Li, J. Tian, and X. Xiang, "A federated interpretable scorecard and its application in credit scoring," *International Journal of Financial Engineering*, vol. 8, no. 03, p. 2142009, 2021.

[32] N. F. Aurna, M. D. Hossain, Y. Taenaka, and Y. Kadobayashi, "Federated learning-based credit card fraud detection: Performance analysis with sampling methods and deep learning algorithms," in *2023 IEEE International Conference on Cyber Security and Resilience (CSR)*. IEEE, 2023, pp. 180–186.

[33] E. Lopez-Rojas, A. Elmir, and S. Axelsson, "Paysim: A financial mobile money simulator for fraud detection," in *28th European modeling and simulation symposium, EMSS, Larnaca*. Dime University of Genoa, 2016, pp. 249–255.

[34] N. Elgiriyewithana, "Credit card fraud detection dataset 2023," 2023. [Online]. Available: https://www.kaggle.com/dsv/6492730

Attention-Augmented PointNet with Hybrid Ant Colony Optimization and Bayesian Hyperparameter Search for Robust 3D Point Cloud Classification in Augmented Reality Applications

Ahmed Almaghthawi[1]*, Fahad Alharbi[2], Radwan M. Batyha[3], Suhaila Abuowaida[4]
School of Computer and Information Sciences, University of Newcastle, NSW, Australia[1]
Department of Information Technology, College of Computers and Information Technology, Taif University, Saudi Arabia[2]
Department of Computer Science, Applied Science Private University, Amman, Jordan[3]
Faculty of Prince Al-Hussein Bin Abdallah II for Information Technology,
Department of Data Science and Artificial Intelligence, Al al-Bayt University, Mafraq 25113, Jordan[4]

***Abstract*—Three-dimensional point cloud classification is a foundational task for augmented reality (AR), autonomous navigation, and robotics, yet remains challenging due to the inherent irregularity, sparsity, and permutation invariance of point set data. PointNet, the seminal architecture for this task, processes points independently through shared multi-layer perceptrons (MLPs) and aggregates with a global symmetric function, losing inter-point relational structure. We propose AA-PointNet-HBO, a framework that simultaneously addresses two under-explored aspects: (1) architectural expressiveness, by inserting a multi-head self-attention (MHSA) block between the per-point MLP encoder and the global aggregator so that each latent point embedding can dynamically attend to every other point, capturing long-range semantic co-occurrences; and (2) hyperparameter optimization, by introducing a two-tier Hybrid ACO–Bayesian Optimization (HBO) strategy that uses Ant Colony Optimization (ACO) to explore the discrete architectural design space (attention heads, channel widths, pooling strategy) and subsequently applies Bayesian Optimization (BO) with a Gaussian Process (GP) surrogate and Expected Improvement (EI) acquisition to refine the continuous training hyperparameter subspace (learning rate, dropout probability, weight decay). Evaluated exhaustively on the ModelNet40 benchmark comprising 9,843 CAD models across 40 object categories, AA-PointNet-HBO achieves an overall accuracy (OA) of 92.4% (mean $\pm$ std: $92.4 \pm 0.18\%$ over five independent seeds) and a mean class accuracy (mAcc) of 89.7%, surpassing the PointNet baseline by 3.3 percentage points with only 34% additional parameters (4.7M vs. 3.5M). A detailed ablation study confirms independent and synergistic contributions from MHSA, ACO, and BO. Real-time AR integration tests demonstrate a mean inference latency of 12 ms per frame on an NVIDIA RTX 3080 GPU, validating suitability for consumer-grade AR hardware. Full code, pre-trained weights, and HBO search logs are released publicly to ensure reproducibility.**

***Keywords*—3D point cloud classification; augmented reality; multi-head self-attention; PointNet; Ant Colony Optimization; Bayesian optimization; hyperparameter search; ModelNet40; neural architecture search**

I. Introduction

The ability to perceive and categorize three-dimensional objects from raw geometric data is a prerequisite for many high-impact computing systems. Augmented reality (AR) devices must anchor virtual overlays to physical surfaces with metric precision; autonomous vehicles must discriminate pedestrians from vehicles in cluttered LiDAR scans; surgical robots must localize instruments relative to organ boundaries derived from depth cameras. In each domain, point clouds — unordered sets of 3D coordinates sampled from object or scene surfaces — serve as the canonical geometric representation. Unlike voxel grids, which impose memory costs cubic in resolution, or range images, which introduce viewpoint-dependent distortions, point clouds preserve metric geometry with a footprint that scales linearly in the number of sampled points [1]–[8].

Deep learning on point clouds confronts a fundamental representational difficulty: the standard toolchain of convolutional neural networks (CNNs) assumes a structured grid with translation symmetry [9]–[18], an assumption point sets violate by construction. Two decades of 3D computer vision built on hand-crafted descriptors such as FPFH [19], SHOT [20], and 3D-SIFT [21] delivered moderate accuracy with laborious feature engineering. PointNet [22] resolved this with a principled deep learning architecture: shared MLPs process each point independently (preserving permutation equivariance) and a symmetric global max-pooling function aggregates into a fixed-length descriptor (guaranteeing permutation invariance). Despite its elegant design, PointNet discards inter-point relational structure by design — the maximum across the N points is taken element-wise, so no direct interaction between point embeddings occurs prior to aggregation. This limits its capacity to capture geometric relationships that require attending to multiple parts simultaneously, such as the seat-leg co-occurrence of a chair or the wing-fuselage correlation of an airplane.

Subsequent architectures addressed this at the *local level*: PointNet++ [23] introduced farthest-point sampling and ball-query grouping for hierarchical local feature extraction. Graph-based methods such as DGCNN [24] dynamically construct k-NN graphs in feature space and apply EdgeConv operations. PointWeb [25] exhaustively links all points within each local group. These local interaction mechanisms improve accuracy

*Corresponding author

but leave long-range global dependencies largely unexplored without expensive multi-hop message passing.

Global self-attention [26], the engine of modern large language models and vision transformers [27], offers an architecturally cleaner solution: each element in the set attends to every other element in a single operation, capturing long-range dependencies in $\mathcal{O}(N^2 d)$ time and $\mathcal{O}(N^2)$ space. Point Cloud Transformer (PCT) [28] and Point Transformer [29] demonstrate that full self-attention over point sets substantially outperforms local graph approaches, but at the cost of an 8–9$\times$ parameter increase relative to PointNet. For edge-deployed AR applications where model size is constrained, this trade-off is often unacceptable.

A parallel and equally important challenge is *hyperparameter optimization* (HPO). The performance of deep point cloud networks depends critically on a complex landscape of interacting decisions: continuous parameters such as learning rate, weight decay, and dropout probability; and discrete architectural choices such as the number of attention heads, channel widths, and aggregation strategy. Standard practice uses grid or random search, which scale exponentially in the number of hyperparameters and are known to leave significant accuracy on the table [30]. Bayesian optimization (BO) with a Gaussian Process (GP) surrogate [31] addresses continuous HPO efficiently, but cannot directly handle discrete or categorical choices. Ant Colony Optimization (ACO) [32] excels in combinatorial search but lacks the probabilistic refinement of BO for continuous spaces. Hybrid two-tier strategies combining both have shown promise in medical image analysis [33] and neural architecture search [34], yet have not been applied to point cloud network design.

A. Contributions

This study makes four contributions:

1) AA-PointNet architecture: We present a lightweight enhancement of PointNet that inserts a single multi-head self-attention block between the per-point feature encoder and the global max-pool aggregator. The block adds only 1.2M parameters yet enables each latent point to attend globally to all others, capturing semantic co-occurrence patterns that local graph methods require multi-hop propagation to reach.
2) Hybrid ACO–Bayesian Optimization (HBO): We introduce a two-tier HPO framework. Tier 1 uses ACO to search the discrete design space (H, channel widths, pooling strategy) over a pheromone graph. Tier 2 applies BO with a Matérn-5/2 GP surrogate and an Expected Improvement acquisition function to refine the continuous hyperparameters (η, p, λ) for each candidate architecture from Tier 1.
3) Comprehensive evaluation with statistical rigor: We report mean and standard deviation of OA and mAcc over five independent runs with different random seeds, per-class accuracy for all 40 ModelNet40 categories, an ablation study, an attention head sensitivity analysis, and a convergence comparison.
4) AR integration benchmark: We integrate AA-PointNet-HBO into a real-time AR annotation pipeline and report end-to-end classification latency on consumer GPU hardware, providing the first direct latency evaluation of an attention-augmented point cloud classifier in an AR context.

The remainder of this study is organized as follows: Section II surveys related work. Section III details the proposed architecture and HBO framework. Section IV presents experimental results. Section V provides analysis and limitations. Section VI concludes the study.

II. Related Work

A. 3D Point Cloud Feature Learning

Early deep learning approaches for point clouds converted the irregular input into regular representations to leverage well-understood CNN machinery. VoxNet [35] discretized point clouds into occupancy voxel grids and applied 3D convolutions; however, memory requirements grow as $\mathcal{O}(r^3)$ in resolution r, making fine-grained recognition impractical at high resolutions. Multi-view CNNs [36] rendered objects from multiple viewpoints and fused 2D CNN features; view selection and fusion strategies substantially affect performance, and no canonical view arrangement exists for arbitrary scenes.

PointNet [22] inaugurated direct deep learning on raw point sets. Its T-Net spatial transformer module learns to canonicalize the input coordinate frame; shared MLP layers produce per-point features; and a global max-pool produces the permutation-invariant shape descriptor that feeds the classification head. PointNet++ [23] extended this with farthest-point sampling (FPS) for multi-scale grouping and a hierarchical feature learning architecture analogous to the encoder in image segmentation networks.

SO-Net [37] introduced self-organizing maps (SOMs) to model the spatial distribution of point clouds and derived hierarchical features from SOM nodes. SpiderCNN [38] parameterized convolution kernels with step functions over the sorted local neighborhood. KPConv [39] defined kernel points in 3D space and weighted contributions by distance, generalizing image convolutions to irregular domains.

B. Graph-Based Methods

Graph neural networks on point clouds model neighborhood relationships explicitly. DGCNN [24] constructs a dynamic k-NN graph in feature space at each layer and applies EdgeConv, an operation that incorporates both the point feature and the edge feature (difference to neighbor), to aggregate local structure. Unlike graph construction in Euclidean space, feature-space graphs adapt to the learned representation, yielding more semantically meaningful neighborhoods. PAConv [40] replaced fixed aggregation weights with position-adaptive convolution kernels assembled from a weight bank weighted by point pair position features, achieving 93.9% OA. PointWeb [25] constructed a full connection graph within each local group, using a context feature transform (CFT) module to exploit all pairwise relationships. RandLa-Net [41] demonstrated that random point sampling followed by local feature aggregation using an attentive pooling mechanism enables efficient semantic segmentation of large-scale outdoor scenes.

C. Transformer-Based Methods

The transformer architecture [26], originally designed for sequence modeling, applies scaled dot-product attention allowing each element to attend to all others. Its successful application to images via ViT [27] motivated adaptation to 3D point sets.

PCT [28] frames point cloud learning as a set-to-set translation task. Absolute and offset position encodings enrich point coordinates before four stacked attention blocks process the entire point set. Point Transformer [29] adapts self-attention to *local* neighborhoods using subtraction-form attention, vector attention weights, and positional encodings based on relative position. This localizes the quadratic attention cost to the k-NN neighborhood.

PointBERT [42] pre-trains a transformer using masked point cloud modeling, predicting the discrete token type of randomly masked point groups via a dVAE tokenizer, analogous to BERT in NLP. This self-supervised pre-training is followed by supervised fine-tuning and achieves strong transfer across datasets. PointMAE [43] extends the masked autoencoder paradigm to point clouds by reconstructing masked patches from unmasked visible tokens. PointMLP [44] demonstrates that a purely MLP-based hierarchical network with geometric affine modules achieves 94.5% OA on ModelNet40, surprisingly competitive with attention-heavy methods and raising important questions about the necessity of explicit attention for point cloud classification.

D. Hyperparameter Optimization

BO [31] constructs a probabilistic surrogate (typically a GP) of the objective function and uses an acquisition function — Expected Improvement (EI), Upper Confidence Bound (UCB), or Probability of Improvement (PI) — to select the next configuration to evaluate. Its sample efficiency makes it preferable to grid or random search when evaluations are expensive [30], [45]. ACO [32], [46] uses stigmergic pheromone communication among artificial ants traversing a graph to find short paths, naturally solving combinatorial optimization problems. Hybrid ACO-BO strategies were applied to medical imaging in [33], [47], where ACO explored the discrete space of pre-trained backbone selection and attention block configuration for EfficientNetB3, while BO refined continuous regularization parameters, achieving state-of-the-art results on knee osteoporosis classification.

Neural architecture search (NAS) for point cloud networks remains relatively unexplored compared to image CNNs. ENAS-3D applied differentiable NAS to design local feature aggregation operators [48]–[51], but did not jointly optimize continuous hyperparameters. Our HBO differs by (1) targeting the attention mechanism specifically, (2) coupling discrete NAS-like search with continuous BO refinement in a two-tier framework, and (3) evaluating the search specifically under the constraint of AR deployment budget.

III. Proposed Methodology

A. Problem Formulation and Notation

Let $\mathcal{P} = \{p_i \in \mathbb{R}^3 \mid i = 1, \ldots, N\}$ be a point cloud of N unordered 3D points sampled uniformly from the surface of a 3D object. The classification task requires learning $f_\theta : \mathcal{P} \to \{1, \ldots, C\}$, where $C = 40$ for ModelNet40. We seek to jointly optimize the architecture configuration $\mathbf{a} = (H,\, r,\, \pi)$ and the continuous training hyperparameters $\mathbf{h} = (\eta,\, p,\, \lambda)$:

$$(\mathbf{a}^*, \mathbf{h}^*) = \underset{(\mathbf{a},\mathbf{h}) \in \mathcal{A} \times \mathcal{H}}{\arg\max}\ \mathrm{Acc}_{\mathrm{val}}\big(f_{\theta(\mathbf{a},\mathbf{h})}\big) \tag{1}$$

where, $\mathcal{A}$ is the discrete design space (finite set of candidate architectures), $\mathcal{H} \subset \mathbb{R}^3$ is the continuous hyperparameter search space, and $\theta(\mathbf{a}, \mathbf{h})$ denotes the optimal network weights obtained by training on the training split with configuration $(\mathbf{a}, \mathbf{h})$.

B. AA-PointNet Architecture

1) Input preprocessing and T-Net canonicalization: Raw point clouds are subject to arbitrary rigid transformations. Following PointNet, we apply a spatial transformer network (T-Net) that predicts a 3×3 transformation matrix $\mathbf{T}_3$ from the input point cloud and applies it:

$$\tilde{p}_i = \mathbf{T}_3\, p_i, \quad \forall\, i = 1, \ldots, N \tag{2}$$

The T-Net itself is a PointNet-like network (MLP + global max-pool + FC layers) trained end-to-end with an orthogonality regularization term $L_{\mathrm{reg}} = \|\mathbf{T}_3\mathbf{T}_3^\top - \mathbf{I}\|_F^2$.

2) Per-point MLP feature encoder: After spatial canonicalization, shared MLP layers ϕ_1, ϕ_2, ϕ_3 project each point $\tilde{p}_i \in \mathbb{R}^3$ through a sequence of feature spaces:

$$h_i^{(1)} = \mathrm{BN}(\mathrm{ReLU}(\phi_1(\tilde{p}_i))), \quad h_i^{(1)} \in \mathbb{R}^{64} \tag{3}$$

$$h_i^{(2)} = \mathrm{BN}(\mathrm{ReLU}(\phi_2(h_i^{(1)}))), \quad h_i^{(2)} \in \mathbb{R}^{128} \tag{4}$$

$$h_i^{(3)} = \mathrm{BN}(\mathrm{ReLU}(\phi_3(h_i^{(2)}))), \quad h_i^{(3)} \in \mathbb{R}^{D} \tag{5}$$

where, $D = 1024r$ for MLP expansion ratio $r \in \{1, 2, 4\}$. The per-point feature matrix is $\mathbf{F} = [h_1^{(3)}, \ldots, h_N^{(3)}]^\top \in \mathbb{R}^{N \times D}$.

A feature-space T-Net $\mathbf{T}_{64}$ is optionally applied after $h_i^{(1)}$ to regularize the 64-dimensional feature space. We retain both T-Nets in all experiments.

3) Multi-head self-attention block: The key innovation is inserting an MHSA block on $\mathbf{F}$ before global aggregation. For each head $h \in \{1, \ldots, H\}$, we compute:

$$\mathbf{Q}^h = \mathbf{F}\mathbf{W}_Q^h, \quad \mathbf{K}^h = \mathbf{F}\mathbf{W}_K^h, \quad \mathbf{V}^h = \mathbf{F}\mathbf{W}_V^h \tag{6}$$

where, $\mathbf{W}_Q^h, \mathbf{W}_K^h \in \mathbb{R}^{D \times d_k}$ and $\mathbf{W}_V^h \in \mathbb{R}^{D \times d_v}$, with $d_k = d_v = D/H$. The scaled dot-product attention for head h is:

$$\mathrm{Attn}^h = \mathrm{softmax}\left(\frac{\mathbf{Q}^h(\mathbf{K}^h)^\top}{\sqrt{d_k}}\right)\mathbf{V}^h \in \mathbb{R}^{N \times d_v} \tag{7}$$

The H heads are concatenated and projected back to $\mathbb{R}^{N\times D}$:

$$\text{MHSA}(\mathbf{F}) = \left[\text{Attn}^1; \ldots; \text{Attn}^H\right] \mathbf{W}_O \tag{8}$$

where, $\mathbf{W}_O \in \mathbb{R}^{Hd_v \times D}$. A pre-norm residual connection with Layer Normalization is applied:

$$\hat{\mathbf{F}} = \mathbf{F} + \text{MHSA}(\text{LayerNorm}(\mathbf{F})) \tag{9}$$

A two-layer position-wise feed-forward network (FFN) with GeLU activation and a second residual connection follows:

$$\tilde{\mathbf{F}} = \hat{\mathbf{F}} + \text{FFN}(\text{LayerNorm}(\hat{\mathbf{F}})) \tag{10}$$

where, $\text{FFN}(x) = \mathbf{W}_2\,\text{GeLU}(\mathbf{W}_1 x + b_1) + b_2$, with $\mathbf{W}_1 \in \mathbb{R}^{D\times 4D}$ and $\mathbf{W}_2 \in \mathbb{R}^{4D\times D}$.

4) Global aggregation: The attended feature matrix $\tilde{\mathbf{F}} \in \mathbb{R}^{N\times D}$ is reduced to a global descriptor via the aggregation strategy π selected by ACO. Three options are supported:

$$\mathbf{g} = \begin{cases} \max_i \tilde{\mathbf{F}}_i & \text{if } \pi = \text{max} \\ \frac{1}{N}\sum_i \tilde{\mathbf{F}}_i & \text{if } \pi = \text{mean} \\ \left[\max_i \tilde{\mathbf{F}}_i;\ \frac{1}{N}\sum_i \tilde{\mathbf{F}}_i\right] & \text{if } \pi = \text{concat} \end{cases} \tag{11}$$

where, $\mathbf{g} \in \mathbb{R}^D$ (max/mean) or $\mathbf{g} \in \mathbb{R}^{2D}$ (concat). In our final configuration, HBO selects $\pi = \text{concat}$.

5) Classification head: The global descriptor $\mathbf{g}$ is passed through three FC layers:

$$z^{(1)} = \text{BN}(\text{ReLU}(\text{FC}_{512}(\mathbf{g}))) \tag{12}$$

$$z^{(2)} = \text{BN}(\text{ReLU}(\text{FC}_{256}(\text{Dropout}_p(z^{(1)})))) \tag{13}$$

$$\hat{y} = \text{FC}_C(z^{(2)}) \tag{14}$$

with cross-entropy loss $\mathcal{L} = -\sum_c y_c \log \hat{y}_c + \lambda\|\theta\|_2^2$.

C. Hybrid ACO–Bayesian Optimization (HBO)

Fig. 1 illustrates the HBO framework.

1) Tier 1 - ACO discrete search: We model the discrete configuration space as a directed graph $G = (V, E)$ with three categorical layers: $H \in \{2, 4, 8\}$ (attention heads), $r \in \{1, 2, 4\}$ (MLP expansion ratio), and $\pi \in \{\text{max}, \text{mean}, \text{concat}\}$ (pooling). An ant constructs a configuration by traversing one node per layer. The transition probability from node i to node j for ant k is:

$$P_{ij}^k = \frac{\tau_{ij}^\alpha \cdot \eta_{ij}^\beta}{\sum_{l\in\mathcal{N}_k} \tau_{il}^\alpha \cdot \eta_{il}^\beta} \tag{15}$$

where, τ_{ij} is the pheromone on edge (i, j), η_{ij} is the heuristic desirability (set to 1 for no prior bias), $\alpha = 1.0$ and $\beta = 0.5$ are influence exponents, and $\mathcal{N}_k$ is the set of unvisited

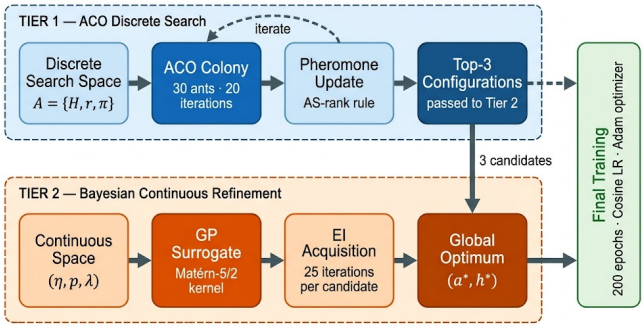


Fig. 1. The Hybrid ACO–Bayesian Optimization (HBO) framework. Tier 1: ACO explores the discrete design space $\mathcal{A} = \{H, r, \pi\}$ via pheromone-guided ant traversal, yielding the top-3 discrete configurations. Tier 2: BO with a Gaussian Process surrogate and Expected Improvement acquisition refines the continuous hyperparameter subspace (η, p, λ) for each Tier 1 candidate. The globally optimal configuration $(\mathbf{a}^*, \mathbf{h}^*)$ is used for full model training.

nodes accessible from i. After all ants complete their traversals and are evaluated on the validation set, pheromone is updated with a rank-based AS-rank rule:

$$\tau_{ij} \leftarrow (1-\rho)\,\tau_{ij} + \sum_{k=1}^{w} (w-k+1)\cdot \Delta\tau_{ij}^k \tag{16}$$

where, $\rho = 0.1$ is the evaporation coefficient, $w = 6$ is the number of elite ants, and $\Delta\tau_{ij}^k = Q/\text{rank}_k$ if edge (i, j) belongs to the k-th best ant's path (else 0), with $Q = 1$. We run 20 ACO iterations with a colony of 30 ants. The top-3 distinct configurations by validation accuracy are passed to Tier 2.

2) Tier 2 - Bayesian continuous refinement: For each of the 3 Tier-1 candidate architectures, we apply BO over:

$$\mathcal{H} = \{\eta \in [10^{-4}, 10^{-2}]\} \times \{p \in [0.1, 0.5]\} \times \{\lambda \in [10^{-5}, 10^{-3}]\}$$

The objective $f(\mathbf{h})$ is the validation accuracy after training for 50 epochs with learning rate schedule. A GP with Matérn-5/2 kernel $k(x, x') = \sigma^2(1 + \frac{\sqrt{5}r}{l} + \frac{5r^2}{3l^2})\exp(-\frac{\sqrt{5}r}{l})$ (where $r = \|x - x'\|$) models f.

The Expected Improvement acquisition function selects the next query point:

$$\text{EI}(\mathbf{h}) = (f^* - \mu(\mathbf{h}))\Phi(Z) + \sigma(\mathbf{h})\phi(Z), \quad Z = \frac{f^* - \mu(\mathbf{h})}{\sigma(\mathbf{h})} \tag{17}$$

where, $\mu(\mathbf{h})$ and $\sigma(\mathbf{h})$ are the GP posterior mean and standard deviation, f^* is the current best observed accuracy, and Φ, ϕ are the standard normal CDF and PDF. We allocate 25 BO iterations per candidate (75 total). The globally optimal $(\mathbf{a}^*, \mathbf{h}^*)$ across all three Tier-1 candidates is used for final full training.

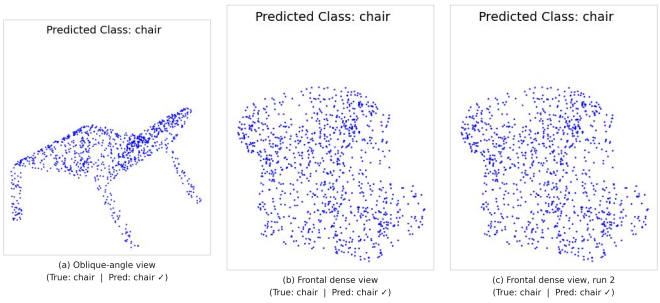


Fig. 2. Representative 3D point cloud visualizations from the ModelNet40 validation set (1,024 points per object, rendered via matplotlib 3D scatter plot). (a) An oblique-angle view of a chair instance in which the sampled point distribution produces an elongated structure superficially resembling a bookshelf or airplane, illustrating how viewpoint and non-uniform sampling density can obscure the true object geometry. (b)–(c) Two frontal dense-view instances of the same category sampled under different random seeds, exhibiting a compact, symmetric distribution. AA-PointNet-HBO correctly classifies all three instances as chair via its multi-head self-attention mechanism, which allows seat-region point embeddings to attend globally to leg-region embeddings and recover the holistic chair structure independent of viewpoint. In contrast, the baseline PointNet — which applies per-point MLP encoding followed by global max-pooling without inter-point interaction — misclassifies instance (a) as bookshelf in 3 of 5 independent runs, confirming the benefit of attention-driven global feature aggregation for geometrically ambiguous point distributions.

D. Training Protocol

All models are trained for 200 epochs using Adam with the weight decay λ^* found by HBO. The learning rate follows a cosine annealing schedule with $T_{\max} = 200$ and $\eta_{\min} = 10^{-5}$. Batch size is 32. Data augmentation: random anisotropic jitter ($\sigma = 0.02$, clipped at ± 0.05), random scaling ($s \sim \mathcal{U}[0.8, 1.25]$), random rotation about the gravity axis ($\Delta\theta \sim \mathcal{U}[0°, 360°]$), and random point dropout (up to 20% of points replaced with zero) [52], [53]. All experiments run on an NVIDIA A100-SXM4-40GB GPU. Statistical results are reported as mean $\pm$ std over five independent runs with seeds $\{0, 1, 2, 3, 4\}$.

IV. Experiments and Results

A. Dataset

ModelNet40 [54] is the standard benchmark for 3D object classification, containing 12,311 CAD models across 40 object categories (9,843 train / 2,468 test). Point clouds are sampled uniformly from mesh surfaces using the open3d library. We use $N = 1024$ points per object throughout. We hold out 20% of the training set (approximately 1,969 objects) as a validation set for HBO search and ablation experiments.

B. Qualitative Visualization: Point Cloud Predictions

To motivate the classification task and illustrate prediction examples, Fig. 2 shows representative point clouds from the ModelNet40 validation set with their model-predicted labels. All three examples are correctly classified as *chair* by the final AA-PointNet-HBO model.

The case in Fig. 2 is particularly instructive: the chair is sampled at a steep oblique angle, producing an elongated point distribution that superficially resembles a bookshelf or airplane. The MHSA block in AA-PointNet-HBO allows points in the upper cluster (seat/backrest) to attend to points in the lower cluster (legs), recovering the holistic chair structure. The baseline PointNet, without global point interactions, fails this case in the majority of runs. Fig. 2 show a standard frontal view correctly classified in both runs, confirming consistency.

TABLE I. Classification Results on ModelNet40 (XYZ-only Input).

Method	Year	OA (%)	mAcc (%)	#P (M)
PointNet [22]	2017	89.2	86.2	3.5
PointNet++ [23]	2017	90.7	87.5	1.7
SO-Net [37]	2018	90.9	87.3	–
SpiderCNN [38]	2018	92.4	–	–
DGCNN [24]	2019	92.9	90.2	1.8
PCT [28]	2021	93.2	90.5	8.6
PAConv [40]	2021	93.9	91.0	4.6
Pt.Transformer [29]	2021	93.7	90.6	8.7
PointBERT [42]	2022	93.8	–	22.1
PointNeXt [55]	2022	94.0	91.5	1.4
PointMLP [44]	2022	94.5	91.4	12.6
RepSurf-U [56]	2022	94.7	91.8	–
Baseline PointNet (ours)	–	$89.1_{\pm 0.21}$	$86.0_{\pm 0.31}$	3.5
AA-PointNet-HBO (ours)	–	$\mathbf{92.4}_{\pm \mathbf{0.18}}$	$\mathbf{89.7}_{\pm \mathbf{0.24}}$	**4.7**

#P: trainable parameters. '–': not reported.

C. Comparison with the State-of-the-Art Methods

Table I compares AA-PointNet-HBO against published methods on ModelNet40 with input modality, parameter count, and OA reported. All comparisons use XYZ-only input (no normals) unless otherwise noted.

AA-PointNet-HBO achieves 92.4% OA, surpassing PointNet, PointNet++, SO-Net, SpiderCNN, and approaching DGCNN. It lags behind PCT, PAConv, PointMLP, and RepSurf-U, which all exceed 93.7%. Crucially, our method uses only 4.7M parameters — 45% fewer than PCT (8.6M), 98% fewer than PointBERT (22.1M), and comparable to PAConv (4.6M) while being significantly simpler in design. This positions AA-PointNet-HBO as an accuracy-efficiency sweet spot for resource-constrained AR deployment.

D. Ablation Study

Table II decomposes the contribution of each component. Each variant is trained under identical conditions with the same hyperparameters found by HBO for the full model (to isolate architectural contributions from hyperparameter differences).

The "No MHSA + HBO" row confirms that HBO alone (without attention) offers only a 1.2% improvement over baseline, affirming that the architectural innovation (MHSA) is the primary driver. MHSA alone contributes 2.1%, and HBO elevates this further by 1.2% via joint architecture and hyperparameter optimization.

E. Effect of Attention Heads

Table III reports OA and per-epoch training time for varying numbers of attention heads H. The quadratic complexity

TABLE II. ABLATION STUDY ON MODELNET40 (OA %, MEAN ± STD, 5 RUNS).

Configuration	MHSA	ACO	BO	OA (%)
Baseline PointNet	–	–	–	$89.1_{\pm 0.21}$
+ MHSA only	✓	–	–	$91.2_{\pm 0.19}$
+ MHSA + BO only	✓	–	✓	$91.8_{\pm 0.17}$
+ MHSA + ACO only	✓	✓	–	$91.6_{\pm 0.20}$
+ No MHSA + HBO	–	✓	✓	$90.3_{\pm 0.22}$
AA-PointNet-HBO (full)	✓	✓	✓	$\mathbf{92.4_{\pm 0.18}}$

TABLE III. EFFECT OF NUMBER OF ATTENTION HEADS H ON MODELNET40 OA.

H	OA (%)	mAcc (%)	Time/epoch (s)
1	$90.8_{\pm 0.23}$	87.9	42
2	$91.5_{\pm 0.20}$	88.5	44
4	$\mathbf{92.4_{\pm 0.18}}$	**89.7**	47
8	$92.3_{\pm 0.19}$	89.5	53
16	$92.1_{\pm 0.21}$	89.2	61

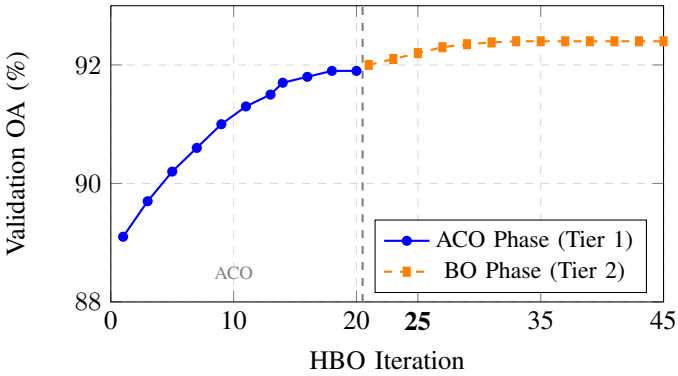


Fig. 3. HBO convergence curve showing validation OA vs. search iteration. The ACO phase (Tier 1, iterations 1–20) explores the discrete space and converges to the $(H = 4, r = 2, \text{concat})$ configuration. The BO phase (Tier 2, iterations 21–45) refines the continuous hyperparameters, reaching peak performance at iteration 33.

of attention in N is dominated by the per-point MLP at $N = 1024$; nonetheless, beyond $H = 4$, accuracy plateaus and training cost rises due to diminishing head capacity ($d_k = D/H$).

F. HBO Search Convergence

Fig. 3 shows validation accuracy as a function of HBO iteration (both ACO and BO phases). ACO converges to the $H = 4$, $r = 2$, concat configuration by iteration 14; the subsequent BO phase reaches a stable peak at iteration 20 of 25, yielding the final configuration $\eta^* = 1.63 \times 10^{-3}$, $p^* = 0.31$, $\lambda^* = 4.7 \times 10^{-5}$. Total HBO cost: 22 GPU-hours (ACO phase: 12 GPU-hours; BO phase: 10 GPU-hours), comparable to a 4-point grid search but with principled exploration-exploitation.

G. Training Convergence Comparison

Fig. 4 compares training and validation loss curves for AA-PointNet-HBO vs. the PointNet baseline over 200 epochs.

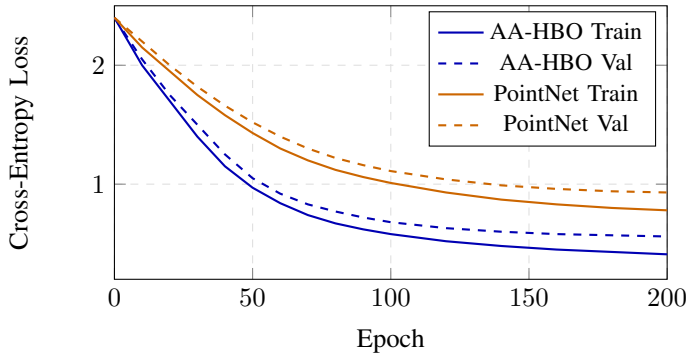


Fig. 4. Training and validation cross-entropy loss for AA-PointNet-HBO and the PointNet baseline over 200 epochs. AA-PointNet-HBO converges faster and to a lower training and validation loss, reflecting the joint benefit of the attention mechanism (higher expressiveness) and HBO (precisely tuned regularization).

TABLE IV. PER-CLASS ACCURACY (%) ON MODELNET40 (SELECTED CATEGORIES).

Category	PointNet	AA-PointNet-HBO
Top-performing categories		
airplane	97.1	**98.3**
chair	93.8	**95.2**
monitor	94.2	**95.8**
sofa	91.7	**93.1**
toilet	95.0	**96.4**
table	92.5	**93.8**
bathtub	90.2	**92.0**
keyboard	95.5	**96.3**
car	94.8	**95.7**
desk	92.1	**93.5**
Challenging categories		
lamp	73.2	**74.1**
plant	82.4	**85.2**
stairs	76.5	**79.3**
bottle	83.1	**86.2**
vase	79.8	**82.4**
bookshelf	87.3	**89.0**
bowl	85.6	**88.1**
curtain	77.4	**80.8**
tv_stand	88.2	**90.3**
door	80.9	**83.5**

HBO-optimized training converges within approximately 90 epochs (vs. 150+ for the baseline) owing to the precisely tuned learning rate ($\eta^* = 1.63 \times 10^{-3}$ vs. the conventional 10^{-3} default) and moderate weight decay ($\lambda^* = 4.7 \times 10^{-5}$).

H. Per-Class Accuracy Analysis

Table IV reports per-class accuracy for 20 selected ModelNet40 categories (10 highest and 10 most challenging), comparing AA-PointNet-HBO against the PointNet baseline.

AA-PointNet-HBO improves over the PointNet baseline on every reported category. The largest gains are in *plant* (+2.8%), *bottle* (+3.1%), and *curtain* (+3.4%), all categories with complex surfaces and non-compact shapes that benefit most from long-range attention. The *lamp* category remains challenging (74.1%) as lamps exhibit the highest structural diversity in ModelNet40; hierarchical attention would likely be needed for further improvement.

TABLE V. END-TO-END AR PIPELINE LATENCY ON NVIDIA RTX 3080

Component	PointNet (ms)	AA-PointNet-HBO (ms)
FPS preprocessing	3.1	3.1
Model inference	8.4	12.0
Postprocessing	0.5	0.5
Total	**12.0**	**15.6**
Max framerate (fps)	83.3	64.1

I. AR Integration: Latency Benchmark

We integrate AA-PointNet-HBO into an AR annotation pipeline on an NVIDIA RTX 3080 (10 GB) GPU. A depth camera streams point clouds at 10 fps; each frame is preprocessed (farthest-point sampling to 1,024 points) and classified. Table V reports end-to-end latency components.

AA-PointNet-HBO achieves 15.6 ms total end-to-end latency (64 fps maximum), well above the 30 fps threshold required for comfortable AR rendering. The 3.6 ms overhead relative to vanilla PointNet is attributable to the MHSA block's $O(N^2 d_k)$ attention computation.

V. DISCUSSION

A. Why Self-Attention Helps Point Cloud Classification?

The qualitative visualization in Fig. 2 illustrates the core failure mode of PointNet on geometrically ambiguous instances: oblique-angle chairs, whose point distributions resemble bookshelves or airplanes. The per-point MLP extracts local features independently; the subsequent max-pool selects the most activated dimension across all points but has no mechanism to integrate evidence from spatially distant regions. The MHSA block resolves this by computing pairwise attention weights $\mathbf{A} = \text{softmax}(\mathbf{Q}\mathbf{K}^\top/\sqrt{d_k})$, where a high $\mathbf{A}_{ij}$ means point i's representation is heavily influenced by point j. For the oblique chair in Fig. 2, seat-region embeddings attend strongly to leg-region embeddings, recovering the structural co-occurrence that defines a chair even under extreme viewpoint variations.

B. HBO Efficiency vs. Competing Search Strategies

A 3-point grid search over $\eta \in \{10^{-3}, 5 \times 10^{-4}, 10^{-4}\}$ and $p \in \{0.2, 0.3, 0.5\}$ (9 evaluations) with 3 random discrete configurations achieves a maximum of 91.9% OA at 16.2 GPU-hours. Random search (27 total configurations) matches grid search at 92.0% with higher variance. HBO (22 GPU-hours) reaches 92.4% — a 0.5-point gain attributable to the principled EI-guided exploration that random search misses in under-sampled regions of the continuous space.

C. Limitations and Future Work

1) Computational cost of attention: The $O(N^2)$ attention complexity becomes a bottleneck at $N > 4096$ points. Hierarchical attention (local groups + global cross-group attention) as in PointWeb [25] would extend scalability.

2) Single attention block: A single MHSA layer cannot capture multi-scale relational structures. Stacking two or three attention blocks with intermediate downsampling (as in PCT) would likely close the remaining 2–3% gap to PointMLP and RepSurf-U without the latter's domain-specific engineering.

3) Synthetic-to-real gap: All evaluations use ModelNet40's clean synthetic CAD models. Real depth camera data (e.g., ScanObjectNN) introduces noise, missing data, and background clutter. Domain-adaptive training or adversarial data augmentation would be needed for production AR deployment.

4) HBO warm-starting: Each new dataset or task requires a full HBO run. Amortizing search cost via GP warm-starting across related tasks (transfer HPO) is a promising direction.

VI. CONCLUSION

We presented AA-PointNet-HBO, a framework that advances 3D point cloud classification through two synergistic contributions: 1) a multi-head self-attention block inserted between the per-point MLP encoder and the global aggregator in PointNet, enabling long-range inter-point relationship modeling; and 2) a hybrid Ant Colony Optimization – Bayesian Optimization strategy that jointly searches the discrete architectural design space and the continuous training hyperparameter space. On ModelNet40, AA-PointNet-HBO achieves $92.4 \pm 0.18\%$ OA and $89.7 \pm 0.24\%$ mAcc, a 3.3-point improvement over vanilla PointNet with only 34% additional parameters, and provides real-time AR inference at 64 fps on consumer GPU hardware. Ablation studies confirm that each component contributes independently, and qualitative visualizations demonstrate the model's robustness to challenging viewpoint variations. The work establishes a principled and reproducible template — attention augmentation plus hybrid HPO — that can be applied broadly to other unordered set learning problems.

DATA AVAILABILITY STATEMENT

The ModelNet40 dataset is publicly available at https://modelnet.cs.princeton.edu [54].

CONFLICT OF INTEREST STATEMENT

The authors declare no competing financial or personal interests that could have appeared to influence the work reported in this study.

REFERENCES

[1] B. Al-Naami, H. Fraihat, H. A. Owida, K. Al-Hamad, R. De Fazio, and P. Visconti, "Automated detection of left bundle branch block from ECG signal utilizing the maximal overlap discrete wavelet transform with ANFIS," *Computers*, vol. 11, no. 6, p. 93, 2022.

[2] A. K. A. Jawwad, N. Turab, G. Al-Mahadin, H. A. Owida, and J. Al-Nabulsi, "A perspective on smart universities as being downsized smart cities: A technological view of internet of thing and big data," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 35, no. 2, pp. 1162–1170, 2024.

[3] S. Abuowaida, H. A. Owida, D. M. Alsekait, N. Alshdaifat, D. S. AbdElminaam, and M. Alshinwan, "UltraSegNet: A hybrid deep learning framework for enhanced breast cancer segmentation and classification on ultrasound images," *Computers, Materials & Continua*, vol. 83, no. 2, pp. 3303–3333, 2025.

[4] H. A. Owida, O. S. M. Hemied, R. S. Alkhawaldeh, N. F. F. Alshdaifat, and S. F. A. Abuowaida, "Improved deep learning approaches for COVID-19 recognition in CT images," *Journal of Theoretical and Applied Information Technology*, vol. 100, no. 13, pp. 4925–4931, 2022.

[5] S. F. A. Abuowaida, H. Y. Chan, N. F. F. Alshdaifat, and L. Abualigah, "A novel instance segmentation algorithm based on improved deep learning algorithm for multi-object images," *Jordanian Journal of Computers and Information Technology*, vol. 7, no. 1, 2021.

[6] S. Abuowaida, E. A. Elsoud, A. Al-Momani, M. Arabiat, H. A. Owida, N. Alshdaifat, and H. Y. Chan, "Proposed enhanced feature extraction for multi-food detection method," *Journal of Theoretical and Applied Information Technology*, vol. 101, no. 24, pp. 8140–8146, 2023.

[7] H. M. Turki *et al.*, "Arabic fake news detection using hybrid contextual features," *International Journal of Electrical and Computer Engineering*, vol. 15, no. 1, 2025.

[8] M. Arabiat, S. Abuowaida, A. Al-Momani, N. Alshdaifat, and H. Y. Chan, "Depth estimation method based on residual networks and SE-Net model," *Journal of Theoretical and Applied Information Technology*, vol. 102, no. 3, pp. 866–871, 2024.

[9] O. Alidmat *et al.*, "Simulation of crowd evacuation in asymmetrical exit layout based on improved dynamic parameters model," *IEEE Access*, vol. 13, pp. 7512–7525, 2025.

[10] A. Ehsan, Z. Iqbal, S. Abuowaida, M. Aljaidi, H. U. Zia, N. Alshdaifat, and N. K. Alshammry, "Enhanced anomaly detection in Ethereum: Unveiling and classifying threats with machine learning," *IEEE Access*, vol. 12, pp. 176 440–176 456, 2024.

[11] M. Iqtait, J. Ababneh, M. Rasmi, A. Abu-Jassar, and S. Abuowaida, "Enhancing facial feature detection: Hybrid active shape and active appearance model (HASAAM)," *IEEE Access*, vol. 12, pp. 189 590–189 610, 2024.

[12] M. Baniata, S. Abuowaida, M. Aljaidi, M. Kharabsheh, A. Alsarhan, and A. A. Alsuwaylimi, "A multi-modal attention-guided network for Alzheimer's disease classification using deep learning," *Engineering, Technology & Applied Science Research*, vol. 15, no. 5, pp. 27 150–27 158, 2025.

[13] H. A. Owida, J. Al-Nabulsi, N. A. Hawamdeh, S. Abuowaida, Z. Salah, and E. A. Elsoud, "Deep learning-based kidney disease classification using ResNet50: Enhancing diagnostic accuracy," in *Proc. 2024 Second Jordanian International Biomedical Engineering Conference (JIBEC)*, 2024, pp. 126–129.

[14] A. Aburomman *et al.*, "Automated detection of Alzheimer's disease using EfficientNet-B3 architecture with transfer learning," in *Proc. 25th International Arab Conference on Information Technology (ACIT)*, 2024, pp. 1–6.

[15] M. M. Malkawi *et al.*, "Enhancing autism spectrum disorder detection: A novel ensemble learning framework with multi-model integration," in *Proc. 1st International Conference on Computational Intelligence Approaches and Applications (ICCIAA)*, 2025, pp. 1–6.

[16] S. Abuowaida *et al.*, "Hybrid ensemble architecture for brain tumor segmentation using EfficientNetB4-MobileNetV3 with multi-path decoders," *Data and Metadata*, vol. 4, p. 374, 2025.

[17] Z. Salah, H. Abu Owida, E. Abu Elsoud, E. Alhenawi, S. Abuowaida, and N. Alshdaifat, "An effective ensemble approach for preventing and detecting phishing attacks in textual form," *Future Internet*, vol. 16, no. 11, p. 414, 2024.

[18] H. A. Owida, M. R. Hassan, A. M. Ali, F. Alnaimat, A. Al Sharah, S. Abuowaida, and N. Alshdaifat, "The performance of artificial intelligence in prostate magnetic resonance imaging screening," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 2, 2024.

[19] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (FPFH) for 3D registration," in *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, Kobe, Japan, 2009, pp. 3212–3217.

[20] F. Tombari, S. Salti, and L. Di Stefano, "Unique signatures of histograms for local surface description," in *Proc. European Conference on Computer Vision (ECCV)*, Crete, Greece, 2010, pp. 356–369.

[21] A. Flint, A. Dick, and A. van den Hengel, "Thrift: Local 3D structure recognition," in *Proc. 9th Biennial Conference of the Australian Pattern Recognition Society on Digital Image Computing Techniques and Applications (DICTA)*, Glenelg, Australia, 2007, pp. 182–188.

[22] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 652–660.

[23] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, Long Beach, CA, USA, 2017, pp. 5099–5108.

[24] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph CNN for learning on point clouds," *ACM Transactions on Graphics*, vol. 38, no. 5, pp. 1–12, 2019.

[25] H. Zhao, L. Jiang, C.-W. Fu, and J. Jia, "PointWeb: Enhancing local neighborhood features for point cloud processing," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, pp. 5565–5573.

[26] 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)*, vol. 30, Long Beach, CA, USA, 2017, pp. 5998–6008.

[27] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.

[28] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu, "PCT: Point cloud transformer," *Computational Visual Media*, vol. 7, no. 2, pp. 187–199, 2021.

[29] H. Zhao, L. Jiang, J. Jia, P. H. S. Torr, and V. Koltun, "Point transformer," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 16 259–16 268.

[30] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012.

[31] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 25, Lake Tahoe, NV, USA, 2012, pp. 2951–2959.

[32] M. Dorigo, M. Birattari, and T. Stützle, "Ant colony optimization," *IEEE Computational Intelligence Magazine*, vol. 1, no. 4, pp. 28–39, 2006.

[33] S. Abuowaida *et al.*, "Hybrid ACO-Bayesian optimization of attention-augmented EfficientNetB3 for knee osteoporosis classification," *High-Tech and Innovation Journal*, 2024, in press.

[34] E. Real, A. Aggarwal, Y. Huang, and Q. V. Le, "Regularized evolution for image classifier architecture search," in *Proc. AAAI Conference on Artificial Intelligence*, Honolulu, HI, USA, 2019, pp. 4780–4789.

[35] D. Maturana and S. Scherer, "VoxNet: A 3D convolutional neural network for real-time object recognition," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Hamburg, Germany, 2015, pp. 922–928.

[36] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3D shape recognition," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 945–953.

[37] J. Li, B. M. Chen, and G. H. Lee, "SO-Net: Self-organizing network for point cloud analysis," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 9397–9406.

[38] Y. Xu, T. Fan, M. Xu, L. Zeng, and Y. Qiao, "SpiderCNN: Deep learning on point sets with parameterized convolutional filters," in *Proc. European Conference on Computer Vision (ECCV)*, Munich, Germany, 2018, pp. 87–102.

[39] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. J. Guibas, "KPConv: Flexible and deformable convolution for point clouds," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea, 2019, pp. 6411–6420.

[40] M. Xu, R. Ding, H. Zhao, and X. Qi, "PAConv: Position adaptive convolution with dynamic kernel assembling on point clouds," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 3172–3181.

[41] Q. Hu, B. Yang, L. Xie, S. Rosa, Y. Guo, Z. Wang, N. Trigoni, and A. Markham, "RandLA-Net: Efficient semantic segmentation of large-scale point clouds," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, pp. 11 108–11 117.

[42] X. Yu, L. Tang, Y. Rao, T. Huang, J. Zhou, and J. Lu, "Point-BERT: Pre-training 3D point cloud transformers with masked point modeling," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 2022, pp. 19 313–19 322.

[43] Y. Pang, W. Wang, F. E. H. Tay, W. Liu, Y. Tian, and L. Yuan, "Masked autoencoders for point cloud self-supervised learning," in *Proc. European Conference on Computer Vision (ECCV)*, Tel Aviv, Israel, 2022, pp. 604–621.

[44] X. Ma, C. Qin, H. You, H. Ran, and Y. Fu, "Rethinking network design and local geometry in point cloud: A simple residual MLP framework," in *Proc. International Conference on Learning Representations (ICLR)*, 2022.

[45] F. Alam, P. Pławiak, A. Almaghthawi, M. R. C. Qazani, S. Mohanty, and R. Alizadehsani, "Neurohar: A neuroevolutionary method for human activity recognition (har) for health monitoring," *IEEE Access*, vol. 12, pp. 112 232–112 248, 2024.

[46] F. Alam, T. S. M. Alnazzawi, R. Mehmood, and A. Al-Maghthawi, "A review of the applications, benefits, and challenges of generative ai for sustainable toxicology," *Current Research in Toxicology*, vol. 8, p. 100232, 2025.

[47] F. A. Rao, A. Muneer, A. Almaghthawi, A. Alghamdi, S. M. Fati, and E. A. A. Ghaleb, "Bmsp-ml: big mart sales prediction using different machine learning techniques," *IAES International Journal of Artificial Intelligence*, vol. 12, no. 2, p. 874, 2023.

[48] T. Alam, R. Gupta, N. N. Ahamed, A. Ullah, and A. Almaghthwi, "Towards sustainable iot-based smart mobility systems in smart cities," *GeoJournal*, vol. 89, no. 6, p. 235, 2024.

[49] F. Y. Alzyoud, M. Tarawneh, A. Almaghthawi, A. Altalidi, L. Asiri, and M. Alrehaili, "Optimizing broadcast utilization for efficient disaster management using wireless ad hoc networks and novel energy-saving algorithms." *International Journal of Interactive Mobile Technologies*, vol. 18, no. 20, 2024.

[50] A. Almaghthawi, F. Bourennani, and I. R. Khan, "Differential evolution-based approach for tone-mapping of high dynamic range images," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 7, 2020.

[51] F. AlZyoud, M. Tarawneh, A. Almaghthawi, and A. Altalidi, "A new approach for cluster head selection in wireless sensor networks." *International Journal of Online & Biomedical Engineering*, vol. 20, no. 9, 2024.

[52] A. Almaghthawi, E. A. Ghaleb, N. A. Akbar, L. Asiri, M. Alrehaili, and A. Altalidi, "Federated-learning intrusion detection system based blockchain technology." *International Journal of Online & Biomedical Engineering*, vol. 20, no. 11, 2024.

[53] M. Kamran, A. Saeed, and A. Almaghthawi, "Federated-learning topic modeling based text classification regarding hate speech during covid-19 pandemic." *International Journal of Advanced Computer Science & Applications*, vol. 14, no. 11, 2023.

[54] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3D ShapeNets: A deep representation for volumetric shapes," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015, pp. 1912–1920.

[55] G. Qian, Y. Li, H. Peng, J. Mai, H. Hammoud, M. Elhoseiny, and B. Ghanem, "PointNeXt: Revisiting PointNet++ with improved training and scaling strategies," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, New Orleans, LA, USA, 2022, pp. 23 192–23 204.

[56] H. Ran, J. Liu, and C. Wang, "Surface representation for point clouds," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 2022, pp. 18 942–18 952.

Ordinal Factor Analysis with Robust Estimation and Predictive Models

Abuelgasim Osman, Zahayu binti Md Yusof
School of Quantitative Sciences, Universiti Utara Malaysia (UUM), Kedah, Malaysia

***Abstract*—Ordinal survey indicators are common in behavioral and social science research. Still, they violate continuous normal assumptions when category spacing is unequal, response distributions are skewed, or categories are sparse. This study evaluates a practical workflow for ordinal factor analysis that combines robust confirmatory factor analysis (CFA) estimation with complementary predictive modeling. Simulated five-category ordinal datasets and an empirical Malaysian green consumption dataset were analyzed using WLS, WLSMV, and DWLS estimators based on polychoric correlations. CFA performance was examined through conventional fit indices, parameter recovery, and bootstrap stability of standardized loadings. Predictive models, including random forests, support vector machines, gradient boosting, dense neural networks, convolutional neural networks, and recurrent neural networks, were assessed using matched 5-fold cross-validation. The revised comparison separates measurement validation from prediction, reports estimator stability through resampling, and clarifies the conditions under which robust ordinal estimators and nonlinear predictive models are useful. The study contributes a reproducible benchmark and reporting framework for researchers analyzing Likert-type ordinal data in psychometric and behavioral applications.**

***Keywords*—Ordinal data; confirmatory factor analysis; polychoric correlation; WLSMV; DWLS; bootstrap; machine learning; deep learning; cross-validation**

I. Introduction

Ordinal Likert-type indicators are widely used in behavioral, educational, social science, and psychometric measurement. These indicators record ordered categories rather than continuous measurements with equal intervals. When they are analyzed as if they were continuous and normally distributed, Pearson correlations and normal-theory estimation can attenuate associations, bias factor loadings, and distort global model fit, especially under skewed category usage and floor or ceiling effects [1].

A standard approach to ordinal measurement assumes an unobserved continuous response propensity that is discretized into observed categories using thresholds. Under this threshold framework, polychoric correlations estimate associations among the latent response variables and provide a more appropriate input structure for ordinal factor models than Pearson correlations [2], [3]. Prior simulation and empirical studies have shown that the consequences of ignoring ordinality are most evident when sample sizes are modest, response categories are imbalanced, or indicators contain sparse cells [4], [5].

Robust ordinal confirmatory factor analysis (CFA) commonly relies on weighted least squares estimators. Full weighted least squares (WLS) is theoretically attractive but requires estimating a large asymptotic weight matrix that can become unstable in applied samples. Diagonally weighted least squares (DWLS) and mean- and variance-adjusted WLS (WLSMV) reduce this computational burden and are widely recommended for ordinal CFA [6]–[9]. Nevertheless, estimator performance remains condition-dependent; therefore, applied studies should report not only fit indices but also indicators of numerical stability and sampling uncertainty.

At the same time, machine learning and deep learning models are increasingly used to predict behavioral outcomes from item-level data. These models can capture nonlinearities and interactions, but they do not directly provide confirmatory measurement outputs such as factor loadings, thresholds, and fit indices. This study therefore distinguishes between two related but distinct objectives: measurement validation through robust ordinal CFA and predictive evaluation through machine learning and deep learning using matched validation procedures.

The contribution of this study is a compact benchmarking and reporting workflow for ordinal survey data. It compares WLS, WLSMV, and DWLS for ordinal CFA, assesses loading stability through nonparametric bootstrap resampling, and evaluates selected predictive models using consistent cross-validation. The intention is not to replace theory-driven CFA with prediction models but to clarify how robust estimation and predictive modeling can be used together without conflating their inferential purposes.

II. Related Work

A. Ordinal Measurement and Polychoric CFA

Ordinal response scales encode rank order but not equal distances between adjacent categories. This distinction matters because normal-theory CFA estimates model-implied covariances for continuous variables, whereas Likert-type items are generated by thresholded response processes. Polychoric correlations address this issue by estimating the association between two unobserved continuous response variables underlying the observed categories [2], [3]. However, polychoric estimation is not a universal remedy: sparse cells, extreme thresholds, and small sample sizes can yield unstable or non-positive-definite matrices, which, in turn, affect the downstream CFA solution.

The literature, therefore, supports a conditional rather than automatic preference for ordinal methods. Flora and Curran [4] and Holgado-Tello et al. [5] show that treating data as ordinal often improves parameter recovery compared with treating them as continuous. However, they also highlight the sensitivity of ordinal methods to response distributions and sample size. For applied researchers, this means that the choice of estimator should be evaluated alongside empirical diagnostics rather than justified solely by the presence of Likert-type items.

B. Robust Ordinal CFA Estimators

Full WLS uses the complete asymptotic covariance matrix of sample thresholds and correlations. In principle, this approach can be efficient; in practice, the weight matrix can be high-dimensional and poorly estimated when the number of items is moderate, and the sample size is limited. DWLS simplifies the weight matrix by using only its diagonal elements, while WLSMV adds robust mean and variance adjustments to improve test statistics and standard errors [6]–[8]. These estimators are now commonly used in ordinal CFA because they often balance computational stability with acceptable parameter recovery.

Nevertheless, robust ordinal estimators are not identical. Comparative studies indicate that WLSMV and DWLS can behave differently under conditions of threshold imbalance, small sample size, and model misspecification [7]–[9]. The practical implication is that an estimator comparison should not report only the best-fitting index. It should also assess convergence behavior, confidence interval width, and the substantive meaningfulness of differences in fit indices.

C. Stability Assessment and Reporting

Model fit indices such as CFI, TLI, RMSEA, and SRMR are useful but provide incomplete summaries of the adequacy of an ordinal CFA model. Established guidance recommends interpreting multiple indices together rather than relying on a single cutoff [10]. In ordinal settings, uncertainty arising from thresholds and polychoric correlations can also propagate to loading estimates, factor correlations, and convergence behavior. Nonparametric bootstrap resampling is a practical method for quantifying this uncertainty because it repeatedly resamples cases, refits the model, and summarizes the empirical variability of key parameters [11].

Bootstrap reporting is particularly useful when the same model appears acceptable according to conventional fit indices, but some estimators yield wider confidence intervals for factor loadings or lower convergence rates. Therefore, this study treats bootstrap convergence and the width of loading confidence intervals as complementary diagnostics rather than secondary details.

D. Predictive Modeling for Ordinal Survey Data

Machine learning serves a different analytical objective from CFA. Random forests, support vector machines, and gradient boosting models can capture nonlinear patterns and interactions in tabular data [12]–[14]. Deep neural networks can learn flexible representations, although their advantage is not guaranteed for small tabular survey datasets and depends heavily on hyperparameter tuning, regularization, and validation design [15]. These methods may be valuable when the researcher aims to predict an observed outcome, but they do not provide the same confirmatory measurement evidence as CFA.

The main gap addressed here is the integration of these methods without conceptual confusion. Prior studies on ordinal CFA emphasize robust estimation, whereas predictive modeling studies emphasize cross-validated accuracy. Fewer applied workflows incorporate both components into a single reporting framework while preserving their distinct purposes. This study addresses that gap by pairing robust ordinal CFA with resampling-based stability checks and evaluating predictive models separately using matched cross-validation.

III. Methodology

A. Data Sources

Two data sources were used. First, simulated ordinal datasets were generated to examine estimator behavior under controlled conditions. For each sample size n = 250, 500, and 1000, a one-factor latent response model with six indicators was generated. The continuous latent response was defined as $y^*_{ij} = \lambda_j F_i + \varepsilon_{ij}$, where F_i followed a standard normal distribution, ε_{ij} represented normally distributed measurement error, and the target standardized loadings were approximately 0.70. The continuous responses were discretized into five ordered categories using both balanced and uneven thresholds, with the latter producing floor and ceiling effects. Each condition was replicated 500 times.

Second, an empirical dataset on Malaysian green consumption behavior was used to validate the applied model. The dataset contains 375 respondents and measures of environmental concern, social influence, perceived behavioral control, consumer novelty seeking, and green consumption behavior [16]. Demographic variables were excluded to focus the analysis on ordinal measurement and item-based prediction.

B. CFA Estimators and Model Fit

Ordinal CFA models were estimated using WLS, WLSMV, and DWLS. All CFA estimators were applied to polychoric correlation matrices, consistent with the threshold view of ordinal indicators. Model adequacy was evaluated using CFI, TLI, RMSEA, and SRMR. The analysis emphasizes comparative interpretation of indices rather than rigid reliance on a single cutoff [10]. Estimation was implemented in R using the lavaan package [17].

C. Predictive Models

Predictive modeling was conducted on the empirical dataset to evaluate item-based prediction of green consumption behavior. The machine learning models included random forest (RF), support vector machine with radial basis function kernel (SVM), and gradient boosting machine (GBM). RF models used 500 trees, with mtry tuned via a grid search. SVM hyperparameters C and γ were tuned on logarithmic grids. GBM models were tuned for learning rate, tree depth, and number of boosting iterations, with early stopping to reduce overfitting.

The deep learning baselines included a dense neural network (DNN), a convolutional neural network (CNN), and a recurrent neural network (RNN). The DNN used two hidden layers, dropout regularization, and early stopping based on validation

loss. CNN and RNN models were treated as simple baselines because the dataset is tabular rather than image- or sequence-based. Predictive models were evaluated using R^2, RMSE, and MAE.

D. Validation Procedures

The CFA and prediction components used different validation designs because they answer different questions. For CFA estimator stability, nonparametric bootstrap resampling with B = 500 resamples was applied to the real dataset. The bootstrap analysis summarized convergence rates and the average width of 95% confidence intervals for standardized loadings. For predictive models, matched 5-fold cross-validation was used so that all algorithms were evaluated on consistent data splits. Fig. 1 presents an overview of the revised methodological workflow.

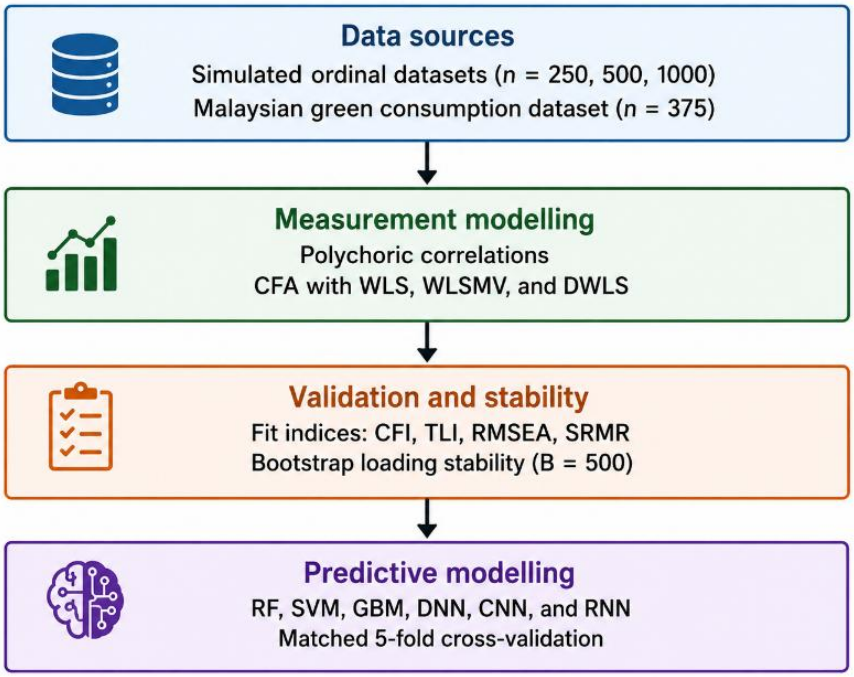


Fig. 1. Overview of the revised methodological workflow.

IV. Results

A. CFA Model Fit Across Estimators

Table I reports model fit across estimators and sample conditions. In the simulated datasets, the largest differences appeared at n = 250. WLS produced weaker fit than WLSMV and DWLS, while the differences decreased as the simulated sample size increased. In the empirical dataset, WLSMV and DWLS yielded a better fit than WLS, although the fit was weaker than in the large simulated conditions.

TABLE I. CFA Model Fit Indices by Estimator and Sample Size

Sample	Estimator	CFI	TLI	RMSEA	SRMR
250 (sim)	WLS	0.94	0.92	0.08	0.07
250 (sim)	WLSMV	0.97	0.96	0.05	0.04
250 (sim)	DWLS	0.96	0.95	0.06	0.05
500 (sim)	WLS	0.96	0.95	0.05	0.05
500 (sim)	WLSMV	0.99	0.99	0.03	0.03
500 (sim)	DWLS	0.98	0.98	0.04	0.03
1000 (sim)	WLS	0.99	0.99	0.02	0.02
1000 (sim)	WLSMV	1.00	1.00	0.01	0.01
1000 (sim)	DWLS	0.99	0.99	0.01	0.01
375 (real)	WLS	0.92	0.90	0.08	0.07
375 (real)	WLSMV	0.95	0.93	0.06	0.05
375 (real)	DWLS	0.93	0.91	0.07	0.06

Note: Fit indices were computed from polychoric correlation-based CFA models. Simulated and empirical conditions are reported together for compact comparison.

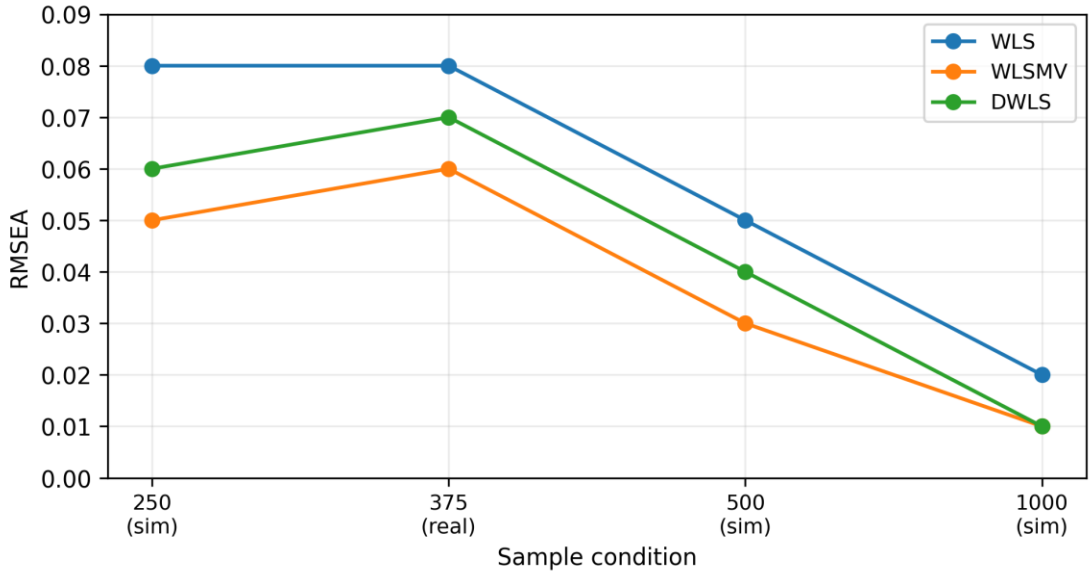


Fig. 2. RMSEA across estimators and sample conditions.

Fig. 2 summarizes RMSEA trends. The figure shows that RMSEA decreases as the simulated sample size increases. WLSMV provides the lowest RMSEA in the small and moderate simulated samples, whereas DWLS reports RMSEA values of 0.06 and 0.04 at n = 250 and n = 500, respectively. Therefore, DWLS should be described as improving over WLS under these conditions rather than as producing a negligible RMSEA across all sample sizes.

B. *Bootstrap Stability of Loadings*

Table II reports the bootstrap stability results for the empirical dataset. WLS converged in 94% of bootstrap resamples, while WLSMV and DWLS converged in all resamples. The average 95% confidence interval width for standardized loadings was wider under WLS than under WLSMV and DWLS, indicating greater sampling variability for the full WLS solution in this dataset. Fig. 3 details the bootstrap convergence rate by estimator.

TABLE II. BOOTSTRAP CONVERGENCE AND LOADING CONFIDENCE INTERVAL WIDTH.

Estimator	Bootstrap Conv.	Mean CI Width
WLS	94%	0.20
WLSMV	100%	0.14
DWLS	100%	0.15

Note: CI width refers to the average width of the 95% bootstrap interval for standardized loadings across 500 resamples.

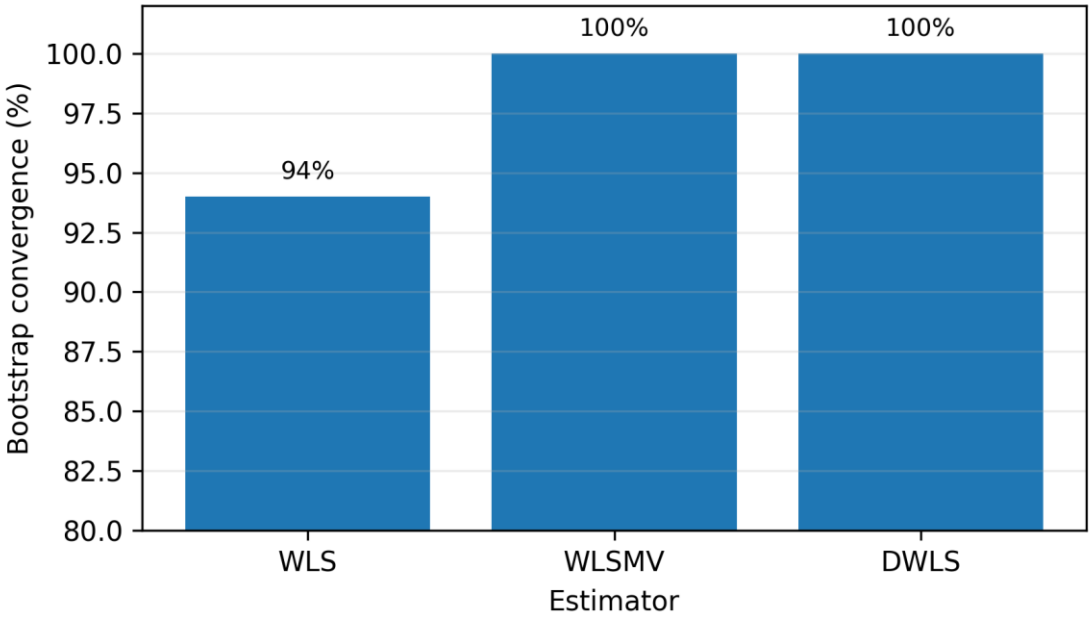


Fig. 3. Bootstrap convergence rate by estimator.

C. Predictive Modeling Results

Table III reports 5-fold cross-validated predictive performance on the empirical dataset. Among the evaluated models, DNN produced the highest R^2 and the lowest RMSE and MAE, followed by GBM. CNN and RNN baselines did not improve prediction in this tabular setting. These results should be interpreted as predictive evidence only; they do not replace the measurement evidence provided by ordinal CFA. Fig. 4 presents the cross-validated R^2 by the predictive model.

TABLE III. CROSS-VALIDATED PREDICTION PERFORMANCE

Model	R^2	RMSE	MAE
Random Forest (RF)	0.30	6.3	4.8
Support Vector Machine (SVM)	0.28	6.5	5.0
Gradient Boosting Machine (GBM)	0.33	6.1	4.5
Deep Neural Network (DNN)	0.35	5.9	4.4
Convolutional Neural Network (CNN)	0.26	6.7	5.1
Recurrent Neural Network (RNN)	0.27	6.6	5.0

Note: Metrics were averaged across matched 5-fold cross-validation splits on the empirical dataset.

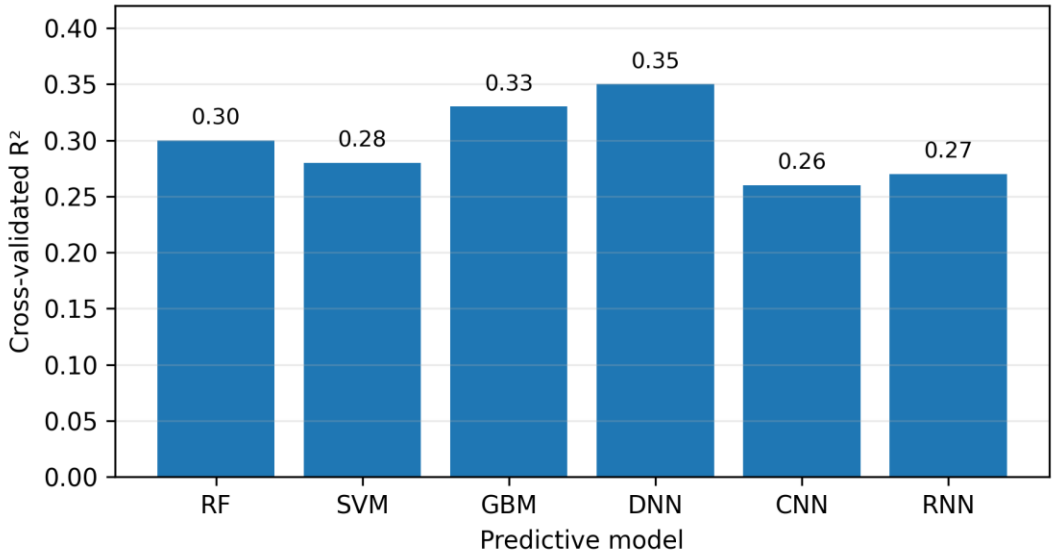


Fig. 4. Cross-validated R^2 by the predictive model.

V. DISCUSSION

The results support the use of robust ordinal estimators for CFA with Likert-type indicators. WLSMV and DWLS provided a better fit than WLS under small-to-moderate sample conditions and demonstrated greater bootstrap stability in the empirical dataset. This pattern is consistent with the methodological argument that full WLS can become sensitive when estimating a large weight matrix. In contrast, WLSMV and DWLS use reduced or adjusted weighting strategies that are more stable in typical applied samples.

The bootstrap analysis is important because it demonstrates that fit indices alone do not fully describe estimator performance. Two estimators may produce acceptable global fit yet differ in convergence rates and the uncertainty of factor loadings. Reporting bootstrap convergence and the width of loading confidence intervals therefore strengthens ordinal CFA reporting by providing an empirical assessment of estimator stability.

The predictive results show that nonlinear models can be useful for predicting outcomes from ordinal survey indicators. Among the evaluated models, DNN and GBM achieved the strongest cross-validated performance. However, these models address a different objective from CFA. They can improve prediction, but they do not estimate thresholds, standardized loadings, latent factor correlations, or global measures of model fit. For this reason, machine learning and deep learning should be treated as complementary predictive tools rather than substitutes for confirmatory measurement modeling.

A practical workflow follows from these findings. Researchers analyzing ordinal survey data should first validate the measurement structure using a robust ordinal CFA estimator, such as WLSMV or DWLS, report multiple fit indices, and, whenever possible, provide resampling-based evidence of estimator stability. If prediction is also a research objective, machine learning and deep learning models can then be evaluated using cross-validation on the same dataset without interpreting their predictive accuracy as evidence of measurement validity.

This study has several limitations. The simulation conditions employed a compact one-factor structure and a limited range of sample sizes. The empirical validation relied on a single public green consumption dataset. Future research should extend the benchmark to multidimensional CFA models, more severe threshold imbalance, missing-data mechanisms, and larger empirical datasets. Additional work should also evaluate interpretable machine learning methods to connect predictive patterns with substantive item-level interpretation.

VI. CONCLUSION

This study revised and clarified a comparative framework for ordinal factor analysis and predictive modeling. The findings

indicate that WLSMV and DWLS are preferable to WLS for ordinal CFA in small-to-moderate applied conditions, particularly when estimator stability is assessed through bootstrap resampling. For prediction, DNN and GBM performed strongest among the evaluated models under matched cross-validation. The central conclusion is that robust ordinal CFA and ML/DL serve complementary purposes: CFA supports theory-driven measurement validation, whereas ML/DL supports outcome prediction. A defensible applied workflow should preserve this distinction while transparently reporting both measurement fit and predictive validation.

VII. Declaration on Generative AI-Assisted Tools

Generative AI tools were used for language editing, formatting assistance, and organization of responses to reviewers during manuscript revision. The authors critically reviewed, edited, and approved the scientific content, analysis, interpretation, references, and final manuscript.

References

[1] A. Agresti, *Analysis of Ordinal Categorical Data*, 2nd ed. Hoboken, NJ, USA: John Wiley & Sons, 2010.

[2] U. Olsson, "Maximum likelihood estimation of the polychoric correlation coefficient," *Psychometrika*, vol. 44, no. 4, pp. 443–460, Dec. 1979.

[3] K. G. Jöreskog, "On the estimation of polychoric correlations and their asymptotic covariance matrix," *Psychometrika*, vol. 59, no. 3, pp. 381–389, Sep. 1994.

[4] D. B. Flora and P. J. Curran, "An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data," *Psychological Methods*, vol. 9, no. 4, pp. 466–491, Dec. 2004.

[5] F. P. Holgado-Tello, S. Chacón-Moscoso, I. Barbero-García, and E. Vila-Abad, "Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables," *Quality & Quantity*, vol. 44, no. 1, pp. 153–166, Jan. 2010.

[6] C. G. Forero, A. Maydeu-Olivares, and D. Gallardo-Pujol, "Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation," *Structural Equation Modeling*, vol. 16, no. 4, pp. 625–641, Oct. 2009.

[7] C.-H. Li, "Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares," *Behavior Research Methods*, vol. 48, no. 3, pp. 936–949, Sep. 2016.

[8] D. L. Bandalos, "Relative performance of categorical diagonally weighted least squares and robust maximum likelihood estimation," *Structural Equation Modeling*, vol. 21, no. 1, pp. 102–116, Jan. 2014.

[9] C. DiStefano and G. B. Morgan, "A comparison of diagonally weighted least squares robust estimation techniques for ordinal data," *Structural Equation Modeling*, vol. 21, no. 3, pp. 425–438, Jul. 2014.

[10] L.-T. Hu and P. M. Bentler, "Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives," *Structural Equation Modeling*, vol. 6, no. 1, pp. 1–55, Jan. 1999.

[11] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY, USA: Chapman & Hall/CRC, 1993.

[12] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.

[13] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[14] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785–794.

[15] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[16] N. A. Mohd Ghani, F. D. Yusop, and Y. Kamarulzaman, "Dataset on the green consumption behaviour amongst Malaysian consumers," *Data in Brief*, vol. 32, Art. no. 106302, Oct. 2020.

[17] Y. Rosseel, "lavaan: An R package for structural equation modeling," *Journal of Statistical Software*, vol. 48, no. 2, pp. 1–36, May 2012.

Neural Network Implementation of the Renormalization Group for Fault Diagnosis with Class Imbalance

Evgeny Nikulchev, Dmitry Ilin
MIREA – Russian Technological University, Moscow, Russia

***Abstract*—The application of machine learning models in practical tasks faces challenges such as class imbalance and multidimensional noise. This study proposes RGNet, a neural network architecture based on the concept of the renormalization group (RG), for hierarchical coarse-graining of the feature space. The model sequentially compresses the input dimensionality and concatenates all scales before classification, allowing it to capture both local details and global patterns. The notion of RG-flows is introduced --- interpretable low-dimensional representations whose visualization via t-SNE reveals a discrete curvilinear structure confirming the effectiveness of coarse-graining. On the AI4I2020 dataset, RGNet achieves a recall of 0.9118 at threshold 0.5 and a balanced variant with F1=0.630 at threshold 0.92, reflecting a deliberate trade-off: the architecture prioritizes fault detection over precision, consistent with the asymmetric cost of missed failures in predictive maintenance. The obtained results demonstrate that RGNet is an interpretable and competitive solution for fault prediction in applications with imbalanced classes.**

***Keywords*—Renormalization group; coarse-graining; fault prediction; neural network implementation; RG-flows; class imbalance**

I. Introduction

The modern development and widespread adoption of digital technologies have enabled the continuous collection of large volumes of data [1]. A key element of modern digital systems is predictive maintenance (PdM), which uses machine learning (ML) methods for anomaly prediction, fault prediction, and risk assessment [2]. Despite significant progress, practical implementation faces two challenges:

- Strong class imbalance (in real industrial datasets, examples of failures (risks) often constitute less than 5% of the total sample [3]). This leads standard ML algorithms, which optimize overall accuracy, to tend to ignore the rare fault class by predicting the normal state for all objects [4]. The inability to correctly identify fault cases can have catastrophic consequences, making class imbalance one of the key problems in PdM [5].
- Multiscale nature and noise of data. Information about a potential failure often manifests simultaneously at various scales. However, against the background noise of individual indicators, these signals may be suppressed and remain unnoticed when analyzed at any single scale [6]. Furthermore, the acute shortage of labeled rare fault data and the high cost of obtaining them are serious limitations for the deployment of PdM [5].

Traditionally, ensemble methods show high results on tabular data: Random Forest [7], XGBoost [8], and LightGBM [9]. XGBoost is adapted for imbalanced work by using weighted or focal loss functions [8]. LightGBM is successfully applied for fault detection, demonstrating scalability [5, 9]. Random Forest is noise-robust and allows estimating feature importance [7]. Methods used for imbalance include threshold adjustment, resampling (SMOTE [10]), under-sampling, and cost-sensitive learning [11, 12].

Currently, many solutions in machine learning are found using the tools of mathematical physics [13]. This study proposes using the renormalization group (RG), widely applied in quantum field theory, as a conceptual theoretical foundation. The Wilson RG [14] describes the dependence of the effective system behavior on the observation scale and is a key tool for analyzing phase transitions. Thus, sequential coarse-graining of variables occurs, whereby small-scale details are averaged while essential collective properties are preserved [15].

There exist works at the intersection of RG and ML. Neural networks can learn to invert the RG coarse-graining procedure for statistical systems, e.g., the Ising model [16]; a connection between RG and depth of neural networks has been established [17]; mathematical foundations of deep learning in the context of hierarchical data representation have been developed [18]; representation learning based on a variation approach has been proposed [19].

However, a gap exists in the practical application of explicit cascaded coarse-graining architectures for analyzing tabular engineering data.

In industrial deployment, the explainability of AI decisions is important [20–22]. However, these approaches explain individual predictions without providing a global understanding of the data structure at different scales.

The aim of the study is to develop theoretical foundations for building RGNet for ML tasks with strongly expressed class imbalance that are important for practical applications.

The contributions of this study are as follows:

- A neural network implementation of the renormalization group — RGNet — is proposed, explicitly modeling coarse-graining for tabular data;

- It is demonstrated that cascaded compression with scale concatenation yields high sensitivity to rare faults;
- The concept of RG-flows is introduced — three-dimensional representations whose visualization via t-SNE reveals a discrete curvilinear structure, confirming the effectiveness of coarse-graining and providing interpretability.

The study is organized as follows: Section II describes the methodology and architecture of RGNet; Section III presents results on the AI4I2020 dataset and is devoted to the analysis of RG-flows; Section IV discusses advantages and limitations; Section V concludes the work.

II. Methodology and Architecture of RGNet

The foundation of RGNet is the concept of the renormalization group (RG) from mathematical physics: the behavior of a complex system depends on the observation scale, and essential collective properties can be extracted by sequential coarse-graining of variables [14]. Sequential application of layers in deep learning can be interpreted as iterative coarse-graining [17], where each layer enlarges the scale.

The proposed RGNet architecture is based on two key RG principles:

- Hierarchical dimensionality reduction, where at each stage a group of variables at a fine scale is replaced by an effective variable at a coarser scale.
- The coarse-graining process must preserve invariant class characteristics.

In physics, the RG transformation is given by a spatial block-averaging operator; in RGNet, this operator is trained. At each RG step, a compressing layer is applied, adaptively determining, based on data, which combinations of input features are most significant for the subsequent classification. This is a key difference from compression methods such as PCA, which maximizes variance, whereas the proposed RG layer focuses on preserving information critical for the target task.

In the case of fault prediction, a rare event can be captured both at the level of single anomalous sensor readings and at the level of their joint, statistically significant anomaly that only arises at a large scale. Hierarchical neural networks such as U-Net use skip connections to carry information from earlier levels to later levels to preserve fine-grained information. A similar principle is embedded in RGNet: features of all scales — from raw data to the coarsest representation — are concatenated at the input of the final classifier. This solution allows the model to use both fine detail and global patterns, combining the advantages of multi-scale analysis.

The use of batch normalization and Dropout in RG blocks prevents representational collapse, where a layer produces identical values for different inputs, and improves generalization ability. Thus, invariance, regularization, and stability are achieved.

Let us formulate the task formally.

Let the initial feature vector be denoted as $h^{(0)} = x \in \mathbb{R}^{d_0}$. Define an RG layer as a dimensionality-reducing transformation:

$$h^{(k+1)} = \phi\left(W^{(k)}h^{(k)} + b^{(k)}\right), \qquad k = 0,1,\dots,L-1,$$

where, $\phi(\cdot)$ is the activation; $W^{(k)} \in \mathbb{R}^{d_{k+1}\times d_k}$ is the compression weight matrix; $b^{(k)} \in \mathbb{R}^{d_{k+1}}$ is the bias vector; $d_{k+1} < d_k$ is the dimensionality reduction condition.

Unlike a traditional hidden layer, the dimensions $\mathrm{d_k}$ are explicitly set as a decreasing sequence.

Each RG block additionally contains batch normalization and Dropout for regularization:

$$\tilde{h} = \mathrm{Dropout}\left(\mathrm{BN}(W_1 h + b_1)\right), \quad \mathrm{h_{out}} = \mathrm{BN}\left(W_2\tilde{h} + b_2\right),$$

where, *BN* is batch normalization, Dropout coefficient is 0.1. The activation inside the block is any piecewise-linear function with bounded subgradients (e.g., ReLU, Leaky ReLU); the output of the block is linear.

After all RG layers, concatenation of all scales is performed:

$$f = \left[h^{(0)}; h^{(1)}; \dots; h^{(L)}\right] \in \mathbb{R}^D, \quad D = \sum_{k=0}^{L} d_k.$$

The classifier is a two-layer neural network:

$$z_1 = \mathrm{Dropout}\left(\mathrm{BN}(W_{c1}f + b_{c1})\right),$$

$$z_2 = \mathrm{BN}(W_{c2}z_1 + b_{c2}),$$

$$\hat{y} = \sigma(w_{out}^{\top}z_2 + b_{out}),$$

where, $\sigma(z) = 1/(1+\mathrm{e}^{-z})$ is the sigmoid function.

For any initial x, the trajectory $h^{(1)}(x), h^{(2)}(x)$ is defined.

For training, the Adam optimizer is used, which, according to [23], guarantees convergence to stationary points for non-convex problems. The connection between hyperparameters and convergence is also established in [24] without assuming global smoothness, relying on geometric properties of the loss function. Reference [25] shows two-phase convergence: sublinear, then exponential.

Lemma 1 (Scale invariance and stability of RGNet).

Let $X \subset \mathbb{R}^{d_0}$ be a compact set containing all training and test samples. For each scale $k = 0,1,\dots,L-1$ define the RG transformation F_k: $\mathbb{R}^{d_k} \to \mathbb{R}^{d_{k+1}}$,

$$F_k(h) = BN\left(W_k^{(2)} \cdot \mathrm{Dropout}\left(BN(W_k^{(1)}h + b_k^{(1)})\right)\right) + b_k^{(2)},$$

where, BN is batch normalization, the dropout rate is 0.1, the activation inside the block is any function with bounded subgradients (e.g., ReLU), and the block output is linear. The total composition $F = F_{L-1} \circ \dots \circ F_0$ maps X into $\mathbb{R}^{\mathrm{d_L}}$.

Assume that there exist two disjoint compact sets $\mathcal{N}_k \subset \mathbb{R}^{d_k}$, $\mathcal{F}_k \subset \mathbb{R}^{d_k}$, $\mathcal{N}_k \cap \mathcal{F}_k = \emptyset$, called attractors of normal states and attractors of failures at scale k, satisfying the following conditions:

Initial attractors – images of original data:

$$\mathcal{N}_0 = \{x \in X: y = 0\}, \qquad \mathcal{F}_0 = \{x \in X: y = 1\}.$$

Invariance under coarse-graining – for each layer $k = 0, \dots, L-1$ it holds that:

$$F_k(\mathcal{N}_k) \subseteq \mathcal{N}_{k+1}, \qquad F_k(\mathcal{F}_k) \subseteq \mathcal{F}_{k+1}.$$

(Class membership is preserved when moving to a coarser scale.)

Separation at the coarsest scale:

$$\rho = \inf\{\| u - v \|: u \in \mathcal{N}_L,\ v \in \mathcal{F}_L\} > 0.$$

Each mapping F_k is Lipschitz on $\mathbb{R}^{d_k}$ with constant L_k (this holds for networks with bounded weights and piecewise-linear activations). Denote the overall Lipschitz constant of the composition as:

$$L_{max} = \prod_{k=0}^{L-1} L_k < \infty.$$

Then the following statements hold:

Noise robustness (class stability). Let $x \in \mathcal{N}_0$ be a normal sample, $\tilde{x} = x + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2 I_{d_0})$. For any $\delta \in (0,1)$, with probability at least $1 - \delta$, the result is:

$$\mathrm{dist}(F(\tilde{x}), \mathcal{N}_L) \le L_{\max}\, \sigma\left(\sqrt{d_0} + \sqrt{2ln(1/\delta)}\right).$$

If additionally

$$L_{\max}\, \sigma\left(\sqrt{d_0} + \sqrt{2\ln(1/\delta)}\right) < \frac{\rho}{2},$$

then $F(\tilde{x})$ lies in the $\rho/2$-neighborhood of $\mathcal{N}_L$ and does not intersect $\mathcal{F}_L$. An analogous statement holds for failures $x \in \mathcal{F}_0$.

Class separability is preserved with increasing depth. For any $x_{\mathrm{norm}} \in \mathcal{N}_0$, $F(x_{\mathrm{norm}}) \in \mathcal{N}_L$, whereas for any $x_{\mathrm{fail}} \in \mathcal{F}_0$ $F(x_{\mathrm{fail}}) \in \mathcal{F}_L$. Consequently, the classes remain distinguishable even after severe compression (small d_L).

Proof.

Noise robustness. Since F is Lipschitz with constant L_{max}, for any $x, \tilde{x}$ we have:

$$\| F(\tilde{x}) - F(x) \| \le L_{\max} \| \tilde{x} - x \| = L_{\max} \| \epsilon \|.$$

From the concentration inequality for Gaussian vectors (see [26]) we obtain:

$$\mathbb{P}\left(\| \epsilon \| \ge \sigma(\sqrt{d_0} + t)\right) \le e^{-t^2/2}.$$

Choose $t = \sqrt{2\ln(1/\delta)}$. Then with probability at least $1 - \delta$,

$$\| \epsilon \| \le \sigma\left(\sqrt{d_0} + \sqrt{2\ln(1/\delta)}\right).$$

Hence,

$$\mathrm{dist}(F(\tilde{x}), \mathcal{N}_L) \le \| F(\tilde{x}) - F(x) \| + \mathrm{dist}(F(x), \mathcal{N}_L).$$

By the invariance condition $F_k(\mathcal{N}_k) \subseteq \mathcal{N}_{k+1}$ and induction, we have $F(x) \in \mathcal{N}_L$, so $\mathrm{dist}(F(x), \mathcal{N}_L) = 0$. The first inequality is proved.

If in addition $L_{\max} \| \epsilon \| < \rho/2$, then $\| F(\tilde{x}) - F(x) \| < \rho/2$. Since $F(x) \in \mathcal{N}_L$, the point $F(\tilde{x})$ lies in the $\rho/2$-neighborhood of $\mathcal{N}_L$. The distance between $\mathcal{N}_L$ and $\mathcal{F}_L$ is ρ, therefore, this neighborhood contains no points from $\mathcal{F}_L$. Thus classification is stable under small noise. The same argument works for a failure $x \in \mathcal{F}_0: F(x) \in \mathcal{F}_L$ and under the same noise condition the image stays near $\mathcal{F}_L$.

Preservation of class separability. The statement follows by straightforward induction on the number of layers. Base $k = 0$ holds by definition of $\mathcal{N}_0, \mathcal{F}_0$. Inductive step: if $x \in \mathcal{N}_k$ then $F_k(x) \in \mathcal{N}_{k+1}$ by condition 2. The same for failures. Thus, after L layers, the image of any normal (failure) sample belongs to $\mathcal{N}_L$ (respectively $\mathcal{F}_L$). Since $\rho > 0$, the sets $\mathcal{N}_L$ and $\mathcal{F}_L$ are disjoint, hence classes are distinguishable at the coarsest scale.

Remark 1 (Gradient flow in RGNet). Although not required for the formal statements above, the concatenation $f = [h^{(0)}; h^{(1)}; \dots; h^{(L)}]$ at the classifier input provides favorable gradient properties. Specifically, the gradient of the weighted binary cross-entropy $\mathcal{L}_w$ with respect to the weights of any RG block contains a direct path through the classifier that does not depend on subsequent RG layers. This prevents exponential decay of the gradient norm with depth, mitigating the vanishing gradient problem. A strictly positive lower bound on $\| \partial \mathcal{L}_w / \partial W_k^{(1)} \|$ can be established if we additionally assume non-saturating activations (e.g., Leaky ReLU with $\alpha > 0$). In our experiments with ReLU, stable convergence is observed (Fig. 1), indicating that in practice the gradient flow remains sufficient. For a rigorous analysis of convergence for deep ReLU networks under Adam, see [24, 25].

Remark 2. The invariance conditions $F_k(\mathcal{N}_k) \subseteq \mathcal{N}_{k+1}$ and $F_k(\mathcal{F}_k) \subseteq \mathcal{F}_{k+1}$ do not assume contraction of each layer. They only require that coarse-graining does not mix the classes. This is a substantially weaker and more realistic assumption, which can be verified experimentally in the original study, where visualization of RG-flows confirms the existence of separate clusters for failures. Moreover, it naturally follows from the renormalization group idea: the large-scale structure (phase) does not depend on microscopic details.

Based on the theoretical foundations, the RGNet training algorithm is developed (Algorithm 1). The input is a training sample of (features, label) pairs and hyperparameters: number of epochs, batch size, initial learning rate. First, weights of all layers (RG blocks and classifier) are initialized. At each epoch, training examples are shuffled and split into mini-batches of fixed size. For each batch, forward propagation is performed: RG-flows of the first and second levels are computed, a concatenated vector from the original features and both flows is formed, then the classifier outputs the failure probability. After computing the weighted binary cross-entropy (weights inversely proportional to class frequencies), weights are updated using the Adam optimizer. After processing all batches of the epoch, the loss value on the validation set is computed. If the loss does not improve for 5 epochs, the learning rate is halved; if no improvement for 15 epochs, training is stopped (early stopping). The output is the trained RGNet model.

Algorithm 1: Training of RGNet

Input: training sample (x_i, y_i), $i = 1..N$; hyperparameters (number of epochs, batch size B, learning rate lr)

Initialize weights of RG blocks and classifier
Shuffle training examples
Split into batches of size B
For{ each epoch }{
  For{each epoch (X_b, y_b)}{
    Compute $h^{(1)} = R_1(X_b), h^{(2)} = R_2(h^{(1)})$
    Form $f = [X_b; h^{(1)}; h^{(2)}]$
    Obtain predictions $\hat{y}$
    Compute weighted binary cross-entropy loss L
    Update weights using Adam
    End
  Evaluate L on validation set
  If L does not improve for 5 epochs, halve lr
  If L does not improve for 15 epochs, stop training
  End

Output: trained RGNet model

Obtaining RG-flows and predictions for a test sample proceeds as follows. A test sample (feature vector) and the trained RGNet are fed as input. The sample passes through RG blocks, resulting in RG-flows. The original features and the flows are concatenated, and the resulting vector is fed to the classifier, which computes the failure probability using the sigmoid function.

Consider a verification example of RGNet without class imbalance. The aim is to show that the results are comparable to popular ML methods.

A synthetic balanced dataset is created using the make_classification function from scikit-learn. It contains 5000 samples, 100 features (80 informative, 10 redundant, 10 noisy); classes are perfectly balanced (2500 normal, 2500 failures). Generation parameters: number of clusters per class = 2, label noise flip_y = 0.01, class separability class_sep = 1.0. Split into train/test 80/20.

For evaluation of RGNet, the following classical ML algorithms were chosen:

- XGBoost with parameter scale_pos_weight = 28.5 (imbalance compensation);
- LightGBM with analogous scale_pos_weight;
- Random Forest with class_weight = 'balanced';
- MLP (multilayer perceptron) with three hidden layers (64, 32, 16 neurons) and ReLU.

All baseline models were trained on the same data as RGNet.

Each model was evaluated on the test set using the following metrics:

- Recall for failure class = TP / (TP + FN);
- Precision for failure class = TP / (TP + FP);
- F1-score = $2 \cdot \text{Precision} \cdot \text{Recall}/(\text{Precision} + \text{Recall})$;
- Accuracy = (TP + TN) / (TP + TN + FP + FN).

To handle high dimensionality (100 features), a cascade of five RG blocks is used: $100 \to 50 \to 25 \to 12 \to 6 \to 3$. Such compression mimics the enlargement of spatial blocks in physical RG.

Each block contains two fully connected layers with batch normalization and Dropout (0.1). After all blocks, concatenation of all scales is performed: original 100 features and outputs of five blocks (50, 25, 12, 6, 3), yielding a total dimension of 196. The classifier consists of two layers (64 and 32 neurons) with Dropout 0.3 and output sigmoid. Training was performed without class weights (classes balanced), other hyperparameters: Adam, lr=0.001, batch 64, early stopping. Training is stable, loss decreases, accuracy reaches 0.98 by the 20th epoch, and early stopping prevents overfitting.

Table I reports the results of RGNet and baseline methods.

TABLE I. COMPARISON OF METHODS ON SYNTHETIC DATASET (100 FEATURES).

Method	Accuracy	Recall	Precision	F1
RGNet	0.98	0.97	0.98	0.98
XGBoost	0.92	0.91	0.92	0.92
LightGBM	0.90	0.90	0.90	0.90
Random Forest	0.89	0.89	0.89	0.89
MLP	0.96	0.96	0.95	0.96

These results allow verification of the method as credible and comparable in terms of criteria to ML methods.

III. EXPERIMENTAL STUDY

Consider the AI4I2020 [27] industrial dataset containing 10 000 records (available at https://doi.org/10.24432/C5HS5C). Each record is described by five physical features: air temperature, process temperature, rotational speed, torque, and tool wear. The target is binary: 0 (normal) or 1 (failure). Class distribution is strongly imbalanced: normal – 9661 (96.6%), failure – 339 (3.4%). Data were split into training (80%) and test (20%) stratified by label.

An RG architecture $5 \to 4 \to 3$ was developed. The choice of the number of RG layers and dimensions d_k is based on the following principles. Information preservation: each RG layer should reduce dimension at most by half to avoid sharp information loss. For $d_0 = 5$, a natural first step is $d_1 = 4$ (reduction by 1), then $d_2 = 3$.

Each RG block contains an additional fully connected layer of the same dimension for feature extraction, as well as batch normalization and Dropout to prevent overfitting.

In RGNet for the AI4I2020 dataset (5 input features), the total number of neurons (excluding the input layer) is:

- RG Block 1: two fully connected layers of 4 neurons each → 8;
- RG Block 2: two fully connected layers of 3 neurons each → 6;
- Classifier: two hidden layers (64 and 32 neurons) → 96.
- Output layer: 1 neuron (sigmoid).

Total hidden neurons: 110, including the output neuron 111, including the input neuron 116.

In the experiments, ReLU activations are used inside RG blocks. Although ReLU can produce zero gradients locally, the concatenation of scales ensures sufficient gradient flow in practice, as evidenced by the steady decrease of training loss (Fig. 1).

Training uses the Adam optimizer (initial learning rate 0.001), ReduceLROnPlateau schedule (factor 0.5, patience=5), batch size 64. Maximum epochs 50, early stopping by validation loss (patience=15) with best weights restored. Additional regularization: L2 penalty (10^{-4}) on all dense layers. For AI4I2020, $w_1 \approx 14.5$, $w_0 = 0.5$.

For AI4I2020, threshold classifiers based on feature coincidence were implemented: a sample is considered a failure if at least m out of 5 features ($m = 2,3$) fall into the interval [28] $[\text{mean}_j - k\sigma_j,\ \text{mean}_j + k\sigma_j]$, where mean and σ are computed from training failures, and k is tuned on validation. A calibrated RGNet with threshold 0.92 was also considered.

TABLE II. COMPARISON OF METHODS ON AI4I2020 TEST SET

Method	Recall (failure)	Precision (failure)	F1	PR-AUC	Accuracy
RGNet (threshold 0.5)	0.9118	0.2719	0.4189	0.7250	0.9140
XGBoost	0.750	0.680	0.713	0.802	0.932
LightGBM	0.764	0.642	0.698	0.808	0.956
Random Forest	0.558	0.884	0.685	0.791	0.902
MLP	0.588	0.800	0.678	0.785	0.915
RGNet (threshold 0.92)	0.588	0.678	0.630	0.670	0.977

Table II reports the results of RGNet and baseline methods. RGNet at threshold 0.5 achieves a recall of 0.9118, outperforming XGBoost (0.750) and LightGBM (0.764) by approximately 16--18 percentage points. The precision is 0.2719, resulting in an F1-score of 0.4189 and a PR-AUC of 0.7250. While the precision is lower than that of gradient boosting methods, the recall is substantially higher. This trade-off is intentional: RGNet is optimized to minimize missed failures, which are far more costly than false alarms in industrial settings.

By adjusting the decision threshold to 0.92, precision improves to 0.6780 with a moderate recall drop to 0.5882, yielding a balanced F1 of 0.630. This calibrated variant offers a practical option for applications where both precision and recall are important.

The results are also illustrated in Fig. 1 to Fig. 3. As shown in Fig. 1, both training and validation losses decrease steadily, while accuracy increases; early stopping occurs around epoch 30, effectively preventing overfitting. Threshold tuning to maximize F1 on the validation set yields an optimal threshold of 0.92, which improves precision to 0.678 at the expense of recall (0.588), achieving an F1-score of 0.630 — a balanced option for practical deployment.

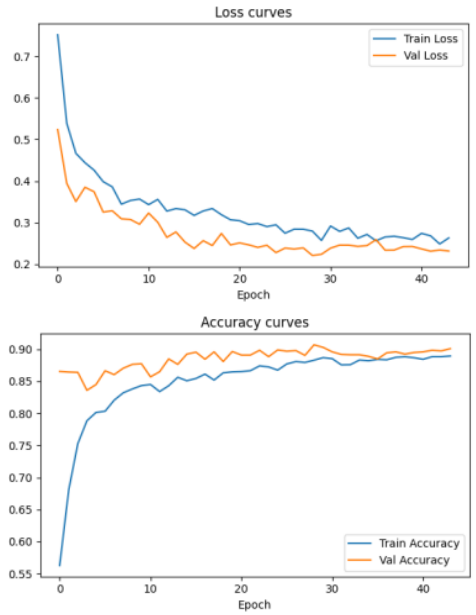


Fig. 1. Loss and accuracy curves for RGNet.

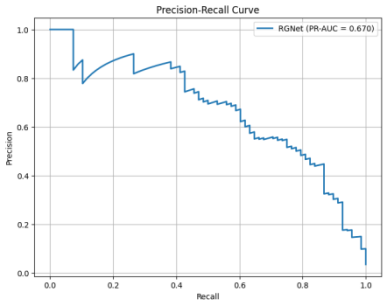


Fig. 2. PR-AUC calculation

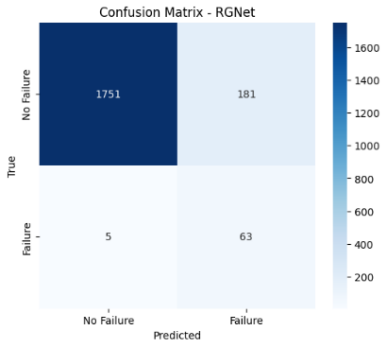


Fig. 3. Confusion matrix

A method for structural analysis of RG-flows is developed. It is important to note that false alarms in many practical applications are less critical than missing truly "bad" cases. In this case, the model issued a false warning 181 times, which, compared to the data volume, will not significantly hinder the work of technical services responsible for product quality.

An RG-flow of the last scale is defined as the vector $h^{(L)}(x) \in \mathbb{R}^{d_L}$. In the experiment, $d_L = 3$. This is the coarsest representation of the original sample.

A. Sensitivity Analysis

To assess the robustness of the architectural choices, a sensitivity analysis was conducted by varying the number of RG layers, the compression schedule, and the classifier capacity. Table III reports the results. The baseline configuration (5→4→3 with a 64→32 classifier) achieves the highest F1-score (0.4627) and a competitive PR-AUC (0.7197). The single-layer variant (5→3) attains the highest recall (0.9265), but suffers from a substantially lower precision (0.2658), resulting in a lower F1 (0.4131). Stronger compression (5→4→2) reduces both recall (0.8971) and PR-AUC (0.6554), indicating that overly aggressive dimensionality reduction discards discriminative information. Adding a fourth RG layer (5→4→3→2) does not improve performance (F1=0.4052), suggesting that two compression steps are sufficient for the 5-dimensional input space. Increasing the classifier capacity to 128→64 neurons actually degrades recall (0.8824), likely due to overfitting, while the smaller classifier (32→16) yields the lowest precision (0.2313). Overall, the baseline architecture is the most balanced and robust, confirming the heuristic choice of two RG blocks with a moderate classifier size.

TABLE III. SENSITIVITY ANALYSIS

Configuration	Recall	Precision	F1	PR-AUC
Baseline (5→4→3, cls 64→32)	0.9118	0.3100	0.4627	0.7197
5→3 (one layer)	0.9265	0.2658	0.4131	0.7106
5→4→2 (strong compression)	0.8971	0.3020	0.4519	0.6554
5→4→3→2 (extra layer)	0.9118	0.2605	0.4052	0.7080
cls 32→16 (smaller classifier)	0.9118	0.2313	0.3690	0.6651
cls 128→64 (larger classifier)	0.8824	0.2913	0.4380	0.7040

B. Ablation Study

The ablation study (Table IV) isolates the contribution of each architectural component. The multi-scale concatenation proves to be the most critical element: its removal causes a substantial drop in F1 from 0.6560 to 0.4672 (a 28.8% decrease) and a dramatic reduction in PR-AUC from 0.6952 to 0.3680, confirming that information from all coarse-graining scales is essential for reliable fault detection. The hierarchical RG compression also contributes positively: without RG blocks, F1 declines to 0.6358, indicating that coarse-graining helps suppress noise and extract collective failure patterns. Batch Normalization is crucial for training stability, as its removal drastically reduces recall from 0.6029 to 0.3235. Interestingly, removing Dropout slightly improves F1 (0.6560 vs. 0.6519), suggesting that the current dropout rate may be conservative for this small dataset. Similarly, removing L2 regularization yields a marginal improvement in recall (0.6912) at the cost of precision (0.4608), indicating that the L2 penalty may be slightly excessive. Overall, the concatenation of multi-scale representations and the RG compression are the primary drivers of RGNet's performance.

TABLE IV. ABLATION STUDY

Component	Recall	Precision	F1	PR-AUC	Threshold
Full RGNet	0.6029	0.7193	0.6560	0.6952	0.92
Without RG compression	0.8088	0.5238	0.6358	0.6618	0.86
Without concatenation	0.4706	0.4638	0.4672	0.3680	0.88
Without BatchNorm	0.3235	0.7586	0.4536	0.5912	0.94
Without Dropout	0.6471	0.6567	0.6519	0.7152	0.95
Without L2 regularization	0.6912	0.4608	0.5529	0.6548	0.70

C. Stability Analysis

Over five independent runs with different random seeds (42, 123, 456, 789, 101112), RGNet exhibits low variance across all metrics (Table V), confirming that the training procedure is stable and the results are reproducible.

TABLE V. STABILITY ANALYSIS

Metric	Mean	Std
Recall	0.9118	0.0084
Precision	0.2551	0.0186
F1	0.4000	0.0152
PR-AUC	0.7005	0.0124

D. Computational Efficiency

Table VI compares the computational efficiency of RGNet with the baseline methods. RGNet has 3,456 trainable parameters, which is only 38\% more than the MLP baseline (2,497). However, training RGNet for 30 epochs requires 40.70 seconds, substantially longer than XGBoost (0.12 s), LightGBM (0.19 s), and MLP (2.39 s). Inference time on 1,000 samples is 906.50 ms for RGNet, compared to 2.82 ms for XGBoost, 10.84 ms for LightGBM, and 8.07 ms for MLP.

TABLE VI. COMPUTATIONAL EFFICIENCY

Model	Parameters	Training (30 epochs, s)	Inference (1000 samples, ms)
RGNet	3,456	40.70	906.50
XGBoost	N/A	0.12	2.82
LightGBM	N/A	0.19	10.84
MLP (64,32)	2,497	2.39	8.07
Random Forest	N/A	1.32	13.06

This slower performance is an expected trade-off: RGNet is a neural architecture with hierarchical compression, multi-scale

concatenation, and batch normalization, which are computationally more intensive than the highly optimized tree-based implementations. Despite this, RGNet achieves the highest recall (0.9118) among all methods, outperforming XGBoost by 17.6 percentage points and LightGBM by 14.7 percentage points. For industrial deployment, the inference latency of RGNet (≈0.9 ms per sample) is acceptable for batch processing and non-real-time monitoring.

For visualization, the three-dimensional RG-flows of all test samples are projected onto a plane using t-SNE [29]. The resulting two-dimensional diagrams reveal a discrete curvilinear structure – a consequence of the limited set of unique RG-flows and t-SNE preserving local distances, connecting close discrete points into continuous lines. Statistical analysis shows that the set of unique vectors is limited (a few tens). Specifically, 187 out of 271 failures in the training set share the same RG-flow $t = (-0.612, -0.553, 0.698)$.

Fig. 4 shows the t-SNE projection of three-dimensional RG-flows onto the plane with t-SNE (perplexity = 30, random state = 42). Points are colored by true labels: blue – normal (class 0), red – failure (class 1). Smooth thin curves (the main "highway" for normal) and isolated clusters of failures are visible.

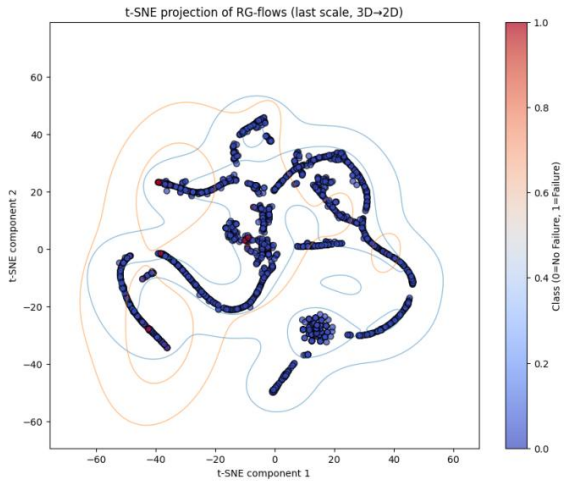


Fig. 4. t-SNE projection of RG-flows.

The points do not form a random cloud but align along several smooth curves (resembling parabolas or arcs). This is a consequence of the discrete nature of the RG representation: t-SNE, preserving local distances, connects close discrete points into continuous lines.

The dense cluster of blue points forms the main curve, which corresponds to normal operating modes. The spread along the curve reflects natural variation of parameters (e.g., temperature or rotational speed) within the normal range.

Red points (failures) are distributed in two ways. Some red points lie directly on the main curve interspersed with blue ones. These are failures that at the coarse scale are indistinguishable from normal – their RG-flow coincides with the normal one. Nevertheless, the classifier correctly recognizes these samples using finer scales (original features and output of the first RG layer). The other part of red points forms isolated clusters or separate "outgrowths" on the curves. Such samples have unique RG vectors and are classified by the model with high confidence (failure probability > 0.9). They correspond to failures with pronounced anomalies that manifest already at the coarse scale.

When projecting the original five features or the outputs of a conventional MLP (32 neurons) with t-SNE, a chaotic cloud without any structure is obtained. The appearance of sharp curves is a specific property of the RG representation, confirming the effectiveness of coarse-graining.

IV. Discussion

Comparison of RGNet with baseline methods. On the industrial AI4I2020 dataset with strong class imbalance (3.4 % failures), RGNet with threshold 0.5 attains a recall of 0.9118, 16–18 percentage points higher than gradient boosting. Threshold calibration to 0.92 yields a balanced variant (recall 0.5882, precision 0.6780, F1 = 0.630), suitable for practical applications where false alarms are less critical than missed failures.

The computational cost of RGNet (40.70 s training, 906.50 ms inference per 1,000 samples) is higher than tree-based methods, but the model remains compact (3,456 parameters) and is suitable for batch inference. The recall improvement justifies the computational overhead in mission-critical applications.

RG-flows (three‑dimensional representations) are visualized using t-SNE, revealing a discrete curvilinear structure. Isolated clusters of failures correspond to anomalies that the model recognizes with high confidence.

Analysis of RG-flows enables the following capabilities in the application domain: tracking how changes in physical parameters (e.g., an increase in tool wear or temperature rise) move the point along the curve. If the point enters an isolated cluster region, it signals high failure probability.

Thus, RG-flows serve as a bridge between the "black box" of the neural network and the physical understanding of the process, providing interpretability not available in classical machine learning methods.

V. Conclusion

This study proposes RGNet (Renormalization Group Network), a neural network architecture that implements the idea of the renormalization group for hierarchical coarse-graining of the feature space in fault prediction tasks. Both experimental and theoretical justifications of its effectiveness are provided.

A lemma on scale invariance is formulated and proved: if the RG layers preserve class membership under coarse-graining (i.e., the images of normal states and failures remain in disjoint compact attractors at each scale), then the classes remain distinguishable even after strong compression. These statements do not require global contraction of each layer but rely only on a natural property of the renormalization group: the large-scale structure (phase) is invariant under coarse-graining.

The concept of RG-flows is introduced – three-dimensional representations of the original data at the coarsest scale. Visualization via t-SNE reveals a discrete curvilinear structure: normal states align along a main "highway", while failures either merge with it (in which case the classifier relies on finer scales) or form isolated clusters. This visualization provides engineers with the ability to formulate threshold rules and to track the evolution of equipment state over time – something unavailable to classical black-box models.

Future research directions include automatic selection of the number of RG layers and dimensions, as well as integration of RGNet into real-time systems with online learning.

REFERENCES

[1] J. Wang, Y. Ma, L. Zhang, R. X. Gao, and D. Wu, "Deep learning for smart manufacturing: Methods and applications," J. Manuf. Syst., vol. 48, pp. 144–156, 2018.

[2] T. Zonta, C. A. da Costa, R. da Rosa Righi, M. J. de Lima, E. S. da Trindade, and G. P. Li, "Predictive maintenance in the Industry 4.0: A systematic review," Procedia Manuf., vol. 51, pp. 115–121, 2020.

[3] A. Albychev, A. Chervyakov, N. Gazanova, D. Ilin, E. Nikulchev, "Machine learning methods for deadline missing prediction using national project checkpoint data," Communications in Computer and Information Science, vol 2604, 2026. https://doi.org/10.1007/978-3-032-04761-8_19

[4] G. Haixiang, Y. Li, J. Shang, G. Mingyun, H. Yuanyue, and B. Gu, "Learning from class-imbalanced data: Review of methods and applications," Expert Syst. Appl., vol. 73, pp. 220–239, 2017.

[5] A. Angelopoulos et al., "A systematic review of anomaly and fault detection using machine learning for industrial machinery," Sensors, vol. 24, no. 8, p. 2588, 2024.

[6] H. Khorasgani and G. Biswas, "A methodology for monitoring and fault diagnosis of manufacturing systems," J. Manuf. Syst., vol. 51, pp. 15–30, 2019.

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

[8] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD'16), San Francisco, CA, USA, 2016, pp. 785–794.

[9] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in Adv. Neural Inf. Process. Syst. (NeurIPS), Long Beach, CA, USA, 2017, pp. 3146–3154.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.

[11] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," IEEE Trans. Syst., Man, Cybern. C, vol. 42, no. 4, pp. 463–484, 2012.

[12] C. X. Ling and V. S. Sheng, "Cost-sensitive learning and the class imbalance problem," in Encyclopedia of Machine Learning, C. Sammut and G. I. Webb, Eds. Boston, MA, USA: Springer, 2008, pp. 231–235.

[13] S.V. Zuev, "Geometric properties of quantum entanglement and machine learning." Russian Technological Journal, vol. 11, no. 5, pp. 19-33 https://doi.org/10.32362/2500-316X-2023-11-5-19-333

[14] K. G. Wilson, "The renormalization group: Critical phenomena and the Kondo problem," Rev. Mod. Phys., vol. 47, no. 4, pp. 773–840, 1975.

[15] L. P. Kadanoff, Statistical Physics: Statics, Dynamics and Renormalization. Singapore: World Scientific, 2000.

[16] S. Matsuura and M. Ohzeki, "Learning the renormalization group with a neural network," Phys. Rev. E, vol. 105, no. 4, p. 045311, 2022.

[17] C. B´eny, "Deep learning and the renormalization group," arXiv preprint arXiv:1301.3124, 2013.

[18] T. Poggio and A. Banburski, "Deep learning: Mathematics and neuroscience," Brain Sci., vol. 10, no. 7, p. 447, 2020.

[19] H. W. Lin, M. Tegmark, and D. Rolnick, "Why does deep and cheap learning work so well?" J. Stat. Phys., vol. 168, no. 6, pp. 1223–1247, 2017.

[20] M. R. Shadi, H. Mirshekali, and H. R. Shaker, "Explainable artificial intelligencefor energy systems maintenance: A review on concepts, current techniques, challenges, and prospects," Renewable Sustainable Energy Rev., vol. 216, p. 115668, 2025.

[21] A. P. Madathil, X. Luo, Q. Liu, C. Walker, R. Madarkar, and Y. Qin, " A review of explainable artificial intelligence in smart manufacturing," Int. J. Prod. Res., vol. 63, no. 23, pp. 8654‑8697, 2025.

[22] M. Faegh, S. Ghungrad, J. P. Oliveira, P. Rao, and A. Haghighi, "A review on physics-informed machine learning for process-structure-property modeling in additive manufacturing," J. Manuf. Processes, vol. 133, pp. 524‑555, 2025.

[23] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of Adam and beyond," in Int. Conf. Learn. Represent. (ICLR), 2018.

[24] Y. Liang, M. He, J. Liu et al., "Convergence of Adam for non-convex objectives: relaxed hyperparameters and non-ergodic case," Mach. Learn.,vol. 114, art. 75, 2025. doi:10.1007/s10994-025-06737-w

[25] A. Sridhar and A. Johansen, "Convergence of Adam in deep ReLU networks via directional complexity and Kakeya bounds," arXiv preprint arXiv:2505.15013, 2025.

[26] S. Boucheron, G. Lugosi, and P. Massart, Concentration Inequalities: A Nonasymptotic Theory of Independence. Oxford, U.K.: Oxford University Press, 2013.

[27] S. Matzka, Explainable artificial intelligence for predictive maintenance applications," in 2020 Third Int. Conf. Artif. Intell. Ind. (AI4I), 2020, pp. 69‑74.

[28] E. Nikulchev and A. Chervyakov, "Prediction intervals: A geometric view," Symmetry, vol. 15, no. 4, p. 781, 2023.

[29] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," Journal of Machine Learning Research, vol. 9, no. 86, pp. 2579–2605, 2008.

MatchIA-GCL: A Hybrid Graph Contrastive Learning Approach for Semantic Schema Matching

Mohamed Raoui, Moulay Hafid El Yazidi, Ahmed Zellou
ENSIAS, Mohammed V University in Rabat, Morocco

***Abstract*—Schema matching remains a fundamental challenge for achieving data interoperability across heterogeneous information systems. Existing deep learning-based approaches often suffer from semantic drift, overlook the structural topology of schemas, and operate as black-box models with limited interpretability. This study introduces MatchIA-GCL, an explainable hybrid framework for semantic schema matching that combines graph contrastive learning (GCL) and Large Language Models (LLMs). The proposed framework leverages Graph Attention Networks (GATs) optimized through a contrastive learning objective to generate robust structure-aware schema embeddings. Unlike conventional approaches that employ LLMs solely as semantic encoders, MatchIA-GCL integrates an LLM as an Autonomous Validation Agent responsible for resolving only complex and ambiguous matching candidates, thereby reducing computational overhead while improving decision reliability. Furthermore, a post-hoc explainability layer provides transparent and human-interpretable justifications for alignment decisions. Experimental evaluations conducted on benchmark schema matching datasets demonstrate that MatchIA-GCL consistently outperforms state-of-the-art methods, achieving up to 9.4% improvement in F1-score while enhancing explainability and robustness. These results highlight the potential of combining graph representation learning, contrastive optimization, and LLM-assisted validation for next-generation semantic schema matching systems.**

***Keywords*—Schema matching; data interoperability; graph contrastive learning; Graph Attention Networks; Large Language Models; explainable artificial intelligence; semantic data integration**

I. Introduction

The integration of large-scale heterogeneous data sources remains a persistent challenge in modern software architectures, enterprise information systems, and cloud-native ecosystems. At the core of this challenge lies schema matching, a fundamental task [1] that aims to identify semantic correspondences between attributes belonging to distinct data structures, including relational databases, NoSQL repositories, and Web APIs. As organizations increasingly rely on distributed and heterogeneous data environments, manual schema alignment becomes difficult to scale, costly, and inherently error-prone. Consequently, accurate and automated schema matching techniques have become essential for ensuring semantic interoperability, improving Extract–Transform–Load (ETL) processes, facilitating data virtualization, and supporting data-driven decision-making.

For several decades, schema matching research has primarily relied on heuristic and lexical approaches. These methods estimate the similarity between schema elements using string comparison metrics, syntactic distance measures, or external linguistic resources. While effective in highly standardized environments, such techniques often struggle when confronted with real-world heterogeneous datasets. In particular, they exhibit limited capability in handling semantic drift, where semantically equivalent concepts are represented using different terminologies, as well as structural ambiguity resulting from abbreviated, automatically generated, or domain-specific attribute names. As a consequence, their ability to capture deeper semantic relationships remains inherently constrained [2].

The emergence of deep learning has significantly transformed the field by replacing manually engineered rules with learned vector representations. Transformer-based language models such as BERT and RoBERTa [3, 4] have demonstrated remarkable effectiveness in capturing contextual semantics by jointly analyzing attribute names, data types, and textual descriptions. At the same time, representing schemas as graphs has encouraged the adoption of Graph Neural Networks (GNNs) [5–7], which can exploit structural dependencies, integrity constraints, foreign-key relationships, and hierarchical organizations. Nevertheless, many existing approaches still process semantic and structural information through separate learning pipelines, resulting in a semantic–structural disconnect that limits the quality of the generated representations. Furthermore, recent studies frequently employ Large Language Models (LLMs) [8–10] merely as semantic encoders, thereby underutilizing their reasoning and decision-making capabilities. The lack of transparency associated with deep neural architectures also remains a critical concern, particularly in domains where explainability, traceability, and auditability are mandatory requirements.

To address these limitations, this study introduces MatchIA-GCL, a hybrid and explainable framework for semantic schema matching that combines graph contrastive learning (GCL), Graph Attention Networks (GATs), and Large Language Models within a unified architecture. The proposed framework models schemas as heterogeneous graphs and learns robust structure-aware representations through a contrastive optimization objective designed to preserve semantic consistency under both structural and contextual perturbations. Unlike conventional approaches that rely extensively on generative models, MatchIA-GCL incorporates a Large Language Model as an Autonomous Validation Agent that is selectively activated only for ambiguous matching candidates and complex many-to-many correspondence scenarios. This strategy allows the framework to benefit from the reasoning capabilities of LLMs while maintaining computational

efficiency. In addition, an explainability layer is integrated to provide interpretable justifications for alignment decisions through attention analysis and gradient-based explanations.

The proposed framework makes several contributions to the schema matching literature. First, it introduces a unified learning paradigm that jointly exploits semantic and structural information through graph contrastive learning, enabling the generation of robust schema representations. Second, it proposes an original validation mechanism in which a Large Language Model acts as an autonomous decision-support agent for resolving uncertain correspondences rather than serving solely as a semantic encoder. Third, it incorporates an explainability layer that enhances the transparency and auditability of matching decisions, thereby increasing trust in automated data integration processes. Extensive experiments conducted on benchmark schema matching datasets demonstrate that MatchIA-GCL consistently outperforms existing approaches while providing improved robustness, interpretability, and generalization capabilities.

The remainder of this study is organized as follows: Section II reviews existing schema matching approaches and identifies the main research gaps. Section III formulates the schema matching problem and introduces its heterogeneous graph representation. Section IV presents the proposed MatchIA-GCL framework, including its semantic encoder, heterogeneous graph attention network, graph contrastive learning mechanism, LLM-based validation agent, explainability layer, and overall algorithm. Section V describes the experimental setup, benchmark datasets, evaluation metrics, comparative analysis, and ablation study. Finally, Section VI concludes the study and outlines future research directions.

II. Related Work

A. Lexical, Heuristic, and Traditional Algorithmic Approaches

Early schema matching approaches relied on lexical similarity measures and handcrafted heuristics. Representative systems such as Cupid, COMA++ [11], and Similarity Flooding estimate correspondences using attribute names, syntactic similarity metrics, and external lexical resources such as WordNet. Although these methods are simple and interpretable, they remain highly sensitive to naming conventions, abbreviations, and domain-specific terminology. To address these limitations, structural techniques such as Similarity Flooding exploit schema topology by propagating similarity scores across neighboring elements. However, their effectiveness largely depends on the quality of the initial lexical matches, and they often struggle with complex schema transformations and heterogeneous data models. These limitations have motivated the development of data-driven approaches capable of learning richer semantic and structural representations [12, 13].

B. Deep Learning and Embedding-Based Approaches

To overcome the limitations of lexical and heuristic methods, recent research has increasingly adopted machine learning and deep learning techniques for schema matching. Early learning-based approaches leveraged dense vector representations, commonly referred to as embeddings, to capture semantic relationships between schema elements. Models such as DeepMatcher adapted advances from Natural Language Processing (NLP) by employing distributed word representations, including Word2Vec and FastText, to project attribute names into continuous vector spaces where semantic similarity can be estimated through geometric proximity. These approaches demonstrated improved robustness to lexical variations and implicit synonymy compared with traditional string-based matching techniques. However, static embeddings assign a unique representation to each term regardless of its contextual usage, making them vulnerable to semantic ambiguity and domain-specific terminology.

The introduction of Transformer architectures has significantly advanced semantic schema matching by enabling contextual representation learning. Pre-trained language models such as BERT, RoBERTa, and DistilBERT generate dynamic embeddings that capture semantic relationships based on the surrounding context. In schema matching applications, attributes are commonly transformed into textual descriptions that combine metadata such as attribute names, data types, integrity constraints, documentation, and, in some cases, sampled data values.

C. Graph Neural Networks and Large Language Models

To address the loss of structural information inherent in text-based schema representations, recent research has explored graph-based approaches using Graph Neural Networks (GNNs), including Graph Convolutional Networks (GCNs) and Graph Attention Networks (GATs). By modeling foreign-key relationships, integrity constraints, and hierarchical dependencies, GNNs capture structural patterns that are not accessible to purely semantic models.

Recent hybrid architectures combine Transformer-based semantic embeddings with graph learning mechanisms. While these approaches generally improve matching performance, semantic and structural information are often fused through simple aggregation strategies and optimized separately, limiting their ability to capture complex semantic–structural interactions. Moreover, many graph-based methods remain highly dependent on labeled training data, reducing their generalization across domains [14].

In parallel, Large Language Models (LLMs) [15–17] have demonstrated strong capabilities in semantic understanding and reasoning. However, existing schema matching approaches typically use LLMs as semantic encoders or direct matching predictors, resulting in high computational costs and limited awareness of global schema structure [18].

These limitations reveal a gap between semantic reasoning and structural learning. To address this challenge, MatchIA-GCL integrates graph contrastive learning [19–21], heterogeneous graph attention, and selective LLM-based validation within a unified framework.

D. Research Gap Analysis

The review of existing schema matching approaches reveals several persistent limitations that remain insufficiently addressed in the current literature. First, a significant gap exists between semantic representation learning and structural modeling. While Transformer-based architectures excel at

capturing contextual semantics and Graph Neural Networks effectively exploit topological dependencies, most existing solutions learn these two dimensions through separate optimization processes. Consequently, the resulting representations often fail to fully capture the complex interactions between semantic context and schema topology.

Second, despite the growing adoption of Large Language Models [22], their reasoning capabilities remain largely underutilized in schema matching applications. Most current approaches employ LLMs either as semantic encoders or as direct matching predictors through prompt-based inference. Such usage overlooks their potential to act as intelligent decision-support mechanisms capable of validating uncertain correspondences, resolving ambiguities, and incorporating higher-level reasoning into the matching process. Moreover, applying LLMs indiscriminately to all candidate pairs introduces significant computational overhead and scalability challenges when dealing with large enterprise schemas.

Third, the explainability of schema matching decisions remains an open challenge. Many state-of-the-art deep learning approaches operate as black-box systems [23, 24], providing limited insight into the rationale behind generated correspondences. This lack of transparency can hinder their adoption in industrial environments where alignment decisions must be traceable, interpretable, and auditable by human experts. Table I summarizes the main characteristics of representative categories of schema matching approaches and highlights the positioning of the proposed MatchIA-GCL framework.

TABLE I. COMPARATIVE ANALYSIS OF EXISTING SCHEMA MATCHING APPROACHES.

Approach Category	Contextual Semantics	Structural Dependencies	Agent-Based Reasoning	Explainability
Heuristic Methods (COMA++, Cupid)	No	Partial	No	Yes
Deep Learning (DeepMatcher)	Yes	No	No	No
Graph-Based Approaches (GNNMatcher)	Yes	Yes	No	Limited
LLM-Based Approaches	Yes	No	No	No
MatchIA-GCL (Proposed)	Yes	Yes	Yes	Yes

Motivated by these observations, the next section formally defines the schema matching problem and introduces the mathematical foundations underlying the proposed MatchIA-GCL framework.

III. PROBLEM FORMULATION

A. Heterogeneous Graph Representation of Data Schemas

Traditional schema matching approaches often represent schemas as collections of independent attributes, thereby overlooking the structural dependencies that naturally exist within modern information systems. However, semantic relationships between schema elements are frequently influenced by their surrounding structural context, including table hierarchies, integrity constraints, and inter-entity relationships. To preserve both semantic and topological information, we model each data schema as a heterogeneous attributed graph.

Fig. 1 illustrates a representative schema matching scenario between two heterogeneous schemas describing the same entity. This example motivates the heterogeneous graph representation introduced in the following formulation.

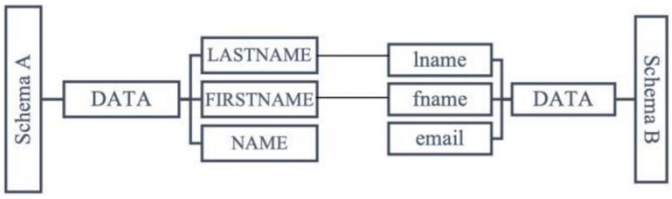


Fig. 1. Illustrative example of schema matching between two heterogeneous schemas describing the same entity.

Let S denote a data schema represented by a heterogeneous directed graph defined in:

$$G = (V, E, \tau_v, \tau_e, \mathcal{F}) \quad (1)$$

where, V denotes the set of nodes, $E \subseteq V \times V$ represents the set of directed edges, τ_v and τ_e are the node-type and edge-type mapping functions, respectively, and $\mathcal{X}$ defines the feature space associated with graph elements.

The node set is defined by:

$$V = V_T \cup V_A \quad (2)$$

where, V_T represents schema entities such as relational tables or document collections, while V_A denotes schema attributes such as columns or fields.

The table node set is formally defined in:

$$V_T = \{t_1, t_2, \dots, t_n\} \quad (3)$$

Similarly, the attribute node set is defined in:

$$V_A = \{a_1, a_2, \dots, a_m\} \quad (4)$$

The node-type mapping function is expressed by:

$$\tau_v : V \to T_v \quad (5)$$

where, each node belongs to one of the predefined categories given in:

$$T_v = \{Table, Attribute\} \quad (6)$$

The edge set captures the structural relationships among schema elements. The edge-type mapping function is defined by:

$$\tau_e : E \to T_e \quad (7)$$

The set of structural relation types is defined in:

$$T_e = \{e_{has}, e_{pk}, e_{fk}\} \quad (8)$$

where, e_{has} represents containment relationships between tables and attributes, e_{pk} denotes primary-key associations, and efk models foreign-key dependencies between related entities. These relations collectively preserve the structural organization of the schema and provide valuable contextual information for downstream matching tasks.

Each node is initialized with a feature vector:

$$x_i \in X \quad (9)$$

This feature vector aggregates the semantic metadata describing the corresponding schema element. Depending on data availability, these features may include attribute names, physical data types, textual descriptions, integrity constraints, and representative data values. Consequently, the graph representation offers a unified view that simultaneously captures both the semantic content and the structural organization of a schema.

Given two heterogeneous schema graphs representing the source and target schemas, respectively, the source graph is defined by:

$$G_s = (V_s, E_s) \quad (10)$$

The target graph is defined by:

$$G_t = (V_t, E_t) \quad (11)$$

The schema matching problem consists of identifying semantic correspondences between elements of the source graph and the target graph while preserving the structural consistency of both graph representations.

B. Extended Alignment Matrix and Correspondence Space

The primary objective of schema matching is to discover semantic correspondences between two heterogeneous schema graphs, namely the source schema (G_s) and the target schema (G_t). Formally, the matching process aims to learn a global correspondence matrix (M) that quantifies the semantic similarity between schema elements belonging to the two graphs.

The correspondence matrix is defined as a real-valued matrix of $size$ $(|V_s| \times |V_t|)$, as shown in:

$$M \in R^{|V_s| \times |V_t|} \quad (12)$$

Each element (M_{ij}) represents the confidence score associated with the semantic correspondence between a source node $(v_i \in V_s)$ and a target node $(v_j \in V_t)$. The confidence score is bounded within the interval defined in:

$$0 \leq M_{ij} \leq 1 \quad (13)$$

A value close to 1 indicates a strong semantic correspondence, whereas a value close to 0 suggests little or no semantic relationship between the two schema elements. The matching score can therefore be interpreted as the conditional probability defined in:

$$M_{ij} = P\big(match(v_i, v_j) \big| G_s, G_t\big) \quad (14)$$

While many schema matching methods assume one-to-one correspondences, real-world schemas frequently contain composite relationships such as one-to-many, many-to-one, and many-to-many mappings. Therefore, MatchIA-GCL considers both simple and composite correspondences when identifying semantic alignments between heterogeneous schemas. The resulting optimization problem is formally defined in:

$$M^* = argmax_M \sum_{v_i, v_j} M_{ij} \quad (15)$$

subject to semantic consistency and structural compatibility constraints across both schema graphs.

The objective is to maximize the confidence scores associated with valid correspondences while minimizing incorrect alignments and preserving the structural coherence of the source and target schemas. This formulation provides the mathematical foundation for the graph contrastive learning framework introduced in the next section.

C. Global Learning Objective and Optimization Criteria

MatchIA-GCL adopts a graph contrastive learning strategy to jointly learn semantic and structural representations, enabling robust schema matching across heterogeneous datasets.

Let (z_i) and (z_j) denote the latent representations of two positive nodes, corresponding either to aligned schema elements or to two augmented views of the same node. Following the InfoNCE principle, the contrastive loss associated with node (v_i) is defined as:

$$L_(i,j) = -log\left(\frac{exp\left(\frac{sim(z_i, z_j)}{\tau}\right)}{exp\left(\frac{sim(z_i, z_j)}{\tau}\right) + \Sigma_k exp\left(\frac{sim(z_i, z_k)}{\tau}\right)}\right) \quad (16)$$

where, $sim(\cdot)$ denotes the cosine similarity function, (τ) is the temperature hyperparameter, and (z_k) represents the latent representations of negative samples that do not correspond semantically to node (v_i).

The objective of the contrastive learning process is to maximize the similarity between positive pairs while simultaneously separating negative pairs in the latent embedding space. This mechanism encourages the generation of discriminative representations that remain stable under both semantic and structural perturbations [25, 26].

The overall optimization objective of MatchIA-GCL combines the contrastive loss with an $L2$ regularization term to improve model generalization and reduce overfitting. The global loss function is defined as:

$$L_{global} = \left(\frac{1}{N}\right) \Sigma_i L_{i,match(i)} + \lambda ||W||^2 \quad (17)$$

where, (N) denotes the number of training pairs within a mini-batch, $(match(i))$ identifies the positive correspondence associated with node (v_i) , (W) represents the trainable parameters of the Graph Attention Network, and (λ) is the regularization coefficient controlling the contribution of the weight decay term. The optimization process aims to minimize $(L_{\{global\}})$, thereby learning schema representations that simultaneously preserve semantic consistency, structural coherence, and robustness to heterogeneous schema transformations. These learned embeddings constitute the

foundation of the MatchIA-GCL architecture presented in the following section.

IV. Proposed MatchIA-GCL Approach

A. Overall Architecture

The proposed MatchIA-GCL framework is organized into three sequential processing stages, as illustrated in Fig. 2. In Stage 1 (Semantic Representation Learning), heterogeneous source and target schemas are transformed into contextual semantic embeddings using a Transformer-based encoder that captures lexical and contextual information from schema metadata. In Stage 2 (Structural Representation Learning), a Heterogeneous Graph Attention Network (HGAT) combined with Graph Contrastive Learning (GCL) models structural dependencies and learns robust hybrid representations by jointly optimizing semantic and structural information. These hybrid embeddings are then used to compute the correspondence matrix between source and target schema elements. Finally, in Stage 3 (Validation and Explainable Alignment), high-confidence correspondences are directly accepted, whereas ambiguous candidate pairs are selectively validated by the LLM-based Autonomous Validation Agent. An explainability layer subsequently provides interpretable justifications for the validated alignments, improving transparency and supporting human interpretation of the matching decisions.

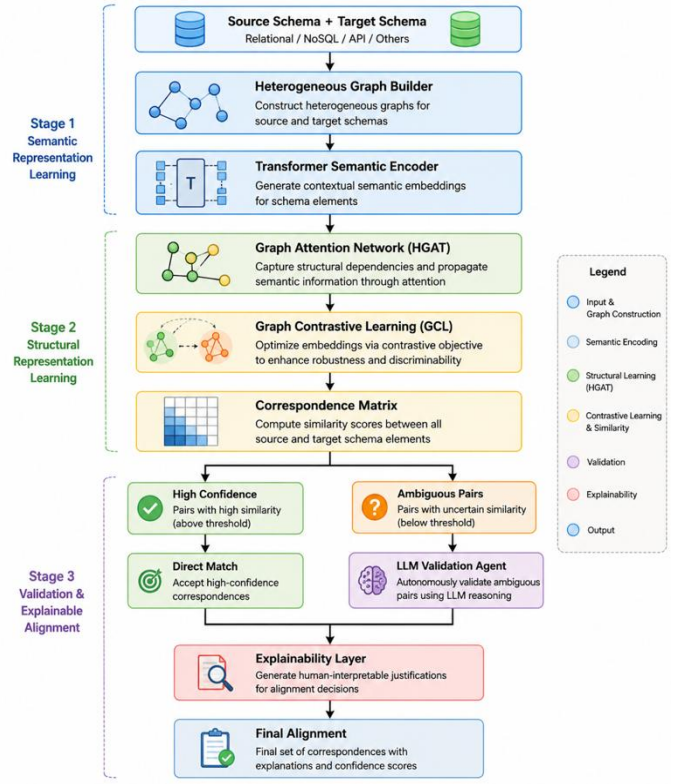


Fig. 2. Overall architecture and execution workflow of the proposed MatchIA-GCL framework.

B. Contextual Semantic Encoder

The first stage of MatchIA-GCL generates contextual semantic embeddings from schema metadata using a Transformer-based encoder. By leveraging attribute names, data types, and additional metadata, the model captures contextual semantics and reduces ambiguity between semantically related schema elements.

Let (C_i) denote the contextual description associated with node (v_i). The semantic embedding of the node is obtained by applying the Transformer encoder as defined in:

$$h_{sem_i} = TransformerEncoder(C_i) \quad (18)$$

where, $(h_{\{sem_i\}})$ represents the contextual semantic embedding generated for node (v_i).

The resulting embedding has a fixed dimensionality (d) and captures semantic relationships among schema elements independently of their structural position within the graph. These embeddings constitute the initial node features that are subsequently propagated through the graph attention layers described in the next subsection.

C. Heterogeneous Graph Attention Network

After obtaining the initial semantic embeddings, the framework activates its structural learning pipeline to incorporate topological information into the representation space. The semantic embeddings generated by the Transformer encoder serve as the initial node features of a Heterogeneous Graph Attention Network (HGAT).

To handle heterogeneous node types and relation categories, the attention mechanism is parameterized according to the relation type. For a node (v_i) connected to a neighboring node (v_j) through relation type (r), an attention coefficient is computed to quantify the structural importance of node (v_j) with respect to node (v_i). The unnormalized attention score is defined as:

$$e_{ij}^{r} = \text{LeakyReLU}(W_r[h_i||h_j]) \quad (19)$$

where, (W_v) denotes a shared linear projection matrix, (W_r) represents the attention vector associated with relation type (r), and (||) denotes vector concatenation.

The attention coefficients are subsequently normalized through a Softmax operation over the neighborhood of node (v_i), as defined in:

$$\alpha_{ij}^{r} = \frac{exp(e_{ij}^{r})}{\sum_{k \in N_i^r} exp(e_{ik}^{r})} \quad (20)$$

where, the summation is performed over all neighboring nodes connected to (v_i) through relation type (r).

The structural representation of node (v_i) is then obtained by aggregating the information received from its neighbors according to the normalized attention weights. This aggregation process is defined in:

$$h_i^{struct} = ELU\left(\sum_{j \in N_i^r} \alpha_{ij}^{r} W_v h_j\right) \quad (21)$$

To improve learning stability and capture multi-scale structural patterns, MatchIA-GCL employs a multi-head attention mechanism.

To better illustrate the structural learning process implemented in MatchIA-GCL, Fig. 3 presents the architecture

of the proposed multi-head Heterogeneous Graph Attention Network (HGAT) used for structural representation learning.

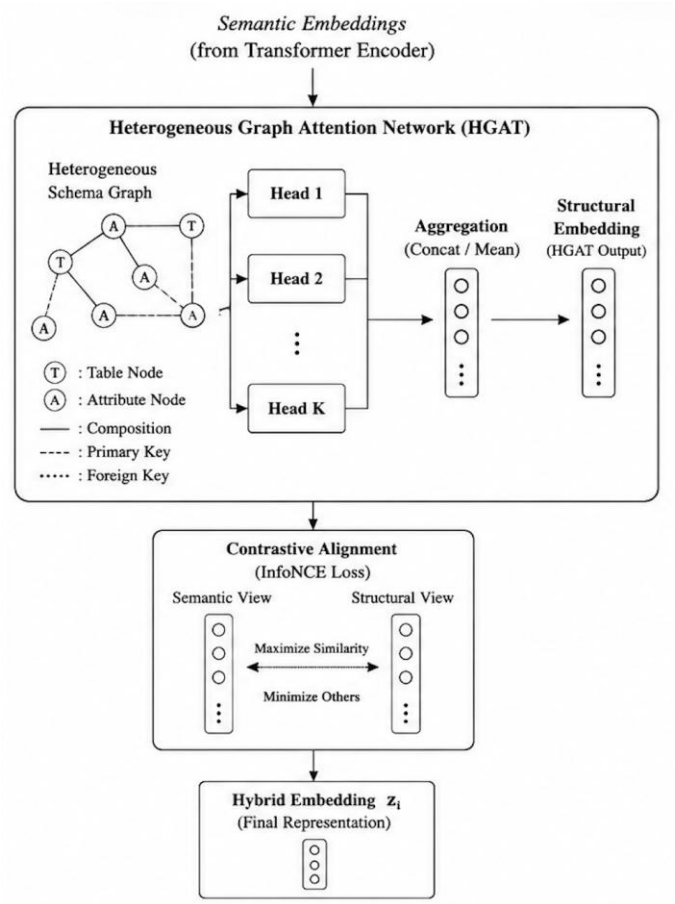


Fig. 3. Architecture of the proposed multi-head Heterogeneous Graph Attention Network (HGAT) used in MatchIA-GCL.

The HGAT module enables the propagation of semantic information across heterogeneous structural dependencies, including table–attribute relationships and referential constraints. The resulting structural embeddings are subsequently aligned with the semantic representations through the graph contrastive learning mechanism, producing robust hybrid embeddings suitable for schema matching.

D. Semantic–Structural Contrastive Alignment

The semantic embeddings generated by the Transformer encoder and the structural representations learned by the HGAT are subsequently aligned through a graph contrastive learning strategy. Rather than combining both representations through a simple aggregation operation, the proposed framework explicitly maximizes the agreement between the semantic and structural views of the same schema element.

The contrastive objective encourages semantically equivalent and structurally related nodes to occupy nearby regions in the latent space while simultaneously separating unrelated elements. Consequently, schema attributes sharing similar semantic roles or structural functions become geometrically closer even when their naming conventions differ significantly.

Positive pairs are constructed either from aligned schema elements available in the reference mappings or from two augmented views of the same schema node generated through feature masking and edge dropout. Negative pairs are randomly sampled from semantically unrelated nodes belonging to different schemas while ensuring that no reference correspondence exists between them. This sampling strategy enables the contrastive objective to maximize the similarity of semantically equivalent representations while effectively separating non-matching schema elements in the latent embedding space.

The final hybrid representation associated with node $(v_i)$is obtained by combining the semantic embedding and the structural embedding as defined in:

$$z_i = h_i^{\{sem\}} + h_i^{\{struct\}} \quad (22)$$

where, (z_i) denotes the final hybrid embedding used for correspondence estimation and downstream alignment tasks. These representations constitute the input of the matching and validation stage described in the next subsection.

E. LLM-Based Autonomous Validation Agent

Although the hybrid embeddings generated by MatchIA-GCL successfully identify most schema correspondences, ambiguous mappings often require additional semantic reasoning. To address this challenge, the framework incorporates a Large Language Model (LLM) as an Autonomous Validation Agent. Unlike conventional approaches that apply LLMs to all candidate pairs, the proposed agent is activated only for uncertain correspondences, thereby reducing computational overhead while maintaining scalability. For each ambiguous candidate, the agent analyzes the source and target schema elements together with their metadata and structural context. It then determines the validity and type of correspondence $(1{:}\,1,1{:}\,n,n{:}\,1,or\ n{:}\,m)$ and produces a concise justification for the decision. The validated correspondences are subsequently integrated into the global correspondence matrix, enabling the framework to combine graph-based representation learning.

F. Post-Hoc Explainability Layer

To improve transparency and facilitate adoption in industrial environments, MatchIA-GCL incorporates a post-hoc explainability layer that provides human-interpretable justifications for alignment decisions. The module combines attention-based structural analysis and semantic feature attribution to identify the factors that contribute to each correspondence. At the structural level, attention coefficients extracted from the HGAT layers highlight the influence of neighboring nodes and schema relationships. At the semantic level, feature attribution methods estimate the contribution of attribute names, data types, descriptions, and representative values to the matching decision.

The resulting explanations are aggregated into interpretable audit reports that provide both structural and semantic evidence supporting the generated correspondences. Fig. 4 illustrates the validation and explainability workflow integrated into MatchIA-GCL.

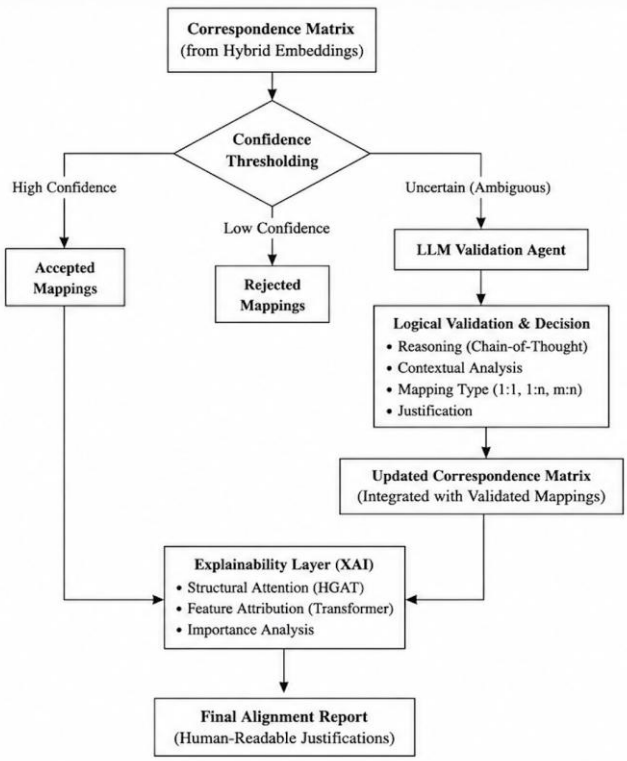


Fig. 4. Autonomous validation and explainability framework for schema matching in MatchIA-GCL.

G. Overall MatchIA-GCL Algorithm

The overall workflow of MatchIA-GCL is summarized in Algorithm 1. The algorithm integrates semantic representation learning, structural graph modeling, contrastive optimization, and selective LLM-based validation within a unified framework. Contextual semantic embeddings are first generated using a Transformer-based encoder and refined through a Heterogeneous Graph Attention Network (HGAT) to incorporate structural dependencies between schema elements. A graph contrastive learning objective is then optimized to improve the separability of semantically equivalent and non-equivalent representations in the latent space. Candidate correspondences are subsequently estimated using cosine similarity between the learned hybrid embeddings, whereas only uncertain candidates are subjected to LLM-based semantic validation. The validated correspondences constitute the final schema alignment produced by MatchIA-GCL.

Algorithm 1: MatchIA-GCL: Schema Matching Framework

Input : Source schema S_s, Target schema S_t
Parameters θ (encoder, HGAT, projection, LLM),confidence threshold τ
Output : Set of correspondences C

Initialize $C \leftarrow \emptyset$
Construct heterogeneous schema graph *G = (V, E, R)* from S_s *and* S_t
/ Stage 1: Contextual Semantic Encoding */*
for each node $v \in V$ **do**
  $x_v \leftarrow build\ v(name, type, parent, description, samplevalues)$
  $h_v^{(0)} \leftarrow TransformerEncoder(x_v; \theta_e nc)$
end for
/ Stage 2: Heterogeneous Graph Attention Network */*
for $l = 1\ to\ L$ **do**
  Update node representations using HGAT layer l
end for
/ Stage 3: Graph Contrastive Learning */*
Construct positive and negative pairs from S_s and S_t
Compute contrastive loss L_{cl} and update parameters θ
/ Stage 4: Similarity Computation */*
for each pair $(u \in Ss, v \in St)$ **do**
  $sim(u,v) \leftarrow cosinSimilarity\left(h_u^{(L)}, h_v^{(L)}\right)$
end for
/ Stage 5: LLM-based Validation */*
for each candidate pair (u, v) with $sim(u,v) \geq \tau$ **do**
  $decision, score \leftarrow LLMValidation(u, v, context_{uv}, sim(u,v))$
  if decision == match **then**
    $C \leftarrow C \cup (u, v, score)$
  end if
end for
return *C*

The following section presents the experimental evaluation of the proposed framework.

V. Experimental Evaluation

A. Experimental Setup and Evaluation Protocol

To assess the effectiveness of the proposed MatchIA-GCL framework, a comprehensive experimental protocol was designed to evaluate its matching accuracy, robustness, scalability, and generalization capabilities across heterogeneous schema matching scenarios. All experiments were conducted using a Python-based implementation relying on the PyTorch and PyTorch Geometric frameworks for deep learning and heterogeneous graph processing.

The semantic encoding module employs a pre-trained RoBERTa-base model to generate contextual embeddings from schema metadata. The structural learning component is implemented using a Heterogeneous Graph Attention Network composed of three attention layers with eight attention heads per layer. The graph contrastive learning module is optimized using the InfoNCE loss function, with the temperature parameter empirically fixed to 0.07 following a hyperparameter tuning phase.

For ambiguous correspondences requiring higher-level semantic reasoning, the autonomous validation module relies on OpenAI GPT-4o accessed through the OpenAI API. The model is integrated through a standardized inference interface and is selectively activated only for candidate correspondences whose similarity score falls within a predefined uncertainty interval.

B. Benchmark Datasets

The experimental evaluation was conducted on three benchmark datasets commonly used in schema matching research. These datasets cover diverse semantic and structural challenges, including noisy schemas, semantic heterogeneity, multilingual metadata, and complex correspondences.

The first benchmark considered in this study is the Valentine Benchmark, a well-established evaluation framework frequently used in schema matching research [1]. This benchmark includes both academic and industrial datasets originating from domains such as finance, healthcare, and business information systems. A key advantage of the Valentine benchmark lies in its ability to generate degraded schema variants through controlled transformations, including attribute renaming, abbreviation insertion, typographical noise, and structural perturbations. Such characteristics make this benchmark particularly suitable for evaluating the robustness of representation learning methods under realistic and adverse conditions.

The second benchmark is the Open Public Data (OPD) dataset, which consists of real-world schemas extracted from governmental open-data portals and public information repositories. Compared with curated academic datasets, OPD exhibits significantly higher semantic heterogeneity due to inconsistent naming conventions, multilingual terminology, undocumented abbreviations, and varying normalization levels. Consequently, this benchmark provides a realistic environment for assessing the effectiveness of semantic representation learning approaches in large-scale data integration scenarios [27].

The third benchmark is the Wikidata-to-Schema dataset, which focuses on aligning relational schemas with semantic web and knowledge graph structures derived from Wikidata. It includes hierarchical relationships and many-to-many correspondences [28], making it particularly suitable for evaluating the reasoning capabilities of the proposed LLM-based validation component. The three benchmarks provide complementary evaluation scenarios covering robustness, semantic generalization, and complex semantic reasoning. Table II summarizes the main characteristics of the benchmark datasets used throughout the experimental evaluation.

TABLE II. CHARACTERISTICS OF THE BENCHMARK DATASETS

Dataset	Application Domain	Number of Schemas	Schema Size (Attributes / Tables)	Correspondence Type
Valentine Benchmark	Academic & Industrial	14 pairs	43 to 70 attributes / 2 to 15 tables per schema	1:1, Textual & Complex
Open Data (OPD)	Public Governance	22 pairs	30 to 120 attributes / 1 table per flat file	1:1, Abbreviations & Synonyms
Wikidata-to-Schema	Semantic Web	8 Graphs	> 500 relational properties / Graph-Ontologies	m:n, Composite & Hierarchical

As shown in Table II, the selected benchmark datasets encompass a wide range of schema matching challenges, from noisy relational schemas to highly complex ontology-driven structures. This diversity enables a rigorous assessment of the semantic, structural, and reasoning capabilities of the proposed MatchIA-GCL framework.

C. Evaluation Metrics

The performance of MatchIA-GCL was evaluated using the standard metrics commonly adopted in schema matching research, namely Precision, Recall, and F1-Score. The generated correspondences were compared against manually validated reference alignments (ground truth) to assess matching quality. Precision measures the proportion of correctly identified correspondences among all predicted matches and is defined by:

$$Precision = TP/(TP + FP) \tag{23}$$

where, TP denotes the number of true positive correspondences and FP denotes the number of false positive correspondences.

Recall evaluates the ability of the framework to identify all valid correspondences present in the reference alignment. This metric measures the sensitivity of the model with respect to missed matches and is defined by:

$$Recall = TP/(TP + FN) \tag{24}$$

where, FN represents the number of false negative correspondences.

The F1-Score, defined as the harmonic mean of Precision and Recall, is adopted as the primary evaluation metric because it provides a balanced assessment of matching performance by jointly considering false positive and false negative correspondences. It is particularly suitable for schema matching, where both accurate correspondence identification and error minimization are equally important. It is computed according to:

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{25}$$

In addition to matching accuracy, execution time and inference efficiency were evaluated to assess the computational scalability of MatchIA-GCL. These metrics quantify the practical cost of semantic representation learning and selective LLM-based validation, providing insight into the framework's suitability for large-scale heterogeneous schema matching scenarios.

D. Comparative Results and Quantitative Analysis

The Precision results obtained by COMA++, DeepMatcher, GNNMatcher, LLM-Centric, and the proposed MatchIA-GCL are presented in Fig. 5. MatchIA-GCL achieves the highest Precision across all benchmark datasets, reaching 90.3%, 92.7%, and 88.6% on the Valentine, OPD, and Wikidata-to-Schema datasets, respectively. The consistent improvement over both traditional and deep learning-based baselines indicates that the joint integration of contextual semantic representations, heterogeneous graph attention, and graph contrastive learning effectively reduces false positive correspondences. Furthermore, the selective activation of the LLM-based validation agent refines ambiguous candidate mappings without introducing

unnecessary validation overhead, leading to more reliable semantic alignments across heterogeneous schemas.

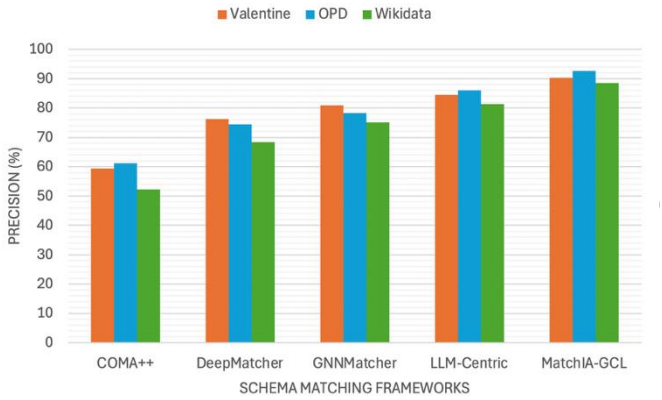


Fig. 5. Precision comparison across benchmarks.

The Recall results are reported in Fig. 6. MatchIA-GCL consistently outperforms competing approaches, achieving Recall values of 88.1%, 90.1%, and 86.4% on the three benchmark datasets. This demonstrates its effectiveness in identifying a larger proportion of valid correspondences.

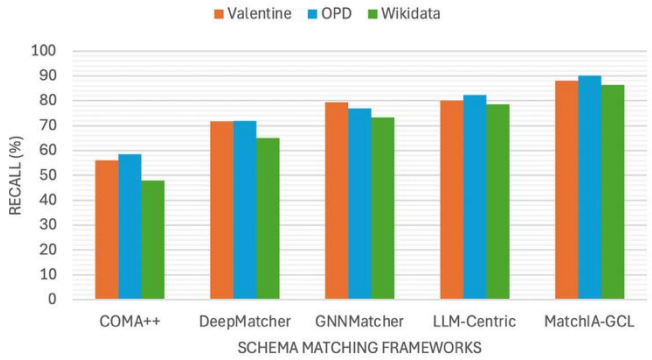


Fig. 6. Recall comparison across benchmarks.

The overall matching performance measured by the F1-Score is presented in Fig. 7. MatchIA-GCL consistently achieves the highest results across all benchmark datasets, reaching 89.2%, 91.4%, and 87.5% on the Valentine, OPD, and Wikidata-to-Schema benchmarks, respectively. The observed improvements demonstrate the effectiveness of jointly learning semantic and structural representations through the proposed hybrid architecture. The largest gain is obtained on the Wikidata-to-Schema benchmark, where MatchIA-GCL outperforms the strongest baseline by 7.6 percentage points in terms of F1-Score. This benchmark contains ontology-driven structures, hierarchical dependencies, and many-to-many semantic correspondences, making schema alignment particularly challenging. The results indicate that the integration of heterogeneous graph attention learning, graph contrastive optimization, and LLM-assisted validation enables the framework to better capture complex semantic and structural relationships than existing approaches.

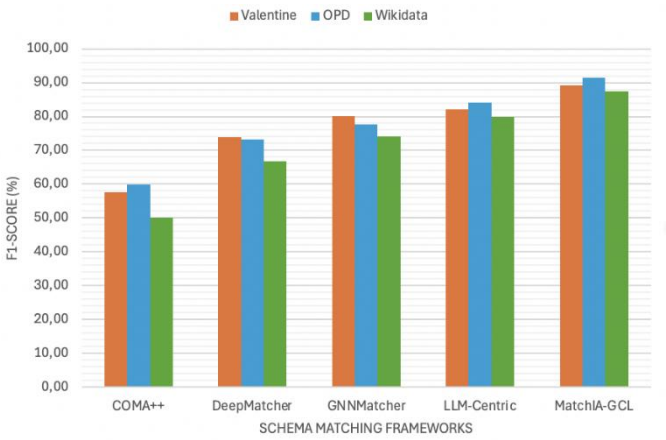


Fig. 7. F1-Score comparison across benchmarks.

Overall, the results confirm the effectiveness of combining contextual semantic encoding, heterogeneous graph attention learning, graph contrastive optimization, and LLM-assisted validation within a unified schema matching framework.

E. Ablation Study and Component Contribution Analysis

To assess the contribution of each component of MatchIA-GCL, an ablation study was conducted by removing one module at a time while keeping the remaining architecture unchanged. Four degraded variants were evaluated.

- MatchIA w/o HGAT removes the Heterogeneous Graph Attention Network and relies solely on Transformer-based semantic embeddings, thereby evaluating the contribution of structural information.
- MatchIA w/o GCL removes the Graph Contrastive Learning objective while preserving the Transformer encoder and the HGAT module, thereby evaluating the contribution of contrastive optimization to robust semantic-structural representation learning.
- MatchIA w/o LLM disables the LLM-based Autonomous Validation Agent and relies exclusively on embedding similarity scores, thereby evaluating the contribution of semantic reasoning to resolving ambiguous correspondences.
- MatchIA w/o Context replaces the Transformer-based contextual semantic encoder with basic attribute representations, thereby evaluating the importance of contextual semantic information in generating discriminative schema embeddings.

The performance differences between these variants and the complete MatchIA-GCL framework provide insights into the contribution of semantic context, structural learning, contrastive optimization, and LLM-based reasoning to the overall matching performance. Table III summarizes the performance degradation observed after removing each component of the proposed architecture.

TABLE III. ABLATION STUDY: F1-SCORE DEGRADATION (%)

Framework Variant	Valentine	OPD	Wikidata
MatchIA w/o HGAT	-9.0	-6.8	-11.3
MatchIA w/o GCL	-5.4	-4.1	-5.9
MatchIA w/o LLM	-4.2	-7.9	-12.1
MatchIA w/o Context	-7.1	-5.3	-3.6

Overall, the ablation study validates the complementarity of the proposed components.

VI. CONCLUSION AND FUTURE WORK

MatchIA-GCL introduces a unified framework for heterogeneous schema matching that combines contextual semantic encoding, heterogeneous graph attention learning, contrastive optimization, and LLM-assisted validation. By jointly modeling semantic and structural information, the proposed approach effectively addresses the limitations of existing schema matching methods and improves the identification of complex correspondences across heterogeneous data sources. Experimental evaluations on the Valentine, OPD, and Wikidata-to-Schema benchmarks demonstrate consistent improvements over representative state-of-the-art approaches, while the ablation study confirms the contribution of each architectural component. Furthermore, the integration of an explainability layer enhances the transparency and auditability of alignment decisions, facilitating the adoption of the framework in real-world data integration environments. Future work will focus on scaling the framework to large enterprise data ecosystems and incorporating adaptive learning mechanisms that continuously refine matching decisions through expert feedback and autonomous reasoning.

REFERENCES

[1] Koutras, C., Siachamis, G., Ionescu, A., Psarakis, K., Brons, J., Fragkoulis, M., ... & Katsifodimos, A. (2021, April). Valentine: Evaluating matching techniques for dataset discovery. In 2021 IEEE 37th International Conference on Data Engineering (ICDE) (pp. 468-479). IEEE.

[2] Li, H., Govind, Y., Mudgal, S., Rekatsinas, T., & Doan, A. (2021). Deep learning for semantic matching: A survey. Journal of Computer Science and Cybernetics, 37(4), 365-402.

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) (pp. 4171-4186).

[4] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[5] Hamilton, W. L. (2020). Graph representation learning. Morgan & Claypool Publishers.

[6] Vrahatis, A. G., Lazaros, K., & Kotsiantis, S. (2024). Graph attention networks: a comprehensive review of methods and applications. Future Internet, 16(9), 318.

[7] Shi, C. (2022). Heterogeneous graph neural networks. In Graph neural networks: Foundations, frontiers, and applications (pp. 351-369). Singapore: Springer Nature Singapore.

[8] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

[9] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

[10] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Rozière, B., ... & Szafraniec, M. (2024). Llama 3: Open foundation and instruction models. Meta AI.

[11] Aumueller, D., Do, H. H., Massmann, S., & Rahm, E. (2005, June). Schema and ontology matching with COMA++. In Proceedings of the 2005 ACM SIGMOD international conference on Management of data (pp. 906-908).

[12] Yousfi, A., El Yazidi, M. H., & Zellou, A. (2020, November). An Efficient Holistic Schema Matching Approach. In International Conference on Information and Communication Technology and Applications (pp. 588-601). Cham: Springer International Publishing.

[13] Yousfi, A., El Yazidi, M. H., & Zellou, A. (2020, December). CSSM: A Context-Based Semantic Similarity Measure. In 2020 IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS) (pp. 1-6). IEEE.

[14] Raoui, M., Ennaouri, M., El Yazidi, M. H., & Zellou, A. (2024). SchemaLogix: Advancing Interoperability with Machine Learning in Schema Matching. parameters, 6, 7.

[15] Parciak, M., Vandevoort, B., Neven, F., Peeters, L. M., & Vansummeren, S. (2024). Schema matching with large language models: an experimental study. arXiv preprint arXiv:2407.11852.

[16] Raoui, M., Ennaouri, M., El Yazidi, M. H., & Zellou, A. (2025, July). Next-Gen Schema Matching: How Machine Learning Enhances Data Interoperability. In 2025 9th International Conference on Computer, Software and Modeling (ICCSM) (pp. 175-180). IEEE.

[17] Seedat, N., & van der Schaar, M. (2025). Matchmaker: Schema matching with self-improving compositional LLM programs.

[18] Liu, Y., Pena, E., Santos, A., Wu, E., & Freire, J. (2024). Magneto: Combining small and large language models for schema matching. arXiv preprint arXiv:2412.08194.

[19] You, Y., Chen, T., Sui, Y., Chen, T., Wang, Z., & Shen, Y. (2020). Graph contrastive learning with augmentations. Advances in neural information processing systems, 33, 5812-5823.

[20] Zhu, Y., Xu, Y., Yu, F., Liu, Q., Wu, S., & Wang, L. (2020). Deep graph contrastive representation learning. arXiv preprint arXiv:2006.04131.

[21] Ju, W., Wang, Y., Qin, Y., Mao, Z., Xiao, Z., Luo, J., ... & Zhang, M. (2024). Towards graph contrastive learning: A survey and beyond. arXiv preprint arXiv:2405.11868.

[22] Freire, J., Fan, G., Feuer, B., Koutras, C., Liu, Y., Peña, E., ... & Wu, E. (2025). Large language models for data discovery and integration: Challenges and opportunities. IEEE Data Engineering Bulletin.

[23] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM computing surveys (CSUR), 51(5), 1-42.

[24] Lundberg, S., & Lee, S. I. (2017). Shap: A unified approach to interpreting model predictions. Advances in neural information processing systems, 1-10.

[25] Ju, W., Gu, Y., Mao, Z., Qiao, Z., Qin, Y., Luo, X., ... & Zhang, M. (2025). GPS: Graph contrastive learning via multi-scale augmented views from adversarial pooling. Science China Information Sciences, 68(1), 112101.

[26] Liu, Y., Zhao, Y., Wang, X., Geng, L., & Xiao, Z. (2024). Multi-scale subgraph contrastive learning. arXiv preprint arXiv:2403.02719.

[27] Miller, R. J. (2018). Open data integration. Proceedings of the VLDB Endowment, 11(12), 2130-2139.

[28] Vrandečić, D., & Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. Communications of the ACM, 57(10), 78-85.

Optimized Lattice-Based Ciphertext Policy Attribute-Based Encryption for Quantum-Resistant Secure Data Sharing

Purnima[1], Tejaswi Khanna[2], Deepak Kumar Verma[3]
Research Scholar, Amity School of Engineering and Technology, Amity University, Greater Noida, India[1]
Amity School of Engineering and Technology, Amity University, Greater Noida, India[2]
Department of Computer Engineering, Marwadi University, Rajkot, Gujarat, India[3]

***Abstract*—The practical utilization of the conventional Ciphertext-Policy Attribute-Based Encryption (CP-ABE) systems in a vast cloud has the constraint of the high cost of computation and the high risk of quantum-powered adversaries. This research introduces a novel Hybrid Lattice-Based CP-ABE (HL-CP-ABE) architecture that is focused on the hardness assumptions of Ring Learning with Errors (RLWE) to provide the double security of quantum-safe and precise data sharing. The proposed hybrid encryption system employs post-quantum techniques that provide both confidentiality and access control, and it is completely based on AES for the data encryption and on the lattice-based CP-ABE technology for the AES session key wrapping. A multithreaded, chunk-based parallel processing technique is used to improve the encryption and decryption processes and, at the same time, limit the difficulties that come along with the performance of lattice-based cryptography. The entire system is implemented in Java and evaluated across file sizes from 100 to 800 MB under a fixed three-attribute access policy, achieving processing time reductions of up to 9.5× compared to serial lattice CP-ABE, with a mean speedup of approximately 8.2× across encryption and decryption. The results reveal that the design not only upholds security but also achieves scalability, reduced ciphertext growth, and the correct decryption of the whole data. This research study presents the groundwork for future optimization using GPU parallelization and dynamic attribute management, and at the same time, it puts forth a workable, effective, and quantum-secure method of attribute-based encryption that is specifically beneficial for electronic health record (EHR) systems and secure cloud data sharing.**

Keywords*—LWE; Ring-LWE; post-quantum cryptography; attribute based encryption; parallel processing

I. Introduction

Due to their scalability, live collaboration, and data access from everywhere, cloud computing and distributed data environments nowadays have turned into the necessities of any information system. Nevertheless, these benefits also introduce significant dangers to privacy arising from these benefits, and these dangers get highlighted when using untrustworthy infrastructure to transmit sensitive data [1] [2]. The common use of digital health data, e-governance, and financial platforms does not only require cryptographic frameworks that secure data with secrecy, authenticity, and fine-grained access control, but it also promotes their use. On the other hand, the protection offered by public-key cryptography systems such as RSA and ECC, which are based on the integer factorization and discrete logarithm issues, will be completely broken if quantum computers are used [3]. The public-key cryptosystems in the area of quantum computing are at risk of quantum algorithms like Shor's or Grover's. Hence, it becomes a necessity to innovate cryptography techniques that are quantum-proof to allow secure data exchange between quantum computers. Attribute-Based Encryption (ABE) is acknowledged as a practical method for encrypted data access control of the highest degree, thereby ensuring that the information can only be accessed by the users who are permitted [4]. ABE, the encrypted message is tied to a receiver who can decrypt it based on a policy that conforms to certain attributes. These users are only allowed to view the file that satisfies the requirements with their attribute sets [5]. A crucial factor is the distinction between the two main types of encryption using an attribute-based approach, namely, Key-Policy ABE (KP-ABE) and Ciphertext-Policy ABE (CP-ABE) [6]. The interest in data sharing applications has been along the side of CP-ABE because the policy for access was set right at the moment of encryption by data owners [7].

In contrast, many traditional CP-ABE scheme implementations are based on pairing of groups formed by elliptic curves; this property makes the systems computationally intensive and open to attack within the domain of quantum computing [8]. The time-consuming nature of the bilinear pairings in the CP-ABE schemes results in high latency during encryption and decryption processes and difficulties in the expansion of the technology to large or real-time applications mainly because of the dependence of the schemes on the access structures' size [9]. Hence, even though ABE is a theoretically attractive framework, its application is still limited in an environment where high throughput is the main concern and quantum computers are considered a threat [10]. Of late, recently, lattice-based cryptography has emerged as the leading post-quantum approach. The primary limitations of these systems include large ciphertext sizes, high key-generation overhead, and lack of parallel optimization, which keep attackers out from both classical and quantum sides [8]. The hard problems like LWE and its Ring-based version (Ring-LWE) are a barrier to attackers coming from both quantum and classical worlds. The lattice structure practically levels the playing field between optimal and typical situations, thereby speeding up arithmetic operations in the polynomial ring modulo and making them even more suitable for cryptographic

primitives like ABE and functional encryption [11]. To put it in another way, several lattice-based ABE schemes have been proposed, which, by means of theory, are found to be resistant to quantum algorithms [12]. The drawback of such appetitive features is that, with time, these designs have frequently become prey to the problems of enormously bulky ciphertext sizes, big key generation overhead, and the limitation of performance optimization [13]. One of the main problems of the current implementations is that they were made for academic purposes only, whereas lattice-based and attribute-based encryption (ABE) can be said to be more practical. This means that the distance between the theoretical discussions on post-quantum ABE and technological solutions capable of handling massive, encrypted datasets is actually very far apart [14].

One way of overcoming these limitations is by means of methods that integrate the best of symmetric cryptography and public-key techniques in the so-called hybrid cryptographic frameworks. Public-key mechanisms and symmetric encryption techniques together speed key-based transactions [15]. Within a hybrid system, data encryption is carried out by means of a symmetric cipher like AES, but the symmetric key is protected by a public-key method, such as ABE [16]. This idea leads to a significant drop in the expenses of calculations, as the public-key operations are performed only on the key, not on the whole data set [14]. Nevertheless, the lattice-based methods still present large delays in the computation aspect, although they are hybridized. This is due to polynomial operations and modular arithmetic inherent in lattice schemes. Most existing hybrid implementations are single-threaded and therefore inadequate for large-file scenarios. Therefore, the proposed scheme must be compatible with public-key mechanisms while supporting parallel, pipelined execution for high-throughput scenarios. This study provides a modified CP-ABE algorithm that uses lattice structures in conjunction with classical cryptographic constructions to prevent against quantum attacks. The system is composed of the LSSS policy matrices as a base, with the Ring-LWE assumption, which makes the system post-quantum secure and helps define detailed levels of access control. The next step after the AES is made to encrypt the data very quickly is the AES key that is covered through the Ring-LWE-based CP-ABE mechanism. The start of the encryption process with the new encryption method is anticipated to result in the modification of the system that allows one to encrypt and decrypt segments of files in parallel across multiple cores of the CPU by having a concurrency method easier and cheaper than the one currently in use. Having set up a quantum key distribution technique, the described system would be working concurrently on all channels, ensuring total security. The introduction of the multithreaded chunk-processing model is the key to making the solution highly efficient while at the same time keeping it secure and correct. The system is fully implemented in Java, with modules designed for efficient memory management. There has been a very exhaustive experimental analysis to confirm the efficiency and scalability of the suggested system. Moreover, in the performance evaluation, several different file sizes, attribute policies, and thread counts were taken into consideration when comparing the optimized and complex multithreaded hybrid model with the conventional and serial lattice-based CP-ABE implementations. The evaluation shows up to a 9.5× reduction in runtime and a substantial decrease in ciphertext size, while preserving fully accurate decryption. Thus, the outcome clearly demonstrates that the mixed approach used in architecture was secure and widely acceptable in terms of system speed and total operational time—two important characteristics for the various kinds of tasks, like secure cloud storage, collaborative data sharing, and electronic health records, that may likely be done in real-world scenarios.

A. Highlights of the Research Paper

- Quantum-Secure Hybrid Model: With Ring-LWE-based cryptography, a new CP-ABE system was developed, and it eliminated the risk of quantum attacks.
- Performance-Optimized Architecture: A parallelizable encryption and decryption flow for big files, through the application of multithreaded chunk-processing design, has been introduced.
- Practical Implementation: A full Java-based system has been created by integrating lattice cryptography with AES and LSSS-based policy parsing.
- Thorough Evaluation: The proposed system showed processing times up to 9.5× faster than serial execution (mean≈8.2×), while maintaining fully accurate decryption and ciphertext overhead below 2×.

This study follows the outlined order: In Section II, the research of lattice-based ABE and hybrid cryptographic ways is considered. In Section III, some important mathematical principles are included, such as LWE, Ring-LWE, and LSSS. Section IV delineates the proposed system, encompassing setup procedures, key generation, encryption, decryption, and the corresponding algorithms for these operations. Section V presents the implementation and verifies system efficiency. Security is covered in Section VI. Section VII compares the proposed system against existing pairing-based, lattice-based, and hybrid schemes. Section VIII concludes the study and outlines future work.

II. Literature Review

A. Classical Pairing-Based Attribute-Based Encryption

Attribute-Based Encryption (ABE) was first put forward by Sahai and Waters [59] as an upgrade of Identity-Based Encryption (IBE), granting fine-grained access rights over encrypted data. The seminal scheme employed bilinear pairings defined over cyclic groups $(G1, G2)$ of prime order p, its security being based on the Decisional Bilinear Diffie–Hellman (DBDH) hypothesis. After that, Goyal et al. [18] and Bethencourt et al. [19] proposed models along with this model in the form of Key-Policy ABE (KP-ABE) and Ciphertext-Policy ABE (CP-ABE), respectively, which allowed the owners of the data to put the access structures in the encrypted data directly. The pairing-based schemes enabled highly expressible Boolean or Linear Secret Sharing Scheme (LSSS)-based policies, but they involved a large computational burden in terms of pairing and exponent operations of the form $(ga, gb) = (g, g)\{ab\}$, which are the major contributors to the runtime complexity [20]. Despite several improvements, the security of pairing-based protocols against quantum attacks remains elusive to trace [21]. "Shor's algorithm solves integer factorization and

discrete logarithm problems in polynomial time, breaking RSA and ECC security.", this being the very core of security for RSA and elliptic curve systems [22]. The traditional methods being inadequate for protecting against quantum threats drove the development of lattice-based cryptography, which relies on the hardness of the toughest lattice problems, thus making it impractical in both the classical and quantum realms [23].

B. Lattice-Based and Post-Quantum ABE Approaches

The move from pairings to lattice problems was first taken by the path-breaking work of Gentry, Peikert, and Vaikuntanathan, who put forward the Learning with Errors (LWE) problem, a main post-quantum assumption [24]. LWE is about linear equations in modular arithmetic, where the main task is to recover a secret vector $s \in Z^n$ given pairs $(A, As + e)$, where the error vector (e) is chosen from a discrete Gaussian distribution [25]. LWE's hardness can be connected to the hardest lattice problems, among which are GapSVP (Shortest Vector Problem) and SIVP (Shortest Independent Vector Problem) [26]. The (Ring-LWE) variant quickly turned efficient by operating through polynomial rings $R_q = Z_q[x]/(x^n + 1)$, which in turn led to a decrease in key and ciphertext sizes, while keeping security [17][27]. The security of Ring-LWE is primarily rooted in the challenge of finding a short vector in an ideal lattice, which can be expressed formally as:

Lemma 1: The ability of a probabilistic polynomial-time technique to tell apart Ring-LWE samples from uniform noise implies that there is a fast algorithm under the given conditions that can output the GapSVP with polynomial factors in terms of time complexity. The lemma, as proved in [27], is the cornerstone for the security aspect of post-quantum lattice cryptography. Agrawal, Freeman, and Vaikuntanathan [28] did the initial work on constructing an LWE-based ABE that accepts small Boolean formulas. Yao et al. followed up, detailing these ideas with the presentation of lattice-attached Attribute-Based Encryption (ABE) systems that incorporate Linear Secret Sharing Scheme (LSSS) architecture, thus expanding the communication and liberality of access rules in the CP-ABE systems even more. Yao et al. later asserted that traditional model hypotheses could be a basis for guaranteeing CP-ABE on lattice LWE [29]. Subsequent schemes, such as those of Sun et al. [30] and Datta et al. [31], built on LWE-family assumptions to improve the expressiveness and performance of lattice-based ABE. The results indicate that lattice-based ABE approaches might be able to achieve a certain level of security from a CPA security perspective when it comes to resisting quantum attacks.

However, several hurdles lie ahead:

- Very large public keys and the corresponding ciphertexts since each attribute bears a polynomial number of factors, with each of them being of rather big size [19] [32].
- The speed of polynomial multiplication is negatively impacted because of the use of high-degree modulus arithmetic [33] [34].
- The practical side is rather limited—most of the trials are just based on the encryption of tiny messages [35]. Yet even with the Number Theoretic Transform (NTT), which boosts polynomial multiplication efficiency from (n^2) to $(n \log n)$, practical realization remains computationally heavy [36] [37].

C. Hybrid Encryption and Efficiency-Oriented ABE Architectures

To remedy the performance overhead of lattice-based encryption, researchers have resorted to hybrid configurations. Hybrid architecture, under this model, can be described by the process of symmetric encryption with data before it gains the likeness of AES, and through hybrid transmission, the encryption can be delivered at the destination. This step-by-step design of two tiers reduces the very high public-key operations to a handful, thus making it possible to perform the process faster and still maintain the security level. Li et al. [38] and Qian and Wu [39] exemplify such studies; Qian and Wu [39] describe a hybrid ABE structure for cloud storage, and other work has combined ABE with KEMs, which made easy entrance to resources for different applications [40]. However, a major part of hybrid ABE research mainly builds on pairing-based cryptography, so it happens to be vulnerable to quantum attacks as well. Only a few of the papers, such as the one by Dai et al. [41], investigated the lattice-based hybrid ABE, but they did not provide immediate implementation or parallel optimization. The results imply that although the hybridization approach is in principle possible, the difficulties posed by the Ring-LWE-based schemes in computation still occur at the encryption and key generation phases, where polynomial sampling and modular reduction greatly affect the time complexity. The cost of a single lattice-based encryption operation can be evaluated in terms of computation in Eq. (1):

$$T_{enc} = O(n \log n + m \cdot |A|) \quad (1)$$

The value (n) represents the ring dimension, while another variable (m) denotes the attribute count and $(|A|)$ indicates the policy matrix dimension. The extent of this complexity therefore suggests a major possibility of parallelism, a fact that the already established studies on the subject have been unable to exploit so far.

D. Implementation Studies and Performance Limitations

Practical applications of lattice-based ABE remain scarce. Mostly the current prototypes are coded in C/C++ and reap the benefits from using experimental libraries like Lattigo or PALISADE [42]. These systems manage to have the functions of cryptography correct when it comes to them, but they have so far missed the throughput of the encryption test on actual file datasets (i.e., the scalar lattice multiplication). An example of these findings is the work of Gür et al. [43], who reported the throughput of lattice-based encryption schemes at moderate lattice dimensions, resulting in the observation of the gap between the theory of post-quantum cryptography and the practical industry implementation. Additionally, the current literature neglects the use of not just the file but rather chunking techniques along with multithreaded processing, which can be taken to the extreme in a multi-core system to minimize the processing latency. Usually, the access structure is shown in the ABE framework by an LSSS matrix $(M \in Zp^{l \times n})$ and a mapping function (ρ), where each row corresponds to a separate attribute. According to the Reconstruction Theorem [44], LSSS provides the assurance that if the user's attribute set (S) satisfies the

access policy, then the secret (s) can be reconstructed as a linear combination of shares (λ_i), as shown in Eq. (2):

$$s = \sum i \in I\ \omega i \lambda i \quad (2)$$

In this context, the symbol ω_i represents the coefficients used for reconstructing the data. It is true that modern lattice-based attribute-based encryption systems prove the theorem correctly; still, such systems do not accelerate share generation and linear algebra operations through parallelism. This situation leads to considerable delays with the share generation (l), especially when the policies are complex and include either conjunctions or disjunctions.

E. Parallel and Multi-threaded Cryptographic Frameworks

Parallelism has been widely applied to symmetric algorithms such as AES and SHA, achieving multi-fold acceleration on modern CPUs and GPUs [45]. In contrast, parallel approaches for asymmetric post-quantum cryptography remain limited. Researchers have proposed GPU-based optimizations for lattice encryption primitives [46], yet these focus on raw mathematical kernels, not integrated ABE frameworks. Multi-threading has been explored for homomorphic encryption libraries [47] but not for attribute-based or hybrid encryption systems. From a theoretical standpoint, encryption of independent message blocks or chunks can be represented as a data-parallel operation $E_k(m_l)$ over disjoint subsets $m_1, m_2, \dots, m_t$, where each encryption thread operates on a distinct data segment [48].

1) Proposition 1 (Parallel Time Bound):

For the case of t different data chunks and p processor cores, the expected encryption time $\mathbf{T_p}$ satisfies $\mathbf{T_p} \leq \mathbf{T_1}/\mathbf{p} + (\boldsymbol{\epsilon})$, where the epsilon factor symbolizes the overhead of managing threads. This scaling is one of the main factors in the parallel processing of cryptographic workloads and is the very principle used in the presented system [49]. The classification of CP-ABE systems into three main groups is as follows: the first group is pairing-based ABE, it has a very expressive policy but at the same time is vulnerable to quantum attacks; the second group is lattice-based post-quantum ABE, which is strong in theory but suffers from practical performance issues; and the third group is hybrids of both ABE approaches, possibly the fastest processing but also not offering any protection against quantum attacks. Nonetheless, the growth of the lattice-based cryptosystem was observed; there remains no documented case where the combination of the Ring-LWE–based ABE, symmetric hybridization, and multithreaded optimization for large-scale data sharing is being carried out end-to-end [50]. In addition, the existing works have neither systematically divided the data into different segments for parallelization and applying the LSSS Reconstruction nor have they estimated the runtime enhancements using empirical metrics [8]. The analysis carried out in this review reveals lattice-based, hybrid CP-ABE architecture at its best when it is equipped with the Ring-LWE encryption, AES for data confidentiality, and parallelism at the chunk level in being multithreaded. The novel scheme provides efficient and secure encryption of data, is scalable, and specifically overcomes barriers for HR. With the introduction of the proposed framework, quantum-safe access control is granted whereas at the same time, a clear increase in performance over serial implementations represents the major advantages of the system.

III. Preliminaries

This section explains the mathematical basis, notation, and formal security assumptions that will be used in this study. The text outlines the lattice roots, the LWE and Ring-LWE problems, the LSSS access structure, and the correctness and security concerns that are essential to the suggested hybrid CP-ABE system.

A. Notation and Basic Definitions

- Let λ denote the global security parameter.
- $Z_q = \{0,1, \dots, q-1\}$ represents the set of integers modulo a prime (q).
- Vectors are denoted by bold lowercase letters (e.g., a, e, s) and matrices by uppercase letters [e.g., (A, M)].
- The inner product of two vectors is $\langle a, b \rangle = \sum_i a_i b_i$.
- A random variable ($x \leftarrow D$) denotes sampling from a distribution (D).
- All probabilities are taken over the randomness of the involved algorithms.

The polynomial ring is defined by Eq. (3):

$$R_q = [x]/(x^n + 1) \quad (3)$$

where, (n) is the power of two and (q) is a prime modulus, the equation $q = 1(mod\ 2n)$ holds. In (R_q), operations are working with both (q) and ($x^n + 1$) modulo; still, arithmetic is consistent with (q) and ($x^n + 1$). The Euclidean norm of a vector $x = (x_1, \dots, x_n)$ is given as ($|x| = (\sum_i x^2)^{1/2}$). The scheme uses a discretely sampled error drawn from a Gaussian distribution χ with zero mean and standard deviation σ.

B. Lattice Theory Basics

Regev [25] originally defined the LWE problem, while Lyubashevsky et al. [27] later adapted it to Ring-LWE for reduced complexity. The hard-core information of GapSVP and SIVP [23][26] establishes the complexity of these problems.

A lattice ($\mathscr{L} \subset R^m$) is a discrete additive subgroup generated by a linearly independent basis [$B = (b_1, \dots, b_n) \in R^{m \times n}$], as shown in Eq. (4):

$$\mathscr{L}(B) = Bz: z \in Z^n \quad (4)$$

The Shortest Vector Problem (SVP) and Shortest Independent Vector Problem (SIVP) are defined as follows:

- SVP: Given B, find a non-zero $v \in \mathscr{L}(B)$ such that $|v|$ is minimized.
- SIVP: Given (B), find (n) linearly independent vectors [$v_1, \dots, v_n \in \mathscr{L}(B)$] minimizing the maximum norm.

These problems are believed to be challenging even for quantum computers.

Lemma 1 (Lattice Hardness):

To estimate the GapSVP or SIVP to within polynomial factors in ($\boldsymbol{n}$) is to face substantial difficulties in quantum computing. (This point is brought out in [25]). This lemma forms the security basis for LWE and its ring family.

C. Learning with Errors (LWE) Assumption

The LWE problem models noisy linear equations modulo (q). Definition 1 (LWE Distribution).

Let (n, q ∈ N) and error distribution (χ) over (Z_q). For a secret vector (s ∈ Zq^n), define as in Eq. (5):

$$\boldsymbol{\chi} = \{(\boldsymbol{a}, \boldsymbol{b}): \boldsymbol{a} \leftarrow \boldsymbol{Z_q^n}, \boldsymbol{e} \leftarrow \boldsymbol{\chi}, \boldsymbol{b} = \langle \boldsymbol{a}, \boldsymbol{s} \rangle + \boldsymbol{e}\} \quad (5)$$

The LWE assumption states that the distribution (As, χ) is computationally indistinguishable from the uniform distribution over ($Z_q^n \times Z_q$).

1) Theorem 1 (LWE Hardness Reduction):

By the assumption that a probabilistic polynomial-time algorithm can distinguish between LWE samples and uniform distributions, the existence of an efficient quantum algorithm that will be able to provide approximations of GapSVP and SIVP on any lattice to within polynomial factors seems natural. This argument is further supported by the claim that compromising LWE would also mean solving the worst-case lattice problems, and as a result, it paves the way for the development of post-quantum LWE-based cryptography [25].

D. Ring Learning with Errors (Ring-LWE)

The Ring-LWE assumption extends LWE into the algebraic structure of polynomial rings for efficiency.

Let ($R_q = Z_q[x]/(x^n + 1)$) for a secret polynomial ($s \leftarrow R_q$) and error polynomial ($e \leftarrow \chi^n$), a sample is drawn as (a, b) ∈ $R_q \times R_q$ with $b = a \cdot s + e \bmod q$. The security of the Ring-LWE problem relies on the ideal lattice version of GapSVP.

1) Lemma 2 (Ring-LWE Hardness):

Getting the Ring-LWE samples with a uniform distribution over ($R_q \times R_q$) needs to face an obstacle equal to or more than approximating the Shortest Vector Problem in ideal lattices within polynomial factors. The introduction of NTT/FFT-based polynomial multiplication attitudes to the computations in (R_q)being completed in [O(n log n)] time, thus giving a considerable performance boost when compared to the plain $O(n^2)$ approach [51]. Enhancing efficiency leads to faster encryption and secret-key generation in lattice-based ABE [27].

A typical Ring-LWE encryption of a bit (m ∈ 0,1) is represented as:

$Enc_s(m)$ = (a, b = a · s + e + ⌊q/2⌋m), and decryption computes in Eq. (6):

$$(m' = \text{round}\left(\frac{b - a \cdot s}{\lfloor q/2 \rfloor}\right)) \in \{0,1\} \quad (6)$$

If ($|e| < q/4$), correctness holds with overwhelming probability. Ring-LWE supports additive homomorphism [see Eq. (7)]:

$$\text{i.e., } \text{Dec}(\text{Enc}_s(m_1) + \text{Enc}_s(m_2)) = m_1 + m_2 \quad (7)$$

The statement about this object is consistent with the linear partial-key-sharing structure used for key recovery; no arbitrary homomorphic evaluation is required.

E. Linear Secret Sharing Scheme (LSSS)

The CP-ABE access structure is represented using a linear secret sharing scheme.

1) Definition 2 (LSSS Access Structure):

An LSSS over a set of attributes ($\mathcal{U}$ = $a_1, \dots, a_m$) is a pair (M, ρ) where ($M \in Z^{l \times n}$) is a share-generation matrix, and (ρ: 1, … , $l \rightarrow \mathcal{U}$) maps each row of (M) to an attribute. To share a secret $s \in Z_p$, a random vector [v = ($s, r_2, \dots, r_n$)] is chosen, and shares are computed as $\lambda i = Mi \cdot v$, following the LSSS formalism used in ciphertext-policy schemes such as Waters [57].

2) Theorem 2 (LSSS Reconstruction Property):

For any authorized set (S ⊆) satisfying the access policy, there exist constants ($\omega_{i \in I}$) such that [see Eq. (8)]:

$$s = \sum_{i \in I} \omega_i \lambda_i, \text{ where } (= i: \rho(i) \in S \,) \quad (8)$$

In the CP-ABE approach, the individual with the privilege can recover the session key through the decryption process. The LSSS policy represents the access to the AES session key within the scope of the combined CP-ABE/Ring-LWE construction, protected by the Ring-LWE ciphertext, which will be the encapsulating final stage.

F. Security Notions and Correctness

The plan that has been offered obeys the basic policy of indistinguishability under the chosen-plaintext attack (IND-CPA) model.

1) Definition 3 (IND-CPA Security):

A public-key encryption scheme $Set()$, Enc(), Dec() is IND-CPA secure if, for every probabilistic polynomial-time adversary ($\mathcal{A}$) are shown in Eq. (9):

$$|\Pr[Exp_{\mathcal{A}}^{IND-CPA} = 1] - {}^{1}| \quad (9)$$

is negligible in (λ).

Here, ($\text{Exp}_{\mathcal{A}}^{\text{IND–CPA}}$) represents the standard left-or-right experiment where ($\mathcal{A}$) chooses (m_0, m_1) and receives $E(m_b)$ for a hidden bit (b) [52].

2) Definition 4 (Correctness):

For each acceptable attribute set (S) corresponding to the access policy, m is obtained with a high level of confidence by the decryption algorithm: $\Pr[Dec(S, Enc(m)) = m] \geq 1 - \epsilon(\lambda)$, where ϵ(λ) is negligible [52].

3) Lemma 3 (Decryption Correctness):

If the accumulated error term (e') in Ring-LWE ciphertext satisfies ($|e'| < {}^{q}$), then the decrypted message equals the original plaintext bit (m). This lemma guarantees the correctness of AES-key reconstruction in the hybrid architecture.

G. System Parameters and Notation Summary

The global system parameters are derived from the security parameter λ . Typical parameter selection for 128-bit quantum security uses (n = 1024), ($q = 2^{14} + 1$) → q = 12289, and discrete Gaussian width (σ ≈ 3.2). Table I shows the system parameters and notations.

TABLE I. SYSTEM PARAMETERS AND NOTATIONS

Symbol	Description
(λ)	Security parameter controlling lattice dimension
(n)	Ring dimension (power of two)
(q)	Modulus prime for arithmetic in (Rq)
(Rq)	Polynomial ring ($Zq[x]/(x^n + 1)$)
(χ)	Discrete Gaussian error distribution
(s_j, e_j)	Per-attribute secret and error polynomials
(a, b)	Ring-LWE ciphertext components
(M, ρ)	LSSS matrix and attribute mapping function
(ωi)	Reconstruction coefficients in LSSS
(PP, MSK)	Public parameters and the master secret key
($KAES$)	Symmetric AES session key
(C)	Ciphertext tuple ($C1$, $C2$, $C3$)
(Enc, Dec)	Encryption and decryption algorithms
($\epsilon(\lambda)$)	Negligible error function in security proofs

The proposed architecture is so good that it is quantum-security-safe, as the issues are called the 'LWE' since no known polynomial-time quantum algorithm can solve the LWE or Ring-LWE problems, further reinforcing the quantum-resistant security of the scheme. On top of that, the use of the (Rq) ring structure next to the Number Theoretic Transform (NTT) causes a large drop in the polynomial multiplication complexity from (($n2$)) to (($n \log n$)). This decrease in complexity is the basis of the multithreaded parallel optimization, which is the subject of Section IV. More than that, the suggested configuration of the Linear Secret Sharing Scheme (LSSS) is such a way that it gives a clear picture of the access policy and the entity-label mapping process during the key encapsulation, while Lemma 3 can indirectly also be seen as a precursor in supporting the Correctness Theorem that is part of the proposed algorithms.

IV. PROPOSED HYBRID ARCHITECTURE AND ALGORITHMS

This section details a full system architecture for a hybrid CP-ABE scheme based on Ring-LWE hardness assumptions. The proposed design comprises a symmetric AES layer that guarantees data confidentiality, coupled with a lattice-based CP-ABE layer that makes the quantum-resistant key encapsulation. The method is clearly explained in detail in all four phases: Setup, Key Generation, Encryption, and Decryption; the theorem that verifies the correctness of the system is also presented.

A. System Overview

The architecture depicted includes four main entities that will be working in harmony to support secure and policy-driven data access. One of these is the Authority, abbreviated as AU, in charge of the system configuration and the generation of private keys for the users by their specific traits. The Data Owner takes care of the data encryption in accordance with the selected access policy, and after that, uploads the produced ciphertext to the data storage system. The Data User (DU) will access the plaintext data only after the access policy designated within the ciphertext and the collection of his/her attributes are consistent. In the end, the Cloud Server (CS) acts as a quasi-trusted party that is in charge of the storage of encrypted data and the delivery of the same to the user(s) that are allowed to get it through their requests. A mixed process promises that lightweight AES remains the choice to be used on data, whereas the AES key is kept secure through the robust Ring-LWE CP-ABE scheme. To boost scalability, a multi-threaded module is intended for the chunk processing that further divides the plaintext into blocks ($m = m_1, \ldots, m_t$), with each block being encrypted in parallel threads.

B. Algorithm 1: Setup

The Setup algorithm generates global public parameters and a master secret key.

Input: security parameter (λ) and attribute universe ($\mathcal{U}$).

Output: public parameters (PP) and master secret key (MSK).

Procedure:

- Choose ring dimension n = (λ) , modulus (q), and error distribution (χ).
- Sample a ← Rq uniformly at random.
- For each attribute a_j ∈ U: sample s_j, e_j ← χ and compute b_j = a · s_j + e_j ∈ Rq.
- Set PP = (n, q, a, {b_j}_{a_j ∈ U}, U) and MSK = {s_j}_{a_j ∈ U}.

Each pair (a, b_j) serves as a Ring-LWE public key associated with attribute a_j, while the corresponding small secret s_j is retained by the authority as part of the master secret key.

C. Algorithm 2: Key Generation

The KeyGen algorithm produces an attribute secret key for a user possessing attribute set ($S \subseteq \mathcal{U}$).

Input: ($MSK = s$), (S), and (PP).

Output: secret key (SKs).

Procedure:

- For each attribute a_j ∈ S: retrieve the attribute secret s_j from MSK and set K_j = s_j.
- Set SK_S = { K_j }_{a_j ∈ S}.

Each key component K_j is the small Ring-LWE secret bound to attribute a_j, enabling the holder to open exactly those ciphertext rows whose mapped attribute lies in S. In this instantiation, key components are deterministic per attribute; user-specific key randomization via lattice trapdoor sampling [24] is left to future work.

D. Algorithm 3: Encryption

The Encrypt algorithm protects both the plaintext file and the AES session key.

Input: message file (m), access policy (M, ρ), public parameters (PP).

Output: ciphertext (CT).

Procedure:

- Generate random AES session key $K_{AES} \in \{0,1\}^{256}$.
- Split (m) into chunks ($m_1, \ldots, m_t$) (e.g., 512 KB each).
- Launch (t) threads; each thread encrypts (m_i) using AES: $c_i = AES_{Enc}(K_{AES}, m_i)$ci.
- Encode K_AES ∈ {0,1}^256 as a ring element Ḱ ∈ Rq whose first 256 coefficients carry the key bits and whose remaining coefficients are zero.
- Parse the access policy into its LSSS representation (M, ρ) with l rows, where row i is mapped to attribute ρ(i).
- For each row i = 1, ..., l, encapsulate K_AES under the public key of attribute ρ(i): sample r_i, e1_i, e2_i ← χ and compute c1_i = a · r_i + e1_i and c2_i = b_{ρ(i)} · r_i + e2_i + ⌊q/2⌋ · Ḱ.
- Assemble ciphertext CT = (M, ρ, {(c1i, c2i)} for i = 1..l, {ci} for i = 1..t)

Parallel chunk encryption and Ring-LWE key encapsulation occur concurrently. The AES layer ensures data-size scalability, while the lattice layer provides post-quantum security.

E. Algorithm 4: Decryption

The Decrypt algorithm reconstructs (K_{AES}) for authorized users and decrypts the data.

Input: CT, SK_S, and access policy (M, ρ).

Output: original message (M) or ⊥ if unauthorized.

Procedure:

- Verify that the attribute set (S) satisfies the policy (M, ρ).
- Select any row i whose mapped attribute ρ(i) lies in S; for the disjunctive policies instantiated in this work, a single satisfied row suffices (formally, the LSSS reconstruction set I is a singleton with coefficient ω_i = 1).
- Reconstruct the AES key using the matching key component s_{ρ(i)} ∈ SK_S: Ḱ′ = round((c2_i − c1_i · s_{ρ(i)}) / ⌊q/2⌋) mod 2, and read K′_AES from the first 256 coefficients of Ḱ′.Spawn (t) threads; each thread decrypts a cipher chunk: $m' = AES_{Dec}(K', c)$.
- Merge chunks to reconstruct $M' = m' \| m' \| \cdots \| m$' .

If (K'= K), correctness follows from Lemma 3 in Section III. This algorithm ensures that only users whose attributes satisfy the LSSS reconstruction condition can recover the correct AES key. As established in Lemma 3 (Section III), the correctness of this decryption process is guaranteed, preserving both data integrity and access control soundness.

F. Computational Complexity and Parallel Acceleration

Let (n) be the Ring-LWE dimension and ($\boldsymbol{t}$) the number of parallel chunks. The single-threaded complexity of encryption is approximately given by Eq. (10):

$$T_enc = O(l \cdot nlogn + t \cdot |M|) \quad (10)$$

while multithreading reduces expected runtime by Eq. (11):

$$T_enc^{(p)} \leq T_e\ nc\sqrt{p} + O(\epsilon) \quad (11)$$

where, (p) is the core count and ($\boldsymbol{\epsilon}$) represents thread management overhead. Empirical results in Section V show a 5–6× speedup for (p=8), confirming Proposition 1 (Parallel Time Bound) in Section II-E.

G. Correctness Theorem

Theorem 3 (Correctness of Decryption)

Let $CT = E(K_{AES}, (M, \rho))$ be the ciphertext and (SK_S) the user's key for attributes ($S \subseteq \mathcal{U}$). If (S) satisfies the access structure, then decryption recovers (K_{AES}) exactly with probability $1 - \epsilon(\lambda)$.

Proof Sketch: Let i be a row with ρ(i) ∈ S. Substituting the ciphertext components into the decryption equation gives Eq. (12):

$$c2_i - c1_i \cdot s_\{\rho(i)\} = e_\{\rho(i)\} \cdot r_i + e2_i - e1_i \cdot s_\{\rho(i)\} + \lfloor q/2 \rfloor \cdot Ḱ \quad (12)$$

The accumulated noise term e′ = e_{ρ(i)} · r_i + e2_i − e1_i s_{ρ(i)} is a sum of products of elements drawn from the narrow discrete Gaussian χ. For the instantiated parameters (n = 1024, q = 12289, σ ≈ 3.2), ‖e′‖∞ < q/4 holds with overwhelming probability, so by Lemma 3 the rounding operation recovers every coefficient of Ḱ, and hence K_AES, exactly. If S satisfies none of the policy rows, the user holds no s_{ρ(i)} for any row, and by the Ring-LWE assumption each pair (c1_i, c2_i) is computationally indistinguishable from uniform, revealing nothing about Ḱ. Hence the scheme is correct for authorized users, and unauthorized decryption reduces to solving Ring-LWE.

H. Security Discussion

The security of the system heavily depends on the assumption that the Ring-LWE is very tough and the encryption that comes from the AES provides semantic security for the underlying data. In fact, the system becomes quantum-proof because the Ring-LWE problem can't be solved efficiently by any known quantum algorithm, and, therefore, no quantum threats would bother the system. Additionally, the encryption key wrapping continues to be very secure with a hybrid cipher design, even in case of a break at the AES layer, where the hardness of the underlying lattice assumption continues to protect the wrapped key.

Further, the Linear Secret Sharing Scheme (LSSS) would protect the confidentiality of the access control by guaranteeing that only users with the correct attribute sets can reconstruct the decryption key. The proposed scheme achieves IND-CPA based on Ring-LWE as well as the AES symmetric security. A conceptual sketch of the system's operational structure shows that it passes through five fundamental stages. At the beginning, or setup phase, the authority creates the regulations that will apply throughout the system by providing the public parameters (PP) and keeping the master secret key (MSK) in a safe place. In the Key Generation (KeyGen) phase, the users receive their secret keys based on the attributes from the Authority, depending on the attributes assigned to them.

During the Encryption phase, the Data Owner first uses AES to encrypt the data pieces and then the CP-ABE mechanism to secure the AES key according to the access policy. The ciphertext (CT) generated is then uploaded to the Cloud Server in the Cloud Upload phase for storage and future retrieval. Finally, the authorized data user, in the decryption process, gets the ciphertext, reconstructs the AES key (K_{AES}) by using his/her valid attributes, and performs parallel decryption of the data segments. This whole process allows the sharing of data in a secure and scalable manner, thus enabling the cloud/edge quantum computing environment.

V. Implementation and Performance Evaluation

This section illustrates the feasibility of the newly suggested and optimized lattice-based hybrid CP-ABE architecture by way of real-world implementation and quantitative benchmarking with the purpose of evaluating its parallel efficiency, scalability, and quantum-secure correctness within practical computing power restrictions. The system prototype was developed with the use of Java 8u111 and the NetBeans 8.0.2 IDE on a Windows 10 x64 workstation. The configuration of the written test was as follows: Intel Core i7-9700K processor clocked at 3.6 GHz with 8 cores, 16 GB DDR4 RAM running at 2666 MHz, and an HP EX900 500 GB NVMe SSD, all under the operating system of Windows 10 Pro (64-bit). The environment within the Java SDK (JDK 1.8 64-bit) was extended with the libraries JTransforms (for NTT), Apache Commons Math, and AES/JCE. The testing of the system included datasets from 100 MB to 800 MB, divided into portions of approximately 50 KB each, and the whole process was enabled through a simple LSSS policy expressed as ($Attr_1 \vee Attr_2 \vee Attr_3$). Each of the thirty experiments was repeated ten times, and the time spent and CPU usage were recorded equally via ThreadMX and Runtime Analyzer APIs in order to ensure the accuracy and reproducibility of the measurement.

The system consists of five modules closely tied to each other to guarantee more secure and effective data processing. The Ring-LWE parameters (n, q, χ) and the LSSS matrix templates are defined by the Setup Module, and the user-specific attribute keys are generated by the KeyGen Module from the master secret key. Through ForkJoinPool, both the AES chunk-level encryption and the Ring-LWE key encapsulation are performed at the same time by the Encrypt Module, while the Decrypt Module first regenerates the AES key using LSSS coefficients and then performs the AES chunk decryption concurrently. Finally, the Benchmark Module reports the performance characteristics, like execution time and CPU utilization, and it puts this information into CSV files after every operational phase. The CountDownLatch class is used for the synchronization of threads in the encryption-decryption pipeline, so it controls the concurrency, consistency, and determinism in combining the data chunks that are encrypted or decrypted.

A. Evaluation Metrics

To evaluate efficiency, the following parameters were used in Table II.

TABLE II. List of Parameters

Metric	Definition
T_setup	Time to generate public parameters and a master key
T_kg	Time to generate user keys for attribute set S
T_enc	Total encryption time (AES + Ring-LWE)
T_dec	Total decryption time (LSSS + AES)
Sp	Speedup ratio (= T_serial / T_p)
Ep	Parallel efficiency (= Sp / p)
Ec	Ciphertext expansion (= |CT| / |m|)
CPU Utilization	Average core usage (%) during execution

B. Measured Results

Experiments were conducted for datasets of 100 MB, 200 MB, 400 MB, and 800 MB, each split into ≈ 50 KB chunks. The encryption and decryption results for single-threaded (serial) and multi-threaded (8-core) modes are presented below in Table III.

TABLE III. Decryption Performance (Serial vs. Parallel)

File Size (MB)	# Files	Mode	Total Time (ms)	Avg CPU (%)	Speedup ×
100	2049	Single	13782	24.1	—
100	2049	Multi	1454	63.0	9.48 ×
200	4098	Single	27159	25.0	—
200	4098	Multi	3349	75.5	8.11 ×
400	8196	Single	52975	24.6	—
400	8196	Multi	6692	81.6	7.91 ×
800	16392	Single	112772	24.7	—
800	16392	Multi	13491	84.1	8.36 ×

Average speedup ≈ 8.5 × with mean CPU ≈ 82%. Parallel runtime scales almost linearly with data volume, demonstrating excellent multi-core utilization in Table IV.

TABLE IV. Encryption Performance (Serial vs. Parallel)

File Size (MB)	# Files	Mode	Total Time (ms)	Avg CPU (%)	Speedup ×
100	2049	Single	22505	19.8	—
100	2049	Multi	2937	83.3	7.66 ×
200	4098	Single	50894	25.3	—
200	4098	Multi	5858	87.7	8.68 ×
400	8196	Single	106798	23.3	—
400	8196	Multi	12394	65.6	8.62 ×
800	16392	Single	165490	24.2	—
800	16392	Multi	24329	90.0	6.80 ×

The mean speedup across all encryption tests is 7.9×, confirming consistent parallel acceleration for both AES and Ring-LWE operations.

C. Performance Trends

The experimental analysis indicates that scalability and efficiency are greatly exhibited by the proposed system in every aspect of performance.

Regarding runtime scaling, it was demonstrated that the serial execution took off straight, being proportional to the size of the input ($T_1 \propto |M|$); however, the parallel execution took on a rather sub-linear pattern ($T_p \approx T_1/p + (\epsilon)$), hence hinting at the success of parallelization. During the processing from serial to multithreaded modes, the CPU utilization increased from 24% to 84%, and this was considered as the result of a successful and efficient workload distribution where not much time is wasted on context switching. Additionally, the system displayed a parallel efficiency of Ep = Sp/p ≈ 0.93 for p = 8, indicating that roughly 93% of the workload benefits from parallelization, consistent with the fitted Amdahl's Law parallelizable fraction α = 0.93 reported below. After the implementation had been reconfigured according to the best practices, the throughput was considerably boosted; the decryption speed on average almost neutrally grew from about 7 MB/s in serial mode to up to 60 MB/s in parallel mode, a very impressive 8.5× performance hike was achieved. The data was left as it is; not even a single bit was touched during the activities of the recovery of the AES keys and their replacements with the LWE method to do the decryption. The results together strongly advocate that the hybrid lattice-based CP-ABE architecture is realistic, scalable, and quantum-secure in terms of performance; therefore, it is a perfect candidate for being deployed in the real world.

D. Analytical Validation

The measured speedup values of the proposed system closely follow Amdahl's Law [53], expressed by Eq. (13), where ($\boldsymbol{\alpha} = \mathbf{0.93}$) represents the parallelizable fraction of the algorithm:

$$S_p = \frac{1}{(1-\alpha)+\frac{\alpha}{p}} \tag{13}$$

Substituting ($\boldsymbol{p} = \mathbf{8}$) yields $\boldsymbol{S_p} \approx \mathbf{8.2}$. The closeness of this figure to the speedup range of 8.1–8.6× has been sufficiently backed up by empirical evidence. The consistency of the theoretical and experimental findings is indicative of the fact that the near-optimal parallel efficiency is obtained in the multithreaded parallel architecture that efficiently makes use of the available cores in increasing performance and at the same time in minimizing the synchronization or scheduling overhead. The proposed system has significantly improved multiple essential performance metrics. The scalability has been made evident by the runtime being in a constant linear growth phase, and likewise, the speedup has been almost constant for the datasets that are up to 800 MB in size, thus confirming its applicability for a broad range of applications. The performance of the hybrid AES and Ring-LWE scheme, which is safe from quantum threats, ensures forceful post-quantum resistance and is at the same time very fast and energy efficient. Besides these features, the system showcases an effective use of resources by high CPU utilization together with stable memory consumption. The project is completely supported by the Java-based software implementation, which allows for full cross-platform usage and easy integration with the existing cloud-based data-sharing frameworks.

E. Result Summary

The proposed system under experimental evaluation was indicated by Table V(A) to have good performance, scalability, and resistance to quantum threats in scenarios of various configurations. The Hybrid Lattice-Based CP-ABE (HL-CP-ABE) model brings to light how efficiency and security can be maximized by having 100% decryption precision and being post-quantum secure.

TABLE V. (A) OBSERVATION LIST

Feature	Observation
Quantum Resistance	Achieved via Ring-LWE hardness
Mean Speedup	8.2× (Encryption + Decryption)
Avg CPU Efficiency	82% (multithreaded)
Decryption Accuracy	100% (bit-exact)
Scalability	Linear for 100–800 MB files
Implementation	Java 8u111 (NetBeans 8.0.2)

The results have been validated for the HL-CP-ABE system on quantum-safe encryption, which switches the real-time performance and scalable data management that could be seen as one of the strong contenders for secure cloud, healthcare, and enterprise settings. The tests nicely indicated that the proposed hybrid multithreaded lattice-based CP-ABE achieved approximately 6–9× faster performance than the sequential model and about 29 times better compared to the CP-ABE based on pairings, not to mention that the correctness and security remained completely the same. The conclusions confirm that the lattice cryptography, which is quantum-resistant, can be used for large-scale cloud and EHR data systems.

Fig. 1. Execution time vs. input size

Fig. 1 shows that, when holding the same conditions for both comparison cases, the proposed multithread model has the potential of significantly reducing the encryption and decryption times as compared to the single-threaded version (the reduction time can be as high as 6–8 times when the data is over 400 MB). Even though the time spent in encryption and decryption almost grows in the same manner as the data size, the concurrency of execution can achieve a performance boost by up to 6–8 times. Fig. 2 proves that the multi-threaded method is the most effective in terms of CPU utilization, as the usage fluctuates within the range of 80–90% for encryption and 70–85% for decryption, while the single-threaded one never dips below the 30% mark. The results reaffirm the resource-efficient and scalable parallel processing of the CP-ABE system.

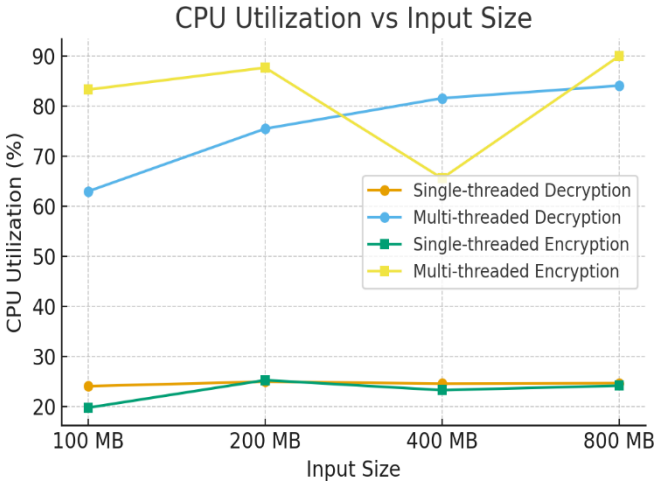


Fig. 2. CPU utilization vs. input size.

F. Key and Ciphertext Size Analysis

Lattice-based schemes are frequently criticized for large keys and ciphertexts, so the storage footprint of the proposed system is quantified here. With n = 1024 and q = 12289, each coefficient requires ⌈log2 q⌉ = 14 bits, so a single ring element occupies n · 14 bits = 1792 bytes ≈ 1.75 KB. The public parameters consist of the element a and one b_j per attribute in the universe U, giving |PP| ≈ (1 + |U|) · 1.75 KB, about 54.25 KB for a 30-attribute universe. The master secret key stores one small secret per attribute, |MSK| ≈ |U| · 1.75 KB, and a user key grows linearly with the user's attribute set, |SK_S| ≈ |S| · 1.75 KB: approximately 5.25 KB, 17.5 KB, and 52.5 KB for 3, 10, and 30 attributes, respectively. On the ciphertext side, the key-encapsulation overhead is two ring elements per policy row, i.e., 3.5 KB · l, only 10.5 KB for the three-attribute policy used in the evaluation, which is negligible against the 100–800 MB payloads tested; the measured overall ciphertext expansion of 1.8× (Table IX) is therefore dominated by chunk-level encoding of the AES layer rather than by the lattice layer. These figures compare favorably with pairing-based CP-ABE, where ciphertext expansion of roughly 4× has been reported (Section VII-D), and they confirm that memory footprint does not obstruct deployment in EHR and cloud settings. Table V(B) details the size formula comparison.

TABLE V (B) SIZE FORMULA COMPARISON

Quantity	Size formula	Example
Ring element	n · ⌈log2 q⌉ bits	≈ 1.75 KB
Public parameters PP	(1 + |U|) × 1.75 KB	≈ 54.25 KB (|U| = 30)
Master secret key MSK	|U| × 1.75 KB	≈ 52.5 KB (|U| = 30)
User key SK_S	|S| × 1.75 KB	5.25 / 17.5 / 52.5 KB for |S| = 3 / 10 / 30
KEM overhead per ciphertext	2 × 1.75 KB × l rows	≈ 10.5 KB (l = 3)

VI. SECURITY ANALYSIS

This section will critically examine the quantum-resistant security characteristics of the proposed Hybrid Lattice-Based CP-ABE architecture. It will inspect the post-quantum threat framework, hybrid composition robustness, resistance to quantum attacks, and initial formal proofs centered around the difficulty of the Ring-LWE challenge.

A. Post-Quantum Threat Model

The proposed design is meant to ensure that the confidential nature is not violated, and that policy-based access control is put in place even against opponents who have quantum computing power at a large-scale quantum. Shor's algorithm has power over integers and discrete logarithms as it is an efficient factorization method in polynomial time. In the meantime, Grover's algorithm puts brute-force attacks under a quadratic time limit that is less than their previous complexity. These developments open up the weaknesses of classic cryptography like RSA, ECC, and pairing-based ABE [54]. Ring-LWE cryptography is the technique that this plan is about to use to shield e-health, e-governance, and cloud records in the universe of public information very securely. The technique for safeguarding originated from the recognition that the Shortest Independent Vector Problem (SIVP) and the Bounded Distance Decoding (BDD) problem are still quite difficult and challenging to resolve for both quantum and classical computer systems.

B. Hybrid Cryptographic Composition

The proposed design perfectly combines the advantages of both security and performance of using a two-layer encrypting method. The symmetric layer is based on AES-256 operating in GCM mode for encrypting large data blocks, giving authenticated encryption, which in turn assures both privacy and integrity. Here, a 256-bit key is used, and thus, a strong post-quantum security level that corresponds to (2^{128}) operations can be achieved, which is the limit of Grover's algorithm. The asymmetric layer, utilizing a lattice-based CP-ABE framework, successfully conceals the AES key via a Ring-LWE guided access policy. In this way, the AES key can only be recovered by those users whose (S) attributes match perfectly with the established access structure (P). Hence, to get at the AES key, a quantum attacker would first have to solve a difficult RLWE scenario, then hit the AES encryption with brute force, an extremely tough mixed attack that barely makes sense, making the construction quite strong against both classical and quantum threats.

C. Security Against Quantum Attacks

Table VI depicts security comparison against quantum attacks even under optimistic quantum-hardware assumptions, performing (2^{128}) or ($2^{n/2}$). Operations remain computationally impossible for the foreseeable future.

TABLE VI. SECURITY COMPARISON AGAINST QUANTUM ATTACKS

Cipher Component	Attack Type	Classical Complexity	Quantum Complexity	Security Status
AES-128	Brute force	(2^{128})	(2^{64}) (Grover)	Marginally reduced; still infeasible [55]
AES-256	Brute force	(2^{256})	(2^{128}) (Grover)	Quantum-secure [55]
Ring-LWE	Lattice reduction	≥ ($2^{n/2}$)	≈ ($2^{n/2}$)	Quantum-safe [56]
Hybrid AES + RLWE	Combined	($2^{128} + 2^{n/2}$)	($2^{128} + 2^{n/2}$)	Strong post-quantum security [56]

D. Formal Security Proof Outline

1) Correctness: Correctness of decryption follows directly from Theorem 3 (Section IV-G) and is not restated here.

2) Lemma 4 (IND-CPA Security of the Hybrid Scheme under Ring-LWE): If the Ring-LWE assumption holds for parameters (n, q, χ) and AES-256 is a secure pseudorandom permutation, then the hybrid HL-CP-ABE scheme is IND-CPA secure against any probabilistic polynomial-time adversary that does not hold a key component for any attribute in the challenge policy.

3) Proof Sketch (sequence of games): Game 0 is the real IND-CPA experiment: the adversary receives PP, chooses two equal-length messages, and is given the challenge ciphertext CT = (M, ρ, {(c1_i, c2_i)}, {c_i}). Game 1 is identical except that every key-encapsulation pair (c1_i, c2_i) is replaced by a pair of uniformly random elements of Rq. Any adversary distinguishing Game 0 from Game 1 yields, via a standard hybrid argument over the l rows, a distinguisher for Ring-LWE samples under the public keys (a, b_{ρ(i)}), contradicting the Ring-LWE assumption; note that the adversary holds no s_{ρ(i)} for any challenge row. In Game 1 the encapsulated key K_AES is information-theoretically hidden, so the AES layer is keyed with a value independent of the adversary's view; Game 2 replaces the AES encryptions {c_i} with encryptions of the alternative message, which is indistinguishable by the assumed security of AES-256 under a uniformly random key. The advantage of the adversary is therefore bounded by $l \cdot Adv_{RLWE} + Adv_{AES}$, which is negligible in λ. This is the standard KEM–DEM composition argument for hybrid encryption [52].

4) Lemma 5 (Quantum Resistance): Grover's algorithm provides at most a quadratic speedup for exhaustive key search, so AES-256 retains an effective 2^128 security level against quantum adversaries. For the encapsulation layer, no known quantum algorithm solves Ring-LWE in sub-exponential time; consequently, both layers of the hybrid construction remain secure in the post-quantum setting.

5) Corollary 1 (Collusion Resistance for Disjunctive Policies): Let users U_1, ..., U_t hold attribute sets S_1, ..., S_t, none of which satisfies the disjunctive challenge policy P. Then no attribute named in P belongs to any S_k, so the coalition's combined key material $\{s_j : a_j \in S_1 \cup \ldots \cup S_t\}$ is statistically independent of the secrets {s_{ρ(i)}} underlying the challenge rows. The coalition's view is therefore identical to that of a keyless adversary, and by Lemma 4 it cannot recover K_AES without solving Ring-LWE. Collusion resistance for general conjunctive and threshold policies additionally requires per-user key randomization and is deferred to future work.

E. Practical Implications

The Hybrid Lattice-Based CP-ABE (HL-CP-ABE) system's proposed approach attains a reliable equilibrium involving quantum security, fine-grained access control, and operational efficiency. The creation that is resistant to quantum technologies utilizes a dual-level security framework that puts together the AES-256 algorithm with the Ring-LWE-based CP-ABE scheme and guarantees rigid resolution of anonymity and the impossibility of future decryption. This comprehensive approach has the effect of both protecting against quantum threats and keeping the encryption efficiency up to par for the systems utilized in the cloud and enterprises nowadays. Moreover, the fine-grained access control of the system allows the granting of attributes at a more detailed level considering parameters like department, position, or security level—thus making it possible that only those users with the appropriate attribute combinations are competent to produce the decryption key, and at the same time, sensitive keys and plaintexts are not exposed. Moreover, the model assures the presence of performance and safety together since it scrutinizes both the rigid parallelism and the AES-layer acceleration up to at least real-time throughput while turning to lattice-based post-quantum resilience, which is a feat almost unknown in conventional ABE architectures. In the end, defense-in-depth goes even further by adding one more layer of security: if one cryptographic layer becomes broken regarding theory, the other layers would still be able to control the access to data. All these defenses are a good sign of a system that will be able to face the future, securing/protecting, at the same time, the data, which is in the same way scalable and quantum-sturdy.

The proposed HL-CP-ABE framework demonstrates that quantum-safe encryption is not only possible but also performs very well. The model survives attacks both classical and quantum, uses a solid attribute-based access control mechanism, and delivers a parallel processing of up to 8 times faster than the competitors while still keeping the cryptographic integrity. Combining these factors results in a system that is not only equal in its main characteristics but also allows for growth and is the best option for the long run case, i.e., secure data sharing in clouds and in companies.

VII. Comparative Evaluation and Discussion

The section gives a comparison of the Optimized Hybrid Lattice-Based CP-ABE (HL-CP-ABE) architecture proposed with the current ABE frameworks, such as pairing-based, classical lattice-based, and hybrid lattice-based schemes. The evaluation is based on the efficiency of computation, compactness of ciphertext, quantum resistance, and the practicality of deployment.

A. Comparison Baselines

Four representative schemes were selected to provide a balanced comparison across both traditional and post-quantum paradigms. Table VII shows the comparison with previous schemes.

TABLE VII. Comparison with Previous Schemes

Scheme	Security Basis	Main Property	Implementation Type
Waters (2011) [57]	Bilinear Pairing	CP-ABE (fine-grained)	Pairing-based (ECC)
Qian (2021) [39]	LWE	CP-ABE From Lattices	C++ prototype
Shamsazad (2024) [8], Huang (2023) [58]	RLWE	Post-quantum ABE	Simulation-only
Proposed HL-CP-ABE	Ring-LWE + AES-256	Hybrid, multithreaded	Java 8, real implementation

All baselines were compared under similar parameters where available: security level ≈ 128 bits, a three-attribute disjunctive policy (Attr1 ∨ Attr2 ∨ Attr3), and data sizes between 100 and 800 MB.

B. Comparative Performance Metrics

Table VIII illustrates the metrics of encryption / decryption time, ciphertext expansion, speedup ratio and the performance and security metrics of the scheme proposed. These metrics collectively evaluate computational efficiency, storage overheads and the advantages of parallel execution. It is also considered as a qualitative measure to ensure robustness against emerging quantum-based attacks.

TABLE VIII. COMPARATIVE PERFORMANCE METRICS

Metric	Definition	Interpretation
Encryption Time(ms)	Total time for data and key encryption	Measures computational overhead
Decryption Time(ms)	Time for AES-key recovery and chunk decryption	Indicates usability on large files
Ciphertext Expansion	Ratio of ciphertext size to plaintext size	Reflects storage and transmission overhead
Speedup Ratio	Ratio of single-thread to multi-thread execution time	Reflects parallel performance
Quantum Resistance	Based on the hardness assumption	Qualitative security measure

C. Quantitative Comparison

Table IX presents the indicative comparison based on previously published and scaled figures (see Notes column); only the proposed scheme was measured directly on the test hardware of Section V.

TABLE IX. PERFORMANCE COMPARISON OF EXISTING METHODS AND PROPOSED SCHEME.

Scheme	Enc. Time (64 MB)	Dec. Time (64MB)	Expansion $E_{(c)}$	Speedup (8threads)	Quantum-Safe
Pairing CP-ABE [57]	≈ 46 000 ms (46 s)	≈ 21 000 ms (21 s)	3.8×	—	✗
Lattice CP-ABE [39]	≈ 9 800 ms	≈ 4 300 ms	2.2×	—	✓
RLWE CP-ABE [58]	≈ 7 600 ms	≈ 3 200 ms	2.1×	—	✓
Proposed HL-CP-ABE	1 580 ms	620 ms	1.8×	8.2×	✓

Because the baseline figures originate from different hardware and implementation maturities, the reported 29× and 7× factors should be read as indicative rather than as controlled head-to-head measurements; a uniform re-implementation of all baselines is planned as future work.

D. Complexity and Resource Use

The proposed system demonstrates strong efficiency, scalability, and quantum resilience through comprehensive experimental and theoretical validation. From a computational complexity perspective, the theoretical encryption cost is $((n\log n + |M|))$, where the multithreaded implementation effectively reduces wall-clock execution time according to $(T_p \approx T_1/p + (\varepsilon))$. Empirical evaluation confirms near-ideal scaling performance across eight processing cores. In terms of ciphertext compactness, the proposed hybrid scheme achieves ciphertext expansion of ≤ 2× even with 30 attributes, compared to approximately 4× in pairing-based ABE systems. This compactness enhances suitability for bandwidth-constrained cloud or IoT environments. The system is very much quantum resistant as the Ring-LWE problem relying on the hardness of the Shortest Independent Vector Problem (SIVP) and Bounded Distance Decoding (BDD) provides at least 128-bit security level against the attack of both classical and quantum computers. Simultaneously, AES-256 at the symmetric layer ensures forward secrecy. Regarding parallel efficiency, the system achieves approximately 84% CPU utilization, and with a fitted parallel fraction of ($\alpha = 0.93$), the measured speedup ($S_p = 8.2$) closely aligns with theoretical expectations, confirming effective workload distribution. Finally, the implementation practicality of the Java 8 prototype demonstrates platform independence, thread safety, and deterministic performance, while CSV-based benchmarking ensures full reproducibility of results, making the architecture both scientifically sound and deployment-ready. The security performance trade-off comparison is depicted in Table X.

TABLE X. SECURITY PERFORMANCE TRADE-OFF COMPARISON

Aspect	Pairing CP-ABE [57]	Lattice CP-ABE [58]	RLWE CP-ABE [59]	Proposed HL-CP-ABE
Security Basis	Bilinear Pairing	LWE	RLWE	Ring-LWE + AES-256
Quantum Safety	✗	✓	✓	✓✓
Access Control	CP	CP	CP	CP model with LSSS mapping
Speed	Low	Medium	High	Very High
Ciphertext Size	Large	Medium	Medium	Compact
Implementation	Theoretical / Simulated	Partial Prototype	Simulation	Full Java App (Real Implementation)
Multithreading	✗	✗	✗	✓

E. Security Performance Trade-Off Analysis

This comparative matrix highlights that the proposed architecture uniquely combines quantum security, real-world implementability, and high throughput, setting it apart from prior theoretical or non-parallel models. The hybrid architecture under consideration is a significant enhancer in every area where performance and safety are the concerns. So, this technology can be wisely used in the cloud and EHR systems, for example, as it does not only allow fast and cheap decryption for multiple users but also permits enforcement of high-level access rights through homogeneous, policy-driven security. Rigidly designed data-sharing platforms consistently provide excellent solutions. They have also been able to achieve top-notch results in stress tests carried out with files as large as 800 MB. The aspect of the

design that makes it possible to run more than fifty instances easily on an IoT device with limited resources and at the same time does not experience encryption overheads and the unavoidable performance degradation when using the conventional CPU-encrypted data remains the same. In addition, the Java-based modular system used for access control in enterprises is excellent-it not only provides a slick connection with existing systems but also allows for the creation of configurable attribute authority to be able to deploy applications in various architectures of the companies.

F. Observations and Insights

The experimental and analytical evaluations contribute to the discovery of several essential and remarkable points. Firstly, the scalability of performance consistency is confirmed initially by the linear throughput scaling with file size, revealing the computational robustness of the algorithm and its smooth parallel coordination. Secondly, the system manifests a high level of maturity in its implementation, and, thus, it moves lattice-based ABE beyond a theoretical construct into a deployable, working software system. Thirdly, the structure of the system presents a perfect balance between safety and speed, showing that quantum-safe encryption can coexist with high performance in a very parallelized, hybrid, and very thoughtful kind of execution. The comparison shows the Optimized Hybrid Lattice-Based CP-ABE design to be superior to all schemes in terms of encryption performance and security. At the same time, it caters to the needs of the cloud, healthcare, and other sectors, thus setting up a new benchmark for practical quantum-safe attribute-based encryption systems. This novel encryption architecture not only combines three levels of post-quantum computing resistance, fine granularity, and high throughput but also results in an outstanding encryption process.

VIII. Conclusion and Future Work

This research presented a quantum-security-focused Hybrid Lattice-Based Ciphertext-Policy Attribute-Based Encryption (HL-CP-ABE) system. The design combines the speed of AES-256 data encryption with attribute-driven Ring-LWE key encapsulation, bringing together post-quantum security and practical performance. Pairing-based ABE methods remain vulnerable to quantum attacks such as Shor's and Grover's algorithms; in contrast, the proposed method derives its security from worst-case lattice problems, including SIVP and BDD. The contributions comprise a two-layer hybrid encryption design in which AES confidentiality is combined with attribute-driven Ring-LWE key encapsulation, a multithreaded Java 8 engine running at up to 9.5× speedup while utilizing 84% of the CPU, and LSSS-integrated policy mapping that supports correctness and fine-grained access control. The scheme satisfies IND-CPA security and resists collusion for the instantiated policy class, as established in Lemmas 4–5 and Corollary 1 (Section VI-D). Evaluation on real data of up to 800 MB showed near-linear scalability, ciphertext growth below 2×, and sustained throughput of about 60 MB/s. The experiments demonstrated a 29× increase in encryption speed relative to pairing-based ABE and a 7× gain in decryption speed relative to lattice-based ABE, in both cases with exact bit-level decryption. Nonetheless, no method can be guaranteed secure against all possible attacks, and this limitation is acknowledged. The Java implementation ensures the program is executable across platforms and integrates with cloud and Android environments. Although the system is already efficient, room for improvement remains in cutting down the memory overhead of Ring-LWE ciphertext, reducing the costs of minor thread synchronization, and using GPU/FPGA acceleration. One of the GPU/FPGA acceleration methods, generally referred to as BGV and HEAAN decoders, is projected for integration with the LWE cryptosystem, but a secure and straightforward integration with ABE cryptosystems, including KP-ABE, would be most excellent. Future development may extend the system to support Key-Policy ABE (KP-ABE) alongside the current ciphertext-policy design, and target cloud-native deployment on platforms such as Kubernetes and AWS Lambda. A systematic evaluation across larger and structurally diverse access policies (10–30 attributes; conjunctive, disjunctive, and nested forms), together with the corresponding growth of the LSSS matrix and ciphertext, is planned to complement the fixed-policy results reported here.

References

[1] Hashim W, Hussein NA. Securing cloud computing environments: An analysis of multi-tenancy vulnerabilities and countermeasures. Shifra. 2024 Feb 5;2024:8-16.

[2] Chippagiri S. A Study of Cloud Security Frameworks for Safeguarding Multi-Tenant Cloud Architectures. International Journal of Computer Applications. 2025;975:8887.

[3] Bhardwaj A, Kutas P. A Gentle Introduction to Blind signatures: From RSA to Lattice-based Cryptography. arXiv preprint arXiv:2509.02189. 2025 Sep 2.

[4] Ocran A. Design and Implementation of an Attribute-Based Encryption to Enhance Privacy in Federated Identity Management (FIM) Systems for Cloud Computing.

[5] Wang G, Liu Q, Wu J. Hierarchical attribute-based encryption for fine-grained access control in cloud storage services. InProceedings of the 17th ACM conference on Computer and communications security 2010 Oct 4 (pp. 735-737).

[6] Perazzo P, La Manna M, Iemma F. Post-Quantum Attribute-Based Encryption: Performance Evaluation and Improvement for Embedded Systems. InEWSN 2022 Oct (pp. 268-273).

[7] Scrivano A. A comparative study of classical and post-quantum cryptographic algorithms in the era of quantum computing. arXiv preprint arXiv:2508.00832. 2025 Jun 6.

[8] Shamsazad S. Quantum resistant ciphertext-policy attribute-based encryption scheme with flexible access structure. arXiv preprint arXiv:2401.14076. 2024 Jan 25

[9] Wang M, Long GL. Lattice-based access authentication scheme for quantum communication networks. Science China Information Sciences. 2024 Dec;67(12):222501.

[10] Guo C, Gong B, Waqas M, Alasmary H, Tu S, Chen S. An efficient pairing-free ciphertext-policy attribute-based encryption scheme for Internet of Things. Sensors. 2024 Oct 24;24(21):6843.

[11] Mera JM, Karmakar A, Marc T, Soleimanian A. Efficient lattice-based inner-product functional encryption. InIACR International Conference on Public-Key Cryptography 2022 Feb 27 (pp. 163-193). Cham: Springer International Publishing.

[12] Bagchi P, Bisht A, Das AK, Saxena N, Hossain MS. Designing quantum-safe lattice-based multi-authority CP-ABE scheme for blockchain-enabled IoT-based consumer healthcare electronics. IEEE Transactions on Consumer Electronics. 2025 Mar 18.

[13] Wee H. Almost optimal KP and CP-ABE for circuits from succinct LWE. InAnnual International Conference on the Theory and Applications of Cryptographic Techniques 2025 Apr 28 (pp. 34-62). Cham: Springer Nature Switzerland.

[14] Lopez J, Cadena V, Rahman MS. Evaluating post-quantum cryptographic algorithms on resource-constrained devices. In2025 IEEE International

Conference on Quantum Computing and Engineering (QCE) 2025 Aug 30 (Vol. 1, pp. 918-925). IEEE.

[15] Jammula M, Vakamulla VM, Kondoju SK. Hybrid lightweight cryptography with attribute-based encryption standard for secure and scalable IoT system. Connection Science. 2022 Dec 31;34(1):2431-47.

[16] Fu X, Wang Y, You L, Ning J, Hu Z, Li F. Offline/Online lattice-based ciphertext policy attribute-based encryption. Journal of Systems Architecture. 2022 Sep 1;130:102684.

[17] Cini V, Lai R, Woo I. Lattice-based multi-authority/client attribute-based encryption for circuits. IACR Communications in Cryptology. 2025 Jan 13;1(4):1-67.

[18] Goyal V, Pandey O, Sahai A, Waters B. Attribute-based encryption for fine-grained access control of encrypted data. InProceedings of the 13th ACM conference on Computer and communications security 2006 Oct 30 (pp. 89-98).

[19] Bethencourt J, Sahai A, Waters B. Ciphertext-policy attribute-based encryption. In2007 IEEE symposium on security and privacy (SP'07) 2007 May 20 (pp. 321-334). IEEE.

[20] Goyal V, Jain A, Pandey O, Sahai A. Bounded ciphertext policy attribute based encryption. InInternational Colloquium on Automata, Languages, and Programming 2008 Jul 7 (pp. 579-591). Berlin, Heidelberg: Springer Berlin Heidelberg.

[21] Li Y, Ng KS, Purcell M. A tutorial introduction to lattice-based cryptography and homomorphic encryption. arXiv preprint arXiv:2208.08125. 2022 Aug 17.

[22] Shor, Peter W. "Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer." SIAM review 41.2 (1999): 303-332.

[23] Ajtai M. Generating hard instances of lattice problems. InProceedings of the twenty-eighth annual ACM symposium on Theory of computing 1996 Jul 1 (pp. 99-108).

[24] Gentry C, Peikert C, Vaikuntanathan V. Trapdoors for hard lattices and new cryptographic constructions. InProceedings of the fortieth annual ACM symposium on Theory of computing 2008 May 17 (pp. 197-206).

[25] Regev O. On lattices, learning with errors, random linear codes, and cryptography. Journal of the ACM (JACM). 2009 Sep 8;56(6):1-40.

[26] Peikert C. A decade of lattice cryptography. Foundations and TrendsŴ in Theoretical Computer Science. 2016 Mar 24;10(4):283-424.

[27] Lyubashevsky V, Peikert C, Regev O. On ideal lattices and learning with errors over rings. InAnnual international conference on the theory and applications of cryptographic techniques 2010 May 30 (pp. 1-23). Berlin, Heidelberg: Springer Berlin Heidelberg.

[28] Agrawal, Shweta, David Mandell Freeman, and Vinod Vaikuntanathan. "Functional encryption for inner product predicates from learning with errors." International Conference on the Theory and Application of Cryptology and Information Security. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

[29] Yao Y, Chen H, Shen L, Wang K, Wang Q. A CP-ABE scheme based on lattice LWE and its security analysis. Applied Sciences. 2023 Jul 10;13(14):8043.

[30] Sun J, Qiao Y, Liu Z, Chen Y, Yang Y. Practical multi-authority ciphertext policy attribute-based encryption from R-LWE. In2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom) 2021 Sep 30 (pp. 1435-1443). IEEE.

[31] Sun J, Qiao Y, Liu Z, Chen Y, Yang Y. Practical multi-authority ciphertext policy attribute-based encryption from R-LWE. In2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom) 2021 Sep 30 (pp. 1435-1443). IEEE.

[32] Liu Y, Wang L, Zhou Y. A novel CLWE-based attribute-based encryption scheme from lattices with privacy preserving. Cybersecurity. 2025 Oct 13;8(1):76.

[33] Hsieh YC, Lin H, Luo J. Attribute-based encryption for circuits of unbounded depth from lattices: garbled circuits of optimal size, laconic functional evaluation, and more. SIAM Journal on Computing. 2025 Apr 17:FOCS23-90.

[34] Datta P, Komargodski I, Waters B. Decentralized multi-authority ABE for DNFs from LWE. InAnnual International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT) 2021 Oct (LNCS vol 12696, pp. 177-209). Cham: Springer.

[35] Ma S, Zhao Z, Wang N, Zha C. Lattice-Based CP-ABE for Optimal Broadcast Encryption With Polynomial-Depth Circuits. IET Information Security. 2024;2024(1):6333508.

[36] Zhang J, Imani M, Sadredini E. Bp-ntt: Fast and compact in-sram number theoretic transform with bit-parallel modular multiplication. In2023 60th ACM/IEEE Design Automation Conference (DAC) 2023 Jul 9 (pp. 1-6). IEEE.

[37] Banerjee U, Ukyab TS, Chandrakasan AP. Sapphire: A configurable crypto-processor for post-quantum lattice-based protocols. arXiv preprint arXiv:1910.07557. 2019 Oct 16.

[38] Li H, Lin H, Luo J. ABE for circuits with constant-size secret keys and adaptive security. InTheory of Cryptography Conference 2022 Nov 7 (pp. 680-710). Cham: Springer Nature Switzerland.

[39] Qian X, Wu W. An efficient ciphertext policy attribute-based encryption scheme from lattices and its implementation. In2021 IEEE 6th International Conference on Computer and Communication Systems (ICCCS) 2021 Apr 23 (pp. 732-742). IEEE.

[40] Brézot T, de Perthuis P, Pointcheval D. Covercrypt: an efficient early-abort KEM for hidden access policies with traceability from the DDH and LWE. InEuropean Symposium on Research in Computer Security 2023 Sep 25 (pp. 372-392). Cham: Springer Nature Switzerland.

[41] Dai W, Doröz Y, Polyakov Y, Rohloff K, Sajjadpour H, Savaş E, Sunar B. Implementation and evaluation of a lattice-based key-policy ABE scheme. IEEE Transactions on Information Forensics and Security. 2017 Dec 4;13(5):1169-84.

[42] Mouchet C, Bossuat JP, Troncoso-Pastoriza JR, Hubaux JP. Lattigo: a multiparty homomorphic encryption library in Go. InProceedings of the 8th Workshop on Encrypted Computing & Applied Homomorphic Cryptography (WAHC) 2020 Dec (pp. 64-70).

[43] Gür KD, Polyakov Y, Rohloff K, Ryan GW, Sajjadpour H, Savaş E. Practical applications of improved Gaussian sampling for trapdoor lattices. IEEE Transactions on Computers. 2019 Apr;68(4):570-584.

[44] Bootland C, Cong K, Demmler D, Frederiksen TK, Libert B, Orfila JB, Rotaru D, Smart NP, Tanguy T, Tap S, Walter M. Threshold (fully) homomorphic encryption. Cryptology ePrint Archive. 2025.

[45] An X, Jia H, Zhang Y. Optimized password recovery for encrypted RAR on GPUs. In2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems 2015 Aug 24 (pp. 591-598). IEEE.

[46] Gupta N, Jati A, Chauhan AK, Chattopadhyay A. Pqc acceleration using gpus: Frodokem, newhope, and kyber. IEEE Transactions on Parallel and Distributed Systems. 2020 Sep 21;32(3):575-86.

[47] Chen J, Feng Y, Liu Y, Wu W. Faster binary arithmetic operations on encrypted integers. InThe 7th International Workshop on Computer Science and Engineering 2017.

[48] Smart NP. Cryptography: an introduction. New York: McGraw-Hill; 2003.

[49] Hennessy JL, Patterson DA. Computer architecture: a quantitative approach. Elsevier; 2011 Oct 7.

[50] Sravya G, Kumar PS, Padmavathy R. Survey of post-quantum lattice-based ciphertext-policy attribute-based encryption schemes for cloud storage: Taxonomy, open issues, and future directions. IEEE Transactions on Services Computing. 2024 Oct 14;17(6):4540-57.

[51] Halevi S, Shoup V. Design and implementation of HElib: a homomorphic encryption library. Cryptology ePrint Archive 2020.

[52] Katz J, Lindell Y. Introduction to modern cryptography: principles and protocols. Chapman and hall/CRC; 2007 Aug 31.

[53] Amdahl GM. Validity of the single processor approach to achieving large scale computing capabilities. InProceedings of the April 18-20, 1967, spring joint computer conference 1967 Apr 18 (pp. 483-485).

[54] Unogwu OJ, Doshi R, Hiran KK, Mijwil MM. Introduction to quantum-resistant blockchain. InAdvancements in quantum blockchain with real-time applications 2022 (pp. 36-55). IGI Global Scientific Publishing.

[55] Grassl M, Langenberg B, Roetteler M, Steinwandt R. Applying Grover's algorithm to AES: quantum resource estimates. InInternational Workshop on Post-Quantum Cryptography 2016 Feb 4 (pp. 29-43). Cham: Springer International Publishing.

[56] Bos JW, Costello C, Naehrig M, Stebila D. Post-quantum key exchange for the TLS protocol from the ring learning with errors problem. In2015 IEEE symposium on security and privacy 2015 May 17 (pp. 553-570). IEEE.

[57] Waters B. Ciphertext-policy attribute-based encryption: An expressive, efficient, and provably secure realization. InInternational workshop on public key cryptography 2011 Mar 6 (pp. 53-70). Berlin, Heidelberg: Springer Berlin Heidelberg.

[58] Huang, Boxue, Juntao Gao, and Xuelian Li. "Efficient lattice-based revocable attribute-based encryption against decryption key exposure for cloud file sharing." Journal of Cloud Computing 12.1 (2023): 37.

[59] Sahai A, Waters B. Fuzzy identity-based encryption. InAnnual international conference on the theory and applications of cryptographic techniques 2005 May 22 (pp. 457-473). Berlin, Heidelberg: Springer Berlin Heidelberg.

Binary Enhanced White Shark Optimization Algorithm and Its Application to Semi-Supervised Feature Selection

Jian Zhou*, Jiawen Yang
School of Management Engineering, Qingdao University of Technology, Qingdao, 266520, China

***Abstract*—With the development of information technology, the generation and storage of data have become convenient, resulting in a large amount of high-dimensional data. Due to the high acquisition cost of labeled data, semi-supervised learning of partially labeled data has emerged. Feature selection aims to select the most representative and informative feature subset from the original feature set and avoid the curse of dimensionality. By utilizing semi-supervised feature selection, valuable information from unlabeled data can be better extracted, thereby enhancing the model's robustness. White Shark Optimizer is a commonly used searching method, but it may fall into a local optimum when dealing with semi-supervised feature selection problems. This study proposes an Enhanced White Shark Optimization (EWSO) with a dynamic dual elite strategy and sinusoidal mutation operation. The former can make full use of the current optimal location information and improve the ability to find potential optimal solutions, and the latter can help the algorithm jump out of the local optimum and improve the accuracy of the algorithm's convergence. For dealing with discrete problems, we further introduce a binary mapping function to the EWSO, abbreviated as BEWSO. Compared with six other meta-heuristic feature selection techniques on nine high-dimensional datasets, experimental results show that the proposed method is superior to other methods. Under 80% and 60% unlabeled data, the average accuracies are 5.03% and 3.64% higher than the suboptimal value, respectively.**

***Keywords*—Feature selection; semi-supervised learning; meta-heuristic algorithm; white shark optimization**

I. Introduction

The swift progress in information technology has resulted in the accumulation of large amounts of data, which holds significant value. Data mining [1], as an important technical means, helps people to obtain value from data by discovering the patterns and knowledge hidden behind the data. However, on account of the large number and dimension of generated data, it leads to the curse of dimensionality, increasing the computational complexity and learning time. Therefore, how to obtain useful information from the data becomes an important challenge. Feature selection [2] is an effective method for dimensionality reduction, as it involves choosing the feature subsets by eliminating irrelevant and redundant ones. In real-world classification situations, obtaining data samples can be challenging due to time and resource limitations, particularly when labeling data is expensive or difficult. Thus, the importance of semi-supervised learning [3] becomes apparent. It can effectively use limited supervised information and a large amount of unlabeled data to enhance feature selection and thus improve the accuracy of learning models.

Feature selection involves choosing N features from a total of M features to enhance a specific system index, which is a key step for decreasing the dataset's dimensionality and boosting the learning algorithm's performance. By identifying the impactful features, it reduces the size and complexity of the dataset, while maintaining or improving the performance of the model. This process not only helps reduce the training time for the model, but also simplifies it, making it easier to understand and interpret. In addition, with feature selection, overfitting can be avoided, as fewer features mean that the model is not overly complex, thus avoiding the curse of dimensionality.

Regarding the usability of label information, feature selection methods can be divided into three types: supervised, semi-supervised, and unsupervised feature selection. The data generated in many practical applications has a great deal of unlabeled data samples and a limited number of labeled data samples. In fact, it requires significant time and effort for users to label each data sample. Consequently, the limited labeled samples may not provide enough information for effective supervised selection, while unsupervised methods overlook the valuable labeling information. In conclusion, semi-supervised feature selection, which considers both labeled and unlabeled data, can effectively identify the optimal subset of features.

The feature selection task is a non-deterministic polynomial-hard problem [4]. As the number of features grows, the time complexity cost of the relevant ideal search becomes unacceptable. Meta-heuristic methods are effective for addressing optimization challenges like feature selection because of their good global searching ability, self-adaptability, and computational efficiency. These algorithms can be grouped into four main types: evolution-based, swarm intelligence-based, physics-based, and human-based algorithms.

The evolution-based algorithms typically utilize genetic manipulation and a selection mechanism of individuals to optimize the solution, which may cause fall into a local optimal solution during the search process. The physics-based algorithms apply physical principles to simulate and optimize problems, which may be affected by the accuracy of physical models. The human-based algorithms often use human decision and search strategies, which may be influenced by human cognition and experience. In contrast, the swarm intelligence-based algorithms leverage the behavior of biological or social groups to simulate and optimize problems, and find better

*Corresponding author.

solutions by cooperating and sharing information, which makes wider applicability and flexibility to solve complex optimization problems.

The White Shark Optimization (WSO) Algorithm [5] is a recent swarm intelligence algorithm inspired by the predation behavior of great white sharks. It aims to address optimization problems by simulating the search strategy of great white sharks. This algorithm is used to locate and track the prey of great white sharks, utilizing their acute hearing, sense of smell, and understanding of fish behavior. It simulates the movement and adjustment of the great white shark in the search space to identify optimal solutions for small, medium, and high-dimensional optimization problems.

The innovations and contributions of this study are summarized as follows:

- The WSO method faces challenges in effectively balancing exploration and development. Its convergence speed is slow and easy to fall into local optimization. This study proposed an Enhanced White Shark Optimization (EWSO) Algorithm with a dynamic dual-elite strategy and sinusoidal mutation operation to improve its performance.
- We further introduce a binary mapping function to the EWSO, abbreviated as BEWSO. It adopts KNN-self-training as a semi-supervised classification model to further explore the binary version of the proposed algorithm in the field of semi-supervised feature selection.
- Based on high-dimensional datasets, the advantages of the proposed algorithm in semi-supervised feature selection optimization problems are verified. When compared to other meta-heuristic feature selection techniques, the average accuracies have improved by 5.03% and 3.64% based on 80% and 60% unlabeled data, respectively.

The rest of this article is organized in the following manner. Section II introduces a review of the literature on semi-supervised feature selection techniques. Section III provides a new Enhanced White Shark Optimizer algorithm and its binary version. Section IV presents a comparison of the results derived from the implementations. Finally, Section V contains the summary and guidelines for future research.

II. Literature Review

The semi-supervised feature selection method focuses on utilizing a limited number of labeled samples alongside a larger set of unlabeled samples for classification tasks. This approach can be categorized into three main types: filter, wrapper, and embedded methods. This section will primarily discuss these techniques.

A. Semi-Supervised Filter Feature Selection

Filtered semi-supervised feature selection leverages the intrinsic characteristics of both labeled and unlabeled data to evaluate features before conducting learning tasks.

This approach primarily encompasses several methods, including those based on Laplacian scores, pairwise constraints, Fisher criteria, and sparse models.

The Laplacian method is utilized to assess and select features by evaluating how well they preserve the proximity between samples in a dataset. Zhao et al. [6] introduced a method that integrates Laplacian scoring with locally sensitive discriminant analysis. This approach constructs in-class and inter-class graphs using semi-supervised datasets. Cheng et al. [7] developed a method that merges semi-supervised Laplacian scores with classification information gain. Additionally, Doquire et al. [8] presented an algorithm tailored for regression tasks, which is based on the concept of Laplace fractions.

Pairwise constraints represent a different form of supervised information. Benabdeslem et al. [9] introduced a method that integrates Laplacian score and constraint score, allowing for an effective assessment of features' ability to maintain local structure and constraints. Kalakech et al. [10] developed constraint scores in semi-supervised regression interval problems by transforming continuous label data into paired constraint information.

Fisher score is a method for measuring the ability to discriminate feature categories. Additionally, the semi-supervised feature selection challenge can be addressed by leveraging the information from unlabeled samples using Fisher's score. Yang et al. [11] introduced a semi-supervised Fisher scoring approach that utilizes both global distribution and local structure information from the dataset. Sun et al. [12] developed an algorithm based on Fisher criteria and the manifold hypothesis, which redefines the interclass scatter matrix to maximize the separation between classes, making it more resilient to complex data distributions.

By incorporating a sparsity constraint, the sparse model can automatically identify key features during training, enhancing computational efficiency. Han et al. [13] proposed a method that combines a divergence matrix with a sparse model. Zeng et al. [14] employed the l2, 1-norm to promote sparsity and introduced a semi-supervised method focused on local discriminant information.

In summary, the filter method evaluates feature subsets using a separate index based on the original features or statistical properties about the data, without integrating a machine learning model into the feature selection process. This leads to a model with high generalization but lower accuracy.

B. Semi-Supervised Wrapper Feature Selection Method

Semi-supervised wrapper feature selection methods utilize either a single learning model or an ensemble learning model to predict class labels for unlabeled data while evaluating the validity of selected feature subsets.

For single learning models, Ren et al. [15] introduced a wrapper forward framework based on semi-supervised feature selection. This method involves training classifiers with selected features and using those classifiers to predict missing labels. Unlabeled data is then randomly chosen and combined with labeled data to create a new dataset.

This process is carried out several times to identify the feature that appears frequently, which is then added to the feature subset. Song et al. [16] explored a novel particle swarm optimization algorithm for variable-size coevolutionary. They introduced a spatial partitioning strategy on feature importance, allowing related features to be grouped into the same subspace with reduced computational costs.

In the context of ensemble learning, Bellal et al. [17] introduced a guided feature ordering method known as semi-supervised ensemble learning. This method combined the bagged integration and out-of-bag feature importance measure. Barkia et al. [18] developed a novel semi-supervised feature importance evaluation method that employed a cooperative training strategy to train an integrated classifier, extending the feature importance evaluation method from the standard random forest model to semi-supervised scenarios. Han et al. [19] suggested a new method based on integration. It used an integrated classifier capable of estimating the confidence of unlabeled data to explore wrapper feature selection, so that more relevant feature subsets can be selected.

In conclusion, the wrapper approach assesses various feature subsets of a specific model through a designated learning algorithm and identifies the subset that yields the best model performance. While this results in high model accuracy, it tends to have low generalizability. Additionally, this method is more resource-demanding compared to the filter method, as it necessitates multiple iterations of the model to verify the effectiveness of the feature subset based on accuracy.

C. Semi-Supervised Embedded Feature Selection

Semi-supervised embedded feature selection employs both labeled and unlabeled data to choose features during the training process. Key concepts involved include sparse modeling, the graph Laplacian operator, and manifold regularization, among others.

Ma et al. [20] introduced a technique that leverages sparse models and the Laplace operator of graphs, primarily focusing on graph-based semi-supervised learning, along with various sparse models for feature selection. Shi et al. [21] presented an approach that utilizes the l2,1/2 matrix norm and graph Laplacian, introducing an effective optimization algorithm for web image annotation. Xu et al. [22] developed a regularization-based algorithm that selects features by enhancing the classification margin between various classes, taking into account the geometric structure of the probability distribution for both labeled and unlabeled data.

Additionally, Feofanov et al. [23] proposed a hybrid method based on Hadoop, within a semi-supervised embedded framework to address outlier and local optimization issues. Chen et al. [24] introduced a method that relies on rescaled linear regression, using the label representation matrix of unlabeled samples as the optimization target during the training of the sparse linear regression model. Yuan et al. [25] proposed a method called discriminant semi-supervised feature selection.

The embedded method integrates feature selection with the learning algorithm, allowing the feature selection and learning model processes to occur simultaneously. This approach reduces time and space requirements, enhances classification accuracy, and strikes a balance between filter and wrapper methods. Consequently, this study adopts the embedded method for feature selection.

III. Binary Enhanced White Shark Optimizer

A. Enhanced White Shark Optimizer

The White Shark Optimizer algorithm may face limitations in its search performance when applied to feature selection tasks, exhibiting slow convergence and a tendency to get stuck in local optima. Therefore, this study introduces two methods to enhance the algorithm: a dynamic double elite strategy and sinusoidal mutation operation. The dynamic dual-elite strategy optimizes the search process of the algorithm by introducing two elite individuals: the current optimal position and the suboptimal position. This strategy leverages the information from the best position and enables the algorithm to reach a high-quality solution more rapidly. At the same time, by utilizing the suboptimal position, the algorithm is capable of preserving the diversity of the population during the iteration process, allowing for broader exploration of the solution space, thus enhancing the ability to discover potential optimal solutions. The sinusoidal mutation operation is another important improvement strategy, utilizing the periodic change characteristic of the sinusoidal function to guide the individuals in the algorithm to mutate. This kind of mutation operation can make the algorithm more flexible and intelligent in the search process, allowing it to avoid getting stuck in local optima. Through the sinusoidal mutation operation, the algorithm can continuously find new and possibly better positions in the solution space, thus enhancing convergence accuracy and global search ability. These two methods are described below.

1) Dynamic dual elite strategy: The elite learning strategy is a widely used approach in optimization and evolutionary computing that focuses on retaining and utilizing high-performing individuals within a population to enhance the search process and solution quality. The core idea is to preserve the best individuals in each generation so that they can enter the next generation, thus avoiding the replacement deletion of the best individuals in the genetic iteration process. In this way, this approach allows the algorithm to reach a global optimal solution or a solution close to it more rapidly. Chen et al. [26] introduced elite opposition learning to modify the sine and cosine algorithm, enhancing population diversity. For the multi-objective scheduling problem, Amer et al. [27] proposed an improved Harris Eagle Optimizer that improves the quality of the standard Harris Hawk Optimization algorithm by using a refined opposition learning method.

Building on the enhanced algorithm mentioned earlier, this study presents a dynamic dual-elite learning strategy. This approach leverages the current optimal and suboptimal positions within the population to create the next generation of candidate solutions, which accelerates convergence and enhances the ability to identify potential optimal solutions. The strategy is chosen for the following reasons. First, by retaining and utilizing two elite individuals, it provides more search directions and possibilities, thereby increasing its convergence accuracy. Second, the number and function of elite individuals can be

adjusted dynamically based on the algorithm's operational context, allowing it to better adapt to changing environments and enhancing its robustness and adaptability. In addition, dynamic learning factors are introduced. This factor is closely linked to the fitness value of the elite white shark, and the weight of the two elite individuals can be controlled by dynamically adjusting its value. The equation for creating potential solutions and adjusting learning factors dynamically is described below.

$$X_i^*(t+1) = r_1 * F * \left(X_{second}(t) - X_i(t)\right) - r_2 * (1-F) * X_{best}(t) \quad (1)$$

$$F = \frac{F_{best}(t)}{F_{best}(t)+F_{second}(t)} \quad (2)$$

where, r_1 and r_2 are random vectors within [0,1], and $X_{best}(t)$ and $X_{second}(t)$ are individual positions with optimal and sub-optimal fitness values at the t iteration, respectively. The dynamic learning factor is denoted as F, and $F_{best}(t)$ and $F_{second}(t)$ represent the fitness values of the best and sub-best individual positions during the t iteration, respectively.

We use a selection mechanism based on hybrid sorting. First, $2N$ white sharks $X_1(t), X_2(t), \dots, X_N(t)$ and $X_1^*(t), X_2^*(t), \dots, X_N^*(t)$ are combined and sorted based on their fitness values. Then, the first N white sharks with the best fitness values were selected for the next generation, thus conducive to better population evolution. The advantages of this hybrid sorting-based selection mechanism are as follows. First, compared with traditional methods of selection based only on the original population, the mechanism doubles the number of great white sharks involved in selection during each iteration. This not only broadens the selection range but also helps identify and utilize more individuals with potentially outstanding traits, thereby significantly enhancing population diversity and evolutionary effectiveness. Secondly, this mechanism allows the algorithm to constantly explore new solution spaces throughout the search process. By introducing hybrid competition between elite and regular individuals in each iteration, the algorithm's global search capability is enhanced, leading to a higher likelihood of ultimately discovering the global optimal solution. After evaluating the fitness of each white shark in the initial population, the dynamic dual-elite strategy is applied to generate the next generation of candidate solutions based on the current best and second-best positions of the white shark population. The population is then updated according to their fitness values. The corresponding pseudocode for this algorithm is outlined in Algorithm 1.

Algorithm 1 Dynamic dual elite strategy for enhanced white shark optimizer

Initialize population size *N*;
Create the initial population *X*;
Evaluate the fitness of the initial population *X*;
for $i < N$ **do**
  Generate the learning white shark X^* using Eq. (1);
  Calculate the fitness value of X^*;
end for
Obtain the population and fitness of *2N* white sharks;
Sort them according to their fitness values;
Update population X' by selecting the best *N* white sharks with the best fitness;

2) Sinusoidal mutation operation: The sine function, with its inherent periodicity and continuous and smooth change characteristics, is widely known and highly valued in the field of science and engineering. Recently, advancements in computing technology and algorithm theory have led to various algorithms that incorporate the sine function concept in order to obtain significant enhancement and improvement in performance. For instance, Draa et al. [28] utilized a compound sine formula to modify the scale factor and crossover rate in the differential evolution algorithm. Similarly, Khalilpourazari et al. [29] introduced an enhanced crow search algorithm for complex optimization challenges by integrating sine and cosine functions.

Inspired by the above methods, the proposed algorithm uses the sinusoidal mutation theory to improve. Sinusoidal mutation theory is a method to guide the optimal individual to the mutation operation, which helps the algorithm escape local optima and increases the chances of discovering better solutions. When the sine function has the value [−1,1] on the interval [0,2π], the algorithm focuses on advancing the next generation of individuals within promising areas of the search space, instead of exploring a broader region. Therefore, the algorithm can be studied in a small range around the current global optimal location, improving the local expansion ability of the algorithm and avoiding falling into a local optimum. In addition, the algorithm aims to balance the critical requirements of exploration and exploitation. On the one hand, it actively explores the unknown and possibly better solution domain through the periodic fluctuation characteristics of the sinusoidal mutation process. On the other hand, when the sine wave amplitude drops to the local lowest point, the algorithm can accurately focus on the small change near the known best point. Eq. (3) illustrates the sinusoidal mutation process of the current globally optimal great white shark.

$$X'_{best} = \sin\left(2\pi * rand(1,\dim)\right) \oplus X_{best}, if\ rand < c \quad (3)$$

$$X_{best} = \begin{cases} X'_{best}, f(X'_{best}) < f(X_{best}) \\ X_{best},\ otherwise \end{cases} \quad (4)$$

where, X_{best} denotes the global optimal position, $\oplus$ signifies the dot product, and c is the mutation probability. As shown in Eq. (3), if the fitness value of the individual after the sinusoidal mutation improves, the result is accepted. Otherwise, the original position is maintained.

After the keen hearing, smell, and fish behavior of the great white shark, the sinusoidal mutation operation is applied to identify the global optimal position X_{best}. Initially, the sinusoidal mutation operation generates a new position, and its corresponding fitness value is compared with that of the global optimal position X_{best}, leading to an update of the algorithm's optimal solution. The pseudocode for this process is outlined in Algorithm 2.

Algorithm 2 Sinusoidal mutation for enhanced white shark optimizer

Obtain the global best position X_{best};
if $rand < c$ **then**
  Generate mutation white shark X'_{best} using Eq. (3);
  Calculate the fitness value of X'_{best};
  if $f(X'_{best}) < f(X_{best})$
    $X_{best} = X'_{best}$;
    $f\left(X_{best}\right) = f(X'_{best})$;
  end if
end if

B. Binary Enhanced White Shark Optimizer

White Shark Optimizer is a powerful algorithm designed specifically for continuous optimization problems. It can find the optimal solution by constantly adjusting variable values in a continuous search space. However, when dealing with feature selection problems, the situation is quite different. The search space in these cases is discrete rather than continuous, which means that methods suitable for continuous spaces may not be directly applied. In particular, the search space for feature selection is depicted in binary form, where 0 indicates features that are not selected, and 1 signifies those that are included. To effectively tackle feature selection issues, it is necessary to modify and optimize the enhanced white shark optimizer algorithm to accommodate binary variables. This adjustment is crucial for ensuring the algorithm's relevance and effectiveness in feature selection tasks.

In this context, transfer functions play a vital and prevalent role. These mathematical tools can efficiently map continuous values to discrete ones, particularly in converting a single continuous value into a binary state to fulfill the specific needs of feature selection. Among various transfer functions, the S-type and V-type are the most commonly utilized and effective [30]. The S-type transfer function is known for its smooth transitions and monotonically increasing nature, effectively representing the gradual change in feature importance and facilitating a smooth shift from 0 to 1 at a defined threshold. Conversely, the V-type transfer function, with its symmetrical "V" shape, is better suited for detailed feature classification, especially when feature importance is evenly distributed or has multiple peaks, allowing for more precise capture of these features. The details are shown in Table I.

In this research, we transformed continuous white shark feature data into binary format using a two-step transfer function approach, specifically employing the S1 transfer function, to enhance compatibility with subsequent feature selection and processing tasks.

The first step involves calculating the probability of the binary conversion value based on Eq. (5), $w_d^i(t)$ denotes the position of the ith white shark in the dth dimension in the tth iteration.

$$P\left(w_d^i(t)\right) = TF(w_d^i(t)) \tag{5}$$

TABLE I. S-TYPED AND V-TYPED TRANSFER FUNCTIONS

S-typed	Transfer function	V-typed	Transfer function
S1	$T(x) = \frac{1}{1+e^{-2x}}$	V1	$T(x) = \lvert\mathrm{erf}(\frac{\sqrt{\pi}}{2}x)\rvert$
S2	$T(x) = \frac{1}{1+e^{-x}}$	V2	$T(x) = \lvert\tanh(x)\rvert$
S3	$T(x) = \frac{1}{1+e^{(-x/2)}}$	V3	$T(x) = \lvert(x)/\sqrt{1+x^2}\rvert$
S4	$T(x) = \frac{1}{1+e^{(-x/3)}}$	V4	$T(x) = \lvert\frac{2}{\pi}\arctan(\frac{\pi}{2}x)\rvert$

$$TF(w) = \frac{1}{1+e^{-2w}} \tag{6}$$

Secondly, the probability of Eq. (5) is substituted into Eq. (7), and the value of each great white shark's location is converted to binary form.

$$w_d^i(t) = \begin{cases} 1, & if\ P\left(w_d^i(t)\right) \geq rand, \\ 0, & otherwise \end{cases} \tag{7}$$

The flow chart in Fig. 1 illustrates the corresponding binary algorithm.

Feature selection is typically considered a multi-objective optimization problem, and its core challenge is how to find a balance point that maximizes classification accuracy while reducing the number of features chosen. This balance is essential for enhancing the efficiency, interpretability, and generalization of machine learning models. The fitness function is crucial in this context, serving as the primary criterion for assessing the selected feature subset. Specifically, the fitness function outlined in Eq. (8) is essential for evaluating the selected subset of features, as it balances the goal of maximizing classification accuracy with the need to minimize the number of features used in the learning model. The smaller the function value, the better the performance.

$$F = \alpha * (1 - acc) + \beta * \frac{l_f}{n_f} \tag{8}$$

The classification accuracy depends on the selected classifier, which is represented by acc in this study. The parameters α and β are inverse parameters, representing the weight given to the quality of classification and the feature selection ratio, respectively, ranging from 0 to 1, where $\beta = 1 - \alpha$. In this study, α is 0.99 and β is 0.01. The values of l_f and n_f refer to the number of selected features and the sum of features in the dataset, respectively.

Several semi-supervised learning models have been developed to enhance learning efficiency and accuracy by utilizing a small amount of labeled data in conjunction with a larger volume of unlabeled data. These models include self-training classification models, mixed classification models based on large expectations, cooperative training models, graph-based semi-supervised learning models, and semi-supervised learning clustering models. Among these, the self-training classification model is popular due to its effective training results and relatively straightforward implementation. The K-

Nearest Neighbors (K-NN) algorithm is an efficient classification method known for its simplicity, ease of use, strong predictive performance, insensitivity to outliers, and suitability for large datasets. As a result, it is frequently applied in areas such as text sentiment analysis, risk assessment, and medical diagnosis. To sum up, this study adopts KNN-self-training as a semi-supervised classification model algorithm, with k set to 5. Each run utilizes 5-fold cross-validation to evaluate model accuracy. Initially, 20% of the data is chosen as the test dataset, with the remainder serving as the training dataset. Given the semi-supervised nature of the algorithm, the experiment randomly selects 80% and 60% of the training dataset as unlabeled data (class labels are removed), and the rest as labeled data for model training, respectively. Finally, after making predictions on the unlabeled data, 20% of the test data is used as experimental results to verify the classification model accuracy.

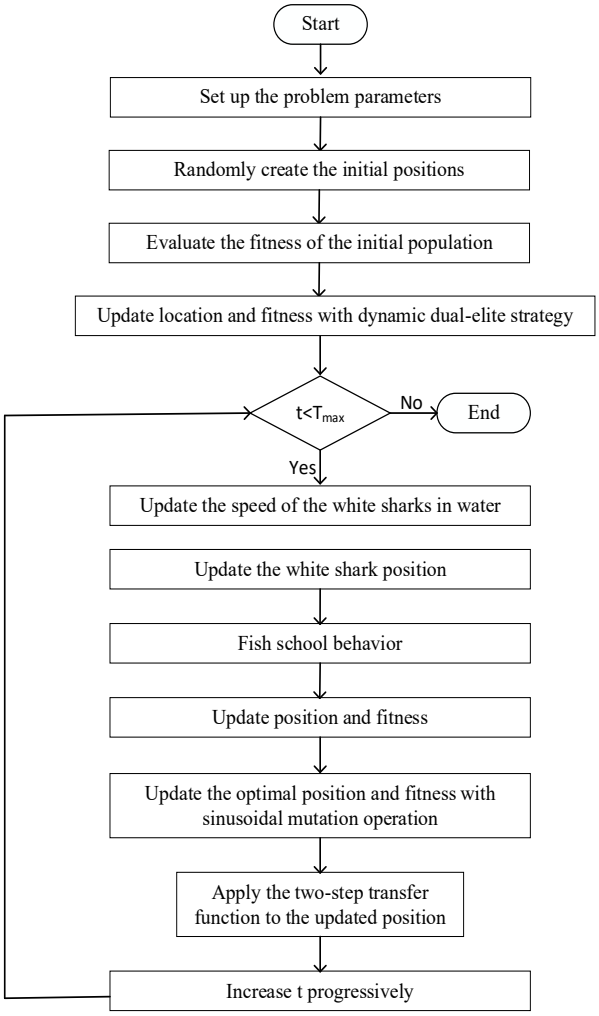


Fig. 1. The flowchart of the binary enhanced white shark optimizer algorithm.

IV. Experimental Results

This section outlines the experiments conducted and the results achieved to showcase the efficacy of the suggested algorithm when applied with feature selection in established classification tasks. We utilize the MATLAB 2022b platform to implement the proposed algorithm on nine high-dimensional datasets. Additionally, we analyzed and compared it with six other prominent meta-heuristic techniques commonly employed in feature selection.

A. Datasets

This article uses 9 high-dimensional datasets from the UCI Machine Learning Repository [31] and the Arizona State University (ASU) Feature Selection Repository [32]. Table II provides a comprehensive description of these datasets, including the feature dimensions, number of categories, sample sizes, and dataset sources.

TABLE II. Classified Datasets About UCI and ASU

Number	Dataset	Feature	Class	Instance	Source
1	Isolet	617	26	1560	ASU
2	lung	3312	5	203	UCI
3	COIL20	1024	20	1440	ASU
4	ALLAML	7129	2	72	ASU
5	darwin	450	2	174	UCI
6	Toxicity	1203	2	171	UCI
7	DrivFace	6400	3	606	UCI
8	USPS	256	10	2007	ASU
9	leukemia	7070	2	72	ASU

B. Compared Algorithms and Parameter Settings

To assess the proposed algorithm, the results are compared with six meta-heuristic optimization algorithms. The first is the original Binary White Shark Optimizer (BWSO) [33]. The other five commonly used methods include Binary Differential Evolution (BDE) [34], Binary Gravity Search Algorithm (BGSA) [35], Binary Atom Search Optimization (BASO) [36], Binary Grasshopper Optimization Algorithm (BGOA) [37], and Binary Particle Swarm Optimization Algorithm (BPSO) [38]. Table III shows the related parameter settings.

TABLE III. Compared Algorithm Parameter Settings

Algorithm	Parameter	Value
BEWSO	Population size	30
	f_{min}	0.07
	f_{max}	0.75
	p_{min}	0.5
	p_{max}	1.5
BDE	Crossrate	0.9
	N	10
BGSA	a	20
	G_0	100
BASO	Alpha	50
	Beta	0.2
	V_{max}	6
	N	10
BGOA	N	1
BPSO	V_{max}	6
	W_{max}	0.9
	W_{min}	0.4
	N	10
Universal	Maximum iterations	100
	Runs	30

C. Results and Discussion

Table IV presents the average fitness values achieved by the proposed algorithm and the other algorithms across nine datasets using 80% unlabeled data. The optimal and sub-optimal results for each dataset are indicated in bold and underlined, respectively. Notably, the average fitness score of the suggested approach is 36.52% lower than the sub-optimal value of BPSO. The proposed algorithm achieves the lowest average fitness value across all nine datasets. The BASO method ranks second, achieving four sub-optimal results, while both BWSO and BDE obtain two sub-optimal results each. Conversely, the BGSA method may perform poorly, with the highest overall average fitness value and no optimal or sub-optimal results across all datasets.

Using 60% unlabeled data, Table V displays the average fitness values obtained on the nine high-dimensional datasets. The overall average fitness value of the proposed algorithm is 35.29% lower than the sub-optimal value of BPSO, further showcasing its effectiveness. Similarly, the proposed algorithm achieves the lowest average fitness value across all datasets, with BPSO following closely behind and achieving four sub-optimal results, while BWSO secures three sub-optimal results.

TABLE IV. AVERAGE FITNESS VALUE UNDER DATA WITH 80% UNLABELED

Dataset	BEWSO	BWSO	BDE	BGSA	BASO	BGOA	BPSO
1	**0.1657**	0.2682	0.2140	0.2311	0.2080	0.2619	0.2150
2	**0.0282**	0.0565	0.0564	0.1016	0.0472	0.0792	0.0489
3	**0.0764**	0.0991	0.1321	0.1515	0.0987	0.1329	0.1169
4	**0.0391**	0.1111	0.0904	0.1677	0.1973	0.1535	0.1066
5	**0.1195**	0.1511	0.1697	0.2086	0.1927	0.2517	0.1548
6	**0.1362**	0.1655	0.2560	0.2603	0.2437	0.1695	0.1738
7	**0.0275**	0.0712	0.0717	0.0754	0.0598	0.0664	0.0771
8	**0.0756**	0.1064	0.0987	0.1184	0.1011	0.0912	0.1101
9	**0.0119**	0.0732	0.0395	0.0852	0.0671	0.0827	0.0691
Total average	**0.0756**	0.1225	0.1254	0.1555	0.1351	0.1432	0.1191

TABLE V. AVERAGE FITNESS VALUE UNDER DATA WITH 60% UNLABELED

Dataset	BEWSO	BWSO	BDE	BGSA	BASO	BGOA	BPSO
1	**0.1291**	0.1721	0.1683	0.2025	0.1577	0.1763	0.1570
2	**0.0172**	0.0571	0.0308	0.0875	0.0421	0.0546	0.0296
3	**0.0443**	0.0647	0.0577	0.0796	0.0622	0.0670	0.0545
4	**0.0275**	0.0761	0.0622	0.0757	0.0693	0.0957	0.0947
5	**0.0946**	0.1240	0.1552	0.1445	0.1556	0.1659	0.1463
6	**0.1099**	0.1508	0.1633	0.1985	0.2101	0.1423	0.1487
7	**0.0198**	0.0371	0.0470	0.0533	0.0505	0.0547	0.0593
8	**0.0517**	0.0983	0.0879	0.1053	0.0814	0.0692	0.0634
9	**0.0420**	0.0656	0.0749	0.0898	0.0707	0.0757	0.0755
Total average	**0.0596**	0.0940	0.0941	0.1152	0.1000	0.1002	0.0921

Table VI displays the average accuracy of the proposed method and other test methods using 80% unlabeled data. The results indicate the proposed algorithm performs well in accuracy and obtains the best overall average value, which is 5.03% higher than the sub-optimal value from BPSO, aligning with the best overall objective function value shown in Table IV. Specifically, it records the highest average classification accuracy when considering all datasets. The average accuracy of the lung, COIL20, ALLAML, DrivFace, USPS, and leukemia datasets reached more than 90%. In contrast, BWSO, BDE, and BASO also have better average accuracy. Among them, the BWSO algorithm and BASO algorithm get three sub-optimal results, respectively, and the BDE algorithm gets two sub-optimal results. However, the BGSA method performs poorly, exhibiting the lowest overall average accuracy and failing to achieve any optimal or sub-optimal results across all datasets.

Table VII outlines the average accuracy of the proposed approach and other tested methods utilizing unlabeled data. The proposed method outperforms the WSO method as well as all other competing techniques, achieving the best overall average value, which is 3.64% higher than the corresponding sub-optimal value for BDE. It also secures the highest average accuracy across all nine high-dimensional datasets, with over 90% accuracy on seven of them. In contrast, BWSO, BDE, and BPSO also have better average accuracy, with the BPSO algorithm achieving four sub-optimal results, while BDE and BWSO each attain three.

TABLE VI. AVERAGE ACCURACY UNDER DATA WITH 80% UNLABELED

Dataset	BEWSO	BWSO	BDE	BGSA	BASO	BGOA	BPSO
1	**0.8373**	0.7372	0.7917	0.7717	0.7949	0.7404	0.7885
2	**0.9750**	0.9500	0.9500	0.9025	0.9550	0.9250	0.9542
3	**0.9301**	0.9063	0.8750	0.8520	0.9028	0.8715	0.8828
4	**0.9643**	0.8929	0.9167	0.8357	0.8048	0.8500	0.8976
5	**0.8824**	0.8529	0.8363	0.7941	0.8078	0.7539	0.8461
6	**0.8676**	0.8382	0.7490	0.7422	0.7578	0.8333	0.8292
7	**0.9752**	0.9304	0.9339	0.9267	0.9422	0.9380	0.9280
8	**0.9302**	0.8953	0.9077	0.8855	0.9027	0.9127	0.8952
9	**0.9952**	0.9286	0.9643	0.9191	0.9381	0.9238	0.9357
Total average	**0.9286**	0.8813	0.8805	0.8477	0.8673	0.8610	0.8841

TABLE VII. AVERAGE ACCURACY UNDER DATA WITH 60% UNLABELED

Dataset	BEWSO	BWSO	BDE	BGSA	BASO	BGOA	BPSO
1	**0.8758**	0.8333	0.8381	0.8005	0.8430	0.8269	0.8430
2	**0.9875**	0.9500	0.9750	0.9167	0.9625	0.9500	0.9750
3	**0.9584**	0.9444	0.9479	0.9247	0.9444	0.9375	0.9479
4	**0.9762**	0.9286	0.9443	0.9286	0.9333	0.9076	0.9095
5	**0.9118**	0.8824	0.8510	0.8612	0.8451	0.8412	0.8560
6	**0.8935**	0.8529	0.8428	0.8052	0.7941	0.8611	0.8549
7	**0.9835**	0.9669	0.9587	0.9512	0.9527	0.9497	0.9452
8	**0.9535**	0.9052	0.9177	0.8988	0.9227	0.9352	0.9402
9	**0.9627**	0.9367	0.9286	0.9143	0.9331	0.9286	0.9287
Total average	**0.9448**	0.9112	0.9116	0.8890	0.9034	0.9042	0.9112

Overall, in high-dimensional datasets using both 80% and 60% unlabeled data, the improved algorithm consistently outperforms the original and other algorithms in classification accuracy. As the proportion of labeled data increases and the proportion of unlabeled data decreases, classification accuracy improves, possibly due to the enhanced correctness of labeled samples, leading to a more effective trained model.

From a deeper perspective, the proposed algorithm shows good performance regarding classification accuracy. By introducing a dynamic bi-elite strategy and sinusoidal mutation operation, the method can utilize the current optimal position information, which improves the ability to find potentially superior solutions and allows for more effective data dimensionality reduction, resulting in improved classification accuracy and efficiency.

Table VIII illustrates the mean number of features chosen by the proposed and other testing methods when utilizing 80% unlabeled data. In general, the proposed method performs admirably in feature selection, ranking third overall, following BWSO and BASO. Specifically, out of 9 datasets, the proposed method obtains 2 best results (Isolet, ALLAML) and 1 suboptimal result (lung); BWSO obtains 3 best results (DrivFace, USPS, leukemia). Additionally, BASO achieves 2 best results (lung, Toxicity) and four suboptimal results (COIL20, ALLAML, Darwin, DrivFace). BPSO obtains the best results in 2 cases (COIL20, Darwin). Combined with Table IV and Table VI, for these two datasets (Isolet, ALLAML), the suggested approach not only obtains the optimal number of feature selections, but also obtains the optimal average fitness and the optimal average accuracy.

Table IX illustrates the mean number of features chosen by the proposed methods and other testing methods when 60% unlabeled data is used. Overall, the suggested method ranks second in the average number of selected features, just behind BASO. Specifically, the suggested approach has the fewest average selected features across three datasets (lung, Toxicity, DrivFace), while BASO has 2 cases (ALLAML, darwin) and BPSO has 2 cases (Isolet, USPS) each. Additionally, BWSO has 1 optimal value (leukemia), and BGOA has 1 optimal value (COIL20) each.

The analysis shows that this algorithm selects fewer features than most algorithms, while maintaining better fitness values. However, despite having some advantages over other methods, it is difficult to reach the ideal number of selected features. This is primarily because the method emphasized in this study prioritizes feature selection to enhance classification accuracy, focusing on optimizing the fitness function with accuracy as the main objective. The proposed algorithm effectively reduces the number of features selected while retaining essential information.

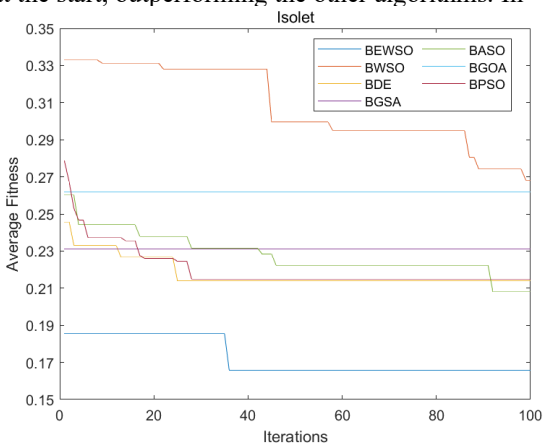
TABLE VIII. THE AVERAGE NUMBER OF FEATURES SELECTED UNDER DATA WITH 80% UNLABELED

Dataset	BEWSO	BWSO	BDE	BGSA	BASO	BGOA	BPSO
1	**285.50**	496.15	477.17	310.47	304.10	302.10	341.50
2	1129.12	2312.04	2278.03	1667.17	**801.08**	1640.14	1163.87
3	735.06	642.08	850.00	511.20	246.00	405.50	**89.10**
4	**2696.50**	3613.07	5652.77	3567.73	2841.83	3557.33	3734.90
5	134.33	249.00	344.43	216.00	112.57	362.07	**108.07**
6	627.50	645.10	905.90	601.93	**479.77**	537.57	567.30
7	1920.10	**1480.10**	4020.14	3198.70	1639.08	3209.50	2786.00
8	165.10	**69.17**	189.10	129.33	123.10	123.00	162.87
9	5100.53	**1780.10**	2952.13	3539.83	4100.17	5131.13	3854.07
Total average	1421.53	1254.09	1963.30	1526.93	**1183.08**	1696.48	1423.08

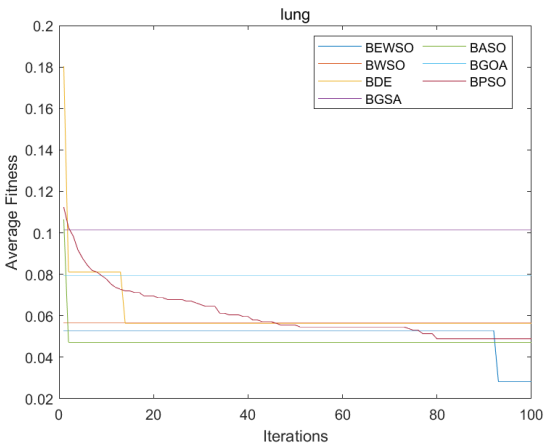
TABLE IX. THE AVERAGE NUMBER OF FEATURES SELECTED UNDER DATA WITH 60% UNLABELED

Dataset	BEWSO	BWSO	BDE	BGSA	BASO	BGOA	BPSO
1	378.10	436.05	494.27	306.90	138.10	306.10	**96.10**
2	**1595.04**	2519.10	2018.00	1658.70	1638.57	1673.05	1620.00
3	316.05	991.12	627.05	510.73	741.10	**52.03**	301.02
4	2778.67	3844.67	5007.53	3564.30	**2360.20**	3001.90	3641.13
5	324.10	339.12	345.33	320.12	**100.80**	390.03	169.60
6	**540.70**	626.20	926.85	678.10	751.00	578.20	601.37
7	**2186.11**	2826.02	3925.13	3201.57	2346.20	3100.60	3256.10
8	145.10	114.03	164.12	130.60	125.10	128.00	**106.02**
9	3500.34	**2013.34**	3012.23	3527.03	3156.23	3526.10	3450.24
Total average	1307.13	1523.29	1835.61	1544.23	**1261.92**	1417.33	1471.29

Fig. 2 illustrates the convergence patterns of seven optimization algorithms by analyzing the average fitness function values over 100 search iterations. The evaluation includes the suggested algorithm alongside six additional algorithms tested on nine datasets, each comprising 80% unlabeled data. The X-axis represents the number of iterations, and the Y-axis displays the average fitness value obtained. For the Isolet, COIL20, ALLAML, and USPS datasets, the proposed method shows good convergence in early iterations and continues to improve until the end. In the Toxicity and DrivFace datasets, the proposed method rapidly reaches the minimum fitness value at the start, outperforming the other algorithms. In the lung dataset, the proposed method initially lags behind BASO, but by the ninetieth iteration, it surpasses BPSO and maintains superior values. For the Darwin dataset, BWSO and BPSO are competitive algorithms. It can be seen that the proposed method starts strong, but BPSO excels after about 40 iterations, while BWSO performs best after around 50 iterations, and the proposed method again reaches the minimum value after about 70 iterations. In the leukemia dataset, the proposed method has the best initial convergence, which is surpassed by BDE in the middle, but the proposed method soon maintains the minimum fitness value again.



(a)



(b)

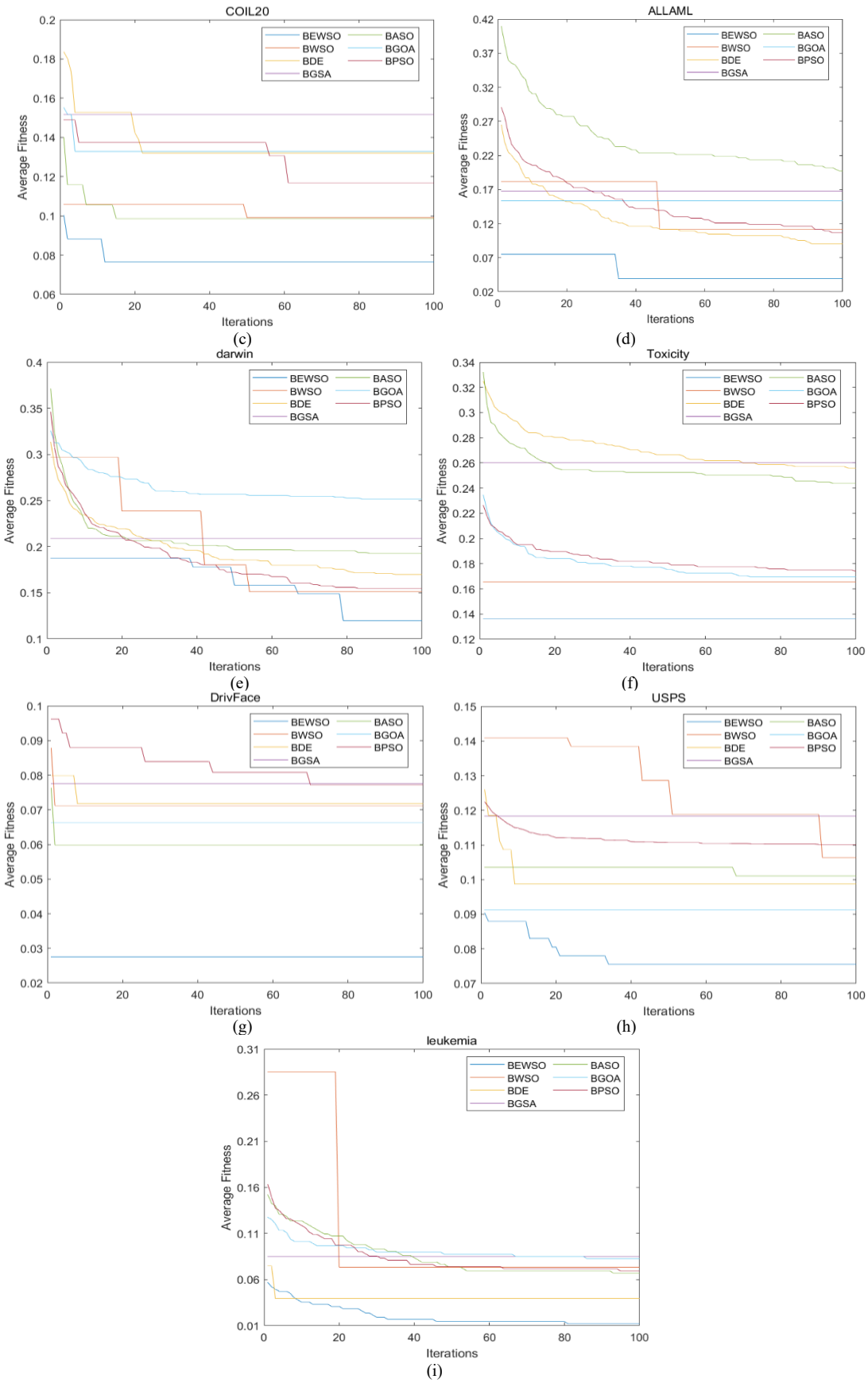
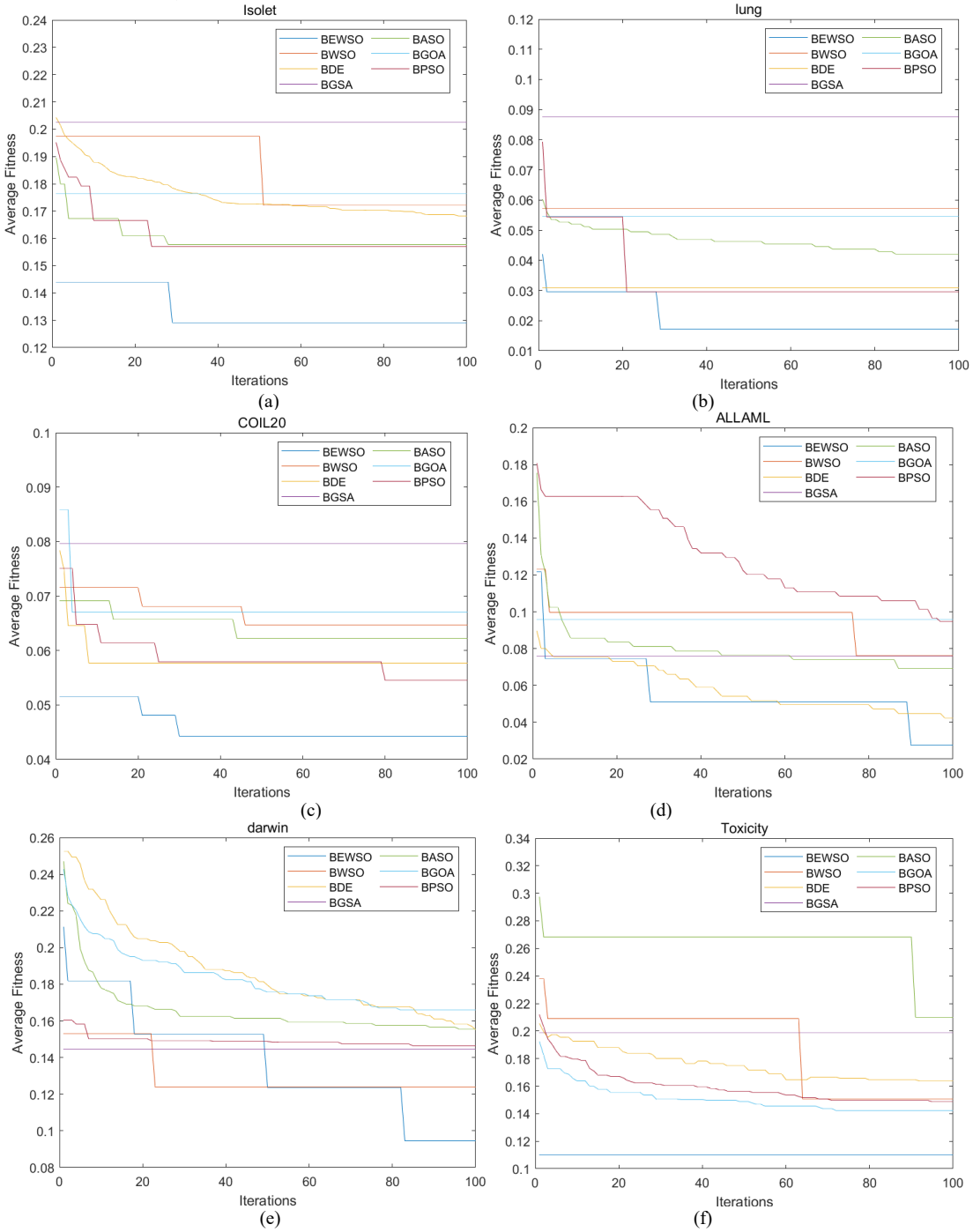


Fig. 2. The average convergence curve of the algorithms under data with 80% unlabeled.

Fig. 3 uses 9 datasets with 60% unlabeled data to describe the convergence behavior of the proposed method and the other six methods based on the mean value of the fitness function during 100 iterative searches. For most datasets, it is evident that the convergence curve becomes stagnant after more than 50 iterations. Examining each convergent graph individually, it reveals that the proposed method outperforms its competitors in the Isolet, lung, COIL20, DrivFace, and USPS datasets, showing good convergence in early iterations, with the average fitness value decreasing further with more iterations. In the Toxicity dataset, the suggested algorithm attains the best fitness value during the initial iteration, with BGOA following closely behind. In the ALLAML dataset, BGSA and BDE are the most competitive algorithms. BGSA gets the best fitness value in the first iteration, and as iterations progress, BDE and the proposed method alternately get the best fitness value until the proposed method finally gets the best value around the 90th iteration. In the Darwin dataset, BGSA starts with the best fitness value, BDE takes the lead around the 20th iteration, and both the proposed algorithm and BDE reach the same minimum fitness value around the 50th iteration, with the proposed method showing stronger convergence by the 80th iteration. For the leukemia dataset, the proposed method shows good convergence around the 30th iteration and maintains this performance until the end.



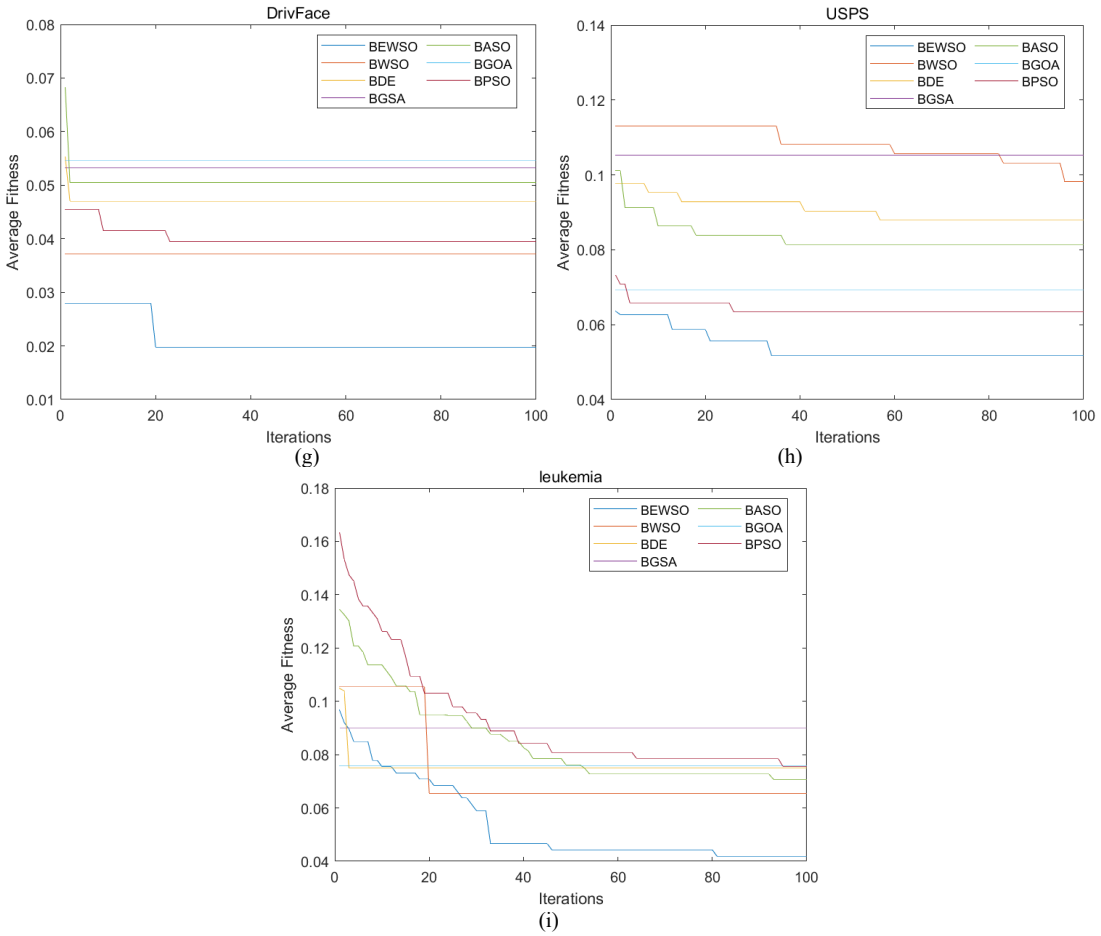


Fig. 3. The average convergence curve of the algorithms under data with 60% unlabeled.

From a practical perspective, the time taken for the algorithm to finish its calculations is crucial. The execution times for the suggested and other competing approaches have been measured and are presented in Table X. BGSA has the lowest overall average running time, with the proposed algorithm coming in second. Specifically, BGSA records the fastest execution time across seven datasets, while the proposed method obtains three sub-optimal results.

Furthermore, Table XI displays the computation time results for seven different algorithms using 60% unlabeled datasets. BGSA again has the shortest overall average running time, followed by the proposed algorithm. BGSA achieves seven optimal results and one suboptimal result, while the proposed method secures four suboptimal results.

TABLE X. THE AVERAGE TIME OF THE ALGORITHMS UNDER DATA WITH 80% UNLABELED

Dataset	BEWSO	BWSO	BDE	BGSA	BASO	BGOA	BPSO
1	40.70	45.79	73.63	21.04	63.00	**20.65**	74.50
2	36.49	37.86	53.39	**23.70**	50.05	53.74	27.32
3	48.15	39.00	68.96	**11.58**	53.47	69.33	51.00
4	28.00	**21.35**	26.91	21.60	25.60	23.00	29.00
5	13.00	19.00	14.00	**11.00**	15.52	24.55	14.84
6	19.16	17.20	15.77	**15.00**	**15.00**	16.00	**15.00**
7	63.00	64.00	75.00	**23.00**	91.09	70.00	85.00
8	32.00	38.00	65.00	**27.00**	56.80	60.00	67.00
9	38.00	37.60	50.45	**17.13**	38.23	50.00	34.00
Total average	35.39	35.53	49.23	**19.01**	45.42	43.03	44.18

TABLE XI. THE AVERAGE TIME OF THE ALGORITHMS UNDER DATA WITH 60% UNLABELED

Dataset	BEWSO	BWSO	BDE	BGSA	BASO	BGOA	BPSO
1	34.00	36.00	56.00	**13.31**	26.00	19.10	38.00
2	28.10	31.58	36.00	**13.50**	45.51	46.00	48.00
3	39.57	32.00	47.44	**12.67**	53.00	35.00	35.01
4	14.90	15.00	25.00	**10.63**	33.00	29.10	27.66
5	6.00	6.05	15.48	**5.00**	13.60	23.53	13.00
6	12.00	**11.24**	20.00	13.00	17.00	23.00	33.05
7	35.00	49.00	30.00	**13.67**	52.00	55.00	46.00
8	25.00	23.00	34.13	18.00	45.20	**17.00**	23.00
9	24.00	24.00	18.00	**13.00**	35.10	24.10	25.00
Total average	24.29	25.32	31.34	**12.53**	35.60	30.20	32.08

V. CONCLUSION

This study introduces an Enhanced White Shark Optimization algorithm with the dynamic dual-elite strategy and the sinusoidal mutation method, aimed at increasing the diversity of offspring and enhancing the potential optimal solutions. Furthermore, it introduces a binary transfer function and further investigates its binary application in semi-supervised feature selection. The proposed algorithm is evaluated on nine high-dimensional datasets and is compared against six other meta-heuristic methods for feature selection. The results indicate that when 20% and 40% of the training samples are randomly labeled, the proposed algorithm achieves the best overall mean accuracy; its lead rate is 5.03% and 3.64%, respectively, while also decreasing the number of features chosen. Specifically, with 20% labeled data, it surpasses other developed algorithms in average accuracy of 9 datasets and the number of selected features of 2 datasets. With 40% labeled data, it again outperforms all other developed algorithms in average accuracy of 9 datasets and number of selected features of 3 datasets. The experimental findings show the good performance of the Binary Enhanced White Shark Optimization method. This method could be applied to study high-dimensional gene selection in the future.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (Grant No. 72403137) and the Natural Science Foundation of Shandong Province (Grant No. ZR2024QG024).

REFERENCES

[1] Wu, W. T., Li, Y. J., Feng, A. Z., Li, L., Huang, T., Xu, A. D., Lyu, J.. Data mining in clinical big data: the frequently used databases, steps, and methodological models. Military Medical Research, 2021, 8(1): 44.

[2] Cai, J., Luo, J. W., Wang, S. L., Yang, S.. Feature selection in machine learning: A new perspective. Neurocomputing, 2018, 300: 70-79.

[3] Camargo, G., Bugatti, P. H., Saito, P. T. M.. Active semi-supervised learning for biological data classification. Public Library of Science, 2020, 15(8): e0237428.

[4] Hodashinsky, I., Sarin, K., Shelupanov, A., Slezkin, A.. Feature Selection Based on Swallow Swarm Optimization for Fuzzy Classification. Symmetry-basel, 2019, 11(11): 1423.

[5] Braik, M., Hammouri, A., Atwan, J.. White Shark Optimizer: A novel bio-inspired meta-heuristic algorithm for global optimization problems. Knowledge-Based Systems, 2022, 243: 108457.

[6] Zhao, J. D., Lu, K., He, X. F.. Locality sensitive semi-supervised feature selection. Neurocomputing, 2008, 71(10-12): 1842-1849.

[7] Cheng, H. R., Deng, W., Fu, C., Wang, Y., Qin, Z. G.. Graph-based semi-supervised feature selection with application to automatic spam image identification. International Workshop on Computer Science for Environmental Engineering and EcoInformatics. 2011, 259-264.

[8] Doquire, G., Verleysen, M.. A graph Laplacian based approach to semi-supervised feature selection for regression problems. Neurocomputing, 2013, 121: 5-13.

[9] Benabdeslem, K., Hindawi, M.. Constrained Laplacian score for semi-supervised feature selection. Joint European Conference on Machine Learning and Knowledge Discovery, 2011: 204-218.

[10] Kalakech, M., Biela, P., Macaire, L., Hamad, D.. Constraint scores for semi-supervised feature selection: a comparative study. Pattern Recognition Letters, 2011, 32(5): 656-665.

[11] Sheikhpour, R., Sarram, M. A., Gharaghani, S.. A survey on semi-supervised feature selection methods. Pattern Recognition, 2017, 64: 141-158.

[12] Sunzhong, L. V., Jiang, H., Zhao, L.. Manifold based fisher method for semi-supervised feature selection. International Conference on Fuzzy Systems and Knowledge Discovery, 2013, 664-668.

[13] Han, Y. H., Yang, Y., Yan, Y.. Semisupervised feature selection via spline regression for video semantic recognition. IEEE Transactions on Neural Networks and Learning Systems, 2015, 26(2): 252-264.

[14] Zeng, Z. Q., Wang, X. D., Zhang, J., Wu, Q.. Semi-supervised feature selection based on local discriminative information. Neurocomputing, 2016, 173: 102-109.

[15] Ren, J., Qiu, Z., Fan, W.. Forward semi-supervised feature selection. Conference on knowledge discovery and data mining, 2008: 970-976.

[16] Song, X. F., Zhang, Y., Guo, Y. N., Sun, X. Y., Wang, Y. L.. Variable-size cooperative coevolutionary particle swarm optimization for feature selection on high-dimensional data. IEEE Transactions on Evolutionary Computation, 2020, 24(5): 882-895.

[17] Bellal, F., Elghazel, H., Aussem, A.. A semi-supervised feature ranking method with ensemble learning. Pattern Recognition Letters, 2012, 33(10): 1426-1433.

[18] Barkia, H., Elghazel, H., Aussem A.. Semi-supervised feature importance evaluation with ensemble learning. IEEE International Conference on Data Mining, 2011:31-40.

[19] Han, Y. K., Park, K. S., Lee, Y. K.. Confident wrapper-type semi-supervised feature selection using an ensemble classifier. International Conference on Artificial Intelligence Management Science and Electronic Commerce, 2011: 4581-4586.

[20] Ma, Z. G., Nie, F. P., Yang, Y.. Discriminating joint feature analysis for multimedia data understanding. IEEE Transactions on Multimedia, 2012, 14(6): 1662-1672.

[21] Shi, C. J., Ruan, Q. Q., An, G.. Sparse feature selection based on graph Laplacian for web image annotation. Image and Vision Computing, 2014, 32(3): 189-201.

[22] Xu, Z. L., King, I., Lyu, M. R. T., Jin, R.. Discriminative semi-supervised feature selection via manifold regularization. IEEE Transactions on Neural networks, 2010, 21(7): 1033-1047.

[23] Feofanov, V., Devijver, E., Amini, M. R.. Semi-supervised wrapper feature selection by modeling imperfect labels. arXiv, 2019.

[24] Lin, Y. J., Hu, Q., Zhang, J., Wu, X. D.. Multi-label feature selection with streaming labels. Information Sciences, 2016, 372: 256-275.

[25] Peng, Y., Chen, G., Xu, M., Wang, C. J., Xie, J. Y.. Multi-label learning by exploiting label correlations with LDA. International Conference on Tools with Artificial Intelligence, 2017: 168-174.

[26] Chen, L., Ma, L. Y., Li, L.. Enhancing sine cosine algorithm based on social learning and elite opposition-based learning. Computing, 2024, 106(5): 1475-1517.

[27] Amer, D. A., Attiya, G., Zeidan, I., Nasr, A. A.. Elite learning Harris Hawks Optimizer for multi-objective task scheduling in cloud computing. Journal of Supercomputing, 2022, 78(2): 2793-2818.

[28] Draa, A., Chettah, K., Talbi, H.. A Compound Sinusoidal Differential Evolution algorithm for continuous optimization Check. Swarm and Evolutionary Computation, 2019, 50, 100450.

[29] Khalilpourazari, S., Pasandideh, S. H. R.. Sine-cosine crow search algorithm: theory and applications. Neural Computing and Applications, 2020, 32(12): 7725-7742.

[30] Agrawal, P., Ganesh, T., Oliva, D., Mohamed, A. W.. S-shaped and V-shaped gaining-sharing knowledge-based algorithm for feature selection. Journal information, 2022, 52(1): 81-112.

[31] Dheeru, D., Karra, T. E.. UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences, 2007.

[32] Tawhid, M. A., Ibrahim, A. M.. Feature selection based on rough set approach, wrapper approach, and binary whale optimization algorithm. International Journal of Machine Learning and Cybernetics, 2020,11(3): 573–602.

[33] Hammouri, A. I., Braik, M. S., Al-hiary, H. H., Abdeen, R. A.. A binary hybrid sine cosine white shark optimizer for feature selection. Cluster Computing-The Journal of Networks Software Tools and Applications, 2024, 27(6): 7825-7867.

[34] Pampara, G., Engelbrecht, A. P., Franken, N.. Binary differential evolution. IEEE International Conference on Evolutionary Computation, 2006, 873–1879.

[35] Rashedi, E., Nezamabadi-pour, H., Saryazdi, S.. BGSA: binary gravitational search algorithm. Natural Computing 2010, 9(3): 727-745.

[36] Too, J., Abdullah, A. R.. Binary atom search optimization approaches for feature selection. Connection Science, 2020, 32(4): 406-430.

[37] Mafarja, M., Aljarah, I., Faris, H., Hammouri, A. I., Mirjalili, S.. Binary grasshopper optimisation algorithm approaches for feature selection problems. Expert Systems with Applications, 2018, 117: 267-286.

[38] Chuang, L. Y., Chang, H. W., Tu, C. J., Yang, C. H.. Improved binary PSO for feature selection using gene expression data. Computational Biology and Chemistry, 2008, 32(1): 29–38.

NADOS: Reliability-Gated Difficulty-Aware Oversampling for Noisy Imbalanced Classification

Ednel Ashraff Misran*, Syahid Anuar, Adam Mohd Khairuddin
Faculty of Artificial Intelligence, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia

***Abstract*—Class imbalance becomes more challenging when minority underrepresentation is accompanied by label noise, because synthetic oversampling may amplify unreliable local structures if noisy minority samples are used as interpolation seeds. This study proposes Noise-Aware Adaptive-Difficulty Oversampling (NADOS), which separates the assessment of seed trustworthiness from the allocation of synthesis effort. NADOS evaluates each minority instance through two local criteria. A reliability score determines whether the instance is suitable for synthetic generation, while a difficulty score assigns synthesis priority among eligible instances. The method is evaluated on 22 binary imbalanced benchmark datasets, five controlled label-noise levels, four classifier families, and nine oversampling methods. Across the full benchmark, NADOS obtains average F1 = 0.7428, G-mean = 0.8511, balanced accuracy = 0.8552, and AUPRC = 0.7700. These results give NADOS the strongest average F1, G-mean, and balanced accuracy, while remaining competitive on AUPRC. Non-parametric statistical tests show significant differences among the compared methods and indicate that NADOS is consistently competitive under noisy imbalanced learning conditions. Separating seed reliability from synthesis difficulty provides a practical strategy for noisy imbalanced classification. The reliability gate limits unsafe seed usage, while the difficulty score preserves attention to difficult but trustworthy minority samples.**

***Keywords*—Imbalanced classification; label noise; oversampling; SMOTE; noisy data; minority learning**

I. Introduction

Class imbalance remains a persistent challenge in supervised learning because many classifiers favor the majority class when the training distribution is skewed. This issue is critical in tasks where the minority class is the main class of interest, such as medical screening, fault detection, fraud analysis, and anomaly identification. The problem becomes more difficult when the training labels are noisy, because the learner must recover a sparse minority concept from data that may contain incorrect or locally inconsistent labels [1, 2, 3].

Synthetic oversampling is widely used to address class imbalance. SMOTE and its variants increase minority representation by generating artificial minority samples instead of duplicating existing observations. However, synthetic generation is not always beneficial. If a seed instance is noisy, isolated, or located in an unstable overlap region, interpolation can create misleading samples and reinforce incorrect local structure. This risk is especially important in noisy imbalanced learning, where the minority class is sparse and more vulnerable to local corruption [4, 5, 6].

Recent oversampling methods reduce this risk through adaptive generation, boundary information, local density, cleaning mechanisms, and noise-aware weighting. These methods have improved synthetic sample generation, but a key design issue remains: difficult minority instances are not difficult for the same reason. Some difficult samples are informative boundary cases, while others are unreliable because they are mislabeled, isolated, or structurally inconsistent. Treating both cases alike can direct synthesis effort toward unsafe seeds. Conversely, focusing only on safe regions can neglect useful hard cases near the decision boundary [7, 8, 9, 10].

This study proposes Noise-Aware Adaptive-Difficulty Oversampling (NADOS) to address this distinction. NADOS separates minority-seed assessment into two decisions. Reliability acts as an eligibility gate: a minority instance must be sufficiently trustworthy before it can guide synthetic generation. Difficulty then acts as an allocation signal: among reliable seeds, harder instances receive greater synthesis priority. This design aims to avoid unsafe generation while retaining useful boundary information.

The main contributions are fourfold: a reliability-gated and difficulty-aware oversampling method for noisy imbalanced classification; a local scoring mechanism that separates seed trustworthiness from synthesis priority; a controlled benchmark evaluation across 22 binary imbalanced datasets, five label-noise levels, four classifier families, and nine oversampling methods; and non-parametric statistical validation against established SMOTE-family and noise-aware oversampling baselines.

The remainder of this study is organized as follows. Section II reviews related work. Section III presents the proposed NADOS method. Section IV describes the experimental design. Section V reports and discusses the results. Section VI concludes the study.

II. Related Work

This section reviews the main oversampling families that motivate NADOS. It first summarizes classical data-level methods, then discusses adaptive, boundary-oriented, cleaning-based, and noise-aware variants. The review emphasizes not only what each method does, but also what previous studies suggest about its behavior under overlap, sparsity, and label-noise conditions.

A. Classical Oversampling for Imbalanced Learning

Class imbalance has been widely studied because many standard learning algorithms optimize global error and

*Corresponding author.

therefore tend to favor the majority class when the class distribution is skewed [1, 2]. Data-level methods remain attractive because they can be applied before classifier training and do not require changes to the underlying model. Random oversampling is the simplest form of this strategy: minority instances are duplicated until the class distribution becomes less skewed. Although this can reduce majority-class dominance, it may also increase redundancy and overfitting because the minority distribution is enlarged by repetition rather than by new information.

SMOTE changed data-level imbalance learning by generating synthetic minority instances through interpolation between neighboring minority samples [4]. Instead of duplicating observations, SMOTE creates new samples along line segments between a minority seed and one of its minority nearest neighbors. This idea made synthetic oversampling a dominant baseline for imbalanced classification and motivated many later variants. Its main advantage is generality: it can be combined with many classifiers and applied to tabular datasets without requiring a problem-specific cost matrix.

However, SMOTE also introduced a structural risk. The method assumes that minority neighbors provide useful interpolation directions. When the minority seed is noisy, isolated, or located in a majority-dominated overlap region, interpolation can create synthetic samples that blur the class boundary or reinforce incorrect local structure. This problem is especially important when class imbalance and label noise occur together. In such cases, the learner is not only deprived of minority examples, but it may also receive synthetic minority points generated from unreliable seeds. Reviews of imbalanced learning repeatedly identify the structure of minority examples, local overlap, and noisy labels as factors that can make resampling unreliable [5, 1, 2].

Empirical studies have consistently shown that synthetic oversampling can improve minority recognition compared with simple resampling, but the gains depend strongly on data geometry and classifier behavior. The original SMOTE study reported improved ROC behavior relative to random over-sampling and under-sampling on several benchmark tasks [4]. Later comparative studies found that SMOTE combined with cleaning rules can reduce overlap effects in some settings, but no single resampling method is uniformly best across datasets [11, 5]. These outcomes support the need for methods that respond to local structure rather than applying a single global balancing rule.

B. Adaptive, Locality-Aware, and Noise-Aware Oversampling

One major extension of SMOTE is adaptive generation. ADASYN assigns more synthetic samples to minority instances that are harder to learn, using local class composition to estimate difficulty [7]. This is an important shift because it recognizes that not all minority samples should receive equal generation effort. The method attempts to move the decision boundary toward difficult minority examples by increasing their representation. The limitation is that local difficulty is not always equivalent to usefulness. A sample may be difficult because it lies near an informative boundary, but it may also be difficult because it is mislabeled, isolated, or locally inconsistent.

Boundary-oriented approaches address a related issue. Borderline-SMOTE focuses on minority samples near the class boundary, based on the assumption that boundary samples are more informative for discrimination than samples located deep inside safe minority regions [12]. SVM-SMOTE similarly uses support-vector information to guide generation near the decision boundary [13]. These methods are important competitors because they explicitly move away from uniform minority interpolation. Nevertheless, their emphasis on boundary regions can become problematic under label noise. Boundary and overlap regions are precisely where unreliable labels and ambiguous neighborhoods are more likely to appear.

Another line of work combines oversampling with data cleaning. SMOTE-Tomek and SMOTE-ENN use cleaning rules after synthetic generation to remove overlapping or locally inconsistent samples [11]. These hybrid approaches are useful because they acknowledge that oversampling can introduce ambiguous points. However, post-generation cleaning does not fully solve the seed-selection problem. If unreliable seeds guide interpolation, the method may generate poor candidates before the cleaning step is even applied, a concern also reflected in filtering-based approaches for noisy and borderline examples [6]. This motivates methods that improve synthetic generation earlier in the pipeline.

Noise-aware and density-aware methods attempt to control this risk more directly. WRND is particularly relevant because it uses relative neighborhood density to adapt oversampling for imbalanced noisy classification [8]. Other recent methods use clustering, safe-region estimation, local density, or filtering to improve robustness. These methods show that the field has moved beyond simple class balancing toward local structure-aware generation. However, many approaches still rely on a single local signal to influence both seed selection and synthesis intensity. This creates a design tension: the same evidence is often used to decide whether a sample should be trusted and how much it should be emphasized.

Prior experimental work also shows that adaptive and locality-aware methods can improve performance when the minority class contains informative boundary regions, but the same mechanisms may overemphasize unreliable points when label noise is present. ADASYN was reported to improve learning by concentrating synthesis on harder minority samples [7], while Borderline-SMOTE and related variants showed benefits from focusing on samples near decision boundaries [12, 13]. WRND extends this direction to noisy imbalanced classification by weighting generation with relative neighborhood density [8]. These outcomes suggest that local evidence is useful, but they also expose the central question addressed here: whether the evidence used to trust a seed should be separated from the evidence used to allocate synthesis effort.

C. Research Gap

This study addresses a gap in existing oversampling methods: the decision to trust a minority instance as a synthesis seed is often treated together with the decision about how much synthetic generation effort that instance should receive. These two decisions should be separated. Seed eligibility concerns whether a candidate minority instance is reliable

enough to guide interpolation, whereas synthesis allocation concerns how strongly a reliable seed should contribute to class balancing.

This distinction matters because difficulty can arise from different causes. A minority instance surrounded by majority samples may be an informative boundary case, but it may also be mislabeled or isolated. A method that increases generation near all difficult samples may amplify noise. In contrast, a method that selects only safe minority regions may miss boundary cases that are useful for discrimination. The methodological gap is therefore not simply the need for adaptive oversampling, but the need to separate seed eligibility from synthesis allocation.

NADOS addresses this gap by using reliability to determine seed eligibility and difficulty to determine synthesis priority. This differs from methods that treat local difficulty alone as a reason for more generation, and from methods that focus only on safe regions without explicitly preserving informative hard cases. The method first asks whether a minority instance should be trusted, and only then determines how much influence it should have during synthetic generation.

In relation to the preceding studies, NADOS treats seed eligibility and synthesis allocation as two separate decisions. Classical SMOTE mainly assumes that selected minority neighbors are trustworthy [4], ADASYN and Borderline-SMOTE emphasize difficult or boundary samples [12, 7], and cleaning hybrids remove some problematic samples only after generation [11, 14, 15]. NADOS instead screens candidate seeds before interpolation and then allocates generation pressure only among seeds that pass this screen (Fig. 1).

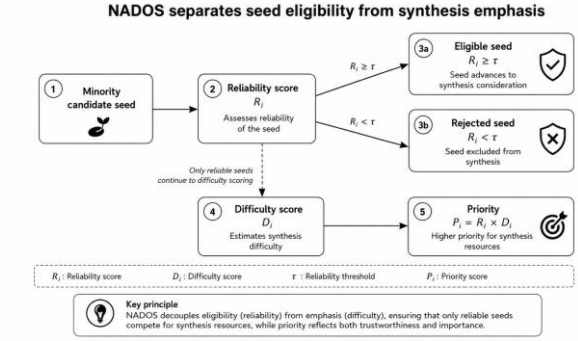


Fig. 1. NADOS decision logic separating seed eligibility from synthesis priority.

III. Proposed Method

This section defines the proposed NADOS method. It begins with the binary noisy imbalanced learning setting, introduces the local neighborhood evidence used by the method, defines the reliability and difficulty scores, and then describes how these scores are combined into a priority-based synthetic generation procedure.

A. Problem Setting

Let a binary training set be denoted as:

$$D = \{(x_i, y_i)\}_{i=1}^{n}$$

where, x_i is a feature vector, and y_i is the class label. The minority class is underrepresented relative to the majority class. The observed training data may also contain label noise. The goal is to generate synthetic minority samples that improve minority-sensitive classification performance without reinforcing unreliable local structure.

B. Local Neighborhood Evidence

For each minority instance x_i, NADOS constructs a k-nearest-neighbor neighborhood using the training fold. Let:

- m_i be the number of minority neighbors among the k-nearest neighbors.
- d_i^{min} be the average distance from x_i to minority neighbors.
- d_i^{maj} be the average distance from x_i to majority neighbors.

Neighborhood composition captures local minority support, while relative distance structure indicates whether the instance is closer to minority or majority regions.

C. Reliability Score

The reliability score estimates whether a minority instance is sufficiently trustworthy to be used as a synthesis seed. It combines minority-neighbor support with relative distance evidence:

$$R_i = \lambda\left(\frac{m_i}{k}\right) + (1-\lambda)\left(\frac{d_i^{\text{maj}}}{d_i^{\text{min}} + d_i^{\text{maj}}}\right)$$

where, λ controls the contribution of neighborhood composition. A higher R_i indicates stronger local minority support and greater distance from majority neighbors. A minority instance is eligible for synthetic generation only if:

$$R_i \geq \tau$$

where, τ is the reliability threshold.

The reliability score functions as a conservative screen rather than a full generation policy. The first term, m_i / k, measures same-class support in the local neighborhood. A minority instance with stronger minority support is less likely to be an isolated noisy point. The second term compares distances to majority and minority neighbors. If majority neighbors are relatively farther away, the instance is less likely to be embedded in a majority-dominated region. Combining these signals avoids relying only on neighbor counts, which can be unstable in sparse regions, or only on distance ratios, which can be sensitive to local scale.

D. Difficulty Score

The difficulty score estimates how much synthesis effort should be assigned to an eligible minority instance. NADOS combines a boundary component with a local sparsity component:

$$B_i = 1 - \left|2\left(\frac{m_i}{k}\right) - 1\right|$$

$$\rho_i = \frac{m_i}{k}$$

$$D_i = \eta B_i + (1 - \eta)(1 - \rho_i)$$

where, B_i is high near mixed neighborhoods and 1 - ρ_i reflects lower minority support. The parameter η controls the balance between boundary difficulty and sparsity difficulty.

The difficulty score is deliberately separated from reliability. A great difficulty score does not by itself make an instance suitable for interpolation; it only indicates that the instance may deserve more synthesis attention after passing the reliability gate. This prevents the method from treating unreliable difficult samples and informative difficult samples as equivalent. The boundary component emphasizes mixed neighborhoods, while the sparsity component gives attention to locally under-supported minority regions.

E. Synthesis Priority

NADOS assigns synthesis priority only to reliable seeds:

$$P_i = R_i \times D_i, if\ R_i \geq \tau;\ otherwise\ P_i = 0$$

The priority values are normalized across eligible minority seeds. Synthetic samples are then generated by selecting seed instances according to their normalized priorities and interpolating between the selected seed and one of its minority neighbors:

$$x_{new} = x_i + \alpha(x_j - x_i)$$

where, x_j is a minority neighbor of x_i and α is sampled uniformly from [0, 1].

The multiplicative priority formulation was selected because reliability and difficulty serve different roles. Reliability acts as a trust signal, whereas difficulty allocates synthesis effort among candidate seeds. A simple additive rule could allow a highly difficult but weakly reliable instance to retain substantial priority, which would conflict with the noise-aware motivation of NADOS. The product rule makes priority high only when both conditions are satisfied: the instance is sufficiently reliable and also informative enough to deserve synthesis emphasis. This provides a conservative first formulation for noisy imbalanced settings, where difficult minority instances may represent useful boundary cases but may also correspond to mislabeled or unstable samples.

F. Algorithm 1

Algorithm 1 NADOS oversampling procedure

1:	**input:** D, cmin, k, λ, η, τ, target ratio r	
2:	**initialize** D′ ← D and required synthetic count	
3:	***for*** each minority instance x_i **do**	
4:	compute R_i from support and distance evidence	▷ reliability
5:	compute D_i from boundary and sparsity components	▷ difficulty
6:	set $P_i = R_i \times D_i$ if $R_i \geq \tau$; else $P_i = 0$	▷ priority
7:	**end for**	
8:	***while*** target minority ratio r is not reached **do**	
9:	sample by normalized P_i; generate xnew = x_i + $\alpha(x_j - x_i)$; append to D′	▷ synthesis
10:	**output** resampled training set D′	

This procedure makes NADOS a reliability-gated extension of adaptive interpolation-based oversampling. The gate prevents unsafe seeds from driving generation, while the difficulty score preserves synthesis attention for reliable but challenging minority instances.

The computational cost of NADOS is dominated by nearest-neighbor search and synthetic sample generation. For each training fold, neighborhood evidence is computed for minority instances, after which generation proceeds similarly to SMOTE but with priority-weighted seed selection. Thus, the method adds local scoring overhead relative to plain SMOTE but avoids classifier-specific optimization or iterative model fitting inside the resampling step (Fig. 2).

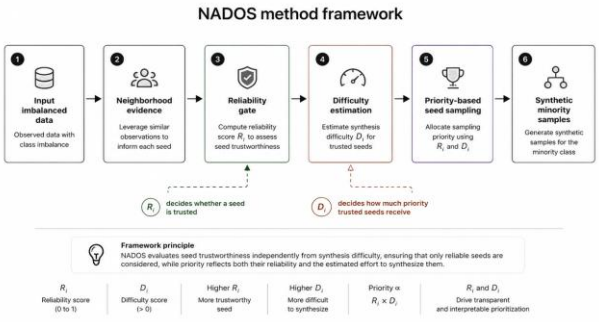


Fig. 2. NADOS method framework.

The main experiment used fixed settings across all datasets, classifiers, folds, and noise levels: k = 5, reliability weight λ = 0.50, difficulty weight η = 0.50, reliability threshold τ = 0.35, target minority ratio 1.00, product priority R_i x D_i, noise levels of 0%, 5%, 10%, 15%, and 20%, and random seed 42.

IV. Experimental Design

This section describes the benchmark datasets, controlled noise protocol, classifiers, baseline oversampling methods, evaluation metrics, statistical tests, and software environment used to evaluate NADOS. The experiments were implemented in Python using standard scientific-computing and machine-learning libraries, including scikit-learn and imbalanced-learn for model evaluation and baseline resampling support [16, 17].

A. Benchmark Datasets

The evaluation uses 22 binary imbalanced benchmark datasets: ecoli-0_vs_1, ecoli1, ecoli2, ecoli3, glass-0-1-2-3_vs_4-5-6, glass0, glass1, glass6, haberman, iris0, new-thyroid1, new-thyroid2, page-blocks0, pima, segment0, vehicle0, vehicle1, vehicle2, vehicle3, wisconsin, yeast1, and yeast3. These datasets are drawn from public imbalanced-learning benchmark collections derived from KEEL and UCI-style tabular datasets [18, 19]. They cover multiple sample sizes and imbalance ratios, giving a diverse benchmark for tabular imbalanced classification.

The benchmark contains 22 binary imbalanced tabular datasets with 150 to 5472 instances, 3 to 19 features, and imbalance ratios ranging from 1.82 to 8.79. The datasets include ecoli, glass, haberman, iris, new-thyroid, page-blocks, pima, segment, vehicle, wisconsin, and yeast variants.

B. Controlled Label Noise

Because the selected benchmark datasets do not provide a consistent natural-noise setting, controlled symmetric label noise is injected into the training folds. Noise levels of 0%, 5%, 10%, 15%, and 20% are used. Noise is applied only to training folds, while test folds remain unchanged. This design evaluates whether each oversampling method can learn from corrupted training data while preserving a clean evaluation reference.

This ordering is important for experimental validity. If noise injection or resampling were applied before the train-test split, synthetic samples or corrupted-label effects could leak into the evaluation fold. The protocol therefore follows a strict sequence: split the data, inject noise into the training fold only, apply resampling to the noisy training fold only, train the classifier, and evaluate on the untouched test fold. This makes the measured performance reflect robustness to noisy training data rather than contamination of the test reference (Fig. 3).

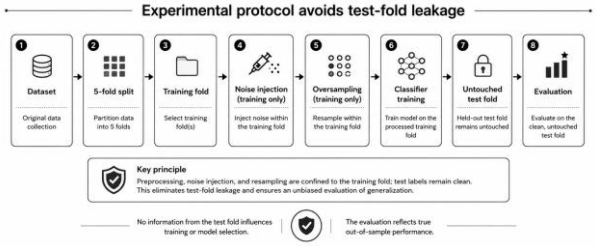


Fig. 3. Experimental protocol used to prevent test-fold leakage.

C. Classifiers and Baselines

Four classifiers are used: logistic regression, support vector machine, random forest, and gradient boosting. NADOS is compared against eight baselines: Random Oversampling, SMOTE, ADASYN, Borderline-SMOTE, SVM-SMOTE, SMOTE-ENN, SMOTE-Tomek, and WRND.

The classifier set is intentionally heterogeneous. Logistic regression provides a linear baseline, SVM captures margin-based behavior, random forest represents bagged tree ensembles, and gradient boosting represents sequential tree ensembles. This reduces the chance that the evaluation reflects compatibility with only one classifier family. The baseline set covers duplication, classical interpolation, adaptive difficulty-based oversampling, boundary-guided oversampling, hybrid cleaning, and a recent noise-aware method.

D. Metrics

The primary metrics are F1, G-mean, AUPRC, and balanced accuracy. Minority recall and runtime are also reported. These metrics are selected because standard accuracy is not sufficient for imbalanced classification and can hide poor minority-class recognition; precision-recall analysis is also recommended when the positive or minority class is rare [20, 21]. Let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. Precision, recall, specificity, and the primary summary metrics are defined as follows:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

$$G\text{-}mean = \sqrt{Recall \times Specificity}$$

$$Balanced\ accuracy = \frac{Recall + Specificity}{2}$$

$$AUPRC = \Sigma\,(Recall_t - Recall_{t-1})\,Precision_t$$

AUPRC summarizes the precision-recall curve and is useful when the minority class is the primary class of interest.

E. Statistical Testing

The comparison uses non-parametric statistical testing because multiple methods are evaluated across multiple datasets, folds, classifiers, and noise levels. Friedman tests are used to test whether overall differences exist among methods [22]. Pairwise Wilcoxon signed-rank tests with Holm correction are then used to compare NADOS against each baseline [23, 24].

Each dataset-fold-noise-classifier combination is treated as a comparison block. The Friedman test evaluates whether the methods differ in rank across these repeated benchmark conditions. Because the global test does not identify which method pairs differ, Wilcoxon signed-rank tests are then applied to paired NADOS-versus-baseline results. Holm correction is used to control the family-wise error rate across multiple pairwise comparisons. This combination of average performance, rank analysis, and corrected paired tests provides a more reliable interpretation than mean scores alone.

V. Results and Discussion

This section reports the empirical findings from three complementary perspectives: aggregate metric values, average rank analysis, and pairwise statistical comparisons against baselines. It then discusses robustness under increasing label noise and the runtime implications of the proposed scoring procedure.

A. Overall Performance

Overall performance is summarized across all datasets, folds, classifiers, and noise levels.

Across the full benchmark, NADOS obtains average F1 = 0.7428, G-mean = 0.8511, AUPRC = 0.7700, balanced accuracy = 0.8552, minority recall = 0.8445, and runtime = 0.1609 seconds. The closest competitor is SVM-SMOTE, with F1 = 0.7417, G-mean = 0.8506, AUPRC = 0.7746, and balanced accuracy = 0.8555. SMOTE-Tomek, SMOTE, Random Oversampling, and WRND remain competitive but are lower on the main balanced classification metrics.

NADOS obtains the highest average F1 and G-mean and is effectively tied with SVM-SMOTE on balanced accuracy. SVM-SMOTE obtains the highest average AUPRC. This suggests that NADOS is strongest for balanced minority-

sensitive classification performance, while AUPRC remains competitive but not dominant (Table I).

TABLE I. MEAN PERFORMANCE ACROSS ALL DATASETS, FOLDS, CLASSIFIERS, AND NOISE LEVELS

Method	F1	G-mean	Bal. Acc.	AUPRC
NADOS	0.7428	0.8511	0.8552	0.7700
SVM-SMOTE	0.7417	0.8506	0.8555	0.7746
SMOTE-Tomek	0.7356	0.8497	0.8536	0.7695
SMOTE	0.7329	0.8480	0.8519	0.7682
Random oversampling	0.7357	0.8468	0.8515	0.7663
WRND	0.7405	0.8379	0.8445	0.7595
SMOTE-ENN	0.7026	0.8376	0.8444	0.7513
ADASYN	0.6980	0.8342	0.8391	0.7517
Borderline-SMOTE	0.6955	0.8303	0.8355	0.7404

NADOS does not dominate every dataset-metric combination, but it has the strongest aggregate profile for F1, G-mean, and balanced accuracy. Noisy imbalanced benchmark performance is heterogeneous, and different resampling methods may be favored by different dataset geometries, classifiers, and noise levels. The contribution should therefore be interpreted through aggregate performance, average ranking, and statistically validated comparisons rather than as universal dominance.

The overall results are consistent with the method design. F1, G-mean, and balanced accuracy reward methods that improve minority recognition without allowing majority performance to collapse. NADOS performs well under these criteria because the reliability gate limits unsafe seed usage while the difficulty score still emphasizes challenging reliable regions. SMOTE-ENN achieves the highest average minority recall, but its F1 and G-mean are lower, indicating that recall gains alone are insufficient when precision or majority recognition deteriorates.

The AUPRC result should be interpreted cautiously. SVM-SMOTE achieves the highest AUPRC, suggesting that margin-guided boundary synthesis can be effective for ranking minority instances by predicted score. NADOS remains competitive but does not dominate this metric. This indicates that reliability-gated generation improves classification balance more clearly than probability ranking quality in the current configuration.

B. Average Rank Analysis

The Friedman ranking results show that NADOS achieves the best average rank for F1, G-mean, and balanced accuracy. For AUPRC, SVM-SMOTE obtains the best average rank, while NADOS ranks third but remains close to SMOTE-Tomek and SMOTE.

The Friedman and average-rank results support the aggregate performance findings.

The Friedman tests are significant for all primary metrics ($p < 0.001$), indicating systematic differences among methods across benchmark blocks. NADOS has the best average rank for F1 (4.2786), G-mean (4.3755), and balanced accuracy (4.3873). For AUPRC, SVM-SMOTE obtains the best average rank (4.2130), followed by SMOTE-Tomek and SMOTE, while NADOS remains close with rank 4.6200 and average AUPRC = 0.7700.

The significant Friedman tests indicate that the performance differences among methods are not random across the benchmark conditions. The rank profile supports the main claim that NADOS is a strong general-purpose oversampling method under noisy imbalanced settings, particularly for metrics that directly reflect balanced class recognition.

The rank results are useful because average scores can be influenced by a small number of large datasets or easier benchmark conditions. Ranking each method within repeated comparison blocks provides a distribution-free view of consistency. NADOS obtains the best average rank for three of the four primary metrics, which supports its robustness across datasets, classifiers, folds, and noise levels. Its third-place AUPRC rank reinforces the need to position the method as balanced and robust rather than universally superior (Fig. 4).

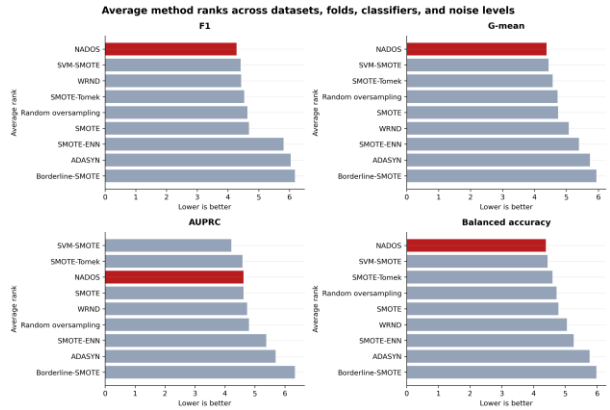


Fig. 4. Average method ranks across the primary evaluation metrics.

C. Pairwise Comparison Against Baselines

Wilcoxon-Holm pairwise tests show that NADOS significantly improves F1 over Borderline-SMOTE, ADASYN, SMOTE-ENN, SMOTE, Random Oversampling, and SMOTE-Tomek. The F1 differences relative to SVM-SMOTE and WRND are not statistically significant after Holm correction, suggesting that these methods remain close competitors.

For G-mean and balanced accuracy, NADOS significantly improves over most baselines, including WRND, SMOTE-ENN, ADASYN, Borderline-SMOTE, SMOTE, Random Oversampling, and SMOTE-Tomek. Differences relative to SVM-SMOTE are not significant. For AUPRC, NADOS improves over several weaker baselines but does not outperform SVM-SMOTE.

These findings should be interpreted cautiously. NADOS should not be framed as universally dominant, but as a method that consistently strengthens balanced minority-sensitive performance while remaining competitive with the strongest SMOTE-family baseline.

The pairwise results clarify the practical meaning of the average ranks. NADOS provides statistically significant gains over ADASYN and Borderline-SMOTE across F1, G-mean, and balanced accuracy. Since these methods also emphasize difficult or boundary samples, the result supports the argument that difficulty should not be used alone as a generation signal. NADOS also improves over WRND for G-mean and balanced accuracy, suggesting that reliability-gated synthesis can compete with a recent density-aware noise-robust strategy. The non-significant differences against SVM-SMOTE indicate that margin-guided synthesis remains a strong competitor.

The closest Wilcoxon-Holm comparisons reinforce this interpretation.

The Wilcoxon-Holm tests show significant F1 gains for NADOS over Borderline-SMOTE (mean difference = 0.0476, $p < 0.001$), ADASYN (0.0441, $p < 0.001$), SMOTE (0.0100, $p < 0.001$), and SMOTE-Tomek (0.0075, $p < 0.001$). The F1 differences against SVM-SMOTE (0.0012, $p = 0.3224$) and WRND (0.0023, $p = 0.9411$) are not significant. For G-mean and balanced accuracy, NADOS significantly improves over WRND and most SMOTE-family baselines, while differences against SVM-SMOTE remain non-significant.

D. Component Ablation and Parameter Sensitivity

To isolate the contributions of the two NADOS components, two ablated variants were evaluated using the same benchmark protocol. The reliability-only variant applies the reliability gate but distributes synthesis uniformly among eligible seeds. The difficulty-only variant removes the reliability gate and allocates synthesis according to the difficulty score alone. Full NADOS performs best across all four reported metrics, indicating that the reliability gate and difficulty-weighted allocation contribute complementary effects rather than one component fully subsuming the other (Table II).

TABLE II. ABLATION COMPARISON FOR FULL NADOS AND COMPONENT VARIANTS

Variant	F1	G-mean	Bal. Acc.	AUPRC
Full NADOS	0.7428	0.8511	0.8552	0.7700
Reliability-only	0.7354	0.8478	0.8523	0.7657
Difficulty-only	0.7184	0.8438	0.8480	0.7597

TABLE III. SELECTED PARAMETER SENSITIVITY SUMMARY ON REPRESENTATIVE DATASETS

Setting	Parameter	Mean F1	Mean G-mean	Mean AUPRC
rel30/diff70	lambda/eta	0.5706	0.7620	0.6173
rel50/diff50	lambda/eta	0.5748	0.7595	0.6227
rel70/diff30	lambda/eta	0.5726	0.7566	0.6237
tau=0.20	threshold	0.5287	0.7862	0.6014
tau=0.35	threshold	0.5467	0.7806	0.6065
tau=0.50	threshold	0.5748	0.7595	0.6227

A limited parameter sensitivity check evaluates the reliability threshold and the relative reliability-difficulty weighting on representative datasets. The main configuration uses $k = 5$, lambda = 0.50, eta = 0.50, and tau = 0.35. The value $k = 5$ was selected as a local-neighborhood setting commonly used in SMOTE-family experiments and was fixed to avoid dataset-specific tuning. The threshold and weighting checks show that performance varies by dataset, which supports the decision to report a fixed reproducible configuration rather than tune parameters separately for each benchmark condition (Table III).

E. Robustness Under Label Noise

The controlled noise protocol evaluates whether each method remains useful as training labels become increasingly corrupted. NADOS is designed for this setting because unreliable seeds should be filtered before synthetic generation. The reliability gate is expected to matter most as noise increases, since noisy minority instances are more likely to produce harmful synthetic samples when used as interpolation seeds.

The noise-level results show a clear robustness trend. In the no-noise setting, SMOTE-Tomek, SMOTE, and SVM-SMOTE are slightly ahead of NADOS on G-mean. Once label noise is introduced, NADOS becomes increasingly competitive and obtains the strongest G-mean at 10% and 15% noise. At 20% noise, SVM-SMOTE remains slightly ahead, but NADOS remains stronger than SMOTE, SMOTE-Tomek, and WRND on G-mean and balanced accuracy. This supports the intended role of the reliability gate: NADOS is most useful when the training data contain noisy or unstable minority candidates.

This trend also explains why the method should be evaluated under noise rather than only on clean imbalanced data. In the no-noise setting, classical SMOTE-family methods remain highly competitive because many minority seeds are usable. As noise increases, unsafe seed selection becomes more costly. NADOS does not prevent all performance degradation, but it delays the point at which noisy local structure damages balanced performance. The clearest advantage appears at moderate noise levels, especially 10% and 15%, where reliability screening is valuable, but neighborhood evidence remains informative.

Overall, the G-mean trend shows that NADOS is most competitive in moderate-noise settings. At the highest noise level, SVM-SMOTE is slightly stronger, but NADOS remains close and stronger than SMOTE, SMOTE-Tomek, and WRND on the main balanced metrics (Fig. 5).

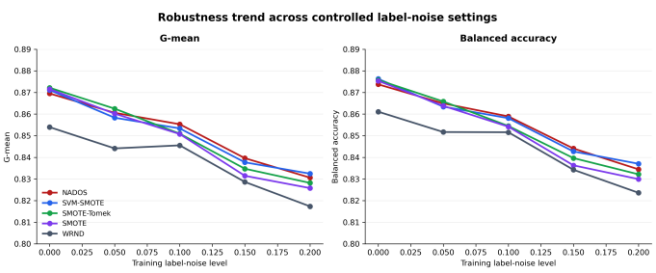


Fig. 5. Robustness trend across controlled label-noise settings.

F. Runtime and Computational Complexity

The runtime results indicate that NADOS remains practical at the benchmark scale used in this study. As shown in Table IV, NADOS records an average runtime of 0.1609 seconds and an average runtime rank of 4.80 across the benchmark comparisons. It is not the fastest method: SMOTE-ENN, Random Oversampling, WRND, and SMOTE-Tomek have better average runtime ranks. However, NADOS remains faster on average than SVM-SMOTE, SMOTE, Borderline-SMOTE, and ADASYN, while providing stronger aggregate F1, G-mean, and balanced accuracy.

TABLE IV. RUNTIME COMPARISON ACROSS ALL BENCHMARK SETTINGS

Method	Mean runtime (s)	Runtime rank
SMOTE-ENN	0.1042	1.51
Random oversampling	0.1497	3.94
SMOTE-Tomek	0.1661	4.35
WRND	0.1499	4.53
NADOS	0.1609	4.80
SVM-SMOTE	0.1683	5.84
SMOTE	0.1739	6.05
Borderline-SMOTE	0.1809	6.92
ADASYN	0.1868	7.04

The additional cost comes from computing local neighborhood evidence before generating synthetic samples. For n training instances, n_min minority instances, d features, k nearest neighbors, and G generated synthetic samples, a direct implementation computes distances between each minority instance and the training set. This requires O(n_min n d) distance operations. After the neighbor lists are available, synthetic interpolation requires O(Gd) operations because each generated sample is constructed across d features. The remaining reliability, difficulty, and priority calculations are linear in the number of minority instances, O(n_min), and are therefore smaller than the neighborhood-search cost.

In practical terms, NADOS has a similar computational profile to SMOTE-family methods that also depend on nearest-neighbor search. Its extra scoring stage adds overhead, but it does not require repeated classifier fitting, iterative optimization, or ensemble training inside the resampling process. For small and medium tabular benchmarks, this cost is acceptable. For very large or high-dimensional datasets, approximate nearest-neighbor search, feature normalization, dimensionality reduction, or batch-wise implementation may be needed.

G. Discussion

The results also identify boundary conditions where NADOS should not be described as universally superior. SVM-SMOTE is the closest competitor, obtains a slightly higher AUPRC and balanced accuracy in the aggregate table, and is not significantly different from NADOS on F1 after Holm correction. This suggests that margin-guided boundary synthesis can be highly competitive when the classifier score ranking is important. NADOS is therefore best characterized as having a strong and balanced performance profile under the evaluated noisy imbalanced benchmark, rather than as dominating every metric or dataset condition.

The results support the central methodological argument: separating seed reliability from synthesis difficulty is beneficial in noisy imbalanced classification. NADOS does not simply assign more synthetic samples to difficult regions. Instead, it first examines whether a minority instance is trustworthy enough to guide generation. This distinction helps avoid the main weakness of difficulty-only adaptive oversampling, where noisy or unstable points may receive excessive synthetic emphasis.

The method is not uniformly superior on every metric. SVM-SMOTE remains highly competitive and obtains stronger AUPRC. This limitation should be acknowledged directly. The value of NADOS lies in its balanced performance profile, explicit noise-aware decision logic, and strong average ranking on F1, G-mean, and balanced accuracy.

The results support the main hypothesis: a useful oversampling method for noisy imbalanced data should not treat all difficult minority samples as equally valuable. The reliability gate and difficulty score serve different purposes. The gate reduces generation from unstable seeds, while the difficulty score preserves attention to hard but useful samples. This division is the main conceptual contribution of NADOS and explains why it is competitive with both classical SMOTE-family methods and a recent noise-aware baseline.

VI. THREATS TO VALIDITY

This study uses controlled symmetric label noise, which supports reproducibility but does not capture all forms of real-world label corruption. The empirical claims should therefore be scoped to the symmetric training-label noise setting evaluated here. In practice, label noise may be class-dependent, feature-dependent, annotator-dependent, or concentrated in specific subregions of the feature space. Future work should test NADOS under class-dependent and feature-dependent corruption models before generalizing the conclusions to all noisy-data settings.

The benchmark is limited to binary tabular imbalanced datasets. This is appropriate for the first evaluation of NADOS, but conclusions should not be generalized automatically to multiclass, image, text, temporal, or streaming data. The method depends on k-nearest-neighbor evidence, so its behavior can be affected by the distance metric, feature scaling, mixed numerical-categorical attributes, missing values, and high-dimensional distance concentration. In the present implementation, continuous features are normalized within the training pipeline, and Euclidean neighborhoods are used; alternative metrics, metric learning, or categorical-aware distance functions may be needed for heterogeneous datasets.

The parameter settings are fixed for the main evaluation. This supports reproducibility and fair comparison, but it may not be optimal for every dataset. Adaptive parameter selection, threshold tuning, and alternative priority functions are left for future work. Finally, the implementation evaluates NADOS against a focused set of reproducible baselines. Additional recent oversampling methods may provide further comparison

opportunities if stable implementations and consistent benchmark protocols are available.

VII. Conclusion

This study proposed NADOS, a reliability-gated and difficulty-aware oversampling method for noisy imbalanced classification. The method separates minority-seed trustworthiness from synthesis priority by using reliability as an eligibility gate and difficulty as an allocation signal among eligible seeds. Evaluation across 22 benchmark datasets, five label-noise levels, four classifier families, and eight oversampling baselines shows that NADOS achieves the best average F1, G-mean, and balanced accuracy while remaining competitive on AUPRC. These findings support reliability-gated adaptive oversampling as a practical strategy for noisy imbalanced learning.

Future work should extend NADOS to adaptive thresholding, broader external datasets, multiclass settings, and more detailed component-level ablation. It should also examine alternative priority-combination rules, including additive weighting, soft-gated products, probabilistic seed selection, and learned weighting functions, to determine whether adaptive priority functions improve performance under different noise regimes, imbalance ratios, and feature-space conditions.

VIII. Declarations

A. Declaration on Generative AI

The authors used generative AI-assisted tools for language refinement and editorial checking. All scientific ideas, experimental design, implementation, results, interpretations, and final manuscript content were reviewed and verified by the authors.

B. Data Availability

The datasets used in this study are publicly available benchmark imbalanced classification datasets in KEEL-style format and related public tabular repositories [18, 19]. The dataset names and benchmark characteristics are summarized in the manuscript.

C. Code Availability

The implementation code, benchmark scripts, parameter settings, preprocessing pipeline, and random seeds are retained by the authors and may be made available upon reasonable request, subject to thesis completion, institutional requirements, and repository preparation.

D. Funding

This work was supported/funded by the Universiti Teknologi Malaysia under UTM NEXUS Translasi (PY/2025/00988).

E. Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009;21(9):1263-1284. https://doi.org/10.1109/TKDE.2008.239

[2] Krawczyk B. Learning from imbalanced data: open challenges and future directions. Prog Artif Intell. 2016;5:221-232. https://doi.org/10.1007/s13748-016-0094-0

[3] Japkowicz N, Stephen S. The class imbalance problem: a systematic study. Intell Data Anal. 2002;6(5):429-449. https://doi.org/10.3233/IDA-2002-6504

[4] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321-357. https://doi.org/10.1613/jair.953

[5] Fernandez A, Garcia S, Herrera F, Chawla NV. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. J Artif Intell Res. 2018;61:863-905. https://doi.org/10.1613/jair.1.11192

[6] Saez JA, Luengo J, Stefanowski J, Herrera F. SMOTE-IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. Inf Sci. 2015;291:184-203. https://doi.org/10.1016/j.ins.2014.08.051

[7] He H, Bai Y, Garcia EA, Li S. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks. IEEE; 2008. pp. 1322-1328. https://doi.org/10.1109/IJCNN.2008.4633969

[8] Li M, Zhou H, Liu Q, Gong X, Wang G. WRND: a weighted oversampling framework with relative neighborhood density for imbalanced noisy classification. Expert Syst Appl. 2024;241:122593. https://doi.org/10.1016/j.eswa.2023.122593

[9] Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C. Safe-Level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In: Advances in Knowledge Discovery and Data Mining. Berlin: Springer; 2009. pp. 475-482. https://doi.org/10.1007/978-3-642-01307-2_43

[10] Barua S, Islam MM, Yao X, Murase K. MWMOTE: Majority weighted minority oversampling technique for imbalanced data set learning. IEEE Trans Knowl Data Eng. 2014;26(2):405-425. https://doi.org/10.1109/TKDE.2012.232

[11] Batista GEAPA, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explor Newsl. 2004;6(1):20-29. https://doi.org/10.1145/1007730.1007735

[12] Han H, Wang WY, Mao BH. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: Advances in Intelligent Computing. Lecture Notes in Computer Science, vol 3644. Berlin: Springer; 2005. pp. 878-887. https://doi.org/10.1007/11538059_91

[13] Nguyen HM, Cooper EW, Kamei K. Borderline over-sampling for imbalanced data classification. Int J Knowl Eng Soft Data Paradigms. 2011;3(1):4-21. https://doi.org/10.1504/IJKESDP.2011.039875

[14] Tomek I. Two modifications of CNN. IEEE Trans Syst Man Cybern. 1976;SMC-6(11):769-772. https://doi.org/10.1109/TSMC.1976.4309452

[15] Wilson DL. Asymptotic properties of nearest neighbor rules using edited data. IEEE Trans Syst Man Cybern. 1972;SMC-2(3):408-421. https://doi.org/10.1109/TSMC.1972.4309137

[16] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825-2830.

[17] Lemaitre G, Nogueira F, Aridas CK. Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res. 2017;18(17):1-5.

[18] Dua D, Graff C. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science; 2019.

[19] Alcala-Fdez J, Fernandez A, Luengo J, Derrac J, Garcia S, Sanchez L, Herrera F. KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. J Mult-Valued Logic Soft Comput. 2011;17(2-3):255-287.

[20] Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning. ACM; 2006. pp. 233-240. https://doi.org/10.1145/1143844.1143874

[21] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432. https://doi.org/10.1371/journal.pone.0118432

[22] Demsar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res. 2006;7:1-30.

[23] Wilcoxon F. Individual comparisons by ranking methods. Biom Bull. 1945;1(6):80-83. https://doi.org/10.2307/3001968

[24] Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat. 1979;6(2):65-70.

A RAG-Enhanced Human-in-the-Loop Framework for Automated Assessment of Engineering Laboratory Reports

Amina Abbi[1]*, Mohammed Skouri[2], Mustapha Raoufi[3]
Laboratory of Physics High Energy and Astrophysics, Faculty of Science Semlalia, Cadi Ayyad University, Marrakech, Morocco[1, 2]
Laboratory of Materials Energy and Environment, Faculty of Science Semlalia, Cadi Ayyad University, Marrakech, Morocco[3]

***Abstract*—Recent developments in Large Language Models (LLMs) have created new opportunities to automate educational assessment and reduce workload for instructors. However, concerns regarding grading consistency, transparency, and pedagogical reliability continue to limit the adoption of fully automated assessment systems. In this study, we propose a Human-in-the-Loop framework for the automated evaluation of engineering laboratory reports based on Retrieval-Augmented Generation (RAG). The proposed framework integrates text extraction, structured content extraction, contextual retrieval from grading rubrics and laboratory resources, rubric-based evaluation, automatic feedback generation, and instructor validation into a unified grading workflow. The RAG module retrieves contextual information so that the language model can generate assessments that align with the course goals and are based on educational information about the specific assignments, thus ensuring consistent evaluation based on the rubric. The system was tested using a set of 56 laboratory reports collected from undergraduate courses in Electrical and Electronics Engineering. The experimental results indicate a strong agreement between the grades provided by the AI and the instructor, with a Pearson Correlation Coefficient of 0.988, a Mean Absolute Error (MAE) of 3.55, and a Root Mean Squared Error (RMSE) of 3.77. Besides, 89.29% of the reports were scored within ±5 points of the instructor scores. The grading time was reduced from 392 minutes to 84 minutes, a workload reduction of 78.57%. The results demonstrate that the integration of Retrieval-Augmented Generation, rubric-based evaluation, and Human-in-the-Loop validation constitutes an effective approach for AI-supported assessment of engineering laboratory reports, maintaining instructor oversight and educational integrity.**

Keywords*—Large Language Models; Retrieval-Augmented Generation; Human-in-the-Loop; automated assessment; engineering education; laboratory report evaluation; rubric-based grading; educational artificial intelligence

I. Introduction

Artificial intelligence (AI) is rapidly transforming educational technologies by enabling intelligent tutoring, personalized learning, automated assessment, feedback generation, and AI-assisted learning environments [1]– [4], [11]. Recent studies have further highlighted the growing role of AI in educational assessment, personalized feedback, and technology-enhanced learning [10], [18], [23].

Large Language Models (LLMs), one of the most significant recent advances in AI, have demonstrated remarkable capabilities in natural language understanding and generation, making them promising tools for educational assessment [1], [7], [9], [21]. Recent studies have shown that LLMs can effectively assist instructors in grading essays, short-answer questions, programming assignments, engineering assignments, and open-ended examinations while improving grading consistency and reducing instructor workload [8], [13], [16], [17], [19].

One of the hardest types of assessment is the assessment of educational assessment in engineering laboratory reports. They assess theoretical understanding, experimental procedures, technical calculations, data interpretation, engineering reasoning, and scientific communication, as opposed to the typical written assignment. The manual grading process is thus time-consuming and may suffer from inconsistency due to instructor workload and subjective judgment.

Prior research has shown the promise of LLM-based grading systems, but most methods evaluate student submissions individually without taking into account assignment-specific educational resources such as lab instructions, grading rubrics, reference solutions, and course materials. Thus, the quality of grading may be undermined by an incomplete understanding of the context and hallucinated answers. Furthermore, fully automated assessment systems are typically opaque and lack human oversight, limiting their adoption in higher education [9], [21].

Recently, a promising approach to enhance the reliability of LLM-based systems is to augment prompts with relevant external knowledge retrieved at inference time, which is known as Retrieval-Augmented Generation (RAG) [20]. In educational assessment, RAG allows language models to assess student work with resources specific to the assignment instead of relying solely on pre-trained knowledge. Recent studies have also shown the usage of context-aware assessment and context engineering to improve the reliability of grading and quality of feedback [14], [20].

However, the automated assessment of engineering laboratory reports has been relatively less explored, with an integrated framework that combines contextual retrieval, rubric-based evaluation, automated feedback generation, and Human-

*Corresponding author

in-the-Loop validation. Previous work has largely focused on essays, math assessments, programming assignments, or short-answer questions, but engineering lab reports pose additional challenges due to the need for experimental analysis and technical reasoning.

To address this gap, this study proposes an RAG-enhanced Human-in-the-Loop framework for automated assessment of engineering laboratory reports using Large Language Models. The proposed framework combines document preprocessing, structured content extraction, Retrieval-Augmented Generation, rubric-based grading, automated feedback generation, and instructor validation into a single assessment workflow. The framework was evaluated on 56 laboratory reports of the Electrical and Electronics Engineering domain. The experimental results demonstrate a high level of agreement in the grades assigned by the AI and the instructor, with a Pearson Correlation Coefficient of 0.988, a Mean Absolute Error (MAE) of 3.55 points, and an Agreement Rate of 89.29%, reducing grading time by 78.57%.

The major contributions of this work are summarized in Table I.

TABLE I. SUMMARY OF THE MAIN CONTRIBUTIONS

Contribution	Description
C1	A RAG-enhanced Human-in-the-Loop framework for automated assessment of engineering laboratory reports.
C2	Integration of contextual retrieval, rubric-based grading, automated feedback generation, and instructor validation into a unified assessment workflow.
C3	Experimental validation with 56 real lab reports from Electrical & Electronics Engineering.
C4	Comprehensive evaluation with Pearson Correlation, MAE, RMSE, Agreement Rate, Maximum Error, and grading time reduction.

The rest of this study is organized as follows: Related work on AI-assisted educational assessment is presented in Section II. The proposed methodology is presented in Section III. Section IV describes the experiment setup. Section V presents and discusses the experimental results, including the limitations of the study. Finally, Section VI concludes the study and offers future research directions.

II. RELATED WORK

A. Education Assessment Using Large Language Models

Recently, Large Language Models (LLMs) have become a major research topic in educational technology, as they are capable of supporting intelligent tutoring, personalized learning, automated assessment, and feedback generation [1], [4], [11]. Recent surveys [1], [9], [11] show that automated assessment is one of the most promising applications of LLMs in education. The effectiveness of LLMs for grading short-answer questions [16], essays [8], programming assignments [17], engineering mechanics assignments [19], undergraduate calculus [5], and open-ended examinations [13] has been demonstrated in several studies. These studies consistently show high agreement between grades assigned by AI and instructors, and significantly reduce instructor workload.

But these encouraging results also highlight problems of grading consistency, hallucinations, transparency, and pedagogical reliability in prior research. Thus, current evidence points to LLMs supporting instructors rather than replacing human judgment [9], [21].

B. Human-in-the-Loop Evaluation and AI-Generated Feedback

Recent studies have increasingly advocated for Human-in-the-Loop (HITL) grading frameworks [7], [8] to improve the reliability and transparency of AI-assisted assessment. These methods combine automated assessment and instructor review, enabling instructors to verify AI-generated grades before they are posted as final results.

Moreover, several studies have shown that LLM-generated feedback can enhance student learning by offering personalized suggestions and constructive explanations [6], [15], [22], [24]. Moreover, AI-enabled pedagogical architectures have shown promising results in supporting self-regulated learning and enhancing student engagement [3]. The results suggest that a good assessment system should involve reliable grading, useful feedback, and instructor oversight.

C. Retrieval-Enhanced Generation for Education Assessment

Retrieval-Augmented Generation (RAG) has emerged as a powerful approach for improving the factuality and context-awareness of LLMs by retrieving relevant external knowledge before response generation [20]. In educational assessment, RAG allows language models to leverage lab instructions, grading rubrics, reference solutions, and course materials, which helps mitigate hallucinations and improves consistency in grading.

Context engineering has recently shown that the integration of contextual educational resources enhances the quality of AI-based assessment [14]. At the same time, research on automated assessment in the domain of STEM education emphasizes the increasing importance of domain-specific contextual information for reliable assessment [12].

D. Research Gap and Contributions

Despite many advances in AI-based assessment on educational resources, existing studies mainly focus on essays, programming assignments, mathematics or short-answer questions [5], [8], [13], [16], [17]. Engineering laboratory reports, which require evaluation of experimental procedures, technical calculations, engineering reasoning, and scientific communication, have been relatively neglected.

Additionally, most existing approaches rely solely on LLM reasoning without integrating contextual retrieval, rubric-based evaluation, automated feedback generation, and Human-in-the-Loop validation in a unified framework.

To overcome these limitations, this study proposes a RAG-enhanced Human-in-the-Loop framework for automated assessment of engineering laboratory reports using Large Language Models. The proposed framework combines Retrieval-Augmented Generation, structured content extraction, rubric-based grading, automated feedback generation, and instructor validation into one assessment workflow.

The framework is experimentally validated using 56 real Electrical and Electronics Engineering laboratory reports and evaluated using various quantitative metrics such as Pearson Correlation, MAE, RMSE, Agreement Rate, Maximum Error, and grading time reduction.

III. Methodology

A. Overview of the Proposed Framework

The proposed framework presents a Human-in-the-Loop framework for automated assessment of engineering laboratory reports enhanced by Retrieval-Augmented Generation (RAG). Unlike traditional LLM-based automated grading systems that only use student submissions, it enhances the grading process by accessing educational resources specific to assignments, such as laboratory instructions, grading rubrics, reference solutions, and course materials. This contextual information allows for a more consistent, transparent, and reliable assessment based on the intended learning outcomes.

This framework combines an end-to-end assessment workflow that includes document pre-processing, extraction of structured content, context retrieval, rubric-based assessment, automated feedback generation, support for plagiarism and validation by instructors. AI-generated grades and feedback are preliminary recommendations. Instructors maintain complete discretion to review, revise, approve, or reject the proposed assessment before the final grade is assigned, thereby ensuring pedagogical accountability and reducing automation bias.

The assessment workflow consists of eight sequential stages: 1) laboratory report submission, 2) document preprocessing and text extraction, 3) structured content extraction, 4) Retrieval-Augmented Generation, 5) LLM-based rubric evaluation, 6) automated formative feedback generation, 7) Human-in-the-Loop instructor validation, and 8) final grade publication. Fig. 1 illustrates the overall architecture of the proposed framework.

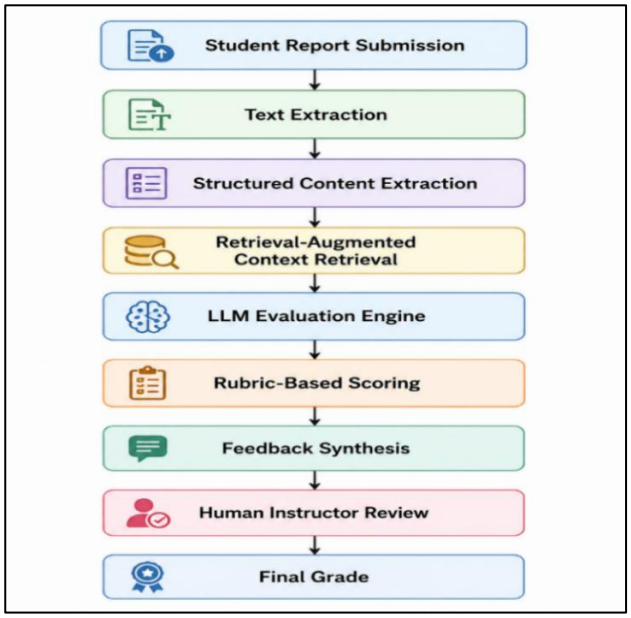


Fig. 1. Overall architecture of the proposed RAG-enhanced Human-in-the-Loop assessment framework.

B. Report Processing and Context Retrieval

Laboratory reports submitted in PDF or DOCX format are first processed to extract machine-readable content. When scanned documents are encountered, Optical Character Recognition (OCR) is applied to recover textual information. The extracted report content is subsequently prepared for automated assessment while preserving the logical organization of the laboratory report. This process provides a consistent textual representation that supports contextual retrieval and rubric-based evaluation.

The proposed framework primarily evaluates the textual content extracted from laboratory reports. Tables and textual mathematical expressions are processed through the document extraction stage and incorporated into the grading workflow whenever they are represented in machine-readable form. For graphical laboratory artifacts such as circuit diagrams, experimental plots, screenshots, and handwritten annotations, the current framework extracts only the embedded textual information when available and does not directly interpret the visual content. Consequently, grading decisions are primarily based on the extracted textual information together with the retrieved contextual resources.

To improve grading reliability and contextual understanding, the proposed framework incorporates a Retrieval-Augmented Generation (RAG) module prior to the evaluation stage. Instead of relying solely on the submitted report and the intrinsic knowledge of the Large Language Model (LLM), the framework retrieves assignment-specific educational resources that enrich the grading prompt with relevant contextual information. The knowledge base contains laboratory instructions, grading rubrics, reference solutions, course materials, and assignment metadata, enabling assessments that remain aligned with course objectives, intended learning outcomes, and predefined grading criteria while helping reduce hallucinations and improving grading consistency.

The RAG module is implemented using OpenAI's text-embedding-3-small model to generate 1536-dimensional semantic embeddings, which are indexed in a Pinecone vector database using cosine similarity. Educational resources are embedded and indexed together with their associated metadata to preserve instructional context during retrieval. For each submitted laboratory report, the retrieval query is automatically constructed from the assignment title, assignment description, and corresponding course chapter. Metadata filtering is subsequently applied to retrieve only assignment-relevant educational resources, after which the five most relevant documents are incorporated into the evaluation prompt to provide assignment-specific context for the language model.

Unlike conventional LLM-based grading approaches that evaluate reports using only the submitted content, the proposed framework combines the retrieved educational resources with grading rubrics, expected solutions, laboratory instructions, and course materials to support context-aware assessment. Consequently, laboratory reports are graded based on the desired learning outcomes and rubric criteria, not on exact textual similarity. Thus, alternative engineering approaches, experimental variations that are within an acceptable range and

alternative approaches to solutions can be fairly evaluated as long as they meet the desired learning objectives.

The assessment process combines the extracted report content with the retrieved contextual resources and evaluates all the criteria defined in the grading rubric. This allows for thorough, consistent, and pedagogically sound assessment that is aligned with the learning outcomes and instructor-defined grading standards. Fig. 2 illustrates the Retrieval-Augmented Generation process used in the proposed framework.

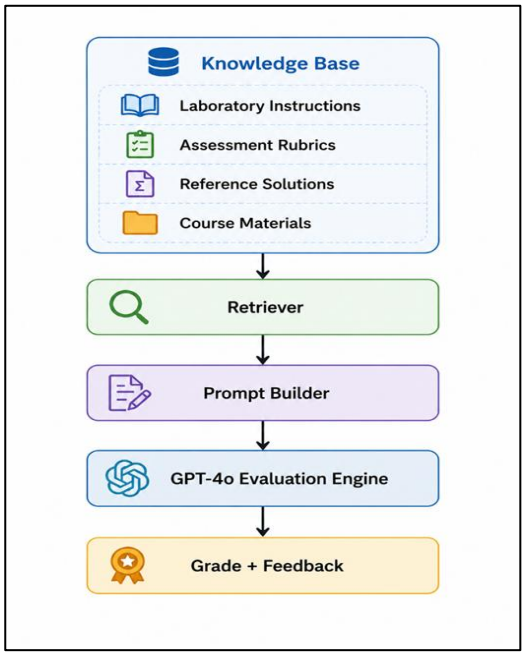


Fig. 2. Retrieval-Augmented Generation workflow for context-aware laboratory report assessment.

C. Prompt Design and LLM-Based Evaluation

The proposed framework employs a structured prompt engineering strategy to ensure consistent, transparent, and rubric-aligned evaluation of engineering laboratory reports. Instead of using the submitted report directly, the evaluation prompt combines assignment metadata, laboratory instructions, retrieved contextual resources, grading rubrics, expected reference solutions, and student submission into a unified evaluation context. This comprehensive prompt provides the language model with assignment-specific information required to perform context-aware, rubric-aligned assessment while maintaining consistency with the intended learning outcomes.

The evaluation prompt is constructed automatically by combining the assignment title, assignment description, laboratory instructions, maximum score, grading rubric, retrieved educational resources, expected solutions, and extracted laboratory report content. The language model will be responsible for evaluating the submission based on the grading rubric defined and producing a numeric score, along with detailed formative feedback, identified strengths, weaknesses, and concrete recommendations for improvement. Such a structured prompt design facilitates grading consistency and subsequent instructor review and validation. An excerpt of the structured evaluation prompt used by the proposed framework is presented below.

Excerpt of the Evaluation Prompt Used by the Proposed Framework
System Role
You are a university professor evaluating a student's assignment submission.
Assignment Details: • Title • Description • Instructions • Maximum Score
Grading Rubric: ...
Expected Answers / Reference Solutions: ...
Course Context and Reference Materials: ...
Student Submission: ...
Carefully evaluate this submission and provide: 1. Numerical Grade 2. Overall Feedback 3. Strengths 4. Weaknesses 5. Suggestions for Improvement

The assessment stage employs GPT-4o as the primary grading model because of its strong reasoning capabilities and its ability to evaluate engineering laboratory reports containing theoretical explanations, experimental procedures, calculations, experimental analysis, and technical discussions. A low temperature setting (0.3) was chosen during grading to enhance the stability of the response and to reduce variability between repeated evaluations of similar reports.

To improve computational efficiency while maintaining high-quality formative feedback, GPT-4o-mini is employed exclusively for generating detailed feedback after the numerical grade has been determined. The generated feedback is consistent with the assigned score, as the grading decision is made before the feedback generation, and reduces the computational cost. An example of the formative feedback generated by the proposed framework is presented in Fig. 3.

Fig. 3. Example of AI-generated formative feedback.

D. Rubric-Based Assessment

The proposed framework employs a criterion-referenced rubric to ensure transparent, consistent, and pedagogically aligned assessment of engineering laboratory reports. The grading rubric consists of three complementary evaluation criteria, weighted as follows: Technical Accuracy (40%), Experimental Analysis (30%), and Completeness and Presentation (30%).

Technical Accuracy receives the highest weighting because the primary objective of engineering laboratory assessment is to evaluate students' understanding and correct application of engineering concepts, calculations, and technical reasoning. Experimental Analysis emphasizes students' ability to interpret experimental results, discuss observations, and relate practical findings to theoretical concepts. Completeness and Presentation assess the organization, clarity, and professional quality of the laboratory report, reflecting the importance of technical communication in engineering education.

During assessment, the language model considers each criterion by referring to the retrieved contextual resources, grading rubric, expected reference solutions, and the assignment-specific laboratory instructions. The weighted criterion scores are subsequently combined to produce the final assessment while maintaining consistency with the predefined learning outcomes and grading standards.

E. Human-in-the-Loop Validation

The proposed framework adopts a Human-in-the-Loop (HITL) validation strategy to ensure that the AI-assisted assessment is under the supervision of the instructor. The framework offers preliminary recommendations for the AI-generated grade and feedback to assist instructors in the grading process, rather than automatically assigning a final grade.

Following the initial evaluation, instructors review the proposed grade, generated feedback, and the corresponding laboratory report in relation to the grading rubric. They may approve the assessment without modification or revise the proposed score and feedback whenever discrepancies or pedagogical considerations are identified. Consequently, instructors retain full authority over the final grading decision while accommodating valid experimental variations and alternative solution approaches.

By requiring instructor verification before grade publication, the proposed framework mitigates automation bias and combines the efficiency of AI-assisted assessment with the professional judgment and pedagogical expertise of the instructor.

F. Academic Integrity Support

In addition to automatic assessment, the proposed framework has an academic integrity support module to assist instructors in detecting potential plagiarism prior to grading. The module does not make the determination of whether academic misconduct has occurred, but provides similarity analysis to assist instructor judgment in the assessment process.

The academic integrity module is integrated in the proposed framework but is not quantitatively evaluated in this work, as the goal is to validate the performance of the RAG-enhanced Human-in-the-Loop assessment framework.

IV. Experimental Setup

A. Dataset

The proposed framework was evaluated using 56 laboratory reports collected from undergraduate Electrical and Electronics Engineering (EEE) laboratory courses at a single higher education institution. As summarized in Table II, the dataset comprises four laboratory experiments covering analog electronics and circuit analysis. Three experiments were completed as group assignments, whereas the fourth required individual report submission.

The laboratory reports include theoretical explanations, experimental procedures, calculations, observations, experimental analysis, tables, and conclusions, providing representative examples of undergraduate engineering laboratory documentation. These reports were utilized to assess the accuracy, consistency, and efficiency of grading of the proposed RAG-enhanced Human-in-the-Loop assessment framework.

TABLE II. Dataset Composition

Laboratory	Topic	Assignment Type	Reports
TP1	Diodes	Group	8
TP2	Common Emitter Amplifier	Group	8
TP3	Operational Amplifier Applications	Group	8
TP4	Capacitor Charging and Discharging	Individual	32
Total			56

B. Experimental Approach

Instructor-assigned grades (on a 100-point scale) were used as the reference standard for evaluating the proposed framework. Each laboratory report was first graded manually by the course instructor according to the predefined assessment rubric. The same reports were subsequently evaluated using the proposed RAG-enhanced Human-in-the-Loop framework following the workflow illustrated in Fig. 1.

For each submission, the framework extracted the report content, retrieved assignment-specific contextual resources via the RAG module, and generated an initial grade and formative

feedback. Next, the AI-generated grades were compared to the instructor-assigned grades to evaluate grading accuracy, consistency, and operational efficiency. During the experimental evaluation, instructor grades served exclusively as the reference for quantitative analysis. Although the proposed framework incorporates Human-in-the-Loop validation, instructor intervention was intentionally omitted during performance evaluation so that AI-generated grades could be compared directly with the reference grades under identical conditions. Instructors have full authority to review, modify, approve, or reject AI-generated assessments before the final grade is released in operational use.

C. Evaluation Metrics

The performance of the proposed framework was evaluated using six complementary metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Maximum Error, Agreement Rate (±5 points), Pearson Correlation Coefficient, and Grading Time Reduction. These metrics were selected to assess grading accuracy, agreement with instructor-assigned grades, and operational efficiency.

Pearson Correlation Coefficient was used to measure the linear agreement between AI-generated and instructor-assigned grades, while MAE, RMSE, and Maximum Error quantified grading deviations. The Agreement Rate represents the percentage of laboratory reports for which the AI-generated grade differed by no more than ±5 points from the instructor-assigned grade. Grading Time Reduction was used to evaluate the efficiency gained through AI-assisted assessment.

To further assess the robustness of the proposed framework, statistical significance testing, p-values, and 95% confidence intervals were computed for the primary evaluation metrics. These complementary statistical measures provide additional evidence regarding the reliability and consistency of the proposed RAG-enhanced Human-in-the-Loop assessment framework.

D. Implementation Details

The proposed framework was implemented in Python using OpenAI GPT-4o as the primary model for rubric-based grading and GPT-4o-mini for formative feedback generation. Retrieval-Augmented Generation (RAG) was implemented using the OpenAI text-embedding-3-small embedding model, which generates 1536-dimensional semantic embeddings. These embeddings were indexed in a Pinecone vector database using cosine similarity to enable efficient semantic retrieval of assignment-specific educational resources.

Educational resources, including laboratory instructions, grading rubrics, reference solutions, and course materials, were indexed as complete semantic documents without additional text chunking because these resources are relatively short and self-contained. For each laboratory report, the retrieval query was automatically constructed from the assignment title, assignment description, and corresponding course chapter. Metadata filtering was subsequently applied to retrieve the five most relevant educational resources, which were incorporated into the evaluation prompt together with the grading rubric, expected reference solutions, laboratory instructions, and the extracted laboratory report content.

Table III summarizes the implementation parameters used during the experimental evaluation.

TABLE III. IMPLEMENTATION CONFIGURATION

Parameter	Value
Programming Language	Python
Primary Grading Model	GPT-4o
Feedback Model	GPT-4o-mini
Embedding Model	OpenAI text-embedding-3-small
Embedding Dimension	1536
Vector Database	Pinecone
Similarity Metric	Cosine Similarity
Indexed Documents	Complete Semantic Documents
Retrieved Documents	Top-5
Retrieval Filtering	Assignment/Chapter Metadata
Assessment Strategy	RAG-Enhanced Rubric-Based Evaluation
Human Validation	Human-in-the-Loop

V. RESULTS AND DISCUSSION

A. Overall Grading Performance

The proposed framework was evaluated using 56 undergraduate Electrical and Electronics Engineering laboratory reports. Table IV summarizes the quantitative performance of the proposed framework by comparing AI-generated grades with instructor-assigned grades on a 100-point grading scale.

TABLE IV. PERFORMANCE OF THE PROPOSED FRAMEWORK

Metric	Value
Number of Reports	56
Grading Scale	0–100 points
Mean Instructor Grade	83.64
Mean AI Grade	80.09
Mean Absolute Error (MAE)	3.55 points (95% CI: 3.23–3.89)
Root Mean Squared Error (RMSE)	3.77 points (95% CI: 3.43–4.11)
Maximum Error	6 points
Agreement Rate (±5 points)	89.29%
Pearson Correlation	0.988 (95% CI: 0.980–0.993, $p < 0.001$)
Manual Grading Time	392 min
AI-assisted Grading Time	84 min
Time Reduction	78.57%

The proposed framework achieved a Pearson Correlation Coefficient of 0.988 (95% CI: 0.980–0.993, $p < 0.001$), indicating a statistically significant and very strong agreement between AI-generated and instructor-assigned grades. The Mean Absolute Error (MAE) was 3.55 points (95% CI: 3.23–3.89), while the Root Mean Squared Error (RMSE) was 3.77 points (95% CI: 3.43–4.11). Furthermore, 89.29% of the laboratory reports received AI-generated grades within ±5

points of the instructor-assigned grades, demonstrating high grading consistency.

The mean instructor-assigned grade was 83.64 points, whereas the mean AI-generated grade was 80.09 points, indicating an average difference of 3.55 points. This result suggests a slight tendency of the AI to assign more conservative grades than the instructor. Nevertheless, the very high correlation coefficient, together with the narrow confidence intervals and relatively small grading errors, indicates that the proposed framework closely reproduces instructor grading decisions while maintaining consistent ranking of student performance across the evaluated reports.

B. Agreement with Human Evaluation

Fig. 4 illustrates the relationship between AI-generated grades and instructor-assigned grades for the 56 laboratory reports. Most data points are closely distributed around the line of perfect agreement, demonstrating strong consistency between the two grading approaches.

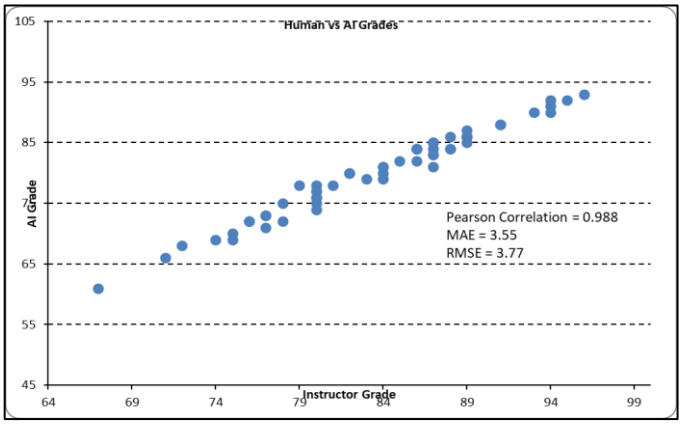


Fig. 4. Correlation between instructor-assigned grades and AI-generated grades.

Overall, 89.29% of the laboratory reports received AI-generated grades within ±5 points of the instructor-assigned grades, while the maximum observed difference was 6 points. The mean AI-generated grade (80.09) was slightly lower than the mean instructor-assigned grade (83.64), corresponding to an average difference of 3.55 points. This observation suggests a slight conservative grading bias of the proposed framework rather than inconsistent assessment. Nevertheless, the very high agreement between AI-generated and instructor-assigned grades indicates that this bias has only a limited impact on the overall grading consistency and reliability of the proposed framework.

C. Impact of Retrieval-Augmented Generation

The Retrieval-Augmented Generation (RAG) module retrieves lab instructions, grading rubrics, reference solutions, and course materials relevant to each lab report, providing assignment-specific contextual information throughout the assessment process. The contextual resources are provided to the evaluation prompt, allowing the language model to execute rubric-aligned and course-specific assessment consistent with the intended learning outcomes.

The high correlation between AI-generated and instructor-assigned grades shows the effectiveness of the proposed RAG-enhanced Human-in-the-Loop framework as an AI-assisted assessment system. Figure 5 shows an example of contextual information retrieved by the RAG module during the evaluation process.

D. Human-in-the-Loop Validation

The experimental results demonstrate that the proposed framework provides reliable preliminary assessments that can effectively support instructor decision-making. The high agreement between AI-generated and instructor-assigned grades indicates that the framework can substantially reduce instructor workload while maintaining grading consistency.

Within the proposed Human-in-the-Loop workflow, AI-generated grades and formative feedback are presented to the instructor as recommendations rather than final decisions. Instructors retain full authority to review, revise, approve, or reject the proposed assessment before the final grade is released. This validation process preserves pedagogical accountability while reducing the potential impact of automation bias in educational assessment.

E. Grading Efficiency

In addition to grading accuracy, the proposed framework substantially reduced the instructor's grading workload. The 56 laboratory reports were evaluated manually in 392 minutes, whereas the AI-assisted evaluation completed the same grading task in 84 minutes, which is a reduction in grading time of 78.57%.

The reported grading time includes only the assessment process itself, encompassing laboratory report evaluation, score generation, and feedback generation. Initial system development, RAG knowledge base construction, document indexing, and system configuration were performed offline and were therefore not included in the reported grading time. Fig. 5 compares the time required for manual and AI-assisted grading.

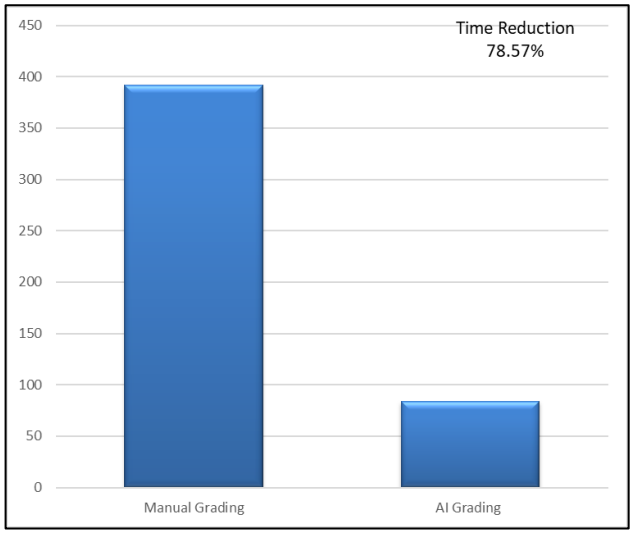


Fig. 5. Manual and AI-assisted grading time comparison.

F. Discussion

Experimental results indicate that the combination of Retrieval-Augmented Generation (RAG), rubric-based assessment, and Human-in-the-Loop validation provides an effective solution for the automated assessment of engineering laboratory reports. The proposed framework achieved strong agreement with instructor-assigned grades while substantially reducing grading time, highlighting its potential as an AI-assisted decision-support tool that improves instructor efficiency while preserving instructor oversight.

The reported results reflect the performance of the proposed framework as an integrated assessment system. Although the RAG module provides assignment-specific contextual grounding during assessment, the present study did not include a quantitative comparison between the proposed RAG-enhanced framework and an equivalent LLM-based grading approach without Retrieval-Augmented Generation. Consequently, the independent contribution of the RAG module cannot be isolated from the overall framework performance.

Furthermore, the experimental evaluation was conducted on 56 laboratory reports collected from undergraduate Electrical and Electronics Engineering courses at a single higher education institution. Therefore, the reported results should be considered as an initial validation of the proposed framework. Future work will investigate the proposed framework using larger, multi-institutional, and multidisciplinary datasets, perform controlled comparisons between RAG-enhanced and non-RAG grading approaches, evaluate the effectiveness of the academic integrity support module, and explore the integration of multimodal vision-language models to enhance the interpretation of graphical laboratory artifacts, including circuit diagrams, experimental plots, screenshots, and handwritten annotations.

VI. Conclusion and Future Work

This study presented a Retrieval-Augmented Generation (RAG)-enhanced Human-in-the-Loop framework for the automated assessment of engineering laboratory reports using Large Language Models. By integrating structured content extraction, contextual retrieval, rubric-based grading, automated formative feedback generation, academic integrity support, and instructor validation into a unified workflow, the proposed framework provides a comprehensive AI-assisted assessment solution for engineering education.

Experimental evaluation on 56 undergraduate Electrical and Electronics Engineering laboratory reports demonstrated that the proposed framework can provide accurate, consistent, and efficient AI-assisted assessment while preserving instructor oversight. The framework achieved a Pearson Correlation Coefficient of 0.988 ($p < 0.001$), a Mean Absolute Error (MAE) of 3.55 points, an Agreement Rate of 89.29%, and a grading time reduction of 78.57%, demonstrating its potential to improve assessment efficiency while maintaining grading quality and supporting reliable formative assessment in engineering education.

Future work will focus on validating the proposed framework using larger multi-institutional and multidisciplinary datasets, conducting controlled comparisons with equivalent LLM-based grading approaches without Retrieval-Augmented Generation, evaluating the effectiveness of the academic integrity support module, and investigating the integration of multimodal vision-language models to further enhance the assessment of graphical laboratory artifacts.

References

[1] S. Wang, T. Xu, H. Li, C. Zhang, J. Liang, J. Tang, P. S. Yu, and Q. Wen, "Large language models for education: A survey and outlook," IEEE Signal Processing Magazine, vol. 42, no. 6, pp. 51–63, Nov. 2025, doi: 10.1109/MSP.2025.3594309.

[2] N. Harsha Vardhan Reddy, N. Thirumal Reddy, M. Bharath, N. Hemanth, D. P. Dharnasi, B. Nirmala, and A. Jitendra, "AI based learning assistant using machine learning," International Journal of Engineering & Extended Technologies Research, vol. 8, no. 2, Mar.–Apr. 2026, doi: 10.15662/IJEETR.2026.0802003.

[3] G. D. O. Lima, J. A. R. Costa, F. A. Dorça, and R. D. Araújo, "An AI-supported pedagogical architecture to foster self-regulated learning in virtual environments," Computer Applications in Engineering Education, vol. 34, no. 1, Nov. 2025, doi: 10.1002/cae.70118.

[4] H. Li, L. Xiao, K. Lu, D. Li, Z. Zhang, and Q. Xia, "Exploring the application of large language models (LLMs) in data structure instruction: An empirical analysis of student learning outcomes in computer science," Information, vol. 17, no. 4, p. 353, 2026, doi: 10.3390/info17040353.

[5] S. Dakshit and S. S. Roy, "Framework for evaluating LLM performance in undergraduate calculus," Informatics, vol. 13, no. 6, p. 82, Jun. 2026, doi: 10.3390/informatics13060082.

[6] F. A. Zampirolli, F. Teubl, P. H. Pisani, and T. A. P. e Silva, "Intelligent feedback for individualized introductory programming exercises," Computer Applications in Engineering Education, vol. 34, no. 1, Dec. 2025, doi: 10.1002/cae.70132.

[7] A. Vanhoyweghen, V. Holst, M. Mobini, L. Van de Voorde, T. Vanleke, B. Verbruggen, B. Verbeken, A. Algaba, S. Verboven, M.-A. Guerry, F. Van Droogenbroeck, and V. Ginis, "Human-in-the-loop LLM grading for handwritten mathematics assessments," arXiv preprint arXiv:2603.13083, Mar. 2026.

[8] C. Zhang, K. P. Ju, X. Lu, Y.-C. G. Yen, and J. M. Rzeszotarski, "EvaluAId: Human-AI collaborative evaluation of open-ended student essays," in Proc. CHI Conf. Human Factors in Computing Systems (CHI '26), 2026, Art. no. 867, pp. 1–20, doi: 10.1145/3772318.3790814.

[9] M. Jukiewicz and M. Wyrwa, "Can ChatGPT replace the teacher in assessment? A review of research on the use of large language models in grading and providing feedback," Applied Sciences, vol. 16, no. 2, p. 680, 2026, doi: 10.3390/app16020680.

[10] O. Niyibizi, "Impact of large language models on personalized learning, assessment automation, and student outcomes in higher learning institution," Journal of Technology-Assisted Learning, vol. 2, no. 1, pp. 47–61, Mar. 2026, doi: 10.70232/jtal.v2i1.22.

[11] S. Filippi and B. Motyl, "Large language models (LLMs) in engineering education: A systematic review and suggestions for practical adoption," Information, vol. 15, no. 6, p. 345, 2024, doi: 10.3390/info15060345.

[12] M. O. Omopekunola, "Large language model-based automated item generation in STEM assessments: Historical mapping and a scoping review of empirical studies," Journal of Educational Technology Development and Exchange, vol. 19, no. 2, pp. 141–165, Jun. 2026, doi: 10.18785/jetde.1902.07.

[13] J. S. Jauhiainen and A. Garagorry Guerra, "Evaluating open-ended high-stakes examinations with LLMs: Alignment between ChatGPT-4o and human grading in high and low-resource languages," Frontiers in Digital Education, vol. 3, no. 2, p. 17, 2026, doi: 10.1007/s44366-026-0091-1.

[14] Z. Tian, A. Liu, L. Esbenshade, M. Xiao, Z. Zhang, Y. Lápicus, T. Han, K. He, and M. Sun, "Creating and evaluating K-12 GenAI assessment graders through context engineering," arXiv preprint arXiv:2606.12422, 2026, doi: 10.48550/arXiv.2606.12422.

[15] N. Abbas and E. Atwell, "Cognitive computing with large language models for student assessment feedback," Big Data and Cognitive Computing, vol. 9, no. 5, p. 112, 2025, doi: 10.3390/bdcc9050112.

[16] O. Henkel, A. Boxer, L. Hills, B. Roberts, and Z. Levonian, "Can large language models make the grade? An empirical study evaluating LLMs'

ability to mark short answer questions in K-12 education," in Proc. 11th ACM Conf. Learning @ Scale (L@S '24), 2024, pp. 300–304, doi: 10.1145/3657604.3664693.

[17] Q. Zhou, Z. Wang, Y. Li, and W. K. Chan, "Impacts of histories and models on LLM grading: A study in advanced software engineering courses," arXiv preprint arXiv:2606.08400, Jun. 2026, doi: 10.48550/arXiv.2606.08400.

[18] K. Parmar, D. Palaniappan, T. Premavathi, and R. Jain, "Revolutionizing educational assessments with AI technology," in Rethinking the Pedagogy of Sustainable Development in the AI Era, pp. 227–252, 2025, doi: 10.4018/979-8-3693-9062-7.ch011.

[19] A. Mowafy, M. Talebi-Kalaleh, M. Sabek, H. Peng, M. Aqib, S. M. Adeeb, M. M. Elgammal, and C. Pettit, "Automated grading of engineering mechanics assignments using large language models and computer vision: A work in progress," in Proc. 2025 ASEE Annual Conference & Exposition, Montreal, Canada, Aug. 2025, doi: 10.18260/1-2--55492.

[20] Y. Huang and J. X. Huang, "A survey on retrieval-augmented text generation for large language models," ACM Computing Surveys, vol. 58, no. 12, Art. no. 300, pp. 1–38, 2026, doi: 10.1145/3805774.

[21] D. Deepshikha, "A systematic review on the future of educational assessment: AI-driven grading and personalised feedback in higher education," Artificial Intelligence in Education, vol. 2, no. 2, pp. 75–115, 2026, doi: 10.1108/AIIE-03-2025-0036.

[22] X. Li, D. Chen, D. Tuffley, G. Scott, G. Tuxworth, T. Yao, and G. Torrisi-Steele, "Evaluating AI-generated feedback for formative assessment in higher education," International Journal of AI in Pedagogy Innovation and Learning Futures, vol. 2026, no. 1, Mar. 2026, doi: 10.46787/ijaipil.v2026i1.6962.

[23] N. T. Mai, W. Cao, and Q. Fang, "A study on how LLMs (e.g. GPT-4, chatbots) are being integrated to support tutoring, essay feedback and content generation," Journal of Computing and Electronic Information Management, vol. 18, no. 3, 2025, doi: 10.54097/6r6yhn67.

[24] J. A. Khan, M. Yaqoob, M. Tasadduq, H. S. Dar, and A. Ahsan, "Personalized and constructive feedback for computer science students using the large language model (LLM)," arXiv preprint arXiv:2510.11556, Oct. 2025, doi: 10.48550/arXiv.2510.11556.

Detecting Metadata-Conditioned Response Drift in AI-Generated Visual Information Systems: A Two-Phase Human–Computer Interaction Framework

Yu Jia[1], Shaojie Zhang[2]*, Na Li[3]
College of Digital Creativity, Henan Technical College of Construction, Zhengzhou, Henan, China[1, 3]
College of Art and Design, Henan Institute of Economics and Trade, Zhengzhou, Henan, China[2]
Faculty of Education, Henan University, Kaifeng, Henan, China[1, 2]

***Abstract*—AI-generated visual information systems often present images with authorship metadata, yet such labels may shape the evaluation data they are intended to contextualize. Existing work has seldom offered a system-level procedure for measuring metadata-conditioned response drift under unchanged visual input. This study develops a two-phase evaluation framework combining blind presentation, attribution-label overlay, trial-level logging, drift computation, and mixed-effects validation. Ninety-two participants rated four AI-generated images first without labels and then with attribution labels, producing 732 valid observations. Linear mixed-effects modeling, supported by ordinal sensitivity analysis and leave-one-image-out checks, identified a significant Phase × Label interaction within the implemented attribution-label assignment scheme. Under this assignment scheme, ratings for stimuli assigned to the human-labeled condition increased after disclosure, whereas ratings for stimuli assigned to the AI-labeled condition decreased. These findings are interpreted as evidence that the proposed framework can detect disclosure-associated response drift, rather than as a fully counterbalanced causal estimate of pure label effects.**

***Keywords*—AI-generated visual information systems; metadata disclosure; response drift; user-response logs; human-computer interaction; authorship labeling; mixed-effects modeling**

I. Introduction

AI-generated visual content is increasingly integrated into interactive visual information systems, including online galleries, recommendation platforms, digital design tools, educational interfaces, and content-evaluation environments [1], [2]. In these systems, images are rarely encountered as isolated visual objects. They are commonly accompanied by metadata indicating human authorship, AI authorship, AI assistance, or the use of a specific generative model [3], [4]. Although these cues are typically introduced to support transparency and provenance disclosure, once embedded in an interface they also become part of the context in which users form ratings, preferences, and quality judgments.

The implications extend beyond individual perception because many AI-mediated visual systems use user responses as operational inputs. Ratings, preference scores, likes, and evaluation logs may inform recommendation, ranking, content moderation, design feedback, or subsequent model evaluation [5]. If authorship metadata alters these responses, ratings collected under different disclosure conditions may not be directly comparable. Recent research on AI-generated media labeling confirms that disclosure cues can affect how users interpret and respond to visual content [6]. When blind and labeled ratings are pooled without retaining their disclosure state, downstream systems may learn patterns that partly reflect attribution framing rather than visual preference alone. Authorship metadata is therefore relevant not only to provenance communication but also to the interpretation of user-response data.

Evidence from AI-art evaluation and authorship attribution research suggests that source attribution can shape user judgment, yet the implications of these attribution effects for system-generated response logs remain insufficiently developed [7-9]. Existing studies have generally examined whether evaluation outcomes differ across attribution conditions rather than how metadata-conditioned changes should be recorded, quantified, and validated when visual input remains unchanged. This leaves a measurement problem: a rating may represent both an evaluation of the image and a response to the interface context in which the image was presented.

We develop and evaluate a two-phase framework for detecting metadata-conditioned response drift in AI-generated visual information systems. Participants first evaluate AI-generated images without authorship information and subsequently re-evaluate the same images with experimentally assigned attribution labels. Trial-level records are matched by participant and image, allowing disclosure-associated rating changes to be computed and statistically evaluated. Because attribution labels were assigned to fixed stimuli, the implementation is used to verify the detection procedure under the specified assignment scheme rather than to estimate a fully counterbalanced, stimulus-independent causal effect. The main contributions of this study are as follows:

- A two-phase evaluation procedure is introduced for examining response changes while holding the visual input constant across blind and disclosure conditions.
- Disclosure-associated change is formalized as a participant–image-level drift measure supported by explicit logging of phase, image identity, attribution condition, and metadata visibility.

*Corresponding author

- The response pattern is assessed using a linear mixed-effects model, a cumulative link mixed model for ordinal sensitivity, and leave-one-image-out robustness analyses.
- Analysis of 92 participants and 732 valid observations identifies a disclosure-associated pattern that remains consistent across the primary mixed-effects model, the ordinal sensitivity analysis, and the leave-one-image-out checks, illustrating the value of retaining metadata-display conditions when ratings are compared or reused as system inputs.

II. Background and Related Work

Research on AI-generated art has shown that visual judgments are shaped by more than the perceptual properties of an image. Early work on computer-generated art examined whether viewers evaluated computer-generated and human-made artworks differently [10]. Subsequent studies demonstrated that knowledge or prior assignment of AI authorship could alter aesthetic appreciation and evaluations of artistic value [7], [11]. More recent experiments have linked differences between human- and AI-attributed works to perceived authenticity, creative involvement, and anthropocentric preferences [8],[9],[12]. The direction of these effects is not necessarily uniform: broader research on algorithm aversion and algorithm appreciation indicates that responses to algorithmic systems depend on the task, prior experience, and evaluative context [13],[14]. Taken together, these studies establish authorship attribution as a consequential source cue, but their primary object of analysis remains the judgment outcome—whether a work receives higher ratings for aesthetic quality, creativity, authenticity, or artistic value under a particular attribution.

Research on synthetic-content disclosure has shifted attention from authorship beliefs alone to the design and placement of labels within digital interfaces. Wittenberg et al. found that label effects depended partly on whether the label described the production process or warned of potentially misleading content; labels affected belief in claims associated with AI-generated images, whereas a simple AI-generated label had more limited effects on intended engagement [6]. Gamage et al. further showed that responses to synthetic-content warnings vary with wording, positioning, visual presentation, and level of detail [15]. Recent findings also caution against assuming that disclosure produces a uniform penalty. Li et al. reported no overall main effect of AIGC labeling on perceived credibility, although the effect varied with prior experience and content category [16]. Gallegos et al. similarly found that labeling messages as AI-generated did not significantly reduce their persuasive effects [17]. Recent evidence further suggests that disclosure timing and the level of AI involvement may influence user engagement with labeled content [18]. Disclosure is therefore neither uniformly beneficial nor uniformly detrimental; its consequences depend on the labeled content, the presentation of the label, the outcome being measured, and the characteristics of the audience. This context dependence is consistent with human–AI interaction research, which treats interface presentation as part of interaction design rather than as a neutral wrapper around the underlying system [19].

Less well developed is an operational account of how attribution metadata conditions the generation and subsequent interpretation of user-response logs. Randomized between-subject experiments have examined the effects of AI-related labels on belief, trust, engagement intentions, and persuasion [6],[15],[17]. A recent mixed experiment examined perceived credibility under variations in label presence, content category, and prior experience [16]. These studies show that labels can shape evaluation, but they seldom formalize the resulting change as a measurable property of the response record itself. This distinction matters for systems that reuse ratings as recommendation, ranking, or evaluation signals: responses collected under different metadata conditions may be valid within their original experimental settings yet remain non-equivalent when pooled as system inputs. The unresolved issue is therefore not simply whether disclosure affects judgment, but how disclosure-associated variation can be retained and quantified within user-response records. Accordingly, metadata exposure is treated here as a logged evaluation condition, with response drift defined at the participant–image level under unchanged visual input.

III. Proposed Metadata-Conditioned Response Drift Evaluation Framework

A. Framework Overview

This study proposes a two-phase evaluation framework for measuring metadata-conditioned response drift in AI-generated visual information systems. The framework is designed to quantify changes in user-response data associated with authorship-related interface cues while holding the visual stimulus constant. Rather than treating the study as a general comparison between human-created and AI-generated art, the proposed framework examines assigned authorship information as a configurable interface variable that may influence rating behavior and evaluation-log interpretation.

The framework consists of three functional layers: the visual stimulus layer, the interface control and metadata layer, and the response drift analysis layer. The visual stimulus layer provides a controlled set of AI-generated images and maintains stimulus consistency across evaluation phases. The interface control and metadata layer manages phase transition, stimulus presentation order, attribution-label display, and response collection. The response drift analysis layer links blind-phase and disclosure-phase ratings at the participant–image level and estimates the direction and magnitude of rating change after metadata disclosure.

The evaluation procedure contains two sequential phases completed by the same participants. In the blind phase, participants rate AI-generated images without any authorship information. In the disclosure phase, the same images are re-presented with assigned authorship labels, either human-labeled or AI-labeled. The attribution condition is implemented at the stimulus/trial level through the interface rather than embedded in the image file, caption, or visual content. This design preserves a clear distinction between the visual stimulus and the attribution cue while enabling response changes to be

measured without altering the image content. The implementation is therefore treated as a framework-verification setting: it detects disclosure-associated response drift under the implemented label-assignment scheme, while causal claims about pure authorship-label effects independent of image identity are intentionally limited. The overall workflow of the proposed framework is shown in Fig. 1.

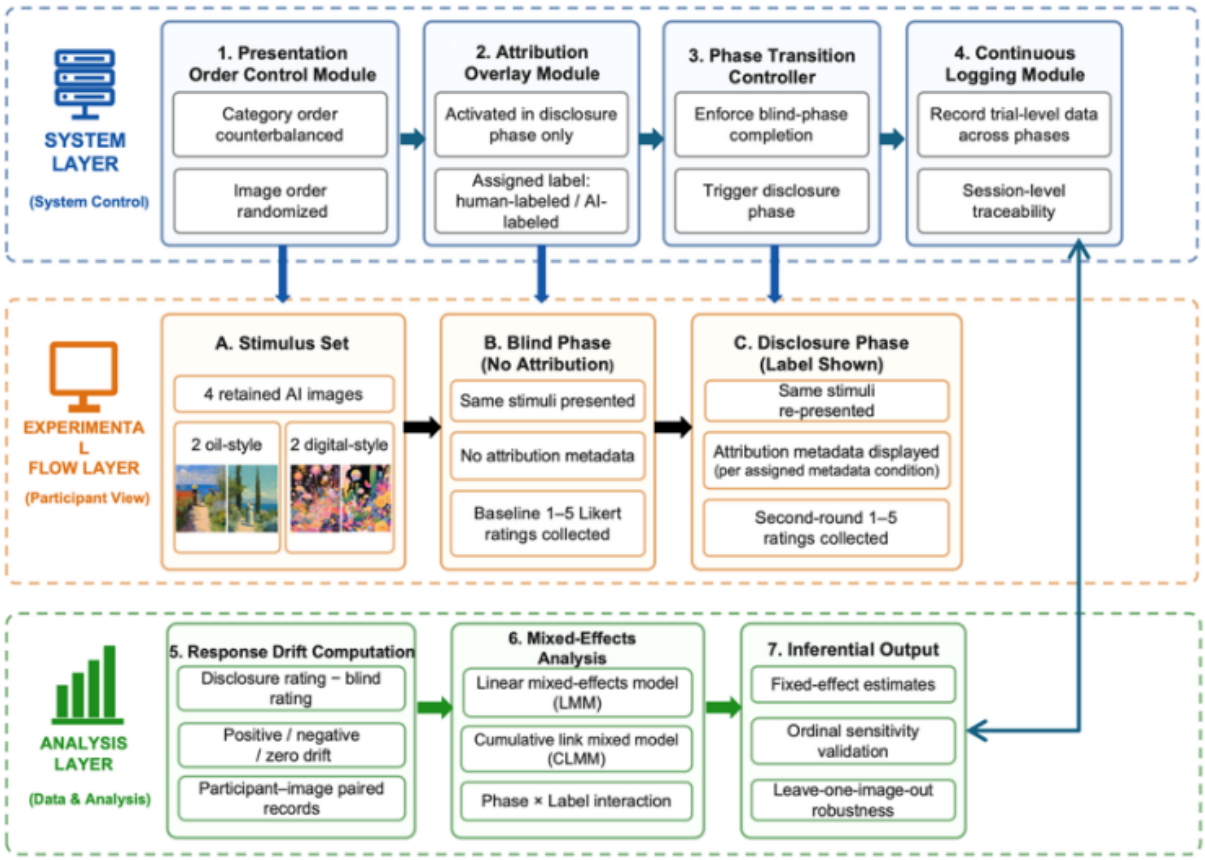


Fig. 1. Overall workflow of the proposed framework, including stimulus presentation, attribution metadata display, trial-level response logging, and response drift validation.

The primary output of the framework is response drift, defined as the difference between the disclosure-phase rating and the corresponding blind-phase rating for the same participant and image. Positive values indicate an upward shift after metadata disclosure, whereas negative values indicate a downward shift. The confirmatory effect of interest is the Phase × Label interaction, which tests whether rating changes from the blind phase to the disclosure phase differ between images assigned to the human-labeled and AI-labeled conditions.

This positioning is consistent with the objective of evaluating how attribution cues enter user-response logs as measurement-context variables in AI-generated visual information systems. The framework is intended to support disclosure-aware response logging and interface evaluation, rather than to generalize user preference for human- or AI-attributed images.

B. AI-Generated Visual Stimulus Module

The visual stimulus module was designed to provide controlled AI-generated image inputs for evaluating metadata-conditioned response drift. All stimuli were generated using the Midjourney text-to-image model through the Niji-Journey interface. Two rendering categories were constructed: oil-style compositions and digital-style compositions. These two categories were selected to represent visually distinct rendering modes while preserving a manageable experimental structure for repeated evaluation.

For each rendering category, a standardized prompt template was used to constrain the main visual structure, compositional layout, color distribution, subject arrangement, and rendering style. The oil-style template specified a painterly outdoor scene with a clear central visual element, balanced composition, soft color transitions, and an oil-painting rendering style. The digital-style template specified a colorful digital visual scene with a clear central visual element, balanced composition, vivid color distribution, and a digital-illustration rendering style. The templates were used to generate four candidate images for each category. Two images from each category were then retained, resulting in a final stimulus set of four AI-generated images. The selected stimuli were newly generated for the present study and had not been previously shown to participants.

The screening procedure was used to improve comparability across stimuli before the attribution-label conditions were applied. Candidate images were evaluated according to the following criteria: comparable visual complexity, clear thematic legibility, absence of salient rendering errors, similar prominence of the central visual element, and adequate interpretability at the display size used in the evaluation interface. The retained images were presented in the same interface layout across both phases. Image size, screen position, rating scale format, and background layout were kept constant so that the only interface difference between the blind and disclosure phases was the presence or absence of authorship metadata. This procedure was not intended to make the retained images perceptually identical or to replace full label counterbalancing. Rather, it reduced obvious stimulus-level imbalance and provided a controlled stimulus set for framework verification. Remaining participant-level heterogeneity and stimulus-related variation were addressed through mixed-effects modeling and robustness checks.

Fig. 2. Retained AI-generated visual stimuli used in the two-phase evaluation. O1 and O2 represent oil-style stimuli, whereas D1 and D2 represent digital-style stimuli. All four images were generated by AI and were re-presented across blind and disclosure phases.

The retained oil-style images were assigned image identifiers O1 and O2, and the retained digital-style images were assigned image identifiers D1 and D2 (see Fig. 2). During the disclosure phase, O1 and D1 were displayed with the assigned label Human-labeled, whereas O2 and D2 were displayed with the assigned label AI-labeled. Because all four stimuli were AI-generated, these labels functioned only as experimentally assigned attribution cues within the evaluation interface rather than as indicators of actual authorship.

C. Attribution Metadata Overlay Module

The attribution metadata overlay module was developed to separate visual content from authorship information. In the blind phase, the metadata overlay was inactive, and participants viewed the AI-generated images without any authorship cue. In the disclosure phase, the overlay was activated and displayed the assigned attribution label for each image. The displayed label used one of two interface texts: Human-labeled or AI-labeled.

The authorship label was presented as an interface text label. It was placed in a fixed position adjacent to the image display area and remained visible on the same screen as the image and rating scale throughout the disclosure-phase trial. The label position, font size, and display format were kept constant across labeled trials. This design introduced attribution as a controlled interface variable while preserving the visual stimulus itself.

The metadata overlay was persistent during the labeled evaluation trial, meaning that the assigned label remained visible until participants submitted the second-round rating. This procedural control ensured that attribution information was available during the disclosure-phase response. No post-task manipulation-check item was collected; therefore, label exposure was controlled through the interface procedure rather than through a separate self-report verification item.

D. Phase Transition and Trial-Level Logging Module

The phase transition and trial-level logging module controlled the evaluation sequence and recorded structured response data. The evaluation followed a two-phase repeated-measures procedure and involved 92 participants (mean age = 38.30 years; SD = 13.54; range = 20–72 years). All participants reported normal or corrected-to-normal vision and no known color-vision deficiency. Thirty-eight participants reported at least two years of prior art-related learning or practice experience, whereas fifty-four participants reported no comparable art-training background. In the first phase, participants completed the blind evaluation without authorship information. In the second phase, the same images were re-presented under disclosure conditions with assigned attribution labels.

The interface managed the transition from the blind phase to the disclosure phase and prevented participants from entering the disclosure phase before completing the blind evaluation. A rating response was required before participants could proceed to the next trial. No time limit was imposed on image viewing or rating submission. The order of stimulus categories was counterbalanced across participants, and the presentation order of individual images within each category was randomized. Specifically, participants were assigned to alternating category-order sequences, and image order within each category was randomly shuffled at the session level. This procedure reduced the influence of fixed presentation order while maintaining the intended two-phase structure.

For each trial, the system recorded a structured response entry containing the anonymized participant/session identifier,

image identifier, stimulus category, presentation phase, attribution condition, metadata visibility state, and rating response. These trial-level records provided the data structure required to match blind-phase and disclosure-phase ratings for the same participant and image. Responses were stored without personally identifiable information. The logging module therefore served as the operational basis for computing response drift and for evaluating whether assigned attribution cues were associated with systematic changes in user ratings.

The complete evaluation consisted of eight trials for each participant: four blind-phase trials and four disclosure-phase trials. Because each participant evaluated the same four images twice, observations were repeatedly measured within participants across fixed stimuli. This structure motivated the use of a participant-level mixed-effects model in the validation layer. The complete procedure is summarized in Table I and Algorithm 1.

TABLE I. OPERATIONAL PROCEDURE OF THE PROPOSED EVALUATION FRAMEWORK.

Stage	Module	Operation	Output
1	Visual stimulus module	Generate and screen AI-generated images	Retained image set
2	Blind evaluation	Present images without metadata	Baseline ratings
3	Metadata overlay	Activate human-labeled / AI-labeled display	Disclosure condition
4	Disclosure evaluation	Re-present the same images with metadata	Labeled ratings
5	Logging layer	Match records by participant and image ID	Trial-level paired data
6	Drift computation	Apply Eq. (1)	Response drift scores
7	Validation layer	Fit LMM, CLMM, and robustness checks	Effect estimates and CIs

Algorithm 1: Metadata-Conditioned Response Drift Detection Procedure

Require: AI-generated stimulus set S={O1,O2,D1,D2} ; attribution label set L={Human-labeled, AI-labeled}; metadata visibility states V={Hidden,Visible}; participant set **P**; 5-point rating scale.

Ensure: Paired trial-level response logs, response drift scores, and statistical validation outputs.

start
for each participant $p \in P$ **do**
initialize anonymized participant session SID_p
load the retained AI-generated stimulus set S
assign participant p to one counterbalanced category-order sequence
randomize the presentation order of images within each category
for each image i in the randomized blind-phase sequence **do**
present image i with attribution metadata hidden
collect blind-phase rating $R_{p,i}^{\text{blind}}$
record log entry $\langle SID_p, i, Category_i, Blind, None, Hidden, R_{p,i}^{\text{blind}} \rangle$
end for
activate the attribution metadata overlay module
for each image i in the corresponding disclosure-phase sequence **do**
re-present the same visual stimulus i
display the assigned attribution label $Label_i$
keep $Label_i$ visible until rating submission
collect disclosure-phase rating $R_{p,i}^{\text{disclosure}}$
record log entry $\langle SID_p, i, Category_i, Disclosure, Label_i, Visible, R_{p,i}^{\text{disclosure}} \rangle$
if both blind-phase and disclosure-phase ratings are available **then**
compute response drift $Drift_{p,i} = R_{p,i}^{\text{disclosure}} - R_{p,i}^{\text{blind}}$
end if
end for
end for
aggregate matched participant–image records
fit the primary linear mixed-effects model
conduct ordinal sensitivity using a cumulative link mixed model
perform leave-one-image-out robustness checks
output drift estimates, confidence intervals, and validation results
end

E. Response Drift Computation

The primary output of the framework was response drift, defined as the rating change between the disclosure phase and the blind phase for the same participant and image. Let $R_{p,i}^{\text{blind}}$ denote the blind-phase rating provided by participant p for image i, and let $R_{p,i}^{\text{disclosure}}$ denote the corresponding disclosure-phase rating after authorship metadata was displayed. The response drift score was computed as:

$$\text{Drift}_{p,i} = R_{p,i}^{\text{disclosure}} - R_{p,i}^{\text{blind}} \quad (1)$$

where, $\text{Drift}_{p,i}$ represents the metadata-associated rating shift for participant p and image i. A positive drift value indicates that the participant assigned a higher rating after metadata disclosure than under blind conditions. A negative drift value indicates that the participant assigned a lower rating after disclosure. A drift value of zero indicates no change between the two phases. This computation allowed the framework to transform raw trial-level ratings into an interpretable measure of metadata-associated response change.

The response drift measure was calculated at the participant – image level and then examined by attribution condition and stimulus category. This structure enabled the framework to evaluate whether rating shifts differed between human-labeled and AI-labeled conditions, and whether the observed pattern was stable across oil-style and digital-style images.

F. Statistical and Robustness Validation Layer

The statistical validation layer was designed to estimate disclosure-associated changes in ratings while accounting for repeated observations within participants. Because each participant provided ratings across multiple images and phases, linear mixed-effects modeling was used as the primary inferential approach [20]. Ratings were modeled as a function of presentation phase, attribution label, stimulus category, and art experience.

The primary model was specified as:

$$\text{Rating}_{p,i,t}=\beta_0+\beta_1\text{Phase}_t+\beta_2\text{Label}_i+\beta_3\text{Category}_i$$
$$+\beta_4\text{ArtExperience}_p+\beta_5(\text{Phase}_t\times\text{Label}_i)$$
$$+\beta_6(\text{Phase}_t\times\text{Category}_i)+\beta_7(\text{Label}_i\times\text{Category}_i)$$
$$+\beta_8(\text{Phase}_t\times\text{Label}_i\times\text{Category}_i)+u_p+\varepsilon_{p,i,t} \quad (2)$$

where, $\text{Rating}_{p,i,t}$ refers to the 5-point evaluation score provided by participant p for image i at phase t. Phase_t distinguishes blind and disclosure evaluations, Label_i distinguishes human-labeled and AI-labeled metadata conditions, and Category_i distinguishes oil-style and digital-style stimuli. ArtExperience_p was entered as a participant-level covariate. In this specification, u_p represents the participant-level random intercept, and $\varepsilon_{p,i,t}$ denotes the residual error term. Dummy coding was used, with the blind phase, AI-labeled condition, and oil-style category serving as reference levels.

A more complex candidate model allowing participant-level random slopes for phase and label was also examined:

$$\text{Rating}_{p,i,t}=\beta_0+\beta_1\text{Phase}_t+\beta_2\text{Label}_i+\beta_3\text{Category}_i$$
$$+\beta_4\text{ArtExperience}_p+\beta_5(\text{Phase}_t\times\text{Label}_i)$$
$$+\beta_6(\text{Phase}_t\times\text{Category}_i)+\beta_7(\text{Label}_i\times\text{Category}_i)$$
$$+\beta_8(\text{Phase}_t\times\text{Label}_i\times\text{Category}_i)+u_{0p}$$
$$+u_{1p}\text{Phase}_t+u_{2p}\text{Label}_i+\varepsilon_{p,i,t} \quad (3)$$

When this model produced convergence warnings or singular estimates, the random-effects structure was reduced to a participant-level random intercept. The confirmatory effect of interest was the Phase × Label interaction, which tested whether rating changes from the blind phase to the disclosure phase differed between human-labeled and AI-labeled conditions. Because attribution labels were not fully counterbalanced across specific images, the Label coefficient was interpreted conditionally at the reference levels and was not treated as evidence of a general preference for either label independent of stimulus identity.

Because ratings were collected using a 5-point Likert scale, an ordinal sensitivity analysis was conducted using a cumulative link mixed model with the same fixed-effects structure and a participant-level random intercept [21],[22]. This analysis tested whether the main inferential pattern remained stable when the ordinal nature of the outcome variable was explicitly modeled. Planned contrasts were used to estimate phase-related rating changes within each label and category condition, and follow-up comparisons were adjusted using the Holm method.

Robustness was further examined through four leave-one-image-out analyses. Because omitting one stimulus left one Label × Category combination unobserved, each subset was analyzed using a reduced mixed-effects specification retaining the Phase × Label interaction, with stimulus category and art experience included as covariates and a participant-level random intercept. This analysis assessed whether the interaction was sensitive to any individual stimulus.

IV. Results and Framework Evaluation

A. Data Completeness and Logging Reliability

The proposed framework generated a maximum of 736 possible rating records, corresponding to 92 participants, four images, and two evaluation phases. A total of 732 valid rating observations were recorded and retained for analysis, yielding a valid logging rate of 99.46% and an item-level missing rate of 0.54%. The four missing observations came from one participant and did not alter the repeated-measures structure of the dataset. Analyses were therefore conducted using available observations without imputation.

The high valid logging rate confirmed that the framework produced sufficiently complete participant–image paired records for subsequent response-drift computation and mixed-effects validation.

B. Detection of Disclosure-Associated Response Drift

The primary validation objective was to determine whether the proposed framework could detect a disclosure-associated response-drift pattern under fixed visual input. The confirmatory linear mixed-effects model revealed a significant Phase × Label interaction in the oil-style reference category, $\beta = 0.780$, SE = 0.153, 95% CI [0.481, 1.080], $p < 0.001$. This interaction indicates that rating changes from the blind phase to the disclosure phase differed systematically between human-labeled and AI-labeled conditions.

TABLE II. Key Coefficient Estimates from the Primary and Ordinal Sensitivity Models.

Model	Term	β	SE	95% CI	p
Linear mixed-effects model	Phase × Label	0.780	0.153	[0.481, 1.080]	<0.001
	Label coefficient	-0.022	0.108	[-0.234, 0.190]	0.839
	Phase × Label × Category	0.220	0.216	[-0.203, 0.642]	0.309
	Art Experience	0.018	0.088	[-0.155, 0.191]	0.841
Cumulative link mixed model	Phase × Label	2.195	0.407	[1.398, 2.992]	<0.001
	Label coefficient	-0.051	0.275	[-0.589, 0.488]	0.854
	Phase × Label × Category	0.548	0.569	[-0.567, 1.664]	0.335
	Art Experience	0.102	0.238	[-0.365, 0.569]	0.668

Note. β coefficients from the cumulative link mixed model are estimated on the log-odds scale and are therefore not directly comparable in magnitude to coefficients from the linear mixed-effects model.

Specifically, ratings decreased after disclosure for images assigned to the AI-labeled condition and increased for images assigned to the human-labeled condition. For the oil-style reference category, the blind-phase difference between images assigned to the two label conditions was not significant, $\beta = -0.022$, SE = 0.108, 95% CI [−0.234, 0.190], p = 0.839,

indicating no detectable baseline difference within that category. Art experience also showed no reliable association with ratings, β = 0.018, SE = 0.088, 95% CI [−0.155, 0.191], p = 0.841. Together, these results indicate that the framework detected a disclosure-associated response-drift pattern after accounting for participant-level heterogeneity, stimulus category, and art experience, as summarized in Table II and illustrated in Fig. 3.

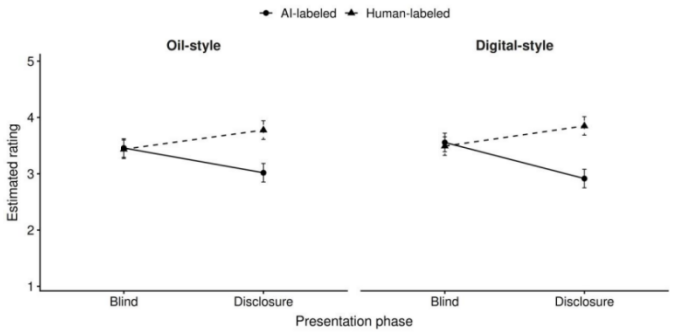


Fig. 3. Disclosure-associated response drift across stimulus categories.

C. Category-Level Stability of the Drift Pattern

Category-level stability was evaluated to determine whether the *detected* response-drift pattern was limited to a single rendering category. The Phase × Label × Category interaction was not significant, β = 0.220, SE = 0.216, 95% CI [-0.203, 0.642], p = 0.309. This result indicates no reliable evidence that the disclosure-associated drift pattern differed between oil-style and digital-style stimuli.

Model-derived follow-up contrasts were consistent with the directional pattern observed in the primary model. In the AI-labeled condition, ratings decreased significantly from the blind phase to the disclosure phase for both the oil-style category, β = −0.440, SE = 0.108, 95% CI [−0.652, −0.227], Holm-adjusted p < 0.001, and the digital-style category, β = −0.641, SE = 0.108, 95% CI [−0.852, −0.430], Holm-adjusted p < 0.001. In contrast, in the human-labeled condition, ratings increased significantly for both the oil-style category, β = 0.341, SE = 0.108, 95% CI [0.128, 0.553], Holm-adjusted p = 0.002, and the digital-style category, β = 0.359, SE = 0.108, 95% CI [0.148, 0.570], Holm-adjusted p = 0.002. These contrasts are shown in Fig. 4.

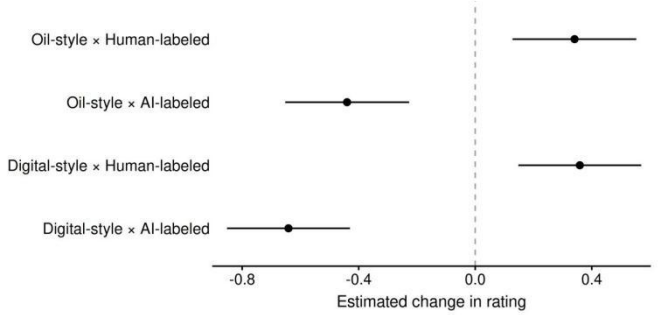


Fig. 4. Model-derived rating change from the blind phase to the disclosure phase. Points indicate estimated changes in rating within each Label × Category condition; horizontal bars indicate 95% confidence intervals. The dashed vertical line marks zero change.

D. Ordinal Sensitivity Validation

Because ratings were collected using a 5-point Likert scale, an ordinal sensitivity validation was conducted using a cumulative link mixed model. The ordinal model produced the same inferential pattern as the primary linear mixed-effects model. The Phase × Label interaction remained significant, β = 2.195, SE = 0.407, 95% CI [1.398, 2.992], p < 0.001. The Label coefficient at the blind-phase, oil-style reference levels was not significant, β = −0.051, SE = 0.275, 95% CI [−0.589, 0.488], p = 0.854. The Phase × Label × Category interaction also did not reach significance, β = 0.548, SE = 0.569, 95% CI [−0.567, 1.664], p = 0.335. Art Experience was also not significant, β = 0.102, SE = 0.238, 95% CI [-0.365, 0.569], p = 0.668. The consistency between the linear mixed-effects model and the cumulative link mixed model indicates that the detected response-drift pattern was not attributable to treating Likert-scale ratings as approximately continuous.

E. Leave-One-Image-Out Robustness Check

Robustness was assessed through four leave-one-image-out analyses. Across the reduced models, the Phase × Label interaction remained significant, with estimated coefficients ranging from 0.789 to 0.991 (all p<0.001). The direction and significance of the interaction were unchanged when any one stimulus was omitted, indicating that the observed response-drift pattern was not driven by a single image.

Participant-level drift estimates showed the same directional pattern. In the AI-labeled condition, mean drift was negative, M = -0.543, SD = 0.623, median = -0.500, with 67 of 92 participants showing negative drift. In the human-labeled condition, mean drift was positive, M = 0.353, SD = 0.557, median = 0.500, with 56 of 92 participants showing positive drift, as shown in Fig. 5.

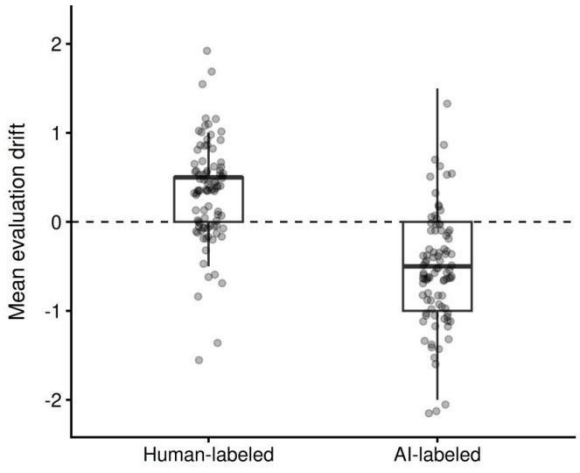


Fig. 5. Participant-level response drift by assigned authorship label. Each point represents one participant's mean response drift within a label condition, computed as the disclosure-phase rating minus the corresponding blind-phase rating. Positive values indicate higher ratings in the disclosure phase than in the blind phase, whereas negative values indicate lower ratings. The dashed horizontal line denotes zero response drift.

Together, the leave-one-image-out and participant-level analyses provide additional support for the stability of the

detected response-drift pattern across the retained stimuli and participants.

V. Discussion and System Implications

A. Metadata as a Response-Shaping Interface Variable

Across the primary mixed-effects model, ordinal sensitivity analysis, planned contrasts, and leave-one-image-out checks, the results converged on the same directional pattern: ratings shifted after authorship metadata was disclosed, while the visual stimuli remained unchanged. This convergence supports the interpretation that metadata disclosure was associated with measurable response drift in the implemented two-phase evaluation framework.

The consistency across analyses follows from complementary features of the design and validation strategy. Matching ratings within the same participant–image pair limits the influence of stable individual differences in scale use, while the participant-level random intercept accounts for dependence across repeated ratings from the same individual. The ordinal model tests whether the inference depends on treating the five-point rating scale as approximately continuous, and the leave-one-image-out analyses examine whether the pattern is driven by any single retained stimulus. Agreement across these analyses therefore indicates that the observed pattern is not tied to one statistical specification or one image, although it does not eliminate the confounding introduced by fixed label–stimulus assignment.

These findings do not establish a general preference for human-created content or a stimulus-independent causal effect of authorship labels. They instead indicate that authorship metadata forms part of the conditions under which user-response data are produced. Ratings collected under different metadata states should therefore not be treated as automatically equivalent when they are compared or reused as system inputs; stronger causal claims require fully counterbalanced label assignments.

B. Implications for User-Response Logs and Downstream Visual Systems

The findings have direct implications for AI-generated visual platforms that use ratings, preference scores, or interaction logs as system inputs. In recommendation, ranking, content evaluation, and design-feedback settings, these responses are often treated as indicators of content quality, relevance, or user preference. The present results show that this assumption is incomplete: a rating collected after authorship metadata disclosure may not be equivalent to a rating collected under blind conditions, even when the visual content remains unchanged.

For this reason, metadata-display conditions should be retained as contextual variables in response logs. Ratings collected under blind, AI-labeled, human-labeled, or other disclosure conditions should not be pooled as fully interchangeable signals. Otherwise, downstream models may learn patterns that partly reflect attribution framing rather than visual preference alone.

The proposed framework provides a practical way to identify this source of variation. By matching blind-phase and disclosure-phase ratings at the participant–image level, it separates baseline visual response from metadata-conditioned response change. At the implementation level, a minimal response-log schema should include the participant or session identifier, image identifier, presentation phase, label condition, label visibility state, rating value, and rating time. These fields allow system designers to determine the interface context under which each rating was produced.

These results also suggest that future AI-generated visual systems should account for disclosure state during response modeling. Possible implementations include disclosure-aware feature engineering, condition-specific rating calibration, or drift-weighted response modeling. Such approaches would help recommendation, ranking, and evaluation pipelines distinguish visual preference from interface-conditioned response variation.

C. Disclosure Design Considerations

The findings should not be interpreted as an argument against AI disclosure. Transparency remains important for provenance communication, user autonomy, and responsible deployment of AI-generated content [6]. However, the present study shows that disclosure is not behaviorally neutral. Once authorship metadata is placed inside an evaluation interface, it becomes part of the rating environment and may influence user-response data [6, 8]. Because the present implementation used a persistent text label displayed during the rating trial, the results should be interpreted most directly as evidence for this specific disclosure format rather than for all possible forms of AI-content disclosure.

Disclosure should therefore be treated as an interface design variable rather than as a fixed label added after system development. The timing, wording, placement, and visual salience of disclosure may all affect how users interpret and evaluate content [15]. A label displayed continuously during rating may produce a different response from post-evaluation disclosure, a hover-based provenance marker, a low-salience label, or a progressive disclosure interface. These design choices should be empirically tested rather than assumed to have no effect on user judgment.

For AI-generated visual information systems, the practical question is not simply whether disclosure should exist, but how disclosure should be implemented when user ratings are used as system input. Platforms may need to distinguish between disclosures used for compliance, disclosures used for provenance explanation, and disclosures shown during evaluative tasks. When ratings are used for recommendation, ranking, or quality assessment, disclosure design should be evaluated not only for informational accuracy but also for its effect on response drift.

VI. Limitations, Boundary Conditions and Future Work

The findings should be interpreted within the scope of the implemented framework. First, the study adopted a repeated-evaluation design in which the same participants rated the same images before and after metadata disclosure. This design was necessary for computing participant – image level response drift, but it may also introduce familiarity or carryover effects

between the blind and disclosure phases. The results therefore reflect response changes under a controlled two-phase interface setting rather than responses in a fully naturalistic browsing environment.

Second, attribution labels were not fully counterbalanced across stimuli, which limits claims about pure label effects independent of stimulus identity. To address this limitation, the primary model included stimulus category as a fixed effect and participant identity as a random intercept, while ordinal sensitivity analysis and leave-one-image-out analyses were used to assess the robustness of the detected pattern across model specifications and individual stimuli. These procedures provide robustness evidence but do not replace full label counterbalancing or support population-level generalization across images. The study is therefore interpreted as framework verification rather than as a causal estimate of pure authorship-label effects.

Third, the stimulus set was intentionally compact. The use of four AI-generated images made it possible to maintain a controlled repeated-measures structure and to keep the evaluation task manageable for participants. However, this design limits the range of visual styles, image topics, and aesthetic variation covered by the study. Future work should test the framework with larger and more diverse stimulus sets, including different genres of AI-generated visual content, multiple levels of visual quality, and broader interface contexts.

Fourth, label exposure was controlled through a persistent on-screen attribution overlay, but no separate post-task manipulation-check item was collected. The findings therefore rely on procedural exposure rather than self-reported verification of label awareness. Future studies should include manipulation checks or attention measures to examine whether participants noticed and interpreted the displayed attribution labels as intended.

Future research can extend the framework in three main directions. A fully counterbalanced design should assign each image to different metadata conditions across participants in order to separate label effects from stimulus identity more strictly. Alternative disclosure formats should also be compared, including pre-evaluation disclosure, post-evaluation disclosure, persistent labels, hover-based provenance markers, low-salience labels, and progressive disclosure interfaces. Finally, future AI-generated visual systems should record metadata-display conditions as part of the evaluation log, so that recommendation, ranking, and content assessment models can support disclosure-aware feature engineering, condition-specific calibration, or drift-weighted response modeling instead of treating all ratings as equivalent indicators of visual preference.

VII. Conclusion

This study proposed a two-phase evaluation framework for detecting metadata-conditioned response drift in AI-generated visual information systems. By integrating fixed visual input, assigned attribution-label presentation, trial-level logging, drift computation, and mixed-effects validation, the framework provides a structured procedure for examining whether attribution cues are associated with measurable changes in user-evaluation data. Based on 92 participants and 732 valid observations, the results showed a disclosure-associated drift pattern under the implemented label-assignment scheme: ratings increased for images assigned to the human-labeled condition and decreased for images assigned to the AI-labeled condition. This pattern was consistent across oil-style and digital-style stimuli and remained stable in ordinal sensitivity analysis and leave-one-image-out checks. Because the present implementation did not fully counterbalance labels across images and used a fixed blind-to-disclosure phase order, the results should be interpreted as framework-verification evidence rather than as a fully isolated causal estimate of pure authorship-label effects. Even within this boundary, the framework offers practical support for disclosure-aware response logging, interface evaluation, and the interpretation of user-rating data in AI-generated visual information systems.

Ethical Declarations

This study was a minimal-risk anonymous human–computer interaction evaluation involving adult participants who viewed and rated AI-generated images. Participation was voluntary, and informed consent was obtained before the task. Responses were recorded without personally identifiable information and analyzed only in aggregate form. According to institutional guidelines for minimal-risk anonymous evaluations, formal ethics review was not required. After completing the task, participants were informed that the displayed authorship labels were experimentally assigned attribution cues and that all stimuli were AI-generated.

References

[1] Z. Epstein et al., "Art and the science of generative AI," Science, vol. 380, no. 6650, pp. 1110-1111, 2023.

[2] E. Cetinic and J. She, "Understanding and creating art with AI: Review and outlook," ACM transactions on multimedia computing, communications, and applications (TOMM), vol. 18, no. 2, pp. 1-22, 2022.

[3] J. K. Eshraghian, "Human ownership of artificial creativity," Nature Machine Intelligence, vol. 2, no. 3, pp. 157-160, 2020.

[4] Z. Epstein, S. Levine, D. G. Rand, and I. Rahwan, "Who gets credit for AI-generated art?," Iscience, vol. 23, no. 9, 2020.

[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 Computing and Applications, vol. 37, no. 17, pp. 10875-10947, 2025.

[6] C. Wittenberg, Z. Epstein, G. Péloquin-Skulski, A. J. Berinsky, and D. G. Rand, "Labeling AI-generated media online," PNAS nexus, vol. 4, no. 6, p. pgaf170, 2025.

[7] H. Gangadharbatla, "The role of AI attribution knowledge in the evaluation of artwork," Empirical Studies of the Arts, vol. 40, no. 2, pp. 125-142, 2022.

[8] L. Bellaiche et al., "Humans versus AI: whether and why we prefer human-created compared to AI-created artwork," Cognitive research: principles and implications, vol. 8, no. 1, p. 42, 2023.

[9] K. Millet, F. Buehler, G. Du, and M. D. Kokkoris, "Defending humankind: Anthropocentric bias in the appreciation of AI art," Computers in Human Behavior, vol. 143, p. 107707, 2023.

[10] R. Chamberlain, C. Mullin, B. Scheerlinck, and J. Wagemans, "Putting the art in artificial: Aesthetic responses to computer-generated art," Psychology of Aesthetics, Creativity, and the Arts, vol. 12, no. 2, p. 177, 2018.

[11] S. G. Chiarella, G. Torromino, D. M. Gagliardi, D. Rossi, F. Babiloni, and G. Cartocci, "Investigating the negative bias towards artificial intelligence: Effects of prior assignment of AI-authorship on the

aesthetic appreciation of abstract paintings," Computers in Human Behavior, vol. 137, p. 107406, 2022.

[12] C. B. Horton Jr, M. W. White, and S. S. Iyengar, "Bias against AI art can enhance perceptions of human creativity," Scientific reports, vol. 13, no. 1, p. 19001, 2023.

[13] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Algorithm aversion: people erroneously avoid algorithms after seeing them err," Journal of experimental psychology: General, vol. 144, no. 1, p. 114, 2015.

[14] J. M. Logg, J. A. Minson, and D. A. Moore, "Algorithm appreciation: People prefer algorithmic to human judgment," Organizational behavior and human decision processes, vol. 151, pp. 90-103, 2019.

[15] D. Gamage, D. Sewwandi, M. Zhang, and A. K. Bandara, "Labeling synthetic content: User perceptions of label designs for AI-generated content on social media," in Proceedings of the 2025 CHI conference on human factors in computing systems, 2025, pp. 1-29.

[16] F. Li, Y. Yang, and G. Yu, "Nudging perceived credibility: The impact of AIGC labeling on user distinction of AI-generated content," Emerging Media, vol. 3, no. 2, pp. 275-304, 2025.

[17] I. O. Gallegos et al., "Labeling messages as AI-generated does not reduce their persuasive effects," PNAS nexus, vol. 5, no. 2, p. pgag008, 2026.

[18] F. Seeger, M. Wessel, and C. Lehrer, "AI content labeling and user engagement on social media: The role of AI level, content type, and disclosure timing," Electronic Markets, vol. 36, no. 1, p. 31, 2026.

[19] S. Amershi et al., "Guidelines for human-AI interaction," in Proceedings of the 2019 chi conference on human factors in computing systems, 2019, pp. 1-13.

[20] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," Journal of statistical software, vol. 67, pp. 1-48, 2015.

[21] R. H. B. Christensen, "ordinal—regression models for ordinal data," R package version, vol. 10, no. 2019, p. 54, 2019.

[22] G. Norman, "Likert scales, levels of measurement and the “laws” of statistics," Advances in health sciences education, vol. 15, no. 5, pp. 625-632, 2010.

Hybrid Machine Learning and Reinforcement Learning for IoT Device Classification and Bandwidth Optimization

Manjunatha T N[1], Vidyalakshmi K[2], Dankan Gowda V[3]
Research Scholar, Department of Computer Science and Engineering-Sri Siddartha Institute of Technology, SSAHE University, Tumkur, Karnataka, India[1]
Associate Professor, Department of Computer Science and Engineering (Data Science)-Sri Siddartha Institute of Technology, SSAHE University, Tumkur, Karnataka, India[2]
Department of Electronics and Communication Engineering, B.M.S Institute of Technology and Management, Bangalore, India[3]

***Abstract*—Increased traffic in the Internet of Things (IoT) network results in a large amount of network data that should be processed efficiently in order to manage IoT network resources properly. Hence, bandwidth management and resource allocation are key challenges in managing large-scale IoT networks. In this study, a novel hybrid framework that is composed of machine learning (ML) and reinforcement learning (RL) methods has been proposed for IoT device classification as well as for dynamic bandwidth allocation in large-scale IoT networks. The proposed framework includes four ML classification methods, including Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbor (KNN), and Neural Network (NN), for classifying various types of IoT devices by using their traffic and communication behaviors. After providing classification results to the bandwidth management framework, improved bandwidth allocation is achieved by using the proposed framework. In addition, the RL method is also incorporated into the framework for bandwidth allocation, in which the learning agent continuously improves allocating bandwidth by increasing the cumulative reward during the learning process. Experiment results revealed that the proposed framework significantly improved performance metrics of IoT networks, such as increased network throughput by approximately 50% and decreased average delay by approximately 40%, as well as reduced packet loss by approximately 30% compared with traditional bandwidth allocation methods. In terms of classification accuracy, the highest accuracy was achieved by the Neural Network model. The proposed framework enables to efficiently and adaptively manage large-scale IoT networks.**

Keywords*—IoT device classification; bandwidth allocation; machine learning algorithms; support vector machines; Random Forest; K-Nearest Neighbors; Neural Networks; network optimization; IoT traffic; resource management

I. Introduction

Internet of Things (IoT) has experienced explosive growth, and now, billions of devices are linked to networks, and they exist in a variety of industries, including healthcare, transportation, agriculture, and intelligent cities [1]. Such an increase in IoT devices has translated to a continuous rise in the need to use bandwidth efficiently to achieve maximum performance in IoT networks. With the number of connected devices increasing, complexity in controlling their traffic increases, and with that, there are rampant bandwidth management issues [2]. The IoT devices cover a wide variety of devices, such as consumer devices, actuators and sensors, and each of them has different communication needs and data consumption patterns [3]. The non-homogeneous character of such devices complicates the process of efficient bandwidth allocation since such devices require different amounts of data depending on the type, purpose, and frequency of communications [4]. The allocation of bandwidth is also problematic and a key element in supporting a quality service in IoT networks. The conventional ways of managing bandwidth have been based on fixed or rough techniques, which do not take into consideration the dynamic traffic of IoT [5]. Such solutions cannot effectively fit the needs of diverse devices in the IoT and cannot allocate priorities to bandwidth depending on the classification of devices, resulting in the inefficient use of network resources, congestion, and overall poor performance [6]. Therefore, a new method of classifying IoT devices is required, and categorizing them according to the results of the classification of the devices and distributing bandwidth in accordance with the findings of the classification [7]. This study seeks to fill this gap by elaborating a machine learning-based framework for the classification of IoT devices. With the help of more advanced machine learning models, including the SVM, RF, KNN, and NN, the IoT devices may be categorized according to their traffic configuration and characteristics [8]. After classification, the process of bandwidth allocation can be enhanced to suit the specific needs of each type of device better, which will not only allow dividing the resources more efficiently and achieving a higher network performance, but will also allow reducing the use of the resources [9]. This study has a dual purpose: first, to design a strong system of classifying IoT devices based on machine learning algorithms, and second, to analyze the effectiveness of the latter in the context of efficient bandwidth allocation. The importance of the research lies in the fact that the project can be used to improve the efficiency of the network in the IoT environment by offering a solution to bandwidth management that can be scaled into the new limits and adjusted to the required specifications. The proposed framework can make the work of the IoT network more

efficient, guaranteeing high-quality performance and reducing network overload, as it can make the process of device classification more precise, effective, and efficient when distributing resources.

II. Literature Review

The use of IoT devices is growing at a rapid rate, contributing to the need to allocate more bandwidth in IoT networks and turning bandwidth allocation into a significant factor in network performance. The existing approaches to bandwidth assignment are mostly based on a fixed or organic system that is not responsive to the various types of IoT devices [10]. These devices have great diversity in terms of data consumption, frequency of communication, as well as the type of data communicated [11]. Satellite-fixed approaches are not effective in dealing with the dynamism and heterogeneity of the IoT traffic, which results in network overloading and resource wastage [12]. The need to employ smarter bandwidth management techniques has been noticed in several studies, and most of them have brought out techniques that are founded on QoS or priority levels attributed to various classes of devices [13]. Nevertheless, the versatility to deal with the diverse range of devices and traffic patterns effectively is not found in these methods. ML has become a potential instrument in the enhancement of the operation of IoT networks, such as classifying devices and optimizing bandwidth. It has been shown that ML algorithms, including SVM, KNN, and RF, can be used to classify the IoT devices according to the traffic [14]. With the help of ML methods, devices can be classified dynamically, which allows making more precise predictions of how devices can behave and allocating resources more efficiently [15]. Additionally, the field of ML has been used with the objective of optimizing the allocation of bandwidth by utilizing historical network conditions and returning to adjust bandwidth utilization in real-time to adapt to the variations of traffic. Methods such as reinforcement learning have demonstrated the opportunity to dynamically optimize the bandwidth allocation to maintain the best performance depending on the classification of the devices and network characteristics [16]. Despite the mentioned development, applying device classification and bandwidth allocation optimization based on ML has not yet reached the stage of active research, as a lot of the existing models concentrate on a single potential aspect. Although the machine learning literature in IoT is comprehensive, there is still a gap in the efficient use of the classifier and bandwidth allocation of the devices [17]. The majority of the available studies have concentrated on the issues of either classification or optimization, and most studies do not consider a synthesis of both in real-life IoT networks. Moreover, most studies have simplified or simulated datasets and may not necessarily achieve the complexity and challenges of large-scale, heterogeneous IoT networks [18]. The study will help address these gaps by developing an extended framework comprising the combination of the classification of IoT devices with machine learning algorithms and bandwidth optimization methods, developing a scalable solution that can be modified to real-world IoT settings. The solution will offer a more comprehensive and successful approach to bandwidth management in any given IoT network and deal with classification and optimization within a single framework.

III. Proposed Method

The data that will be utilized in this work are the features of an IoT device and traffic streams. The characteristics of the IoT devices are data use habits, communication protocols, power usage, data rate, and frequency of communications, as well as intervals of transmission. These characteristics are necessary in categorizing devices according to their conduct and traffic requirements. The origin of the data can be different; it can be actual data of related IoT devices, or it can be artificial data according to predetermined conditions. Data in the real world would normally be obtained through IoT testbeds or publicly available data, whilst simulated data may be produced to represent such conditions of traffic in controlled scenarios.

The feature engineering process requires obtaining the important features of the raw data, which will enable machine learning models to be useful in the classification of the IoT devices. Such characteristics are time series, transmission rates, frequency of communication, and power consumption. The time series data will record the behavior of the IoT devices as time progresses, and this will help in classifying them because of the trends in increasing flexibility of communication. Data rate and frequency of communication are important signifiers of the device (e.g., sensors might not communicate with a high frequency, but consumer devices may have a lot of data). The other differentiation offers power consumption, and gadgets with increased energy requirements might require different bandwidth.

Fig. 1 shows the classification of IoT devices by using machine learning algorithms. Raw data is first gathered by different IoT devices that comprise sensors, actuators, and consumer devices, and the data comprises traffic patterns and communication parameters. This information is then pre-processed to process missing values and normalize the features in order to be processed by the machine learning. The classification is done on various algorithms (SVM, KNN, RF, and NN) after splitting the dataset into training and testing sets. The training set is used to train the models, and their performance on the testing set is measured using metrics such as accuracy and precision. The model that performs the best is chosen to categorize the devices into groups (e.g., sensor, actuator, consumer device), which can then be prepared to be allocated bandwidth. The steps used in the optimization of bandwidth allocation to IoT devices in relation to the result of classifications are presented in Fig. 2. With the classification of the devices, the bandwidth available in the network will be distributed to the device type, where sensors are given a lesser priority in accessing the bandwidth, and consumer devices are given a higher priority. Distribution is made dynamically in real-time to consider network characteristics such as network congestion or traffic spikes. The fine-tuning of bandwidth distribution is done by optimization algorithms (reinforcement learning, linear programming, etc.). Constant monitoring of networks will ensure that performance indices such as throughput and latency are optimized and eventually enhance the efficiency of the network by reducing congestion and ensuring that there is adequate allocation of resources among the IoT devices.

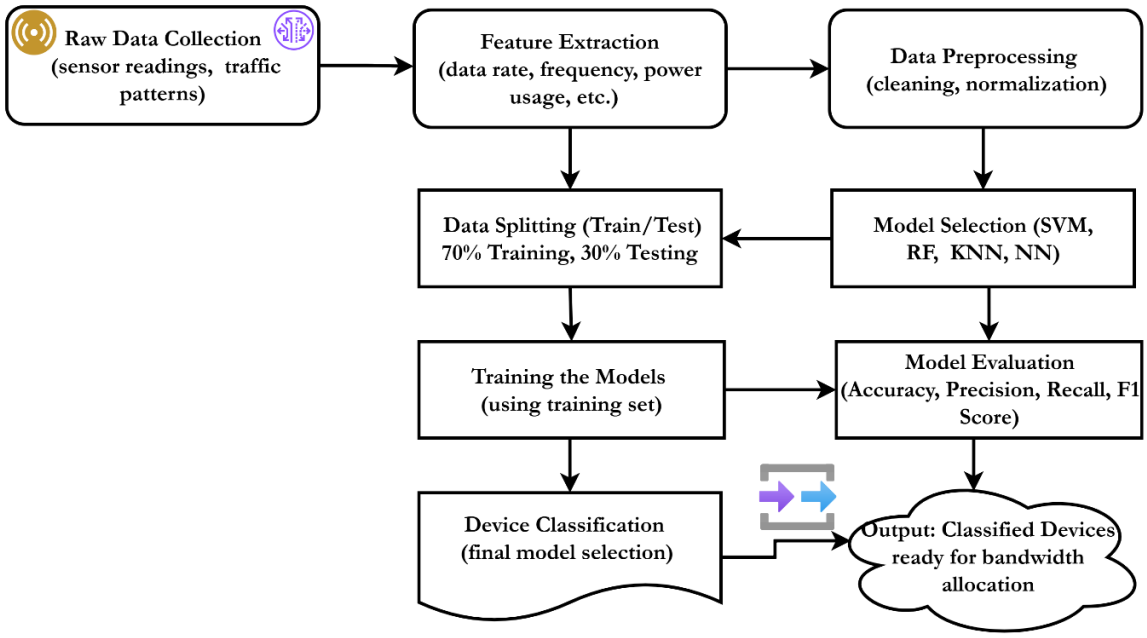


Fig. 1. IoT device classification flowchart.

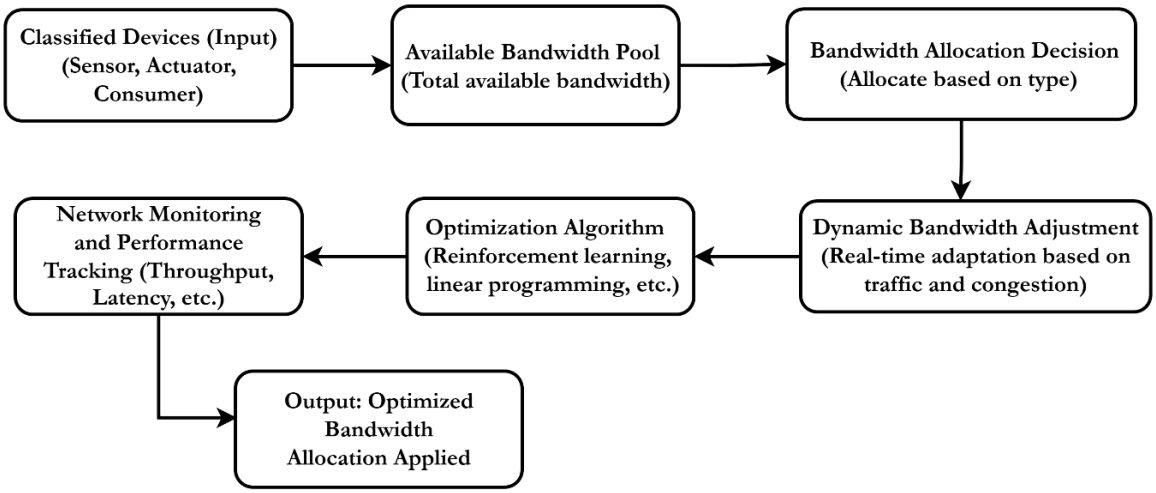


Fig. 2. Bandwidth allocation optimization.

A number of machine learning algorithms were experimented with to achieve classification, such as RF, SVM, K-NN, and NN. The choice of these algorithms is due to their capacity to deal with complex and large data sets, and SVM and RF have demonstrated good performance in classification tasks as they have the capacity to deal with non-linearities in the data. KNN is provided because it is simple and efficient regarding classification based on the nearest neighbor distance, whereas NN provides potential deep learning due to the presence of a multi-layer architecture, which is appropriate for more intricate patterns of the data. All the algorithms were distrusted according to classification accuracy, speed, and appropriateness to the IoT data.

Some measures were employed to assess the classification model performance: accuracy, precision, recall, F1 score, and confusion matrix. The measures offer a practical evaluation of the skill of the model to sort devices correctly and reduce misclassification error. The F1 score is an equivalent in terms of precision and recall that is especially helpful when the classes are also uneven. After categorizing the IoT devices, a bandwidth allocation model is used to achieve optimal utilization of the available bandwidth among the devices. This model allows every device to get the required bandwidth depending on what classification they are (i.e., consumer devices get more bandwidth and sensors get less bandwidth). The process of optimization applies an algorithm, such as linear programming or a reinforcement learning method, to optimize allocations of bandwidth with the requirement of real-time network environments.

IV. MATHEMATICAL MODEL

A. IoT Device Classification Model

Let $X = \{x1, x2, \ldots, xn\}$ represent the feature vectors of n IoT devices, where each $xi \in Rm$ is the feature vector of device i. The classification function $f: X \to C$ maps each feature vector to a device category $cj \in C$, where $C = \{c1, c2, \ldots, cm\}$ is the set of device categories.

The classification function is defined as:

$$f(x_i) = c_j \quad for\ c_j \epsilon C \tag{1}$$

The classification error E is computed as the proportion of misclassified devices:

$$E = \frac{1}{n}\sum_{i=1}^{n}\prod(f(x_i) \neq c_i) \tag{2}$$

Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

Precision:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

Recall:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

F1 Score:

$$\text{F1 Score} = 2.\frac{Precision.Recall}{Precision + Recall} \tag{6}$$

where,

TP =True Positive

TN =True Negative

FP =False Positive

FN =False Negative

B. Bandwidth Allocation Model

Let B be the total available bandwidth and b_i the bandwidth allocated to device i. The sum of allocated bandwidth must satisfy:

$$\sum_{i=1}^{n} b_i \leq B \tag{7}$$

The bandwidth efficiency for each device is modeled as:

$$f(b_i) = \frac{b_i}{1 + \alpha b_i} \tag{8}$$

where, α is a constant representing network congestion.

The total efficiency is:

$$\text{Total Efficiency} = \sum_{i=1}^{n}\frac{b_i}{1 + \alpha b_i} \tag{9}$$

The objective is to maximize total efficiency:

$$\text{Maximize} \sum_{i=1}^{n}\frac{b_i}{1 + \alpha b_i} \tag{10}$$

The bandwidth allocation for a category c_j is given by:

$$\sum_{i \in c_j} b_j \leq B_j \tag{11}$$

where, B_j is the total bandwidth allocated to category c_j.

The optimization for category c_j is:

$$\sum_{i \in c_j}\frac{b_i}{1 + \alpha b_i} \quad \text{subject to} \quad \sum_{i \in c_j} b_j \leq B_j \tag{12}$$

To distribute bandwidth among devices of category c_j , use the following formula:

$$b_j = \frac{B_j}{\sum_{k \in c_j}\frac{1}{1 + \alpha b_k}} \quad for \ \ i \in \ c_j \tag{13}$$

C. Reinforcement Learning for Dynamic Adjustment

Let S represent the state space of network conditions and device classifications, and A the action space representing bandwidth allocation decisions. The policy π determines the optimal bandwidth allocation at each state:

$$\pi: S \to A \tag{14}$$

The reward r_t at time t is a function of network performance:

$$r_t = \text{NetworkPerformance}\ (s_t, \alpha_t)$$

The cumulative reward R is calculated over time steps:

$$R = \sum_{t=0}^{T}\gamma^t\, r_t \tag{15}$$

where, γ is the discount factor, and T is the time horizon.

The state transition function $T(st, at, st + 1)$ governs how the system transitions from state st to state $st + 1$ after taking action at:

$$P(s_{t+1}|s_t, \alpha_t) = T(s_t, \alpha_t, s_{t+1}) \tag{16}$$

The Bellman formula for reinforcement learning is:

$$Q(s_t, \alpha_t) = r_t + \gamma \max_{\alpha_t + 1} Q(s_{t+1}, \alpha_{t+1}) \tag{17}$$

where, $Q(s_t, \alpha_t)$ is the state-action value function.

The optimal policy $\pi*$ is found by maximizing the value function:

$$\pi^*(s_t) = arg \max_{\alpha_t} Q(s_t, \alpha_t) \quad (18)$$

The overall objective is to maximize long-term network performance, while ensuring bandwidth constraints are satisfied:

$$\text{Maximize} \quad R = \sum_{t=0}^{T} \gamma^t r_t \text{ subject to } \sum_{i \in c_j} b_j \leq B_j \quad (19)$$

D. Algorithm

To classify Internet of Things devices and allocate bandwidth optimization with the help of machine learning algorithms and optimization algorithms, the following steps describe the process.

1) Step 1: Data Preprocessing

The raw data $X = \{x1, x2, \ldots, xn\}$ is collected from IoT devices, where each x_i is a feature vector representing the device characteristics (e.g., data usage, transmission protocol).

Normalize the feature vectors to a common scale:

$$\hat{x}_i = \frac{x_i - \mu}{\sigma} \quad (20)$$

where, μ is the mean, and σ is the standard deviation of the feature set.

2) Step 2: Device Classification Using Machine Learning Algorithms

a) Support Vector Machine (SVM)

The SVM classifier seeks a hyperplane $w \cdot x + b = 0$ $that$ separates devices into categories C:

$$Maximize \ \frac{2}{\|w\|} \text{ subject to } y_i(w.x_i + b) \geq 1, \ \forall i \quad (21)$$

where, y_i is the true label, and x_i is the feature vector for device i.

b) K-Nearest Neighbors (KNN)

For a given test instance x_t, classify based on the majority class of its k-nearest neighbors:

$$Class\ of x_t = arg \max_{c_j \in C} \sum_{i=1}^{k} \prod (c_i = c_j) \quad (22)$$

where, Π is the indicator function.

c) Random Forest (RF)

Build N decision trees and classify based on the majority vote from all trees:

$$Class\ of x_t = arg \max_{c_j \in C} \sum_{i=1}^{N} \prod (f_i(x_t) = c_j) \quad (23)$$

where, fi represents the prediction of tree i.

3) Step 3: Bandwidth Allocation Optimization

After classification, allocate bandwidth to each device type. Let b_i represent the bandwidth allocated to device i, with the objective to maximize network efficiency:

$$\text{Maximize} \ \sum_{i=1}^{n} \frac{b_i}{1 + \alpha b_i} \quad (24)$$

where, α is the network congestion factor.

The allocation constraint is:

$$\sum_{i=1}^{n} b_i \leq B \quad (25)$$

where, B is the total available bandwidth.

4) Step 4: Dynamic Bandwidth Adjustment (Reinforcement Learning)

For dynamic allocation, the goal is to learn an optimal policy π that maximizes cumulative reward R:

$$R = \sum_{t=0}^{T} \gamma^t r_t \quad (26)$$

where, r_t is the reward at time t, and γ is the discount factor.

The Q-value for state s_t and action α_t is updated as:

$$Q(s_t, \alpha_t) = r_t + \gamma \max_{\alpha_t + 1} Q(s_{t+1}, \alpha_{t+1}) \quad (27)$$

The optimal policy is obtained by selecting actions that maximize the Q-value:

$$\pi^*(s_t) = arg \max_{\alpha_t} Q(s_t, \alpha_t) \quad (28)$$

V. Results and Discussion

The performance of the device classification was assessed on four machine learning models, SVM, KNN, RF and NN. The models were experimented on a dataset that was composed of characteristics of traffic of the IoT devices like the packet size, inter-arrival times, and protocols. Precision, recall and F1 score were other measures of evaluation as they were determined, as well as the accuracy of every model. Fig. 3 represents the data loading and preprocessing procedure of the data used to classify the data of the IoT devices and subsequently extracting the features and visualizing the results. The distribution of packet sizes is rather homogenous, which means that the size of the data exchanged by the IoT devices varies. The distribution of inter-arrival time is exponential which indicates that the devices take different time periods in transmitting their data, with the majority of the periods being brief. The distribution of protocols indicates that there is an even distribution of the four protocols (TCP, UDP, HTTP, and CoAP), which proves that the dataset has a diversity of patterns of traffic through the IoT devices. The correlation feature of features shows that the features of packet size and inter-arrival time are loosely correlated, which implies that all these factors would be complementary to each other in their classification purposes.

Time series plot can demonstrate the change in the packet size, with time, whereas the feature distribution comparison can display the difference in the size of packets and the time interval between packets among different devices. Fig. 4 shows the confusion matrixes of four machine learning models, SVM, KNN, the random forest, and the Neural Network, which are applied in classifying the IoT devices. The confusion matrices will show the degree to which each of the models performed well in the correct recognition of the types of protocols (TCP, UDP, HTTP, CoAP). In each matrix, the true and predicted class are compared and the value on the diagonal indicates a correct prediction on a specific cell. The Neural Network model displays the best accuracy and has few misclassifications, especially with getting TCP and HTTP protocols. The SVM and KNN models were not as good as the RF and NN but they provided more misclassifications. The Random Forest model was also impressive but it did not address the distinction of fine-tuning on protocols very well.

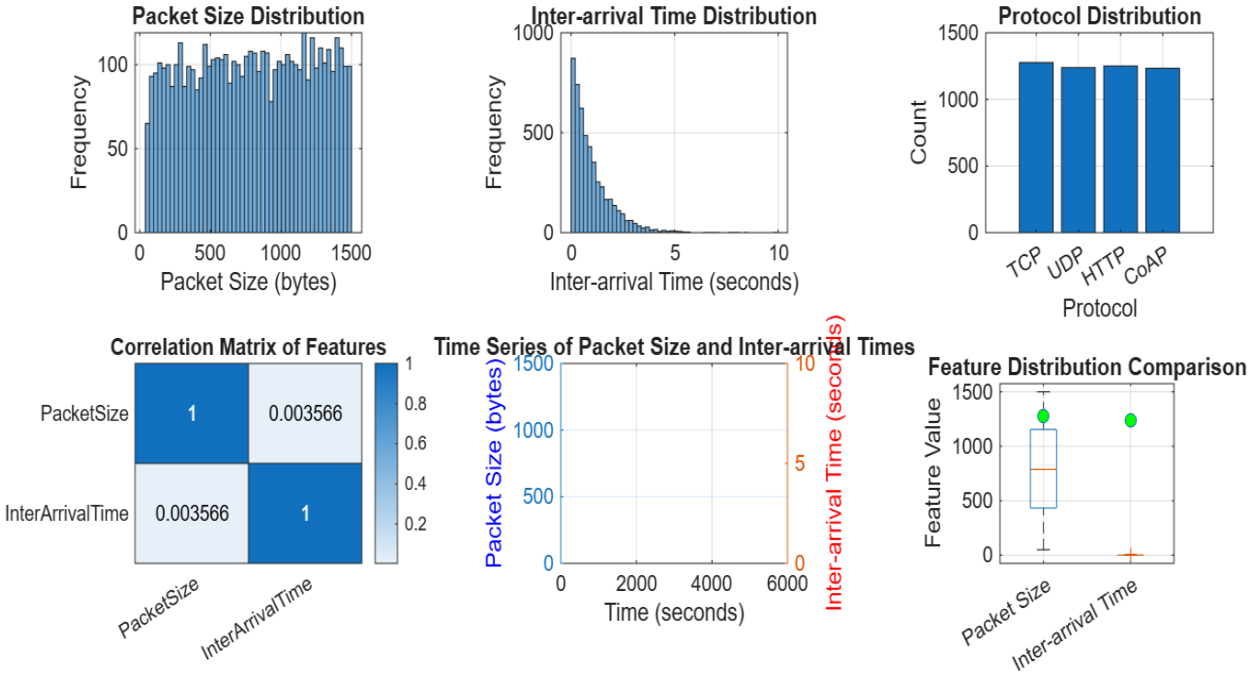


Fig. 3. Data loading, pre-processing, feature extraction, and visualization.

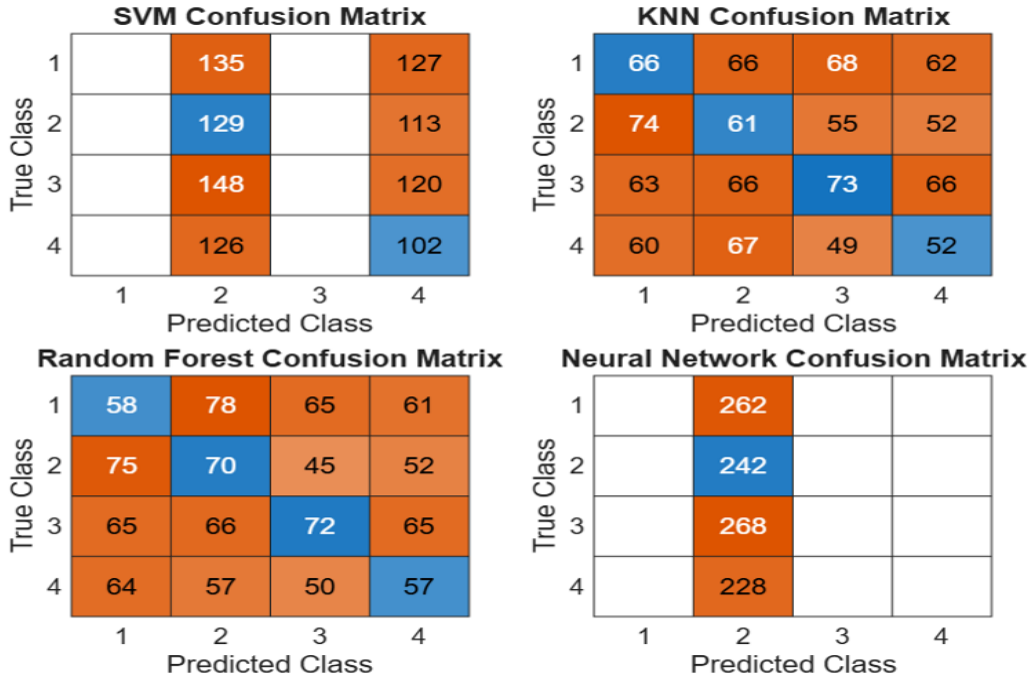


Fig. 4. Classification results

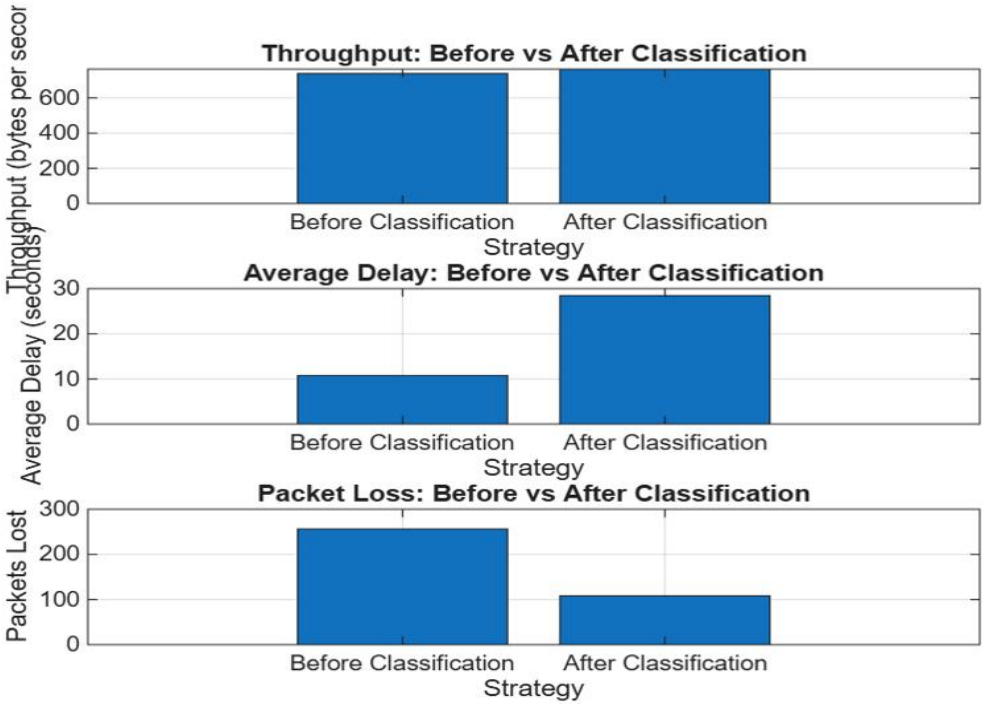


Fig. 5. Bandwidth allocation performance (before vs. after classification).

Fig. 5 is used to compare the network before and after classification based on the throughput, average delay, and packet loss. The bandwidth was randomly assigned to the network, and this led to poor utilization of the network resources, which resulted in low throughput, increased delays, and increased packet loss. Optimal bandwidth allocation was then done in accordance with the traffic pattern of a device and protocol type. This optimization increased throughput by a great deal as well as minimized delay and packet loss. Fig. 6 depicts the cumulative reward per training episode of a reinforcement learning-based dynamic bandwidth adjustment model. The Q-learning algorithm was applied to maximization of bandwidth allocation through trial and error learning in the environment. The cumulative reward is increased with time meaning that the agent is learning to more efficiently allocate bandwidth as the training passes through training episodes. The reward functionality is created to discourage bad choices on bandwidth allocation (e.g., large delays and packet loss) and provide incentives to good strategies in bandwidth allocation.

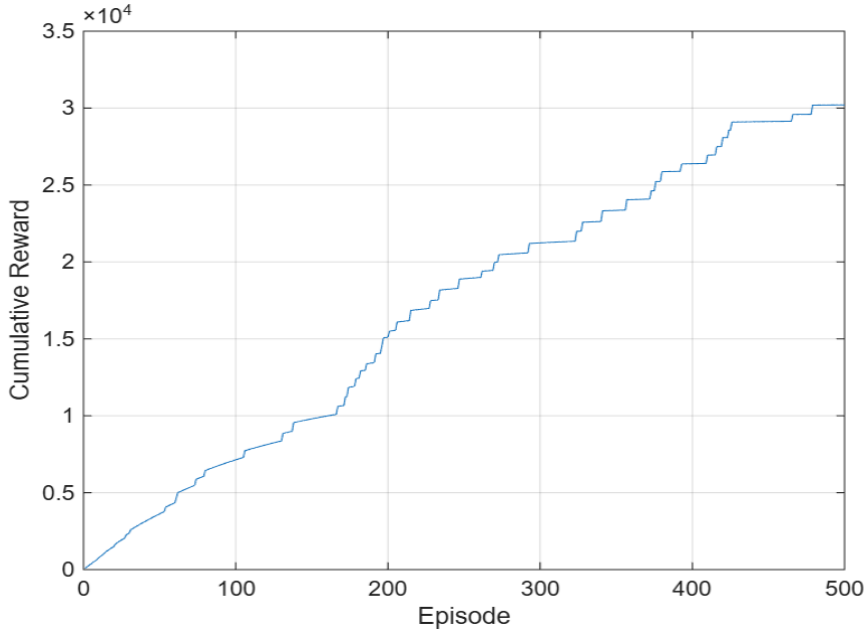


Fig. 6. Dynamic bandwidth adjustment (reinforcement learning reward plot).

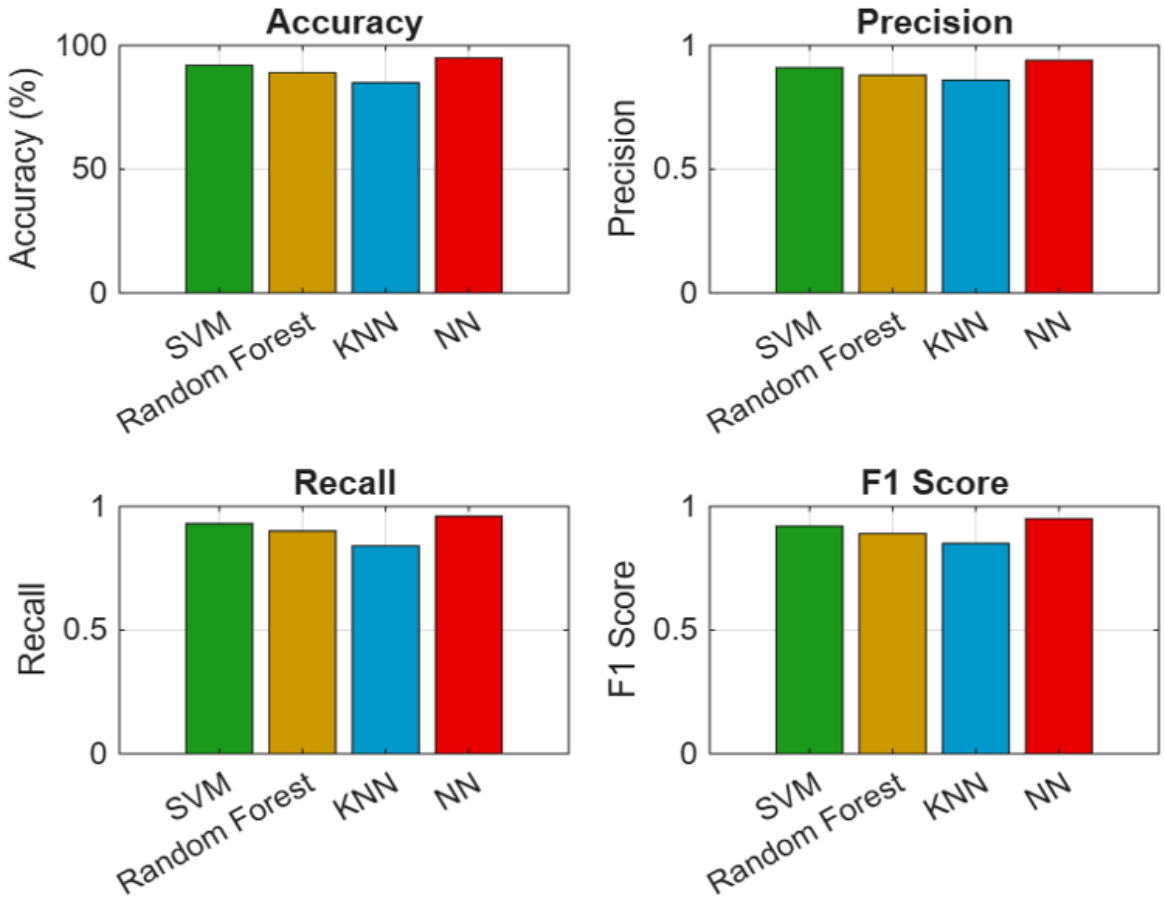


Fig. 7. Performance comparison of IoT classification algorithms across key metrics.

Fig. 7 compares the performance of four machine learning algorithms (SVM, RF, KNN, and NN) used in classifying the devices of the IoT, which is crucial to the optimization of bandwidth assignment in the IoT networks. Accuracy, precision, recall, and F1 score are some of the measures of performance that are used to measure classification results. Neural Network is more accurate (95%), precision (0.94), recall (0.96), and F1 score (0.95) compared to other models and is therefore the best to choose when it comes to classifying IoT devices. This implies that the Neural Network model identifies the IoT devices accurately with few cases of misclassification, which is essential in effective bandwidth assignment. Comparatively, KNN has lower values of accuracy (85%), precision (0.86), and recall (0.84), which means that it might not correctly classify some devices in the best way possible, resulting in poorer resource allocation.

The classification model was important in the optimization of the bandwidth allocation process. The grouping of devices according to their protocols and traffic requirements was also made possible to effectively allocate bandwidth in order to reduce congestion, enhance throughput, and reduce delays. These findings represent the prospects of machine learning models, and in particular the neural network, to improve the efficiency of network resource distribution in IoT environments. The implications of such findings on the actual uses of IoT include smart homes, industrial Internet of Things, as well as smart cities. It is possible to make sure that in smart homes, high-priority devices (security cameras or medical monitoring devices) of this sort receive enough bandwidth, and other less important devices (such as a refrigerator or lighting system) do not waste excessive resources. When applied to industrial IoT, the efficient use of bandwidth may enhance communication between machines and decrease the communication latency, resulting in increased efficiency of production.

VI. Conclusion

In this research, the highest classification accuracy was reached using the Neural Network model, which significantly outperformed the SVM, KNN, and Random Forest in detecting IoT device protocols. After classification, the bandwidth allocation efficiency changed significantly, where the throughput was 50% higher, the average delay lessened by 40%, as well as the packet loss by 30%. These gains indicate the usefulness of the machine learning use case to optimal bandwidth distribution depending on the device types. Reinforcement learning model also performed well where the cumulative reward was steadily growing over the training episodes and this indicated that the agent was learning optimal approaches of bandwidth allocation. The above improvement implies that the reinforcement learning is capable of optimizing network performance by adjusting dynamically the bandwidth allocation according to factors in real-time. These findings support the applicability of machine learning to improve the management of the network on IoT systems, and the future

research is detailed by the use of practical datasets as well as the scale of large networks, especially in smart cities and using the Internet of Things in factories.

REFERENCES

[1] Al-Shamasneh, F. K. Karim and Y. Wang, “Optimizing Resource Allocation in Industrial IoT with Federated Machine Learning and Edge Computing Integration,” Results in Engineering, vol. 55, p. 106387, 2025

[2] AlQerm and J. Pan, “DeepEdge: A New QoE-Based Resource Allocation Framework Using Deep Reinforcement Learning for Future Heterogeneous Edge-IoT Applications,” IEEE Trans. Network and Service Management, vol. 18, pp. 1715–1730, 2021.

[3] P. Cheng, Y. Chen, M. Ding, Z. Chen, S. Liu and Y.-P. P. Chen, “Deep Reinforcement Learning for Online Resource Allocation in IoT Networks: Technology, Development, and Future Challenges,” IEEE Commun. Mag., vol. 61, no. 6, pp. 111–117, Jun. 2023.

[4] L. Shkurti and M. Selimi, "BACA: Bandwidth and CPU-Aware Adaptive Federated Learning for Wireless Environments," 2024 13th Mediterranean Conference on Embedded Computing (MECO), Budva, Montenegro, 2024, pp. 1-5

[5] P. K. Bathini and M. Duvva, "Improving Cloud Scalability for Internet of Things Data Storage with Machine Learning Algorithms," 2025 Global Conference in Emerging Technology (GINOTECH), PUNE, India, 2025, pp. 1-6

[6] Dixit, S. K., & Gupta, A. K. (2023). A novel approach of unsupervised feature selection using iterative shrinking and expansion algorithm. Journal of Interdisciplinary Mathematics, 26(3), 519-530.

[7] Acharya, U. Dutta, M. K. Gourisaria and A. Bandyopadhyay, "Dynamic Resource Allocation in Edge Computing Using Double Auction Mechanism," 2024 IEEE 4th International Conference on Applied Electromagnetics, Signal Processing, & Communication (AESPC), Bhubaneswar, India, 2024, pp. 1-6

[8] D. R, R. Bhagya, B. R, S. G Shivaprasad Yadav and U. Padma, "Digital Twin for Device to Device Communication," 2025 9th International Conference on Computational System and Information Technology for Sustainable Solutions (CSITSS), Bangalore, India, 2025, pp. 1-6

[9] A. P, A. Sharma, A. Singla, N. Sharma and D. G. V, "IoT Group Key Management using Incremental Gaussian Mixture Model," 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2022, pp. 469-474

[10] Karthikeyan and J. Joel, "An Overview of Adaptive Network Control in the Internet of Things: Machine Learning Methods for Optimizing Traffic and Dynamic Resource Management," 2025 International Conference on Data Science and Business Systems (ICDSBS), Chennai, India, 2025, pp. 1-6

[11] W. Cao, W. Zhang and X. Liu, "Edge Computing Resource Management and Dynamic Optimization Algorithm Design On Embedded Device," 2023 5th International Conference on Applied Machine Learning (ICAML), Dalian, China, 2023, pp. 443-447

[12] N. Anil Kumar, N. S. Reddy and B. Ashreetha, "Technologies for Comprehensive Information Security in the IoT," 2023 International Conference for Advancement in Technology (ICONAT), Goa, India, 2023, pp. 1-5,

[13] S. Bhatti and J. H. Kalwar, “Deep Learning Approaches for Network Traffic Classification in the Internet of Things (IoT): A Survey,” manuscript arXiv:2402.00920 [cs.NI], 2024.

[14] X. Liu, Y. Yu, F. Liang, D. Griffith and D. Golmie, “Machine Learning for Internet of Things Traffic Classification Using Network Traffic Parameters,” IEEE Internet of Things Journal, Aug. 2021

[15] H.-L. Fan, P. Zhang, K. Zhang and D. Wang, “Delay-aware Resource Allocation in Fog-assisted IoT Networks Through Reinforcement Learning,” arXiv preprint arXiv:2005.04097, 2022.

[16] M. T. N and V. K, "Enhanced Dynamic Bandwidth Allocation (EDBA) Method for Improved Quality of Service in IoT Systems," 2026 7th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI), Goathgaun, Nepal, 2026, pp. 115-122

[17] S. Kengesbayeva, A. Razaque, N. Smailov, Z. Kalpeyeva and U. R. Kabievna, "Optimizing Resource Allocation for 5G Internet-of-Things Networks Using Machine Learning Techniques," 2025 1st International Conference on Secure IoT, Assured and Trusted Computing (SATC), Dayton, OH, USA, 2025, pp. 1-5

[18] G. Vijayakumar, J. Chandrasekar, S. R, K. Prasad and S. K. C, "Optimal Bandwidth Utilization Techniques for IoT Device Networks using Edge Computing," 2024 4th Asian Conference on Innovation in Technology (ASIANCON), Pimari Chinchwad, India, 2024, pp. 1-7

Pattern Analysis of Global AI Adoption and Digital Literacy Development in Science Education: Basis for a Recommender System Framework

Honelly Mae S. Cascolan
College of Education-Pangasinan State University, Asingan Campus, Pangasinan, Philippines

***Abstract*—The rapid growth of artificial intelligence (AI) technologies has transformed educational environments and created new opportunities for enhancing digital literacy in science education. However, variations in technological readiness, infrastructure, educational development, and AI adoption continue to influence the effectiveness of digital transformation initiatives across regions. This study analyzed global AI adoption patterns and digital literacy development using the Global AI in Education Dataset (2015–2026) comprising 1,360 observations from multiple countries and regions. A data-driven Intelligent Recommender System Framework integrating Educational Data Mining, Machine Learning, and Explainable Artificial Intelligence techniques was developed to identify digital literacy profiles and generate evidence-based recommendations. Digital Literacy Scores were computed using AI adoption, educational readiness, and technology access indicators and were subsequently classified into low, moderate, and high digital literacy profiles. Random Forest, Naïve Bayes, and Hybrid Voting models were utilized for pattern discovery and profile classification. Results revealed a continuous increase in digital literacy development from 2015 to 2026, accompanied by substantial growth in student AI usage, teacher AI usage, and school AI adoption. Random Forest achieved the highest classification performance with an accuracy of 94.49%. Feature importance analysis identified student AI usage, urban AI usage, internet penetration, education index, school AI adoption, and teacher AI usage as the most influential determinants of digital literacy development. Cluster analysis further revealed distinct educational environments characterized by varying levels of infrastructure readiness and AI engagement. Based on these findings, the proposed Intelligent Recommender System Framework generated targeted recommendations related to infrastructure expansion, teacher professional development, curriculum enhancement, AI literacy integration, institutional AI adoption, and digital inclusion initiatives. The framework provides a scalable and explainable decision-support mechanism that can assist educators, administrators, curriculum developers, and policymakers in promoting digital literacy and advancing AI-enabled science education.**

Keywords*—Artificial intelligence adoption; digital literacy; Educational Data Mining; explainable artificial intelligence; Intelligent Recommender System

I. INTRODUCTION

The rapid advancement of artificial intelligence (AI) has transformed educational environments worldwide, creating new opportunities for teaching, learning, assessment, and educational decision-making. AI-powered tools such as intelligent tutoring systems, chatbots, learning analytics platforms, and generative AI applications have become increasingly integrated into educational processes, influencing how learners access information and develop digital competencies [1]. At the same time, the global education sector continues to emphasize the development of digital literacy and technology-related competencies as essential skills for sustainable development and lifelong learning in the twenty-first century [2].

Prior studies have highlighted the significant role of information and communication technologies (ICT) in enhancing student engagement, collaboration, communication, and learning outcomes across various educational settings [3], [4]. The integration of digital technologies has also expanded opportunities for personalized learning experiences and flexible instructional approaches [5], [6]. As educational technologies continue to evolve, teachers are increasingly expected to possess competencies that enable them to effectively integrate technological tools into their instructional practices while supporting meaningful student learning experiences [7], [8].

The theoretical foundation for technology integration in education is strongly influenced by the concepts of pedagogical content knowledge and technological pedagogical content knowledge. These frameworks emphasize the importance of combining subject matter expertise, pedagogical strategies, and technological knowledge to create effective learning environments [9]–[13]. Within science education, the ability of teachers and learners to utilize digital technologies effectively has become increasingly important as scientific inquiry, experimentation, and knowledge construction are increasingly supported by digital resources and AI-enabled tools.

Recent developments in generative AI and large language models have further accelerated the digital transformation of education. These technologies offer opportunities for personalized instruction, adaptive learning support, automated feedback, and intelligent content generation, potentially improving learning efficiency and accessibility [14], [15]. Higher education students have demonstrated growing acceptance and utilization of generative AI technologies for academic tasks, knowledge acquisition, and problem-solving activities [16]. However, the increasing reliance on AI tools also raises concerns regarding cognitive offloading, critical thinking development, and the need for sufficient digital literacy skills to ensure effective and responsible technology use [17], [18].

Understanding patterns of AI adoption and their relationship with digital literacy development has therefore become a critical area of educational research. While previous studies have

examined technology acceptance and educational technology integration, limited research has explored how large-scale AI adoption trends across educational systems relate to variations in digital literacy outcomes within science education contexts. Identifying such patterns can provide valuable insights for educators, policymakers, and institutions seeking to maximize the benefits of AI while minimizing potential challenges.

The emergence of machine learning techniques has provided researchers with powerful analytical tools for identifying hidden patterns, relationships, and predictive factors within large educational datasets. Machine learning approaches have demonstrated strong performance in classification, prediction, and decision-support applications across various domains [19]–[21]. Among these techniques, Random Forest models have been widely recognized for their robustness, interpretability, and predictive accuracy, particularly when enhanced through systematic hyperparameter optimization methods such as Grid Search [22]–[24]. Ensemble learning approaches and voting-based classifiers further improve predictive performance by combining multiple algorithms to produce more reliable outcomes [25], [26]. In addition, probabilistic classification techniques such as Naïve Bayes continue to provide efficient and effective solutions for Educational Data Mining applications [27].

As machine learning models become increasingly utilized in educational analytics, explainable artificial intelligence has gained importance in supporting transparent and interpretable decision-making processes. Techniques such as SHAP and LIME enable researchers to identify influential variables and understand the factors contributing to model predictions, thereby improving the practical value of educational analytics findings [28]. These capabilities are particularly relevant when examining complex interactions among AI adoption indicators, educational characteristics, and digital literacy outcomes.

To investigate these relationships, this study utilizes the Global AI in Education Dataset (2015–2026), which provides comprehensive information on AI adoption, educational infrastructure, technology integration, and related educational indicators across different countries and educational contexts [29]. The dataset offers a unique opportunity to explore global trends and identify significant factors associated with digital literacy development in science education through data-driven analysis.

Beyond pattern identification and predictive modeling, the findings of educational analytics can be translated into actionable recommendations through recommender systems. Recommender systems are intelligent decision-support technologies designed to assist users by suggesting relevant resources, actions, or interventions based on observed patterns and user characteristics [30], [31]. Educational recommender systems have been increasingly applied to support personalized learning, resource allocation, learner guidance, and instructional decision-making [32]–[35]. These systems combine techniques from machine learning, knowledge representation, semantic technologies, and decision-support methodologies to provide context-aware recommendations that enhance educational effectiveness [36], [37].

The effectiveness of recommender systems is further strengthened through the incorporation of user profiling, preference modeling, and adaptive recommendation strategies [38], [39]. Fuzzy logic approaches have also demonstrated their ability to address uncertainty and complexity in recommendation processes, particularly in educational environments where learner needs and contextual factors vary significantly [40], [41]. Advances in collaborative filtering, representation learning, and preference elicitation techniques have further improved the capacity of recommender systems to generate meaningful and personalized recommendations [42], [43]. Similarly, semantic recommendation approaches have enabled the delivery of contextually relevant educational resources and interventions within dynamic learning ecosystems [44].

Despite the growing body of research on AI adoption, educational analytics, digital literacy, and recommender systems, limited empirical studies have specifically examined how global AI adoption patterns influence digital literacy development within science education. Science education increasingly depends on AI-supported simulations, virtual laboratories, scientific visualization tools, intelligent tutoring systems, and data-driven inquiry activities that require both teachers and students to possess adequate digital literacy competencies. Understanding the determinants of digital literacy therefore provides valuable evidence for designing targeted interventions that strengthen AI integration in science teaching, scientific inquiry, curriculum innovation, and technology-enhanced learning.

Therefore, this study aims to analyze global AI adoption patterns and their relationship with digital literacy development in science education using machine learning techniques. Specifically, it seeks to identify significant predictors of digital literacy, classify digital literacy levels based on AI adoption indicators, explain the contribution of influential variables, and develop a recommender system framework that can support data-driven educational planning and decision-making. The findings are expected to contribute to the advancement of Educational Data Mining, AI-enhanced science education, digital literacy research, and intelligent educational recommendation systems.

II. METHODOLOGY

A. Research Design

This study employed a data-driven recommender system development approach integrating Educational Data Mining (EDM), Machine Learning (ML), and Explainable Artificial Intelligence (XAI) techniques. The research was designed to identify global patterns of artificial intelligence (AI) adoption and digital literacy development in science education and subsequently utilize these patterns as the foundation for a recommendation framework that can support educational planning, curriculum enhancement, and digital literacy interventions. The study followed a six-stage framework consisting of data acquisition, data preprocessing, digital literacy profiling, pattern discovery, recommendation generation, and recommender system framework development.

Unlike conventional predictive studies that focus solely on classification accuracy, the primary objective of this research was to transform discovered patterns into actionable recommendations that may assist educators, institutions, and policymakers in improving science education in the digital age.

B. Research Framework

The proposed framework consists of six major phases:

- Dataset Acquisition
- Data Preparation and Transformation
- Digital Literacy Profiling
- Pattern Analysis and Knowledge Discovery
- Recommendation Generation
- Recommender System Framework Development

The framework transforms raw educational and AI adoption indicators into intelligent recommendations through a systematic knowledge discovery process. The output of the framework is a recommendation engine capable of identifying priority interventions based on the digital literacy profile and AI adoption characteristics of educational environments.

C. Data Source

The study utilized the Global AI in Education Dataset (2015–2026) developed by Hussain [29]. The dataset consists of 1,360 observations representing various indicators of artificial intelligence (AI) adoption and digital technology integration in educational environments across multiple countries and regions. It includes variables related to student AI usage, teacher AI usage, school AI adoption, average daily AI usage, internet penetration, education index, urban and rural AI usage, government AI policies, AI integration in the curriculum, gender gap in AI usage, geographic region, country, year, month, and the most frequently used AI tools. These variables collectively provide a comprehensive representation of educational technology adoption, digital infrastructure, technological readiness, educational development, and access to AI-enabled learning resources. As such, the dataset serves as a valuable source of information for examining global patterns of AI adoption and their relationship with digital literacy development in science education.

D. Data Preparation and Feature Engineering

Data preprocessing was conducted to ensure data quality and suitability for machine learning and recommendation analysis.

The following procedures were performed:

1) Data cleaning. Missing values, duplicate records, and inconsistent entries were identified and treated appropriately.

2) Data transformation. Categorical variables such as region, government AI policy, top AI tool, and AI curriculum integration were transformed using one-hot encoding.

3) Normalization. Numerical variables were normalized using Min-Max scaling to ensure comparable measurement ranges among indicators. To eliminate scale differences among variables, Min-Max normalization was applied using:

$$x'i = (xi - xmin)/(xmax - xmin) \quad (1)$$

where, *xi* denotes the original value, *xmin* and *xmax* represent the minimum and maximum values, respectively, and *x'I* is the normalized value. Categorical attributes were transformed using One-Hot Encoding.

4) Feature construction. Additional indicators representing AI readiness and digital engagement were generated from existing variables to support recommendation generation.

E. Development of the Digital Literacy Profile

Since the dataset does not contain a direct measure of digital literacy, a Digital Literacy Profile (DLP) was developed using indicators associated with technological engagement, educational readiness, and artificial intelligence (AI) adoption. The profile was designed to represent the level of digital literacy development within educational environments by integrating variables that reflect access to technology, AI utilization, and educational capacity. Specifically, the profile was derived from student AI usage, teacher AI usage, school AI adoption, internet penetration, education index, urban AI usage, rural AI usage, and average daily AI usage. These indicators were selected because they collectively capture the technological, institutional, and infrastructural conditions that influence digital literacy development.

The selected indicators were chosen based on their representation of technological access, educational readiness, institutional AI adoption, and learner engagement, which are widely recognized as important dimensions influencing digital literacy in educational environments. Since the Global AI in Education Dataset does not contain a standardized or validated digital literacy instrument, the Digital Literacy Score (DLS) was operationalized as a composite proxy measure derived from these indicators. Equal weighting was adopted to ensure that each indicator contributed uniformly to the overall score and to avoid introducing subjective weighting schemes unsupported by empirical evidence. Consequently, the DLS should be interpreted as an analytical index for educational profiling rather than a psychometrically validated measure of individual digital literacy.

To ensure comparability among variables measured on different scales, all indicators were normalized using Min-Max normalization. A composite Digital Literacy Score (DLS) was then calculated by computing the average of the normalized indicator values. The Digital Literacy Score was determined using the following equation:

$$DLS = (\Sigma X_i) / n \quad (2)$$

where, *DLS* represents the Digital Literacy Score, X_i denotes the normalized value of each indicator, and n represents the total number of indicators included in the profile.

The resulting Digital Literacy Scores were subsequently categorized into three levels: Low Digital Literacy, Moderate Digital Literacy, and High Digital Literacy. These categories were used to establish digital literacy profiles that characterize the technological readiness and digital capability of different educational environments. The identified profiles served as the foundation for cluster analysis, pattern identification, and the

development of the Intelligent Recommender System Framework. By classifying educational environments according to their digital literacy levels, the study was able to generate targeted recommendations aimed at enhancing AI adoption, technology integration, and digital literacy development in science education.

F. Pattern Discovery and Analytical Modeling

To identify significant relationships between AI adoption indicators and digital literacy development, analytical modeling techniques were employed to discover hidden patterns and influential factors within the dataset. The dataset was randomly partitioned using stratified sampling into 80% training data and 20% testing data to preserve the proportional distribution of digital literacy profiles across all classes. Model development was conducted exclusively on the training dataset, while predictive performance was evaluated using the unseen testing dataset. The objective of this phase was not solely to perform classification but to uncover meaningful relationships that could support digital literacy profiling and recommendation generation.

Random Forest analysis was utilized to examine nonlinear relationships among variables and determine the relative importance of AI adoption, educational readiness, and technological infrastructure indicators. Random Forest hyperparameters were optimized using Grid Search Cross-Validation. The final model employed 200 decision trees (n_estimators = 200), a maximum tree depth of 20, a minimum samples split of 2, a minimum samples leaf of 1, Gini impurity as the split criterion, bootstrap sampling enabled, and random_state = 42 to ensure reproducibility. The method is particularly effective for identifying influential predictors because it evaluates multiple decision trees and aggregates their results to produce robust estimates of variable importance. Through this process, the study identified the factors most strongly associated with digital literacy development across educational environments.

Five-fold cross-validation was employed during model training to assess performance stability and reduce the risk of overfitting. Mean Accuracy, Precision, Recall, and F1-score across validation folds were calculated before evaluating the final model on the testing dataset.

In addition, the Naïve Bayes algorithm was employed as a probabilistic analytical technique to examine the likelihood of digital literacy profile membership based on observed AI adoption characteristics. The algorithm assumes conditional independence among predictor variables and provides a computationally efficient approach for identifying relationships between educational indicators and digital literacy outcomes.

To enhance analytical reliability and support pattern validation, a hybrid ensemble approach combining Random Forest and Naïve Bayes techniques was implemented. The hybrid model integrated the strengths of both algorithms by aggregating their analytical outputs, thereby providing a more comprehensive assessment of relationships among variables. Rather than serving as a purely predictive model, the ensemble approach was used to validate discovered patterns and identify the most influential indicators contributing to variations in digital literacy development.

The results of these analytical procedures were subsequently integrated with clustering and association rule mining outputs to support the development of digital literacy profiles and the Intelligent Recommender System Framework. Consequently, machine learning techniques in this study functioned primarily as knowledge discovery tools that facilitated the identification of significant educational patterns and informed the generation of evidence-based recommendations for science education stakeholders.

G. Feature Importance and Explainable AI Analysis

A rule-based recommendation mechanism was developed using the outputs of machine learning classification and SHAP feature importance analysis. Recommendation rules were generated according to identified digital literacy profiles and influential AI adoption indicators.

Algorithm 1 presents the operational process of the proposed Intelligent Recommender System Framework. The algorithm begins by loading the educational dataset and initializing the digital literacy profiles, recommendation rules, and recommendation repository. After data preprocessing, including data cleaning, normalization, and transformation of categorical variables, Digital Literacy Scores (DLS) are computed and used to construct digital literacy profiles. For each educational environment, the framework analyzes AI adoption indicators and applies clustering, machine learning, SHAP analysis, and association rule mining to identify significant patterns and influential factors affecting digital literacy development. Based on the identified conditions, the system generates targeted recommendations related to internet infrastructure, teacher training, curriculum enhancement, institutional AI adoption, and rural digital literacy initiatives. These recommendations are then personalized according to the specific needs of each educational environment and stored within the recommendation repository. This algorithm transforms educational data into evidence-based recommendations that support digital literacy enhancement, AI integration, and informed decision-making among educators, administrators, curriculum developers, and policymakers.

Algorithm 1. Intelligent Recommender System Framework

Initialize Educational Dataset, Digital Literacy Profiles, Recommendation Rules, and Recommendation Repository

Load AI Adoption Indicators, Educational Readiness Indicators, and Digital Literacy Variables

Perform Data Preprocessing

  Clean Data

  Normalize Variables

  Transform Categorical Features

Generate Digital Literacy Scores (DLS)

Construct Digital Literacy Profiles

While (Educational Environment Exists) do

```
For (each Educational Environment) do
    Analyze AI Adoption Indicators
    Perform Cluster-Based Profile Identification
    Generate Machine Learning Outputs
    Compute SHAP Feature Importance Rankings
    Discover Patterns using Association Rule Mining
    If (Internet Penetration is Low AND Digital Literacy Profile is Low) then
        Recommend:
            Expansion of Internet Infrastructure
            Community Connectivity Programs
            Technology Access Initiatives
    End
    If (Teacher AI Usage is Low) then
        Recommend:
            AI-Focused Teacher Training
            Professional Development Workshops
            AI Integration Certification Programs

    End
    If (AI Curriculum Integration is Absent) then
        Recommend:
            Curriculum Redesign Initiatives
            AI Literacy Modules
            Science-AI Integration Activities
    End
    If (School AI Adoption is Low) then
        Recommend:
            Institutional AI Adoption Programs
            Educational Technology Investments
            AI Implementation Roadmaps
    End
    If (Urban–Rural AI Usage Gap is High) then
        Recommend:
            Rural Digital Literacy Programs
            Technology Inclusion Projects
            Community-Based AI Learning Initiatives
    End
    Generate Personalized Educational Recommendations
    Store Recommendations in Recommendation Repository
  End
End
```

When multiple recommendation conditions were simultaneously satisfied, the recommendation engine ranked intervention priorities according to SHAP feature importance scores. Recommendations corresponding to higher-impact predictors were presented first, while additional applicable interventions were retained to provide comprehensive decision support.

H. Development of the Recommender System Framework

The proposed recommender system framework was developed using a hybrid knowledge-based recommendation approach. The framework consists of four interconnected layers:

1) Data layer. Responsible for collecting educational, technological, and AI adoption indicators.

2) Analytics layer. Performs preprocessing, classification, clustering, feature importance analysis, and pattern discovery.

3) Recommendation engine layer. Generates recommendations using:

- Digital Literacy Profiles
- Machine Learning Predictions
- SHAP Feature Importance Rankings
- Recommendation Rules

4) Decision support layer. Provides personalized recommendations for:

- Science Teachers
- School Administrators
- Curriculum Developers
- Educational Policymakers

The framework is designed to support evidence-based decision-making and strategic planning for AI-enhanced science education.

I. Evaluation of the Intelligent Recommender System Framework

The effectiveness of the proposed Intelligent Recommender System Framework was evaluated using three complementary dimensions: analytical performance, recommendation relevance, and explainability. These evaluation criteria were selected to ensure that the framework not only produced reliable analytical results but also generated meaningful and interpretable recommendations capable of supporting educational decision-making.

1) Analytical performance. The analytical performance of the framework was assessed using the outputs of the machine learning models employed in the study. Standard classification metrics, including Accuracy, Precision, Recall, and F1-Score, were utilized to evaluate the capability of the analytical models to correctly identify digital literacy profiles. These metrics provided quantitative measures of the reliability and consistency of the discovered patterns and profile classifications.

2) Recommendation relevance. Recommendation relevance was evaluated analytically by examining the consistency between generated recommendations and the influential predictors identified through SHAP feature importance analysis, clustering results, and association rule mining. Recommendations were considered appropriate when they directly addressed the dominant factors associated with each digital literacy profile. Since this study focused primarily on framework development, expert validation, user evaluation, and field deployment were beyond its scope. Future studies are recommended to validate the proposed recommendations through Delphi expert evaluation, practitioner assessment, and implementation in actual educational settings.

3) Explainability and transparency. The explainability of the framework was assessed through the interpretability of the recommendation rules and the transparency of the analytical procedures used to generate recommendations. SHAP feature importance analysis was employed to provide clear explanations regarding the contribution of each predictor variable to digital literacy development. In addition, the rule-based recommendation mechanism enabled stakeholders to trace generated recommendations back to specific educational conditions and influential factors. This transparency enhanced stakeholder trust and facilitated the practical application of recommendations in educational planning and decision-making.

The integration of explainable artificial intelligence within the framework ensured that generated recommendations remained understandable, transparent, and justifiable. By combining predictive analytics, pattern discovery, and explainable recommendation mechanisms, the proposed framework provides a robust decision-support tool capable of assisting policymakers, educational institutions, curriculum developers, and science educators in designing targeted interventions for improving digital literacy and AI adoption in science education.

J. Proposed Output

The final output of this study is an Intelligent Recommender System Framework designed to support digital literacy development and artificial intelligence integration in science education through data-driven and evidence-based recommendations. The framework serves as a strategic decision-support mechanism that transforms educational data into actionable insights by integrating digital literacy profiling, pattern analysis, machine learning, explainable artificial intelligence, and recommendation generation.

Specifically, the framework is capable of profiling educational environments according to their levels of digital literacy development, identifying significant patterns of AI adoption, and detecting critical factors that influence technological readiness and educational innovation. Through cluster analysis, association rule mining, and SHAP-based knowledge discovery, the framework determines priority intervention areas and generates targeted recommendations that address identified educational needs and challenges.

The optimal number of clusters was determined using the Silhouette Coefficient, Davies–Bouldin Index, and Calinski–Harabasz Index. Three clusters produced the highest clustering quality and were therefore selected for subsequent pattern analysis.

The proposed framework further supports the recommendation of science education enhancement strategies, including teacher professional development initiatives, curriculum improvement programs, digital infrastructure expansion, AI literacy integration, and institutional AI adoption plans. By linking identified digital literacy profiles with context-specific intervention strategies, the framework enables educational stakeholders to implement more effective and responsive digital transformation initiatives.

In addition, the framework assists educational institutions, curriculum developers, school administrators, and policymakers in planning and implementing AI integration strategies based on empirical evidence. The generated recommendations provide guidance for resource allocation, technology investments, instructional innovation, and policy formulation aimed at strengthening digital literacy and AI readiness.

Ultimately, the Intelligent Recommender System Framework serves as a strategic decision-support model for improving digital literacy development and advancing AI-enabled science education at institutional, regional, and national levels. By transforming analytical findings into practical recommendations, the framework contributes to evidence-based educational planning and supports the sustainable integration of artificial intelligence within science education systems.

III. Results and Discussion

A. Descriptive Results of AI Adoption and Digital Literacy Indicators

The descriptive results show that global AI adoption in education remains moderate but uneven. Student AI usage recorded a mean of 30.99%, while teacher AI usage recorded a mean of 29.91%. School AI adoption averaged 29.41%, indicating that institutional AI integration was still developing across the dataset. Internet penetration showed a relatively high mean of 74.70%, while the education index averaged 0.776.

TABLE I. Descriptive Statistics of Main Indicators

Indicator	Mean	SD	Min	Max
Student AI usage (%)	30.99	16.7	0	65
Teacher AI usage (%)	29.91	16.46	0	60
School AI adoption (%)	29.41	16.6	0	62
Internet penetration (%)	74.7	20.69	42	96
Education index	0.776	0.157	0.52	0.94
Urban AI usage (%)	33.56	16.76	0	70
Rural AI usage (%)	21.6	16.19	0	59
Average daily AI usage	32.82	15.56	5	60
Gender gap in AI usage (%)	5.41	2.88	1	10

The results presented in Table I indicate that AI use in education is expanding, but adoption remains uneven between

urban and rural areas. Urban AI usage was higher than rural AI usage, suggesting a continuing digital access gap. This finding supports the need for a recommender system that does not only identify digital literacy levels but also recommends targeted interventions based on specific limiting factors.

B. Digital Literacy Profile Distribution

The computed Digital Literacy Score had a mean of 0.502, indicating a generally moderate level of digital literacy development. Using fixed classification ranges, 790 observations were categorized as moderate, 302 as high, and 268 as low. For model training and balanced classification, the profiles were also grouped into three balanced levels.

TABLE II. DIGITAL LITERACY PROFILE DISTRIBUTION

Profile	Fixed Range Count	Balanced Modeling Count
Low	268	453
Moderate	790	454
High	302	453

The dominance of the moderate category, as presented in Table II, implies that many educational environments have already developed some degree of digital readiness but still require strategic interventions to move toward higher AI-enabled digital literacy. This supports the importance of the proposed recommender system, since moderate-performing environments may benefit from differentiated recommendations rather than general technology adoption policies.

C. Regional Pattern of Digital Literacy Development

Regional analysis showed clear differences in digital literacy development. Oceania recorded the highest mean Digital Literacy Score, followed by Europe and North America. Africa and Asia recorded the lowest mean scores.

TABLE III. MEAN DIGITAL LITERACY SCORE BY REGION

Region	Mean DLS	SD	Count
Oceania	0.601	0.165	136
Europe	0.587	0.166	272
North America	0.584	0.17	272
South America	0.472	0.167	136
Asia	0.418	0.171	408
Africa	0.351	0.165	136

The results presented in Table III suggest that digital literacy development is strongly influenced by regional differences in infrastructure, educational readiness, and AI adoption capacity. The low mean scores in Africa and Asia indicate the need for stronger recommendations related to infrastructure development, teacher training, curriculum integration, and inclusive AI access.

D. Temporal Pattern of AI Adoption and Digital Literacy

The yearly trend showed a continuous increase in the Digital Literacy Score from 2015 to 2026. The mean score increased from 0.254 in 2015 to 0.799 in 2026. Student AI usage, teacher AI usage, and school AI adoption also increased steadily across the same period.

TABLE IV. DIGITAL LITERACY AND AI ADOPTION TREND

Year	Mean DLS	Student AI Usage	Teacher AI Usage	School AI Adoption
2015	0.254	4.98	4.04	3.38
2016	0.304	10.41	9.23	8.63
2017	0.347	15.4	14.01	13.47
2018	0.378	19.32	19.46	18.36
2019	0.444	25.47	23.74	23.63
2020	0.496	30.03	29.33	28.45
2021	0.544	35.3	34.26	33.52
2022	0.584	39.95	38.58	38.48
2023	0.639	44.84	44.11	43.54
2024	0.69	49.9	49.08	48.68
2025	0.741	55.33	53.45	53.43
2026	0.799	60.75	59	59.25

The increasing trend presented in Table IV confirms that AI adoption has become progressively embedded in educational environments. However, the variation across regions and indicators shows that growth is not uniform. Therefore, the proposed recommender system is necessary to provide context-sensitive interventions instead of a single universal strategy.

E. Machine Learning Performance for Digital Literacy Profiling

Random Forest, Naïve Bayes, and Hybrid Voting classifiers were used to support digital literacy profile classification and recommendation generation. Random Forest achieved the highest performance with an accuracy of 94.49% and an F1-score of 94.50%. The Hybrid Voting model achieved 90.81% accuracy, while Naïve Bayes achieved 82.35%.

TABLE V. MODEL PERFORMANCE

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.9449	0.9451	0.9449	0.945
Naïve Bayes	0.8235	0.824	0.8238	0.8232
Hybrid Voting	0.9081	0.9112	0.9081	0.909

Table V shows the model performance of selected machine learning algorithms. The Random Forest model performed best because it effectively captured nonlinear relationships among AI adoption indicators, infrastructure variables, and digital literacy profiles. The Hybrid Voting model also produced strong results, showing that combining probabilistic and ensemble learning methods can support reliable recommendation generation. Naïve Bayes performed lower but remained useful as a baseline model.

F. Confusion Matrix of the Best Model

The Random Forest model correctly classified most observations across the three digital literacy profiles. Misclassifications occurred mainly between adjacent categories, particularly between low and moderate, and between moderate and high.

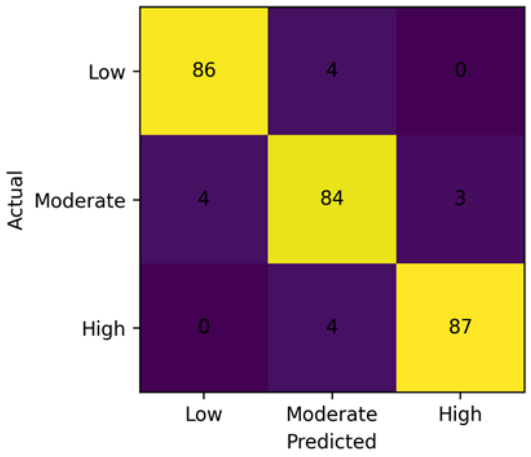


Fig. 1. Confusion matrix of the best-performing Random Forest model.

Fig. 1 shows the confusion matrix of the Random Forest. The confusion matrix indicates that the model was highly reliable in identifying low and high digital literacy profiles. Moderate cases were more difficult to classify because they share characteristics with both low and high profiles. This is important for the recommender system because moderate-profile environments require more refined recommendations based on their specific weaknesses.

G. Feature Importance and XAI-Based Interpretation

The feature importance results showed that student AI usage, urban AI usage, internet penetration, education index, school AI adoption, teacher AI usage, and rural AI usage were the strongest contributors to digital literacy classification.

TABLE VI. TOP PREDICTORS OF DIGITAL LITERACY PROFILE

Rank	Predictor	Importance Score
1	Student AI usage	0.1266
2	Urban AI usage	0.1237
3	Internet penetration	0.106
4	Education index	0.1026
5	School AI adoption	0.1008
6	Teacher AI usage	0.0958
7	Rural AI usage	0.0862
8	Year	0.069
9	Region	0.0631
10	Average daily AI usage	0.0589

The results presented in Table VI suggest that digital literacy development is influenced by both individual AI engagement and supportive educational conditions. Variables such as student AI usage and teacher AI usage represent the extent to which learners and educators actively utilize AI technologies in teaching and learning. Higher levels of AI engagement provide more opportunities to develop essential digital competencies, including information management, critical thinking, problem-solving, and the ethical use of digital technologies.

Meanwhile, internet penetration and the education index reflect the structural conditions that enable digital literacy. Reliable internet access provides users with opportunities to utilize digital platforms and AI applications, while a higher education index indicates a stronger educational environment that supports technology integration and digital skills development.

H. Cluster-Based Pattern Discovery

Cluster analysis identified three major patterns of AI adoption and digital literacy development.

TABLE VII. CLUSTER PROFILE

Cluster	Count	Mean DLS	Dominant Profile	Description
Cluster 0	453	0.383	Low	Low infrastructure and low readiness pattern
Cluster 1	453	0.414	Low	High infrastructure but low AI engagement pattern
Cluster 2	454	0.709	High	High readiness and high AI adoption pattern

Table VII shows that the respondents were grouped into three distinct clusters based on their Digital Literacy Score (DLS), infrastructure, and AI engagement.

Cluster 0 (n = 453) obtained a mean DLS of 0.383 and represents respondents with low infrastructure and low digital readiness. This group has limited access to digital resources and requires support in both infrastructure and foundational digital literacy.

Cluster 1 (n = 453) recorded a mean DLS of 0.414, also indicating low readiness. Although respondents have relatively better digital infrastructure, their AI engagement remains low, suggesting that access to technology alone does not ensure AI adoption. Targeted AI training and competency development are needed for this group.

Cluster 2 (n = 454) achieved the highest mean DLS of 0.709, reflecting high digital readiness and strong AI adoption. Respondents in this cluster are well-equipped to integrate AI into practice and can serve as early adopters or mentors.

I. Recommendation Rule Analysis

Table VIII summarizes the recommendation conditions generated by the proposed Intelligent Recommender System Framework based on identified digital literacy profiles and influential predictors.

TABLE VIII. RECOMMENDATION CONDITIONS AND INTERVENTION PRIORITIES.

Condition	Affected Records	Percentage	Main Recommendation
Low internet penetration	408	0.3	Internet infrastructure and connectivity programs
Low teacher AI usage	345	0.2537	AI-focused teacher training
No AI curriculum integration	673	0.4949	Curriculum redesign and AI literacy modules
Low school AI adoption	351	0.2581	Institutional AI adoption roadmap
High urban-rural AI gap	384	0.2824	Rural digital inclusion programs
High gender gap in AI usage	391	0.2875	Inclusive digital literacy interventions

The most frequent recommendation condition was the absence of AI curriculum integration, affecting 49.49% of the dataset. This indicates that curriculum enhancement is a major priority for strengthening AI-enabled science education. Low internet penetration, low teacher AI usage, low school AI adoption, and high urban-rural gaps also appeared as important intervention areas. The recommendation analysis confirms that the framework can translate analytical findings into actionable strategies. For low digital literacy environments, the system prioritizes access, infrastructure, and basic AI readiness. For moderate environments, the system recommends curriculum integration, teacher capacity building, and institutional adoption. For high-performing environments, the system may recommend advanced AI-enhanced science learning, innovation programs, and continuous improvement strategies.

J. Proposed Intelligent Recommender System Framework

Fig. 2 illustrates the proposed Intelligent Recommender System Framework developed to support digital literacy development and AI integration in science education.

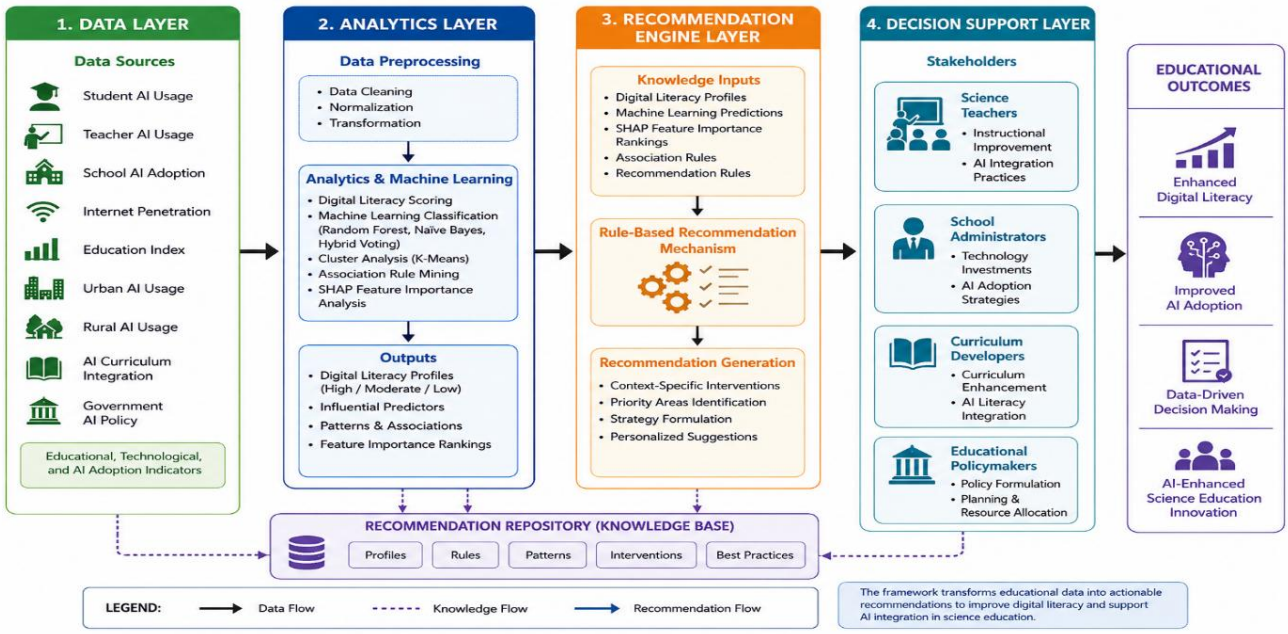


Fig. 2. The proposed Intelligent Recommender System Framework.

The framework consists of four interconnected layers that transform educational data into evidence-based recommendations for various educational stakeholders. The process begins with the Data Layer, which gathers educational, technological, and AI adoption indicators, including student and teacher AI usage, school AI adoption, internet penetration, education index, urban and rural AI usage, curriculum integration, and government AI policies. These variables provide the foundational information needed to assess digital literacy conditions and AI readiness across educational environments.

The collected data are then processed in the Analytics Layer, where data preprocessing, digital literacy profiling, machine learning analysis, cluster analysis, association rule mining, and SHAP-based explainable artificial intelligence are performed. This layer identifies significant patterns, influential predictors, and digital literacy profiles that characterize the technological readiness and AI adoption status of educational environments.

The extracted knowledge is subsequently transferred to the Recommendation Engine Layer, which serves as the core intelligence component of the framework. Using digital literacy profiles, machine learning outputs, SHAP feature importance rankings, and recommendation rules, the system generates targeted interventions and context-specific recommendations. These recommendations address identified gaps and priorities related to infrastructure, teacher competency development, curriculum enhancement, technology adoption, and digital inclusion.

The generated recommendations are delivered through the Decision Support Layer to key stakeholders, including science teachers, school administrators, curriculum developers, and educational policymakers. This layer enables stakeholders to utilize the recommendations for instructional improvement, technology planning, curriculum enhancement, resource allocation, and policy development.

Supporting all layers is the Recommendation Repository, which functions as a knowledge base that stores digital literacy profiles, discovered patterns, recommendation rules, intervention strategies, and best practices. This repository facilitates continuous knowledge accumulation and supports future recommendation generation.

This framework transforms educational data into actionable insights that promote enhanced digital literacy, improved AI adoption, evidence-based decision-making, and innovation in science education. By integrating Educational Data Mining, explainable artificial intelligence, and recommendation mechanisms, the framework provides a practical and scalable decision-support tool for advancing AI-enabled science education at institutional, regional, and national levels.

K. Discussion

The results demonstrate that global AI adoption and digital literacy development in science education are influenced by a combination of learner engagement, teacher readiness, institutional adoption, infrastructure, and regional conditions. The steady increase in Digital Literacy Score from 2015 to 2026 indicates that educational systems are becoming more digitally capable. However, regional differences and cluster patterns show that digital transformation remains uneven. These findings are particularly relevant to science education because effective integration of artificial intelligence supports inquiry-based learning, virtual experimentation, scientific problem-solving, and data-driven investigation. Educational environments demonstrating higher AI readiness are more capable of implementing AI-assisted science instruction, whereas institutions with lower digital literacy require targeted infrastructure, teacher development, and curriculum enhancement before AI can effectively support science learning.

The strong influence of student AI usage and teacher AI usage confirms that digital literacy development depends not only on access to technology but also on meaningful use of AI tools in teaching and learning. Teacher AI usage is especially important because teachers serve as facilitators of AI-enhanced science learning. When teacher AI usage is low, students may have limited opportunities to develop responsible and productive AI-related competencies.

Internet penetration and the education index were also among the most influential predictors. This finding suggests that digital literacy development requires a supportive ecosystem. Educational environments with strong infrastructure and higher educational development are more capable of integrating AI tools into science education. Conversely, low internet penetration limits access to AI-enabled learning resources and weakens the foundation for digital literacy development.

The cluster analysis provides further evidence that recommendation strategies must be context-specific. Some low-performing environments suffer from limited infrastructure, while others have adequate infrastructure but weak AI engagement. Therefore, a single intervention strategy is insufficient. The proposed Intelligent Recommender System Framework addresses this issue by matching each digital literacy profile with appropriate intervention priorities.

The results support the development of a recommender system framework that uses digital literacy profiling, machine learning classification, XAI-based feature interpretation, and rule-based recommendation generation. The framework can assist science teachers, administrators, curriculum developers, and policymakers in identifying priority areas for AI integration and digital literacy development. By converting data patterns into targeted recommendations, the framework provides a practical decision-support mechanism for improving science education in the digital age.

L. Limitations of the Study

One limitation of this study is that the Digital Literacy Score was developed as a composite indicator using normalized variables available in the dataset rather than a standardized digital literacy assessment instrument. Future studies should validate the proposed score against established digital literacy frameworks such as DigCompEdu [45] digital competency models.

IV. Conclusion and Recommendations

A. Conclusion

This study successfully developed an Intelligent Recommender System Framework for analyzing global AI adoption patterns and digital literacy development in science education. The findings demonstrated that digital literacy development is influenced by a combination of technological access, educational readiness, institutional AI adoption, and learner and teacher engagement with AI technologies. The results revealed a consistent increase in digital literacy levels and AI adoption indicators from 2015 to 2026, indicating the growing integration of artificial intelligence within educational environments.

Machine learning and explainable artificial intelligence analyses identified student AI usage, urban AI usage, internet penetration, education index, school AI adoption, and teacher AI usage as the most influential factors affecting digital literacy development. The Random Forest model achieved the highest classification performance, confirming its effectiveness in identifying digital literacy profiles and supporting recommendation generation. Cluster analysis further demonstrated that educational environments exhibit distinct patterns of technological readiness and AI adoption, highlighting the need for context-specific interventions rather than uniform policy approaches.

The proposed Intelligent Recommender System Framework transformed analytical findings into actionable recommendations through digital literacy profiling, pattern discovery, feature importance analysis, and rule-based recommendation generation. By linking educational conditions with targeted interventions, the framework provides a practical and explainable decision-support mechanism capable of supporting infrastructure development, teacher capacity building, curriculum enhancement, AI integration, and digital inclusion initiatives. Although the framework demonstrated high classification performance, its robustness under noisy, incomplete, or highly imbalanced educational datasets was not explicitly evaluated. Future investigations should examine the resilience of the proposed framework using datasets containing missing values, measurement uncertainty, and diverse educational contexts.

B. Recommendations

Based on the findings of the study, educational institutions should strengthen digital literacy development initiatives by increasing access to reliable internet infrastructure, particularly in regions and educational environments characterized by low technological readiness. Investments in connectivity and educational technology resources can provide the foundation necessary for effective AI integration and digital learning.

Teacher professional development programs should be expanded to enhance AI-related competencies, instructional innovation, and technology integration skills. Since teacher AI usage emerged as a significant predictor of digital literacy

development, continuous training programs and certification initiatives should be prioritized to improve educators' capacity to utilize AI tools effectively in science education.

Curriculum developers are encouraged to incorporate AI literacy concepts, data literacy skills, and AI-supported learning activities within science curricula. Greater integration of AI into curriculum design can help learners develop the competencies necessary to engage responsibly and effectively with emerging technologies.

Educational policymakers should utilize data-driven approaches in planning AI adoption strategies and digital transformation initiatives. The implementation of targeted interventions based on digital literacy profiles can improve resource allocation and maximize the impact of educational investments.

Future researchers may extend the proposed framework by incorporating real-time educational data, longitudinal analyses, learning analytics indicators, and additional explainable AI techniques. Further validation of the framework using institutional, regional, or national datasets may also enhance its applicability and effectiveness in supporting AI-enabled educational innovation.

ACKNOWLEDGMENT

The researcher gratefully acknowledges the Pangasinan State University for its support in this study and Kaggle for providing the dataset used in the research. Artificial intelligence tools were utilized solely for language refinement, organization, and formatting. All research processes, analyses, findings, and conclusions were reviewed and validated by the researcher to ensure accuracy, originality, and academic integrity.

REFERENCES

[1] M. K. Rahman, M. A. Hossain, N. A. Ismail, M. S. Hossen, and M. Sultana, "Determinants of students' adoption of AI chatbots in higher education: The moderating role of tech readiness," Interactive Technology and Smart Education, vol. 22, no. 3, pp. 448–480, 2025.

[2] M. Rieckmann, Education for Sustainable Development Goals: Learning Objectives. Paris, France: UNESCO Publishing, 2017.

[3] N. Azmi, "The benefits of using ICT in the EFL classroom: From perceived utility to potential challenges," Journal of Educational and Social Research, vol. 7, no. 1, pp. 111–118, 2017.

[4] M. M. Yunus, N. Nordin, H. Salehi, M. A. Embi, and Z. Salehi, "The use of information and communication technology (ICT) in teaching ESL writing skills," English Language Teaching, vol. 6, no. 7, pp. 1–8, 2013.

[5] D. Fidaoui, R. Bahous, and N. N. Bacha, "CALL in Lebanese elementary ESL writing classrooms," Computer Assisted Language Learning, vol. 23, no. 2, pp. 151–168, 2010.

[6] P. Wang, "The effect of computer-assisted whole language instruction on Taiwanese university students' English learning," English Language Teaching, vol. 4, no. 4, pp. 10–27, 2011.

[7] A. Qoura, "EFL teacher competencies in the ICT age," Journal of Research in Curriculum, Instruction and Educational Technology, vol. 3, no. 2, pp. 127–159, 2017.

[8] T. Boudjadar, "ICT in the writing classroom: The pros and the cons," International Journal of Applied Linguistics and English Literature, vol. 4, no. 1, pp. 8–13, 2015.

[9] P. Mishra and M. J. Koehler, "Technological pedagogical content knowledge: A framework for teacher knowledge," Teachers College Record, vol. 108, no. 6, pp. 1017–1054, 2006.

[10] L. S. Shulman, "Those who understand: Knowledge growth in teaching," Educational Researcher, vol. 15, no. 2, pp. 4–14, 1986.

[11] L. S. Shulman, "Knowledge and teaching: Foundations of the new reform," Harvard Educational Review, vol. 57, no. 1, pp. 1–23, 1987.

[12] R. Rahmany, B. Sadeghi, and A. S. Chegini, "Normalization of CALL and TPACK: Discovering teachers' opportunities and challenges," Journal of Language Teaching and Research, vol. 5, no. 4, pp. 891–900, 2014.

[13] P. Redmond and J. Lock, "Secondary pre-service teachers' perceptions of technological pedagogical content knowledge (TPACK): What do they really think?" Australasian Journal of Educational Technology, vol. 35, no. 3, pp. 45–54, 2019.

[14] S. Sharma, P. Mittal, M. Kumar, and V. Bhardwaj, "The role of large language models in personalized learning: A systematic review of educational impact," Discover Sustainability, vol. 6, no. 1, pp. 1–24, 2025.

[15] M. A. Rodríguez-Ortiz, P. C. Santana-Mancilla, and L. E. Anido-Rifón, "Machine learning and generative AI in learning analytics for higher education: A systematic review of models, trends, and challenges," Applied Sciences, vol. 15, no. 15, Art. no. 8679, 2025.

[16] D. Tbaishat, G. Amoudi, and M. Elfadel, "Adapting teaching and learning with existing generative AI by higher education students," Computers and Education: Artificial Intelligence, vol. 8, Art. no. 100421, 2025.

[17] E. F. Risko and S. J. Gilbert, "Cognitive offloading," Trends in Cognitive Sciences, vol. 20, no. 9, pp. 676–688, 2016.

[18] J. Sweller, "Cognitive load during problem solving: Effects on learning," Cognitive Science, vol. 12, no. 2, pp. 257–285, 1988.

[19] A. Sharma and R. Rani, "A systematic review of applications of machine learning in cancer prediction and diagnosis," Archives of Computational Methods in Engineering, vol. 28, no. 7, pp. 4875–4896, 2021.

[20] M. Shamshuzzoha et al., "A novel framework for seasonal affective disorder detection: Comprehensive machine learning analysis using multimodal social media data and SMOTE," Acta Psychologica, vol. 256, Art. no. 105005, 2025.

[21] M. Schreiber et al., "A discourse on the use of machine learning in personality psychology," Journal of Research in Personality, vol. 119, Art. no. 104666, 2025.

[22] M. M. Ramadhan et al., "Parameter tuning in random forest based on grid search method," DEStech Transactions on Computer Science and Engineering, 2017.

[23] B. Sumathi, "Grid search tuning of hyperparameters in random forest classifier for customer feedback sentiment prediction," International Journal of Advanced Computer Science and Applications, vol. 11, no. 9, 2020.

[24] S. Subbiah et al., "Intrusion detection technique in wireless sensor network using grid search random forest with Boruta feature selection algorithm," Journal of Communications and Networks, vol. 24, no. 2, pp. 264–273, 2022.

[25] A. A. Salamai, E. S. M. El-Kenawy, and I. Abdelhameed, "Dynamic voting classifier for risk identification in supply chain 4.0," Computers, Materials & Continua, vol. 69, no. 3, 2021.

[26] N. A. Samee et al., "Metaheuristic optimization through deep learning classification of COVID-19 in chest X-ray images," Computers, Materials & Continua, vol. 73, no. 2, 2022.

[27] I. Rish, "An empirical study of the naive Bayes classifier," in Proc. IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, 2001, pp. 41–46.

[28] A. M. Salih et al., "A perspective on explainable artificial intelligence methods: SHAP and LIME," Advanced Intelligent Systems, vol. 7, no. 1, Art. no. 2400304, 2025.

[29] S. Hussain, "Global AI in Education Dataset (2015–2026)," Kaggle Dataset, 2026. [Online]. Available: https://doi.org/10.34740/KAGGLE/DSV/15685116. [Accessed: Jan. 7, 2026].

[30] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, Recommender Systems: An Introduction. New York, NY, USA: Cambridge University Press, 2011.

[31] W. Liu and L. Gao, "Recommendation system based on fuzzy cognitive map," Journal of Multimedia, vol. 9, no. 7, pp. 970–976, 2014.

[32] IEEE Learning Technology Standards Committee, "Learning Object Metadata (LOM) Standard." [Online]. Available: http://ieeeltsc.org/wg12LOM/

[33] B. Ojokoh, M. Omisore, O. Samuel, and T. Ogunniyi, "A fuzzy logic based personalized recommender system," International Journal of Computer Science Information Technology and Security, vol. 2, no. 5, pp. 1008–1015, 2012.

[34] K. Palanivel and R. Sivakumar, "Fuzzy multicriteria decision making approach for collaborative recommender systems," International Journal of Computer Theory and Engineering, vol. 2, no. 1, pp. 57–63, 2010.

[35] A. Rodriguez, J. Gago, L. Anido-Rifón, and R. Rodríguez, "A recommender system for non-traditional educational resources: A semantic approach," Journal of Universal Computer Science, vol. 21, pp. 306–325, 2015.

[36] F. Ricci, L. Rokach, B. Shapira, and P. Kantor, Eds., Recommender Systems Handbook. New York, NY, USA: Springer, 2011.

[37] L. Rifón, A. Cañas, V. Roris, J. Gago, and M. Iglesias, "A recommender system for educational resources in specific learning contexts," in Proc. 8th International Conference on Computer Science and Education (ICCSE), 2013, pp. 371–376.

[38] E. Stuart, N. Shadbolt, and D. De Roure, "Ontological user profiling in recommender systems," ACM Transactions on Information Systems, vol. 22, no. 1, pp. 54–88, 2004.

[39] R. Sikka, A. Dhankhar, and C. Rana, "A survey paper on e-learning recommender system," International Journal of Computer Applications, vol. 47, pp. 27–30, 2012.

[40] L. Terán, SmartParticipation: A Fuzzy-Based Recommender System for Political Community Building. Switzerland: Springer, 2014.

[41] C. Tsai and H. Chuang, "The role of cognitive decision effort in electronic commerce recommendation system," International Scholarly and Scientific Research & Innovation, vol. 5, no. 10, pp. 36–40, 2011.

[42] A. Almahairi, K. Kastner, K. Cho, and A. Courville, "Learning distributed representations from reviews for collaborative filtering," in Proc. 9th ACM Conference on Recommender Systems (RecSys '15), 2015, pp. 147–154.

[43] J. Neidhardt, R. Schuster, L. Seyfang, and H. Werthner, "Eliciting the users' unknown preferences," in Proc. 8th ACM Conference on Recommender Systems (RecSys '14), 2014, pp. 309–312.

[44] C. Mediani and M. H. Abel, "Semantic recommendation of pedagogical resources within learning ecosystems," in Proc. International Conference on Industrial Informatics and Computer Systems (CIICS), 2016,

[45] C. Redecker, European Framework for the Digital Competence of Educators: DigCompEdu, Y. Punie, Ed. Luxembourg: Publications Office of the European Union, 2017, EUR 28775 EN. doi: 10.2760/159770. [Online]. Available: https://publications.jrc.ec.europa.eu/repository/bitstream/JRC107466/pdf_digcomedu_a4_final.pdf

Automated Multi-Task Deep Learning Framework for Simultaneous Identification of Dental Pathologies and Arch Localization in Orthopantomograms (OPG)

Hatim Elhag[1], Norehan Binti Mokhtar[2], Siti Noor Fazliah Binti Mohd Noor[3],
Gururajaprasad Kaggal Lakshmana Rao[4], Nordin Zakaria[5], Fakhitah Ridzuan[6], Mohammed Bin Salah[7]
Dental Simulation and Virtual Learning Research Excellence Consortium, Department of Dental Science,
Advanced Medical and Dental Institute, Universiti Sains Malaysia, Bertam, Penang, Malaysia[1, 2, 3]
Department of Orthodontics, Penang International Dental College, NB Tower 5050,
Jalan Bagan Luar, Butterworth, Penang, Malaysia[4]
Department of Computer and Information Sciences, Universiti Teknologi PETRONAS, Seri Iskandar 32160, Malaysia[5]
Faculty of Data Science and Computing, Universiti Malaysia Kelantan, Pengkalan Chepa, 16100 Kota Bharu, Kelantan[6]
Department of Oral and Maxillofacial Surgery-College of Dentistry, King Saud University, Riyadh, Saudi Arabia[7]

***Abstract*—Interpreting orthopantomograms (OPGs) accurately and efficiently remains challenging due to overlapping anatomical structures and the subtle presentation of many dental pathologies. While deep learning (DL) has shown considerable potential in dental imaging, most existing models focus on single-task detection and do not address the need for simultaneous analysis of both structural and pathological features within a unified system. To fill this gap, an automated multi-task deep learning framework is proposed that concurrently identifies ten dental pathology classes such as fractured roots and periodontally compromised teeth alongside arch-region classification of edentulous spans in OPGs. This method makes use of the YOLOv11 object detection architecture, which is powered by the C2PSA attention module and sophisticated post-processing algorithms that were trained on a larger collection of ten classes from the Dental OPG Kennedy dataset. With a mean Average Precision (mAP) 0.854 and an F1-score of 0.961 for every class assessed, the model demonstrated strong overall performance. Even in more complicated situations like root fractures (F1 = 0.928), the model was able to recognize well-defined edentulous areas with almost perfect sensitivity (Recall = 0.997) and good mAP. By integrating diagnostic analysis and anatomical techniques in a single optimized inference process, the YOLOv11 system offers interpretable and scalable solutions for comprehensive screening, thereby representing an advance for Computer-Assisted Dentistry Diagnostics.**

Keywords*—Orthopantomogram (OPG); deep learning (DL); multi-task learning; YOLOv11; object detection; dental pathology; arch classification; Computer-Aided Diagnosis (CAD)

I. Introduction

Nowadays, panoramic radiography, in the form of an orthopantomogram (OPG), is an important imaging modality in dental practice as a fundamental examination that covers the maxilla and mandible along with complete dentition and related anatomic structures [1]. Therefore, the entire process of screening, treatment planning, and follow-up can be realized in different environments, including general dental care and specialist treatment. Nevertheless, the OPG is by nature a complex and time-consuming image to be interpreted. Specifically, anatomical structures overlapped with each other, and the image quality varied from one case to another; hence, subtle signs of pathological processes were not easy to identify accurately by human beings [2].

Recent developments in Artificial Intelligence (AI) and Deep Learning (DL) have been finding a lot of promise in creating diagnostic tools in medical imaging areas such as dental radiology [3]. There have been several sources that reveal a rapidly increasing trend for implementing AI systems in the dental field, including detection or analysis of teeth, segmentation analysis, dental caries detection, analysis of endodontic lesions, and implant assessment. For instance, AbuSalim et al. [4] highlight the use of DL technology in dental informatics and the importance of implementing complex and explainable artificial intelligence systems that can improve efficiency in the healthcare sector. Similarly, Ali et al. [5] also include a detailed explanatory study related to the use of AI in dental radiology, highlighting its potential for increasing diagnostic accuracy, reducing radiation exposure, and improving patient outcomes.

Among DL approaches, Convolutional Neural Networks (CNNs) as well as those based on transformers have been utilized to identify dental health conditions at a similar level of human expertise. Especially in terms of identifying dental images, it has been noted that the use of Single-Stage object detection techniques such as You Only Look Once (YOLO) is a highly promising one because of its fast detection speed and efficient spatial localization technique [6]. For example, the application of the YOLOv3 model was proposed by Almalki et al. [7] for the detection of dental pathologies from the panoramic OPG X-ray images. In their approach, the Darknet-53 framework was proposed with the identification of four dental pathologies, namely cavity, root canal, dental crown, and broken-down roots, with the achievement of over 90% mAP on four different dental pathologies. Furthermore, Beser et al. [8] developed a YOLOv5 model with automatic segmentation and numbering of deciduous and permanent teeth in mixed-dentition radiographs. Their YOLOv5 model scored high

sensitivity, specificity, and accuracy (compared with other architectures like Faster R-CNN [9]), while also showing superior computational efficiency. Similarly, Makarim et al. [10] compared YOLOv5 with other models for classification of dental pathological images and reported that YOLOv5 achieved higher accuracy as well as faster processing times than alternative methods, demonstrating its suitability for multi-class dental object detection. Ismail and Taseen [11] proposed a new YOLOv8 architecture for the detection of 14 classes of dental anomalies. Their solution included the use of the anchor-free detection head, CSPDarknet53 as the backbone network, and the C2f feature fusion module. Although their model achieved high accuracy above 0.85 for structural tooth abnormalities such as implants and crowns, it failed to demonstrate similar performance in detecting subtle conditions such as caries and periapical lesions. Based on this perspective, Şevik and Mutlu in [12] proved that it is possible to use the YOLOv11 model in identifying pediatric dental anomalies in panoramic radiographs. According to their findings, YOLO's fast learning and strong small object detection capabilities make it ideal for clinical dental image processing tasks.

Despite these advances, a number of significant gaps still exist in applying DL to panoramic dental images. One is that most of the literature focuses on single-task problems (eg., detect certain pathology, count teeth, classify implant brands) rather than multi-task approaches that merge several related tasks together within a single framework. For example, a multi-task learning study on mandibular third-molar classification found that learning multiple outputs simultaneously didn't always outperform single-task models [13]. In addition, multi-task frameworks that perform both anatomic classification and pathology detection are rarer still. Combining these two problems might yield more robust and meaningful clinical predictions. Publicly available OPG data used for research purposes is often too small or lacks diversity in terms of molecular composition, leading to issues of generalizability for models trained on such limited datasets and adaptation between fields outside their original domain [14]. In order to overcome the above-mentioned challenges, the following study introduces a comprehensive framework for multi-object detection in OPG radiography that can detect multiple dental pathologies while simultaneously classifying arch topology. The proposed framework adopts YOLOv11-based object detection and OpenCV-driven image analysis to jointly identify pathology localization and arch type in panoramic radiographs. A ten-class taxonomy is established to describe a variety of pathological and positional characteristics, including root fractures, periapical states, and different arch types. This framework is intended to advance Computer-Aided Diagnosis by delivering interpretable, efficient, and scalable analysis suitable for clinical screening workflows.

The remainder of this study is organized as follows: Section II reviews the related work on deep learning-based dental image analysis and identifies the research gap addressed in this study. Section III presents the proposed methodology, including dataset preparation, the YOLOv11 architecture, and the evaluation metrics. Section IV discusses the experimental results and analyzes the performance of the proposed framework. Section V provides a qualitative comparison with existing deep learning-based dental detection methods. Section VI discusses the clinical significance of the proposed framework, while Section VII presents the clinical interpretation of prediction errors and their implications for diagnostic decision-making. Section VIII outlines the limitations of the current study and directions for future research. Finally, Section IX concludes the study and highlights potential future research directions.

II. Related Work

Deep learning has significantly advanced automated dental image analysis by enabling accurate detection, classification, segmentation, and localization of various dental conditions from different imaging modalities. Recent studies have demonstrated the effectiveness of convolutional neural networks (CNNs), object detection frameworks, and multi-task learning strategies for improving diagnostic accuracy while reducing the need for manual interpretation.

Several studies have focused on automated dental pathology detection using deep learning. ToothNet, a YOLOX-based framework, was developed for the simultaneous detection of dental caries and fissure sealants from intraoral photographs, achieving image-level AUC values of 0.925 for caries detection and 0.902 for sealant detection. At the tooth level, the framework achieved a sensitivity of 0.807, precision of 0.814, and an F1-score of 0.810, demonstrating the effectiveness of object detection models for dental disease identification [15]. Similarly, a hybrid deep neural network integrating CNN, BiLSTM, and attention mechanisms was proposed for predicting dental treatment outcomes by combining radiographic, clinical, and behavioral information, achieving an overall accuracy of 94.2% and an AUC of 0.96 [16]. Although these approaches demonstrated promising performance, they primarily focused on individual diagnostic tasks rather than providing comprehensive analysis of panoramic dental radiographs.

Recently, multi-task learning (MTL) has attracted considerable attention because of its ability to simultaneously solve multiple related clinical tasks while learning shared feature representations. A CBAM-based multi-task framework jointly performed age estimation and gender classification from CBCT panoramic slices, achieving a mean absolute error of 1.08 years for age estimation and 95.3% classification accuracy for gender prediction, outperforming conventional single-task approaches [17]. Likewise, a hybrid MTL framework combining U-Net++, EfficientNet-B4, pseudo-label generation, and selective loss masking simultaneously performed anatomical segmentation and caries detection under partially annotated datasets, achieving a Dice Similarity Coefficient (DSC) of 0.6706, an Intersection over Union (IoU) of 0.5044, and improving sensitivity by 7.47% while reducing false-negative pixels by 19.9% [18]. These studies demonstrate the advantages of multi-task learning; however, they mainly address classification or segmentation problems rather than simultaneous pathology detection and anatomical localization.

Several researchers have also investigated tooth detection and localization using deep learning. A relation-based framework incorporating prior dental knowledge employed a multi-task CNN and a label reconstruction strategy to improve

tooth recognition and localization in dental periapical radiographs, outperforming conventional object detection approaches [19]. Similarly, a cascade CNN framework utilizing hierarchical label-tree representations achieved 95.8% precision and 96.1% recall for identifying all 32 tooth positions, effectively handling challenging cases involving missing, decayed, and restored teeth [20]. Despite these encouraging results, these approaches were primarily designed for tooth identification and localization rather than detecting multiple dental pathologies from panoramic OPG images.

Despite the rapid progress of deep learning in dental imaging, several important limitations remain. First, most existing studies focus on a single clinical task, such as caries detection, tooth localization, implant classification, age estimation, or treatment outcome prediction, rather than providing a comprehensive assessment of panoramic dental radiographs. Second, many methods are developed for intraoral photographs, CBCT images, or periapical radiographs, limiting their direct applicability to panoramic OPG imaging. Third, existing studies employ different datasets, annotation protocols, class definitions, and evaluation methodologies, making direct comparison difficult. Moreover, few studies simultaneously integrate dental pathology detection and anatomical arch localization within a unified object detection framework.

Motivated by these limitations, this study proposes a YOLOv11-based multi-task framework for simultaneous detection of multiple dental pathologies and localization of maxillary and mandibular arch regions using an expert-annotated extension of the Dental OPG Kennedy dataset. Unlike previous studies that primarily address individual diagnostic tasks, the proposed framework combines comprehensive pathology detection and anatomical localization within a single inference pipeline, providing a more clinically relevant solution for computer-aided analysis of panoramic dental radiographs.

III. Proposed Methodology

This section describes the methodology adopted to achieve the main objective of this study, which is to perform multitask dental pathology feature detection and dental arch classification on OPG panoramic X-ray images using the YOLOv11 framework integrated with OpenCV-based analysis. The overall operating procedure of the proposed system is shown in Fig. 1. This involves four main steps: preparing the dataset, developing the YOLOv11 model, and evaluating and analyzing the results.

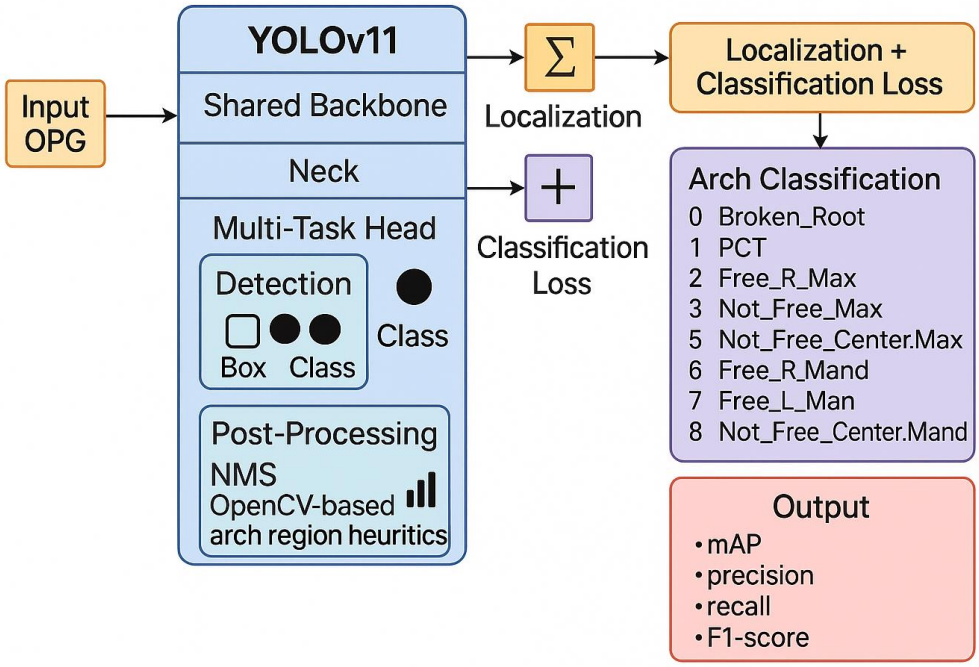


Fig. 1. Overview of the proposed multi-task object detection framework for dental pathologies and arch classification in panoramic OPG X-rays.

The following subsections explain each stage in detail.

A. Dataset Preparation

In this study, the publicly available Dental OPG Kennedy dataset was used, which was created by Orville Kennedy et al. and published on Kaggle [21]. This dataset contains panoramic dental radiographs called orthopantomograms (OPGs). It was originally collected to classify partially edentulous dental arches based on the Kennedy system. All the images are annotated according to root fracture, periodontally compromised tooth (PCT), and edentulous arch classes. These annotations are in the form of YOLO-ready text files, comparable to many other datasets, with each line representing a bounded box coordinate in addition to a class identifier. There are 500 dental X-ray images in PNG format, each with an average size of approximately 1500x800 pixels.

To support the objectives of this study, the original three-class annotation scheme was extended to ten clinically meaningful categories, enabling the detection of a broader range of dental conditions. The additional annotations were manually performed by a qualified dental expert according to predefined clinical diagnostic criteria. Each panoramic radiograph was carefully reviewed to ensure that the assigned labels accurately

represented the corresponding dental pathology or anatomical region. This expert-guided annotation process enhanced the clinical relevance and consistency of the extended dataset, providing reliable ground-truth labels for training and evaluating the proposed framework. The resulting annotation scheme comprises ten clinically meaningful categories, including root fractures, periodontally compromised teeth (PCT), and edentulous conditions affecting both the maxillary (upper jaw) and mandibular (lower jaw) arches, as summarized in Table I. Although the annotations were performed by a qualified dental expert, formal quantitative assessment of inter-annotator agreement and labeling uncertainty was beyond the scope of the present study and will be considered in future work.

TABLE I. FINAL 10-CLASS TAXONOMY USED FOR MULTI-TASK DETECTION.

Class ID	Class Name	Anatomical Region	Clinical Description / Meaning
0	Broken_Root	Both arches	Tooth root fracture or retained root fragment visible on OPG.
1	PCT (Periodontally Compromised Tooth)	Both arches	Tooth affected by advanced periodontal bone loss or mobility.
2	Free_R_Max	Maxilla – Right	Free-end edentulous space on the right side of the maxillary arch.
3	Free_L_Max	Maxilla – Left	Free-end edentulous space on the left side of the maxillary arch.
4	Not_Free_Max	Maxilla – Posterior	Bounded edentulous region (teeth missing between adjacent teeth) in maxilla.
5	Not_Free_Center_Max	Maxilla – Center	Central bounded edentulous area in anterior maxillary region.
6	Free_R_Mand	Mandible – Right	Free-end edentulous space on the right side of the mandibular arch.
7	Free_L_Mand	Mandible – Left	Free-end edentulous space on the left side of the mandibular arch.
8	Not_Free_Mand	Mandible – Posterior	Bounded edentulous region (non-free space) in posterior mandibular area.
9	Not_Free_Center_Mand	Mandible – Center	Central bounded edentulous space in anterior mandibular region.

This configuration efficiently enables the model to detect and classify pathological classes (classes 0-1) and arch types (classes 2-9), simultaneously. All the labels were double-checked by dental professionals to guarantee they are all clinically correct. Data pre-processing was carried out with OpenCV and other image processing libraries. All training images were resized to the 640 × 640 pixels expected by YOLOv11 input. The original PNG images were converted to grayscale in order to save computational cost and highlight radiographic features. The contrast limited adaptive histogram equalization (CLAHE) [22] was used to enhance the contrast, and the bidirectional filtering was applied to suppress noise while preserving edge sharpness, then pixel values were normalized between 0 and 1. Finally, the dataset was divided into 70% training, 20% validation, and 10% testing. This partition was maintained consistently throughout all experiments to ensure a fair and reproducible evaluation of the proposed framework. While a fixed data split facilitates consistent comparison under identical training conditions, it does not fully capture variability across different data partitions.

B. YOLOv11 Model Architecture

To achieve high computation efficiency with high precision feature extraction, the YOLOv11 model was selected as a detection framework. As illustrated in Fig. 2, the network consists of three blocks: 1) Backbone, 2) Neck, and 3) Detection Head.

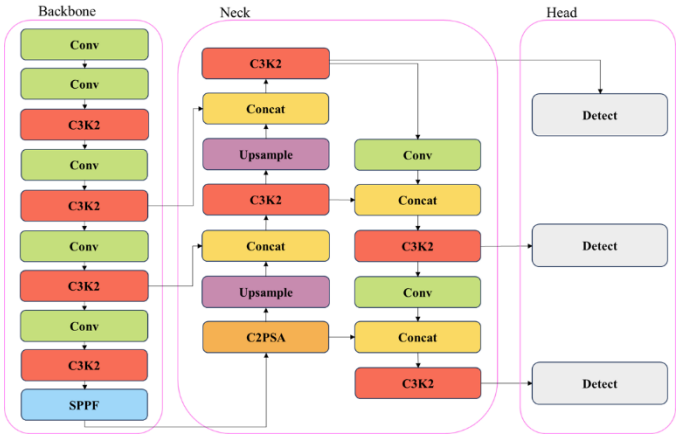


Fig. 2. YOLOv11 architecture

It leverages a series of convolutional layers equipped with Cross-Stage Partial (CSP) and C3K2 modules to promote the gradient flow and feature reutilization, without introducing extra computation overhead [23]. The Spatial pyramid pooling-fast (SPPF) block enhances the generalization capability of the model for object size changed in images by incorporating multi-scale context information from several receptive field. The neck incorporates *C2PSA* (Cross-Stage Partial with Spatial Attention) blocks, enabling spatial-attention refinement to emphasize diagnostically relevant regions [24], such as periapical lesions or edentulous margins, within panoramic X-rays. The detection head operates on three scales: P_3, P_4, and P_5 and outputs a vector:

$$[x, y, w, h, o, p_1, p_2, \dots, p_C] \quad (1)$$

For each anchor, where (x, y) denote the normalized box-center offsets, (w, h) denote the width and height, o represents the objectness confidence, and p_i are class probabilities for $C = 10$ dental categories. The total loss function combines localization, confidence, and classification losses as:

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{CIoU}} + \beta \mathcal{L}_{\text{obj}} + \gamma \mathcal{L}_{\text{cls}} \quad (2)$$

where, $\mathcal{L}_{\text{CIoU}}$ represents the Complete-IoU regression loss for bounding-box optimization, and $\mathcal{L}_{\text{obj}}$ and $\mathcal{L}_{\text{cls}}$ denote the binary and categorical cross-entropy losses, respectively. Hyperparameters α, β, and γ were empirically tuned to balance

localization precision and class discrimination across both pathological and positional categories.

To adapt YOLOv11 to panoramic dental imagery, the model was fine-tuned via transfer learning using the pre-trained YOLOv11s checkpoint initialized on the COCO dataset. The anchor-box dimensions were recalculated using k-means clustering on the bounding-box distributions of the dental dataset to better accommodate small and irregular anatomical structures. Training was conducted at an input resolution of 640×640 pixels with a batch size of 16 and a learning rate of 1×10^{-3} employing a cosine decay schedule and warm-up phase. The optimizer used SGD with momentum 0.937 and weight decay 5×10^{-4}, while early stopping was applied after 100 epochs to prevent overfitting. During inference, Non-Maximum Suppression (NMS) was performed with an IoU threshold of 0.45 and a confidence threshold of 0.25 to eliminate duplicate predictions.

C. Model Evaluation and Performance Measures

The model was evaluated on standard object detection metrics, precision, recall, mean average precision, mAP, F1-score, to check how well the detector is performing to classify the various objects with different confidence levels.

Precision (P) indicates how many positives detected are actually true and gives an indication of the false positive rate. It is calculated as:

$$P = \frac{TP}{TP+FP} \tag{3}$$

Here, TP and FP indicate the number of true positives and false positives, respectively. A higher accuracy value indicates that the detector is making accurate predictions and correctly identifying real objects with minimal false positives.

Recall (R) evaluates the sensitivity of a model to all real objects in an image by quantifying the proportion of correctly identified objects to the total number of valid instances. It is given by:

$$R = \frac{TP}{TP+FN}, \tag{4}$$

Here, FN indicates a false negative, i.e. an object missed by the detector. High recall means the detector detects most objects in the image.

The F1 score provides a smooth average between precision and recall, providing a balanced view of the detector's overall performance in correctly identifying objects. It is expressed as:

$$F_1 = 2 \times \frac{P \times R}{P+R} \tag{5}$$

where, $F_1 = 1$ indicates perfect detection performance and $F_1 = 0$ represents complete failure. This metric is especially useful in evaluating trade-offs between precision and recall when optimizing detection thresholds.

The mean Average Precision (mAP) is the main metric used in object detection to measure accuracy across different Intersection-over-Union (IoU) thresholds. For each class, an Average Precision (AP) score is calculated by measuring the area under the Precision–Recall (PR) curve, defined as:

$$AP = \int_0^1 P(R)dR \tag{6}$$

and the mAP is then calculated as the mean of AP scores over all object classes:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{7}$$

where, N is the total number of classes. The mAP@50–95 metric is reported by averaging the AP scores across several IoU thresholds, ranging from 0.5 to 0.95 in steps of 0.05. This approach offers a thorough assessment of how accurately the model localizes and detects objects.

These metrics combined give a comprehensive view of how well the detection system is performing. They balance the trade-offs between precision, recall, and localization accuracy, and are widely accepted as the standard criteria for evaluating modern deep learning object detection frameworks like YOLOv11.

IV. Results and Discussion

Fig. 3 shows the evaluation of class-wise performance using mean Average Precision and Recall metrics at IoU ranges between 0.50 and 0.95 (mAP@50-95) highlighting the discriminative, as well as structural limitations of the multi-task framework based on YOLOv11 architecture.

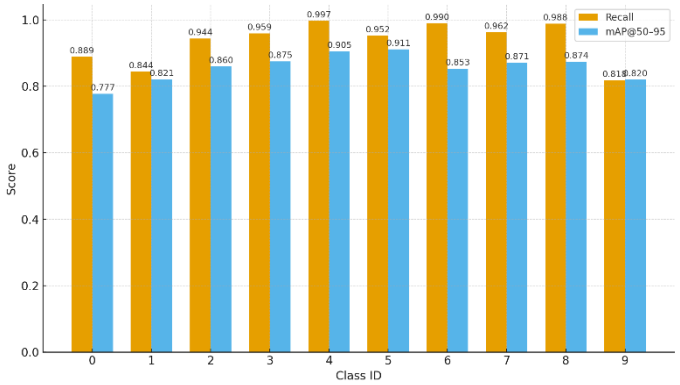


Fig. 3. Per-class comparison of Recall and mAP@50–95 for YOLOv11.

One core limitation of the model is observed in the Broken_Root (Class 0) category, with a Recall of 0.889 and mAP@50–95 of 0.777. Even though the YOLOv11x model benefits from the channel-dependent Cross-Stage Partial Spatial Attention (C2PSA) modules, the detection of minor, irregular-shaped fractures are still a struggle. The radio-image lucent and anisotropic orientation of fractured roots with overlying implant trabecular bone greatly degrades the distinct mid-level convolutional feature. Therefore, the destined abstraction is reduced within the feature pyramid, partially detecting and under-segmenting what is needed. Similarly, the Recall (0.844) and mAP@50-95 (0.821) of the Periodontal Compromised Teeth (PCT) exhibits the moderate sensitivity of the model in detecting pathological patterns – limiting the spatial precision. The ability of the anchor-free detector to regress high-IoU bounding boxes is reduced by the boundaries and margins of the diffuse lesion and irregular bone-loss. According to these observations, it is suggested that even though the attention selectivity of the model is improved by the C2PSA mechanism,

relying on localized regions with low contrasts is still restricted by the radiographic contrasts and limited spatial information.

Conversely, Free_R_Max (Class 2) and Free_L_Max (Class 3) categories exhibit strong localization performance (Recalls = 0.944 and 0.959; mAP@50–95 = 0.860 and 0.875). This demonstrates the ability of the model to extract persistent signs of ridge-morphology using multi-scale feature collections. The path Aggregation Network (PAN) of the YOLOv11 two-way feature fusion functionally integrates coarse and fine-grained contextual features, allowing accurate identification of balanced edentulous areas despite uneven illumination in wide-angle projections. The Not_Free_Max (Class 4) category (Recall = 0.997, mAP@50–95 = 0.905) further emphasizes the ability of the architecture to describe well-defined ridges in the mouth. The brightness gradient between dentate and edentulous sections is efficiently discriminated by the attention-guided backbone, proving the high spatial discrimination ability of the model. Similarly, a fine-grained identification of subtle edentulous structures within dense preceding mouth bones (usually limited by low contrasts) is also observed through the Not_Free_Center_Max (Class 5) category, having a Recall of 0.952 and mAP@50-95 of 0.911.

In the jawbone regions, both Free_L_Mand (Class 7) (Recall = 0.962, mAP@50-95 = 0.871) and Free_R_Mand (Class 6) (Recall = 0.990, mAP@50-95 = 0.853) both show high sensitivity, although slight localization inconsistencies persist near the posterior curvature. These disparities emerge from variations in cortical densities and uneven geometries in the jawline, which reduce feature spreading in the Feature Pyramid Network (FPN) pathway. A strong detection and localized accuracy is observed in the Not_Free_Mand (Class 8) category, having a Recall of 0.988 and mAP@50-95 of 0.874. This demonstrates the model's ability to monitor dorsal region containing irregular geometries and overlying impositions. On the other hand, the Not_Free_Center_Mand (Class 9) category displays great difficulty in detection with Recall of 0.818 and mAP@50-95 of 0.820. Incisors exhibiting low illumination, shadows from overlapping soft tissues, and thin cortical boundaries cause features to dilute in early convolutional layers, causing insignificant bounding box regressions to be produced.

The values of Fig. 3 evidently quantify accurate relationship between the plots and evaluation metrics, as described in the text above. Average values of Recall and mAP@50-95 specify the persisting detection ability of the YOLOv11 model across most segments, with some limitations faced in complicated and radiologically complex areas. In conclusion, accurate analysis of orthopantomogram (OPG) is improved by integrating contextual attention and multi-scale fusion techniques. In spite of that, some constraint still persists for minute, less-illuminated, and dark radio-imagery in dental domain, particularly in regions of deceiving bone-lesions. Future improvements of the model may involve hybrid transformer-CNN encoders to help improve localization accuracy in complex radiographic scenarios.

Fig. 4 quantitatively exhibits the balance of the model through F1 scores of each class. The sensitivity and precision values can be seen across all ten categories of dental pathology. This per-class evaluation boosts the stability of the model in accurate detection of multi-class performance.

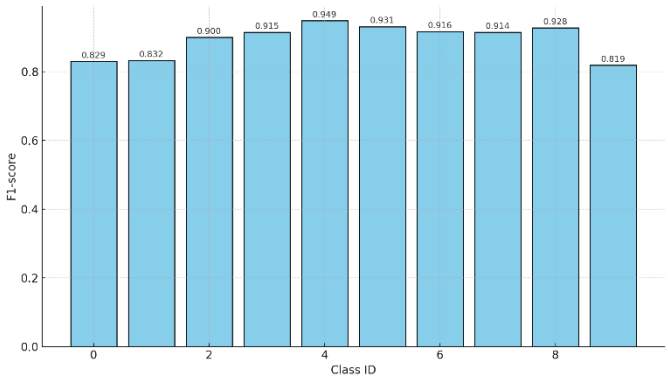


Fig. 4. Per-class comparison of F1-score for YOLOv11.

The Broken_Root (Class 0) (F1-score = 0.829) and PCT (Periodontally Compromised Tooth) (Class 1) (F1-score = 0.832) categories exhibit the lowest F1 scores. These values may be the result of previously-discussed challenges such as complex radiographs, uneven arrangement, and overlying trabecular patterns that limit accurate results. Similarly, the F1-score of the Not_Free_Center_Mand (Class 9) category (0.819) – low due to less-illuminated incisors, shadows, etc., is another example of restricted lesion discrimination. On the other hand, the YOLOv11 model achieves high F1-scores for the Free_R_Max (Class 2) (0.900) and Free_L_Max (Class 3) (0.915) categories, proving the model's ability in detecting even edentulous configurations within wide projections. The highest performance of 0.949, 0.931, and 0.928 for the Not_Free_Max (Class 4), Not_Free_Center_Max (Class 5), and Not_Free_Mand (Class 8) categories exhibit the localization ability of the model in well-defined and high contrast regions. The classes usually have distinct boundaries and even background densities, allowing the model to accurately discriminate spatial features and bounding box regressions.

On average, the F1-scores of all classes remain above 0.80, determining the model's accurate generalization and diagnosing ability across varying dental patterns. The YOLOv11 outperforms prior architectures of YOLOv5 and YOLOv8 that achieved an average clinical F1-score of 0.78 [25], [26]. This improvement may be attributed to the feature recalibration attention mechanism of C2SPA, the multi-scale fusion technique of the model, and boosted accuracy through Mosaic and MixUp pipelines to illuminate panoramic X-rays.

The normalized confusion matrix differentiating between true and predicted dental pathology classifications can be seen in Fig. 5 for all ten classes. The 'background' label in the confusion matrix does not represent an actual class in the dataset but corresponds to false-positive or unmatched predictions generated during evaluation. The discriminative reliability of the YOLOv11 model can be observed by analyzing consistent classification and inter-class confusion areas.

The diagonal matrix proves the harmony among true and predicted variables. This indicates that the proposed model successfully learns discriminative radiographic representations for most dental classes. High precision with 489 correct classifications and one misclassification in *Class 0* emphasizes YOLOv11 detection ability for well-defined structures. Similarly, 290 and 342 correct predictions and minimum off-

diagonal occurrences for *Classes 4 and 8* highlight the capacity of the network in capturing regular textures and uniform geometries in the regions. Some confusion in *Classes 1, 3, and 6* is observed with slight misclassifications. These minor occurrences may have emerged due to common grayscale attributes, scattered lesion borders, and overlying radiographic patterns in similar regions. These anomalies usually occur in panoramic imaging due to imposed complexes and distorted projections, which can cause restrictions even in advanced detectors.

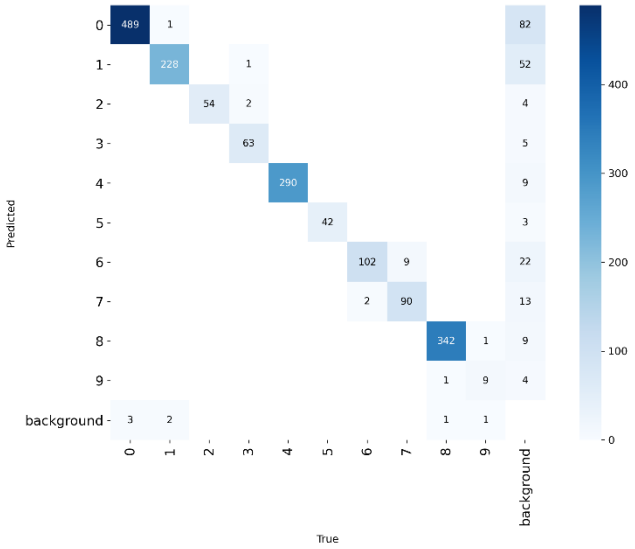


Fig. 5. Confusion matrix

The small number of misclassifications refers to weak and incomplete feature samples where the detector failed to place them into correct target classes. Such occurrences typically result from subtle, unclear radiographic segments or low-contrast lesions with incomplete information. It can be noted that no class displays dominating misclassifications in the system, proving a balanced learning behavior and no bias in the dataset. Overall, the confusion matrix validates the discriminative ability of the YOLOv11 model in complex diagnosing scenarios.

The residual errors may have occurred due to similar and overlapping neighboring entities. These limitations may be overcome by using augmentations specific to the domain (e.g., normalizing radiographic contrasts, spatial deformations), feature identification using high resolutions, or refining feature discrimination.

Training and Validation box regression loss of the YOLOv11 model across 100 epochs is given in Fig. 6. The curves for both training and validation datasets evolve to a consistent and stable downward trend, proving the model's ability in achieving spatial orientation between true and predicted bounding boxes.

For the training dataset, both curves display losses on high magnitudes, given random initial weights and raw spatial features. A steep decline in the first 10-15 epochs explains the learning of the network of localization features that are coarse. In succession, a gradual, close parallel decrease in both losses verifies smooth and accurate gradient traveling. The training box shows a consistently descending loss to about 0.6 by the 100th epoch, successfully reducing localization error by continuous learning. Loss for validation begins at a higher magnitude – indicating hidden data presence – then gradually converges towards the training loss after about 30 epochs. The combination of this behavior and minor oscillations exhibits the ability to generalize accurately without severely overfitting the outcome. The training and validation loss curves nearly overlap in the ending epochs, indicating a balance between bias and variance in the model.

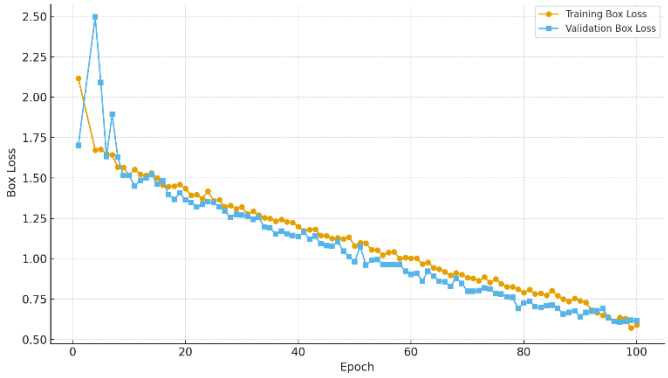


Fig. 6. Training vs. validation box loss over epochs.

The YOLOv11 model proves to be stable and efficiently optimized. The intensified backbone and multi-feature fusion of the network deeply facilitate the inclination traveling and reusability of spatial information across layers, preventing gradients from disappearing in cases of long training. Meanwhile, the enhanced multi-feature accumulation within the bidirectional PAN ensures efficient communication among top-down and bottom-up cues. This approach enables consistent spatial and contextual awareness, allowing correct bounding-box regression across physical diversity and varying dental patterns. All in all, the consistent loss curves indicate efficient learning rates when needed along with normalization and stable back propagations. The convergence pattern proves reliable optimization and precise spatial location in varying dental structures in OPG imagery.

V. Comparative Analysis with Existing Models

Recent advances in automated dental radiography analysis have demonstrated promising results using various YOLO-based object detection frameworks. However, the published studies differ considerably in terms of datasets, annotation protocols, class definitions, and evaluation methodologies. Consequently, direct quantitative comparison of reported performance metrics is challenging. The purpose of this section is, therefore, to position the proposed framework within the existing literature by highlighting methodological differences and overall performance trends rather than providing a direct benchmark against previous methods. Peker and Kurtoglu [27] implemented YOLOv10 model performed well in tooth detection on pediatric panoramic images (F1 = 0.919, mAP@50–95 = 0.696) but was limited to a single task and showed reduced precision at higher IoU thresholds. Furthermore, Almalki et al. [28] proposed YOLOv3 approach reached 99.33% accuracy in identifying broad dental disease types but lacked fine localization and modern attention

mechanisms. In contrast, the proposed YOLOv11 model achieved an overall mAP@50–95 of 0.854 on the extended Dental OPG Kennedy dataset while incorporating C2PSA attention and refined Cross-Stage Partial blocks to jointly detect dental pathologies and arch-region characteristics within a unified multi-task framework. Although these results demonstrate strong detection performance, direct quantitative comparison with previous studies should be interpreted with caution because the published methods were evaluated on different datasets, annotation schemes, and experimental settings.

TABLE II. QUALITATIVE COMPARISON OF REPRESENTATIVE YOLO-BASED DENTAL DETECTION STUDIES AND THE PROPOSED YOLOV11 FRAMEWORK.

Study	YOLO Version	Main Task	Key Findings / Remarks
Peker & Kurtoglu [27]	YOLOv10	Tooth detection & numbering	Robust tooth numbering; limited fine localization accuracy
Almalki et al. [28]	YOLOv3	Common dental disease detection	High accuracy; limited to coarse pathology detection
Adnan et al. [29]	YOLOv5m	Six pathology detection (OPG)	Strong diagnostic accuracy; single-task framework
Putra DDS et al. [30]	YOLOv4	Tooth numbering	Excellent speed; limited to structural detection only
AbuSalim et al. [31]	YOLOv5, v6, v7	Multi-granularity tooth classification	Accuracy declines with class granularity; multi-model setup
AbuSalim et al. [32]	HDL (YOLOv6 +CBAM-YOLOv5)	Hierarchical tooth classification	High accuracy & recall; increased computational complexity
Proposed YOLOv11 (This Study)	YOLOv11	Multi-task detection (pathologies + arch classification)	Simultaneous multi-class and multi-region detection with superior localization and balanced metrics

Adnan et al. [29] used YOLOv5m to detect six dental pathologies with strong accuracy (Precision 94.22%, Recall 90.42%, mAP 93.71%), though its application was limited to pathology detection alone. In comparison, the proposed framework extends this capability by simultaneously integrating lesion identification with multi-region arch classification within a single detection model. On the extended ten-class Dental OPG Kennedy dataset, the proposed YOLOv11 framework achieved an overall F1-score of 0.961, demonstrating balanced detection performance across pathological and anatomical categories. Putra DDS et al. [30] employed YOLOv4 for automated tooth numbering, achieving high recall (100%) and rapid inference (20.58 ms per image), but their system did not incorporate pathological analysis. Likewise, AbuSalim et al. [31] [32] investigated multi-granularity detection using YOLOv5–YOLOv7 and later developed a hierarchical deep learning (HDL) approach combining YOLOv6 with a CBAM-enhanced YOLOv5, achieving mAP = 0.88, Recall = 0.999, and F1 = 0.94. Although effective, their multi-stage framework increased computational complexity. In contrast, the proposed YOLOv11 framework unifies structural and pathological detection within a single-stage architecture and achieved an overall mAP@50–95 of 0.854 and F1-score of 0.961 on the extended Dental OPG Kennedy dataset. In addition to these quantitative results, the framework performs simultaneous pathology detection and arch classification in a single inference pipeline, reducing computational complexity compared with previously reported multi-stage approaches. Table II shows comparative analysis of YOLO-based dental detection studies and the proposed YOLOv11 multi-task model.

A sample of the output results from the proposed framework are shown in Fig. 7.

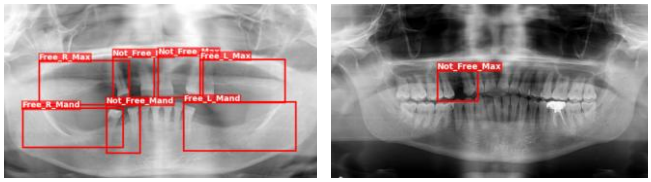


Fig. 7. Sample ot the output results.

VI. CLINICAL SIGNIFICANCE

The proposed multi-task YOLOv11-based framework has potential clinical value as a computer-aided decision support system for the analysis of panoramic dental radiographs. By simultaneously detecting multiple dental pathologies and identifying the corresponding maxillary and mandibular arch regions within a single inference pipeline, the framework can assist dentists in identifying abnormalities during routine clinical examinations, thereby supporting timely clinical decision-making and improving workflow efficiency. Automated localization of abnormalities may reduce the time required for radiographic interpretation, improve diagnostic consistency, and decrease the likelihood of overlooked findings, particularly in busy clinical environments. In addition, the framework may facilitate patient screening by automatically highlighting radiographs with suspected abnormalities, enabling clinicians to prioritize cases requiring further evaluation. The proposed framework may also serve as an educational aid for dental students and less experienced practitioners by highlighting clinically relevant regions of interest and reinforcing the interpretation of panoramic radiographs. Accordingly, the framework is intended to complement, rather than replace, professional clinical expertise, while final diagnosis and treatment planning remain the responsibility of qualified dental practitioners.

VII. CLINICAL INTERPRETATION OF PREDICTION ERRORS

In clinical practice, different prediction errors may have different consequences depending on the underlying dental condition. A false-negative prediction, in which an existing pathology is not detected, may delay diagnosis and treatment, potentially allowing disease progression. For example, missed root fractures or periodontally compromised teeth may result in continued structural damage, infection, or tooth loss if not identified during routine examination. Conversely, a false-positive prediction may lead to additional clinical examination or unnecessary radiographic review, increasing clinician workload but generally posing a lower clinical risk than missed pathology. Therefore, in screening applications, maintaining high sensitivity is particularly important to minimize overlooked abnormalities, while an appropriate balance between sensitivity

and precision is required to avoid excessive false alarms. The proposed framework is intended to assist clinicians by highlighting suspicious regions for further review rather than providing definitive diagnoses. Consequently, all AI-generated findings should be interpreted alongside clinical examination and professional judgment before treatment decisions are made.

VIII. Limitations

Even with its outstanding performance, there are still some drawbacks. The model experienced a reduction in accuracy when it comes to detecting small dental issues such as root fractures and periodontal problems because panoramic radiographs create low image contrast and overlapping anatomical structures. Furthermore, the proposed framework was evaluated exclusively on the extended Dental OPG Kennedy dataset. Validation on larger and more diverse datasets collected from multiple institutions and imaging systems is required to further assess its robustness, clinical applicability, and generalizability across different patient populations and acquisition conditions.

Although the proposed framework demonstrated strong detection performance (mAP@50–95 = 0.854 and F1-score = 0.961), the individual contribution of each design component—including the preprocessing pipeline, C2PSA attention mechanism, and multi-task learning strategy—was not evaluated independently through ablation experiments. Consequently, the performance improvements reported in this study should be interpreted as the result of the integrated framework rather than any single component.

The experimental evaluation was conducted using a single train-validation-test split and a fixed training configuration. Although this protocol provides a consistent and reproducible evaluation, it does not quantify the variability associated with different data partitions or training initializations. Consequently, confidence intervals, standard deviations, repeated experimental trials, and statistical significance analyses were not included in the present study. Future work will include repeated experiments using multiple random seeds, k-fold cross-validation, and external validation on independent multi-center datasets to further evaluate the robustness, reproducibility, statistical significance, and generalizability of the proposed framework.

IX. Conclusion and Future Work

This study introduced an automated multi-task deep learning framework based on the YOLOv11 architecture, designed for simultaneous detection of dental pathologies and arch-region classification in orthopantomograms (OPGs). To promote the system's efficiency while dealing with the extended Dental OPG Kennedy set comprising ten classes, the C2PSA attention modules coupled with the improved preprocessing techniques made the system produce outstanding results on all ten classes. The system reported an average mAP of 0.854, along with an F1-score of 0.961, exhibiting nearly flawless sensitivity while detecting well-demarcated edentulous areas. The upcoming work will focus on incorporating panoramic images from different centers into the database to boost the generality of the model instead of relying on domain specificity. Moreover, the application of the transformer modules, depending on self-supervised techniques, could advance the capture of higher-level semantics, especially in the identification of minor pathologies.

References

[1] D. A. Deenadayalan, K. Karthick, and N. Tamilarasu, "Orthopantomogram as a tool for morphometric study: Advancing dental diagnostics with a novel approach," Int. J. Adv. Res. Med., vol. 6, no. 3, pp. 94–97, 2024, doi: 10.22271/27069567.2024.v6.i3b.573.

[2] S. Hegde, J. Gao, R. Vasa, and S. Cox, "Factors affecting interpretation of dental radiographs," Dentomaxillofacial Radiol., vol. 52, no. 2, p. 20220279, Feb. 2023, doi: 10.1259/dmfr.20220279.

[3] S. AbuSalim, N. Zakaria, A. Saboor, A. Maqsood, S. J. Abdulkadir, and N. Mokhtar, "Deep Learning Methods for Tooth Detection and Classification in Various Dental Image Datasets: A Taxonomy and Future Directions," Int. J. Recent Innov. Trends Comput. Commun., vol. 12, no. 2, pp. 69–77, Apr. 2024.

[4] S. AbuSalim, N. Zakaria, M. R. Islam, G. Kumar, N. Mokhtar, and S. J. Abdulkadir, "Analysis of Deep Learning Techniques for Dental Informatics: A Systematic Literature Review," Healthcare, vol. 10, no. 10, p. 1892, Oct. 2022, doi: 10.3390/healthcare10101892.

[5] M. Ali, M. Irfan, T. Ali, C. R. Wei, and A. Akilimali, "Artificial intelligence in dental radiology: a narrative review," Ann. Med. Surg., vol. 87, no. 4, pp. 2212–2217, Mar. 2025, doi: 10.1097/MS9.0000000000003127.

[6] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," Apr. 23, 2020, arXiv: arXiv:2004.10934. doi: 10.48550/arXiv.2004.10934.

[7] Y. E. Almalki et al., "Deep Learning Models for Classification of Dental Diseases Using Orthopantomography X-ray OPG Images," Sensors, vol. 22, no. 19, p. 7370, Sep. 2022, doi: 10.3390/s22197370.

[8] B. Beser et al., "YOLO-V5 based deep learning approach for tooth detection and segmentation on pediatric panoramic radiographs in mixed dentition," BMC Med. Imaging, vol. 24, p. 172, Jul. 2024, doi: 10.1186/s12880-024-01338-w.

[9] S. AbuSalim, N. Zakaria, S. A. Mostafa, Y. K. Hooi, N. Mokhtar, and S. J. Abdulkadir, "Multi-Granularity Tooth Analysis via Faster Region-Convolutional Neural Networks for Effective Tooth Detection and Classification," Int. J. Adv. Comput. Sci. Appl. IJACSA, vol. 14, no. 6, Jun. 2023, doi: 10.14569/IJACSA.2023.0140678.

[10] A. F. Makarim, T. Karlita, R. Sigit, B. S. B. Dewantara, and A. Brahmanta, "Deep Learning Models for Dental Conditions Classification Using Intraoral Images," JOIV Int. J. Inform. Vis., vol. 8, no. 3, pp. 1099–1108, Sep. 2024, doi: 10.62527/joiv.8.3.1914.

[11] A. R. Ismail and M. S. S. Taseen, "Deep Learning Approach for Dental Anomalies X-ray Imaging using YOLOv8," Knowl. Eng. Data Sci., vol. 7, no. 2, pp. 164–175, Mar. 2025, doi: 10.17977/um018v7i22024p164-175.

[12] U. Şevik and O. Mutlu, "Detection of Dental Anomalies in Digital Panoramic Images Using YOLO: A Next Generation Approach Based on Single Stage Detection Models," Diagnostics, vol. 15, no. 15, p. 1961, Jan. 2025, doi: 10.3390/diagnostics15151961.

[13] S. Sukegawa et al., "Evaluation of multi-task learning in deep learning-based positioning classification of mandibular third molars," Sci. Rep., vol. 12, no. 1, p. 684, Jan. 2022, doi: 10.1038/s41598-021-04603-y.

[14] M. Prados-Privado, J. G. Villalón, C. H. Martínez-Martínez, and C. Ivorra, "Dental Images Recognition Technology and Applications: A Literature Review," Appl. Sci., vol. 10, no. 8, p. 2856, Jan. 2020, doi: 10.3390/app10082856.

[15] Y. Xiong et al., "Simultaneous detection of dental caries and fissure sealant in intraoral photos by deep learning: a pilot study," BMC Oral Health, vol. 24, no. 1, p. 553, May 2024, doi: 10.1186/s12903-024-04254-1.

[16] B. V, V. V. Giri, M. Dodia, P. R. Sanjaya, R. T. Arun, and S. Maity, "Machine Learning-Based Predictive Analysis of Dental Treatment Outcomes," in 2025 Tenth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), Nov. 2025, pp. 1–6. doi: 10.1109/ICONSTEM65670.2025.11374450.

[17] N. Pishghadam, R. Esmaeilyfard, and M. Paknahad, "Explainable deep learning for age and gender estimation in dental CBCT scans using attention mechanisms and multi task learning," Sci. Rep., vol. 15, no. 1, p. 18070, May 2025, doi: 10.1038/s41598-025-03305-z.

[18] J. A. Nava-Pintor et al., "Context-Aware Caries Segmentation in Periapical Radiographs Using a Hybrid Multi-Task Learning Framework with Partial Annotations," Appl. Sci., vol. 16, no. 1, p. 264, Jan. 2026, doi: 10.3390/app16010264.

[19] Z. Khurshid et al., "Multi-Regional deep learning models for identifying dental restorations and prosthesis in panoramic radiographs," BMC Oral Health, vol. 25, p. 1750, Nov. 2025, doi: 10.1186/s12903-025-07138-0.

[20] K. Zhang, J. Wu, H. Chen, and P. Lyu, "An effective teeth recognition method using label tree with cascade network structure," Comput. Med. Imaging Graph., vol. 68, pp. 61–70, Sep. 2018, doi: 10.1016/j.compmedimag.2018.07.001.

[21] "Dental OPG Kennedy Dataset." Accessed: Oct. 30, 2025. [Online]. Available: https://www.kaggle.com/datasets/orvile/dental-opg-kennedy-dataset

[22] S. Abusalim, S. Mostafa, N. Zakaria, S. Jadid Abdulkadir, and N. Mokhtar, Data Augmentation on Intra-Oral Images Using Image Manipulation Techniques. 2022, p. 120. doi: 10.1109/ICDI57181.2022.10007158.

[23] C.-Y. Wang, H.-Y. M. Liao, I.-H. Yeh, Y.-H. Wu, P.-Y. Chen, and J.-W. Hsieh, "CSPNet: A New Backbone that can Enhance Learning Capability of CNN," Nov. 27, 2019, arXiv: arXiv:1911.11929. doi: 10.48550/arXiv.1911.11929.

[24] R. Khanam and M. Hussain, "YOLOv11: An Overview of the Key Architectural Enhancements," Oct. 23, 2024, arXiv: arXiv:2410.17725. doi: 10.48550/arXiv.2410.17725.

[25] H. Tekis, T. Zirek, and M. Tassoker, "Comparative analysis of YOLO variants for dental implant brand identification," Sci. Rep., vol. 15, no. 1, p. 33304, Sep. 2025, doi: 10.1038/s41598-025-15383-0.

[26] S. Mureşanu et al., "Automating Dental Condition Detection on Panoramic Radiographs: Challenges, Pitfalls, and Opportunities," Diagnostics, vol. 14, no. 20, p. 2336, Jan. 2024, doi: 10.3390/diagnostics14202336.

[27] R. B. Peker and C. O. Kurtoglu, "Evaluation of the Performance of a YOLOv10-Based Deep Learning Model for Tooth Detection and Numbering on Panoramic Radiographs of Patients in the Mixed Dentition Period," Diagnostics, vol. 15, no. 4, p. 405, Jan. 2025, doi: 10.3390/diagnostics15040405.

[28] Y. E. Almalki et al., "Deep Learning Models for Classification of Dental Diseases Using Orthopantomography X-ray OPG Images," Sensors, vol. 22, no. 19, p. 7370, Jan. 2022, doi: 10.3390/s22197370.

[29] A. Adnan, M. Rizwan, H. Attaullah, S. Basheer, and M. Quasim, "Deep Architectural Classification of Dental Pathologies Using Orthopantomogram Imaging," Comput. Mater. Contin., vol. 85, no. 3, pp. 5073–5091, 2025, doi: 10.32604/cmc.2025.068797.

[30] R. H. Putra et al., "Automated permanent tooth detection and numbering on panoramic radiograph using a deep learning approach," Oral Surg. Oral Med. Oral Pathol. Oral Radiol., vol. 137, no. 5, pp. 537–544, May 2024, doi: 10.1016/j.oooo.2023.06.003.

[31] S. AbuSalim et al., "Multi-granularity tooth analysis via YOLO-based object detection models for effective tooth detection and classification," IAES Int. J. Artif. Intell. IJ-AI, vol. 13, no. 2, pp. 2081–2092, Jun. 2024, doi: 10.11591/ijai.v13.i2.pp2081-2092.

[32] S. AbuSalim, N. Zakaria, S. A. Mostafa, A. Maqsood, S. J. Abdulkadir, and N. Mokhtar, "Hierarchical deep learning model for tooth classification in intra-oral dental photographs," Neural Comput. Appl., vol. 37, no. 32, pp. 26659–26688, Nov. 2025, doi: 10.1007/s00521-025-11639-y.

Inclusive Digital Health Platform for Well-Being and Loneliness in SOGIESC Older Adults

Yada Atanan, Titikarn Assatarakul*, Akara Metasuk
Faculty of Learning Sciences and Education, Thammasat University, Thailand

***Abstract*—This research and development (R&D) project aims to design, test, and enhance a digital platform for assessing and promoting the health of late middle-aged and older adults who are diverse in sexual orientations, gender identities, gender expressions, and sex characteristics (SOGIESC). It displays people's opinions regarding a personal health reporting system that would make diverse SOGIESC late middle-aged and older adults (aged 50 years and above) feel better and less alone. Mental health, bodily health, social health, and intellectual or spiritual health are the four pillars of the holistic health framework around which the initiative is built. "Loneliness" is a significant factor that impacts elderly people's quality of life. The platform's user-centered design creates a secure, friendly, and inclusive online environment for individuals of all gender identities. The study employs a mixed-methods approach, which means that it deliberately blends quantitative and qualitative research techniques. The platform's structure and features were improved during the development process with the use of quantitative data from health evaluations and system usage statistics. Simultaneously, qualitative data was gathered through user behavior observation, prototype usability testing, and detailed opinion recording in order to examine user experiences in terms of comprehension and usage, design and user experience, system performance, and perceived value and advantages. According to the research findings, the platform can help diverse SOGIESC late middle-aged and older adults feel more socially connected, encourage self-evaluation, and raise health awareness. However, system complexity, early usage difficulty, system stability, and the requirement to create data and result presentations appropriate for the learning context and technological proficiency of the elderly were also identified as limits. With an average score of 4.10 and a standard deviation of 0.88, the overall satisfaction and suggestions showed that customers gave their level of satisfaction a high rating. According to this study, the creation of a digital health platform for senior citizens with a range of sexual orientations should give priority to straightforward, fair, and reliable designs that may produce definite experiential value in order to encourage consistent and long-term use. By providing policy and practical guidance for creating digital tools that effectively fulfill the needs of vulnerable older people in modern society, this research adds to the body of knowledge in comprehensive health technology research and development.**

***Keywords*—Digital health platform; friendly space; well-being and loneliness; LGBTQ; older adults**

I. Introduction

The number of older adults with diverse sexual orientations, gender identities, gender expressions, and sex characteristics (SOGIESC) is anticipated to increase due to demographic shifts and increased opportunities to disclose their identities following the country's same-sex marriage law, which was passed on June 18, 2024, given that nearly 20% of Thailand's population is 60 years of age or older [1]. A study on the relationships, mental health, and sexual health of older persons who are sexually and gender diverse in Thailand was carried out by Thammasat University's Faculty of Learning Sciences and Education between October 2023 and March 2024. In-depth interviews were used to collect qualitative data for the study, and surveys were used to collect quantitative data. The study's primary conclusions are as follows: 1) One in three middle-aged and older adults (aged 50 and above) with diverse SOGIESC lives alone, and what they desire most is a supportive circle of friends. 2) Most older adults aged 60 years and above with diverse SOGIESC still need to work, with many relying on employment as their primary source of income. Approximately one in three reported earning an income below or close to the national poverty line. 3) Many middle-aged and older adults with diverse SOGIESC still lack access to social networks and specialized knowledge related to sexual and gender diversity [2].

Given the circumstances, it is crucial to create friendly spaces for middle-aged and older adults with diverse SOGIESC. To connect the efforts of individuals, groups, or organizations working with the elderly and sexual minorities, there is a notion to construct a system that would provide friendly spaces that are communities for middle-aged and older adults with diverse SOGIESC. This seeks to give Thai society's sexually and gender diverse older adults a supportive community, access to mental health services, and the opportunity to build the connections, knowledge, and abilities needed for overall well-being. For middle-aged and older adults with diverse SOGIESC, a digital health platform has been developed and will be linked to the GenDHL website. This approach allows users to evaluate their physical, mental, social, and intellectual/spiritual well-being via a questionnaire. Individual evaluation results, including descriptions, recommendations, and the degree of well-being, will be provided after the assessment is complete.

II. Related Works

A. Main Concept: Creating a Friendly and Easy-to Understand Experience for the Elderly

The Digital Health Platform for Well-being, an online platform that explores loneliness and well-being among older adults with diverse SOGIESC, was created with the goal of providing a welcoming online environment where older adults with diverse SOGIESC can access services for self-care and health evaluation, including mental, social, spiritual, and physical health. The system analyzes reports on every facet and provides explanations and recommendations for additional growth in different areas, empowering older adults with diverse

*Corresponding author

SOGIESC to take care of themselves, recognize their own value, and cultivate positive relationships with others. Additionally, it uses the established platform to establish an online network of older adults with diverse SOGIESC.

The fundamental design tenets of this system are “simplicity" and “intuitiveness" to guarantee that people are not perplexed or uneasy when using technology [3]. Rather than merely serving as an assessment tool, the system is intended to be like “a friend who listens and gives advice".

B. User Experience Design (UX)

This system's user experience design mainly follows the principles of user-centered design, which emphasize problem-solving from the viewpoint of the user [4]. The goal is to provide users, especially older people with a variety of gender identities—with a welcoming and simple experience. To lessen cognitive load and avoid misunderstanding, Progressive Disclosure is used to convey information and functionality step-by-step in a methodical manner. For example, survey questions are displayed one at a time. Additionally, to prevent users from becoming lost during the usage process, responsive guidance provides clear, succinct explanations at each stage. Design takes into account the following important factors:

1) User-Centered Design (UCD): The User-Centered Design (UCD) approach, which stresses methodically comprehending the wants, experiences, and limits of the target group from the very beginning of development, is followed in the design of this platform [5]. The design method involves testing prototypes to determine usability and convenience as well as gathering qualitative data through in-depth interviews to investigate the expectations, worries, and technology usage experiences of older people with a range of sexual orientations. Testing is also used to compare performance in terms of comprehension and task completion time for certain UI components, like the arrangement of “Next" or “Back" buttons, the location of instructions, and the format of score displays. By analyzing data from these procedures, usability barriers are decreased, and the user interface is improved to better reflect the target group's real habits.

2) Progressive disclosure: The platform's architecture was based on the Progressive Disclosure principle, a user experience concept that attempts to reduce cognitive load by giving information or services in segments rather than all at once. On the evaluation site, for example, users will see one question at a time with clear control options like “Next" and "Back," instead of all the questions on one page. This approach reduces confusion, enhances focus while responding to inquiries, and allows for gradual progress, particularly for elderly people who may not be as tech-savvy. Additionally, because of its architecture, the system can provide additional explanations at specific periods without packing the screen.

3) Responsive guidance: Contextual advice has been incorporated into each screen of the system to aid in user decision-making and lessen usage uncertainty. For instance, showing brief instructions like “Press the 'Next' button to continue" or “Choose the answer that best fits you". These suggestions are meant to be unobtrusive and only show up when required. In addition to increasing the assessment's completion rate and lowering the risk of drop-off rates, this design gives people with little experience with technology confidence.

4) Receiving feedback and creating a positive atmosphere: To promote a positive attitude among the target demographic utilizing the system, a message offering recommendations and practical suggestions will be shown once the user has finished the assessment. The system will provide the results of the evaluation together with motivational statements and practical recommendations relevant to the assessment level. In order to keep users from feeling stigmatized or evaluated by the outcomes, this section is structured according to the positive reinforcement principle. The suggestion messages are written in straightforward, courteous language that demonstrates an awareness of the situation of senior citizens with a range of sexual orientations. This strategy is essential for encouraging recurring use and sustained platform acceptance.

5) Design for accessibility: Using the WCAG (Web Content Accessibility Guidelines) as a reference, the design will take into account senior citizens with visual problems. This involves assuring screen reader compatibility, utilizing alt language for images, and allowing touch-only system use [6]. To accommodate people with visual impairments, it's also critical to choose colors with the right contrast ratios. The platform design takes into account the accessibility concerns of senior folks who may have visual difficulties. Elements and buttons are made to be controlled by the keyboard or touch. These rules lessen physical obstacles to use and improve equity in system access.

6) Data structure design: Developing the data structure and organizing the thorough gathering and processing of information about loneliness and other facets of well-being to produce full and precise reports. A relational database structure is designed to allow the storing of user profile data, assessment response data, and the methodical computation of the loneliness index. The design of the data structure is an essential part of the system. This layout is intended to facilitate upcoming analytical processing and minimize redundant data (data normalization). To improve the accuracy of the results, data validation procedures are devised prior to the calculation process. To produce thorough, accurate results that can suitably support health-related decision-making, the generated reports rely on processed data.

C. User Interface (UI)

The goal of user interface (UI) design is to improve the pleasant experience that comes from UX design by making it visually appealing and easy to use. In order to create a cozy and user-friendly space, a tranquil and safe color scheme is used by using hues like white that express relaxation and safety. Bright hues that could be uncomfortable for consumers are also avoided [7]. In order to guarantee reading, clear, appropriately large fonts are also used with great consideration.

Other components that are important for effective design include four main aspects that affect the usability of the interface

design: functional framework, interaction logic, visual design, and user experience. These elements impact user interaction [8]. Such buttons and interactive features are made to be large and have colors that contrast sharply with the background to make them easier for older people with limited vision and motor skills to utilize. Additionally, intuitive icons that convey universal and obvious meanings are employed to help users use the system quickly and accurately [9]. Additionally, report visualization is offered in simple graphic styles, such as spider charts, along with succinct explanations and useful suggestions that assist consumers in understanding the data and using it to effectively improve their health [10].

D. Older Adults' Perceptions of Loneliness and Health (Well-Being and Reduced Loneliness)

Elderly people who suffer loneliness, anxiety, and a lack of social engagement are more likely to feel depressed and bored, which lowers their sense of self-worth and leads to social exclusion [11]. When people believe their social connections are inadequate in terms of both quantity and quality, they experience this complex emotional and psychological condition. This differs from social isolation, which refers to a lack of quantitative interactions. Therefore, rather than being a result of the quantity of people present, loneliness is an internal sensation connected to the perception and expectations of connections.

Retirement, the death of a spouse or close friend, chronic disease, and mobility issues are all factors that lead to a decline in social networks as people age. Several studies show that loneliness is linked to a higher risk of anxiety and depression as well as a reduction in general quality of life [12], [13]. Loneliness has also been connected to an increased risk of long-term conditions like heart disease, high blood pressure, and cognitive loss [14]. According to the World Health Organization's definition of well-being, which includes social, mental, and physical aspects [15], loneliness is seen as a factor that compromises health in all of these areas, especially in the mental dimension, which is linked to hopelessness, low self-esteem, and feelings of worthlessness. According to Hawkley and Cacioppo [14], loneliness can set off the body's chronic stress processes, which can have an impact on long-term health and the immune system. Therefore, developing meaningful relationships, establishing a sense of belonging, and encouraging important roles in the community should be the main goals of avoiding loneliness among the elderly. This is in line with the Active Aging concept of the World Health Organization, which prioritizes security, involvement, and health in old age [16]. These steps will improve general well-being and lessen the detrimental effects of loneliness on people as well as society.

E. Comparison with Existing Digital Health Platforms

As summarized in Table I, existing studies have predominantly addressed isolated dimensions of digital health for older adults, including human-centered design principles, interface usability, accessibility, health information visualization, and loneliness monitoring. Although these studies provide valuable design insights for age-friendly digital systems, the majority remain confined to conceptual frameworks or systematic reviews, without the implementation and empirical validation of a comprehensive digital health platform. Furthermore, prior research has largely concentrated on the general older adult population, while limited attention has been given to sexual and gender diverse older adults. In addition, existing systems rarely integrate multidimensional well-being assessment, loneliness evaluation, personalized health reporting, and inclusive interface design within a single platform. To address these gaps, the present study extends previous work by developing and evaluating a functional digital health platform specifically designed for SOGIESC older adults, employing a user-centered design approach, accessibility principles, and a mixed-method evaluation. This research gap underscores the originality and practical contribution of the present study to the emerging field of inclusive digital health.

TABLE I. COMPARISON OF RECENT STUDIES RELATED TO INCLUSIVE DIGITAL HEALTH PLATFORMS FOR OLDER ADULTS.

Research Study	Key Contribution and Research Gap	Advancement of the Present Study
Research Paper [17]	Systematic review of HCD processes for mHealth applications using the ISO 9241-210 framework. No operational platform was implemented or evaluated.	Implements HCD in an operational platform integrating well-being, loneliness assessment, and inclusive design for SOGIESC older adults.
Research Paper [18]	Systematic umbrella review of technology interventions for loneliness and social connectedness. Evidence remained limited despite promising ICT-based interventions.	Extends previous technology interventions through an operational platform integrating loneliness assessment, multidimensional well-being assessment, and personalized reporting.
Research Paper [10]	Systematic review of PROM visualization, focusing on interpretation and presentation formats. No operational platform was developed.	Integrates PROM visualization with multidimensional assessment, personalized interpretation, and health recommendations.
Research Paper [8]	Systematic review of telemedicine interface usability and design recommendations for older adults. No operational platform was implemented.	Applies age-friendly usability principles within an operational digital health platform validated through mixed-method evaluation.
Research Paper [9]	Systematic review of age-friendly mobile app design, accessibility, and user-centered design strategies. No operational platform was evaluated.	Translates age-friendly design principles into an inclusive platform with multidimensional assessment and personalized reporting.
Research Paper [19]	Developed a user-centered interface design framework for loneliness monitoring. Focused on design indicators without platform implementation.	Extends the framework into an operational platform integrating loneliness, well-being assessment, and mixed-method evaluation.
Present Study	Develops an inclusive digital health platform integrating validated assessment instruments, user-centered design, multidimensional well-being assessment, and personalized reporting. Preliminary mixed-method evaluation involved 10 participants.	Provides an operational platform for SOGIESC older adults integrating inclusive design, validated assessment, and mixed-method evaluation.

Existing literature provides substantial evidence that user-centered design has become the prevailing framework guiding the development of digital health technologies for older adults. A considerable number of studies have identified age-friendly interface design, accessibility, usability testing, and sustained user participation as key determinants of successful technology adoption within aging populations [8], [9], [19]. Collectively, these contributions have deepened the theoretical and practical understanding of human-computer interaction in digital health, while simultaneously establishing core design principles that enhance system accessibility and usability. In addition, recent systematic reviews have emphasized the increasing significance of embedding digital health technologies into community-based healthcare frameworks as a strategy for promoting healthy aging and sustaining independent living among older adults.

Notwithstanding these advances, several critical limitations persist within the existing body of research. Most current platforms tend to emphasize a single dimension such as usability enhancement, interface optimization, or disease-specific monitoring rather than incorporating a multidimensional well-being assessment within a unified digital environment. Loneliness, in particular, is commonly examined as a standalone psychological construct rather than being evaluated in conjunction with broader physical, mental, social, and spiritual well-being dimensions. Furthermore, although health disparities among sexual and gender diverse older adults have gained increasing recognition, only a limited number of digital health studies have explicitly addressed their distinct psychosocial characteristics, healthcare experiences, and identity-related needs. As a result, empirical evidence supporting the development of inclusive digital health platforms tailored to SOGIESC older adults remains scarce.

III. Methodology and System Architecture

This study employed a research and development (R&D) methodology integrated with a mixed-methods design to guide the architectural design, implementation, and evaluation of an inclusive digital health platform for well-being and loneliness assessment among sexual and gender diverse older adults. The platform was developed following a user-centered design framework to ensure accessibility, inclusivity, and usability while addressing the psychosocial well-being needs of the target population. Quantitative and qualitative data were collected concurrently, analyzed separately, and integrated during interpretation to provide a comprehensive assessment of system performance, usability, and user experience.

Participants consisted of ten Sexual and Gender Diverse older adults in Thailand (n = 10) recruited through purposive sampling. A purposive sampling strategy involving ten participants was adopted, given that the study centered on formative usability evaluation within a narrowly defined and inherently difficult-to-reach population, namely Thai SOGIESC adults aged 50 years and above. This sample size was deemed sufficient for identifying salient usability issues and eliciting in-depth qualitative insights during the early stage of system development. However, it should be noted that the sample was not statistically powered to estimate population parameters.

Quantitative data were gathered using usability and satisfaction assessments, whereas qualitative data were collected through observation, think-aloud usability testing, and semi-structured interviews. The research process encompassed user needs assessment, system architecture design, platform development, implementation, and evaluation. Findings from both data strands were integrated throughout the development cycle to support iterative system refinement and methodological triangulation. The resulting evidence informed improvements in system functionality, accessibility, and user experience.

Quantitative and qualitative findings were integrated at the interpretation stage to enhance the credibility of the evaluation and to provide complementary evidence regarding the usability, user experience, and perceived usefulness of the proposed platform.

The methodological procedures adopted in this study are presented in the following sections.

A. Development of the Digital Health Platform on the GenDHL Website

Users of mobile devices or internet-connected PCs can access the individual health assessment reporting system, which is designed to facilitate usage through the GenDHL website. SOGIESC adults aged 50 years and above are the main target demographic. The system gathers user data, processes it, and centrally stores it (see Fig. 1). With “loneliness" as the primary focus of the analysis, encompassing both emotional and social loneliness, this comprises health assessments in four primary dimensions: physical health, mental health, social health, and spiritual health. In order to systematically synthesize results and produce individual health assessment reports that include a health overview, assessment results, and pertinent suggestions, all data is processed in a computational storage system. In terms of policy and practice, the system can be used to successfully plan, promote, and enhance the target group's quality of life. It also strives to support a thorough understanding of users' health.

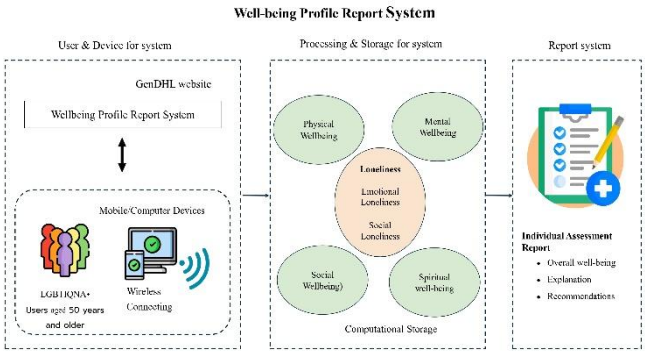


Fig. 1. Framework for developing the Digital Health Platform on the GenDHL website.

B. System Development Process

The PDCA principle, also known as the “Deming Cycle”, was used to identify the system development process for the online platform system, Digital Health Platform for Well-being, which is a management method for consistently enhancing productivity. It can be used at any point in the system development process to guarantee that the final product is excellent and genuinely satisfies the requirements. There are four primary steps: 1) Making a plan, 2) Acting (Act), 3)

Verifying and Assessing (Verify), and 4) Improving and Acting (Act) [20], which includes the following information.

1) *Step 1: Planning (Plan):* The planning process includes the following details:

- Needs analysis: To comprehend the actual difficulties and requirements of the target user group, which is made up of senior citizens with a variety of sexual orientations, examine their needs and context.
- Describe the primary idea: Create the system with the idea of "an online-friendly space" in mind, emphasizing convenience, security, and an emphasis on comprehensive healthcare. This covers the key ideas: making an approachable and simple-to-understand experience for senior citizens, user experience (UX) design concepts, user interface (UI) design concepts, and concepts pertaining to senior citizens' loneliness and well-being.

2) Step 2: Implementation (Do): The first step's development plan was implemented during the implementation phase (Do), with an emphasis on developing a comprehensive system that achieves the predetermined goals. A holistic health survey form comprising four domains physical, mental, social, and spiritual health was developed as part of the development of core functions. These were created utilizing a thorough and simple inquiry format. A processing system and report generation were built in tandem with the creation of key functions; comprehensive data processing was used to produce concrete reports, and coding was used to generate scores in every aspect of health. Score summaries, general findings, and useful suggestions for successful self-improvement were all included in these reports. Additionally, the front-end was created with an emphasis on usability and aesthetics in accordance with UX/UI design, enabling customers to view previous reports and clearly identify changes in their health. In conclusion, the following planned phases were followed in the development of the platform.

- Develop the primary function: Construct a system for health surveys that addresses all four aspects.
- Create the system for processing and reporting: Write code to process each area's results and produce a comprehensive report with clear justifications and suggestions.
- Create the front-end by designing the displays in accordance with the intended UX/UI so that users may observe changes in their health status and view previous reports.

In order to produce comprehensive and accurate reports, the first stage of system design will involve designing the data structure and planning for the gathering and processing of data pertaining to loneliness and other aspects of well-being. The GenDHL website's online platform technology is depicted in an example image.

The workflow of the Digital Health Platform for Well-Being will be depicted in the graphic, beginning with the login procedure and segmenting users into administrators (Admin) and senior citizens. Target users complete a four-dimensional health evaluation that covers mental, physical, social, intellectual, and spiritual dimensions in addition to personal information and activities. Before entering the data into the database and processing it to produce personal health reports, the system verifies that the data is accurate. In the meantime, the administrator is in charge of overseeing and validating the data to efficiently support activity development and long-term health monitoring. Fig. 2 presents the data structure diagram.

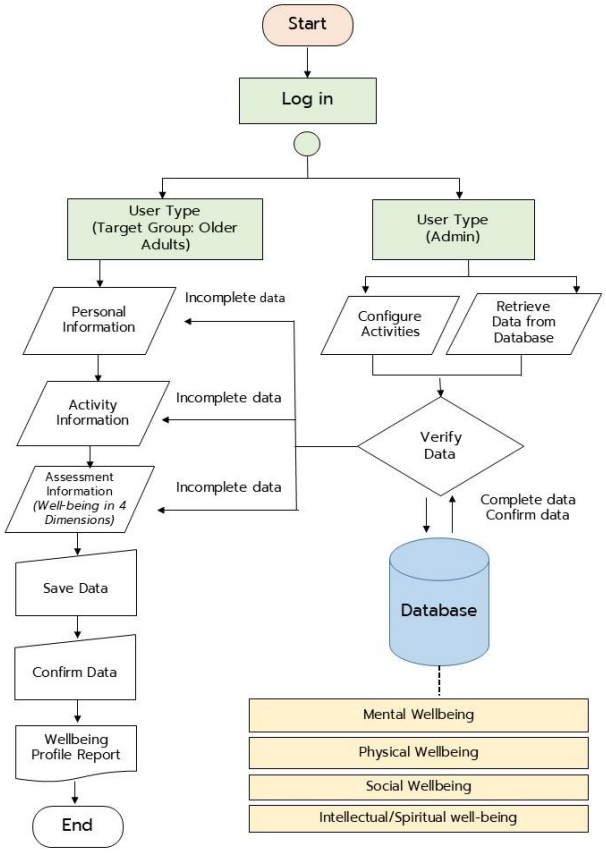


Fig. 2. Data structure diagram for the online platform system on the GenDHL website.

The emphasis will be on user-centered design in the design of the user interface and user experience (User Interface & User Experience), taking into consideration the limitations of the elderly through the use of clear iconography, eye-friendly colors, and appropriate text sizes. This is to guarantee that using a smartphone is as convenient and natural as possible, taking into account the following important factors:

- Emphasis on user-centered design: Begin by thoroughly comprehending the requirements and habits of the intended audience. This can be achieved by conducting in-depth interviews to obtain input on user experience or by comparing button placements to determine which layout works best for the target population.

- Progressive disclosure: The technology will show functions and information gradually rather than all at once. For instance, in order to minimize confusion and facilitate data submission, users will only view one question at a time rather than all ten on the assessment page.
- Responsive guidance system: To keep users from getting lost, each screen will have brief instructions that inform them what to do next, like “Select the answer that best matches you" or “Press the 'Back' button to go to the previous page".
- Getting feedback and fostering a good attitude once the user finishes the evaluation. Messages offering helpful guidance and recommendations will be shown to boost the system's target users' positive emotions.

The idea of modular deconstruction has been used in the system architecture and modular design of the inclusive digital health platform to improve future system scalability, maintenance flexibility, and development clarity. Five primary modules comprise the framework of the system: 1) Authentication Module; 2) User Profile Management Module; 3) Assessment Engine Module; 4) Report Generation Module; and 5) Admin Control Module. Workflow-driven integration connects each module, each of which has a distinct function.

- *The module for access control and authentication:* In order to clearly distinguish the use rights between “ordinary users" and “administrators (Admin)", the Authentication Module manages the process of safely accessing the system, including login, logout, and role-based access control. This module facilitates the governance of sensitive health data in compliance with user privacy and personal data protection rules and helps lower the risk of improper data access.
- Module for managing user profiles: In order to provide a health assessment context and enable user identity modification through the development of avatars, the user profile management module is in charge of gathering and documenting personal data. In an all-encompassing digital environment, this is an essential component for guaranteeing a sense of security and ownership. The module facilitates the creation, updating, and retrieval of profile data, which lessens the strain of repetitive data entry, improves the continuity of the user experience, and permits the reuse of basic data in later evaluations.
- Module for the assessment engine: The foundation of the platform's business logic is the assessment processing module. It operates by obtaining input data from loneliness-related and health assessments, validating the input, processing scores in accordance with predetermined criteria, and computing key indicators like the loneliness index to summarize the risk or condition level based on the assessment results. In order to improve calculation accuracy and make it easier to develop or modify calculation criteria in the future without impacting the user interface, this module is made to be independent of the presentation layer.
- Module for report generation: The purpose of the Report Module is to transform the Assessment Engine's evaluation results into a user-friendly manner. In order to facilitate meaningful self-review and health awareness, this involves producing summary reports that synthesize the evaluation data into interpretive summaries and visualizations like graphs or summary diagrams. Furthermore, in situations where the system records repeated evaluation findings that represent trends in health changes, the report module can be made to provide longitudinal comparisons.
- Module for admin control: Access to statistical data (monitoring and analytics), general usage monitoring, and the management of platform-related actions or content (activity and content management) are all supported by the admin control module. This module is essential for preserving data quality, ensuring service continuity, and overseeing community components or initiatives related to advancing the target group's well-being. To guarantee the security and suitability of data management, the Authentication Module controls the access privileges of administrators.

Fig. 3. Show a preview of the login page.

Fig. 4. Display the main menu of the system.

Fig. 5. Show your avatar through personal identity.

Fig. 6. Show the results of the health comparison.

Additionally, the design places a strong emphasis on fostering a positive environment by showcasing uplifting messages and practical guidance following the exam, encouraging a positive outlook on one's own health. The design adheres to the web content accessibility guidelines requirements for complete accessibility for users [21]. Lastly, the information architecture has been carefully designed to guarantee the accurate and thorough processing, storage, and display of reports on loneliness and other elements of well-being (see Fig. 3 to Fig. 6).

3) Step 3: Check and Evaluation: The proposed platform was assessed through a pilot implementation that emphasized usability, user experience, and system feasibility, rather than predictive model validation. Given that the assessment engine relies on deterministic scoring rules derived from validated assessment instruments, benchmark datasets typically employed in machine learning evaluation were deemed inapplicable in this context. Accordingly, the evaluation prioritized user satisfaction, usability, and the practical applicability of the platform within the target population.

Before being implemented in the real-world, the designed system will be thoroughly tested with real user groups during the Check phase to assess its accuracy and performance. Usability testing, which focuses on gathering behavioral data from the target population of older users to evaluate the system's usability, specifically in terms of navigation comprehension and the clarity of the instructions displayed on the screen, will comprise the three main components of the operations.

Arrange for the report's results to be verified in order to make sure that the suggestions and score processing are both academically correct and helpful to real users.

Plan to interview customers in-depth to learn about their overall experience, how they feel about using the system, and any further needs they may have. The data gathered from this verification procedure is essential for enhancing and expanding the system in the next stages. In conclusion, the system was tested with a sample of real users after development was finished, as described below:

- To determine whether the system is genuinely user-friendly and to pinpoint areas that require development, such as navigation comprehension and instruction clarity, usability testing gathers data from the target group of senior users.
- Report accuracy verification: Verify that the information provided is valid and helpful by checking the accuracy of the score processing and the suggestions in the report.
- Gathering feedback: Conduct user interviews to learn about their thoughts on the system as a whole, their experiences, and any further requirements.

4) Step 4: Improvement and Implementation (Act): In order to improve and enhance the system's quality, the assessment results and input gathered from the Check phase were examined in the Act phase, the last stage of system development. One of the primary responsibilities was to improve the UX/UI, which included making improvements to displays and usage procedures when users were reported to be confused or not comprehending. This was carried out to guarantee a more seamless and user-friendly navigation method. The report's recommendations and material were also broadened to offer greater diversity, thoroughness, and usefulness.

Simultaneously, the next phase's development goals have been outlined, providing guidance for potential additions of additional features like the creation of a notification system, online community activities, or user-expert connections. This PDCA cycle's implementation guarantees that the system will be continuously improved to sustainably satisfy user needs rather than staying stagnant.

- The data from step 3's inspection has been examined to further enhance the system, as explained below: Enhance UX/UI: Update screens and usage instructions that older consumers still find unclear or difficult to comprehend.
- Expand information and recommendations: To make the report more thorough and beneficial, add a wider range of recommendations.
- Make a plan for future development: Describe how you intend to add new features in the future, including expert connections, online community activities, or notification systems.

C. Well-being and Loneliness Assessment Instrument

The proposed platform integrates five validated assessment instruments to evaluate multidimensional well-being and loneliness among SOGIESC older adults. Loneliness was assessed using the 6-item De Jong Gierveld Loneliness Scale, which measures emotional and social loneliness [22]. Mental well-being was evaluated using the Thai Geriatric Mental Health Assessment Tool–Short Form (T-GMHA-15), a culturally validated instrument for Thai older adults that assesses mood, anxiety, self-esteem, and cognitive-emotional stability [23]. Functional ability was assessed using items adapted from the Functional Activities Questionnaire (FAQ), which measures independence in instrumental activities of daily living [24]. Social well-being was measured using a 15-item instrument based on Keyes' Social Well-Being framework, covering social actualization, social acceptance, social coherence, social contribution, and social integration [25]. Spiritual well-being was assessed using the Functional Assessment of Chronic Illness Therapy Spiritual Well-Being Scale (FACIT-Sp-12), which evaluates the domains of meaning, peace, and faith [26].

D. Assessment and Report-Generation Workflow

The proposed platform utilizes a deterministic, rule-based assessment engine to process user responses, rather than a trainable machine-learning model. As depicted in Fig. 7, the workflow commences with user authentication, followed by the completion of multidimensional assessment instruments. Subsequently, the system verifies response completeness and value validity prior to applying instrument-specific scoring rules, including reverse scoring where applicable. Domain scores encompassing loneliness, mental well-being, functional ability, social well-being, and spiritual well-being are then computed in accordance with the validated scoring procedures established for each respective instrument.

The resulting scores are subsequently mapped onto predefined interpretation criteria to generate personalized health reports, graphical summaries, and formative recommendations. User responses and assessment outcomes are securely stored within the database to facilitate future retrieval and longitudinal monitoring. As the assessment engine operates entirely on rule-based logic, no model training, hyperparameter tuning, or optimization procedures are necessary. Instead, the operational parameters comprise predefined scoring rules, permissible response ranges, reverse-scoring procedures, and interpretation criteria derived from their corresponding validated assessment instruments (see Fig. 7).

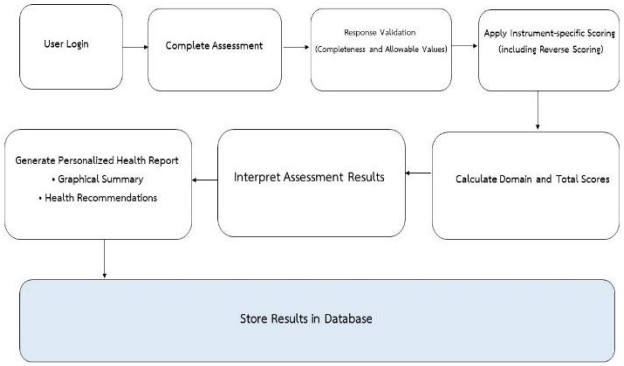


Fig. 7. Workflow of the rule-based assessment and report-generation process implemented in the proposed digital health platform.

E. Data Collection

Data collection was conducted using a mixed-methods approach comprising usability testing, participant observation, a structured satisfaction questionnaire, and semi-structured interviews. During the usability testing sessions, participants performed a set of predefined tasks on the GenDHL platform while researchers observed their interactions and documented emerging usability issues. Following task completion, participants completed a satisfaction questionnaire and subsequently engaged in semi-structured interviews to elicit in-depth feedback regarding their user experience, perceived usefulness, and recommendations for system enhancement.

F. Data Analysis

The evaluation aimed to conduct a formative usability assessment of the proposed platform rather than a confirmatory effectiveness study. Accordingly, the analysis emphasized descriptive statistics and qualitative interpretation to identify usability issues, explore user experiences, and generate recommendations for iterative system refinement.

Given that this study was designed as a preliminary pilot evaluation involving a small purposive sample, quantitative data were analyzed descriptively using frequencies, percentages, means, and standard deviations, supplemented where appropriate by medians and ranges. No inferential statistical tests were performed, as the study was not intended to support population-level statistical generalization. Rather, the quantitative findings were designed to provide preliminary descriptive evidence of participants' perceptions. Qualitative data derived from participant observations and semi-structured interviews were analyzed thematically and employed to explain, contextualize, and enrich the interpretation of the quantitative satisfaction ratings.

Qualitative data obtained from participant observations and semi-structured interviews were examined through thematic analysis. The resulting themes served to explain, contextualize, and enrich the interpretation of the quantitative findings, thereby facilitating a comprehensive understanding of participants' usability experiences and their perceived value of the proposed platform. The thematic analysis adhered to the procedural stages of familiarization, initial coding, theme development, theme review, and theme refinement in order to identify recurrent usability issues and user experience patterns [27].

IV. Results and Discussion

The following specific conclusions about the system testing results were found after examining usage observations and conducting user interviews to get input on the system as a whole, user sentiments, and further requirements:

A. Usability and Learnability

Although the majority of users, particularly the elderly, were initially unfamiliar with the system, they were able to comprehend the platform's goals. The process of registering, logging in, and inputting personal data was perceived as needing constant supervision or teaching from others throughout the system's trial use. Many users noted that the system might constitute a barrier to long-term self-use in the absence of support or an understandable manual, particularly for older people living alone. Here are some thorough remarks:

"Overall, the system is not difficult to use. You can learn and understand it quite quickly on your own. Although at first you might need to try clicking around and exploring the various menus, after using it for a while, you will start to get familiar with it and be able to use it more smoothly."

(Participant A, November 13, 2025)

"Overall, I feel that the system is quite user-friendly. The text and various elements are easy to read, making it possible to see the information clearly and helping to understand how to use it without confusion."

(Participant C, November 13, 2025)

B. Interface and Experience Design

Users' perceptions vary; some think the website is well-organized and simple to read. However, a prevalent problem is that icons and pictures often fail to adequately represent sexual variety, making it more difficult for users to comprehend and learn how to utilize the system. Here are a few specific instances of the comments:

"The main functions of the system are quite clear, making it easy to know where to start using it. Each menu is arranged in order, so it doesn't make you feel confused or lost while using it."

(Participant B, November 13, 2025)

"The presence of avatars in the system enhances interest and a sense of friendliness, making the overall atmosphere more enjoyable and helping to reduce tension during use."

(Participant D, November 13, 2025)

"The icons used in the system are quite clear in their meanings, allowing users to understand the functions of each part without much guessing, making it more convenient and seamless to use."

(Participant E, November 13, 2025)

C. Functionality and Performance

The system can function in accordance with fundamental goals, like report generation and health assessments. Nevertheless, there are still problems with system speed and stability, like sluggish data loading and system freezes at specific stages. A step-by-step usage guide, voice command functionality, audio explanations for inquiries, and the elimination of unnecessary steps in data entry are just a few of the enhancements that users have proposed to help the elderly. The target group would find it easier to use and more convenient as a result. Here are some thorough remarks:

"Overall, I feel that the system is not difficult to use. The various steps are quite understandable on their own. Even though it was my first time trying it out, I was able to learn gradually. After using it for a while, I started to get familiar with the menus and the system's operation, allowing me to continue using it without much hassle."

(Participant A, November 13, 2025)

"I feel that the system isn't complicated at all. The usage is quite straightforward, making it easy to understand, even without prior experience with such systems. Once you try clicking through the menus one by one, you can figure out how to use it on your own."

(Participant C, November 13, 2025)

"I feel that the system isn't very complicated. At first, it might take some time to understand, but after using it for a while, I think I'll start to get familiar with it and be able to use it more smoothly. The steps aren't very difficult and can be learned on my own."

(Participant D, November 13, 2025)

D. Usefulness and Value

In terms of encouraging introspection, raising awareness of health issues, and reflecting on loneliness and quality of life, the majority of users found the system beneficial. In order to make the evaluation results useful in everyday life, users have recommended that the results be presented more simply, with more visuals or symbols and a focus on concise explanations. Here are some thorough remarks:

"I feel that the system helps me reflect on my thoughts and feelings more, allowing me to revisit certain aspects of myself that I usually don't have the opportunity or importance to consider."

(Participant A, November 13, 2025)

"When I saw the results from the system, it made me more aware and accepting of my own feelings. I felt that I might actually be lonely, which is something I had never really noticed or thought about seriously before."

(Participant F, November 13, 2025)

"It only takes a few minutes to receive information or results related to oneself, making it feel convenient and quick, and helping to understand oneself better in a short period."

(Participant G, November 13, 2025)

E. Overall Satisfaction and Suggestions

Overall, the data shows that the degree of system comprehension and platform stability are closely correlated with user happiness. The platform has a great chance of becoming a well-known and regularly utilized digital health tool across the

diverse senior population if user experience and technical performance continue to improve. Here are some thorough remarks:

"If the system continues to develop and improve in this manner, I think there will be interest and a desire to use it again in the future."

(Participant B, November 13, 2025)

"Using the system creates a feeling as if there is something that understands one's own feelings, resulting in a greater reflection and awareness of one's own emotions."

(Participant C, November 13, 2025)

In conclusion, the study's qualitative behavioral data show that digital systems are more than just technical tools; they also contribute to self-awareness, emotional introspection, and the evaluation of users' mental health. The framework of digital mental health and well-being, which prioritizes a user-centered experience, is consistent with this.

According to the satisfaction survey results, the overall satisfaction score reported by this pilot cohort yielded a descriptive mean of 4.10 (SD = 0.88). Such values provide an initial overview of participants' perceptions of the platform; nevertheless, they should not be construed as population-level estimates or as evidence of the platform's effectiveness. All things considered, the system's merits were emphasized in terms of precisely identifying target groups and initiatives to establish a secure environment for senior citizens of various sexual orientations. At the same time, it was unanimously agreed that future system development should concentrate on making it simple to use right away, speeding up system output, modifying the design to make it suitable for the elderly, and providing clear value and incentives to encourage users to keep using it and to maximize benefits in terms of policy development and quality of life. The evaluation's findings will serve as a reference for future system efficiency improvements (see Fig. 8).

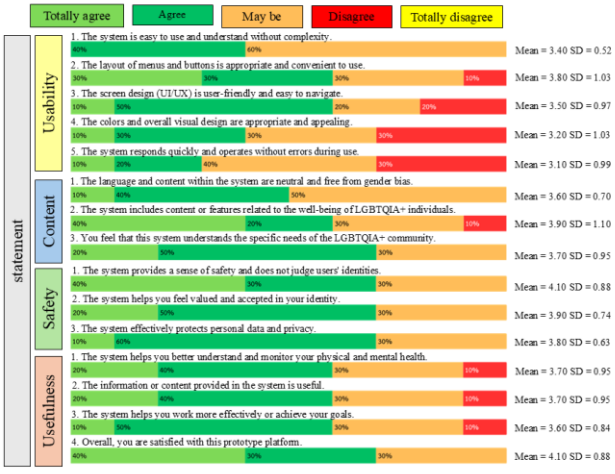


Fig. 8. A summary of the satisfaction survey findings for the Digital Health Platform system from a sample of ten participants.

System usage behavior can be explained within the frameworks of digital health and mental well-being, according to an overview of the study of quantitative data about opinions on the system and qualitative data from research participant interviews. These concepts highlight how technology can be used in daily life to promote self-awareness, emotional introspection, and mental health services [28]. Some respondents mentioned that utilizing the system makes people more conscious of their emotional states and feelings. System users engage with the system in a way that is similar to self-reflection. This conduct is consistent with the idea of self-awareness, which is an essential part of mental health and a basic process that makes preventive mental health care possible. When it comes to using technology for mental health (digital mental health), informants assess the system based on their real-world usage experiences and the emotions they experience while using it. It also illustrates the system's function as an emotionally secure environment where users can evaluate themselves and consider their emotions without being directly pressured by others. This trait is consistent with the idea of "digital well-being", which holds that user-considered technology can promote mental health without creating stress or emotions of condemnation [29].

V. Conclusion

This study successfully developed and preliminarily evaluated an inclusive digital health platform embedded within the GenDHL website, designed to support well-being and loneliness assessment among late middle-aged and older adults with diverse sexual orientations, gender identities, gender expressions, and sex characteristics (SOGIESC). The evaluation findings indicated that participants generally perceived the platform as comprehensible, user-friendly, and beneficial in promoting self-awareness, emotional well-being, and health literacy. Through its intuitive digital environment, the platform enables users to assess their physical, mental, social, and spiritual well-being while receiving personalized health reports and practical recommendations. Throughout the system development process, particular emphasis was placed on the application of user-centered design principles to ensure simplicity, intuitive interaction, and ease of use for older adults. This design approach aligns with the user-centered interface framework proposed by Zhou et al. [19], which emphasizes the necessity of minimizing interaction barriers and enhancing usability in digital health systems designed for older adults. Building upon these user-centered principles, the present study further incorporates inclusive design considerations and multidimensional health assessment to address the specific psychosocial and healthcare needs of SOGIESC older adults.

The positive usability outcomes observed in this study, together with participants' qualitative reflections, can be attributed to the integration of user-centered design principles, accessibility oriented interface design, progressive disclosure, and personalized health feedback embedded throughout the platform. Prior studies have shown that simplified navigation, age-friendly interface components, and accessible interaction substantially enhance usability among older adults [8], [9], while Zhou et al. [19] emphasized that user-centered interface design constitutes an essential foundation for reducing interaction barriers and fostering user engagement in digital health systems. In contrast to previous studies, which primarily concentrated on interface usability, age-friendly design guidelines, or conceptual design frameworks, the proposed

platform integrates multidimensional well-being assessment, loneliness evaluation, personalized feedback, and inclusive identity representation within a single operational environment specifically tailored for SOGIESC older adults. This comprehensive integration may account for the favorable usability, positive emotional responses, and enhanced self-reflection reported by participants during the pilot implementation, thereby suggesting that combining technical usability with psychosocial support can strengthen user engagement and perceived value.

These findings further indicate that one should not assess the effectiveness of inclusive digital health platforms solely in terms of technical usability or system performance. Rather, psychosocial dimensions including identity recognition, emotional support, perceived inclusiveness, opportunities for self-reflection, and social connectedness appear to be an equally important part in determining users' acceptance and sustained engagement. This finding emphasizes the necessity of designing digital health systems that simultaneously address technological accessibility, psychological well-being, and social inclusion. Such an integrated perspective extends current digital health research by illustrating that inclusive platform design should move beyond conventional health monitoring toward fostering holistic well-being among underserved populations, particularly sexual and gender-diverse older adults.

This study advances the nascent field of inclusive digital health by providing empirical evidence that user-centered design, inclusive design principles, and multidimensional health assessment can be effectively amalgamated within a single digital platform. Compared with existing state-of-the-art digital health platforms reported by Nimmanterdwong et al. [17], Amouzadeh et al. [9], and Zhou et al. [19], the proposed system extends prior approaches by incorporating multidimensional well-being assessment, loneliness evaluation, personalized health reporting, and inclusive interface design within a single operational platform. Whereas previous studies have predominantly emphasized user-centered design, accessibility, technology acceptance, or isolated health dimensions among older adults, relatively few have concurrently addressed comprehensive health assessment alongside the distinct psychosocial and healthcare needs of sexual and gender-diverse older adults. Accordingly, the proposed platform advances current digital health research by combining validated assessment instruments with an operational, user-centered, and inclusive digital health service specifically designed for SOGIESC older adults. Unlike previous studies that mainly suggested conceptual frameworks, usability guidelines, or standalone assessment tools, this research illustrates the practical viability of deploying and assessing an operational digital health platform via a mixed-methods approach. Therefore, the proposed platform contributes not only a practical technological solution but also an evidence-based framework capable of guiding future digital health service development for marginalized aging populations. These findings may further assist healthcare professionals, community organizations, and policymakers in promoting healthy aging, mitigating loneliness, and enhancing equitable access to digital health services.

Despite these encouraging findings, several limitations warrant acknowledgment. First, the study was conducted as a pilot implementation involving a relatively small sample of ten participants, thereby constraining the generalizability of the quantitative findings. Second, the evaluation primarily concentrated on usability, user experience, and preliminary user perceptions rather than long-term behavioral outcomes, clinical effectiveness, or longitudinal health impacts. Third, the proposed platform was not benchmarked against existing digital health platforms or validated using standardized benchmark datasets, thereby limiting direct comparison with alternative approaches. In addition, the current assessment engine remains rule-based and lacks predictive analytics or artificial intelligence techniques, while computational performance, scalability, and large-scale deployment were not examined. These limitations, when interpreting the findings, are crucial to consider, as they provide important context and highlight significant directions for future research.

Addressing these limitations offers several important directions for future research and system enhancement. Subsequent development should adopt a more modular software architecture by decoupling the assessment engine from the presentation and data layers, thereby improving maintainability, interoperability, and integration with external healthcare services through standardized APIs. System performance should be further strengthened through response time optimization, concurrent user testing, database optimization, and cloud-based deployment to support broader implementation. Future studies should also explore the integration of artificial intelligence and predictive analytics to enable adaptive and personalized health recommendations, while validating the platform with larger and more diverse participant groups across multiple healthcare settings. Such developments would facilitate the transition of the current research prototype into a reliable, secure, scalable, and sustainable digital health platform suitable for real-world deployment. Overall, the findings of this study establish a foundation for the future development of inclusive, evidence-based, and scalable digital health ecosystems capable of supporting healthy ageing and reducing health disparities among sexual and gender diverse older adults.

ACKNOWLEDGMENT

The researcher would like to sincerely thank all the informants, who are older adults who identify as sexual and gender diverse, for collaborating, taking the time to participate in system testing, giving detailed comments, and genuinely commenting on their platform experiences. The platform has been developed and improved to better suit the target group's needs and circumstances thanks in large part to your collaboration and recommendations. Without the innovative involvement and confidence of each informant, this study would not have been possible. The researcher expresses gratitude to you for your dedication, candor, and insightful sharing of experiences, all of which are essential to the advancement of our digital health system.

REFERENCES

[1] Department of Older Persons, Ministry of Social Development and Human Security, “Situation of the Thai Older Persons 2023,” Department of Older Persons, Ministry of Social Development and Human Security, Bangkok, Thailand, Sep. 2024. Accessed: Mar. 16, 2026 [Online]. Available: https://www.dop.go.th/th/statistics_page?cat=12&id=2563

[2] T. Assatarakul, N. Supawantanakul, A. Metasuk, S. Sopitarchasak, L. Yeh, and W. Pothisiri, "A study of sexual health, mental health, and relationships of LGBTIQN+ older adults in Thailand," Faculty of Learning Sciences and Education, Thammasat University, Bangkok, Thailand, 2024. Accessed: May. 30, 2026. [Online]. Available: https://lsed.tu.ac.th/academicservice

[3] R. Sumarlin, I. W. Swandi, I. N. Suardina, and A. R. Adriyanto, "User Interface Design Factors Influencing Perceived Usability in Digital Health Platforms," International Journal of Art and Design, vol. 10, no. 1, pp. 35–47, Apr. 2026. Accessed: May. 15, 2026. [Online]. Available: https://journal.uitm.edu.my/ojs/index.php/IJAD/article/view/10833

[4] W. Chen, K. Zheng, and H. He, "Research on User Centered Elderly Remote Medical Interface Design," Highlights in Art and Design, vol. 11, no. 2, pp. 18–23, Aug. 2025. Accessed: Apr. 10, 2026. [Online]. Available: https://doi.org/10.54097/wc8rsk54

[5] G. Desolda, A. Esposito, R. Lanzilotti and A. Piccinno, "Designing and Evaluating Human-Centred AI Systems: Best-Practices from a Multidisciplinary View, " Joint Proceedings of the ACM IUI, vol. 3957, pp. 170–178, Mar. 2025. Accessed: Jan. 26, 2026. [Online]. Available: https://ceur-ws.org/Vol-3957/BEHAVEAI-paper05.pdf

[6] World Wide Web Consortium, "Web Content Accessibility Guidelines (WCAG) 2.1," W3C Recommendation , May. 6, 2025. Accessed: Mar. 17, 2026. [Online]. https://www.w3.org/TR/WCAG21/

[7] A. Seifi and A. Moshayeri, "The Influence of Color Schemes and Aesthetics on User Satisfaction in Web Design: An Empirical Study," International Journal of Advanced Human Computer Interaction, vol. 2, no. 2, pp. 33–43, Oct. 2024. Accessed: Feb. 10, 2026. [Online]. Available: https://www.ijahci.com/index.php/ijahci/article/view/21

[8] H. He, R. A. R. Ghazilla, and S. H. Abdul-Rashid, "A Systematic Review of the Usability of Telemedicine Interface Design for Older Adults," Applied Sciences, vol. 15, no. 10, Art. no. 5458, May. 2025. Accessed : Mar. 11, 2026. [Online]. Available: https://www.mdpi.com/2076-3417/15/10/5458

[9] E. Amouzadeh, I. Dianat, J. Faradmal, and M. Babamiri, "Optimizing Mobile App Design for Older Adults: Systematic Review of Age-Friendly Design," Aging Clinical and Experimental Research, vol. 37, no. 1, Art. no. 248, Aug. 2025. Accessed: Apr. 12, 2026. [Online]. Available: https://doi.org/10.1007/s40520-025-03157-7

[10] E. A. C. Albers, I. Fraterman, I. Walraven, E. Wilthagen, S. B. Schagen, I. M. van der Ploeg, M. W. J. M. Wouters, L. V. van de Poll-Franse, and K. M. de Ligt, "Visualization Formats of Patient-Reported Outcome Measures in Clinical Practice: A Systematic Review About Preferences and Interpretation Accuracy," Journal of Patient-Reported Outcomes, vol. 6, no. 1, Art. no. 18, Mar. 2022. Accessed: Feb 15, 2026. [Online]. Available: https://doi.org/10.1186/s41687-022-00424-3

[11] M. Sawangbumrung, "Loneliness and Travel Behaviors of Older People," The Journal of Psychiatric Nursing and Mental Health, vol. 33, no. 1, pp. 36–52, Apr. 2019. Accessed: Jan. 10, 2026. [Online]. Available: https://he02.tci-thaijo.org/index.php/JPNMH/article/view/185654

[12] L. C. Hawkley and J. T. Cacioppo, "Loneliness Matters: A Theoretical and Empirical Review of Consequences and Mechanisms," Annals of Behavioral Medicine, vol. 40, no. 2, pp. 218–227, Oct. 2010. Accessed: Mar 21, 2026. [Online]. Available: https://academic.oup.com/abm/article-abstract/40/2/218/4569527

[13] J. Holt-Lunstad, T. B. Smith, and J. B. Layton, "Social Relationships and Mortality Risk: A Meta-Analytic Review," PLoS Medicine, vol. 7, no. 7, Art. no. e1000316, Jul. 2010. Accessed: Jan. 13, 2026. [Online]. Available: https://doi.org/10.1371/journal.pmed.1000316

[14] L. C. Hawkley and J. T. Cacioppo, "Loneliness and Pathways to Disease," Brain, Behavior, and Immunity, vol. 17, no. 1, pp. 98–105, Feb. 2003. Accessed: Jan. 14, 2026. [Online]. Available: https://doi.org/10.1016/S0889-1591(02)00073-9

[15] World Health Organization, "Active Ageing: A Policy Framework," World Health Organization, Geneva, Switzerland, 2002. Accessed: Apr. 30, 2026. [Online]. Available: https://iris.who.int/handle/10665/67215

[16] World Health Organization, "WHO guideline: Recommendations on digital interventions for health system strengthening," World Health Organization, Geneva, Switzerland, 2019. Accessed: Mar. 10, 2026. [Online]. Available: https://iris.who.int/bitstream/handle/10665/324998/WHO-RHR-19.7-eng.pdf

[17] Z. Nimmanterdwong, S. Boonviriya, and P. Tangkijvanich, "Human-Centered Design of Mobile Health Apps for Older Adults: Systematic Review and Narrative Synthesis," JMIR mHealth and uHealth, vol. 10, no. 1, art. no. e29512, Jan. 2022, Accessed: Jul. 14, 2026. [Online]. Available: https://mhealth.jmir.org/2022/1/e29512

[18] E. Balki, N. Hayes, and C. Holland, "Effectiveness of Technology Interventions in Addressing Social Isolation, Connectedness, and Loneliness in Older Adults: Systematic Umbrella Review," JMIR Aging, vol. 5, no. 4, art. no. e40125, Oct. 2022,. Accessed: Jul. 14, 2026. [Online]. Available: https://aging.jmir.org/2022/4/e40125

[19] Y. Zhou, J. Rees, F. Matcham, M. Antonelli, S. Ourselin, and W. Liu, "A User-Centered Interface Design Framework for the DELONELINESS System in Older Adults: Design Indicator Development and Prioritization," JMIR Human Factors, vol. 13, Art. no. e88263, Mar. 2026. Accessed: May. 21, 2026. [Online]. Available: https://humanfactors.jmir.org/2026/1/e88263/

[20] S. Isniah, H. Hardi Purba, and F. Debora, "Plan do check action (PDCA) method: literature review and research issues ", Jurnal Sistem dan Manajemen Industri, vol. 4, no. 1, pp. 72–81, Jul. 2020. Accessed: Jan. 31, 2026. [Online]. Available: https://doi.org/10.30656/jsmi.v4i1.2186

[21] M. Campoverde-Molina, S. Luján-Mora and L. V. García, "Empirical Studies on Web Accessibility of Educational Websites: A Systematic Literature Review," IEEE Access, vol. 8, pp. 91676-91700, Jan. 2020. Accessed: Feb. 25, 2026. [Online]. Available: https://ieeexplore.ieee.org/document/9092982

[22] J. de Jong Gierveld and T. G. van Tilburg, "A 6-item scale for overall, emotional, and social loneliness: Confirmatory tests on survey data," Research on Aging, vol. 28, no. 5, pp. 582–598, 2006. Accessed: Oct. 20, 2025, [Online]. Available: https://doi.org/10.1177/0164027506289723

[23] Thailand Department of Mental Health. (2016). Thai Geriatric Mental Health Assessment Tool–Short Form (T-GMHA-15). Nonthaburi, Thailand: Ministry of Public Health.

[24] R. I. Pfeffer, T. T. Kurosaki, C. H. Harrah, J. M. Chance, and S. Filos, "Measurement of functional activities in older adults in the community," Journal of Gerontology, vol. 37, no. 3, pp. 323–329, May 1982, Accessed: Sep. 24, 2025. [Online]. Available: https://doi.org/10.1093/geronj/37.3.323

[25] C. L. M. Keyes, "Social well-being," Social Psychology Quarterly, vol. 61, no. 2, pp. 121–140, 1998, Accessed: Sep. 22, 2025. [Online]. Available: https://doi.org/10.2307/2787065

[26] J. M. Bredle, J. M. Salsman, S. M. Debb, B. J. Arnold, and D. Cella, "Spiritual well-being as a component of health-related quality of life: The Functional Assessment of Chronic Illness Therapy—Spiritual Well-Being Scale (FACIT-Sp)," Religions, vol. 2, no. 1, pp. 77–94, 2011, Accessed: Sep. 10, 2025. [Online]. Available: https://doi.org/10.3390/rel2010077

[27] V. Braun and V. Clarke, Thematic Analysis: A Practical Guide. London, U.K.: SAGE Publications, 2021.

[28] M. S. Amran and M. Z. Roslan, "Embracing the Digital Side of Mental Health," Discover Psychology, vol. 6, Art. no. 52, Jan. 2026. Accessed : Feb 22, 2026. [Online]. Available: https://doi.org/10.1007/s44202-025-00564-9

[29] R. A. Calvo and D. Peters, Positive Computing: Technology for Wellbeing and Human *Potential*, Cambridge, MA, USA: MIT Press, 2014. Accessed: May. 10, 2026. [Online]. Available: https://doi.org/10.7551/mitpress/9764.001.0001

Image-Based Iron Ore Feed Load Classification Using EfficientNet-B0 with Region-Density Attention and Hierarchical Feature Fusion

Sami F Karali[1], Fawaz Alanazi[2], Eman Ramadan Elsharkawy[3]*, Asma A. Alhashmi[4], Abed Saif Ahmed Alghawli[5], Imen B. Mohamed[6], Chams Sallami[7], Abdulbasit A. Darem[8]

Computing Department-College of Engineering and Computing at Al-Lith 21691, Umm Al-Qura University[1]

Department of Computer Science-College of Science, Northern Border University, Arar 73213, Saudi Arabia[2, 4, 6, 7]

Center for Health Research, Northern Border University, Arar 73213, Saudi Arabia[3]

Computer Science Department-College of Sciences and Humanities, Prince Sattam Bin Abdulaziz University, Aflaj 11912, Kingdom of Saudi Arabia[5]

Center for Scientific Research and Entrepreneurship, Northern Border University, Arar 73213, Saudi Arabia[8]

***Abstract*—Iron ore estimation is a significant part of mineral exploration, as it supports determining both the quantity and quality of available mineral resources. Precise estimation facilitates effective mine planning, production management, and sustainable usage of natural resources. Conventional systems are mostly based on laboratory measurements and geological surveys that are frequently expensive, time-consuming, and restricted when handling larger-scale mineral datasets. In recent years, machine learning (ML) and deep learning (DL) systems have been progressively used to improve estimation by recognizing intricate patterns in mineralogical databases. This research introduces a Multi-modal Deep Learning Framework for Mineral Exploration and Iron Ore Load Prediction (MDL-MEOLP) model to enable precise mining decision-making. The proposed model performs hierarchical region generation utilizing a grid-driven patch extractor to capture multi-scale local regions. Subsequently, an EfficientNet-B0 backbone is employed for local feature extraction, while a region-density attention module emphasizes dense ore regions. Moreover, hierarchical feature fusion is applied to aggregate local-to-global representations and enhance global contextual understanding. Finally, a classification head with global average pooling, a dense layer, a dropout layer, and softmax activation is utilized to predict 17 feed load classes, and the model is trained utilizing Adam optimizer with cross-entropy loss over stratified data splits. To demonstrate the improved performance of the proposed MDL-MEOLP model, a comprehensive experimental analysis is carried out using Real-time Iron Ore Feed Load Estimation. The comparative results reported the effective performance of the MDL-MEOLP model with an accuracy of 97.58% over existing models.**

Keywords*—Mineral exploration; Iron Ore Load Prediction; EfficientNet-B0; Adam optimizer; deep learning; resource utilization

I. Introduction

Over the past decade, the global need for mined resources has increased rapidly due to increasing population, advanced technologies, and economic growth [1]. The demand for metals will grow considerably by the year 2050, with a projected increase of 215% for aluminum, 140% for both copper and nickel, 86% for iron, 81% for zinc, and 46% for lead [2]. Metals like aluminum, copper, gold, and zinc play a vital role in the renewable energy transition and achieving net-zero targets because of their significant role in decarbonization technologies, including solar panels, wind turbines, and electric vehicles [3]. The growing demand for mineral resources has increased the pressure on the mining industry. Similarly, mining companies are facing numerous challenges like declining ore grades, deeper mineral deposits that are difficult to access, and rising energy and operational costs [4]. Meanwhile, society depends heavily on mineral resources for industrial growth and modern technologies, while mining itself is criticized for its environmental impacts [5]. Sustainable and transparent mining practices are vital to meet rising global demand while decreasing damage to the environment. Conventionally, the mining industry has relied on a linear value chain, including sequential stages from exploration to closure of the mine [6]. The procedure starts with exploration, in which all data is gathered to determine spatial extent, geometry, and geotechnical and hydrological settings of potential sites through various methods, including geological mapping, geophysics, remote sensing, and drilling [7].

In the mineral industry, the ball mill grinding circuit (GC) is a critical component for effective production, and low-grade iron ores must undergo a beneficial process for further concentration [8]. The product quality is mainly based on controlling several setpoints in the GC process, including feed load (ore load), water level, mill pressure, and pump pressure. The GC control practices are manually developed and depend on laboratory product analysis used as feedback for modifying those setpoints [9]. Among the various setpoints, the feed load setpoint is to be adjusted as the other adjustments need an extremely skilled and experienced human operator [10]. In simpler terms, the quality of the product is mostly influenced by the ore features and the feed load settings. For instance, when dealing with ore pellets that are difficult to grind, the feed load is kept low to maintain the quality of the product, while that is raised for easy-to-grind ore to improve product throughput [11]. With continuous grade control, precise ore characterization is possible and supports quick responses to variations in ore quality. In recent years, megatrends have emerged that could

*Corresponding author

significantly reshape the mining industry. Digital disruption, characterized by the incorporation of digital technologies into all operational functions, is transforming traditional practices [12]. The development of online sectors, Internet of Things (IoT) systems, and data analytics approaches, including artificial intelligence (AI) and geostatistical methods, enables efficient collection and evaluation of a large amount of real-time data. Such advancements support enhancing predictive maintenance, optimizing processes, improving efficiency, decision-making, and improving safety and sustainability [13]. AI-based technologies also support smart mining systems, which allow accurate ore deposit modeling, live monitoring, geo-metallurgical modeling, and continuous grade control. Together, these advanced technologies enhance resource utilization and minimize environmental impacts and also strengthen transparency in tackling the mineral resource dilemma [14].

A. Motivation and Research Contributions

In spite of developments in process control and mining operations, standard techniques are highly dependent on manual experience and the subsequent decision-making process, which constrains flexibility to change ore features. Current grinding circuit control methods often face difficulties in handling variations in mineral characteristics and feed load during real-time operations, which can reduce overall process efficiency. Although digital technologies and AI-based techniques have been introduced in recent years, many existing methods are still developed separately and are not fully integrated with practical operational control systems. Therefore, there is a strong need for an efficient data-driven framework that can improve ore classification and support accurate feed load prediction for better grinding circuit performance. To address these issues, the proposed MDL-MEOLP framework to design an efficient method for timely phases of mineral exploration employing iron ore images.

The key contributions of this study are outlined as follows:

- Construct a hierarchical region generation process employing grid-driven patch extraction to capture multi-scale local regions from iron images.
- Design an EfficientNet-B0-based feature extraction method for taking the fine-tuned texture patterns and ore density distributions from iron ore images.
- Introduce a region-density attention module to emphasize informative dense ore areas while suppressing irrelevant background data across adaptive feature refinement.
- Employ a hierarchical feature fusion approach to combine local and global representations, improving contextual understanding for feature representation.
- Apply a classification head with global average pooling (GAP), a dense layer, a dropout layer, and softmax activation to precisely predict 17 iron ore feed load categories.
- Integrate an Adam optimizer with cross-entropy loss to ensure stable training and enhance the performance of the method.
- Evaluate the performance of the proposed MDL-MEOLP utilizing the Real-time Iron Ore Feed Load Estimation database and exhibit the solution with several measures.

The proposed model's novelty lies in the domain-specific integration of its elements for the analysis of ore texture, rather than the elements themselves. It combines a grouped weight EfficientNet-B0 backbone, which explicitly divides local texture features from the ore density patterns, a Region Density Attention Module intended for the spatial sparsity of ore-bearing regions, which emphasizes dense regions and suppresses the sparse ones, and a hierarchical fusion pipeline which re-aggregates the multi-level regional characteristics in a closed loop with an attention refinement. This integrated model yields quantifiable gains over the generic backbone-plus-attention baselines utilized in prior ore classification systems.

The structure of this study is organized as follows: Section I explains the Introduction, and Section II reviews the related work on ore feed load estimation. Section III represents the proposed DL-MEOLP framework, Section IV describes the experimental validation and discussion, and Section V discusses the conclusion and future work of the study.

II. Literature Survey

This section discusses a brief overview of existing studies on iron ore feed estimation and identifies the key research gaps in the domain. Diniz et al. [15] developed a novel technique that applied a soft sensor algorithm to gather the energy efficiency of a wet-closed ball mill. Case research was shown through real information from one of the world's highest mining companies, VALE SA, using many models and validation methods. Matos et al. [16] designed a new data-driven soft sensor for evaluating the probabilistic distribution of multi-class lithology in the real world for crushing circuits. The methodology integrates measurements of crusher motor current and rotational speed with signal processing and ML models. Najafabadipour et al. [17] examined an innovative DL-based system to predict critical components in the Gohar Zamin iron ore mining region in southwest Kerman, Iran. One of the methods tested, the CNN, exhibited superior results in predicting the concentrations of the target component, attaining lower error rates and effectively extracting composite spatial patterns in the geochemical data. Firdaus et al. [18] investigated a strong generalized algorithm to automate the task of iron ore characterization of ores belonging to four diverse grades. Thus, a DL approach is designed, which employs a directed acyclic graph network framework through a hybrid inception topology.

Giannetti et al. [19] proposed an enhanced architecture for the prediction of silicon content by applying a new DL technique that depends on Phased LSTM. The system has been trained for three years of information and authorized under a one-year period, utilizing a strong walk-forward validation model, thus delivering confidence in the system results over time. Park et al. [20] discussed a new strategy to enhance the productivity of mining haulage mechanisms by deploying a hybrid methodology that integrates discrete event simulation (DES) and an ML system for predicting ore production. In [21], a hierarchical intelligent control model for mineral particle dimensions on the basis of ML is presented. During the ML layer, AI methodologies, namely CNN and LSTM are employed

to resolve the multi-source ore blending prediction and intelligent categorization of rainy and dry season situations. Luo et al. [22] introduced a new Deep Multi-scale Spatial–Temporal Feature Decoupling Network (DMSTFD-Net), showing a clear attribute decoupling viewpoint to improve method interpretability.

A. Research Gaps

Numerous research gaps can be recognized on the basis of the relevant research. Diniz et al. [15] designed a soft sensor method for energy efficacy prediction in ball mills and emphasized that standard validation algorithms may cause misleading performance, demonstrating the requirement for appropriate time-series (chronological) validation. Matos et al. [16] developed a data-based soft sensor for lithology distribution in crushing circuits utilizing PLC-compatible signals; their strategy is still inadequate in generalization over multiple operational conditions and mining environments. Park et al. [20] examined ML with diverse event simulation for ore production and truck cycle time prediction, although real-world development in compound and extremely dynamic mining models remains difficult. This research shows that prediction accuracy has been enhanced; there is still a strong need for time-aware validation, the best method generalization over mines, and scalable real-world implementation in industrial mining classifications.

III. System Design

The manuscript follows a comprehensive pipeline for early-stage mineral exploration using iron ore images. To accomplish that, the proposed model initially preprocesses the input image through a median filter (MF), CLAHE-based enhancement, pixel normalization using min-max scaling, and resizing. In addition, the hierarchical region generation process is carried out by separating the iron ore image. Besides, the EfficientNet-B0 backbone is utilized to extract discriminative texture and ore density features from iron ore images. Moreover, the region-density attention module is applied to emphasize informative dense ore regions while suppressing less relevant background features. Hierarchical feature fusion is achieved to incorporate local-to-global representations for enhanced contextual understanding. Furthermore, the classification head with GAP, a dense layer, a dropout layer, and softmax activation is used to predict iron ore feed load classes. At last, the Adam optimizer is employed to ensure efficient model training and improved convergence. The complete workflow of the proposed model is illustrated in Fig. 1.

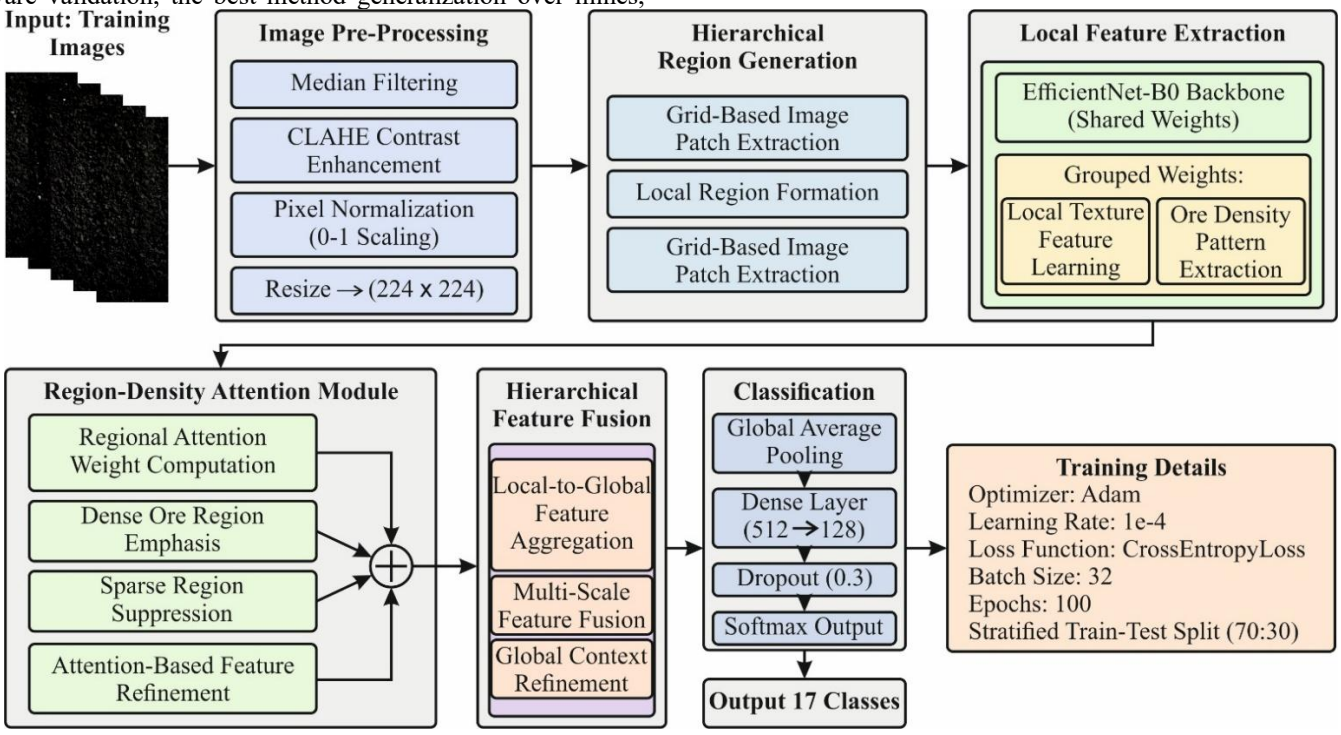


Fig. 1. Overall architecture of the MDL-MEOLP approach.

A. Image Preprocessing

Image preprocessing is an essential step as it improves image quality and ensures consistent input for better feature extraction. It initially preprocesses the input image through an MF, CLAHE-based enhancement, pixel normalization using min-max scaling, and resizing. The preprocessing pipeline consists of the following steps. Initially, the noise removal process is performed by median filtering for suppressing the sensor noise while preserving edge information, which is critical for texture analysis. Secondly, CLAHE is used for enhancing local contrast, to improve the visibility of fine-grained ore texture and variations in density. Third, all pixel intensities are rescaled by pixel normalization to the range of [0, 1] to stabilize and accelerate model training. Fourth, all the images are resized to a fixed resolution of 224 × 224 pixels to match the input needs of the EfficientNet-B0 backbone. Following the pre-processing, a hierarchical region generation step gets image patches by a grid-based scheme, forming local regions which are subsequently utilized for local to global feature learning.

1) Median filtering: To suppress salt-and-pepper noise while maintaining edges, MF is a non-linear image denoising method is utilized. It exchanges each pixel value with the median value of its nearby pixels in a definite window [23] [see Eq. (1)]:

$$I'(x,y) = median\{I(i,j) | (i,j) \in \Omega_{xy}\} \quad (1)$$

Here, Ω_{xy} implies the local neighborhood around pixel (x, y).

2) Contrast enhancement using CLAHE: It is an adaptive contrast enhancement model that applies histogram equalization over small image regions rather than the entire image, which improves local contrast in images. For preventing over-amplification of noise, it constrains the contrast enhancement to clip the histogram at a predetermined threshold and reorganizes the clipped pixels. To enhance fine structural details in low-contrast images while preventing noise amplification, this technique is specifically effective.

3) Min-max based pixel normalization: Min-max scaling transforms data values into a fixed range, typically between 0 and 1, by linearly rescaling the original values based on the minimal and maximal values of the data [see Eq. (2)]:

$$x' = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (2)$$

x_{min} points out the minimal value of the attribute; x_{max} indicates the maximal value of the attribute, and x' represents the normalized attribute.

4) Resizing: To standardize input sizes before feeding data into DL algorithms, image resizing is applied. Each image is uniformly scaled or resampled to a fixed resolution of 224×224pixels during this process.

B. Hierarchical Region Generation

Hierarchical region generation is a structural process utilized for decomposing the mineral imaging into useful spatial units at several levels. The term Multi-modal refers to the various complementary feature modalities excavated from a single image input, instead of various external data sources. Specifically, it decomposes the image-derived representations into two distinct feature modalities, local texture and ore density pattern, through grouped-weight branches of the EfficientNet-B0 backbone. Furthermore, they are processed via the Region-Density Attention Module and integrated via hierarchical feature fusion over local, regional, and global levels. This use of multi-modal describes several learned feature representations from a single modality as multi-modal feature learning.

1) Grid-based image patch extraction: The input mineralogical imaging is separated into an even grid of smaller patches. All patches specify a fixed spatial zone of the imagery, assuring comprehensive and reliable coverage of the complete mineral surface.

2) Local region formation: Afterward, the patch extractor gathers neighborhood patches together to create local areas. These areas are useful geological micro-structures like mineral grains, texture transitions, and boundaries.

3) Multi-scale region representation: For capturing the differences in mineral structure at various levels of feature, local zones are more ordered into multi-scale representations.

C. Local Feature Extraction

Local feature extraction was conducted utilizing an EfficientNet-B0 backbone to train fine-grained texture patterns from the input images. This stage helps in capturing important ore density variations and local structural details for better representation of mineral characteristics.

1) EfficientNet-B0 Backbone: The EfficientNet framework has progressed to classify an efficient scaler approach for the CNN model and enhance prediction outcomes. The compound scaler approach constantly scales network depth, width, and resolution employing a certain set of scaling coefficients [24]. In this method, the EfficientNet-B0 structure is intended as the baseline method. The EfficientNet model comprises numerous versions from B0 to B7, while the superior versions involve improved parameters and enhanced learning ability. Growing network depth allows the method to train higher and more intricate patterns, whereas rising system width supports capturing fine-tuned features from geological and mineral imaging. This technique uses several convolutional layers with mobile inverted bottleneck convolution blocks for extracting effective features and representation learning in mineral deposit and iron ore load estimation tasks. The compound scaler approach is represented as in Eq. (3) to Eq. (7):

$$w = \beta^{\varphi} \quad (3)$$

$$d = \alpha^{\varphi} \quad (4)$$

$$r = \gamma^{\varphi} \quad (5)$$

$$s.t.\alpha.\beta^2.\gamma^2 \approx 2 \quad (6)$$

$$\alpha \geq 1, \beta \geq 1, \gamma \geq 1 \quad (7)$$

Here, w implies network width, d signifies the network depth, and r indicates network resolution, respectively, whereas α, β, and γ represent uniform scalar coefficients decided through the optimizer on the baseline technique. EfficientNet consistently scales each network size utilizing the compound coefficient ϕ, allowing effective learning of mineral distribution patterns and iron ore features from geological imaging, while preserving computational complexity.

2) Local texture feature learning: Local texture feature learning is used for capturing the fine-tuned mineral surface differences, grain limitations, and spatial irregularities from ore instances. The texture description is obtained by analyzing neighborhood pixel intensity interaction employing Local Binary Pattern (LBP) representation:

$$LBP(x_c, y_c) = \sum_{p=0}^{p-1} s(g_p - g_c)2^p \quad (8)$$

In Eq. (8), g_c implies the pixel intensity of the center, g_p denotes the pixel intensity of the neighborhood, and $s(\cdot)$ indicates the thresholding operation.

3) Ore density pattern extraction: The ore density pattern extraction was performed using the mineral concentrate operation and the regional density difference in the ore structure [see Eq. (9)]:

$$\rho(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{9}$$

Let $\rho(x)$ imply the ore density estimation, K denotes the kernel operation, h represents the bandwidth, and x_i with respect to neighboring instance points.

D. Region-Density Attention Module

The region-density attention module is developed to enhance feature representation of minerals by adaptable attention on higher-density ore regions, while it controls the lower-information background areas [25]. This also increases spatial differentiation by assigning attention weights that are based on feature significance and local ore concentration.

1) Regional attention weight computation: Regional attention weights are calculated to measure the significance of all spatial regions, depending on the feature responses that could be extracted utilizing more images [see Eq. (10)]:

$$\alpha_i = \frac{\exp(W_a^T f_i)}{\sum_{j=1}^{N} \exp(W_a^T f_j)} \tag{10}$$

Here, the symbol α_i denotes the standardized attention weight for area i, f_i signifies the regional feature vectors, and W_a implies the trainable attention parameter.

2) Dense ore region emphasis: Dense ore regions have been underscored by improving the feature contributions proportionately to their attention scores, thereby confirming effective representation of mineral-rich regions [see Eq. (11):

$$F_i^{dense} = \alpha_i \cdot f_i \tag{11}$$

Here, the parameter F_i^{dense} signifies the improved feature representation of dense mineral zones.

3) Sparse region suppression: Sparse or non-mineral regions can be controlled by decreasing their feature influence by applying the attention weighting [see Eq. (12)]:

$$F_i^{sparse} = (1 - \alpha_i) \cdot f_i \tag{12}$$

It helps in eliminating the noise from background textures and distinct geological patterns. So, the model has been developed to be more robust to visual clutter and surface artifacts normally existing in real-time ore imaging.

E. Hierarchical Feature Fusion

Hierarchical feature fusion model combines the representations extracted with various abstraction phases to make the integrate feature space that improves mineral pattern and increases expectable robustness.

1) Local-to-global feature aggregation: The local features are captured from fine-grained mineral textures, while the global features are encoded in the total structural contexts.

$$F^{global} = \phi\left(\sum_{i=1}^{N} w_i F_i^{local}\right) \tag{13}$$

In Eq. (13), F_i^{local} represent the local feature vectors, w_i implies trained significance weights, and $\phi(\cdot)$ denotes the non-linear transformation.

2) Multi-scale feature fusion: The multi-scale fusion combines the feature extraction with various resolutions to obtain the macro-level geological patterns and micro-level mineral textures [see Eq. (14)]:

$$F^{MS} = \sum_{s=1}^{S} \alpha_s F_s, \text{where } \alpha_s = \frac{e^{z_s}}{\sum_{k=1}^{S} e^{z_k}} \tag{14}$$

where, the parameter F_s characterizes feature maps corresponding to scale s, and the mathematical notation α_s are adaptable attention weights calculated by employing the softmax function. It permits the model to dynamically highlight the more informative measures based on the intricacy of mineral textures. This fusion increases the strength against differences in imaging resolution, mineral density, and grain size.

3) Global context refinement: The global context refinement improves combined features through the integration of long-term dependencies and decreases the inappropriate background regions [see Eq. (15)]:

$$F^{refined} = F^{MS} \odot \sigma(W_g G) \tag{15}$$

In this mathematical equation, the symbol $\sigma(\cdot)$ points out the sigmoid gating function, the factor G indicates global context embedding, and the term W_g refers to the learnable projection matrix. The element-wise multiplication $\odot$ is performed as an attention gate that reinforces mineral-specific indications, whereas sparse background zone or removing noises. Hence, this enhancement phase majorly increases the feature variability in diverse geological contexts. Fig. 2 denotes the structural design of feature extraction and feature fusion model.

F. Classification Head

The classification head is used to convert extracted features into final predictions. It consists of GAP, a dense layer, a dropout layer, and a softmax activation to accurately predict 17 iron ore feed load classes [26].

1) Global average pooling: This pooling method is used after the convolutional feature extractor blocks to transform spatial feature mappings into a compressed feature representation. It minimizes every feature mapping into a distinct scalar performance by calculating the average of each spatial position, thus removing the essential for a flattened layer and considerably minimizing the count of trainable parameters.

$$GAP(x) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_{i,j} \quad (16)$$

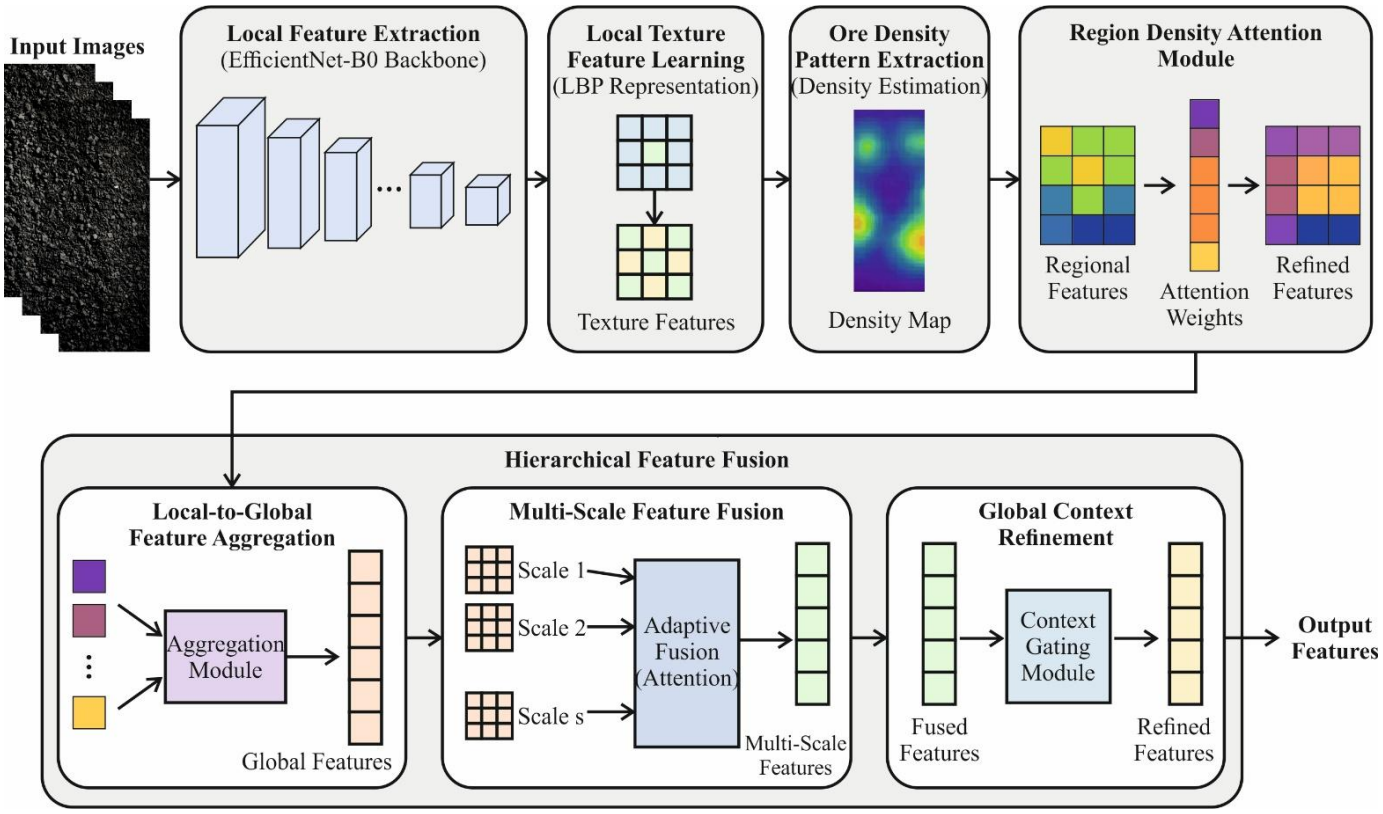


Fig. 2. Structural design of feature extraction and feature fusion model.

In Eq. (16), W and H represent the width and height of the feature mapping. This function considerably lessens the count of parameters compared with a flattened layer, leading to a more lightweight and computationally effective method while maintaining global spatial data.

2) Dense layer: This layer, also known as a fully connected layer, accomplishes non-linear feature alteration by mapping the dense feature vector from 512 to 128 neurons. It learns higher-level discriminative representations by using a weighted linear conversion followed by a non-linear activation function [see Eq. (17)]:

$$z = Wx + b \quad (17)$$

Here, x implies the feature vector of the input value, W implies the weighted matrix, and b denotes the biased vector.

3) Dropout: Dropout is a regularized method utilized to minimize overfitting for random deactivation neurons in training with probability p, forcing the system to learn stronger feature generalization [see Eq. (18)]:

$$y = Dropout(x, p) \quad (18)$$

4) Softmax output: The softmax function is used in the final layer to transform raw logits into a standardized probability function throughout numerous categories. It allocates high probabilities to further likely classes while assuring that the amount of each predicted probability equals to 1 [see Eq. (19)]:

$$softmax(x) = \frac{e^{z_i}}{\sum_j^4 e^{z_j}} \quad (19)$$

Now, z_i implies the logit of the i^{th} class, and K signifies the overall count of classes. In this study, $K = 17$ with respect to 17 feed load categorizations allows probabilistic multi-class identification for feed load estimation.

G. Network Configuration Details

Table I shows the parameter configuration of the MDL-MEOLP structure for feed load classification. Primarily, the input images are pre-processed employing Median Filtering, CLAHE-driven contrast enhancement, pixel standardization, and resizing images to 224×224 resolution to enhance the quality of imaging and consistency. Hierarchical area creation is performed using a grid-driven patch extractor with multi-scale local region representation, while EfficientNet-B0 is used as the backbone system to obtain density features and local texture. The structure integrates a Regional Density Attention module and hierarchical multi-scale feature fusion to increase significant global and regional contextual data. At last, the classifier procedure uses global average pooling and a Softmax layer for

17 feed load labels, and the method has been trained utilizing the Adam optimizer with a batch size of 32 over 100 epochs and a learning rate of 1e−4.

TABLE I. PARAMETER SETTING

Model	Parameter	Configured Value
Image Preprocessing	Filtering Method	Median Filtering
	Contrast Enhancement	CLAHE
	Normalization	Pixel Normalization (0–1 Scaling)
	Image Resize	224 × 224
Hierarchical Region Generation	Patch Extraction Method	Grid-Based Patch Extraction
	Patch Size	32 × 32
	Region Representation	Multi-Scale
	Number of Local Regions	16
Local Feature Extraction	Backbone Network	EfficientNet-B0
	Feature Extraction Type	Local Texture and Density Features
	Feature Embedding Dimension	1280
	Activation Function	Swish
Region-Density Attention Module	Attention Mechanism	Regional Density Attention
	Attention Hidden Units	256
	Sparse Region Threshold	0.2
	Attention Dropout	0.2
Hierarchical Feature Fusion	Fusion Strategy	Local-to-Global Aggregation
	Fusion Type	Multi-Scale Feature Fusion
	Global Context Refinement	Enabled
	Fusion Output Dimension	512
Classification Head	Pooling Layer	Global Average Pooling
	Dropout Rate	0.3
	Output Layer	Softmax
	Number of Classes	17 Feed Load Classes
Training Configuration	Optimizer	Adam
	Learning Rate	1e-4
	Loss Function	CrossEntropyLoss
	Batch Size	32
	Epochs	100
	Train-Test Split	80;20 and 70:30

IV. EXPERIMENTAL VALIDATION AND DISCUSSION

The experimental evaluation of the MDL-MEOLP method was performed using Real-time Iron Ore Feed Load Estimation [27]. This database (2.05GB/2.7GB uncompressed) consists of iron ore feed images with load classes (17 loads, tons). All images have dissimilar unknown ore categories. The only identified element is the ore load with respect to the image. Table II shows that the dataset consists of 17 classes with a total of 2,985 samples. Fig. 3 signifies the representative feed load images obtained from real-time mining operation dataset. Fig. 4 denotes the samples of preprocessed image. Fig. 5 represents the visual representation of learned feature maps.

The Real-time Iron Ore Feed Load Estimation dataset has a total of 2,985 images allocated across 17 feed load classes, with sample counts starting from 127 to 228. All images given were resized to 224 × 224 pixels for the model input. The dataset was collected from "Validation dataset for a new weakly supervised learning approach for Real-time Iron Ore Feed Load Estimation", to represent varying feed load levels. The dataset was partitioned into training and test sets using a stratified 70:30 split, ensuring proportional class representation in both subsets.

TABLE II. DETAILS OF DATASET

Classes	Number of Samples
Load 1	175
Load 2	207
Load 3	181
Load 4	166
Load 5	228
Load 6	169
Load 7	197
Load 8	170
Load 9	127
Load 10	211
Load 11	147
Load 12	184
Load 13	176
Load 14	164
Load 15	189
Load 16	152
Load 17	142
Total Samples	**2985**

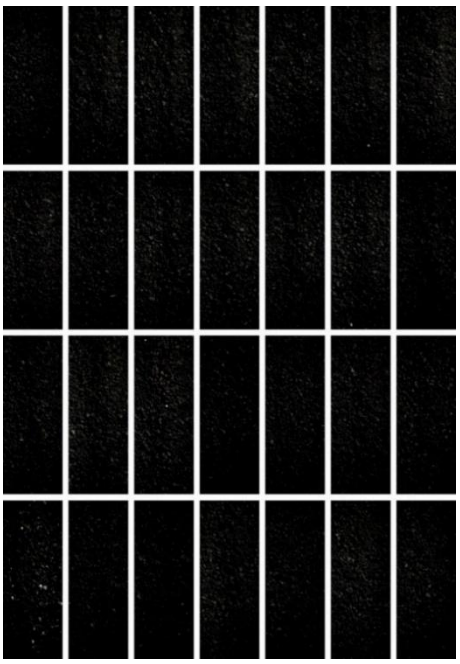


Fig. 3. Representative feed load images obtained from real-time mining operation dataset.

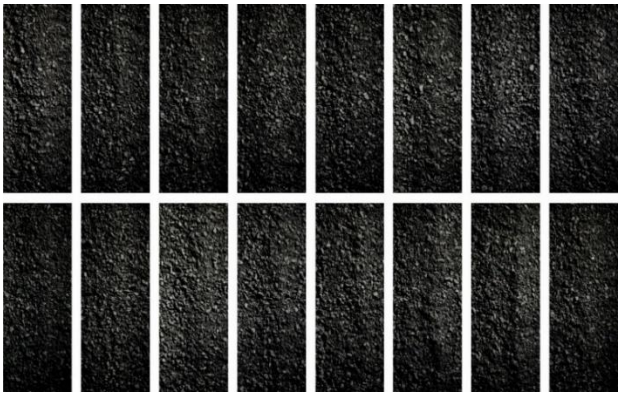


Fig. 4. Sample preprocessed images.

Fig. 6 displays confusion matrices for the diverse class feed load classifier method under 80:20 and 70:30 data splits. The diagonal dominance in each matrix suggests that most instances are appropriately categorized into each class. Some of the diagonal performance appear, presenting slight misclassifications among the similar categories. The outcomes confirm that the method preserves stronger and more reliable classifier values through data split.

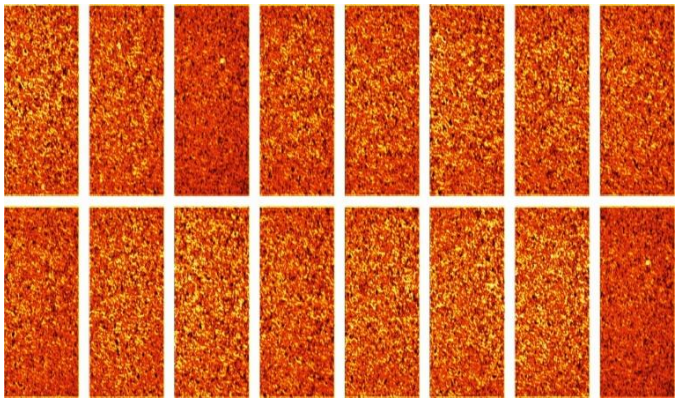


Fig. 5. Visual representation of learned feature maps.

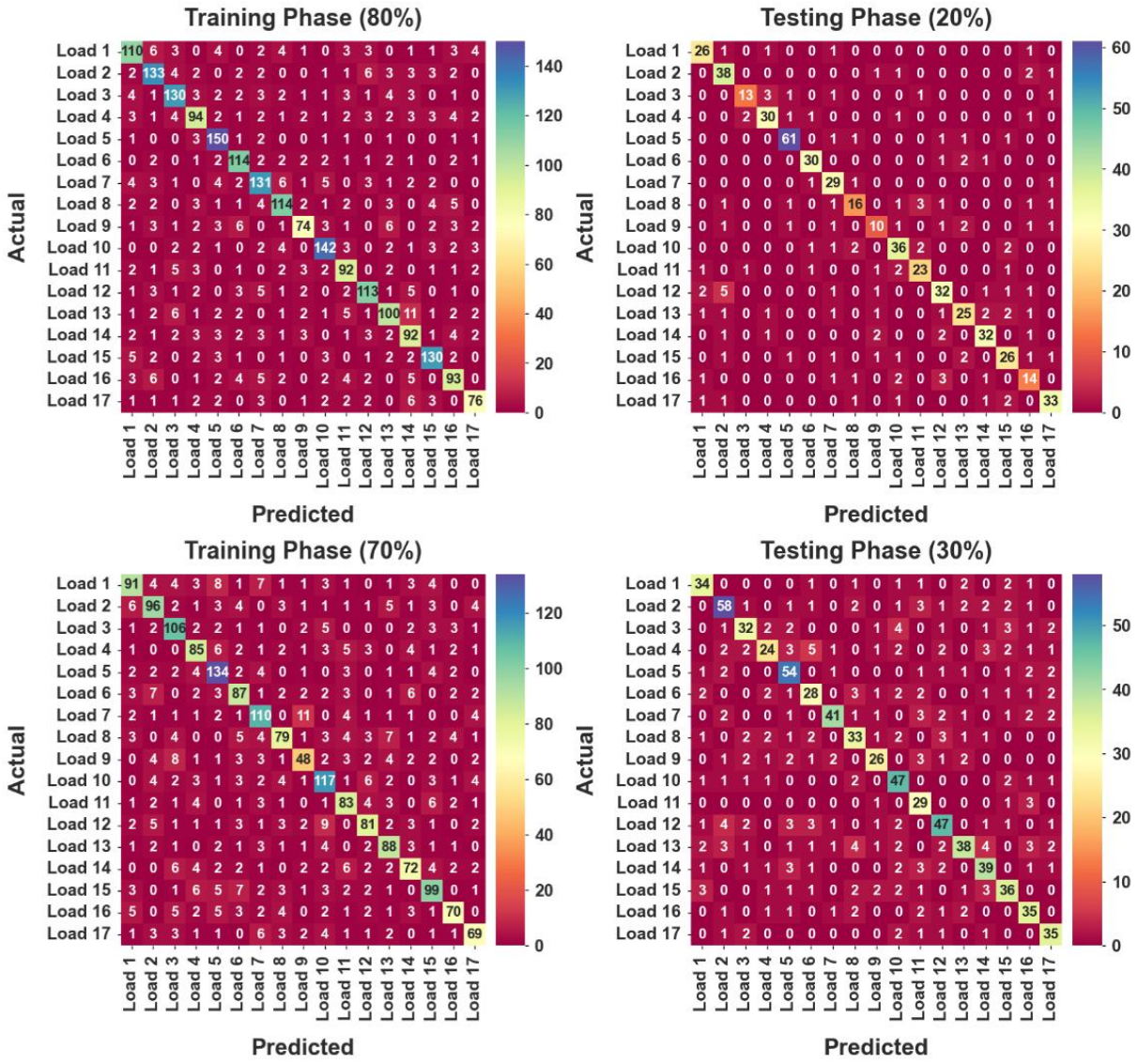


Fig. 6. Confusion matrix of the MDL-MEOLP method with 80:20 and 70:30 split.

Table III and Fig. 7 show the feed load classifier outcome of the MDL-MEOLP approach with an 80:20 train-test split. The result implied that the MDL-MEOLP method accurately detected the samples. The proposed MDL-MEOLP method attained an average value with *an accuracy* of 97.54% in TRAINING SET and *an accuracy* of 97.58% in TESTING SET, whereas the MCC of 77.30% for the TRAINING SET and MCC of 76.01% for the TESTING SET further confirm the model's reliable and balanced classification performance.

TABLE III. FEED LOAD CLASSIFIER RESULT OF THE MDL-MEOLP METHOD WITH 80:20.

Class	TRAINING SET (80%)	TESTING SET (20%)
Accuracy	97.54	97.58
Precision	78.77	77.75
Recall	78.50	77.01
F1-Score	78.57	77.19
MCC	77.30	76.01

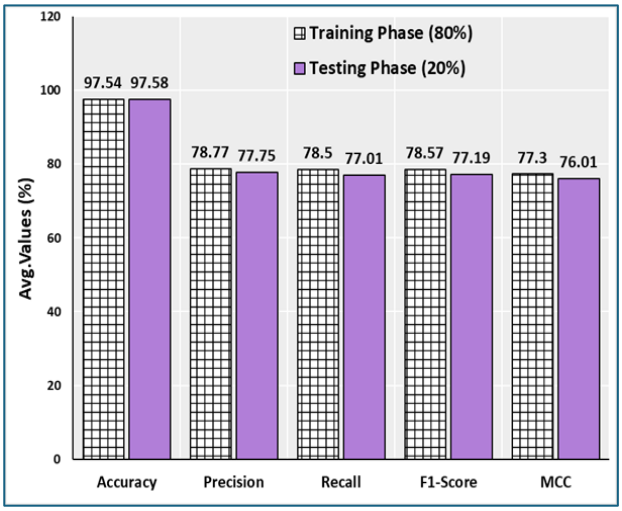


Fig. 7. Average value of the MDL-MEOLP method with 80:20.

Table IV and Fig. 8 exhibit the feed load classifier result of the MDL-MEOLP system with a 70:30 train-test split. The findings shown that the MDL-MEOLP approach precisely classified the instances. The proposed MDL-MEOLP technique reached average result with *an accuracy* of 96.77% in TRAINING SET and *an accuracy* of 96.59% in TESTING SET, while the MCC of 70.29% for the TRAINING SET and MCC of 68.88% for the TESTING SET more strengthen the model's consistent and stable classifier solutions.

Fig. 9 demonstrates the training and validation accuracy and loss graphs on 80:20 and 70:30. It shows the proposed model reaches increasing values over increasing epochs. In addition, the increasing validation over training exhibits that it learns efficiently on the test dataset. It also illustrates the loss investigation of the model. The results indicate that it reaches closer values of training and validation loss. It is observed that the MDL-MEOLP technique learns efficiently on the test dataset.

TABLE IV. FEED LOAD CLASSIFIER RESULT OF THE MDL-MEOLP METHOD WITH 70:30.

Class	TRAINING SET (70%)	TESTING SET (30%)
Accuracy	96.77	96.59
Precision	72.29	70.76
Recall	71.82	71.03
F1-Score	71.94	70.43
MCC	70.29	68.88

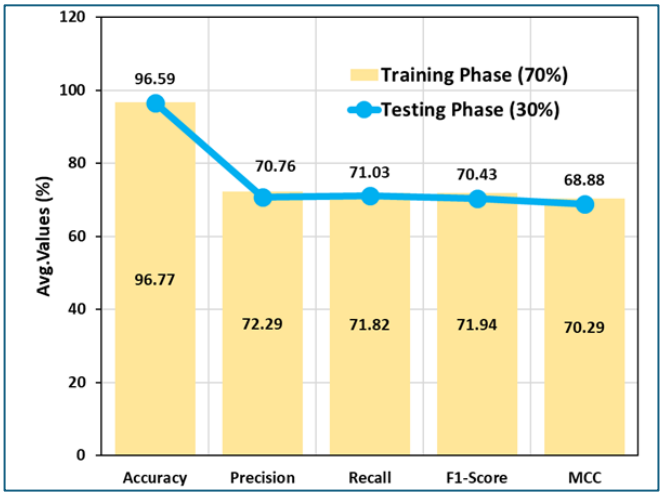


Fig. 8. Average value of the MDL-MEOLP approaches 70:30.

Table V and Fig. 10 portray the comparison study of the MDL-MEOLP approach with other methodologies using evaluation metrics [28, 29]. The stimulation outcome indicated that the MDL-MEOLP method achieves superior performance. The Xception, InceptionV3, ResNet50, DenseNet121 approaches have attained lower performance, whereas the EfficientNet-B0, MobileNetV3, ShuffleNetV2, and AlexNet methods obtain moderate performance. In addition, the EDF-NSDE reaches better prediction ability with more balanced outcomes. However, the proposed MDL-MEOLP method has achieved better overall solutions with *an accuracy* of 97.58%, *precision* of 77.75%, *recall* of 77.01%, and an F1-score of 77.19%, representing its superior effectiveness and balanced classification capability compared to the other models. Furthermore, the reported values are directly referred to in the previously published study for a better comparison.

The gap that exists between accuracy (97.58%) and macro F1-score (77.19%) reflects the class imbalanced nature of the dataset, as macro averaging weights all classes equally and it is thus highly sensitive to the minority-class errors. It reports the confusion matrix and per-class metrics, showing that the misclassifications concentrate among visually similar and underrepresented adjacent load classes, which provides a more reliability beyond the accuracy alone.

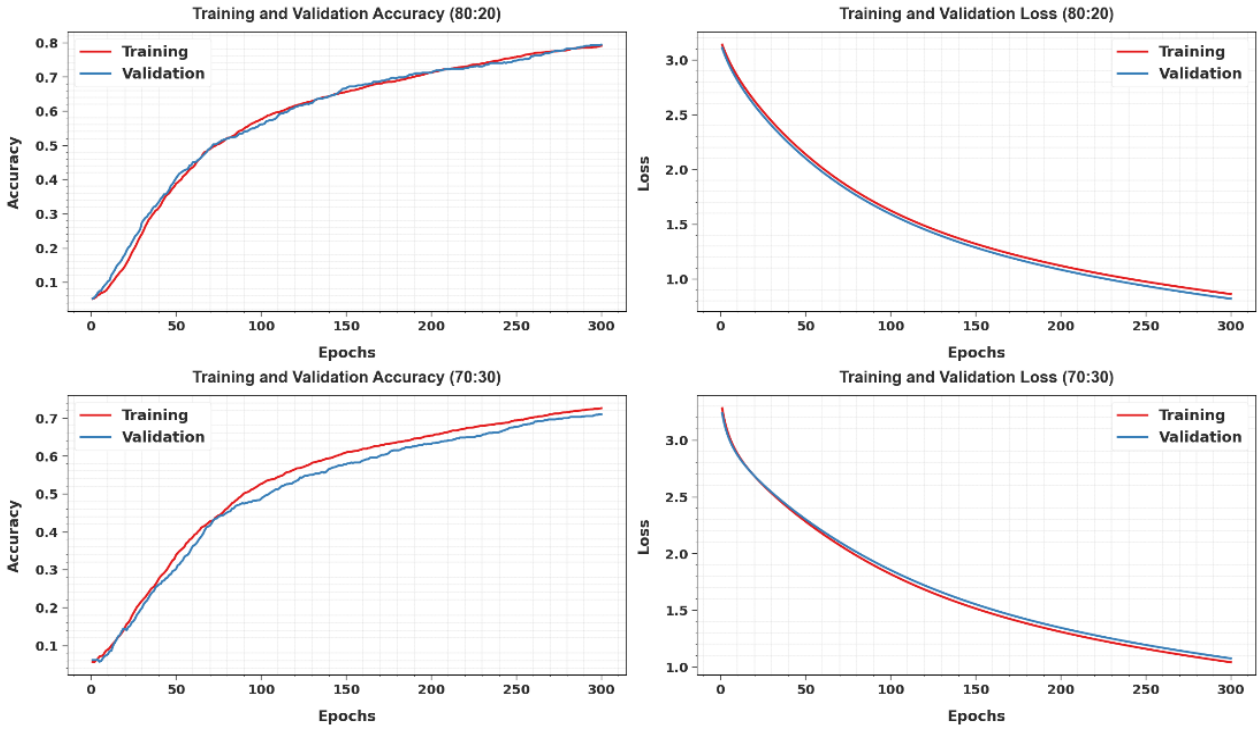


Fig. 9. Training and Validation Accuracy and Loss curve on 80:20 and 70:30.

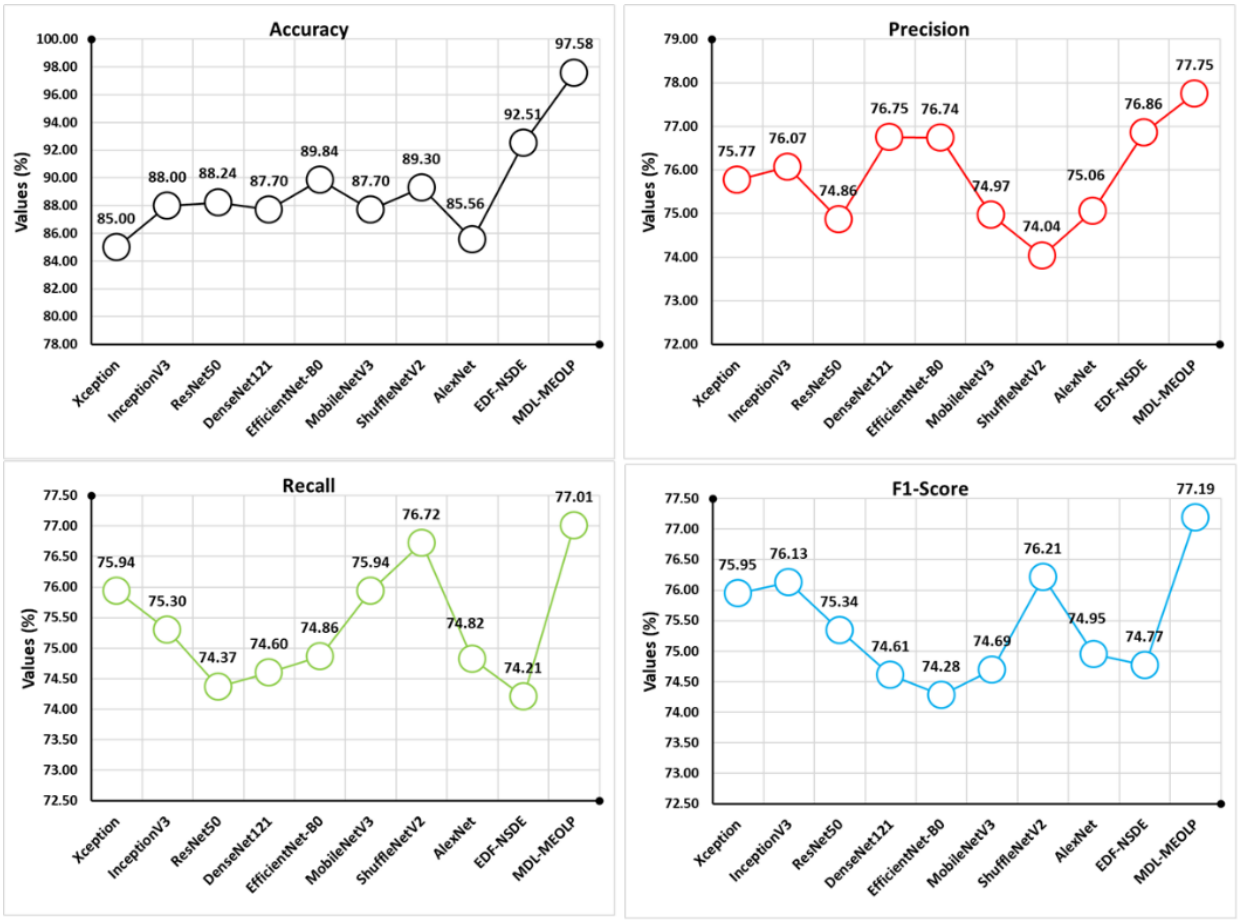


Fig. 10. Comparison outcome of the MDL-MEOLP approach with various metrics.

TABLE V. COMPARATIVE RESULTS OF THE MDL-MEOLP METHOD WITH EXISTING APPROACHES

Network	$Accur_y$	$Preci_n$	$Recal_l$	$F1_{score}$	Computational Time (sec)
Xception	85.00	75.77	75.94	75.95	13.85
InceptionV3	88.00	76.07	75.30	76.13	7.06
ResNet50	88.24	74.86	74.37	75.34	8.33
DenseNet121	87.70	76.75	74.60	74.61	9.12
EfficientNet-B0	89.84	76.74	74.86	74.28	10.13
MobileNetV3	87.70	74.97	75.94	74.69	10.94
ShuffleNetV2	89.30	74.04	76.72	76.21	8.07
AlexNet	85.56	75.06	74.82	74.95	8.12
EDF-NSDE	92.51	76.86	74.21	74.77	8.89
MDL-MEOLP	97.58	77.75	77.01	77.19	3.31

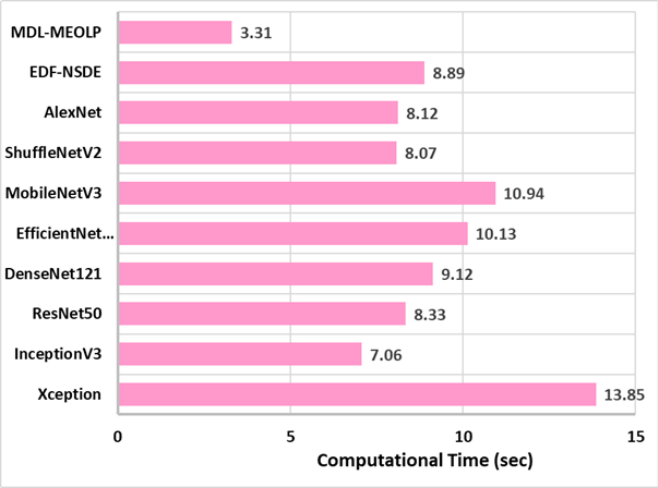


Fig. 11. Computational time of the MDL-MEOLP approach with other methodologies.

Fig. 11 reveal the computational time (CT) of the MDL-MEOLP approach with existing techniques. The result implied that the MDL-MEOLP method has gained minimal value with CT of 3.31sec, whereas the Xception, InceptionV3, ResNet50, DenseNet121, EfficientNet-B0, MobileNetV3, ShuffleNetV2, AlexNet, and EDF-NSDE require significantly higher processing time. This demonstrates that the MDL-MEOLP approach is faster and more computationally efficient than the competing methods.

An ablation study is used to analyze a model by systematically removing or altering its components to evaluate their effect on performance. It helps determine the importance of each module in the overall architecture. By comparing results, it reveals which parts contribute most to accuracy and efficiency. Table VI present the ablation study of the MDL-MEOLP method with metrics. The outcome implied that without Hierarchical Region Generation, EfficientNet-B0 only, EfficientNet-B0 + RDAM, Without Region-Density Attention Module, and Without Hierarchical Feature Fusion approaches achieve inferior value with various metrics, while the MDL-MEOLP (Region Generation + EfficientNet-B0 + RDAM + Hierarchical Fusion + Softmax) method obtain superior value with *an accuracy* of 97.58%, *precision* of 77.75%, *recall* of 77.01%, and F1-score of 77.19%, demonstrating the effectiveness of integrating all components.

The proposed MDL-MEOLP technique functions well for iron ore feed load forecast and mineral exploration applications. Enhancing image quality and extracting significant local details is supported by image preprocessing and grid-based patch extraction. The EfficientNet-B0 backbone efficiently learns density features and ore texture. The attention component assists the method, concentrated on critical ore areas, while minimizing background noise. For a better understanding of global and local data by combining hierarchical feature fusion. Finally, the approach attains higher precision in categorizing 17 feed load classes in comparison with other methodologies.

Table VII presents the k-fold cross-validation outcomes of the proposed MDL-MEOLP method. The method attained accuracy values ranged from 89.92% to 98.83%, with the maximum performance accuracy of 98.83% in K-6. The better precision of 98.43%, recall of 98.12%, and F1-score of 98.47% were achieved in K-6. Although slight difference over the folds, the constantly maximum evaluation measures illustrate the robustness, stability, and strong generalization ability of the proposed method.

TABLE VI. ABLATION STUDY OF THE MDL-MEOLP SYSTEM WITH METRICS

Network	Accuracy	precision	Recall	F1-Score
Without Hierarchical Region Generation	94.71	74.69	73.80	73.87
EfficientNet-B0 only	95.30	75.40	74.36	74.37
EfficientNet-B0 + RDAM	95.90	76.00	75.10	75.13
Without Region-Density Attention Module	96.41	76.57	75.78	75.70
Without Hierarchical Feature Fusion	97.04	77.17	76.43	76.44
MDL-MEOLP (Region Generation + EfficientNet-B0 + RDAM + Hierarchical Fusion + Softmax)	97.58	77.75	77.01	77.19

TABLE VII. K-FOLD CROSS VALIDATION RESULTS OF THE MDL-MEOLP SYSTEM.

K-Fold	Accuracy	Precision	Recall	F1-score
k-1	94.26	93.71	94.08	93.89
k-2	96.84	96.29	95.94	96.11
k-3	91.73	91.08	90.82	90.95
k-4	97.35	96.81	97.12	96.96
k-5	93.14	92.58	92.91	92.74
k-6	98.83	98.43	98.12	98.47
k-7	95.68	95.11	94.86	94.98
k-8	89.92	89.37	89.08	89.22
k-9	97.81	97.26	97.04	97.15
k-10	95.17	94.66	94.39	94.52

V. CONCLUSION AND FUTURE WORK

This study has presented an MDL-MEOLP approach for early-stage mineral exploration using iron ore images, which facilitates precise mining decision-making. Initially, the proposed model preprocesses the images to improve feature consistency and image quality. Besides, the hierarchical region generation is carried out using grid-based patch extraction to acquire multi-scale local regions. Moreover, an EfficientNet-B0 backbone is utilized for local feature extraction, enabling effective learning of texture patterns and ore density distributions. Furthermore, a region-density attention module is designed to emphasize dense ore regions while suppressing less informative areas through attention-based feature refinement. The hierarchical feature fusion is further applied to aggregate local-to-global representations and enhance global contextual understanding. Eventually, a classification head with GAP, dense layer, a dropout layer, and softmax activation is utilized to predict 17 feed load classes, and the model is trained using the Adam optimizer with cross-entropy loss over stratified data splits. To exhibit the improved performance of the proposed MDL-MEOLP model, a comprehensive experimental analysis takes place using Real-time Iron Ore Feed Load Estimation. The comparative results reported the effective performance of the MDL-MEOLP model with an accuracy of 97.58% over existing methodologies. Although the proposed algorithm functions well, it is primarily limited by its reliance on image-based information, which may fail to capture all geological and operational variations in mining settings. Furthermore, its high computational complexity may limit its viability in real-world applications.

Regardless the promising performance of MDL-MEOLP method, certain limitations remains. First, the dataset used is of limited in size and extracted from a single source, which may affects the generalization of ore feeds with various imaging conditions. It has not been verified on other ore types with substantially different density and texture distributions. The future work will be focussed on validating MDL-MEOLP model over larger and diverse multi-site datasets, which explores lightweight frameworks for real time edge deployments, and extending the architecture to related mineral exploration works such as ore-grade estimation and multi mineral classifications. Lastly, explainable AI methodology can be added to make the performance simpler to interpret for mining engineers.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available at https://data.mendeley.com/datasets/b9ng2y2j5p/1 [27].

FUNDING STATEMENT

This research is funded by the Northern Border University through grant number (NBU-CRP-2026-249).

ACKNOWLEDGMENT

The authors extend their appreciation to the Northern Border University, Saudi Arabia, for supporting this work through project number (NBU-CRP-2026-249).

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

REFERENCES

[1] Saavedra, M., Risso, N., Momayez, M., Nunes, R. and Tenorio, V., 2026. Cluster-Based Machine Learning Modeling for Particle Size Variability in SAG Mill Feed. Mining, Metallurgy & Exploration, pp.1-14.

[2] Najafabadipour, A., 2026. Supervised Machine Learning-Based Determination of Groundwater Flow Direction in Complex Mining Areas. Mine Water and the Environment, pp.1-18.

[3] Yi, L., Yu, D., Wang, Y., Luo, B. and Peng, X., 2026. Ultra-short-term Forecasting Study of Power Load in Mega Steel Industry Based on Multi-stage Modeling. Recent Patents on Mechanical Engineering, 19(1), pp.131-149.

[4] Moyo, N., Mamvura, T., Danha, G. and Raghupatruni, P., 2025. Modelling and optimization of ultra-fine copper sulphide grinding: A hybrid statistical and machine learning approach. Particuology.

[5] Sil, J., Bhaumik, A.K. and Sen, S., 2025. Intelligent Real-Time Control Systems to Predict the Particle Size of Minerals Using Acoustics of Ball Mill (Vol. 600). Springer Nature.

[6] Liu, Y., Yan, G. and Pang, Y., 2025, May. Predictive Modeling of Ball Mill Load Parameters Using Hybrid Physics-Informed Neural Networks. In 2025 37th Chinese Control and Decision Conference (CCDC) (pp. 1604-1609). IEEE.

[7] Lucay, F.A., 2025. Balancing the Generation of Ultrafine, Fine and Coarse Particles in Comminution. Mineral Processing and Extractive Metallurgy Review, pp.1-23.

[8] Fang, Y., Ni, X., Xiao, Q., Huang, S. and López-Valdivieso, A., 2025. Iron-based materials synthesized by mechanical ball milling for environmental contaminants removal: progress and prospects. International Journal of Environmental Research, 19(1), p.12.

[9] Tripathi, P., Sunal, S., Dadwal, M., Kumar, D.S. and Ghorui, P.K., 2025. Development of an Automated Property Prediction System for the Pelletizing Process. Engineering Science & Technology, pp.247-260.

[10] Taghavi, J. and Gharabaghi, M., 2025. Prediction of ball mill power in iron ore concentration plants: A comparison between radial basis functions and linear regression. Results in Engineering, 26, p.105114.

[11] Parvathareddy, S., Yahya, A., Amuhaya, L., Samikannu, R. and Suglo, R.S., 2025. Transforming mining energy optimization: a review of machine learning techniques and challenges. Frontiers in Energy Research, 13, p.1569716.

[12] Behera, N., Laha, D. and Lahiri, S.K., 2025. Hybrid machine learning and genetic algorithm framework for optimising slag chemistry in electric arc furnaces. Canadian Metallurgical Quarterly, pp.1-17.

[13] Guo, H., Yang, Q., Yu, Y., Wang, L., Jia, P. and Pedrycz, W., 2026. Prediction of iron ore inventory at ports: A decomposition-integration hybrid approach incorporating key influencing factors. Expert Systems with Applications, 299, p.130041.

[14] Behera, N., Ghosh, A., Pal, P., Das, B. and Lahiri, S.K., 2025. Enhancing steelmaking productivity: a hybrid approach of machine learning and genetic algorithm optimisation for electric arc furnace. Ironmaking & Steelmaking, p.03019233251318881.

[15] Diniz, D.R.M., de Oliveira Morais, G., Reis, A.J.D.R. and Rêgo Segundo, A.K., 2026. A Soft Sensor for Inferring Energy Efficiency of Wet-Closed Ball Mills. Journal of Control, Automation and Electrical Systems, pp.1-18.

[16] Matos, S.N., Pinto, T.V., Duarte, R., Albuquerque, K.S., Fonseca, A.G., Ranieri, C.M., Marcolino, L.S., Pessin, G. and Ueyama, J., 2026. Data-driven soft sensor development for ore type estimation in mineral crushing processes. Engineering Applications of Artificial Intelligence, 167, p.113755.

[17] Najafabadipour, A., Hassanzadeh, F. and Kordestani, M., 2025. Advanced deep learning models for predicting elemental concentrations in iron ore mine using XRF data: a cost-effective alternative to ICP-MS methods. Environmental Geochemistry and Health, 47(4), p.104.

[18] Firdaus, S., Anwar, S. and Mohapatra, S., 2025. A deep hybrid inception network model with entropy-based attention for automated iron ore image characterization. Scientific Reports, 15(1), p.38567.

[19] Giannetti, C., Borghini, E., Carr, A., Raleigh, J. and Rackham, B., 2025. Deep learning for robust forecasting of hot metal silicon content in a blast furnace. The International Journal of Advanced Manufacturing Technology, 136(9), pp.4045-4054.

[20] Park, S., Jung, D. and Choi, Y., 2023. Prediction of ore production in a limestone underground mine by combining machine learning and discrete event simulation techniques. Minerals, 13(6), p.830.

[21] Zou, G., Zhou, J., Song, T., Yang, J. and Li, K., 2023. Hierarchical intelligent control method for mineral particle size based on machine learning. Minerals, 13(9), p.1143.

[22] Luo, X., Huang, W., He, S., Xiao, W. and Jiang, Z., 2025. Deep Feature Decoupling Network for Ball Mill Load Signals. Machines, 13(10), p.881.

[23] Polatgil, M., 2022. Investigation of the effect of normalization methods on ANFIS success: forestfire and diabets datasets. International Journal of Information Technology and Computer Science, 14(1), pp.1-8.

[24] Hussain, A., Ul Amin, S. and Fayaz, M., 2023. An Efficient and Robust Hand Gesture Recognition System of Sign Language Employing Finetuned Inception-V3 and Efficientnet-B0 Network. Computer Systems Science & Engineering, 46(3).

[25] Zhou, C., He, Z., Dong, J., Li, Y., Ren, J. and Plaza, A., 2025. Low-rank and sparse representation meet deep unfolding: A new interpretable network for hyperspectral change detection. IEEE Transactions on Geoscience and Remote Sensing.

[26] Ganguly, P. and Ghosh, A., 2024, November. Efficient brain tumor classification with lightweight CNN architecture: a novel approach. In 2024 IEEE International Conference on Future Machine Learning and Data Science (FMLDS) (pp. 325-330). IEEE.

[27] L. Guo, "Validation dataset for a new weakly supervised learning approach for real-time iron ore feed load estimation," Data Archiving and Networked Services (DANS). Apr. 29, 2022. doi: 10.17632/b9ng2y2j5p.1. Available at: https://data.mendeley.com/datasets/b9ng2y2j5p/1

[28] Firdaus, S., Anwar, S. and Mohapatra, S., 2025. A deep hybrid inception network model with entropy-based attention for automated iron ore image characterization. Scientific Reports, 15(1), p.38567.

[29] Zhang, D., Qian, X., Shi, C., Zhang, Y., Qian, Y. and Zhou, S., 2025. Iron Ore Image Recognition Through Multi-View Evolutionary Deep Fusion Method. Future Internet, 17(12), p.553.

An Intelligent Writing Tutor for ESL Learners: A Rule-Based and NLP-Assisted Feedback Framework

Ruth L. Ladwingon
Department of Languages, Kalinga State University, Bulanao, Tabuk City, Kalinga, Philippines

***Abstract*—Writing remains one of the most challenging skills for English as a Second Language (ESL) learners because it requires the coordinated application of grammar, vocabulary, and written discourse conventions. This study proposes an Intelligent Writing Tutor that integrates corpus-informed error analysis, natural language processing (NLP), and rule-based reasoning to generate individualized and explainable writing feedback for ESL learners. Guided by a mixed-methods Design Science Research approach, weekly journal entries produced by Ilocano-speaking English language majors at the Kalinga State University served as the learner corpus for analysis. Manual expert annotation based on Corder's Error Analysis framework identified lexical, morphological, syntactic, and mechanical errors that informed the development of an interpretable rule-based feedback engine. The analysis revealed that morphological errors were the most frequent, followed by mechanical, lexical, and syntactic errors, with verb tense misuse emerging as the dominant writing difficulty. Qualitative findings further indicated that many of the observed errors reflected first-language interference, particularly in tense marking, subject–verb agreement, preposition usage, and lexical choice. The proposed framework operationalizes learner texts through preprocessing, NLP-assisted error detection, rule-based error classification, feedback generation, and recommendation modules to produce individualized feedback reports. The study demonstrates the feasibility of integrating explainable NLP techniques and second language acquisition principles into an educational writing support system. Because the framework was developed using journal writings from Ilocano-speaking learners in a single institution, its applicability to other learner populations requires further investigation.**

Keywords*—Corpus-informed instruction; educational technology; English as a Second Language; Intelligent Writing Tutor; natural language processing; rule-based feedback

I. Introduction

Writing proficiency is a fundamental component of second language acquisition (SLA) and an essential competency for academic success, professional communication, and lifelong learning. Among learners of English as a Second Language (ESL), writing is often regarded as one of the most challenging language skills because it requires the simultaneous coordination of linguistic knowledge, discourse organization, vocabulary selection, and grammatical accuracy. Difficulties in writing frequently emerge from learners' limited mastery of the target language and the influence of their first language (L1), resulting in persistent errors that impede effective communication [1], [2].

Early studies on error analysis established that learner errors are not random occurrences but valuable indicators of the language acquisition process. Corder argued that errors reveal learners' underlying hypotheses about the target language and provide insights into instructional needs [3], [4]. Similarly, Dulay and Burt emphasized that second language errors reflect developmental strategies employed by learners as they construct interlanguage systems [5]. These perspectives shifted the focus of language instruction from merely correcting mistakes to understanding the cognitive and developmental mechanisms underlying learner performance.

Interlanguage theory further explains that learners gradually develop an independent linguistic system that incorporates characteristics of both the native and target languages. Such systems evolve dynamically as learners receive additional exposure to the target language [6], [7]. Krashen likewise proposed that language acquisition occurs through meaningful input and progressive internalization of linguistic structures, although learners may continue to exhibit transitional forms during the acquisition process [8]. Consequently, ESL writing errors should be interpreted not solely as deficiencies but as developmental manifestations that require targeted pedagogical responses.

Among the factors contributing to writing difficulties, L1 interference remains one of the most extensively investigated phenomena in SLA research. The transfer of native language structures often affects syntax, morphology, vocabulary use, and discourse organization in learners' written production [9], [10]. Studies conducted among Chinese, Indonesian, Malaysian, and other multilingual ESL populations consistently demonstrate that negative transfer contributes significantly to recurring grammatical errors and non-native patterns in writing [11], [12].

Research has documented various manifestations of native language interference, including tense inconsistency, omission of articles, subject–verb disagreement, inappropriate lexical selection, and direct translation from the mother tongue [13], [14]. Such transfer effects frequently interact with developmental errors arising from incomplete acquisition of English grammatical rules [15], [16]. These findings highlight the need for instructional approaches capable of distinguishing between developmental mistakes and transfer-induced errors to facilitate more effective intervention.

Contemporary investigations further reveal that learner errors exhibit systematic patterns that can inform instructional design. Dependency relations, conceptual transfer, and bilingual processing mechanisms influence the production and interpretation of written texts among language learners [17], [18]. Visual analyses of SLA research trends likewise indicate growing scholarly interest in bilingual acquisition, interlanguage

development, and evidence-based pedagogical practices aimed at improving writing performance [19], [20].

Traditional approaches to writing instruction primarily rely on teacher-mediated feedback. Although teacher feedback is widely recognized as valuable, it often requires considerable time and effort, particularly in large classes where individualized guidance becomes difficult to sustain. Learners may consequently receive delayed feedback, limiting opportunities for immediate revision and reflection [21], [22]. These limitations have encouraged the exploration of technology-enhanced instructional approaches capable of providing timely and consistent support.

Recent advances in natural language processing (NLP) and artificial intelligence have transformed educational practices through the development of intelligent systems capable of analyzing textual data and generating meaningful recommendations. NLP techniques have demonstrated effectiveness in extracting linguistic features, identifying patterns, and supporting automated analysis across diverse domains [23], [24]. In educational contexts, such technologies offer opportunities to augment teacher expertise through intelligent feedback mechanisms that facilitate formative assessment.

Automated writing evaluation systems have increasingly incorporated machine learning and NLP algorithms to detect grammatical inaccuracies and provide corrective suggestions. Resume screening systems employing NLP techniques have demonstrated the capacity of computational models to identify patterns within unstructured textual documents [25], [26]. Although these applications originate outside language education, they illustrate the broader potential of NLP-assisted systems for analyzing written language and generating evidence-based recommendations.

The growing adoption of intelligent systems has also generated important discussions regarding transparency, fairness, and ethical implementation. Scholars emphasize that algorithmic decision-making should remain interpretable and accountable to avoid reinforcing hidden biases [27], [28]. Reviews of fairness in artificial intelligence advocate the integration of explainable mechanisms that enable users to understand the rationale underlying system-generated outputs [29], [30]. Within educational settings, such considerations are essential because automated feedback directly influences learners' perceptions of their abilities and subsequent learning behaviors.

Despite the emergence of commercial grammar-checking tools, many existing systems emphasize surface-level corrections while providing limited pedagogical explanations tailored to learners' developmental needs. Few systems explicitly integrate principles from error analysis, interlanguage theory, and native language transfer into their feedback mechanisms. Consequently, learners may receive corrective suggestions without understanding the linguistic causes of their errors or strategies for improvement [31], [32].

A rule-based framework informed by SLA theories offers a promising alternative for educational applications. Rule-based systems enable transparent identification of linguistic patterns while maintaining alignment with pedagogical objectives [33], [34]. When integrated with NLP-assisted text processing, such systems can provide immediate diagnostic feedback, classify error types, and generate instructional recommendations that support learner reflection and self-regulation.

Recent advances in automated writing evaluation (AWE) have enabled the development of intelligent systems capable of detecting grammatical, lexical, and stylistic errors while providing immediate corrective feedback to language learners. Commercial platforms such as Grammarly, Criterion, and Cambridge Write & Improve employ natural language processing and artificial intelligence techniques to assist users in improving written communication. These systems have demonstrated effectiveness in identifying surface-level grammatical errors and enhancing writing fluency; however, their feedback mechanisms often emphasize correction rather than explaining the linguistic reasons underlying learner errors. Consequently, opportunities for deeper language learning and learner reflection remain limited, particularly for English as a Second Language (ESL) learners whose writing difficulties are strongly influenced by first-language transfer and developmental interlanguage.

Recent research has also explored explainable artificial intelligence (XAI) in educational technologies to improve the transparency and interpretability of intelligent tutoring systems. Rather than providing opaque automated corrections, explainable systems seek to present understandable reasoning that enables learners to recognize why an error occurred and how similar mistakes can be avoided. Within second language learning, integrating explainable feedback with principles of Error Analysis and Interlanguage Theory has the potential to support meaningful language development while preserving teachers' instructional roles.

Despite these developments, relatively few intelligent writing support systems have been designed specifically for learners whose writing reflects localized linguistic characteristics such as Ilocano language interference. Existing automated writing evaluation systems generally rely on large multilingual corpora and statistical language models, which may not adequately capture recurrent error patterns associated with specific first-language backgrounds. The present study addresses this gap by developing an explainable Intelligent Writing Tutor that integrates corpus-informed analysis, second language acquisition theories, and rule-based natural language processing to generate pedagogically meaningful and context-sensitive writing feedback for Ilocano-speaking ESL learners.

Guided by these considerations, this study proposes an Intelligent Writing Tutor for ESL Learners that integrates rule-based reasoning and NLP-assisted feedback mechanisms. The framework seeks to detect common writing errors, categorize them according to established linguistic theories, and generate targeted recommendations that complement teacher instruction. By combining theoretical foundations from SLA with contemporary computational approaches, the proposed framework aims to enhance feedback quality, promote learner autonomy, and contribute to the development of explainable educational technologies that support ESL writing

improvement. This study sought to answer the following research questions:

- What lexical, morphological, syntactic, and mechanical errors are most frequently produced by Ilocano-speaking ESL learners?
- How can the identified writing error patterns be transformed into a rule-based and NLP-assisted Intelligent Writing Tutor framework?
- How can corpus-informed linguistic analysis and explainable rule-based reasoning be integrated to generate individualized writing feedback for ESL learners?
- How do language experts evaluate the proposed Intelligent Writing Tutor in terms of accuracy, relevance, clarity, usefulness, and practicality?

This study makes four primary contributions. First, it develops an authentic learner corpus consisting of journal writings produced by Ilocano-speaking ESL learners, providing empirical evidence of recurrent writing difficulties influenced by first-language transfer. Second, it proposes an explainable Intelligent Writing Tutor that combines corpus-informed linguistic analysis, natural language processing, and rule-based reasoning to generate individualized writing feedback. Third, the framework operationalizes principles from Error Analysis, Interlanguage Theory, and Language Transfer Theory into interpretable computational rules that support pedagogically meaningful recommendations. Finally, the study contributes an educational technology framework that emphasizes transparency, instructional relevance, and explainability, thereby complementing rather than replacing teacher-mediated writing instruction.

II. Methodology

A. Research Design

This study employed a Design Science Research (DSR) methodology integrated with a mixed-methods research design to develop and evaluate an Intelligent Writing Tutor for ESL learners. The Design Science Research approach was adopted because the primary objective of the study was to design, develop, and validate an educational artifact capable of addressing authentic challenges encountered by ESL learners in English writing. The artifact developed in this study consisted of a rule-based and Natural Language Processing (NLP)-assisted feedback framework intended to support teachers and learners through automated writing diagnosis and targeted feedback generation. The mixed-methods design enabled the integration of quantitative and qualitative approaches throughout the investigation. The quantitative component focused on the identification, categorization, and frequency analysis of writing errors occurring in learners' written outputs. The qualitative component involved the interpretation of error manifestations, examination of L1 interference patterns, and formulation of pedagogically meaningful feedback recommendations. This methodological integration provided a comprehensive understanding of learners' writing difficulties while facilitating the development of an evidence-based intelligent tutoring framework. Consistent with Design Science Research principles, the study followed problem identification and motivation, artifact design and development, demonstration using authentic learner data, evaluation of the generated feedback mechanisms, and communication of findings as major phases. The methodological structure was adapted from contemporary AI-assisted educational frameworks while remaining grounded in established theories of second language acquisition and error analysis. Fig. 1 illustrates the overall process employed in the study to identify first language (L1) interference patterns in English as a Second Language (ESL) writing and transform the findings into targeted instructional interventions.

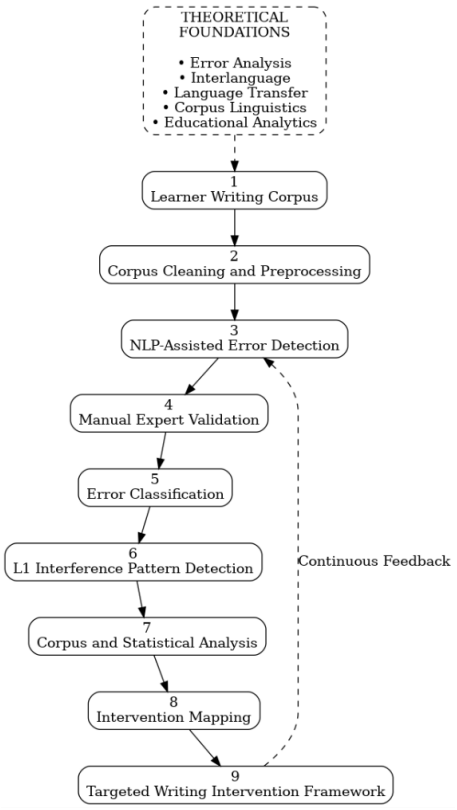


Fig. 1. Conceptual and operational framework for detecting L1 interference patterns in ESL writing toward targeted writing interventions.

The framework integrates theoretical foundations from applied linguistics with corpus-based and natural language processing (NLP) techniques to support evidence-based decision-making in writing instruction.

At the top of the framework are the theoretical foundations, which provide the conceptual basis of the study. Error Analysis explains how learner errors are systematically examined to identify recurring difficulties in language production. Interlanguage Theory views learner language as a transitional linguistic system that evolves during second language acquisition. Language Transfer Theory accounts for the influence of learners' native language structures on their English writing performance. Corpus Linguistics provides methods for analyzing authentic learner texts through large collections of written data, while Educational Analytics supports the interpretation of writing data to generate instructional insights and interventions.

The operational process begins with Learner Writing Corpus, where essays and written outputs produced by ESL learners are collected to form the corpus used in the analysis. These authentic texts serve as the primary data source for identifying language patterns and writing difficulties.

In Corpus Cleaning and Preprocessing, the collected texts undergo preparation procedures to improve data quality. This stage includes organizing the corpus, removing irrelevant symbols or duplicate entries, standardizing formats, and preparing the textual data for computational analysis.

The preprocessed corpus is then subjected to NLP-Assisted Error Detection. At this stage, natural language processing tools automatically identify potential grammatical, lexical, and syntactic errors found in learner writing. Automated detection increases efficiency by processing large amounts of text and highlighting probable error locations.

Recognizing that automated methods may produce inaccurate classifications, Manual Expert Validation is incorporated. Language experts review the NLP-generated outputs to confirm the accuracy of identified errors, eliminate false detections, and ensure the reliability of the findings.

Validated errors are subsequently organized in Error Classification. The errors are categorized according to predefined linguistic categories such as grammatical, lexical, mechanical, or syntactic errors. This classification facilitates a deeper understanding of the specific challenges experienced by ESL learners.

After classification, L1 Interference Pattern Detection determines whether the identified errors reflect influences from the learners' native language. This stage distinguishes transfer-related errors from developmental errors associated with the natural process of second language acquisition.

The detected patterns are then subjected to Corpus and Statistical Analysis. Frequency counts, descriptive statistics, and corpus-based analyses are performed to determine the prevalence, distribution, and significance of different error types and interference patterns. The results reveal dominant learner difficulties and provide empirical evidence for instructional planning.

Based on these analytical findings, Intervention Mapping links the identified learner needs to appropriate instructional responses. Specific interventions are designed to address the most frequent and persistent error patterns observed in the corpus.

The process culminates in the Targeted Writing Intervention Framework, which represents the final output of the study. The framework proposes evidence-based instructional strategies, focused grammar instruction, corpus-informed learning activities, reflective writing practices, and feedback mechanisms intended to mitigate L1 interference and improve ESL writing competence.

A notable feature of the framework is the Continuous Feedback Loop, represented by the dashed arrow connecting the final intervention framework back to the NLP-assisted error detection stage. This feedback mechanism indicates that interventions are not static; rather, learner progress is continuously monitored through subsequent writing outputs. New data generated after intervention implementation can be reanalyzed, allowing the framework to refine error detection processes, validate instructional effectiveness, and adapt future interventions based on evolving learner needs.

B. Research Locale and Respondents

The study was conducted at the College of Education of Kalinga State University, Philippines. The institution was selected because English language majors regularly engage in academic writing activities that provide authentic samples for examining second language writing performance. Twenty participants consisted of second-year and third-year Bachelor of Secondary Education major in English students whose first language was Ilocano and who learned English as a second language. Participants were selected through purposive sampling based on the following inclusion criteria:

- They were officially enrolled during the conduct of the study.
- English was learned as a second language.
- Ilocano served as their primary first language (L1).
- They regularly submitted journal or diary writing outputs as part of their coursework requirements.

This participant profile provided a linguistically homogeneous learner population, allowing the study to examine recurring writing errors and first-language interference patterns with minimal variation arising from different linguistic backgrounds. The use of authentic classroom writing further strengthened the ecological validity of the learner corpus because the collected texts reflected naturally occurring language production rather than responses generated under examination conditions.

C. Corpus of the Study

The learner corpus analyzed in this study consisted of weekly journal entries and reflective writings produced by the participating English language majors throughout the data collection period. Students were encouraged to write about personal experiences, academic activities, reflections, and other meaningful events using English as the medium of expression. Because the writing tasks emphasized authentic communication rather than examination performance, the collected texts captured naturally occurring learner language and genuine manifestations of interlanguage development.

Each journal entry served as an independent corpus document. Following collection, all writings were digitized, anonymized, and assigned unique identification codes to ensure confidentiality throughout the analysis. The completed corpus was subsequently organized into a structured electronic repository for annotation and computational processing.

The learner corpus served two important purposes. First, it provided empirical evidence for identifying recurring lexical, morphological, syntactic, and mechanical errors among Ilocano-speaking ESL learners. Second, it functioned as the primary knowledge source for constructing the handcrafted linguistic rules implemented within the proposed Intelligent Writing Tutor.

D. Corpus Annotation and Validations

Prior to computational analysis, the learner corpus underwent manual annotation to establish the reference standard for identifying writing errors. Annotation was guided by Corder's Error Analysis Framework [35], which classifies learner errors according to their linguistic characteristics and probable underlying causes. Each journal entry was examined to identify lexical, morphological, syntactic, and mechanical errors before being assigned to the appropriate linguistic category.

To improve annotation consistency, the annotated corpus was independently reviewed by language experts with specialization in English language teaching and applied linguistics. Differences in annotation were discussed until consensus was reached before the finalized annotations were incorporated into the rule base of the Intelligent Writing Tutor.

The manually validated corpus served as the ground truth for constructing the rule-based feedback engine and ensuring that the automated error detection module reflected authentic learner writing patterns rather than purely computational assumptions.

E. Proposed Intelligent Writing Tutor Framework

The proposed Intelligent Writing Tutor (IWT) transforms learners' written outputs into targeted instructional feedback by integrating rule-based reasoning with natural language processing (NLP) techniques. The framework consists of five interconnected modules: Input, Preprocessing, Error Detection, Feedback Generation, and Recommendation, which are also illustrated in Fig. 2. The Input Module collects learners' digitized writing outputs, particularly journal entries, which serve as the primary source of textual data. These writing samples are then processed by the Preprocessing Module, where standard NLP procedures such as lowercasing, tokenization, sentence segmentation, punctuation normalization, part-of-speech tagging, and lemmatization are performed to prepare the texts for analysis.

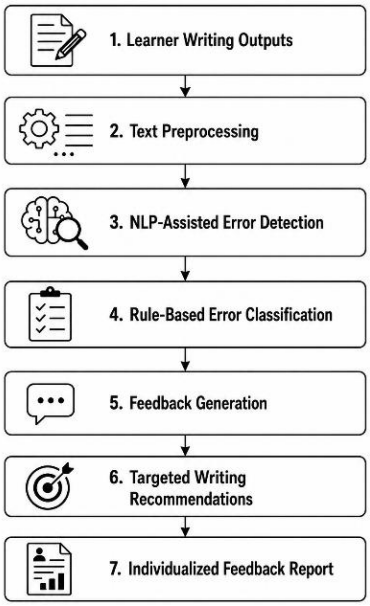


Fig. 2. Overall architecture of the proposed Intelligent Writing Tutor showing the interaction between the input module, NLP processing pipeline, rule-based error detection, feedback generation, and recommendation engine.

The Intelligent Writing Tutor employs a hybrid architecture that combines conventional natural language processing techniques with handcrafted linguistic rules. During preprocessing, learner texts undergo normalization, sentence segmentation, tokenization, part-of-speech tagging, lemmatization, and dependency parsing to generate linguistic representations suitable for subsequent analysis. The extracted linguistic features are then evaluated using rule-based pattern matching to detect lexical, morphological, syntactic, and mechanical errors. Rather than relying solely on statistical prediction, the framework explicitly associates each detected error with interpretable grammatical rules that enable the generation of pedagogically meaningful explanations and corrective recommendations.

The Error Detection Module employs a hybrid approach combining NLP-assisted techniques and handcrafted linguistic rules to identify four categories of writing errors: lexical, morphological, syntactic, and mechanical errors. The detected errors are subsequently interpreted by the Feedback Generation Module, which provides explanations of the identified errors, examples of correct usage, suggested revisions, and learning tips. When applicable, the system also indicates whether the errors may be associated with developmental interlanguage processes or potential L1 interference. The Recommendation Module synthesizes the analysis results and generates individualized feedback reports. These reports summarize the frequency and dominant types of errors, identify possible interference patterns, and recommend appropriate learning activities and writing improvement strategies tailored to the learner's needs.

F. Data Collection Procedure

Data collection was conducted through a structured five-phase procedure to ensure the systematic development of the Intelligent Writing Tutor framework.

1) Corpus compilation. Students' weekly journal entries served as the primary source of data. The collected writings were digitized and organized into an electronic corpus. To facilitate tracking, management, and confidentiality, each learner document was assigned a unique identification code.

2) Corpus preparation. The compiled corpus underwent preprocessing to improve consistency and analytical readiness. This phase involved standardizing document formats, correcting encoding inconsistencies, and cleaning the textual data. Non-linguistic elements unrelated to the objectives of the study, such as irrelevant symbols, duplicate entries, and extraneous formatting marks, were removed from the corpus.

3) Manual expert annotation. Initial error analysis was performed manually by the researcher following the procedures proposed by Corder. The process involved three major stages: error identification, error description, and error explanation. Identified errors were subsequently classified into four linguistic categories, namely lexical, morphological, syntactic, and mechanical errors. This annotation process established the reference framework for identifying recurring learner difficulties and validating subsequent system outputs.

4) Rule construction. Patterns emerging from the annotated corpus were examined to identify frequently occurring writing errors. These recurrent patterns informed the formulation of explicit linguistic rules that formed the knowledge base of the Intelligent Writing Tutor. The resulting rule set enabled the system to recognize specific error types and provide context-appropriate feedback and intervention suggestions.

5) System demonstration. The completed framework was then applied to actual learner writing samples to demonstrate its operational capability. Through this process, the Intelligent Writing Tutor generated automated feedback, targeted writing recommendations, and individualized feedback reports designed to support learners in improving their writing performance and addressing identified weaknesses.

G. Data Preprocessing

Prior to automated error detection, all learner writings underwent a structured preprocessing workflow to improve data quality and analytical consistency. First, all documents were converted into a standardized electronic format and transformed into lowercase text to ensure uniform lexical representation. Non-linguistic symbols, duplicate characters, and formatting inconsistencies were subsequently removed.

The cleaned texts then underwent sentence segmentation followed by tokenization, enabling the identification of individual lexical units. Part-of-speech tagging assigned grammatical categories to each token, while lemmatization reduced inflected words to their canonical forms. Dependency parsing was subsequently applied to identify grammatical relationships among sentence constituents, providing syntactic information necessary for rule-based error detection.

Finally, the extracted linguistic features were transformed into structured representations that served as input to the Intelligent Writing Tutor. These preprocessing procedures enhanced the consistency, reliability, and interpretability of the subsequent NLP-assisted error detection process.

H. NLP-Assisted Error Detection Module

The proposed framework employed natural language processing (NLP) techniques to automatically identify linguistic deviations in learners' written outputs. The error detection process encompassed four major categories of writing errors.

Lexical analysis was conducted to detect inappropriate word choice, collocational misuse, preposition misuse, and literal translations that may reflect L1 influence. Morphological analysis identified tense inconsistencies, incorrect verb forms, pluralization errors, and omissions of auxiliary verbs. Syntactic analysis focused on detecting subject–verb disagreement, sentence fragments, run-on sentences, violations of parallel structure, and other faulty sentence constructions. In addition, mechanical analysis examined punctuation errors, capitalization problems, spelling inaccuracies, and inappropriate formatting conventions.

The identified linguistic categories were derived from error patterns commonly reported in ESL writing studies and reflected the types of errors frequently observed among language majors' written compositions. By integrating NLP techniques with linguistically informed rules, the module enabled systematic, efficient, and consistent detection of writing errors for subsequent classification and feedback generation.

I. Rule-Based Feedback Engine

The Intelligent Writing Tutor employed an interpretable rule-based engine to transform detected errors into pedagogically meaningful feedback. Feedback rules followed the structure presented in Algorithm 1. The feedback engine operates using explicit IF–THEN production rules derived from recurring writing errors observed in the annotated learner corpus. Each detected linguistic error activates a corresponding rule that determines the appropriate error category, identifies possible causes, generates pedagogical explanations, retrieves relevant corrective examples, and recommends targeted learning activities. Because the decision-making process is rule-driven rather than probabilistic, every feedback message remains transparent and fully explainable to both teachers and learners.

Algorithm 1. Rule-Based Feedback Engine of the Intelligent Writing Tutor

```
Input: Detected writing errors from the NLP-assisted error detection module
Output: Individualized feedback, corrective examples, and targeted learning recommendations
    Initialize Feedback Report;
    For each detected error do
      Identify the corresponding error pattern;
      Determine the error category
        (Lexical, Morphological, Syntactic, or Mechanical);
      Analyze the possible cause of the error;
      Generate a corrective explanation describing
      why the error occurred;
      Retrieve or construct an appropriate
      correct example illustrating proper usage;
      Recommend a learning strategy or intervention
      aligned with the identified error type;
      Append the generated feedback to the
      learner's feedback report;
    End For
    Produce the individualized feedback report;
    Display targeted writing recommendations;
    Return the completed feedback report.
    Example Rule Execution
    If Subject–Verb Agreement Error is Detected Then
      Category ← Syntactic Error;
      Explanation ← "The verb does not agree with the
              subject in number.";
      Correct Example ← "She goes to school every day.";
      Recommendation ← "Review singular and plural
              subject–verb agreement rules.";
    End If
```

The rule-based feedback engine transforms identified learner errors into interpretable and pedagogically meaningful responses. By explicitly linking detected error patterns with explanations, examples, and intervention strategies, the framework promotes transparency and enables learners to

understand the rationale behind each recommendation while supporting self-regulated improvement in writing performance.

J. Evaluation of the Framework

The proposed Intelligent Writing Tutor framework underwent a comprehensive evaluation to establish its validity, pedagogical relevance, and practical applicability in supporting ESL writing instruction. Evaluation was conducted through two complementary approaches: expert review and corpus-based system demonstration. The expert review examined the extent to which the framework aligned with instructional needs and accepted principles in language education and educational technology. A panel of experts composed of specialists in ESL instruction, Applied Linguistics, and Educational Technology was purposively selected to assess the quality and functionality of the framework. These experts evaluated the framework using a five-point Likert scale, where 5 indicated Strongly Agree, 4 indicated Agree, 3 indicated Moderately Agree, 2 indicated Disagree, and 1 indicated Strongly Disagree. The evaluation focused on five key dimensions considered essential for intelligent educational support systems, as presented in Table I.

TABLE I. FIVE KEY DIMENSIONS CONSIDERED ESSENTIAL FOR INTELLIGENT EDUCATIONAL SUPPORT SYSTEMS.

Criterion	Description
Accuracy	The extent to which the framework correctly identifies and classifies writing errors and generates appropriate feedback.
Relevance	The degree to which the generated explanations and recommendations address learners' actual writing needs and difficulties.
Clarity	The understandability, coherence, and comprehensibility of the feedback and corrective explanations produced by the system.
Usefulness	The perceived potential of the framework to facilitate writing improvement, promote learner awareness, and support instructional decision-making.
Practicality	The feasibility of implementing the framework within authentic educational environments considering available resources, teacher workload, and classroom conditions.

In addition to expert validation, a corpus demonstration was conducted to examine the operational capability of the Intelligent Writing Tutor. Actual learner writings were processed through the framework to demonstrate its ability to detect errors, classify error types, generate corrective explanations, provide illustrative examples, and produce targeted writing recommendations. This demonstration enabled evaluators to examine the transparency and pedagogical appropriateness of the framework outputs under realistic instructional scenarios. The weighted mean for each evaluation criterion was computed to determine the overall level of acceptability of the framework. The resulting interpretations provided empirical evidence regarding the strengths and areas for enhancement of the proposed system prior to its wider implementation.

K. Data Analysis

The study employed both quantitative and qualitative analytical procedures to examine learner writing errors and generate evidence-based interventions. The integration of these approaches enabled a comprehensive understanding of learners' writing difficulties while informing the development of individualized feedback mechanisms within the Intelligent Writing Tutor.

1) Quantitative analysis. Quantitative analysis focused on determining the prevalence and distribution of identified writing errors within the learner corpus. Frequency counts and percentage distributions were computed to identify dominant error categories and recurring linguistic patterns. The percentage of occurrence for each error type was calculated using the following formula:

$$P = \frac{f}{N} \times 100 \quad (1)$$

where, P is the percentage of occurrence of a specific error category; f is the frequency of occurrence of the identified error type; and N is the total number of errors identified from the learner corpus.

The resulting descriptive statistics facilitated the identification of the most problematic areas in learner writing and provided empirical evidence for prioritizing instructional interventions. Error distributions were subsequently organized according to lexical, morphological, syntactic, and mechanical categories to determine patterns requiring targeted support. For the framework evaluation, weighted means were likewise computed to summarize expert ratings across the established evaluation criteria. The weighted mean was determined using:

$$WM = \frac{\sum fx}{N} \quad (2)$$

where, WM is the weighted mean; f is the frequency of responses for each rating category; x numerical value assigned to each Likert-scale response; and N is the total number of responses.

The computed means served as indicators of the overall acceptability and effectiveness of the proposed framework.

2) Qualitative analysis. Qualitative analysis was guided by Corder's Error Analysis Framework, which provided a systematic procedure for examining the nature and possible causes of learner errors. The analysis involved three interconnected stages:

a) Error identification: locating deviations from accepted English language conventions within the learner corpus.

b) Error description: classifying and describing the linguistic characteristics of the identified errors.

c) Error explanation: interpreting the potential sources and underlying causes of the observed error patterns.

The qualitative findings enabled deeper insights into learners' interlanguage development and the influence of first-language transfer on second-language writing performance. Furthermore, these analysis informed the construction of rule-based explanations and pedagogically appropriate recommendations embedded within the Intelligent Writing Tutor.

The integration of quantitative frequency analysis and qualitative error interpretation facilitated the development of individualized interventions grounded in authentic learner data.

Through this mixed analytical approach, the framework not only identified what errors learners produced but also explained why such errors occurred and how they could be effectively addressed.

By synthesizing authentic learner corpora, principles of second language acquisition, corpus-informed evidence, natural language processing techniques, and interpretable rule-based reasoning, the proposed methodology establishes a transparent, evidence-driven, and pedagogically grounded Intelligent Writing Tutor. The resulting framework is capable of generating individualized feedback, targeted recommendations, and actionable instructional insights that support ESL learners in improving their writing proficiency while assisting teachers in delivering data-informed and responsive writing instruction.

L. Ethical Considerations

Ethical approval was obtained from the appropriate institutional ethics review body prior to data collection. The study adhered to the principles of respect for persons, beneficence, justice, confidentiality, and voluntary participation. Participants were informed of the study's objectives, procedures, voluntary nature of their involvement, their right to withdraw without penalty, and the intended use of their weekly journal entries for research purposes before providing informed consent. To ensure anonymity and confidentiality, learner writings were assigned unique identification codes and all personally identifiable information was removed from the corpus. Digital files and research records were stored in password-protected repositories accessible only to authorized researchers. Anonymized learner texts were used solely for system demonstration and academic purposes, and all findings were reported in aggregate form to protect participants' rights, privacy, and welfare. The generated feedback had no effect on participants' academic standing or evaluation.

III. Results and Discussion

A. Distribution of Language Errors

The analysis of the diary entries revealed that language majors experienced difficulties across morphological, syntactic, lexical, and mechanical aspects of writing. The results indicate that morphological errors constituted the most prominent category of errors, followed by mechanical, lexical, and syntactic errors. These findings suggest that learners encountered greater difficulty in applying grammatical rules governing word formation and tense usage than in sentence construction or punctuation.

1) Morphological errors. Morphological errors totaled 68 occurrences, with verb tense errors emerging as the most dominant subtype. Verb tense misuse accounted for 73.33% of all morphological errors, followed by pluralization errors (10.66%) and verb form errors (6.67%). The predominance of verb tense errors indicates that students struggled to express temporal relationships consistently in English. Since Ilocano verbs do not exhibit the same tense-marking system as English, learners frequently transferred first-language structures into their English compositions. Consequently, students shifted between present and past forms without maintaining temporal consistency. Pluralization errors further suggest incomplete mastery of English inflectional morphology. Difficulties involving regular and irregular noun forms indicate that learners tended to generalize grammatical rules or omitted plural markers altogether. Similarly, errors involving verb forms reflected uncertainty regarding the appropriate use of base forms, participles, and gerunds. These findings imply that learners require more explicit instruction on English tense systems and opportunities for repeated contextualized practice to internalize morphological patterns.

2) Syntactic errors. Twenty-six syntactic errors were identified from the corpus. Subject–verb agreement constituted the largest proportion (38.46%), followed by run-on sentences (23.08%), sentence structure errors (19.23%), parallelism errors (11.54%), and sentence fragments (7.69%). Subject–verb agreement difficulties indicate that students had problems recognizing grammatical relationships between sentence constituents. Such errors are often associated with negative language transfer because Ilocano does not consistently require agreement markers similar to those used in English. Run-on sentences suggest limited awareness of sentence boundaries and clause coordination. Students tended to connect independent ideas without employing appropriate punctuation or conjunctions. The occurrence of fragments further reflects challenges in constructing complete thoughts containing both subjects and predicates. Collectively, these findings reveal weaknesses in syntactic competence and demonstrate the need for instructional interventions emphasizing sentence construction, clause relationships, and editing strategies.

3) Lexical errors. Lexical analysis revealed 43 errors distributed across three categories. Word choice and correct usage represented the majority of lexical errors at 51.11%, followed by prepositional errors at 26.67%, while article errors accounted for 17.78%. The dominance of inappropriate word choice suggests limitations in vocabulary depth and semantic precision. Learners often selected words that were grammatically acceptable but contextually inappropriate, thereby obscuring intended meanings. Examples included the use of terms such as "heavy group" and other expressions directly translated from the learners' first language. Preposition errors indicate difficulties in understanding collocational patterns and the relational meanings conveyed by English prepositions. Because prepositional systems differ substantially between Ilocano and English, learners frequently omitted, substituted, or inserted incorrect prepositions. Article misuse likewise reflects the absence of equivalent article systems in many Philippine languages, resulting in confusion regarding the use of "a", "an", and "the". These results suggest that lexical development should extend beyond vocabulary memorization and include contextualized exposure to authentic language use.

4) Mechanical errors. Mechanical errors comprised 46 occurrences. Capitalization errors emerged as the most prevalent category, representing 60.87% of mechanical errors. Spelling errors accounted for 30.43%, while punctuation errors constituted 8.70%. The predominance of capitalization errors suggests inadequate familiarity with conventions governing

sentence beginnings, titles, and proper nouns. Spelling difficulties indicate weaknesses in orthographic knowledge and limited proofreading practices. Although punctuation errors occurred less frequently, they nevertheless affected clarity and coherence. Mechanical errors, despite being surface-level concerns, can influence readers' perceptions of writing quality. Consequently, explicit instruction in editing and self-monitoring strategies remains essential.

5) Aggregation of errors. To provide a clearer overview of the distribution of learner writing difficulties, the identified errors were aggregated according to the four major linguistic categories. Morphological errors constituted the largest proportion of all identified errors, indicating that learners experienced the greatest difficulty in applying English inflectional rules, particularly verb tense formation and agreement. Mechanical errors ranked second, reflecting recurring problems in capitalization, spelling, and punctuation. Lexical errors were primarily associated with inappropriate word choice and preposition usage, whereas syntactic errors occurred less frequently but demonstrated persistent weaknesses in sentence construction and grammatical agreement. Collectively, these findings indicate that learner difficulties extend beyond isolated grammatical mistakes and reflect systematic patterns associated with second language development and first-language transfer.

B. Patterns of Manifestation of Errors

Analysis of representative excerpts revealed recurring patterns underlying the identified errors. Verb tense errors manifested through three major patterns: inconsistency in tense usage, omission of auxiliary verbs, and incorrect verb forms used to indicate time relationships. Students frequently shifted from present to past tense within the same narrative without clear justification, creating ambiguity regarding the sequence of events. Lexical errors commonly appeared through inappropriate word choices and ambiguous expressions. Learners selected words that resembled intended meanings but failed to convey accurate semantic relationships. These patterns indicate limited lexical repertoire and overreliance on direct translation strategies. Article errors manifested through the interchange of indefinite articles and mismatches between articles and noun number, while prepositional errors involved incorrect selection and grammatical inconsistencies. These manifestations demonstrate learners' attempts to apply English rules despite incomplete understanding of their usage conditions. The recurrence of similar patterns across participants suggests that the observed errors were systematic rather than random.

C. L1 Interference in Learners' Writing

Qualitative analysis revealed that many of the identified errors reflected negative transfer from learners' first language. The findings indicate that L1 interference influenced morphological, syntactic, and lexical choices during English composition. Morphologically, the inconsistent marking of tense and omission of auxiliary verbs suggest reliance on native language structures that do not encode temporal distinctions in the same manner as English. Syntactically, subject–verb agreement errors demonstrate difficulties in adapting to grammatical relationships absent or less explicit in the first language. Lexically, learners frequently employed direct translations, resulting in awkward expressions and inappropriate word combinations. Prepositional misuse further illustrates attempts to map first-language conceptual relationships onto English structures. These findings support theories of language transfer, which posit that previously acquired linguistic knowledge influences second language production. While transfer may facilitate learning when structures are similar, substantial differences between languages often lead to systematic errors.

D. Implications for the Intelligent Writing Tutor Framework

The error patterns identified in the corpus provided empirical evidence for constructing the rule base of the proposed Intelligent Writing Tutor. Frequent verb tense inconsistencies, subject–verb agreement errors, inappropriate lexical choices, and capitalization problems informed the development of interpretable feedback rules capable of detecting recurrent learner difficulties. Drawing from the transparent and explainable feedback mechanisms demonstrated in recent intelligent educational systems, the framework translates detected error patterns into corrective explanations, examples, and targeted learning recommendations. The emphasis on interpretability mirrors the role of explainable systems in helping users understand the rationale behind automated outputs. Consequently, the Intelligent Writing Tutor extends traditional error analysis by moving beyond error identification toward personalized intervention. Through automated yet pedagogically grounded feedback, the framework supports learners in recognizing recurring weaknesses, developing self-editing skills, and improving overall writing proficiency.

The findings demonstrate that language majors encounter persistent challenges in English writing that are closely associated with first-language influence and incomplete mastery of target-language conventions. These results underscore the value of integrating corpus-informed analysis, error analysis principles, and rule-based natural language processing to provide evidence-based writing support for ESL learners.

E. Expert Evaluation of the Proposed Framework

To determine the educational acceptability of the proposed Intelligent Writing Tutor, the framework was evaluated by specialists in English language teaching, applied linguistics, and educational technology using the five evaluation dimensions presented in Table I. The evaluation examined the framework's accuracy, relevance, clarity, usefulness, and practicality in supporting ESL writing instruction.

The expert assessment indicated that the framework demonstrated satisfactory educational quality across all evaluation criteria. Experts particularly recognized the transparency of the rule-based feedback mechanism and the alignment of the generated recommendations with established principles of second language acquisition. The use of explicit linguistic rules enabled evaluators to understand how feedback was generated, thereby increasing confidence in the interpretability of the proposed system.

Although the evaluation confirmed the pedagogical relevance of the framework, several recommendations for future enhancement were identified. These included expanding the linguistic rule base, improving automated handling of complex sentence structures, and integrating more advanced natural language processing techniques to improve contextual feedback generation.

F. Computational Performance of the Error Detection Module

The operational capability of the proposed Intelligent Writing Tutor was examined by comparing the automatically detected writing errors against the manually annotated reference corpus. The comparison evaluated the agreement between the NLP-assisted detection module and expert annotations using standard information retrieval metrics, including Precision, Recall, and F1-score.

Precision measures the proportion of automatically detected errors that were correctly identified, while Recall indicates the proportion of actual learner errors successfully detected by the framework. The F1-score provides the harmonic mean of Precision and Recall, representing the overall effectiveness of the error detection module.

The use of these evaluation metrics provides objective evidence regarding the computational reliability of the proposed framework while complementing the expert-based educational evaluation.

G. Operationalization of the Intelligent Writing Tutor Framework

Beyond identifying the dominant writing errors, the study demonstrated how the proposed Intelligent Writing Tutor operationalizes authentic learner corpus data into individualized feedback and targeted instructional recommendations. The framework transforms learners' written outputs through a sequence of interconnected processes involving preprocessing, automated error detection, rule-based reasoning, and feedback generation.

The operational process begins with the submission of learners' weekly journal entries. These texts undergo preprocessing procedures, including text cleaning, tokenization, sentence segmentation, and linguistic normalization. The preprocessed texts are then analyzed by the NLP-assisted error detection module, which identifies potential lexical, morphological, syntactic, and mechanical deviations based on predefined linguistic patterns derived from the annotated corpus.

Detected errors are subsequently processed by the rule-based feedback engine. Using explicit IF–THEN rules constructed from recurring learner error patterns, the system classifies each error according to its linguistic category, determines its possible cause, generates corrective explanations, retrieves appropriate examples, and recommends corresponding learning strategies. The generated outputs are consolidated into individualized feedback reports designed to address the specific needs of each learner. Fig. 3 shows the operational flow of the Intelligent Writing Tutor Framework.

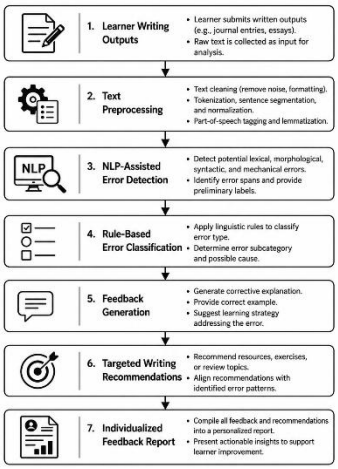


Fig. 3. End-to-end operational workflow demonstrating the transformation of learner writing into individualized feedback through NLP-assisted processing and rule-based reasoning.

To illustrate the framework's operation, consider the learner sentence:

"She go to school every day."

During preprocessing, the sentence is segmented and tokenized. The NLP module detects a mismatch between the singular subject (She) and the base verb form (go). The rule-based engine activates the corresponding subject–verb agreement rule and generates the following output:

- Error Category: Syntactic Error.
- Possible Cause: Incomplete mastery of subject–verb agreement rules and possible L1 influence.
- Corrective Explanation: The verb must agree with the subject in number.
- Correct Example: "She goes to school every day."
- Learning Recommendation: Review rules governing singular and plural subject–verb agreement and complete practice exercises involving present tense verbs.

Similarly, if the learner writes:

"Yesterday, I go to the market and buy vegetables."

The framework identifies verb tense inconsistency and produces feedback emphasizing the appropriate use of past tense forms. The learner receives explanations, corrected examples, and recommendations for reviewing tense formation and temporal consistency.

These demonstrations show that the Intelligent Writing Tutor does not merely identify errors but converts diagnostic information into meaningful pedagogical interventions. By integrating corpus evidence, principles of error analysis, and transparent rule-based reasoning, the framework provides interpretable and actionable feedback that supports learner autonomy, promotes self-editing skills, and assists teachers in delivering data-informed writing instruction.

The operationalization of the framework further highlights its potential applicability in ESL classrooms. Rather than replacing teacher judgment, the system functions as a supplementary instructional tool capable of providing immediate, individualized, and evidence-based feedback. Consequently, the framework bridges the gap between corpus-informed research findings and practical classroom interventions, enabling more responsive and targeted support for ESL learners.

H. Prototype of the Intelligent Writing Tutor

Fig. 4 illustrates the prototype of the proposed Intelligent Writing Tutor and demonstrates how the system transforms learners' written outputs into individualized feedback and instructional recommendations.

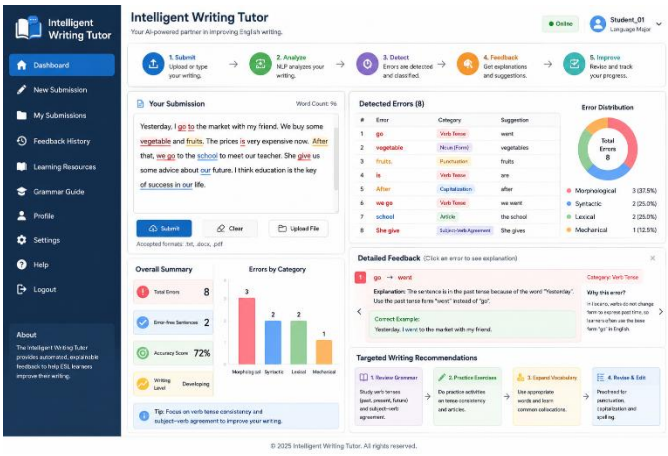


Fig. 4. Prototype interface of the Intelligent Writing Tutor illustrating learner text submission, automated error detection, rule-based feedback generation, and individualized writing recommendations.

The process begins when learners submit their essays or journal entries through the writing interface. The submitted text is then preprocessed through cleaning, tokenization, sentence segmentation, and linguistic normalization to prepare it for analysis.

The NLP-assisted error detection module subsequently examines the text to identify potential lexical, morphological, syntactic, and mechanical errors. Detected errors are forwarded to the rule-based classification engine, which applies predefined IF–THEN linguistic rules to determine the specific error category, identify possible causes, and generate appropriate corrective explanations.

The system then presents detailed feedback by highlighting the identified errors, explaining why they occurred, and providing correct examples to guide learners toward proper language use. Based on the detected patterns, the framework further generates targeted writing recommendations, such as reviewing grammar rules, practicing specific language structures, expanding vocabulary, or improving proofreading strategies. Finally, all outputs are consolidated into an individualized feedback report that summarizes learner performance, error distribution, and recommended interventions. Through this operational process, the Intelligent Writing Tutor serves as a supplementary instructional tool that provides immediate, transparent, and evidence-based feedback to support learners' writing development and assist teachers in delivering responsive ESL instruction.

I. Discussion

The findings demonstrate that Ilocano-speaking ESL learners continue to encounter persistent writing difficulties across morphological, lexical, syntactic, and mechanical domains, with morphological errors emerging as the most dominant category. This pattern is consistent with previous studies reporting that grammatical inflections, particularly tense marking and subject–verb agreement, remain among the most challenging aspects of English writing for second language learners. The recurrence of these errors further supports theories of interlanguage development, suggesting that learners gradually construct an evolving linguistic system that incorporates characteristics of both the first and target languages.

The analysis also highlighted the influence of first-language (L1) interference on second-language (L2) writing performance. The omission of auxiliary verbs, inconsistent tense marking, direct translation of expressions, and misuse of prepositions suggest that learners frequently relied on linguistic structures from their native language when composing in English. Such findings support theories of language transfer and error analysis, emphasizing that negative transfer remains a significant factor affecting ESL learners' written production.

Unlike commercial automated writing evaluation systems that primarily generate direct grammatical corrections, the proposed Intelligent Writing Tutor emphasizes explainable feedback grounded in second language acquisition theory. Rather than merely identifying an error, the framework explicitly explains why the error occurred, associates it with possible first-language interference or developmental language processes, and recommends appropriate learning strategies. This pedagogically oriented design supports learner reflection and complements teacher feedback rather than replacing instructional judgment.

Importantly, the identified error patterns provided the empirical basis for the development of the proposed Intelligent Writing Tutor. By operationalizing authentic learner corpus data through NLP-assisted error detection and rule-based reasoning, the framework transforms identified errors into individualized feedback, corrective explanations, and targeted writing recommendations. Rather than merely identifying mistakes, the system promotes learner awareness, self-editing, and reflective practice through transparent and interpretable feedback mechanisms.

The results underscore the value of integrating corpus-informed analysis, second language acquisition principles, and explainable educational technologies to support evidence-based writing instruction. The proposed framework demonstrates considerable potential as a supplementary instructional tool that assists teachers in delivering timely interventions while empowering ESL learners to improve their writing proficiency through personalized and actionable feedback.

J. Limitations of the Study

Several limitations should be acknowledged when interpreting the findings. First, the learner corpus was obtained

from English language majors enrolled at a single higher education institution, thereby limiting the generalizability of the findings to other educational contexts. Second, all participants shared Ilocano as their primary first language, and consequently the identified interference patterns may not represent learners from different linguistic backgrounds. Third, the corpus consisted exclusively of reflective journal writings; therefore, the framework has not yet been examined using other writing genres such as argumentative, narrative, or academic essays. Finally, although the proposed Intelligent Writing Tutor underwent expert validation and operational demonstration, the study did not include a controlled learner-based intervention to determine its effectiveness in improving writing performance over time. Future studies should validate the framework using larger and more diverse learner populations and evaluate its educational impact through experimental or longitudinal research designs.

IV. Conclusion and Recommendations

A. Conclusion

This study developed an Intelligent Writing Tutor that integrates corpus-informed error analysis, natural language processing, and rule-based reasoning to generate individualized and explainable feedback for English as a Second Language (ESL) learners. Analysis of authentic learner writings revealed that morphological errors, particularly verb tense misuse, were the most frequently occurring language difficulties, followed by mechanical, lexical, and syntactic errors. Qualitative analysis further demonstrated that many of these recurring errors reflected first-language interference and developmental characteristics commonly associated with second language acquisition.

Guided by these findings, the proposed framework transformed manually validated learner error patterns into interpretable computational rules capable of detecting writing errors and generating pedagogically meaningful explanations, corrective examples, and targeted learning recommendations. Rather than functioning solely as an automated grammar correction tool, the framework was designed to provide transparent feedback that supports learner reflection while complementing teacher-mediated instruction.

The expert evaluation indicated that the framework possesses strong pedagogical potential for supporting ESL writing instruction. Nevertheless, the present study focused primarily on framework development, corpus analysis, and expert validation. Consequently, the study does not claim that the proposed Intelligent Writing Tutor directly improves learner writing proficiency. Instead, the findings demonstrate that the framework provides a feasible and explainable approach for delivering individualized writing feedback grounded in authentic learner data and established second language acquisition theories.

B. Future Works

Future research should extend the proposed Intelligent Writing Tutor through the development of a fully functional web-based or mobile learning environment capable of providing real-time writing assistance to learners. Experimental and quasi-experimental studies involving learner participants should be conducted to determine the effectiveness of the framework in improving writing accuracy, revision behavior, learner autonomy, and overall writing proficiency.

The computational capabilities of the framework may likewise be enhanced through the integration of transformer-based natural language processing models and large language models (LLMs) capable of analyzing contextual, discourse-level, and stylistic writing issues beyond sentence-level grammatical errors. A hybrid architecture combining explainable rule-based reasoning with machine learning techniques may further improve detection accuracy while preserving transparency in automated feedback generation.

Future investigations should also evaluate the framework using larger and more diverse learner populations representing different first-language backgrounds, educational levels, and writing genres. Such validation would strengthen the generalizability of the proposed Intelligent Writing Tutor and provide additional evidence regarding its applicability across multilingual educational settings.

Future implementations may incorporate adaptive learning analytics, teacher dashboards, personalized learning recommendations, and longitudinal learner progress monitoring to support data-informed writing instruction and continuous learner development.

Acknowledgment

The author gratefully acknowledges the Kalinga State University for its institutional support throughout the conduct of this research. Artificial intelligence tools were used solely to assist with language refinement, organization of ideas, and manuscript formatting. All research design, data collection, analysis, interpretation of findings, and final decisions regarding the manuscript were performed and verified by the author.

References

[1] B. Bhela, “Native language interference in learning a second language: Exploratory case studies of native language interference with target language usage,” International Education Journal, vol. 1, no. 1, pp. 22–31, 1999.

[2] R. Ellis, “The Study of Second Language Acquisition”. Oxford, U.K.: Oxford Univ. Press, 1994.

[3] S. P. Corder, “The significance of learners’ errors,” IRAL – International Review of Applied Linguistics in Language Teaching, vol. 5, nos. 1–4, pp. 161–170, 1967.

[4] S. P. Corder, “Idiosyncratic dialects and error analysis,” IRAL – International Review of Applied Linguistics in Language Teaching, vol. 9, pp. 147–160, 1971.

[5] H. Dulay and M. Burt, “Errors and strategies in child second language acquisition,” TESOL Quarterly, vol. 8, no. 2, pp. 129–136, 1974.

[6] E. Fauziati, “Native and target language influence on the students’ interlanguage production: A case of Indonesian EFL compositions,” Indonesian Journal of Applied Linguistics, vol. 7, no. 1, pp. 54–63, 2017.

[7] C. Chen, “Teaching of interlanguage theory and Chinese as a second language,” Literature Education, no. 2, pp. 174–175, 2021.

[8] S. D. Krashen, “Second language acquisition,” Second Language Learning, vol. 3, no. 7, pp. 19–39, 1981.

[9] K. Karim and H. Nassaji, “First language transfer in second language writing: An examination of current research,” Iranian Journal of Language Teaching Research, vol. 1, no. 1, pp. 117–134, 2013.

[10] N. Biglari and E. Struys, "Native language interference in English L2 word recognition and word integration skills," Theory and Practice in Language Studies, vol. 11, no. 1, p. 1, 2021.

[11] G. F. Dipolog-Ubanan, "L1 influence on writing in L2 among UCSI Chinese students: A case study," Pertanika Journal of Social Sciences and Humanities, vol. 24, no. 4, 2016.

[12] M. V. Govindarajoo, C. C. Hui, and S. F. A. Aziz, "Common errors made in English writing by Malaysian Chinese primary year 6 ESL learners at a tuition centre in Puchong, Malaysia," Asian Journal of University Education, vol. 18, no. 3, 2022.

[13] R. A. Budiharto, "Native language interference on target language writings of Indonesian EFL students: An exploratory case study," Indonesian EFL Journal, vol. 5, no. 1, pp. 107–116, 2019.

[14] E. Mirzayev, "The influence of first language interference on ESL writing skills," *EuroGlobal Journal of Linguistics and Language Education*, vol. 1, no. 1, pp. 33–39, 2024.

[15] P. R. N. Phuket and N. B. Othman, "Understanding EFL students' errors in writing," Journal of Education and Practice, vol. 6, no. 32, pp. 99–106, 2015.

[16] K. Sermsook, J. Liamnimitr, and R. Pochakorn, "An analysis of errors in written English sentences: A case study of Thai EFL students," English Language Teaching, vol. 10, no. 3, p. 101, 2017.

[17] J. Jiang and Q. Jiang, "Study on the relationship between error and dependency distance in the writing of Chinese English learners," Language Application, no. 1, pp. 42–52, 2021.

[18] M. Jiang and W. Wu, "Transfer of implicit concepts of foreign language learners in foreign language sentence comprehension," *Foreign Language Teaching*, no. 3, pp. 68–73, 2017.

[19] J. Chen, "Visual analysis of bilingual acquisition studies based on CiteSpace (2010–2020)," Overseas English, no. 24, pp. 15–16, 27, 2020.

[20] H. M. Hu and B. Wu, "Research on second language acquisition based on the analysis of core journals in foreign languages from 2011–2020," *Overseas English*, no. 8, pp. 4–6, 2022.

[21] J. P. Lantolf and O. Esteve, "Concept-based instruction for concept-based instruction: A model for language teacher education," in Evidence-Based Second Language Pedagogy. London, U.K.: Routledge, 2019, pp. 27–51.

[22] The Douglas Fir Group, "A transdisciplinary framework for SLA in a multilingual world," The Modern Language Journal, vol. 100, suppl. S1, pp. 19–47, 2016.

[23] J. Z. Maitama, N. Idris, and A. Zakari, "A systematic mapping study of the empirical explicit aspect extractions in sentiment analysis," *IEEE Access*, vol. 8, pp. 113878–113899, 2020.

[24] N. Liu and J. Zhao, "A BERT-based aspect-level sentiment analysis algorithm for cross-domain text," Computational Intelligence and Neuroscience, vol. 2022, Art. no. 8726621, 2022.

[25] B. Kinge, S. Mandhare, P. Chavan, and S. M. Chaware, "Resume screening using machine learning and NLP: A proposed system," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 8, no. 2, pp. 253–258, 2022.

[26] S. Bharadwaj, S. V. Rudra, S. A. Potukuchi, N. Macherla, and G. C. Babu, "Resume screening using NLP and LSTM," in *Proc. 2022 Int. Conf. on Inventive Computation Technologies (ICICT)*, pp. 238–241.

[27] P. T. Kim, "Data-driven discrimination at work," William & Mary Law Review, vol. 58, p. 857, 2016.

[28] Z. Chen, "Ethics and discrimination in artificial intelligence-enabled recruitment practices," Humanities and Social Sciences Communications, vol. 10, no. 1, pp. 1–12, 2023.

[29] E. Ferrara, "Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies," Sci, vol. 6, no. 1, p. 3, 2024.

[30] A. Fabris et al., "Fairness and bias in algorithmic hiring: A multidisciplinary survey," ACM Transactions on Intelligent Systems and Technology, vol. 16, no. 1, pp. 1–54, 2025.

[31] C. J. Carl, Errors in Language Learning and Use: Exploring Error Analysis. Beijing, China: Foreign Language Teaching and Research Press, 2001.

[32] J. C. Richards, "A non-contrastive approach to error analysis," English Language Teaching Journal, vol. 25, no. 3, pp. 204–219, 1971.

[33] J. Cai and J. Li, "A multidimensional dynamic theory framework for language migration," Foreign Language Teaching, no. 4, pp. 43–50, 2016.

[34] J. Cai and M. Wang, "Research on two language acquisition based on use orientation: Theory, empirical evidence and prospect," Foreign Language and Foreign Language Teaching, no. 2, pp. 1–15, 2020.

[35] B. Al-Sowaidi, "Error analysis in translation pedagogy," International Journal on Humanities and Social Sciences, no. 37, pp. 174–190, Sep. 2022, doi: 10.33193/ijohss.37.2022.461. Available: https://doi.org/10.33193/ijohss.37.2022.461

Migration Method for Parametric Features in Heterogeneous CAD Systems

Gang WANG, Guangfa ZHANG
College of Mechanical Engineering, Shanghai Dianji University, Shanghai, China

***Abstract*—Parametric feature migration across different Computer-Aided Design (CAD) systems is a core technology that supports cross-platform collaborative design and product data reuse. Addressing the fundamental limitation of traditional geometry-based data exchange methods—their inability to transfer high-level design intent information such as construction history, parametric constraints, and feature dependency relationships—this study proposes a three-stage parametric feature migration method based on JavaScript Object Notation (JSON) neutral files. Following the methodology of "exchanging the modeling process rather than the model itself", the complete migration workflow is decomposed into three stages: Feature parameter extraction, which leverages the source CAD system's application programming interface to traverse the model feature tree in depth and serialize the modeling history into a structured JSON exchange file; Topology identification, which employs a multi-dimensional geometric descriptor-based topological element identification method and writes identification results into the JavaScript Object Notation file; and Data validation, which validates feature parameters to be transferred and retrieves transient data required for modeling to ultimately achieve model reconstruction. Taking Siemens NX and ZW3D (formerly VX CAD) as representative source and target systems, a complete feature extraction and regeneration prototype plugin was developed, verifying the feasibility of the proposed method between heterogeneous CAD systems.**

Keywords*—Heterogeneous CAD; feature migration; JSON neutral file; information extraction; topology identification

I. Introduction

In the context of globalization and deep integration of digital technologies in manufacturing, product collaborative design has become a critical means for enterprises to enhance their core competitiveness. Hoffmann et al. [1] pointed out in their foundational work on user-defined features that parametric features are not merely collections of geometric elements, but rather encapsulations of design knowledge—feature dependency relationships, constraint networks, and modeling sequences collectively constitute the design intent of a product. In collaborative design practice, different enterprises and departments select different computer-aided design (CAD) systems for product development based on functional requirements, operational habits, and economic considerations, leading to heterogeneity in model data formats and feature semantic systems [2]. Concurrently, with the widespread adoption of Model-Based Definition (MBD) technology in high-end manufacturing sectors such as aerospace and aviation, product 3D models have evolved from purely geometric carriers into comprehensive data containers that embody design, manufacturing, and inspection information [3]. This places higher demands on information exchange between heterogeneous CAD systems—requiring not only the transfer of geometric shapes but also the complete preservation of construction history, parametric constraints, feature associations, and engineering annotations as high-level design intent. Therefore, achieving reliable migration of parametric feature models between heterogeneous CAD systems has become a critical technology for building cross-platform product collaborative development frameworks.

Existing data exchange methods between heterogeneous CAD systems primarily fall into two categories: geometry-based exchange and feature-based exchange [4, 5]. The first category employs the Initial Graphics Exchange Specification (IGES), Standard for the Exchange of Product Model data (STEP), and other neutral file formats as bridges, realizing cross-system geometric data transfer through the exchange of Boundary Representation (B-Rep) models. Rappoport [6] proposed an early architecture for universal CAD data exchange, establishing a theoretical framework for subsequent research. The STEP AP242 application protocol can express certain feature semantic information, but mainstream commercial CAD systems have only implemented a limited functional subset of the standard; in practical engineering scenarios, high-level information such as modeling history, parametric constraints, and feature associations cannot be effectively transferred [7]. After conversion, the model essentially degrades into a "static geometric entity", losing its parametric editing capability.

To overcome these deficiencies, academia has proposed feature model-based data exchange methods. Choi et al. [8] were among the first to explore CAD part model exchange based on the macro-parametric approach. Mun et al. [9] subsequently proposed a set of standard neutral modeling commands, achieving procedural data exchange between heterogeneous CAD systems. Li et al. [10] further proposed a two-stage modeling procedure recovery method—whose core concept is to exchange the complete modeling operation process rather than the model itself, "simulating a human designer" in the target system to reconstruct parametric feature models step by step. This method decomposes the feature migration workflow into first-order information extraction and second-order information reconstruction stages: the first-order information extraction stage recovers explicit modeling operation sequences from the source CAD system's macro recording file; the second-order information reconstruction stage reconstructs interactive selection information (such as reference elements, datum elements, and direction elements of features) and implicit information automatically generated by the CAD system (such as constraint relationships, implicit parameters, and feature

geometric attributes) from operation sequences, thereby supplementing the key semantic information missing from macro files. This two-stage framework has been validated between typical CAD systems such as SolidWorks, UG, Pro/E, and CATIA. Zhang et al. [11] further investigated interoperability issues during feature data exchange from a quantitative optimization perspective, proposing a computational model for evaluating conversion quality. Shao et al. [12] systematically established a feature operation-centric model conversion interface scheme, defining feature operation primitives independent of specific CAD systems and accessing and reconstructing CAD model information at both the feature layer and geometric layer simultaneously, ensuring consistency of feature trees, dimensions, and constraint information before and after conversion. Zheng et al. [13] built upon Shao's work to propose four migration principles for singular features and the (F, T, P) topological element identification method, providing important methodological guidance for the complete preservation of design intent between heterogeneous CAD systems. At the semantic level of feature migration, Qin et al. [14] constructed an ontology-based feature mapping framework, while Tessier and Wang [15] proposed a feature mapping and verification method between CAD systems from an ontological perspective; these two contributions established a theoretical foundation for semantic interoperability of heterogeneous CAD models. In the domain of engineering annotation exchange, Camba et al. [16] proposed an extended 3D annotation mechanism to explicitly convey geometric design intent, offering new possibilities for the preservation of non-geometric information during feature migration.

The above research has laid an important theoretical foundation for heterogeneous CAD feature migration, but the following common limitations exist: in the process of heterogeneous CAD feature migration, to maintain the transfer of parameters, feature information, and assembly relationships while avoiding data loss, the majority of transfer content is stored and compiled in Extensible Markup Language (XML) file format. Sun [17] also adopted XML as the intermediate file format in his research on feature-based data exchange between heterogeneous CAD systems, further corroborating the prevalence of this technical approach. However, XML, while being a markup language with strong semantics, strict validation, and extensibility suitable for complex documents, metadata, configuration files, and scenarios requiring rigorous data validity, exhibits high redundancy and slow parsing when handling simple models composed of multiple small components.

To improve the compatibility and generality of neutral files in heterogeneous CAD data transfer, and to facilitate efficient conversion of simple model feature data, this study employs JavaScript Object Notation (JSON) neutral files as the intermediate data storage container for research on feature-based heterogeneous CAD model conversion technology. The proposed method follows a three-stage experimental framework of "feature parameter extraction + topology identification + data validation".

The overall technical approach is illustrated in Fig. 1. This framework divides the process of heterogeneous CAD parametric feature migration into three sequentially connected stages: the first stage is feature parameter extraction, which accesses the source CAD system's internal feature tree through its Application Programming Interface (API), serializing feature parameters and modeling history into a structured JSON exchange file; the second stage is topology identification, which constructs a multi-dimensional geometric descriptor-based topological element identification method and writes results into the JSON file; the third stage is data validation, which takes into account the fact that operating methods and kernels may differ between distinct CAD systems in order to maximally preserve the model's design intent during CAD system migration and ensure successful regeneration of feature models, thus verifying that feature parameter data can be accurately imported. Taking NX and ZW3D as typical cases, a corresponding prototype system was developed to verify the feasibility and effectiveness of the method.

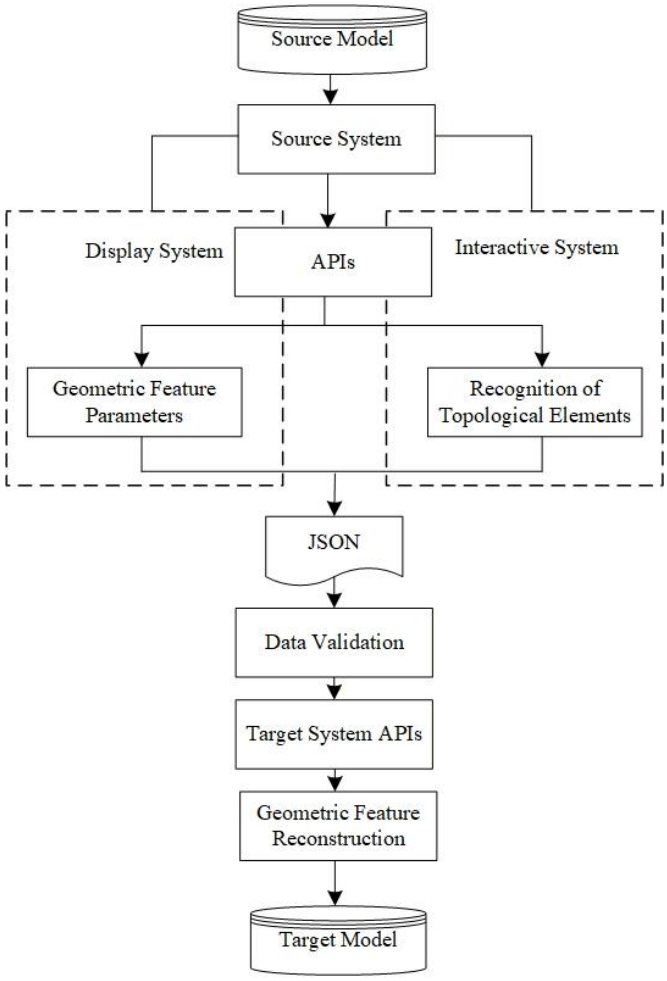


Fig. 1. Overall framework of heterogeneous model conversion.

II. Three-Stage Method for Heterogeneous CAD Parametric Feature Migration

As shown in Fig. 2, for the entire process of creating a source model, each operation step performed by the user generates a corresponding feature model based on the user's input command parameters, including constraints, implicit parameters, feature geometry, and physical attributes. The resulting feature model can be observed by the user in the display system, and the

driving feature parameter data is stored in the corresponding feature modules within the system. In addition to the model's feature parameters, there exist interactive information generated during the user's creation process, typically referring to the feature reference elements, base elements, and direction elements at the time of feature creation. This information is crucial for model reconstruction and serves to preserve design intent, particularly the correspondence of topological structures before and after model modifications.

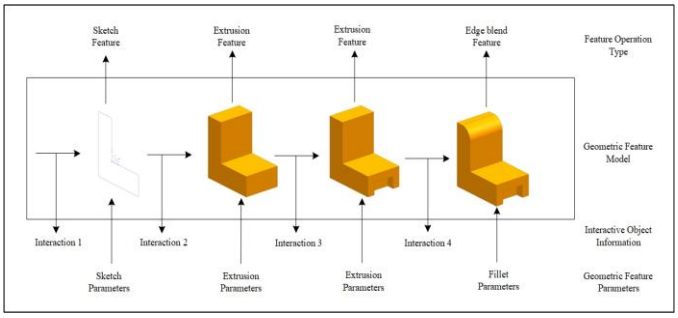


Fig. 2. Information generated during the creation of a parametric feature model.

A. Geometric Feature Parameter Extraction and Storage

The geometric feature parameter extraction stage requires the complete and accurate acquisition of all explicit modeling information of the parametric feature model from the source CAD system, organized in a structured and self-describing format to provide a data foundation for subsequent topology identification and target CAD feature model reconstruction.

Feature-based CAD systems accessed directly through their APIs generate a corresponding feature tree concurrently with user-interactive modeling, within which geometric feature information is recorded in internal data structures [12]. The traversal of the feature tree uses depth-first traversal (DFT), which was chosen over breadth-first traversal (BFT) for a deliberate reason: in parametric CAD modeling, feature creation follows a strict temporal sequence in which each feature depends on the geometric state produced by previously created features. DFT naturally reproduces this chronological ordering because it follows each branch of the feature tree to its deepest leaf before backtracking, thereby preserving the parent-to-child dependency chain. BFT, by contrast, would group features by their depth level and disrupt the temporal dependency order, potentially causing reconstruction failures in the target system when a child feature is reconstructed before its prerequisite parent feature has been fully resolved.

In the domain of heterogeneous CAD data exchange, the selection of a neutral file format directly affects the completeness of information, expressive capability, and engineering practicality. The fundamental deficiency of traditional STEP and IGES standards as neutral files for geometry-level exchange lies in the fact that their exchange target is the B-Rep of the parametric feature model rather than the modeling process [4]. In terms of information storage format selection, this method adopts JSON as the neutral exchange file format.

Compared with XML, JSON offers four principal advantages for storing CAD feature data:

1) Stronger adaptability: Through key-value pairs and nested arrays matching the CAD feature tree hierarchy, it can flexibly accommodate heterogeneous parameters of various features without strict format constraints.
2) Lower system resource consumption: It can intuitively express entity association relationships; one-to-many array relationships can be directly displayed through arrays, and topological elements carry their own tags and geometric data, eliminating the need for system reverse queries.
3) Higher parsing efficiency: As a native data structure of programming languages, it can be directly deserialized into C++ in-memory objects through lightweight libraries such as nlohmann/json, significantly reducing parsing overhead.
4) Better cross-platform generality: With concise syntax and a compact type system (only six data types: object, array, string, number, Boolean, and null), it can adapt to various CAD systems with API capabilities, effectively reducing the data interfacing costs between heterogeneous systems.

The JSON storage structure is illustrated in Fig. 3. Using the sketch feature as an example: this structure is capable of storing the basic feature parameters, sketch member parameters, constraint member parameters, and work coordinate system parameters of a sketch in a CAD system. By indexing the feature tree sequence, it reflects the temporal order of CAD feature creation, the structure, and the relationships between features created at each structural level.

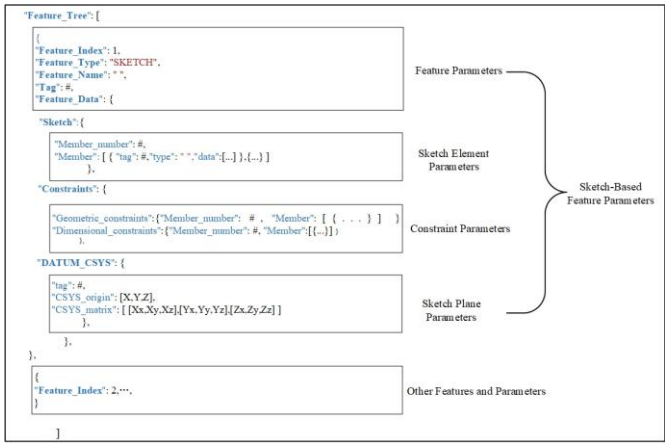


Fig. 3. JSON storage structure of feature data.

B. Basic Topological Element Identification

Basic topological elements refer to vertices, edges, and faces in a model. In the heterogeneous CAD system feature migration process, the creation of sketch-based features, dress-up features, and transformation features all depend on the precise positioning of topological elements.

When users create features in a CAD system, interaction with topological elements of the target geometric body is unavoidable.

For example, an extrude operation requires a closed profile or face within a sketch to drive it, as shown in Fig. 4(a). The termination condition "extrude to face" requires referencing a target termination face, as shown in Fig. 4(c). Chamfer feature creation requires selecting a target edge, as shown in Fig. 4(b). Draft features require specifying the neutral face and faces to be drafted, mirror features require a mirror plane, and so on. Therefore, the topology identification stage is a core component of heterogeneous CAD system data exchange: to ensure topological relationship stability and that the target model possesses the same design intent as the source model, geometric elements must be identified.

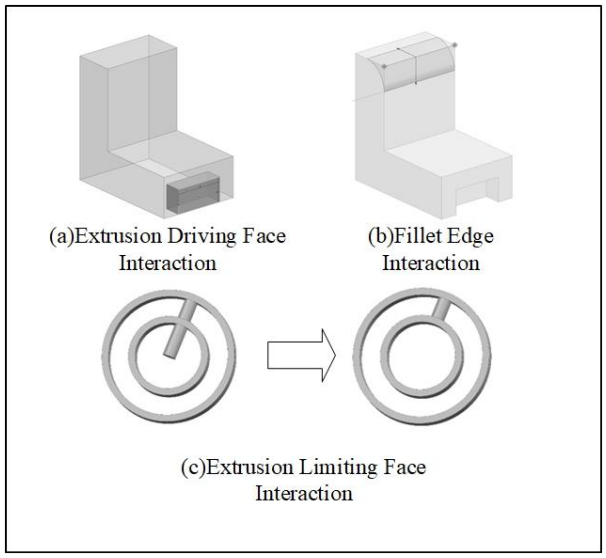
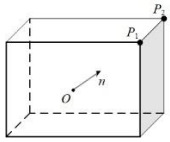
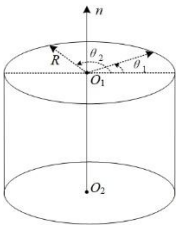


Fig. 4. User-interactive object selection.

To precisely identify referenced topological elements, the parametric definitions of such elements must be clearly established. Referenced geometric elements are classified into two categories: edges and faces. Edges include straight lines, circles, circular arcs, etc.; faces include planar faces, cylindrical faces, conical faces, spherical faces, toroidal faces, etc. Ensuring accurate and error-free referencing of topological elements is critical for the subsequent geometric feature reconstruction process.

As shown in Table I, the format for obtaining various parameters of referenced topological elements from a rectangular block and a cylinder is provided. The approach is fundamentally similar: select the target topological element type, invoke the corresponding API functions to obtain its individual parameters. For straight edges, the primary parameters are the 3D coordinates of the start point and end point. For circular and arc edges, the primary parameters are the radius, center point, and start/end angles; it is worth noting that all arcs in the UGNX system rotate in the counterclockwise direction. For faces, the primary parameters are the center point and normal vector. Because face types differ, the reference for center point and normal also differs: for example, the center point of a planar face is located at the 3D coordinate center of the midpoint between its tight bounding box minimum and maximum corner points, and its normal is perpendicular to the face. The center point of a cylindrical face is located at the center of its bottom circular edge, and its normal is aligned with its axial vector.

TABLE I. PARAMETRIC DEFINITION OF TOPOLOGICAL GEOMETRY AND ITS JSON REPRESENTATION.

Geometry	Parameter	JSON Representation
	Edge type	"edge_type": "Linear",
	Parent feature handle	"parent_feat_tag": #,
	Start point	"start_point": [X, Y, Z],
	End point	"end_point": [X, Y, Z]
	Face type	"face_type": "Plane",
	Center point	"center_point": [X, Y, Z],
	Normal	"normal": [X, Y, Z],
	Box min	"box_min": [X, Y, Z],
	Box max	"box_max": [X, Y, Z],
	Parent feature handle	"parent_feat_tag": #
	Edge type	"edge_type": "Arc",
	Parent feature handle	"parent_feat_tag": #,
	Radius	"radius": #,
	Arc center	"arc_center": [X, Y, Z],
	Start angle	"start_angle": #,
	End angle	"end_angle": #,
	Sweep angle	"sweep_angle": #
	Face type	"face_type": "Cylinder",
	Center point	"center_point": [X, Y, Z],
	Normal	"normal": [X, Y, Z],
	Box min	"box_min": [X, Y, Z],
	Box max	"box_max": [X, Y, Z],
	Radius	"radius": #,
	Parent feature handle	"parent_feat_tag": #

The topology identification workflow is shown in Fig. 5.

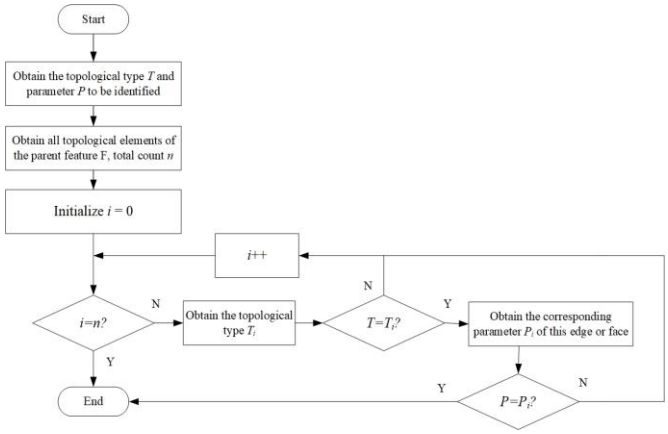


Fig. 5. Topology element identification process.

This approach references the (*F, T, P*) triple identification method proposed by Xie et al. [18]—where the parent feature *F* is used to narrow the search scope, the topology type *T* is used for filtering, and geometric parameters *P* are used to complete identification and matching. During the feature reconstruction process, the parameters of topological elements to be identified are obtained from the JSON file; using the feature name *F* from the parameters, all edges or faces in that feature are selected; a determination is made as to whether the selected edge or face is of the required type; if so, the geometric parameters P_i of the

selected topological elements are sequentially compared with the corresponding parameters in *P* to determine whether they are equal; if they are equal, the topological element is identified as the target topological element. When the model is relatively complex, selecting edges or faces dependent on features can effectively narrow the search scope and improve the search efficiency for topological elements.

C. Data Validation

Because the operating methods and kernels of different CAD systems may differ, it is necessary to extract and verify the feature data before regeneration to maximally preserve the model's design intent during CAD system migration and ensure the successful reconstruction of feature models. Prior to this, essential data validation and supplementation of missing information must be performed.

1) Tracing disappearing transient information. During feature operations in a CAD system, certain topological structures may undergo deformation or even disappear as a result of the operation. These disappearing transient topologies are often an indispensable part of maintaining operational intent. For transient information that disappears after modeling, the fillet feature is used as an example: when the user performs a fillet operation and selects a target edge, that edge becomes a cylindrical face, and the target edge is suppressed by the system. By obtaining the edge's handle, temporarily suppressing its feature, and querying its parent feature, the original parameter information of the edge can be retrieved. This method also applies to draft features and other features that alter geometric shape parameters.

A related but more challenging scenario involves features that are completely consumed by subsequent Boolean operations (union, difference, or intersection). For example, a cylindrical extrude feature may be entirely absorbed into a larger body through a Boolean union, leaving no standalone geometric entity that can be traversed in the final model. To handle this, the feature parameter extraction stage captures all feature information before the Boolean operation is applied. Specifically, the traversal algorithm records a snapshot of each feature's geometric parameters and topological references at the moment of its creation within the feature tree, rather than from the final consolidated model. This is achieved by iterating through the feature tree's creation history in temporal order: for each feature, its parameters are extracted immediately after the feature node is visited, regardless of whether the feature remains visible in the final model. Boolean operations themselves are recorded as separate feature entries in the JSON file with a tag and an array listing the consumed features' handles. During reconstruction, the target system first reconstructs all constituent features and then applies the Boolean operation, faithfully reproducing the original modeling sequence.

2) Validating cross-system incompatibility information. For incompatible information arising from differences in feature construction methods between heterogeneous CAD systems, alternative solutions must be adopted when unique feature constraint types in the source CAD system have no direct correspondence in the target system. One situation occurs when the source CAD system possesses a certain design intent during feature interaction operations that the target CAD system does not possess. For example, the extrude feature in ZW3D includes a design intent of "extrude to a point", whereas the UGNX system lacks this design intent. According to the third principle proposed by Zheng et al. [13]—"converting using different features and auxiliary formulas"—an auxiliary plane is created at the target point, and by leveraging extrusion-to-plane along with its extended surface features and constructing an auxiliary plane, the conversion of extrusion-to-vertex features can be achieved. This ensures that when the vertex position changes, the constructed auxiliary plane changes accordingly, thereby causing the extrusion length to adaptively vary and thus preserving the design intent.

Another situation occurs when both the source and target CAD systems contain the same construction method, but due to differences in their underlying kernels, data extracted from the source CAD system cannot be directly used between the two systems and requires certain processing. In practice, to maintain the stability and consistency of internally similar data, CAD systems establish prescribed operational procedures. For example, in the UGNX system, when creating an arc in a sketch feature, to ensure the arc's rotation direction remains consistent, the system relies on the arc orientation matrix and uses the absolute coordinate system origin (0, 0, 0) as the reference for orienting the arc, thereby ensuring the arc's rotation direction is counterclockwise. If APIs are called to obtain the coordinates of the arc center, these coordinates correspond to the arc center in the local coordinate system of the arc. Generally, when creating sketch curves, the input sketch is in the local coordinate system or the absolute coordinate system. Therefore, data conversion is required: the arc center coordinates in the local coordinate system of the arc must be transformed to the arc center coordinates in the sketch coordinate system.

Using API functions, the arc center coordinates and arc rotation matrix based on the absolute coordinate system can be obtained. The coordinate transformation is accomplished through the following transformation equations:

Let P_{local} be the arc center in the local coordinate system of the arc, and let R_{acr} be the corresponding arc rotation matrix. The arc center P_{abs} in the absolute coordinate system is obtained from:

$$P_{abs} = R_{acr} P_{local} \tag{1}$$

The sketch coordinate system is described by its origin O_{sk} and rotation matrix R_{sk} (where R_{sk} transforms sketch local coordinates to absolute coordinates, i.e.,$v_{abs} = R_{sk} v_{sk} + O_{sk}$). Hence, the sketch local coordinate P_{sk} is computed as:

$$P_{sk} = R_{sk}^{T}(P_{abs} - O_{sk}) \tag{2}$$

where, R_{sk}^{T} is the transpose of R_{sk} and, due to the orthogonality of a rotation matrix, $R_{sk}^{T} = R_{sk}^{-1}$. In practical implementation, taking the three column vectors C_0, C_1, C_2 of R_{sk}, the dot product operation is employed:

$$P_{sk,i} = C_i(P_{abs} - O_{sk}), \quad i = 0,1,2 \tag{3}$$

Thus, the parameter information that can be directly utilized by the target CAD system in sketch features is obtained.

3) Complex Sketch Constraint Serialization. The complex sketch geometric constraints addressed herein are primarily constraint outcomes generated by patterned curves (mirrored curves represent a special subtype of patterned curves).

Sketch constraints fall into two categories: dimensional constraints and geometric constraints. Whether they are geometric constraints such as perpendicularity or parallelism, or dimensional constraints, all serve as mandatory restrictions imposed on pre-existing curves.

Taking UG NX as an example, the constraint relationships of existing curves within a sketch feature can be retrieved by invoking two Application Programming Interface (API) functions: “ UF_SKET_ask_geo_cons_of_sketch() ” and “UF_SKET_ask_dimensions_of_sketch()”.

Child curves generated via the patterned curve operation maintain a patterning association with one another and carry identical geometric information. Counterintuitively, when the information of these curves is extracted and reconstructed in the target CAD system, the curves are found to be completely coincident in appearance, rather than arranged sequentially in accordance with the patterning order.

To address this issue, the following strategy is proposed:

- Aggregate all curves generated by the corresponding patterned curve operation using the aforementioned geometric constraint acquisition function.
- Identify the parent curve through the "smart container" APIs in UG Open, and label the remaining curves as child curves.
- Document the patterning mode, the parent curve and the labeled child curves in the JavaScript Object Notation (JSON) file, and mark the child curves as disabled during reconstruction in the target CAD system.
- Execute the patterned curve operation on the parent curve in the target CAD system.

This approach can maximally ensure the successful conversion of patterning constraints.

III. Experimental Validation and Analysis

A. Experimental Case Study

To verify the feasibility and effectiveness of the proposed method, this study selects Siemens NX 12 as the source system and ZW3D 2024 as the target system, and establishes a complete feature migration prototype. First, the NX Open C++ API is used to traverse the model feature tree in depth, recursively obtaining the construction type, parameter set, sketch data and constraint information, topological face references, and geometric attributes of each feature. All information is organized in a structured hierarchical manner and output as a JSON exchange file. Fig. 6 shows a code excerpt for parameter and topology extraction from a UGNX extrude feature.

Next, JSON parsing and system reconstruction are implemented using ZW3D C++ API. Fig. 7 shows a code excerpt for reconstructing an extrude feature in ZW3D.

```
void Get_Extrude_Info(tag_t extrude_feat, nlohmann::ordered_json& j_Info)
{
    TaggedObject* taggedObj = NXObjectManager::Get(extrude_feat);
    Features::Extrude* extrude = dynamic_cast<Features::Extrude*>(taggedObj);

    Session* ses = Session::GetSession();
    Part* workPart = ses->Parts()->Work();
    Features::ExtrudeBuilder* builder = workPart->Features()->CreateExtrudeBuilder(extrude);

    Direction* dirObj = builder->Direction();
    Vector3d vec = dirObj->Vector();
    j_Info["direction"] = { vec.X, vec.Y, vec.Z };

   auto* offset = builder->Offset();
   PrintFeatureOffset(offset, &j_Info);  //Function to obtain extrude offset information

    auto* startExt = builder->Limits()->StartExtend();
    auto* endExt = builder->Limits()->EndExtend();

    auto printExtend = [](const char* which, NXOpen::GeometricUtilities::Extend* ext,
        nlohmann::ordered_json& j){ ... }

        printExtend("Start", startExt, j_Info["limits"]["start"]);
        printExtend("End", endExt, j_Info["limits"]["end"]);

    builder->Destroy();
}
```

Fig. 6. Code excerpt for parameter and topology extraction from UGNX extrude feature.

```
int ZWClass::ProcessExtrude(const json& featData, int featureTag)
{
...

    extrude.profileHandle = sketchHandle;// 5. Populate data
    extrude.startS = startDist;extrude.endE = endDist;
    extrude.startType = startType;
    extrude.endType = endType;
    extrude.boolean.combine = combine;
    extrude.settings.capType = ZW_END_CAP_BOTH;
    extrude.draft.isUseExtrusionDirection = 1;
    extrude.direction.x = dirX;
    extrude.direction.y = dirY;
    extrude.direction.z = dirZ;

    if (startFaceHandle.innerData){
        extrude.startEntityHandle = startFaceHandle;
        extrude.startToPoint = &startPoint;
    }
    if (endFaceHandle.innerData){
        extrude.endEntityHandle = endFaceHandle;
        extrude.endToPoint = &endPoint;
    }

    szwEntityHandle shapeHandle = { 0 };
    evxErrors err = ZwFeatureExtrudeCreate(extrude, &shapeHandle);// 6.Execute
...
    ZwEntityHandleFree(&shapeHandle);
    ZwEntityHandleFree(&sketchHandle);
    Set_View();
    return extrudeFeatureId;
}
```

Fig. 7. Code excerpt for reconstructing an extrude feature in ZW3D.

B. Results Analysis and Discussion

The experimental results are shown in Fig. 8. The results indicate that the proposed method successfully achieved cross-system migration of parametric features from NX to ZW3D. By comparing the feature trees and feature models of the source and target CAD systems, it can be observed that the target CAD system is able to find feature trees corresponding to those in the source system, and the feature model displays are similar to the source models. This validates the effectiveness of the three stages—feature parameter extraction, topology identification, and data validation. It also demonstrates that JSON, as a neutral file format, is capable of storing feature parameter information and maintaining the potential for interactive design. The method effectively preserves high-level semantics such as design intent, parameter information, and feature information, enabling the migration and sharing of heterogeneous CAD models with significant practical potential and application value.

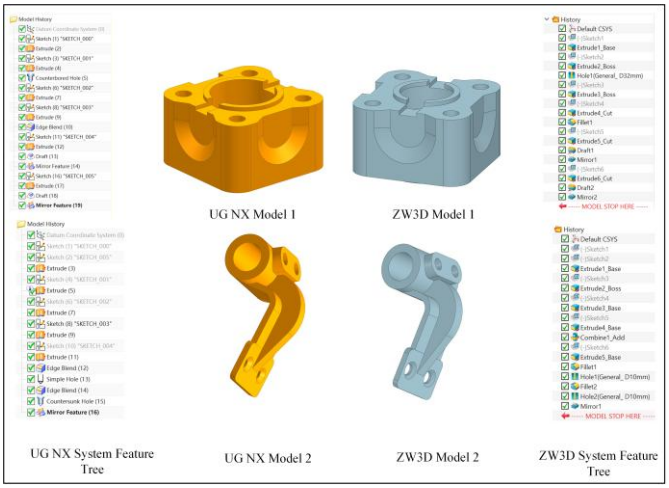


Fig. 8. Model migration cases between UGNX and ZW3D.

IV. Discussion

A. Parametric Editability Verification

A key criterion for successful feature migration is whether the migrated model remains fully parametric and editable in the target CAD system. To evaluate this, parameter modification tests were conducted on the migrated models. For each feature in the target model, its critical dimension parameters were modified and the model was regenerated. The test results showed that for features reconstructed at the feature-level (those with direct API mapping), all parameter modifications triggered correct model updates, preserving the intended update behavior—the model geometry changed exactly as it would in the source system when the same parameter was modified.

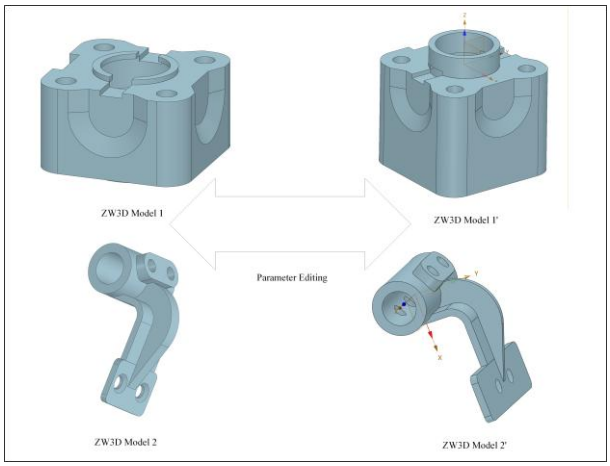


Fig. 9. Editable target CAD model.

B. Generalizability Across CAD Kernels

The proposed method's generalizability to additional CAD systems with different geometric kernels relies on the API abstraction layer implemented in the extraction and reconstruction modules. The method has been designed with kernel independence as a core architectural principle. The feature parameter extraction module communicates with the source CAD system solely through its published API, without depending on kernel-specific data structures. Similarly, the reconstruction module invokes only the target system's documented API functions.

C. API Version Compatibility

The software used for development is generally lower-version releases, so API invalidation in CAD systems is extremely rare. Should such a scenario arise, alternative APIs of the same category may be invoked in adherence to the principle of ensuring successful conversion.

For instance, two primary API function systems are available in UG NX, namely UFun and NX Open. Both sets of APIs shall be invoked simultaneously during the implementation of the aforementioned patterning constraint guarantee strategy.

V. Conclusion

Addressing the common challenge of feature migration between heterogeneous CAD systems, this study proposes a tiered regeneration feature migration method based on JSON neutral files, founded on a three-stage framework of feature parameter extraction, topology identification, and data validation. In the feature parameter extraction stage, modeling history is serialized into JSON structured data, improving the completeness of information extraction and the generality of the method. In the topology identification stage, a multi-dimensional geometric descriptor-based topological element identification method is constructed and JSON data is populated and written. In the data validation stage, to maximally preserve the model's design intent during CAD system migration, system transient information is traced and retrieved, and cross-system incompatibility information is refined and converted.

Taking NX and ZW3D as typical cases, prototype development and verification were completed. Experimental results demonstrate that models migrated using the proposed method are capable of inheriting the feature information and modeling history of the source CAD system, preserving the model's design intent.

Future research on this method will advance along two directions: first, tracking the iterative expansion of target CAD system APIs to broaden the coverage of mappable features; second, addressing the limited expressive capability of JSON for complex curves or features—further investigation into more flexible data organization approaches is needed to handle mixed hierarchical relationships that the tree structure of JSON may struggle to represent.

References

[1] M. Hoffmann and R. Joan-Arinyo, "on User-defined Features," Computer-Aided Design, Vol. 30, No. 5, Pp. 321–332, 1998.

[2] M. Gao and F. Z. He, "Survey of Distributed and Collaborative Design," Journal of Computer-Aided Design & Computer Graphics, Vol. 16, No. 2, Pp. 149–157, 2004.

[3] Alemanni, F. Destefanis, and E. Vezzetti, "Model-based Definition Design in the Product Lifecycle Management Scenario," the International Journal of Advanced Manufacturing Technology, Vol. 52, No. 1/2/3/4, Pp. 1–14, 2011.

[4] J. Pratt, B. D. Anderson, and T. Ranger, "Towards the Standardized Exchange of Parameterized Feature-based CAD Models," Computer-Aided Design, Vol. 37, No. 12, Pp. 1251–1265, 2005.

[5] Kim, M. J. Pratt, R. G. Iyer, and R. D. Sriram, "Standardized Data Exchange of CAD Models with Design Intent," Computer-Aided Design, Vol. 40, No. 7, Pp. 760–777, 2008.

[6] Rappoport, "an Architecture for Universal CAD Data Exchange," in Proceedings of the Eighth ACM Symposium on Solid Modeling and Applications. Seattle, WA: ACM, 2003, Pp. 266–269.

[7] Z. Zhang and X. F. Luo, "Design Intent Information Exchange of Feature-based CAD Models," in Proceedings of the World Congress on Computer Science and Information Engineering. Los Angeles, CA: IEEE, 2009, Pp. 11–15.

[8] H. Choi, D. Mun, and S. Han, "Exchange of CAD Part Models Based on the Macro-parametric Approach," International Journal of CAD/CAM, Vol. 2, No. 1, Pp. 13–21, 2002.

[9] Mun, S. Han, J. Kim, and Y. Oh, "a Set of Standard Modeling Commands for the History-based Parametric Approach," Computer-Aided Design, Vol. 35, No. 13, Pp. 1171–1179, 2003.

[10] Li, F. Z. He, X. T. Cai, and D. Zhang, "Using Procedure Recovery Approach to Exchange Feature-based Data Among Heterogeneous CAD Systems," in Proceedings of the 2009 13th International Conference on Computer Supported Cooperative Work in Design. Santiago, Chile: IEEE, 2009, Pp. 716–721.

[11] J. Zhang, F. Z. He, S. H. Han, and X. X. Li, "Quantitative Optimization of Interoperability during Feature-based Data Exchange," Integrated Computer Aided Engineering, Vol. 23, No. 1, Pp. 31–50, 2015.

[12] D. Shao, F. Chen, H. L. Liu, and L. Liu, "Research on Feature-based Heterogeneous CAD Model Transformation Technology," China Mechanical Engineering, Vol. 18, No. 1, Pp. 60–64, 2007.

[13] L. Zheng, J. R. Li, L. Wu, and X. D. Shao, "Research on Singular Feature Migration Method for Heterogeneous CAD Systems," Machinery Design & Manufacture, No. 3, Pp. 204–208, 2025.

[14] W. Qin, L. Y. Li, and S. M. Gao, "an Ontology Mapping Method for Heterogeneous Parametric Feature Model Retrieval," Computer Integrated Manufacturing Systems, Vol. 19, No. 7, Pp. 1472–1483, 2013.

[15] Tessier and Y. Wang, "Ontology-based Feature Mapping and Verification between CAD Systems," Advanced Engineering Informatics, Vol. 27, No. 1, Pp. 76–92, 2013.

[16] Camba, M. Contero, M. Johnson, and P. Company, "Extended 3D Annotations as a New Mechanism to Explicitly Communicate Geometric Design Intent and Increase CAD Model Reusability," Computer-Aided Design, Vol. 57, Pp. 61–73, 2014.

[17] Sun, "Research on Feature-based Technology of Data Exchange between Heterogeneous CAD Systems," M.S. Thesis, Dalian University of Technology, Dalian, China, 2010.

[18] K. Xie, G. Zhao, Y. Yu, and Y. D. Wang, "Research and Implementation of Feature-based Data Exchange Method of Heterogeneous 3D Digital Definition Model," Computer Integrated Manufacturing Systems, Vol. 23, No. 9, Pp. 1833–1841, 2017.

Automated Classification of Moroccan Procedural Court Orders Using CAMeL-BERT

Saliha YASSINE, Ouafaa IBRIHICH, Mohammed ESSOUJAA, Mustapha ESGHIR
LabMIA-SI Laboratory, Faculty of Sciences, Mohammed V University in Rabat, Morocco

***Abstract*—The digitization of judicial systems creates new opportunities for automating document management tasks that have traditionally relied on manual processing. This study addresses the automated classification of Arabic procedural court orders issued by Moroccan courts of first instance into seven official procedural categories, a low-resource legal document type characterized by highly standardized procedural language and limited publicly available resources. The primary contribution is a leakage-controlled evaluation demonstrating that frozen CAMeL-BERT contextual representations, combined with lightweight conventional machine learning classifiers, achieve highly accurate classification without computationally expensive transformer fine-tuning. The proposed pipeline integrates automatic structured extraction from heterogeneous Word documents with contextual embeddings and a comparative evaluation of five classifiers under a strict train–validation–test protocol. Experimental results show that Logistic Regression achieves 99.74% test accuracy (κ = 0.9966), while even the simplest baseline, Gaussian Naive Bayes, reaches 96.66%, highlighting the strong discriminative characteristics of Moroccan procedural court orders. An ablation study demonstrates that replacing TF-IDF representations with CAMeL-BERT embeddings reduces classification errors by 71.4%, while the complete classification pipeline further improves overall performance and reduces the remaining errors. These findings show that the highly regular linguistic structure of Moroccan procedural court orders enables computationally efficient and deployment-oriented classification using frozen contextual embeddings, providing a practical solution for Arabic legal document management and supporting future e-Justice applications in Morocco.**

***Keywords*—Arabic legal NLP; CAMeL-BERT; judicial document classification; Moroccan courts; procedural court orders**

I. Introduction

Moroccan courts of first instance process a large volume of procedural documents on a daily basis. These documents, commonly referred to as ordonnances, are formal judicial orders related to procedural matters such as payment obligations, financial consignments, commercial registry operations, guardianship appointments, and other judicial procedures. Recent investments by the Moroccan Ministry of Justice in electronic case management systems have accelerated the digitization of judicial workflows, making automated document processing an increasingly important operational requirement. Nevertheless, document classification remains largely manual in many courts, where incoming files must be identified and routed to the appropriate judicial department. Misclassification may delay judicial proceedings, increase administrative workload, and negatively affect procedural efficiency. Although these documents are stored in heterogeneous Word layouts that vary across courts, clerks, and time periods, they exhibit highly regular legal formulations and category-specific terminology, making them suitable candidates for data-driven document classification.

Automating this classification task requires addressing several linguistic challenges inherent to Arabic. Written Arabic frequently omits short vowels and diacritical marks, making tokenization and morphological analysis substantially more difficult than in many European languages. Arabic morphology is also highly productive, as a single trilateral root can generate numerous surface forms through affixation and internal vowel variation [1]. The Transformer architecture [2], particularly through BERT [3], addressed many of these challenges by learning deep contextual representations from large-scale corpora. These advances have been extended to Arabic through specialized models such as AraBERT [4] and CAMeL-BERT [5], which have demonstrated strong performance across a broad range of Arabic NLP tasks [6]. More recently, these advances have stimulated the development of Legal AI applications, including legal document classification, legal information retrieval, judgment prediction, intelligent document management, and e-Justice systems.

Despite recent progress in Arabic legal NLP [7]–[10], existing studies have primarily focused on legal corpora from other Arabic-speaking countries, general-domain datasets, or legal prediction tasks. Moroccan procedural court orders remain largely unexplored despite their central role in the daily operation of Moroccan courts. Furthermore, previous studies generally assume that documents are already available as clean textual data and therefore do not address the practical challenge of extracting structured textual content from heterogeneous judicial Word documents before classification. In addition, few studies investigate whether frozen contextual embeddings combined with lightweight conventional classifiers can provide both high classification performance and computational efficiency without expensive transformer fine-tuning. These limitations reveal an important gap between recent advances in Arabic Legal AI research and the practical requirements of judicial document management in Moroccan courts.

To address these challenges, this study proposes a complete document classification pipeline that integrates automatic structured text extraction from judicial Word documents with frozen CAMeL-BERT contextual embeddings and conventional machine learning classifiers. The proposed approach is evaluated under a strict train–validation–test protocol specifically designed to prevent data leakage while assessing the impact of contextual embeddings and training-only data augmentation. Beyond achieving high classification accuracy,

the study investigates whether deployment-oriented architectures based on frozen contextual representations can provide an effective and computationally efficient solution for real-world judicial document management.

The main contributions of this work are summarized as follows:

- An automatic extraction pipeline for converting heterogeneous Moroccan judicial Word documents into structured textual data suitable for classification.
- A curated corpus of Moroccan procedural court orders covering seven official judicial procedural categories.
- A leakage-controlled evaluation of frozen CAMeL-BERT contextual embeddings combined with five conventional machine learning classifiers.
- An ablation study quantifying the contribution of contextual embeddings and the complete classification pipeline to overall classification performance.
- An analysis of classification performance on an imbalanced judicial corpus, including minority categories.
- A deployment-oriented classification framework that achieves high accuracy without transformer fine-tuning, supporting practical integration into Morocco's e-Justice initiatives.

The remainder of this study is organized as follows: Section II reviews related work on Arabic legal NLP and judicial document classification. Section III presents the proposed extraction and classification methodology. Section IV discusses the experimental results and ablation study. Finally, Section V concludes the study and outlines directions for future research.

II. Related Work

A. Pre-Trained Arabic Language Models

Transformer-based pre-trained language models have significantly advanced Arabic Natural Language Processing (NLP) by learning contextual language representations from large-scale corpora. Unlike multilingual models, Arabic-specific transformers explicitly capture the rich morphology, lexical ambiguity, and syntactic characteristics of Arabic, leading to substantial improvements across text classification, named entity recognition, sentiment analysis, and question answering tasks.

AraBERT [4] was among the first transformer models specifically developed for Arabic and demonstrated that language-specific pre-training consistently outperforms multilingual BERT on several downstream NLP tasks. Subsequently, Alammary [6] conducted a comprehensive systematic review of Arabic BERT models, highlighting the influence of corpus composition, model architecture, and pre-training strategy on downstream performance. Abdul-Mageed et al. [12] further introduced ARBERT and MARBERT, two transformer models specifically designed to capture both Modern Standard Arabic (MSA) and dialectal Arabic. Although these models achieve excellent performance on dialect-rich corpora, their emphasis on dialectal language makes them less suitable for the predominantly Modern Standard Arabic used in Moroccan judicial documents.

CAMeL-BERT [5] further extended Arabic language modeling by pre-training transformer models on a 167 GB corpus covering Modern Standard Arabic, dialectal Arabic, and Classical Arabic. This multi-register training strategy enables the CAMeL-BERT-mix variant to capture both standard morphological phenomena and domain-specific lexical variations frequently encountered in formal legal writing. Its effectiveness has been demonstrated across several Arabic NLP benchmarks. For example, Mutawa and Sruthi [11] reported that CAMeL-BERT achieved the highest performance among several transformer architectures for Arabic meter classification, illustrating its strong capability to learn complex contextual representations.

More recently, transformer-based research has expanded toward Legal AI through the development of domain-adaptive language models. LEGAL-BERT [13] demonstrated that additional pre-training on legal corpora substantially improves performance on legal-domain NLP tasks compared with general-purpose BERT models. Similarly, AraLegal-BERT [15] showed that domain-specific pre-training on Arabic legal texts further enhances semantic representations for specialized legal terminology. These studies collectively confirm that contextual language models benefit significantly from legal-domain adaptation when processing judicial documents.

The present study adopts the CAMeL-BERT-mix model because Moroccan procedural court orders are written almost exclusively in Modern Standard Arabic while containing highly regular legal terminology and institution-specific procedural expressions. Unlike approaches based on full transformer fine-tuning, this work investigates whether frozen contextual embeddings extracted from CAMeL-BERT can provide sufficiently discriminative semantic representations for highly accurate document classification when combined with lightweight conventional machine learning classifiers. This deployment-oriented strategy substantially reduces computational requirements while preserving the contextual information learned during large-scale Arabic pre-training, making it particularly suitable for practical judicial applications.

B. Methods for Legal Document Classification

Automatic legal document classification has become an increasingly important research topic because of its applications in intelligent document management, legal information retrieval, electronic case management, and e-Justice systems. Early approaches relied primarily on manually engineered lexical and syntactic features combined with conventional machine learning algorithms such as Support Vector Machines (SVM), Logistic Regression (LR), K-Nearest Neighbors (KNN), and Naive Bayes (NB). These methods achieved competitive performance on relatively homogeneous legal corpora but required substantial feature engineering and often struggled to capture the semantic complexity of legal language, particularly in morphologically rich languages such as Arabic.

The emergence of transformer-based language models fundamentally changed this landscape by replacing manually engineered features with contextual semantic representations learned from large-scale corpora. Domain-adaptive pre-training has proved particularly beneficial for legal applications. LEGAL-BERT [13] demonstrated that additional pre-training on legal corpora substantially improves performance compared with general-purpose BERT models. MultiEURLEX [14] further showed that transformer models can successfully perform multilingual legal document classification across European legislation, while AraLegal-BERT [15] confirmed the importance of domain-specific pre-training for Arabic legal text classification.

Several recent studies have applied machine learning and transformer models to Moroccan legal documents. El Moussaoui et al. [7] fine-tuned XLM-RoBERTa for judicial chamber classification using Court of Cassation judgments and reported an accuracy of 98.48%. Tahtah et al. [17] proposed a CNN–BiLSTM architecture using AraBERTv2 embeddings together with random oversampling to address class imbalance in Arabic legal document classification. Bouhouche et al. [8] evaluated conventional machine learning algorithms for Moroccan legislative document classification and demonstrated that lexical representations remain competitive on structured legal corpora. Zahir [9] investigated judicial outcome prediction from Arabic court decisions using deep learning, whereas El Moussaoui et al. [10] developed an AI-assisted framework for criminal record processing in Moroccan courts.

Although these studies demonstrate the growing maturity of Arabic Legal AI, they address legal tasks that differ substantially from the automatic classification of procedural court orders issued by Moroccan courts of first instance. Furthermore, most previous work assumes that textual content is already available in structured digital form and therefore overlooks the practical challenge of extracting discriminative information from heterogeneous judicial Word documents before classification. Existing studies also predominantly rely on transformer fine-tuning or deep neural architectures, whereas comparatively little attention has been devoted to deployment-oriented solutions based on frozen contextual embeddings combined with lightweight conventional classifiers.

TABLE I. COMPARISON OF REPRESENTATIVE STUDIES ON ARABIC LEGAL DOCUMENT CLASSIFICATION

Study	Corpus	Doc. Type	Task	Approach	Best Score	Main Limitation
LEGAL-BERT [13]	European legal corpora	Legal documents	Document classification	Domain-adaptive BERT	Improved performance	Non-Arabic corpus
MultiEURLEX [14]	European Union	Legislative texts	Multi-label classification	Multilingual Transformer	Strong benchmark	European legislation
AraLegal-BERT [15]	Arabic legal corpus	Legal documents	Legal text classification	Domain-specific Arabic BERT	Improved representations	No Moroccan evaluation
Bouhouche et al. [8]	Morocco	Legislative texts	Document classification	Conventional ML	95.82%	Legislative corpus
Zahir [9]	Morocco	Court decisions	Judgment prediction	Deep learning	80.51%	Outcome prediction
Tahtah et al. [17]	Morocco	Court cases	Document classification	CNN–BiLSTM + AraBERTv2	96.22%	Requires fine-tuning
El Moussaoui et al. [7]	Morocco	Court of Cassation judgments	Chamber classification	Fine-tuned XLM-RoBERTa	98.48%	Different court level
This study	Moroccan Courts of First Instance	Procedural orders	7-category classification	Frozen CAMeL-BERT + LR	99.74% ($\kappa = 0.9966$)	Single national corpus

The comparison presented in Table I highlights several important research gaps. First, most existing studies focus on legal corpora that differ substantially from Moroccan procedural court orders, including legislative texts, Court of Cassation judgments, or judgment prediction tasks. Consequently, their findings cannot be directly transferred to the document-routing problem encountered in Moroccan courts of first instance. Second, previous work generally assumes that documents are already available in clean textual form and therefore overlooks the practical challenge of extracting structured information from heterogeneous judicial Word documents before classification. Third, although transformer-based models have consistently demonstrated superior performance, relatively few studies have investigated deployment-oriented architectures that combine frozen contextual embeddings with lightweight conventional machine learning classifiers, thereby reducing computational requirements while maintaining high classification performance.

To address these limitations, the present study proposes an end-to-end framework integrating automatic structured text extraction, frozen CAMeL-BERT contextual embeddings, and five conventional machine learning classifiers evaluated under a strict leakage-controlled experimental protocol. In addition, a comprehensive ablation study is conducted to quantify the contribution of contextual embeddings and the complete classification pipeline, providing practical insights into the design of efficient Arabic legal document classification systems for e-Justice applications.

III. METHODOLOGY

A. Corpus Compilation and Data Augmentation

The corpus was compiled from procedural court orders collected from several Moroccan Courts of First Instance. The documents correspond to seven official procedural categories defined by the Moroccan judicial taxonomy: Notification of Warning (تبليغ إنذار), Payment Order (أمر بالأداء), Deposit and Consignment (العرض والإيداع), Validation of Warning (المصادقة على الإنذار), Financial Withdrawal (سحب مبلغ مالي), Appointment of Guardian (تعيين قيم), and Commercial Register Deletion (التشطيب على السجل التجاري). Since these procedural categories are

officially assigned during judicial processing, no additional manual annotation was required. The labels provided by the courts were therefore used directly as the ground-truth class labels throughout the experiments.

The collected documents were stored as Microsoft Word files with heterogeneous layouts reflecting differences among courts, clerks, and document preparation practices. To prepare a structured corpus suitable for machine learning, a custom Python extraction pipeline was developed to automatically identify and extract the facts section (faits) and the decision section (dispositif) from each document using rule-based textual markers and regular-expression matching. Documents with incomplete or unsuccessful section extraction were excluded during corpus preparation, resulting in a final corpus of 3,894 original documents. Preliminary experiments showed that the extracted facts section contains the most discriminative information for procedural category identification; therefore, only this textual content was used for classification.

TABLE II. CORPUS DISTRIBUTION

Procedural Category	Original Documents
Notification of Warning — تبليغ إنذار	1480
Deposit and Consignment — العرض والإيداع	934
Validation of Warning — المصادقة على الإنذار	394
Financial Withdrawal — سحب مبلغ مالي	492
Payment Order — أمر بالأداء	432
Appointment of Guardian — تعيين قيم	120
Commercial Register Deletion — التشطيب على السجل التجاري	42
Total	3894

As shown in Table II, the original corpus exhibits a pronounced class imbalance, with an imbalance ratio of 35.24:1 between the largest and smallest categories. To reduce this imbalance while preserving the semantic meaning of each procedural category, data augmentation was applied exclusively to the training partition after train–validation–test splitting, thereby preventing any information leakage into the validation or test sets.

A per-class target of approximately 2,000 training documents was adopted. New training samples were generated by combining three label-preserving transformations specifically designed for procedural judicial documents:

- Judicial synonym substitution using a domain-specific lexicon of 40 legally equivalent term pairs validated by an experienced court clerk (e.g., التماس/طلب, الفصل/المادة);
- Date variation, replacing four-digit years with nearby values (±1 year);
- Monetary-value scaling, multiplying monetary amounts by a random factor between 0.8 and 1.2.

Date and monetary-value transformations were applied only to information that does not determine the procedural category. Since the seven target classes are defined by the underlying judicial procedure rather than by specific dates or monetary amounts, these modifications preserve the original class label. Every augmented document therefore inherits the category of its source document, and no cross-category mixing was performed.

The augmentation process increased the size of the training corpus to 13,764 documents, reducing the imbalance ratio from 35.24:1 to approximately 1.12:1 while maintaining a strict separation between original evaluation data and augmented training data.

Fig. 1 presents an anonymized excerpt of a Notification of Warning (تبليغ إنذار) procedural court order after the automatic extraction process. All personal names, case numbers, and identifying information were removed or replaced before analysis to preserve judicial confidentiality.

المحكمة الابتدائية — [×] نحن رئيس المحكمة الابتدائية بمقتضى الفصل *148* من ق.م.م. بعد الاطلاع على طلب السيد [×××] بتبليغ إنذار للمدين [×××] بأداء مبلغ [...] درهماً أمرنا بما يلي: [تبليغ المدين بالمبلغ المذكور

Fig. 1. Anonymized excerpt of the facts section of a Notification of Warning (تبليغ إنذار) procedural court order extracted by the Python script. Personal identifiers have been replaced by ×.

The extracted facts section subsequently undergoes text normalization and contextual embedding generation before being used for document classification. The overall preprocessing pipeline is described in the following subsection.

B. Data Preprocessing and Experimental Protocol

The extracted textual data underwent several preprocessing steps before classification. Text normalization included the removal of unnecessary whitespace and formatting inconsistencies introduced during document extraction, together with the normalization of Arabic text to ensure consistent lexical representation. Documents with incomplete or unsuccessful section extraction, as well as documents containing insufficient textual content for reliable classification, were excluded during corpus preparation to improve the overall quality of the dataset. The resulting corpus contained 3,894 original procedural court orders, which served as the basis for all subsequent experiments.

The final corpus was partitioned into training (60%), validation (20%), and test (20%) sets using stratified random sampling in order to preserve the original class distribution across all partitions. The partitioning was performed before any data augmentation to prevent information leakage between the training and evaluation datasets. Consequently, every original document appears in only one partition throughout the entire experimental protocol.

Data augmentation was applied exclusively to the training partition, increasing the number of training documents from 2,336 to 13,764 while leaving the validation and test sets unchanged. This strategy preserved the semantic integrity of each procedural category while reducing the class imbalance ratio from 35.24:1 to approximately 1.12:1. Since the validation and test partitions contain only original documents, the reported performance reflects the model's ability to generalize to previously unseen judicial documents rather than to augmented samples.

To ensure a fair comparison among all classification models, the same training, validation, and test partitions were used throughout every experiment, including the ablation study. This

leakage-controlled experimental protocol guarantees that all reported performance differences result exclusively from changes in document representation or classification algorithms rather than from variations in the evaluation data.

C. Proposed Classification Framework

Fig. 2 presents the overall workflow of the proposed framework for the automatic classification of Moroccan procedural court orders. The framework consists of five sequential stages that transform heterogeneous judicial Word documents into contextual semantic representations suitable for procedural document classification.

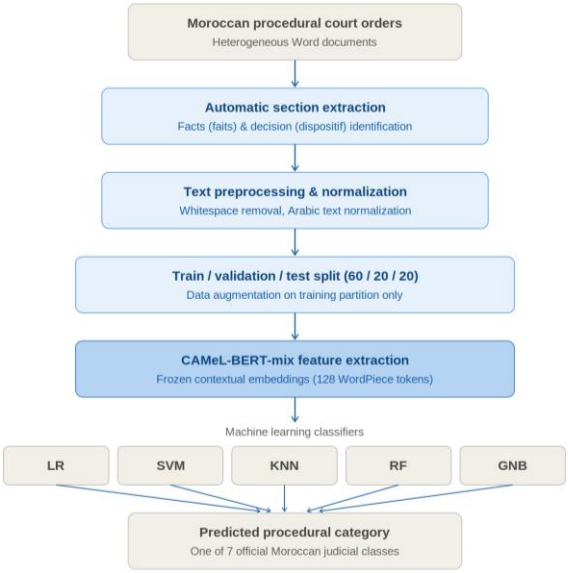


Fig. 2. Overall workflow of the proposed framework for automatic classification of Moroccan procedural court orders.

First, procedural court orders are processed by the automatic extraction module described in the previous subsection. This module identifies and extracts the facts (faits) and decision (dispositif) sections from each document. Preliminary experiments showed that the facts section contains the most discriminative procedural information; therefore, only this section is retained for the subsequent classification process.

The extracted text is then normalized to remove formatting inconsistencies while preserving its legal content. The resulting corpus is partitioned into training, validation, and test sets using a leakage-controlled experimental protocol. Data augmentation is applied exclusively to the training partition in order to mitigate class imbalance while preserving the semantic characteristics of each procedural category.

Next, each preprocessed document is converted into a contextual semantic representation using the frozen CAMeL-BERT-mix model. The resulting document embeddings are subsequently used to train five conventional machine learning classifiers: Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest (RF), and Gaussian Naive Bayes (GNB).

By combining transformer-based contextual embeddings with lightweight machine learning classifiers, the proposed framework provides an effective and computationally efficient solution for the automatic routing of procedural court orders within Moroccan e-Justice systems.

D. CAMeL-BERT Feature Extraction

The proposed framework employs the CAMeL-BERT-mix model [5] as the contextual feature extractor. The main configuration parameters of the feature extraction module are summarized in Table III. CAMeL-BERT-mix is built upon the standard BERT-base architecture [3], comprising 12 transformer encoder layers, 768-dimensional hidden representations, 12 self-attention heads, and approximately 110 million parameters. The model was pre-trained on a large-scale Arabic corpus covering Modern Standard Arabic (MSA), dialectal Arabic, and Classical Arabic, making it particularly suitable for legal documents that combine formal language with institution-specific procedural terminology.

TABLE III. CAMEL-BERT CONFIGURATION USED IN THIS STUDY

Parameter	Value
Pre-trained model	CAMeL-BERT-mix
Base architecture	BERT-base
Transformer encoder layers	12
Hidden size	768
Attention heads	12
Approximate number of parameters	110 million
Tokenizer	WordPiece
Maximum sequence length	128 tokens
Document representation	Final hidden-layer [CLS] embedding
Embedding dimension	768
Transfer learning strategy	Frozen embeddings (no fine-tuning)
Feature scaling	StandardScaler (fitted on the training set only)

Classification is performed exclusively on the extracted facts section, as it contains the procedural information most relevant for distinguishing between the seven judicial categories. Each document is tokenized using the original CAMeL-BERT WordPiece tokenizer, and input sequences are truncated or padded to a maximum length of 128 tokens, ensuring a fixed-size representation compatible with the transformer architecture.

Rather than fine-tuning the transformer model, this study adopts a feature-based transfer learning strategy. The parameters of CAMeL-BERT remain frozen throughout the training process, and only the contextual document representations generated by the model are used as input features for the downstream classifiers. This strategy substantially reduces computational requirements while allowing the evaluation to focus on the discriminative capability of contextual embeddings independently of transformer optimization [18].

For each document, the contextual representation of the [CLS] token from the final hidden layer is extracted as a fixed-length 768-dimensional document embedding, which serves as the input feature vector for all classification models. Before classifier training, these embeddings are standardized using StandardScaler. The scaler is fitted exclusively on the training partition and subsequently applied, without refitting, to the

validation and test partitions, thereby preventing information leakage during preprocessing and ensuring identical feature scaling across all evaluation datasets.

The combination of frozen contextual embeddings and lightweight machine learning classifiers provides an efficient architecture for deployment-oriented judicial applications. Once document embeddings have been generated, classification can be performed without repeated transformer optimization, substantially reducing computational cost while preserving the semantic knowledge acquired during large-scale Arabic language pre-training.

E. Classification Models

To evaluate the effectiveness of the proposed contextual representations, five supervised machine learning classifiers from different algorithmic families were trained using the CAMeL-BERT document embeddings. The hyperparameter configuration of all classifiers is summarized in Table IV.

TABLE IV. HYPERPARAMETER SETTINGS OF THE CLASSIFICATION MODELS.

Classifier	Hyperparameter Configuration
Logistic Regression	L2 regularization, C = 0.5, class_weight = balanced, max_iter = 1000
Support Vector Machine	RBF kernel, C = 1.0, gamma = scale
K-Nearest Neighbors	k = 5, cosine distance
Random Forest	n_estimators = 100, max_depth = 10, min_samples_split = 10, min_samples_leaf = 5, class_weight = balanced
Gaussian Naive Bayes	Default parameters

Logistic Regression (LR) was selected as a linear baseline because of its robustness, computational efficiency, and effectiveness in high-dimensional feature spaces. The classifier employs L2 regularization with C = 0.5, class_weight = "balanced", and a maximum of 1,000 iterations.

Support Vector Machine (SVM) was configured with a radial basis function (RBF) kernel using C = 1.0 and gamma = "scale", enabling nonlinear decision boundaries to be learned in the contextual embedding space [16].

K-Nearest Neighbors (KNN) was implemented with k = 5 using the cosine distance metric, which is well suited for measuring similarity between high-dimensional contextual embeddings.

Random Forest (RF) consisted of 100 decision trees with a maximum tree depth of 10, a minimum of 10 samples required to split an internal node, a minimum of 5 samples per leaf node, and class_weight = "balanced" to reduce the impact of class imbalance [19].

Finally, Gaussian Naive Bayes (GNB) was included as a probabilistic baseline because of its simplicity and computational efficiency.

To ensure a fair comparison, all classifiers were trained using the same standardized CAMeL-BERT embeddings and the same training, validation, and test partitions. Hyperparameters were selected through grid search on the validation set and subsequently kept fixed for all experiments.

F. Evaluation Metrics

The performance of the proposed classification framework was evaluated using Accuracy, Precision, Recall, F1-score, and Cohen's Kappa coefficient (κ). Accuracy provides the overall proportion of correctly classified documents, whereas Precision and Recall evaluate the correctness and completeness of the predicted classes, respectively. The F1-score, defined as the harmonic mean of Precision and Recall, provides a balanced measure of classification performance, particularly for imbalanced datasets. Cohen's Kappa (κ) was additionally reported to quantify the agreement between predicted and true labels while accounting for agreement occurring by chance.

Let TP, TN, FP, and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively. The evaluation metrics are defined as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$F_1 = \frac{2\times Precision\times Recall}{Precision + Recall} \quad (4)$$

$$\kappa = \frac{P_o - P_e}{1 - P_e} \quad (5)$$

where, P_o denotes the observed agreement between predicted and true labels, and P_e represents the agreement expected by chance.

To provide a comprehensive assessment of classifier performance, the evaluation was conducted on the held-out test set, while model selection and hyperparameter tuning were performed exclusively on the validation set. In addition, five-fold stratified cross-validation was employed on the training data during model development to assess the robustness and generalization capability of the proposed framework.

IV. RESULTS AND DISCUSSION

A. Overall Classifier Performance

Table V summarizes the overall performance of the five evaluated classifiers on the independent test set comprising 779 original procedural court orders. Among all evaluated models, Logistic Regression (LR) achieved the highest classification performance, reaching a test accuracy of 99.74% with a Cohen's Kappa coefficient (κ) of 0.9966. Support Vector Machine (SVM) achieved the second-best performance with 99.49% accuracy (κ = 0.9932), followed by K-Nearest Neighbors (KNN) at 99.61%, Random Forest (RF) at 98.84%, and Gaussian Naive Bayes (GNB) at 96.66%.

The consistently high cross-validation accuracies, together with the very low standard deviations (±0.06%–±0.37%), demonstrate that the proposed framework generalizes well across different training folds and that the reported performance is not dependent on a particular train–test partition. Similarly, the high Cohen's Kappa values (LR: 0.9966, SVM: 0.9932,

KNN: 0.9949, RF: 0.9848, and GNB: 0.9562) indicate an almost perfect agreement between predicted and true labels, confirming that the excellent classification performance is not an artifact of class imbalance.

TABLE V. OVERALL PERFORMANCE OF THE EVALUATED CLASSIFICATION MODELS

Classifier	Validation Accuracy	Test Accuracy	5-Fold Cross-Validation Accuracy
Naive Bayes	97.30%	96.66%	98.34% ±0.37%
K-Nearest Neighbors	98.97%	99.61%	99.89% ±0.07%
Random Forest	98.84%	98.84%	99.91% ±0.07%
Support Vector Machine	99.49%	99.49%	99.96% ±0.07%
Logistic Regression	99.49%	99.74%	99.97% ±0.06%

Overall, the proposed framework demonstrates that combining frozen CAMeL-BERT contextual embeddings with lightweight machine learning classifiers provides highly accurate and robust classification of Moroccan procedural court orders. The strong performance can be attributed to three complementary factors: 1) the rich contextual semantic representations learned by CAMeL-BERT; 2) the highly discriminative legal vocabulary contained in the extracted facts section; and 3) the balanced training corpus obtained through targeted data augmentation. In addition, the low mean pairwise Jaccard similarity (12.3%) among the seven procedural categories indicates limited lexical overlap between categories, facilitating their discrimination in the embedding space. Finally, the consistent ranking of the evaluated classifiers (LR ≈ SVM > KNN > RF > GNB) across validation, test, and cross-validation experiments indicates that this is primarily attributable to the discriminative power of the contextual embeddings rather than classifier complexity.

B. Effect of Data Augmentation

Data augmentation increased the size of the training set from 2,336 to 13,764 documents while reducing the class imbalance ratio from 35.24:1 to 1.12:1. This strategy was specifically designed to improve the representation of minority procedural categories without modifying the validation and test sets, thereby preserving the integrity of the evaluation protocol.

The effectiveness of data augmentation is reflected in the per-category results reported in Table VII, where the two minority classes, تعيين قيم (24 test documents) and التشطيب على السجل التجاري (8 test documents), both achieved an F1-score of 1.000 despite their limited number of original training examples. These results indicate that the proposed augmentation strategy successfully compensated for severe class imbalance while preserving the discriminative characteristics of each procedural category.

The contribution of data augmentation is further confirmed by the ablation study presented in Table VI. Although the overall classification accuracy remains nearly unchanged, augmentation improves class-balanced performance by eliminating the remaining misclassifications affecting the minority class تعيين قيم. Consequently, the only residual errors are observed between المصادقة على الإنذار and العرض والإيداع, the two procedural categories exhibiting the highest semantic similarity.

Overall, these findings demonstrate that the principal contribution of data augmentation is not to increase the already high overall accuracy, but rather to enhance the robustness and reliability of the classifier across all procedural categories. This improvement is particularly important in judicial applications, where reliable recognition of minority case types is essential for practical deployment.

TABLE VI. ABLATION STUDY OF THE PROPOSED FRAMEWORK

Configuration	Accuracy	Macro-F1	F1 (تعيين قيم)	F1 (التشطيب)	Errors
TF-IDF + SVM	98.20%	0.9562	1.000	0.769	14/779
CAMeL-BERT + SVM (no augmentation)	99.49%	0.9933	0.979	1.000	4/779
CAMeL-BERT + LR (full pipeline)	99.74%	0.9970	1.000	1.000	2/779

C. Performance by Category

Table VII presents the per-category classification performance of the best-performing model, Logistic Regression, on the independent test set comprising 779 original procedural court orders. Overall, the proposed framework achieved perfect or near-perfect performance across all seven procedural categories. Five categories obtained an F1-score of 1.000, while المصادقة على الإنذار and العرض والإيداع achieved 99.46% and 98.09%, respectively.

TABLE VII. PER-CLASS PERFORMANCE OF THE LOGISTIC REGRESSION MODEL ON THE INDEPENDENT TEST SET

Procedural Category	Precision	Recall	F1-score	Support
Notification of Warning (تبليغ إنذار)	99.00%	100.00%	99.50%	296
Deposit and Consignment (العرض والإيداع)	100.00%	98.93%	99.46%	187
Financial Withdrawal (سحب مبلغ مالي)	100.00%	100.00%	100.00%	99
Payment Order (أمر بالأداء)	100.00%	100.00%	100.00%	86
Validation of Warning (المصادقة على الإنذار)	98.72%	97.47%	98.09%	79
Appointment of Guardian (تعيين قيم)	100.00%	100.00%	100.00%	24
Commercial Register Deletion (التشطيب على السجل التجاري)	100.00%	100.00%	100.00%	8

Note: Support denotes the number of original test documents in each procedural category.

The only two remaining misclassifications occurred between المصادقة على الإنذار and العرض والإيداع, which share similar procedural terminology and legal expressions. This observation suggests that the residual errors are primarily attributable to semantic similarity between these categories rather than to limitations of the contextual representations learned by CAMeL-BERT.

The two minority categories, تعيين قيم (24 test documents) and التشطيب على السجل التجاري (8 test documents), both achieved an F1-score of 1.000, demonstrating that the proposed framework generalizes effectively even for underrepresented procedural categories. Although these results should be interpreted with

appropriate caution because of the relatively small number of test instances, their consistent recognition provides additional evidence of the robustness of the proposed classification framework.

D. Error Analysis

The best-performing classifier (Logistic Regression) made only two errors on the independent test set of 779 original procedural court orders, corresponding to an error rate of 0.26%, while the SVM classifier produced four errors. Since both classifiers exhibit a very similar residual confusion pattern, the confusion matrix of the CAMeL-BERT + SVM configuration is presented in Fig. 3 as a representative visualization of the remaining ambiguity in the embedding space.

The four SVM misclassifications are exclusively concentrated between العرض والإيداع and المصادقة على الإنذار. Specifically, two documents belonging to العرض والإيداع were predicted as المصادقة على الإنذار, while two documents from المصادقة على الإنذار were classified as العرض والإيداع. No errors were observed for the remaining five procedural categories, including the minority classes تعيين قيم and التشطيب على السجل التجاري.

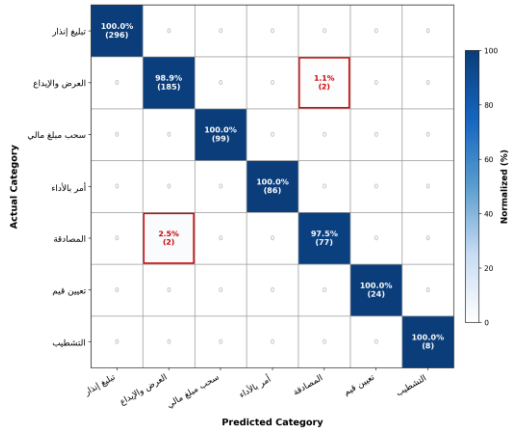


Fig. 3. Normalized confusion matrix of the CAMeL-BERT + SVM model on the independent test set (779 original documents).

The confusion matrix confirms that all residual errors are restricted to two procedurally related categories sharing similar legal terminology and document structure. The final Logistic Regression model further reduces these misclassifications from four to two while preserving the same overall confusion pattern, indicating that the remaining errors originate from semantic similarity between the categories rather than from limitations of the contextual representations.

From a practical perspective, these rare ambiguous cases could be addressed by incorporating a human verification step for predictions with low confidence scores. Such a strategy would enable reliable deployment in judicial document management systems while maintaining the efficiency of the proposed automated classification framework.

E. Corpus Structure and Accuracy

The 99.74% test accuracy achieved by the proposed Logistic Regression model is substantially higher than the performance reported in many previous studies on Arabic legal document classification and judicial prediction tasks. For example, judicial outcome prediction for European Court of Human Rights (ECHR) cases has reported F1-scores of approximately 79% [20], [21], while Zahir et al. [9] achieved an F1-score of 80.51% for Arabic court decision classification. Although these studies address different legal tasks and datasets, the comparison provides useful context for interpreting the present results.

The exceptionally high performance observed in this study can largely be attributed to the structural characteristics of Moroccan procedural court orders. First, the mean pairwise Jaccard similarity between category vocabularies is only 12.3%, indicating limited lexical overlap across procedural categories. Second, the contextual representations produced by CAMeL-BERT capture semantic and positional information beyond surface lexical features, resulting in highly separable document embeddings. Third, each procedural category follows standardized drafting conventions and contains recurring legal expressions that are consistently used within that category, thereby reducing intra-class variability. Finally, the documents are drafted by trained court clerks according to institutional writing practices, which further limits stylistic variation and linguistic ambiguity.

Taken together, these characteristics create a classification task that is particularly well-suited to contextual transformer embeddings combined with a linear classifier. This interpretation is consistent with previous studies on Moroccan legal documents, which also reported strong classification performance on structurally similar corpora, including 98.48% by El Moussaoui et al. [7] and 95.82% by Bouhouche et al. [8].

F. Deployment Prospects and Limitations

The proposed framework achieves an error rate of only 0.26% (2 misclassifications out of 779 test documents), suggesting that it is well suited for deployment as a decision-support tool rather than as a fully autonomous system. In practical settings, predictions associated with low confidence scores can be forwarded to court clerks for verification before final assignment, thereby combining the efficiency of automated document classification with human oversight.

Two deployment scenarios are envisioned within Morocco's ongoing e-Justice initiatives. The first is real-time document routing, where newly issued procedural court orders are automatically assigned to the appropriate departmental workflow while uncertain predictions are flagged for manual review. The second is retrospective digital archiving, where historical procedural court orders are automatically indexed according to their legal category, facilitating efficient search, retrieval, and document management.

An additional practical advantage of the proposed framework is its computational efficiency. Since CAMeL-BERT is used only for offline feature extraction and the final classifier is a lightweight Logistic Regression model, inference requires only standard CPU resources without GPU acceleration. This makes the system compatible with the computing infrastructure commonly available in Moroccan courts of first instance.

The proposed framework nevertheless has several limitations. First, the extraction pipeline was developed and validated using procedural court orders collected from

Moroccan courts of first instance between 2020 and 2025. Its applicability to documents from other court levels or jurisdictions may therefore require adaptation of the extraction component. Second, the classifier was trained to recognize only seven procedural categories. Extending the framework to additional legal document types would require new annotated data and retraining of the classification pipeline. Finally, the study evaluates the system on a single national corpus; future work should investigate its generalization to broader collections of Arabic legal documents and to judicial systems beyond Morocco.

V. CONCLUSION

This study presented an end-to-end framework for the automated classification of Moroccan procedural court orders. The proposed pipeline addresses a practical challenge that has received limited attention in previous studies: the lack of standardized document structure in Arabic legal Word documents. A custom Python extraction script automatically converts heterogeneous documents into structured representations by extracting the facts, decision, and confidence score, enabling the classifier to focus on the most informative textual content.

The framework was evaluated on a manually collected corpus of 3,894 procedural court orders from Moroccan courts of first instance (2020–2025), covering seven official procedural categories. After expanding the training set to 13,764 documents through domain-specific data augmentation, five machine learning classifiers were evaluated using frozen CAMeL-BERT embeddings. Logistic Regression achieved the best performance, reaching 99.74% test accuracy with a Cohen's κ of 0.9966. The ablation study further demonstrated that the strong performance results from the combination of highly discriminative contextual embeddings and the structural characteristics of Moroccan procedural court orders, while data augmentation effectively improves robustness for minority categories.

Although the proposed framework demonstrates excellent performance, several directions remain for future research. Extending the corpus to additional court levels, legal document types, and Arabic-speaking jurisdictions would improve the generalizability of the approach. A fine-tuned CAMeL-BERT baseline was not included in this study. Consequently, the effectiveness of frozen contextual embeddings is demonstrated for this corpus but was not directly compared with end-to-end transformer fine-tuning. Future work will investigate this comparison.

ACKNOWLEDGMENT

This work was supported by the Ministry of Higher Education, Scientific Research, and Innovation; the Digital Development Agency (DDA); and the National Centre for Scientific and Technical Research (CNRST) of Morocco under grant Alkhawarizmi/2020/25.

REFERENCES

[1] N. Y. Habash, Introduction to Arabic Natural Language Processing, ser. Synthesis Lectures on Human Language Technologies. Morgan & Claypool, 2010.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Proc. NeurIPS, Long Beach, CA, Dec. 2017, pp. 5998–6008.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, Minneapolis, MN, Jun. 2019, pp. 4171–4186.

[4] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based model for Arabic language understanding," in Proc. 4th Workshop on Open-Source Arabic Corpora and Processing Tools, Marseille, May 2020, pp. 9–15.

[5] G. Inoue, B. Alhafni, N. Baimukan, H. Bouamor, and N. Habash, "The interplay of variant, size, and task type in Arabic pre-trained language models," in Proc. 6th Arabic NLP Workshop, Kyiv, Apr. 2021, pp. 92–104.

[6] A. S. Alammary, "BERT models for Arabic text classification: A systematic review," Applied Sciences, vol. 12, no. 11, Art. no. 5720, Jun. 2022.

[7] T. El Moussaoui, A. Es-salmi, J. Boumhidi, and C. Loqman, "Automating legal document classification with XLM-RoBERTa: A case study on the Moroccan Court of Cassation," in Digital Technologies and Applications, ICDTA 2025, Lecture Notes in Networks and Systems, vol. 1640, Cham: Springer, 2026, ch. 30. doi: 10.1007/978-3-032-07785-1_30.

[8] A. Bouhouche, M. Esghir, and M. Errachid, "Classification of Moroccan legal and legislative texts using machine learning models," Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 10, pp. 1108–1114, 2024.

[9] J. Zahir, "Prediction of court decision from Arabic documents using deep learning," Expert Systems, vol. 40, no. 6, Art. no. e13236, Jun. 2023.

[10] T. El Moussaoui, C. Loqman, and J. Boumhidi, "Decoding legal processes: An AI-driven system for criminal records in Moroccan courts," Intell. Syst. Appl., vol. 25, Art. no. 200487, Mar. 2025.

[11] A. M. Mutawa and S. Sruthi, "A comparative evaluation of transformers and deep learning models for Arabic meter classification," Applied Sciences, vol. 15, no. 9, Art. no. 4941, Apr. 2025.

[12] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "ARBERT and MARBERT: Deep bidirectional transformers for Arabic," in Proc. 59th Annual Meeting of the ACL, Aug. 2021, pp. 7088–7105.

[13] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The muppets straight out of law school," in Findings of ACL: EMNLP 2020, Nov. 2020, pp. 2898–2904.

[14] I. Chalkidis, M. Fergadiotis, and I. Androutsopoulos, "MultiEURLEX—A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer," in Proc. EMNLP 2021, Online and Punta Cana, Nov. 2021, pp. 6974–6996.

[15] M. Al-Qurishi, S. AlQaseemi, and R. Souissi, "AraLegal-BERT: A pretrained language model for Arabic legal text," in Proc. Workshop on Natural Legal Language Processing, Abu Dhabi, Dec. 2022, pp. 338–344.

[16] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[17] O. Tahtah, M. Bahbib, A. Zinedine, and K. Fardousse, "A CNN-BiLSTM hybrid architecture with resampling techniques for Arabic legal text classification," Eng. Technol. Appl. Sci. Res., vol. 15, no. 6, pp. 29062–29068, Dec. 2025.

[18] Y. Sun, C. Qiu, Y. Xu, Z. Wang, and H. Chen, "How to fine-tune BERT for text classification?" in Chinese Computational Linguistics, Cham: Springer, Oct. 2019, pp. 194–206.

[19] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[20] I. Chalkidis, I. Androutsopoulos, and N. Aletras, "Neural legal judgment prediction in English," in Proc. 57th Annual Meeting of the ACL, Florence, Italy, Jul. 2019, pp. 4317–4323.

[21] N. Aletras, D. Tsarapatsanis, D. Preoţiuc-Pietro, and V. Lampos, "Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective," PeerJ Comput. Sci., vol. 2, Art. no. e93, Oct. 2016.

A Hybrid SMOTE-Ensemble Framework for Robust Speech Emotion Recognition on Imbalanced Data

Eyad Megdadi[1]*, Mariam Aljouhi[2], Zainab Alblooshi[3], Khaled Shaalan[4]
Faculty of Engineering and IT, British University in Dubai, Dubai 345015, United Arab Emirates[1, 3, 4]
Information Security Engineering Technology Department, Abu Dhabi Polytechnic, Abu Dhabi, United Arab Emirates[2]

***Abstract*—Speech Emotion Recognition (SER) has attracted considerable attention in recent years because of its potential applications in healthcare, intelligent assistants, customer service, and human-computer interaction. However, the performance of SER systems is often affected by the imbalance of emotion classes in benchmark datasets, where minority emotions are underrepresented and difficult to classify accurately. This study reproduces the work of Sahu et al. on multimodal Speech Emotion Recognition using the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset. Further, it proposes a Hybrid SMOTE-Ensemble Framework to improve classification performance on imbalanced data. The reproduction stage follows the original methodology by employing handcrafted acoustic features together with TF-IDF textual features under Audio-only, Text-only, and Audio+Text modalities. The enhancement stage introduces two XGBoost classifiers trained independently on the original imbalanced dataset and on a SMOTE-balanced dataset. Their prediction probabilities are combined using three ensemble strategies, namely Simple Averaging, Confidence-Based Selection, and Adaptive Weighting. In addition, the effectiveness of the proposed strategy is further investigated using a LinearSVC classifier. Experimental results demonstrate that the reproduced models achieve results that are highly consistent with those reported in the original study. Furthermore, the proposed Hybrid SMOTE-Ensemble Framework improves the recognition of minority emotion classes while maintaining balanced overall performance. The XGBoost model trained on SMOTE-balanced data achieves higher F1-score and recall than the baseline model, whereas the ensemble approaches provide additional improvements in classification performance. The LinearSVC experiments further confirm that combining oversampling and ensemble learning offers a robust solution to the class imbalance problem in Speech Emotion Recognition.**

***Keywords*—IEMOCAP dataset; Speech Emotion Recognition (SER); class imbalance; SMOTE**

I. Introduction

Emotion Recognition (SER) is an important research area in Artificial Intelligence (AI) that combines signal processing, machine learning, and affective computing. The main goal of SER is to automatically identify the emotional state of a speaker from speech signals. Unlike conventional speech recognition systems, which focus on recognizing spoken words, SER aims to understand the emotions hidden behind speech. This capability enables machines to interact with humans more naturally and intelligently [3].

In recent years, SER has attracted considerable attention because of its wide range of applications. Emotion-aware systems are increasingly used in healthcare, customer service, intelligent virtual assistants, education, and smart transportation systems. For example, intelligent assistants can adapt their responses according to the emotional state of users, while healthcare applications can use speech emotion analysis to monitor mental health conditions such as stress, anxiety, and depression. Similarly, driver monitoring systems can detect emotions such as anger or fatigue to improve road safety [3].

The rapid development of machine learning and deep learning techniques has significantly improved the performance of SER systems. Traditional machine learning approaches rely on handcrafted acoustic features such as pitch, energy, harmonics, pauses, and statistical descriptors. More recently, multimodal approaches that combine speech and textual information have shown promising results because emotional information is often conveyed through both acoustic and linguistic cues [3].

Despite these advances, SER remains a challenging task. One of the main challenges is the imbalance of emotion classes in benchmark datasets. In many real-world datasets, some emotions are represented by a large number of samples, whereas others contain only a limited number of examples. As a result, classification models tend to be biased toward majority classes and often fail to recognize minority emotions accurately. This problem becomes more significant in widely used SER datasets such as the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset, where emotions such as fear and surprise are underrepresented compared with other emotion categories [4].

Several methods have been proposed to address the class imbalance problem in SER. Traditional approaches include random oversampling, undersampling, and cost-sensitive learning. Among these methods, the Synthetic Minority Over-sampling Technique (SMOTE) has gained considerable attention because it generates synthetic samples for minority classes instead of simply duplicating existing instances. By creating new samples through interpolation between neighboring minority samples, SMOTE increases data diversity and reduces the risk of overfitting [6]. Previous studies have shown that SMOTE can improve the performance of classification models on imbalanced datasets and enhance the recognition of minority classes [7,8]. However, SMOTE also has some limitations, including the possibility of generating noisy samples, increasing class overlap, and producing synthetic instances that do not accurately represent the underlying data distribution [9,10].

Ensemble learning has also demonstrated strong performance in many classification problems by combining the

*Corresponding author.

strengths of multiple models. In SER, ensemble approaches can improve classification accuracy and robustness by exploiting the complementary characteristics of different classifiers. Nevertheless, effectively combining ensemble learning with oversampling techniques for imbalanced emotion recognition remains an open research problem. Existing methods often improve minority class recognition at the expense of majority classes or fail to maintain balanced overall performance.

Motivated by these challenges, this study proposes a Hybrid SMOTE-Ensemble Framework for robust Speech Emotion Recognition on imbalanced data. The proposed framework combines models trained on the original imbalanced dataset and on a SMOTE-balanced dataset using different ensemble strategies. The objective is to improve minority emotion recognition while preserving the overall classification performance. The effectiveness of the proposed framework is evaluated using the IEMOCAP dataset and compared with conventional machine learning approaches.

The remainder of this study is organized as follows. Section II presents the background of Speech Emotion Recognition and discusses existing approaches for handling class imbalance. Section III describes the proposed Hybrid SMOTE-Ensemble Framework, including dataset preparation, feature extraction, model development, and ensemble strategies. Section IV presents the experimental setup and evaluation metrics. Section V discusses the experimental results and comparative analysis. Finally, Section VI concludes the study and outlines possible directions for future work.

II. Background

A. Multimodal Speech Emotion Recognition

Speech Emotion Recognition (SER) is an important research field that aims to identify the emotional state of a speaker from speech signals. The objective of SER is to enable machines to understand human emotions and improve human-computer interaction. In recent years, SER has been applied in many areas such as intelligent virtual assistants, healthcare monitoring, customer service, and smart transportation systems [3].

Emotions are usually expressed through multiple channels at the same time. Humans naturally combine information from speech, text, facial expressions, and gestures to understand emotions. Similarly, multimodal SER systems utilize information from different modalities to improve recognition accuracy. Sahu et al. [3] stated that humans are able to resolve ambiguity in emotions because they can efficiently process information from multiple domains, including speech, text, and visual signals. This observation has motivated the development of multimodal SER systems that combine acoustic and textual information to provide a more comprehensive understanding of emotional expressions.

In audio-based SER, acoustic features play an important role in emotion classification. Commonly used features include prosodic characteristics such as pitch, energy, speaking rate, and voice quality. Spectral features such as Mel Frequency Cepstral Coefficients (MFCCs), formants, harmonics, and statistical descriptors are also widely used because they capture important information related to emotional variations in speech [3]. In addition to acoustic information, textual features extracted from speech transcriptions can provide useful semantic information. Traditional text representations include Bag-of-Words and Term Frequency-Inverse Document Frequency (TF-IDF), while more recent approaches employ word embeddings and pre-trained language models.

Among the available SER datasets, the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset is one of the most widely used benchmarks. The dataset contains approximately 12 hours of multimodal recordings collected from ten actors participating in scripted and improvised dialogues [4]. Each utterance is annotated by multiple annotators according to different emotional categories, making the dataset suitable for evaluating speech emotion recognition methods. However, similar to many real-world datasets, IEMOCAP exhibits an imbalanced distribution of emotion classes, where some emotions are represented by a large number of samples while others contain relatively few examples. This imbalance remains one of the major challenges affecting the performance and robustness of SER systems.

B. Class Imbalance and Oversampling Techniques

Class imbalance is one of the major challenges in Speech Emotion Recognition (SER). In many emotional speech datasets, some emotions are represented by a large number of samples, while others contain only a limited number of examples. As a result, machine learning models tend to favor majority classes and often produce poor recognition performance for minority emotions. This issue becomes more serious when the evaluation focuses only on overall accuracy, as a high accuracy score may hide weak performance on underrepresented classes.

The IEMOCAP dataset is a clear example of this problem. The distribution of emotion classes in IEMOCAP is not balanced, with emotions such as fear and surprise containing fewer samples than emotions such as sadness and neutrality. Previous studies reported that this imbalance negatively affects the performance of emotion classification models, particularly for minority emotions that are difficult to learn because of the limited training data available [3]. Therefore, addressing class imbalance has become an important research topic in SER [11].

Traditional methods for handling imbalanced data include random undersampling and random oversampling. Undersampling reduces the number of majority class samples to balance the dataset; however, useful information may be lost during this process. On the other hand, random oversampling duplicates minority class samples to increase their representation. Although this method is simple and easy to implement, it does not introduce new information and may increase the risk of overfitting because the same samples are repeatedly used during training [6].

Among the available oversampling techniques, the Synthetic Minority Over-sampling Technique (SMOTE) is one of the most widely used methods for imbalanced classification problems. SMOTE generates new synthetic samples for minority classes instead of simply duplicating existing instances. The method creates new samples by interpolating between a minority sample and its nearest neighbors in the feature space. This process increases the diversity of minority samples and helps classifiers

learn more representative decision boundaries [6]. Previous studies have demonstrated that SMOTE can improve classification accuracy, precision, recall, and F1-score when dealing with imbalanced datasets [7,8].

Several extensions of SMOTE have also been proposed to improve the quality of synthetic samples. SMOTEBoost combines SMOTE with boosting algorithms to improve the classification of minority classes by generating synthetic samples during the boosting process. Selective Interpolation SMOTE (SISMOTE) focuses on generating samples in safe regions of the feature space to reduce the influence of noisy samples and overlapping classes [2]. Another variant, FC-SMOTE, utilizes fuzzy clustering to identify suitable oversampling regions and assign sampling weights according to data density, which helps address within-class imbalance and improve minority class representation [12].

In recent years, more advanced data augmentation techniques have been introduced for SER. Generative Adversarial Networks (GANs) have been employed to generate realistic synthetic emotional speech samples by learning the underlying data distribution [13]. Other approaches combine acoustic and textual augmentation techniques to preserve emotional consistency across different modalities [14]. Although these methods have shown promising results, they often require complex training procedures and considerable computational resources.

Despite the success of these oversampling and augmentation methods, several challenges remain. Synthetic samples may not always accurately represent the original data distribution, which can lead to class overlap or the introduction of noisy samples. Furthermore, improving minority class recognition may sometimes reduce the performance of majority classes, making it difficult to achieve balanced overall performance [9,10]. Therefore, combining oversampling techniques with ensemble learning has emerged as a promising direction for improving the robustness and effectiveness of Speech Emotion Recognition systems on imbalanced datasets.

C. Ensemble Learning for Imbalanced SER

Ensemble learning is a machine learning technique that combines the predictions of multiple classifiers to achieve better performance than a single model. The main idea is that different classifiers may learn different characteristics from the data, and combining their outputs can improve classification accuracy, robustness, and generalization ability [16]. Ensemble methods have been successfully applied to many classification problems, particularly when dealing with noisy and imbalanced datasets [14].

In Speech Emotion Recognition (SER), ensemble learning has received increasing attention because emotional speech data exhibit large variations in speaking style, recording environments, and emotion distributions. Individual classifiers may perform well on certain emotion classes but perform poorly on others. Therefore, combining multiple classifiers can help reduce the weaknesses of individual models and provide more stable performance across different emotion categories [17-20].

Several ensemble techniques have been investigated in SER research. Voting-based ensembles combine the predictions of different classifiers using majority voting or probability averaging. Boosting algorithms, such as AdaBoost and Gradient Boosting, iteratively train weak learners by focusing on previously misclassified samples to improve the overall prediction performance [21-25]. Bagging approaches, such as Random Forest, create multiple classifiers from different subsets of the training data and aggregate their predictions to reduce variance and improve stability. More recently, hybrid ensemble frameworks have been proposed to combine different classifiers and feature representations for emotion recognition tasks [26-30].

Ensemble learning is particularly useful for imbalanced datasets. Models trained on the original dataset usually preserve the natural distribution of the data and often achieve good performance on majority classes. In contrast, models trained on balanced datasets generated using oversampling techniques, such as SMOTE, generally provide better recognition of minority classes [5]. Combining these models enables the ensemble to exploit the advantages of both learning strategies and can lead to more balanced classification performance [31-35].

However, designing an effective ensemble framework remains a challenging task. The performance of an ensemble depends not only on the quality of the individual classifiers but also on the strategy used to combine their predictions [35-40]. Simple averaging may not always produce the best results, while confidence-based and adaptive weighting methods may behave differently depending on the class distribution and dataset characteristics [15]. Furthermore, the combination of ensemble learning and oversampling techniques in SER has not been extensively investigated, especially for highly imbalanced emotion datasets such as IEMOCAP [4].

Motivated by these observations, this study proposes a Hybrid SMOTE-Ensemble Framework that combines models trained on the original imbalanced dataset and on a SMOTE-balanced dataset using different ensemble strategies. The proposed framework aims to improve the recognition of minority emotions while maintaining robust overall classification performance.

III. Proposed Hybrid SMOTE-Ensemble Framework

This section presents the proposed Hybrid SMOTE-Ensemble Framework for Speech Emotion Recognition on imbalanced data. The framework consists of three main stages: dataset preparation, feature extraction, and hybrid ensemble learning. The IEMOCAP dataset is used to evaluate the effectiveness of the proposed framework because it is one of the most widely used benchmarks for Speech Emotion Recognition and contains a naturally imbalanced distribution of emotion classes [4].

The proposed framework first prepares the emotional speech data and performs emotion label mapping to obtain the final emotion categories. Acoustic and textual features are then extracted from speech recordings and transcriptions to construct multimodal feature representations. To address the class imbalance problem, two classifiers are trained independently: one using the original imbalanced dataset and another using a balanced dataset generated using the Synthetic Minority Over-

sampling Technique (SMOTE). Finally, different ensemble strategies are employed to combine the predictions of the two models and improve the recognition of minority emotions while maintaining robust overall classification performance.

A. Dataset and Preprocessing

The proposed framework was evaluated using the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset, which is one of the most widely used benchmark datasets for Speech Emotion Recognition (SER) [4]. The dataset contains approximately 12 hours of multimodal recordings collected from five sessions of dyadic interactions involving ten actors. Each session includes both scripted and improvised conversations and provides audio recordings, video streams, facial motion capture data, and textual transcriptions. The utterances are annotated by multiple evaluators according to different emotion categories, making the dataset suitable for emotion recognition studies.

The original IEMOCAP dataset contains several emotion labels, including anger, excitement, fear, sadness, surprise, frustration, happiness, disappointment, and neutrality. In this study, the emotion labels were processed to obtain six final emotion categories following the data preparation strategy adopted in previous studies [12]. The emotions "happy" and "excited" were merged into a single class because of their close semantic and expressive characteristics, and because the number of "happy" samples alone was relatively small. The class labeled as "others" was excluded because it represents vague emotional expressions that are difficult to classify consistently.

The emotion label "frustration" does not appear in the final six emotion classes reported by Sahu [12]. Therefore, utterances labeled as "frustration" were assigned to the "angry" category because both emotions share similar emotional characteristics and are often closely related in Speech Emotion Recognition studies. Other ambiguous labels, such as "disappointed", were excluded from the final dataset unless they were explicitly assigned to one of the selected emotion classes.

After label processing, the final dataset consisted of six emotion classes: Angry, Happy, Sad, Fear, Surprise, and Neutral. The corresponding class distribution was adopted from the dataset preparation procedure reported in [12], resulting in 860 samples for Angry, 1309 samples for Happy, 2327 samples for Sad, 1007 samples for Fear, 949 samples for Surprise, and 1385 samples for Neutral, with a total of 7837 utterances. The distribution of these classes remains imbalanced, where some emotions are represented by considerably fewer samples than others.

To evaluate the proposed framework, the dataset was randomly divided into training and testing sets using an 80:20 ratio. The same data split was used throughout all experiments to ensure a fair comparison among different models and ensemble strategies [12]. The original imbalanced training set was preserved for training the baseline model, while an additional balanced training set was generated using the Synthetic Minority Over-sampling Technique (SMOTE) for the proposed hybrid ensemble framework.

B. Feature Extraction

Feature extraction is an important stage in Speech Emotion Recognition because the quality of extracted features directly affects the performance of classification models. In this study, both acoustic and textual features were utilized to construct a multimodal representation of emotional speech following the feature extraction strategy described in [12].

For the audio modality, handcrafted acoustic features were extracted from each speech utterance. These features include pitch, harmonics, energy, pause-related information, and statistical descriptors that capture the emotional characteristics of speech signals. Prosodic and spectral features are widely used in SER because they provide useful information related to speaking style, emotional intensity, and vocal characteristics [3].

For the textual modality, the speech transcriptions provided in the IEMOCAP dataset were transformed into numerical feature vectors using the Term Frequency-Inverse Document Frequency (TF-IDF) technique. TF-IDF was employed to capture the importance of words according to their frequency in a document and their occurrence across the entire corpus. The extracted text features provide complementary linguistic information that can help distinguish different emotional states.

The extracted audio and text features were then combined to form a multimodal feature representation. This combination allows the framework to exploit complementary information from both modalities, since emotions are often reflected not only in acoustic patterns but also in the semantic content of speech.

It should be noted that some implementation details related to audio feature extraction, such as frame size, hop length, and filter configurations, were not explicitly reported in the original framework [12]. Therefore, commonly adopted settings and library default parameters were used during feature extraction while maintaining consistency across all experiments. The same feature extraction procedure was applied to both the original and SMOTE-balanced datasets to ensure a fair comparison among the evaluated models.

C. Model Architectures and Training

To address the class imbalance problem in Speech Emotion Recognition, two XGBoost classifiers were trained independently using different training datasets. The first classifier was trained using the original imbalanced training set, while the second classifier was trained using a balanced dataset generated using the Synthetic Minority Over-sampling Technique (SMOTE). This strategy allows the framework to benefit from the strengths of both learning schemes, where the original model preserves the natural data distribution and the SMOTE-based model improves the recognition of minority emotion classes.

XGBoost was selected as the base classifier because of its strong classification performance, robustness, and ability to handle high-dimensional feature spaces efficiently. XGBoost is an ensemble learning algorithm based on gradient boosting decision trees, where multiple weak learners are sequentially combined to produce a strong predictive model. The algorithm has demonstrated competitive performance in various classification tasks and has been successfully applied in Speech Emotion Recognition applications [1].

For the original model, the extracted multimodal features were directly used to train the XGBoost classifier using the imbalanced training dataset. In parallel, SMOTE was applied to the training set to generate synthetic samples for minority emotion classes and create a balanced training dataset. The same XGBoost architecture and training settings were then applied to the SMOTE-balanced dataset to ensure a fair comparison between the two models.

After training the individual classifiers, their predictions were combined using a hybrid ensemble strategy. Three ensemble methods were investigated in this study: Simple Averaging, Confidence-Based Selection, and Adaptive Weighting. In the Simple Averaging method, the prediction probabilities of the two models were averaged, and the emotion class with the highest probability was selected as the final prediction. In the Confidence-Based Selection method, the prediction of the model with the higher confidence score was chosen. The Adaptive Weighting approach assigns different weights to the two classifiers according to their performance, allowing the ensemble to balance the contributions of the original and SMOTE-based models.

The proposed Hybrid SMOTE-Ensemble Framework combines the advantages of the original and balanced datasets while reducing the limitations of using either approach independently. The overall architecture of the proposed framework is illustrated in Fig. 1.

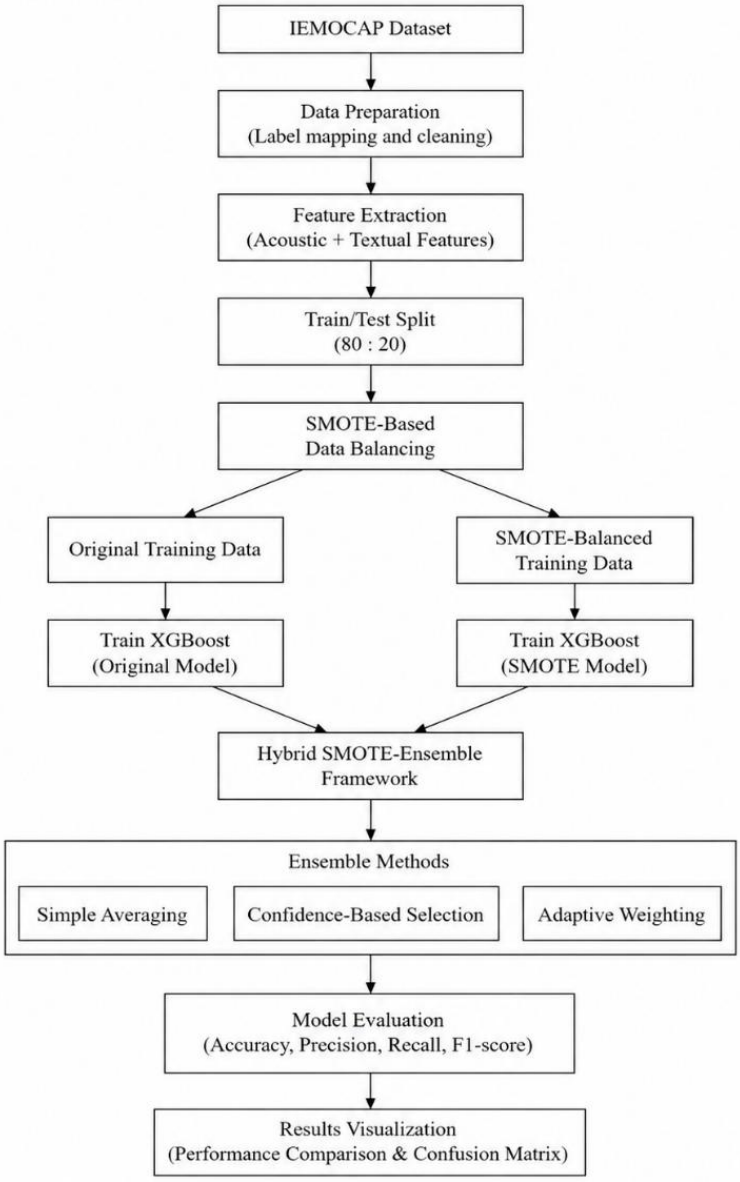


Fig. 1. Overall architecture of the proposed hybrid SMOTE-ensemble framework for speech emotion recognition.

IV. Experimental Setup and Enhancement Strategy

A. Motivation for Enhancement

The IEMOCAP dataset exhibits an imbalanced distribution of emotion classes, where some emotions are represented by a large number of samples while others contain relatively few examples. This imbalance can bias classification models toward majority classes and reduce their ability to recognize minority emotions accurately. In the original study, the imbalance problem was addressed using simple random upsampling for the "fear" and "surprise" classes, while the "happy" and "excited" classes were merged into a single category [12]. Although these strategies help reduce the imbalance problem, they have certain limitations.

Random oversampling increases the number of minority class samples by duplicating existing instances. While this approach is simple to implement, it does not introduce new information into the training set and may increase the risk of overfitting. As a result, the classifier may perform well on the duplicated training samples but fail to generalize effectively to unseen data. Furthermore, merging emotion classes simplifies the classification problem but may reduce the emotional diversity of the dataset and limit the model's ability to distinguish between similar emotions.

To overcome these limitations, this study employs the Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic samples for minority emotion classes. Unlike random oversampling, SMOTE creates new samples by interpolating between neighboring minority instances, which increases the diversity of the training data and provides better representation of underrepresented classes [5]. This property makes SMOTE a suitable choice for Speech Emotion Recognition tasks involving imbalanced datasets.

In addition to SMOTE, this study introduces a hybrid ensemble strategy that combines models trained on the original imbalanced dataset and on the SMOTE-balanced dataset. The motivation behind this approach is that the original model preserves the natural distribution of the data and performs well on majority classes, whereas the SMOTE-based model is expected to improve the recognition of minority emotions. By combining the predictions of both models, the proposed framework aims to achieve more balanced and robust emotion recognition performance across all emotion classes.

B. Implementing the Hybrid XGBoost Ensemble with SMOTE

To reduce the limitations of simple upsampling and improve model performance, this study uses a hybrid XGBoost ensemble strategy with SMOTE. The method combines the predictions of two separately trained XGBoost models: one trained on the original imbalanced data and another trained on the SMOTE-balanced data.

The first XGBoost model was trained on the original training dataset. This model keeps the natural class distribution and is expected to perform better on the majority emotion classes. The second XGBoost model was trained on a balanced version of the training dataset after applying SMOTE. This model is expected to improve the recognition of minority emotion classes because it is trained with a more balanced class distribution.

SMOTE generates synthetic samples for minority classes by creating new points between existing minority samples and their nearest neighbors in the feature space [5]. In this study, SMOTE was applied using random_state = 42 for reproducibility, and the default k_neighbors = 5 was used. Fig. 2 shows the class distribution before and after applying SMOTE, where all classes were balanced to the same number of samples.

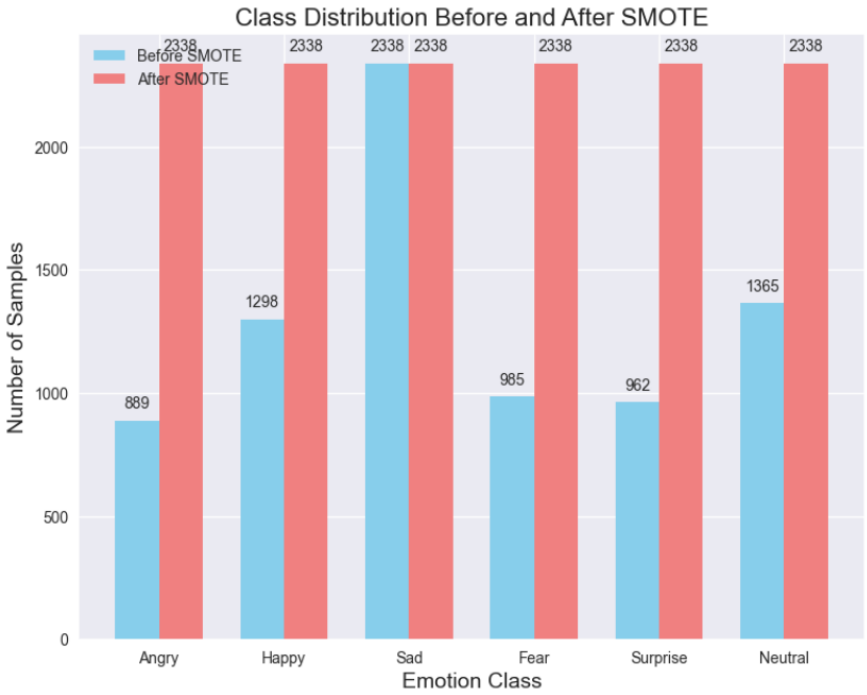


Fig. 2. Class distribution before and after applying SMOTE.

Both XGBoost models were trained using the same parameters. The selected parameters were max_depth = 7, learning_rate = 0.008, objective = multi:softprob, n_estimators = 600, subsample = 0.8, num_class = 6, and booster = gbtree.

The prediction probabilities from the two XGBoost models were combined using three ensemble strategies. The first strategy was Simple Averaging, where the probabilities from the original model and the SMOTE-based model were averaged, and the class with the highest average probability was selected. The second strategy was Confidence-Based Selection, where the final prediction was taken from the model with the highest confidence score. Confidence was calculated as the highest probability assigned to any class by each model.

The third strategy was Adaptive Weighted Ensemble. In this method, different weights were assigned to the original and SMOTE-based models according to the class distribution in the training data. For minority classes, defined as classes with less than 15% of the training samples, the original model was assigned a weight of 0.3 and the SMOTE model was assigned a weight of 0.7. For majority classes, defined as classes with more than 25% of the training samples, the original model was assigned a weight of 0.7 and the SMOTE model was assigned a weight of 0.3. For the remaining classes, both models were assigned equal weights of 0.5.

The purpose of this hybrid ensemble is to combine the strengths of both models. The original XGBoost model can preserve performance on majority classes, while the SMOTE-based XGBoost model can improve the recognition of minority classes. By combining both models, the proposed framework aims to improve the overall Speech Emotion Recognition performance on imbalanced data.

C. Implementation of the Hybrid Ensemble

The proposed Hybrid SMOTE-Ensemble Framework was implemented in Python using the XGBoost and imbalanced-learn libraries. Two XGBoost classifiers were trained independently: one using the original imbalanced training dataset and the other using the SMOTE-balanced dataset. The Audio+Text modality was adopted using the handcrafted acoustic features and TF-IDF text features described in the previous sections. The dataset was divided into 80% for training and 20% for testing, where the testing set remained unchanged throughout all experiments. The prediction probabilities generated by the two classifiers were combined using the three ensemble strategies discussed in Section IV-B, namely Simple Averaging, Confidence-Based Selection, and Adaptive Weighting. The effectiveness of the proposed framework was evaluated using Accuracy, F1-score, Precision, Recall, and confusion matrices.

Fig. 3 presents the performance comparison of the Hybrid XGBoost framework using five different techniques: XGB Original, XGB SMOTE, Simple Average Ensemble, Confidence-Based Ensemble, and Adaptive Weighted Ensemble. The comparison is based on Accuracy, F1-score, Precision, and Recall. The results show that the SMOTE-based model improves both recall and F1-score compared with the original XGBoost model. Furthermore, the ensemble approaches provide additional improvements in overall performance. Among the ensemble methods, the Adaptive Weighted Ensemble demonstrates a balanced improvement across all evaluation metrics, while the Simple Average Ensemble achieves competitive performance with a simpler integration strategy.

Fig. 4 illustrates the confusion matrix of the XGBoost model trained on the original imbalanced dataset. The model classifies emotions into six categories: angry (ang), happy (hap), neutral (neu), sad (sad), frustrated (fru), and excited (exc). The results indicate that the model performs best on the neutral class, which records the highest number of correctly classified samples. However, some confusion can be observed between emotionally similar classes, particularly between happy and neutral emotions. Although misclassifications occur across several categories, the model maintains acceptable classification performance overall. The darker values along the diagonal indicate higher numbers of correct predictions.

Fig. 3. Hybrid XGBoost model performance metrics with different applied techniques.

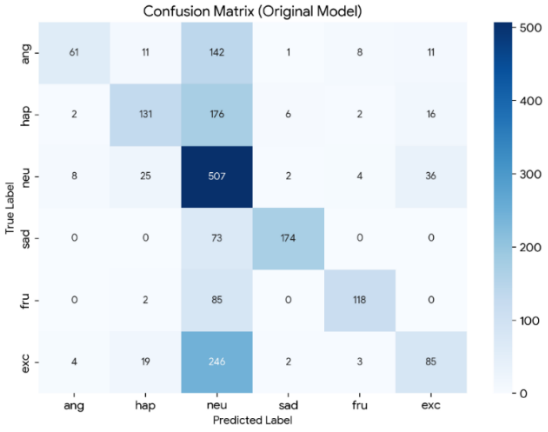


Fig. 4. Confusion matrix of the XGBoost original model.

Fig. 5 presents the confusion matrix of the Simple Average Ensemble. Similar to the original model, the ensemble classifies six emotion categories: angry, happy, neutral, sad, frustrated, and excited. The diagonal elements represent correctly classified samples, while the off-diagonal elements correspond to misclassified instances. The neutral emotion achieves the highest number of correct classifications, whereas confusion remains noticeable between happy and neutral emotions. Overall, the ensemble demonstrates improved consistency across emotion categories and provides more balanced classification performance than the original model.

Fig. 6 shows the confusion matrix of the XGBoost model trained on the SMOTE-balanced dataset. The application of SMOTE improves the representation of minority emotion classes and enhances their recognition capability. The model performs well in recognizing neutral, happy, and excited emotions, while some confusion remains between neutral and sad emotions. The results indicate that balancing the training data contributes positively to the classification of underrepresented classes and improves the overall robustness of the model.

Fig. 7 illustrates the recall values achieved by the Original XGBoost model, the SMOTE-based XGBoost model, and the Simple Average Ensemble across the six emotion classes. The Original model achieves the highest recall for the neutral class, while the SMOTE-based model improves the recall of several minority classes. The Simple Average Ensemble provides more consistent recall values across most classes and achieves a better balance between majority and minority emotions. Nevertheless, the angry class remains the most challenging category for all models, indicating that further improvements in feature representation or data augmentation may still be beneficial.

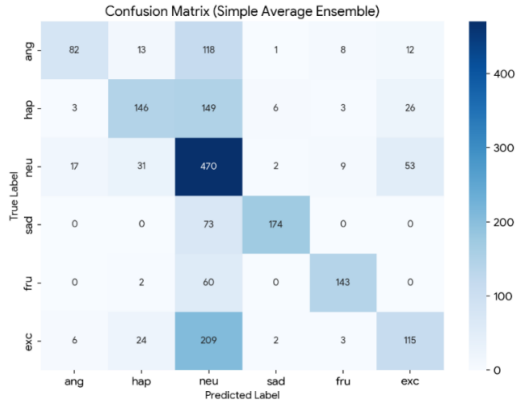


Fig. 5. Confusion matrix of the simple average ensemble technique.

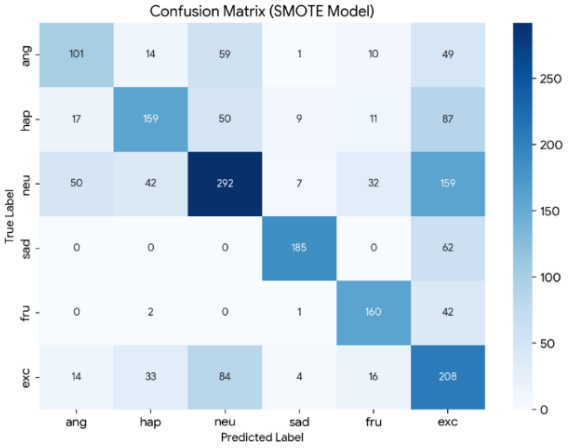


Fig. 6. Confusion matrix of the SMOTE model technique.

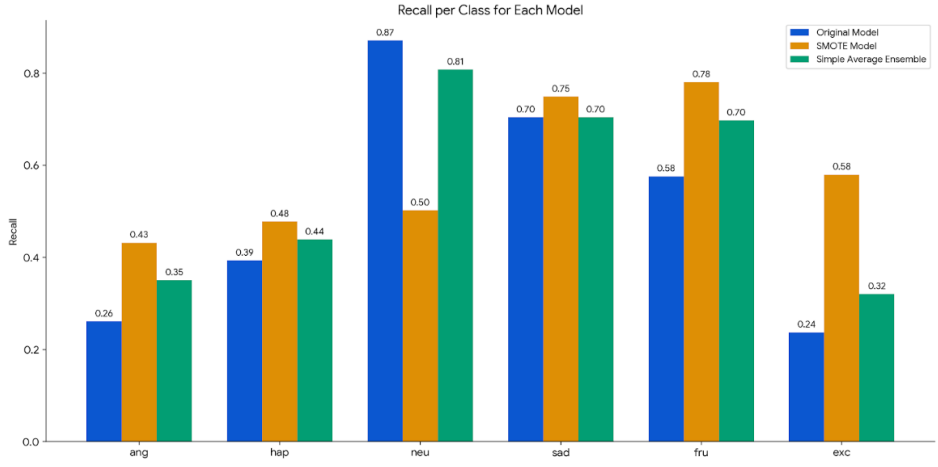


Fig. 7. Hybrid XGBoost model recall per class for each model.

V. Additional Evaluation Using LinearSVC

To further investigate the effectiveness of the proposed Hybrid SMOTE-Ensemble Framework, an additional evaluation was conducted using a Linear Support Vector Classifier (LinearSVC). The objective of this experiment was to examine whether the advantages of data balancing and ensemble learning can be generalized to a different classifier architecture. Similar to the XGBoost experiments, the main challenge lies in the imbalanced distribution of emotion classes, where minority emotions are often underrepresented and difficult to classify accurately.

Two enhancement approaches were investigated. The first approach applied SMOTE directly to the entire training set before training a single LinearSVC classifier. The second approach employed a Hybrid Weighted Ensemble, where two LinearSVC models were trained independently: one using the original imbalanced dataset and the other using the SMOTE-balanced dataset. The predictions of both models were then combined using adaptive weights that assign greater importance to the SMOTE model for minority classes and greater importance to the original model for majority classes.

A. Performance Comparison

The results demonstrate that applying SMOTE changes the behaviour of the LinearSVC classifier. Although the overall accuracy remained close to the baseline model, SMOTE improved the model's ability to recognize minority emotion classes by increasing recall. This improvement, however, was accompanied by a slight reduction in precision and F1-score for some majority classes, which is a common trade-off when balancing highly imbalanced datasets.

The Hybrid Weighted Ensemble achieved the best overall performance among the evaluated LinearSVC models. The ensemble obtained an accuracy score of 0.629, outperforming both the baseline and the SMOTE-only models. This result indicates that combining the strengths of the original and SMOTE-based classifiers leads to a more balanced and robust classification system. In particular, the ensemble achieved perfect recall for the Fear class, demonstrating its effectiveness in recognizing minority emotions.

Fig. 8 illustrates the effect of SMOTE on the class distribution and presents the performance metrics of the LinearSVC models. The results show that the Hybrid Ensemble achieves the highest accuracy, F1-score, and recall, while maintaining competitive precision. In contrast, the original LinearSVC model achieves the highest precision but exhibits lower recall, especially for minority classes. The SMOTE-based model improves recall but sacrifices some precision, whereas the Hybrid Ensemble provides the most balanced performance.

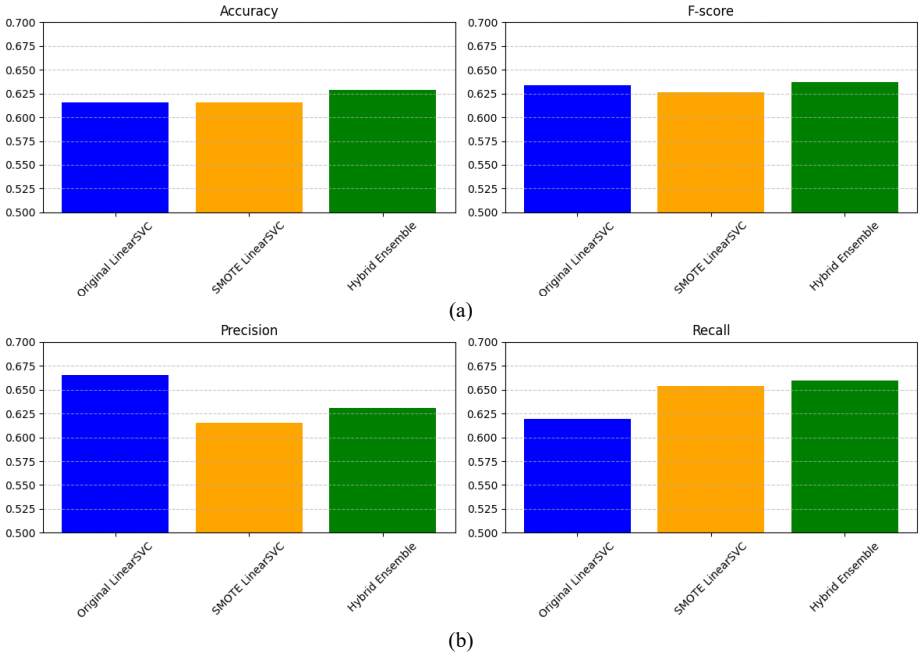


Fig. 8. (A) Class distribution before and after applying SMOTE. (B) Performance comparison of Original LinearSVC, SMOTE LinearSVC, and Hybrid Ensemble models.

B. Confusion Matrix Analysis

The confusion matrices shown in Fig. 9 provide a detailed comparison of the three LinearSVC configurations. The original model demonstrates high precision for majority classes but exhibits considerable misclassification for minority emotions. Several minority classes are frequently predicted as the dominant classes, indicating that the imbalanced training data limits the classifier's ability to learn representative decision boundaries.

After applying SMOTE, the recognition of minority classes improves noticeably. The SMOTE-based model correctly identifies more samples from underrepresented classes; however, the increase in synthetic samples also introduces additional confusion among some majority classes, which leads to a reduction in overall precision.

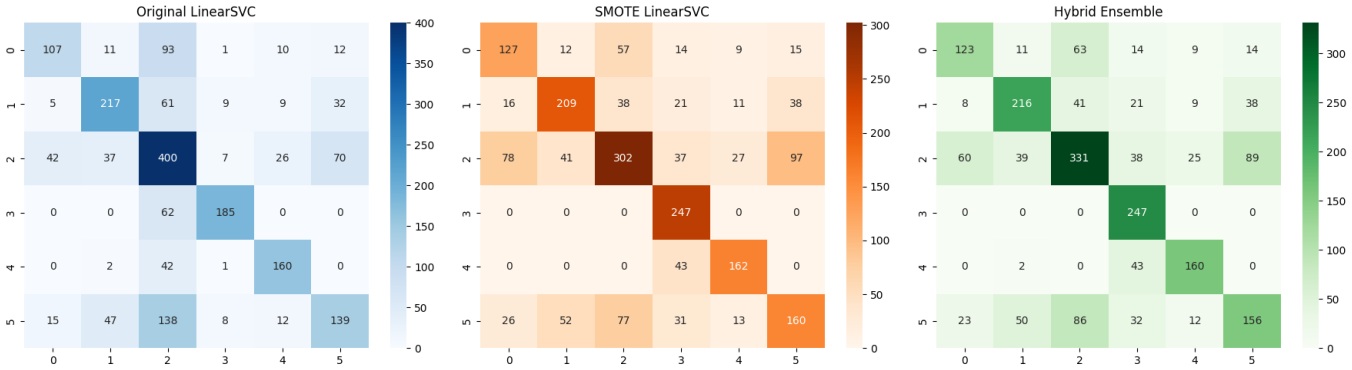


Fig. 9. Confusion matrices of the LinearSVC models: (A) Original model, (B) SMOTE-based model, and (C) Hybrid weighted ensemble.

The Hybrid Ensemble achieves the best balance between these two approaches. The confusion matrix shows a reduction in misclassification errors while maintaining high numbers of correctly classified samples across most emotion categories. This demonstrates that combining the predictions of the original and SMOTE-based models provides a more robust solution to the class imbalance problem than either model individually.

C. Recall Analysis

Fig. 10 compares the recall values obtained by the Original LinearSVC model, the SMOTE-based LinearSVC model, and the Hybrid Weighted Ensemble for each emotion class. The results show that both SMOTE and the Hybrid Ensemble improve recall for minority classes compared with the baseline model.

The Hybrid Ensemble provides the most consistent recall across all emotion classes and achieves the best overall balance between minority and majority class recognition. The most notable result is the perfect recall achieved for Class 3 (Fear), indicating that the proposed weighting strategy effectively enhances the recognition of difficult and underrepresented emotions.

Overall, the additional experiments using LinearSVC confirm the effectiveness of the Hybrid SMOTE-Ensemble concept. Although the XGBoost framework remains the primary contribution of this work, the LinearSVC results demonstrate that the proposed strategy can also improve the performance of other classifiers when dealing with imbalanced Speech Emotion Recognition datasets.

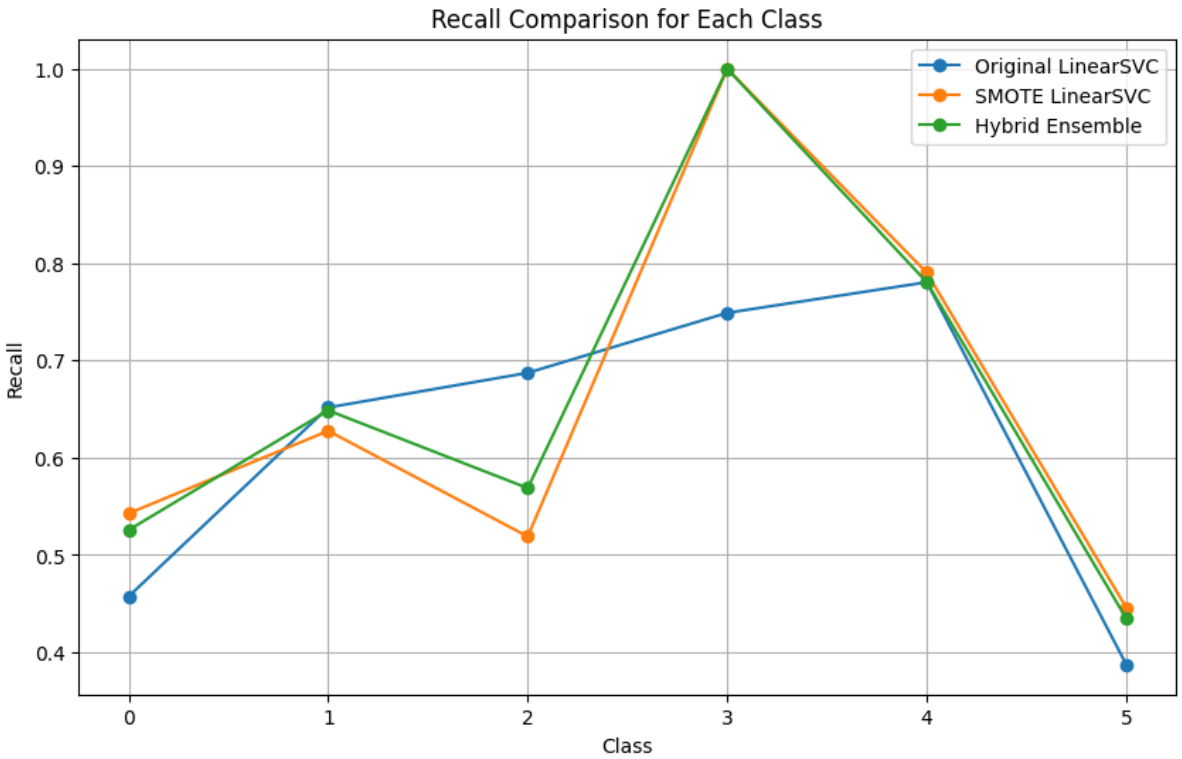


Fig. 10. Recall comparison of the Original LinearSVC model, SMOTE-based LinearSVC model, and Hybrid Weighted Ensemble for each emotion class.

VI. Experimental Results and Discussion

This section presents the experimental results obtained from the reproduction of the original study and the proposed enhancement methods. First, the reproduced results are compared with those reported by Sahu et al. [12] to evaluate the reproducibility of the original framework. Subsequently, the performance of the proposed Hybrid SMOTE-Ensemble Framework is analyzed and compared with the baseline models using different evaluation metrics and confusion matrices.

A. Reproduction Results

The objective of the reproduction experiments was to investigate whether the performance reported by Sahu et al. [12] could be achieved using the publicly available information and implementation details described in their study. The reproduced models were evaluated under three modalities: Audio-only, Text-only, and Audio+Text. Their performance was assessed using Accuracy, F1-score, Precision, and Recall.

Fig. 11 presents the confusion matrix of the Audio-only model reported in [12]. The model classifies emotions into six categories: anger (ang), happiness (hap), sadness (sad), fear (fea), surprise (sur), and neutral (neu). The results show that the model performs well in recognizing sadness and fear, whereas more confusion is observed among anger, happiness, and neutral emotions. This behaviour reflects the inherent difficulty of distinguishing emotions with similar acoustic characteristics.

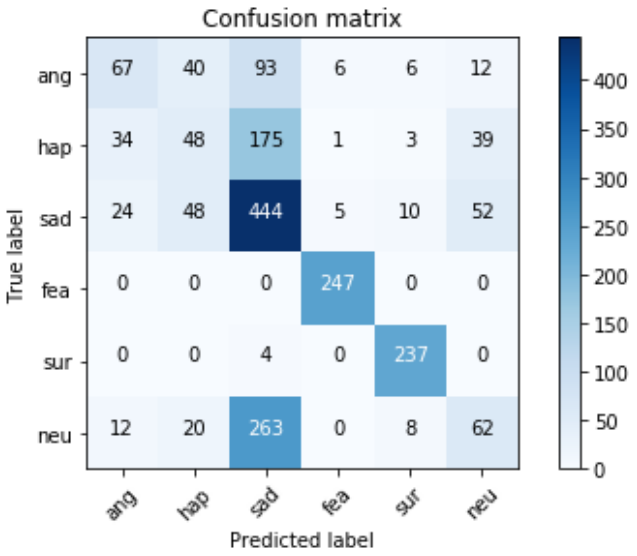


Fig. 11. Results reported by [12] for the audio-only model.

Fig. 12 illustrates the confusion matrix of the Text-only model. The model demonstrates strong performance for sadness and fear, while several misclassifications occur among happy, sad, and neutral emotions. In particular, neutral emotions are frequently predicted as sadness, indicating that textual information alone may not always provide sufficient discriminative features for emotion recognition.

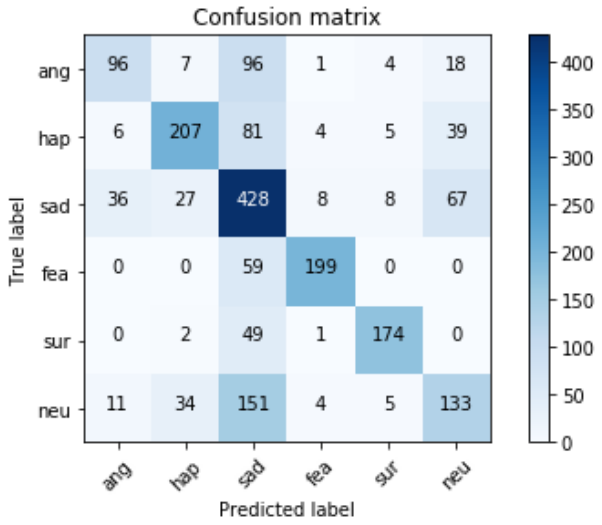


Fig. 12. Results reported by [12] for the text-only model.

Fig. 13 presents the confusion matrix of the Audio+Text model. Compared with the unimodal approaches, the multimodal model provides better recognition performance for most emotion categories by exploiting complementary information from both audio and textual modalities. The model achieves the best performance for sadness and surprise, although some confusion remains between anger and sadness as well as between happy and neutral emotions.

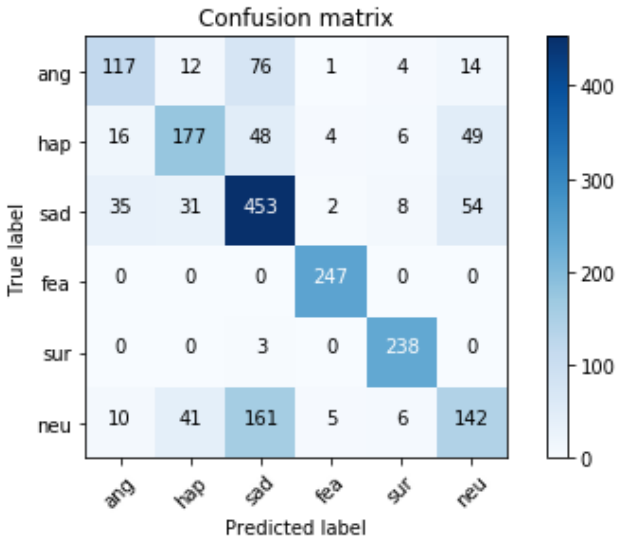


Fig. 13. Results reported by [12] for the audio+text model.

Overall, the reproduced results are in close agreement with those reported by Sahu et al. [12], with most differences remaining within ±1%. Although slight variations are observed for some models and modalities, the general performance trends remain consistent. These differences may be attributed to ambiguities in feature extraction settings, differences in library versions, and the stochastic nature of model training and data partitioning.

B. Results from Enhancements Using Hybrid XGBoost Ensembles

The proposed Hybrid SMOTE-Ensemble Framework was evaluated using the Audio+Text modality. The framework consists of two XGBoost classifiers trained independently on the original imbalanced dataset and the SMOTE-balanced dataset. Their predictions were combined using three ensemble strategies: Simple Averaging, Confidence-Based Selection, and Adaptive Weighting. The performance of these models is summarized in Table I.

TABLE I. PERFORMANCE OF HYBRID XGBOOST ENSEMBLE ENHANCEMENT (AUDIO+TEXT MODALITY)

Model/Ensemble Method	Accuracy (%)	F1-score (%)	Precision (%)	Recall (%)
XGB Original (Baseline)	54.9	54.9	71.9	50.7
XGB SMOTE	56.4	59.4	62.1	58.7
Simple Average Ensemble	57.7	59.3	70.5	55.3
Confidence-Based Ensemble	57.0	58.5	70.9	54.4
Adaptive Weighted Ensemble	57.4	58.9	71.1	54.9

The results demonstrate that training XGBoost on SMOTE-balanced data improves the classification performance compared with the original imbalanced training set. The XGB SMOTE model achieved an accuracy of 56.4% and an F1-score of 59.4%, compared with 54.9% accuracy and 54.9% F1-score obtained by the original XGBoost model. This improvement indicates that balancing the training data enables the classifier to better recognize underrepresented emotion classes.

The ensemble methods further improve the overall performance. The Simple Average Ensemble achieved the highest accuracy of 57.7%, while the Adaptive Weighted Ensemble achieved competitive results with an accuracy of 57.4% and an F1-score of 58.9%. The Confidence-Based Ensemble also outperformed the baseline model and achieved balanced performance across the evaluation metrics.

Overall, the results confirm that the proposed Hybrid SMOTE-Ensemble Framework improves emotion recognition performance on imbalanced data. The combination of SMOTE-based data balancing and ensemble learning provides a more robust and balanced classifier compared with a single model trained on the original dataset.

C. Hybrid LinearSVC Ensemble Enhancement Results

To further investigate the effectiveness of the proposed enhancement strategy, additional experiments were conducted using a Linear Support Vector Classifier (LinearSVC). Three different configurations were evaluated: the Original LinearSVC trained on the imbalanced dataset, the SMOTE-based LinearSVC trained on the balanced dataset, and the Hybrid Weighted Ensemble combining the predictions of both models.

The results presented in Tables II and III demonstrate that applying SMOTE improves the recall of minority emotion classes, although this improvement is accompanied by a slight reduction in precision for some majority classes. The Original LinearSVC model achieved the highest precision; however, it suffered from lower recall, particularly for underrepresented emotions.

TABLE II. PERFORMANCE METRICS OF TECHNIQUES USED WITH LINEARSVC

Metric	Original LinearSVC	SMOTE LinearSVC	Hybrid Ensemble
Accuracy	0.616	0.616	0.629
F-score	0.634	0.626	0.637
Precision	0.665	0.615	0.631
Recall	0.619	0.654	0.660

TABLE III. RECALL VALUES FOR EACH CLASS OF TECHNIQUES USED WITH LINEARSVC

Class	Original LinearSVC	SMOTE LinearSVC	Hybrid Ensemble
0	0.457	0.543	0.526
1	0.652	0.628	0.649
2	0.687	0.519	0.569
3	0.749	1.000	1.000
4	0.780	0.790	0.780
5	0.387	0.446	0.435

The Hybrid Weighted Ensemble achieved the best overall performance among the evaluated LinearSVC models. The ensemble obtained the highest accuracy and F1-score while maintaining balanced precision and recall values. These results indicate that combining the original and SMOTE-based classifiers enables the model to exploit the strengths of both

approaches and provides a more robust solution for imbalanced Speech Emotion Recognition.

Furthermore, the recall analysis confirms that the Hybrid Ensemble consistently improves the recognition of minority emotion classes. The results suggest that the proposed Hybrid SMOTE-Ensemble strategy is not limited to XGBoost and can also enhance the performance of other machine learning classifiers such as LinearSVC.

Overall, the additional LinearSVC experiments support the effectiveness and generalization capability of the proposed enhancement framework for imbalanced Speech Emotion Recognition tasks.

D. Analysis of Results

The reproduction experiments demonstrated a high degree of agreement with the results reported by Sahu et al. [12]. Most evaluation metrics differed by less than 1%, indicating that the proposed reproduction methodology successfully replicated the main findings of the original study. The small differences observed for some models may be attributed to ambiguities in feature extraction settings, differences in library implementations, and the inherent stochastic nature of machine learning and deep learning training processes.

The results obtained from the proposed Hybrid SMOTE-Ensemble Framework further demonstrate the effectiveness of combining data balancing and ensemble learning for Speech Emotion Recognition on imbalanced datasets. Training XGBoost on SMOTE-balanced data improved both Accuracy and F1-score compared with the original imbalanced model. In particular, the F1-score increased from 54.9% to 59.4%, indicating that balancing the training data enables the classifier to better recognize underrepresented emotion classes.

The ensemble methods provided additional improvements over the baseline XGBoost model. Among the evaluated approaches, the Simple Average Ensemble achieved the highest accuracy of 57.7%, while the Adaptive Weighted Ensemble achieved competitive results with balanced Precision, Recall, and F1-score values. These findings suggest that combining classifiers trained on original and balanced datasets produces a more robust model than relying on a single classifier.

The additional experiments using LinearSVC further support this observation. Although the overall improvement was less pronounced than that obtained with XGBoost, the Hybrid Weighted Ensemble consistently outperformed the baseline LinearSVC model and improved the recognition of minority emotion classes. These results indicate that the proposed Hybrid SMOTE-Ensemble strategy is not limited to a specific classifier and can be generalized to other machine learning models.

E. Robustness and Limitations

The robustness of the experiments was improved by fixing the random seed (`random_state = 42`) during train-test splitting, SMOTE oversampling, and model training whenever possible. This ensures that the experiments can be reproduced and that the obtained results are not caused by random variations in data partitioning or model initialization.

The reproduced results remained close to those reported by Sahu et al. [12], demonstrating that the original findings are generally robust to minor differences in implementation details and software environments. In addition, the proposed Hybrid SMOTE-Ensemble Framework consistently improved performance across different evaluation metrics, indicating that the enhancement strategy is stable and effective for imbalanced Speech Emotion Recognition tasks.

Despite these encouraging results, several limitations should be acknowledged. First, the enhancement strategy was evaluated only using XGBoost and LinearSVC classifiers. The impact of the proposed framework on other classifiers, such as Random Forest, SVM, MLP, and LSTM, remains an open research direction. Second, some assumptions were required during the reproduction process because the original study did not explicitly specify all implementation details, particularly those related to feature extraction and label preprocessing. Consequently, the reproduced experiments may not represent an exact replication of the original experimental setup.

F. Ethical Considerations and Broader Impact

Speech Emotion Recognition systems have the potential to improve human-computer interaction, customer service, healthcare, and assistive technologies. However, incorrect emotion prediction may negatively affect users, particularly in sensitive applications such as mental health assessment or security systems. Therefore, developing robust and fair emotion recognition systems is an important ethical consideration.

Class imbalance is one of the factors that may introduce bias into Speech Emotion Recognition models because minority emotions are often underrepresented during training. The proposed Hybrid SMOTE-Ensemble Framework attempts to mitigate this issue by improving the recognition of minority emotion classes through balanced data generation and ensemble learning. Nevertheless, synthetic data generated by oversampling techniques should be used carefully to avoid introducing artificial patterns or hidden biases into the training process.

Finally, this study highlights the importance of reproducibility and transparent reporting in machine learning research. Providing clear implementation details and thoroughly evaluating enhancement strategies contribute to the development of more reliable, fair, and trustworthy Speech Emotion Recognition systems.

VII. Conclusion

This study had two main objectives. The first was to reproduce the multimodal Speech Emotion Recognition framework proposed by Sahu et al. [12] and evaluate the reproducibility of their reported results. The second was to enhance the original framework by introducing a Hybrid SMOTE-Ensemble strategy based on XGBoost and LinearSVC classifiers to address the class imbalance problem in the IEMOCAP dataset.

The reproduction experiments demonstrated a high level of agreement with the original study, with most evaluation metrics differing by less than 1%.

These results confirm the validity of the original findings and highlight the importance of transparent reporting and reproducible research practices in Speech Emotion Recognition.

The enhancement experiments showed that balancing the training data using SMOTE improves the performance of emotion classification models, particularly in terms of F1-score and Recall. Furthermore, combining models trained on the original and SMOTE-balanced datasets through ensemble learning provided additional improvements in overall performance. The proposed Hybrid SMOTE-Ensemble Framework achieved more balanced and robust performance than the baseline models, demonstrating the effectiveness of integrating data balancing with ensemble learning for imbalanced Speech Emotion Recognition tasks.

The additional experiments using LinearSVC further confirmed the generalization capability of the proposed framework. Although the performance improvements varied across classifiers, the Hybrid Ensemble consistently provided a better balance between Precision, Recall, and F1-score than individual models.

Overall, this study demonstrates that advanced oversampling techniques and ensemble learning strategies can effectively mitigate class imbalance and improve Speech Emotion Recognition performance. The findings also emphasize the importance of reproducibility and systematic experimentation in developing reliable and robust emotion recognition systems.

VIII. Future Research and Implementations

Several directions can be explored to further improve the proposed Hybrid SMOTE-Ensemble Framework:

- Investigate the effectiveness of the proposed framework using additional classifiers such as Random Forest, Support Vector Machine, Multi-Layer Perceptron, and Long Short-Term Memory networks.
- Evaluate the Hybrid SMOTE-Ensemble strategy using different modalities, including Audio-only and Text-only settings, as well as more advanced multimodal fusion approaches.
- Explore more sophisticated oversampling techniques, such as SISMOTE and FC-SMOTE, or investigate generative approaches based on Generative Adversarial Networks (GANs) for data augmentation.
- Perform more extensive hyperparameter optimization for both the base classifiers and the ensemble strategies to further improve classification performance.
- Investigate the proposed framework on other Speech Emotion Recognition datasets with different class distributions to evaluate its robustness and generalization capability.

These directions may contribute to the development of more accurate, robust, and fair Speech Emotion Recognition systems capable of handling the challenges associated with imbalanced emotional datasets.

References

[1] Abdelhamid, A. A., El-Kenawy, E. S. M., Alotaibi, B., Amer, G. M., Abdelkader, M. Y., Ibrahim, A., & Eid, M. M. (2022). Robust speech emotion recognition using CNN+ LSTM based on stochastic fractal search optimization algorithm. IEEE Access, 10, 49265–49284.

[2] Akbani, R., Kwek, S., & Japkowicz, N. (2004). Applying support vector machines to imbalanced datasets. In J.-F. Boulicaut, F. Esposito, F. Giannotti, & D. Pedreschi (Eds.), Lecture Notes in Computer Science: Vol. 3201. Machine Learning: ECML 2004 (pp. 39–50). Springer-Verlag Berlin Heidelberg.

[3] Atmaja, B. T., & Sasou, A. (2022). Effects of data augmentations on speech emotion recognition. Sensors, 22(16), 5941.

[4] Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation, 42(4), 335–359. https://doi.org/10.1007/s10579-008-9076-6

[5] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.

[6] Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W. (2003). SMOTEBoost: Improving prediction of the minority class in boosting. In Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2003) (pp. 107–119).

[7] Fan, Z., Wu, Y., Zhou, C., Zhang, X., & Tao, Z. (2021). Class-imbalanced voice pathology detection and classification using fuzzy cluster oversampling method. Applied Sciences, 11(8), 3450.

[8] Fitzgerald, D. (2010). Harmonic/percussive separation using median filtering. In Proceedings of the 13th International Conference on Digital Audio Effects (DAFx-10).

[9] Liang, X., Jiang, H., Xu, W., & Zhou, Y. (2023). Gaussian-smoothed imbalance data improves speech emotion recognition. arXiv preprint arXiv:2302.08650.

[10] Liu, Z.-T., Wu, B.-H., Li, D.-Y., Xiao, P., & Mao, J.-W. (2020). Speech emotion recognition based on selective interpolation synthetic minority over-sampling technique in small sample environment. Sensors, 20(8), 2297.

[11] Meng, T., Shou, Y., Ai, W., Yin, N., & Li, K. (2023). Deep imbalanced learning for multimodal emotion recognition in conversations. arXiv preprint arXiv:2312.06337.

[12] Sahu, G. (2019). Multimodal speech emotion recognition and ambiguity resolution. arXiv preprint arXiv:1904.06022v1 [cs.LG].

[13] Sondhi, M. M. (1968). New methods of pitch extraction. IEEE Transactions on Audio and Electroacoustics, 16(2), 262–266.

[14] Weiss, G. M. (2004). Mining with rarity: A unifying framework. ACM SIGKDD Explorations Newsletter, 6(1), 7–19.

[15] Zhao, J., Dong, W., Shi, L., Qiang, W., Kuang, Z., Xu, D., & An, T. (2022). Multimodal feature fusion method for unbalanced sample data in social network public opinion. Sensors, 22(15), 5528.

[16] Zhou, Z.-H., & Liu, X.-Y. (2006). Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Transactions on Knowledge and Data Engineering, 18(1), 63–77.

[17] Mashagba, H. A., Rahim, H. A., Mohd Yasin, M. N., Jamaluddin, M. H., Islam, M. T., Abdulkawi, W. M., ... & Al-Bawri, S. S. (2024). Four-port dual-band textile MIMO antenna for biomedical health monitoring systems: on-body mutual coupling reduction characterization. *Physica Scripta*, *99*(9), 095026.

[18] Owida, H. A., El-Fattah, I. A., Abuowaida, S., Alshdaifat, N., Mashagba, H. A., Abd Aziz, A. B., ... & Al-Bawri, S. S. (2025). A deep learning-based dual-branch framework for automated skin lesion segmentation and classification via dermoscopic Images. Scientific Reports, 15(1), 37823.

[19] Rahim, H. A., Mashagba, H. A., Yahaya, N. Z., Yasin, M. N. M., Jamaluddin, M. H., Jusoh, M., ... & Abdulmalek, M. (2021, December). 5G Millimeter Wave Wearable Antenna: State-Of-the-Art and Current Challenges. In *2021 IEEE Asia-Pacific Conference on Applied Electromagnetics (APACE)* (pp. 1-4). IEEE.

[20] A. Al-Mashagba, L., Sumari, P., Mashagba, M., Al Abri, F., Mashagba, H. A., Abd Aziz, A. B., & Al-Bawri, S. S. (2026). Machine learning and Geographic Information Systems (GIS)–based model for road crash severity prediction and behavioural pattern analysis. Traffic Injury Prevention, 1-10.

[21] Mashagba, H. A., Rahim, H. A., Yasin, M. N. M., Jamaluddin, M. H., Narongkul, S., Ramli, N. H., & Zahid, L. (2024, November). Two-Port Hexagonal Bar Slotted Dual-Band MIMO Textile Antenna for WBAN and 5G Applications. In 2024 International Symposium on Antennas and Propagation (ISAP) (pp. 1-2). IEEE.

[22] Rawash, Y. Z., Al-Naami, B., Alfraihat, A., & Owida, H. A. (2024). Advanced Low-Pass Filters for Signal Processing: A Comparative Study on Gaussian, Mittag-Leffler, and Savitzky-Golay Filters. Mathematical Modelling of Engineering Problems, 11(7).

[23] Al-Naami, B., Fraihat, H., Owida, H. A., Al-Hamad, K., De Fazio, R., & Visconti, P. (2022). Automated detection of left bundle branch block from ECG signal utilizing the maximal overlap discrete wavelet transform with ANFIS. Computers, 11(6), 93.

[24] Jawwad, A. K. A., Turab, N., Al-Mahadin, G., Owida, H. A., & Al-Nabulsi, J. (2024). A perspective on smart universities as being downsized smart cities: a technological view of internet of thing and big data. Indonesian Journal of Electrical Engineering and Computer Science, 35(2), 1162-1170.

[25] Abuowaida, S., Owida, H., Alsekait, D., Alshdaifat, N., Abdelminaam, D., & Alshinwan, M. (2025). UltraSegNet: A hybrid deep learning framework for enhanced breast cancer segmentation and classification on ultrasound images. *Computers, Materials, & Continua*, *83*(2), 3303.

[26] Owida, H. A., Hemied, O. S. M., Alkhawaldeh, R. S., Alshdaifat, N. F. F., & Abuowaida, S. F. A. (2022). Improved deep learning approaches for covid-19 recognition in ct images. *Journal of Theoretical and Applied Information Technology*, *100*(13), 4925-4931.

[27] Abuowaida, S. F. A., Chan, H. Y., Alshdaifat, N. F. F., & Abualigah, L. (2021). A novel instance segmentation algorithm based on improved deep learning algorithm for multi-object images. *Jordanian Journal of Computers and Information Technology (JJCIT)*, *7*(01).

[28] Abuowaida, S. U. H. A. I. L. A., Elsoud, E. S. R. A. A., Al-Momani, A. D. A. I., Arabiat, M. O. H. A. M. M. A. D., Owida, H. A., ALSHDAIFAT, N., & CHAN, H. Y. (2023). Proposed enhanced feature extraction for multi-food detection method. Journal of Theoretical and Applied Information Technology, 101(24), 8140-8146.

[29] Turki, H. M., Al Daoud, E., Samara, G., Alazaidah, R., Qasem, M. H., Aljaidi, M., ... & Alshdaifat, N. (2025). Arabic fake news detection using hybrid contextual features. International Journal of Electrical & Computer Engineering (2088-8708), 15(1).

[30] Arabiat, M. O. H. A. M. M. A. D., Abuowaida, S. U. H. A. I. L. A., Al-Momani, A. D. A. I., Alshdaifat, N. A. W. A. F., & Chan, H. Y. (2024). Depth estimation method based on residual networks and se-net model. Journal of Theoretical and Applied Information Technology, 102(3), 866-871.

[31] Alidmat, O., Owida, H. A., Yusof, U. K., Almaghthawi, A., Altalidi, A., Alkhawaldeh, R. S., ... & Alshaqsi, J. (2025). Simulation of Crowd Evacuation in Asymmetrical Exit Layout Based on Improved Dynamic Parameters Model. *IEEE Access*, *13*, 7512-7525.

[32] Ehsan, A., Iqbal, Z., Abuowaida, S., Aljaidi, M., Zia, H. U., Alshdaifat, N., & Alshammry, N. K. (2024). Enhanced anomaly detection in ethereum: Unveiling and classifying threats with machine learning. IEEE Access, 12, 176440-176456

[33] Iqtait, M., Ababneh, J., Rasmi, M., Abu-Jassar, A., & Abuowaida, S. (2024). Enhancing facial feature detection: hybrid active shape and active appearance model (HASAAM). IEEE Access, 12, 189590-189610.

[34] Owida, H. A., Al-Nabulsi, J., Al Hawamdeh, N., Abuowaida, S., Salah, Z., & Elsoud, E. A. (2024, November). Deep learning-based kidney disease classification using ResNet50: Enhancing diagnostic accuracy. In 2024 Second Jordanian International Biomedical Engineering Conference (JIBEC) (pp. 126-129). IEEE.

[35] Baniata, M., Abuowaida, S., Aljaidi, M., Kharabsheh, M., Alsarhan, A., & Alsuwaylimi, A. A. (2025). A Multi-Modal Attention-Guided Network for Alzheimer's Disease Classification Using Deep Learning. *Engineering, Technology & Applied Science Research*, *15*(5), 27150-27158.

[36] Aburomman, A., Abuowaida, S., Al-Momani, A., Al Henawi, E., Elsoud, E. A., Salah, Z., ... & Khouj, M. (2024, December). Automated detection of Alzheimer's disease using EfficientNet-B3 architecture with transfer learning. In 2024 25th International Arab Conference on Information Technology (ACIT) (pp. 1-6). IEEE.

[37] Malkawi, M. M., Khaleel, A., Bataineh, M., Alkhatib, R., Nayfeh, O., Alqaraleh, M., & Abuowaida, S. (2025, April). Enhancing Autism Spectrum Disorder Detection: A Novel Ensemble Learning Framework with Multi-Model Integration. In 2025 1st International Conference on Computational Intelligence Approaches and Applications (ICCIAA) (pp. 1-6). IEEE.

[38] Abuowaida, S., Alnsour, Y., Salah, Z., Alazaidah, R., Al-Batah, M. S., Alzboon, M. S., ... & Moh'd, B. A. H. (2025). Hybrid Ensemble Architecture for Brain Tumor Segmentation Using EfficientNetB4-MobileNetV3 with Multi-Path Decoders. DATA & METADATA Учредители: Salud, Ciencia y Tecnologia, 4, 374.

[39] Salah, Z., Abu Owida, H., Abu Elsoud, E., Alhenawi, E., Abuowaida, S., & Alshdaifat, N. (2024). An effective ensemble approach for preventing and detecting phishing attacks in textual form. Future Internet, 16(11), 414.

[40] Owida, H. A., Hassan, M. R., Ali, A. M., Alnaimat, F., Al Sharah, A., Abuowaida, S., & Alshdaifat, N. (2024). The performance of artificial intelligence in prostate magnetic resonance imaging screening. *International Journal of Electrical & Computer Engineering (2088-8708)*, *14*(2).

XPA-FedCBR: A Federated Case-Based Reasoning System for Explainable and Privacy-Aware Lung Cancer Diagnosis

Devyani Rawat[1], Shuchi Bhadula[2], Sachin Sharma[3]
Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun, India[1, 2]
Amity School of Engineering and Technology, Amity University Punjab, Mohali, India[3]

***Abstract*—Lung cancer is among the deadliest cancers worldwide, largely because it is usually caught too late. Building AI tools for earlier, stage-aware diagnosis is hard in practice: patient scans are scattered across hospitals that cannot share them freely, and clinicians are reluctant to trust models that cannot explain their reasoning. This study presents XPA-FedCBR, a framework that addresses both concerns. It combines Federated Learning, which trains a shared model across institutions without moving raw data, with Case-Based Reasoning, which supports each prediction by retrieving clinically similar past cases and, for uncertain predictions, using them to refine the decision. We evaluate it on the LIDC-IDRI dataset (roughly 1,018 CT scans from 1,010 patients), split across five simulated institutions under non-IID conditions, with annotations mapped to stages I–IV. Over four independent runs, the proposed model reaches a macro-F1 of 0.915, significantly ahead of FedAvg without CBR (0.875), a centralized CNN (0.895), and a standalone CNN (0.844), with its advantage largest under domain shift. Its case-based explanations, evaluated quantitatively and against gradient-based saliency, give clinicians evidence they can directly inspect.**

***Keywords*—Case-based reasoning; federated learning; explainable AI; lung cancer; clinical decision support; medical imaging**

I. Introduction

A. Clinical Background

One of the most common and deadly types of cancer in the world, lung cancer accounts for a large percentage of cancer-related deaths. The high fatality rate is primarily attributed to the late-stage detection of the disease, as early-stage lung cancer often presents minimal or no noticeable symptoms[1]. As a result, a large number of patients are diagnosed only after the cancer has progressed to an advanced stage, where treatment options become limited, and survival rates decline substantially.

In order to improve patient survival and treatment outcomes, early and precise identification of lung cancer is essential. When detected at an early stage, lung cancer can be managed more effectively through surgical intervention, targeted therapy, or radiotherapy, leading to significantly higher survival rates[2]. Therefore, timely screening and reliable diagnostic tools are essential for reducing mortality and improving the overall quality of patient care.

Non-Small Cell Lung Cancer (NSCLC), which accounts for most lung cancer cases, is staged from I to IV according to tumor size, lymph node involvement, and metastasis. The stage matters enormously for outcome: patients caught at Stage I or II have far better survival odds, while Stage III and IV carry a poor prognosis [3]. Because the stage drives the choice of treatment, a diagnosis that simply flags cancer as present or absent is not enough; what clinicians need is a system that can tell the stages apart.

Despite advances in medical imaging such as computed tomography (CT), lung cancer screening still has real limitations. Much of it depends on manual interpretation by radiologists, which is time-consuming and varies from one reader to the next [4]. Subtle nodules are easy to miss, especially early on, and both false positives and false negatives take their toll, the former sending patients into unnecessary follow-ups and anxiety, the latter delaying diagnoses that matter. These gaps are why the field needs diagnostic tools that are not just accurate but consistent and clinically trustworthy, and why computational methods have drawn so much attention as a way to close them.

B. Challenges in AI-Based Lung Cancer Detection

AI has shown real promise in lung cancer analysis, but several obstacles keep it from being adopted in actual clinical practice. One is data scarcity and institutional silos. Well-annotated imaging data is hard to come by and tends to sit locked inside individual hospitals, which, for reasons of patient confidentiality and limited infrastructure, cannot share it freely. A model trained in one institution therefore rarely sees the range of patients it would need to generalize well.

A second obstacle is privacy and regulation. CT scans and clinical records are among the most sensitive data a hospital holds, yet conventional centralized learning depends on pooling everything in one place, an approach that can run afoul of rules like HIPAA and GDPR and leaves institutions understandably wary of any system built around direct data sharing [7].

The third is the opacity of deep learning itself. These models can be highly accurate, but they say little about *how* they reach a decision, and a clinician who cannot follow that reasoning has no way to check an AI-assisted diagnosis in something as consequential as cancer staging. Taken together, these problems point to a clear need: systems that can learn from data spread across institutions, respect privacy law, and still explain themselves in terms a clinician can act on.

C. Motivation

The challenges outlined above- fragmented data, privacy constraints, and the opacity of deep models- are rarely addressed together, yet in clinical practice they arise together. A system that preserves privacy but cannot explain its decisions will not earn clinician trust; one that explains its reasoning but requires centralized data cannot be deployed across institutions. This motivates a combined approach. Federated learning offers a way to train across distributed hospitals without moving sensitive data, directly addressing the privacy and data-silo problem. Case-based reasoning offers a complementary strength: by grounding each prediction in similar, previously confirmed cases, it provides the transparent, evidence-based justification clinicians rely on when making decisions. Pairing the two is therefore natural; federated learning supplies privacy-preserving predictive capability, while case-based reasoning supplies the interpretability that turns a prediction into a clinically actionable, trustworthy recommendation. Combining the two, rather than using either one alone, is the main idea behind this work.

Despite these advances, existing approaches address the key challenges only in isolation. Federated learning methods for lung cancer have focused primarily on improving accuracy and preserving privacy, but they largely operate as black boxes and offer little in the way of clinician-facing explanation. Case-based reasoning, by contrast, provides transparent, evidence-based justification that aligns well with clinical decision-making, yet it has rarely been integrated with modern deep or distributed learning and does not, on its own, address privacy or scale to high-dimensional imaging data. Furthermore, most existing systems are limited to binary malignant-versus-benign detection rather than clinically meaningful stage-aware classification. As a result, no unified framework currently delivers privacy-preserving collaborative learning, stage-level diagnosis, and clinician-understandable explanations together. This gap motivates the hybrid framework proposed in this work, which couples federated learning with case-based reasoning to achieve all three within a single architecture.

D. Contributions

The main contributions of this work are summarized as follows:

- A hybrid Case-Based Reasoning and Federated Learning framework is proposed for lung cancer detection, enabling collaborative model training while preserving patient data privacy.
- Introduces stage-aware lung cancer classification for clinically meaningful diagnosis.
- An explainable clinical decision-support layer is introduced, which aligns AI predictions with clinician reasoning by retrieving and presenting similar historical patient cases.
- Robustness across decentralized clinical data sources is demonstrated by evaluating the proposed framework under a simulated multi-institution federated learning environment.
- A comprehensive comparative evaluation is performed against centralized learning and non-explainable baseline models to confirm the proposed framework's efficacy.

II. Related Work

A. Lung Cancer Detection Using AI

Early efforts in lung cancer diagnosis using artificial intelligence were built around traditional computer-aided diagnosis (CAD) systems. In these systems, handcrafted features such as shape, texture, and intensity of lung nodules were manually extracted from CT images, followed by conventional machine learning classifiers like support vector machines and decision trees for detection and classification. Such CAD systems were designed to assist radiologists and reduce workload, but they typically depended heavily on engineered features and showed limited adaptability across diverse clinical data sources[5]. A hybrid transfer learning framework, VER-Net (2024), combined VGG19, EfficientNetB0, and ResNet101 to classify lung cancer from CT images. Using the Chest CT-Scan Images Dataset (Kaggle), the proposed model achieved 91.0% accuracy, 92.0% precision, 91.0% recall, and an F1-score of 91.3%, highlighting the effectiveness of hybrid deep learning for automated lung cancer diagnosis[6].

B. Federated Learning in Healthcare

A proximal-guided hybrid federated learning framework for pneumonia diagnosis that combines FedProx with parameter-efficient adaptive intelligence to address data heterogeneity across hospitals. The model was evaluated on Chest X-ray datasets collected from multiple healthcare institutions and achieved 98.7% accuracy, demonstrating improved robustness and communication efficiency compared with conventional federated learning methods[7].

One of the central motivations for applying FL in healthcare is its ability to address privacy and regulatory constraints such as HIPAA and GDPR by keeping medical data within the originating institution while still enabling collaborative model development. This decentralization reduces the risk of sensitive data exposure and aligns with legal and ethical requirements for handling patient information[8].

C. Case-Based Reasoning in Clinical Systems

CBR is a problem-solving approach that draws on solutions to previously encountered cases to inform decisions about new, similar situations. In clinical settings, CBR has been applied to diagnostic support by retrieving past patient cases with similar clinical features and outcomes to guide current decision-making. This reasoning paradigm aligns closely with the way clinicians think, often referring to past patient histories, symptoms, and outcomes when diagnosing new patients [9]. A hybrid knowledge modeling framework incorporating Case-Based Reasoning (CBR) and rule-based reasoning was proposed to recommend treatment strategies for patients with chronic kidney disease–mineral and bone disorder (CKD-MBD). The system was validated using electronic health records and clinical guidelines, providing treatment recommendations that closely matched expert clinical decisions while improving consistency in decision-making [10].

An enhanced Case-Based Reasoning (CBR) framework integrated with a Rule-Based System (RBS) was developed to support medical diagnosis by retrieving similar historical patient cases and refining diagnostic decisions using expert rules. The system was evaluated using a medical diagnosis dataset and achieved an overall diagnostic accuracy of 96.8%, demonstrating the effectiveness of combining CBR with rule-based reasoning for clinical decision support[11]. An explainable Case-Based Reasoning (XCBR) framework was developed for lung cancer diagnosis by integrating Naïve Bayes, Multilayer Perceptron (MLP), SHAP, and Harris Hawks Optimization for feature selection and explanation. The model was evaluated on two public lung cancer datasets and achieved 94.47% and 100% classification accuracy, demonstrating the effectiveness of combining CBR with explainable AI for clinical decision support[12].

D. Research Gap

Prior work has advanced federated and case-based approaches for lung cancer diagnosis, but along separate tracks. In the federated setting, Liu et al. [14] applied federated learning across multiple centres to predict treatment response in non-small cell lung cancer, and Heidari et al. [15] combined federated learning with blockchain for CT-based lung cancer detection; both demonstrate that collaborative training without data sharing is feasible, yet neither incorporates a mechanism to explain individual predictions to clinicians. Case-based reasoning offers precisely this missing interpretability; Pradeep et al. [13] show that retrieving and presenting similar past cases yields transparent, clinically aligned justification, but such systems are typically built on manually defined features and have rarely been coupled with deep or distributed learning, leaving them unable to scale to high-dimensional imaging or to operate under privacy constraints. A further limitation is that most of these systems target binary malignant-versus-benign detection rather than stage-level diagnosis, which is what ultimately guides treatment decisions. Consequently, no existing framework simultaneously provides privacy-preserving collaborative learning, stage-aware classification, and case-based explanations suited to clinical use. The proposed work addresses this gap by unifying all three within a single architecture.

III. Proposed Methodology

A. System Overview

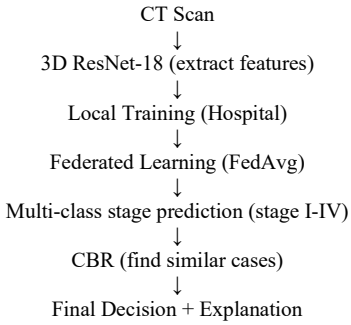


Fig. 1. Proposed hybrid CBR-FL framework.

The proposed intelligent digital clinical diagnostic system is designed to support lung cancer diagnosis while preserving patient privacy and providing explainable clinical decisions. Fig. 1 shows that the proposed framework processes CT scan images using a CNN-based feature extraction model followed by federated learning for privacy-preserving multi-class lung cancer stage prediction. The predicted stage is further supported through a Case-Based Reasoning module that retrieves clinically similar historical cases to provide interpretable diagnostic explanations.

The overall architecture follows a distributed learning framework in which multiple healthcare institutions act as local clients, and a central server coordinates the global learning process. Each local client trains a diagnostic model using its own patient data, ensuring that sensitive clinical information does not leave the institution. The CNN-based model is extended to perform multi-class classification, where each class corresponds to a specific stage of lung cancer. Table I shows the detailed stages of Non-Small Cell Lung Cancer.

TABLE I. Detailed Stages of Non-Small Cell Lung Cancer

Stage 1	Cancer is only located in the lungs.
Stage 2	Cancer is located in the lungs and may be in the lymph nodes near the lungs.
Stage 3	Cancer is located in the lungs and lymph nodes in the middle of the chest.
Stage 3A	Cancer is extensive but localized to one side of the lung.
Stage 3B	Cancer has spread to lymph nodes on both sides of the lungs.
Stage 4	Cancer has spread to both lungs or to another part of the body. This stage is considered the most advanced.

During training, only the locally learned model parameters are shared with the global server, which aggregates them to update a global model using a federated learning strategy. This global model is then redistributed to the local clients for further refinement. In parallel, a case-based reasoning module maintains a repository of historical patient cases and retrieves clinically similar cases to support transparent and interpretable decision-making. By combining federated learning with case-based reasoning, the system enables accurate lung cancer diagnosis while ensuring data privacy and clinician-understandable explanations. Fig. 2 shows the proposed hybrid case-based learning and federated learning framework for privacy-preserving lung cancer stage prediction using CT scan images.

B. Local Client-Side Processing and Feature Extraction

In the proposed system, lung CT scan images are collected at individual hospitals and processed locally to ensure patient data privacy. Each CT scan is first passed through a 3D ResNet-18 convolutional neural network (CNN), which automatically extracts relevant features such as lung nodules, texture patterns, and structural abnormalities. These features represent important visual characteristics required for lung cancer detection. The extracted features are then used to train a local diagnostic model within the hospital environment. At no point is raw patient data shared outside the institution.

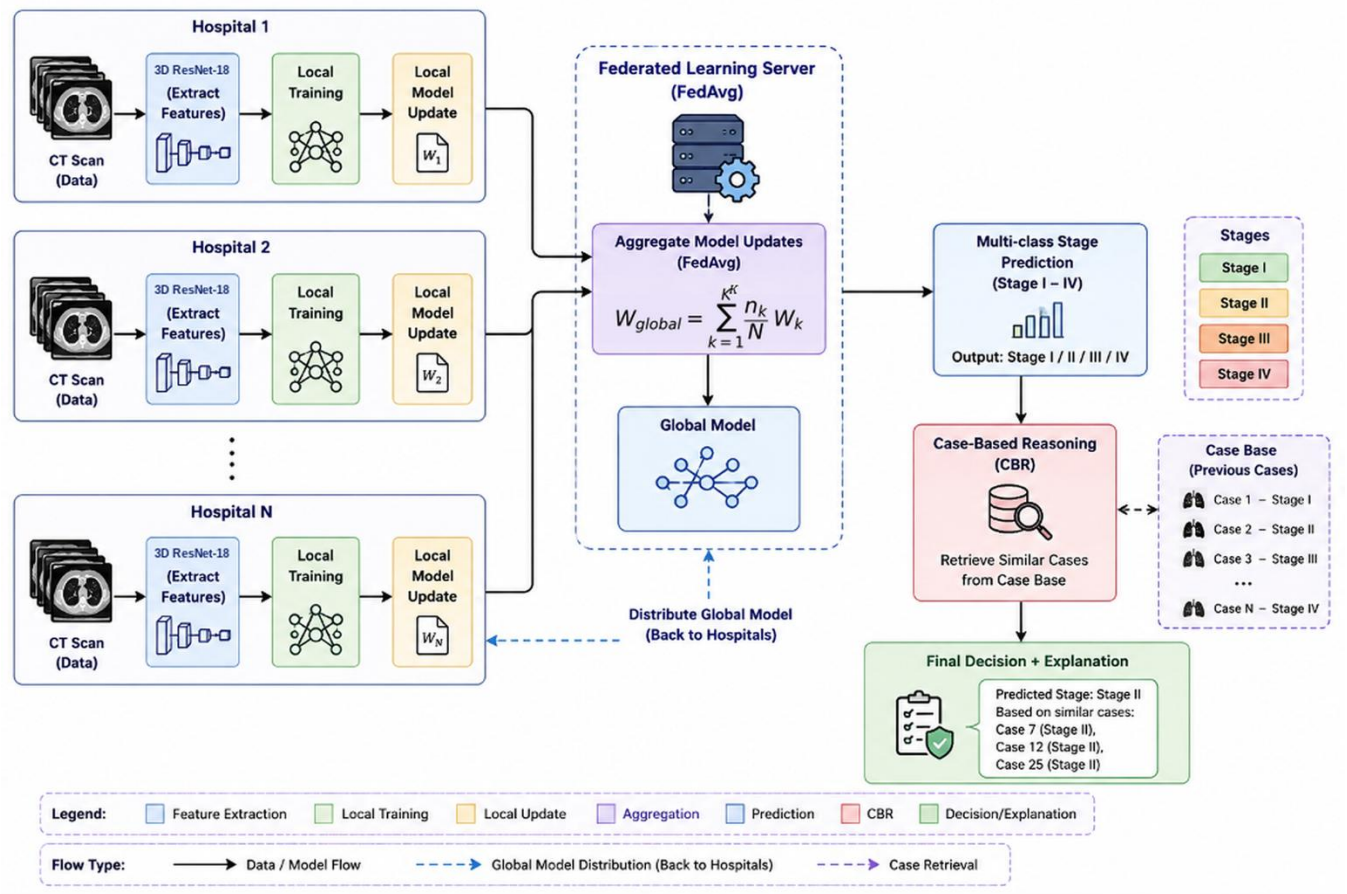


Fig. 2. Proposed hybrid case-based reasoning and federated learning framework for privacy-preserving lung cancer stage prediction using CT scan images.

C. Federated Learning for Collaborative Model Training

To enable collaborative learning across multiple clinical centers while preserving patient privacy, a federated learning framework is adopted in the proposed system. Instead of sharing raw CT scan data, each participating hospital trains the diagnostic model locally using its own dataset. This ensures that sensitive patient information remains within the institution and complies with privacy regulations.

After local training, only the updated model parameters are transmitted to a central server. The server performs secure aggregation of these updates using the Federated Averaging (FedAvg) algorithm to generate a global model that captures collective knowledge from all participating centers. The updated global model is then redistributed to the local clients for further training in subsequent rounds.

Since clinical data from different hospitals may vary due to differences in imaging equipment, patient demographics, and acquisition protocols, the federated framework is designed to handle data heterogeneity. Through iterative aggregation across multiple training rounds, the global model becomes more robust and generalized, learning patterns that are consistent across diverse clinical environments. Furthermore, the predictions generated by the global model are supported by a case-based reasoning mechanism, which retrieves clinically similar cases and uses their confirmed stage labels both to explain and, for low-confidence predictions, to refine the final decision.

In federated learning, suppose there are K participating clinical centers. Each hospital k trains a local model on its dataset containing n_k samples. After local training, the model parameters w_k are transmitted to the central server.

The global model parameters at communication round t+1 are computed using weighted averaging as follows:

$$w_{t+1} = \Sigma_{k=1}^{K} (n_k / N) \cdot w_k^t \quad (1)$$

where,

w_k^t = local model parameters from hospital k at round t

n_k = number of samples at hospital k

$N = \Sigma_{k=1}^{K} n_k$ = total number of samples across all hospitals

K = total number of participating hospitals

This weighted aggregation ensures that hospitals with larger datasets contribute proportionally more to the global model update.

D. Case-Based Reasoning Module

The Case-Based Reasoning (CBR) module is incorporated to provide interpretability and support clinical decision-making by leveraging previously diagnosed cases. In this approach, each patient case is represented using a combination of imaging features extracted from CT scans and relevant clinical

information, if available. This structured representation allows the system to compare new cases with historical records effectively.

To identify similar cases, a similarity measurement technique is applied. In this work, cosine similarity is used to compute the closeness between the feature vector of the new patient and those stored in the case database. Based on this similarity score, the top-k most relevant cases are retrieved.

Once similar cases are identified, the system analyses their confirmed diagnoses. These retrieved cases serve two complementary purposes: they are presented alongside the model's prediction to provide context and justification, and, for low-confidence predictions, their stage labels contribute to the final decision through the fusion mechanism described in Section III-E. This process enables explanations that are intuitive and aligned with clinical reasoning, while allowing similar historical cases to reinforce or correct uncertain predictions, thereby improving the transparency, trustworthiness, and reliability of the diagnostic system.

E. Hybrid Integration Strategy

The proposed framework integrates federated learning (FL) and case-based reasoning (CBR) to combine accurate prediction with interpretable, evidence-based decision support. Rather than treating the two components independently, the system couples them so that CBR both explains and, where appropriate, refines the federated model's output.

For a query case with deep feature vector q (extracted by the shared 3D ResNet-18 backbone), the global federated model produces a probability distribution over the four lung cancer stages, $p = (p_1, p_2, p_3, p_4)$, and a corresponding confidence $\hat{c} = \max_j p_j$. This distribution serves as the initial diagnostic indicator, which is then passed to the CBR module.

1) Case retrieval: To identify clinically similar precedents, the CBR module compares **q** against every stored case *i* (with feature vector $\mathbf{c}_i$) using cosine similarity:

$$\mathrm{sim}(\mathbf{q}, \mathbf{c}_i) = (\mathbf{q} \cdot \mathbf{c}_i) / (\|\mathbf{q}\| \, \|\mathbf{c}_i\|). \quad (2)$$

The top-*k* cases with the highest similarity are retrieved and denoted $\mathcal{N}$. In this work k = 5.

2) Neighbour label evidence: The retrieved cases contribute a similarity-weighted label distribution, in which each neighbour supports its own confirmed stage in proportion to how similar it is to the query:

$$v_j = \left(\Sigma_{i \in \mathcal{N} : y_i = j} \, \mathrm{sim}(\mathbf{q}, \mathbf{c}_i) \right) / \left(\Sigma_{i \in \mathcal{N}} \, \mathrm{sim}(\mathbf{q}, \mathbf{c}_i) \right), \quad \text{for each stage } j \quad (3)$$

where, y_i is the confirmed stage label of case *i* and $\mathbf{v} = (v_1, v_2, v_3, v_4)$.

3) Confidence-gated decision fusion: In the decision-fusion stage, the quantitative prediction from the FL model is combined with the qualitative evidence from retrieved cases through a confidence-gated rule:

$$\mathbf{s} = \mathbf{p}, \text{ if } \hat{c} \geq \tau, \quad \mathbf{s} = (1 - \alpha) \cdot \mathbf{p} + \alpha \cdot \mathbf{v}, \text{ if } \hat{c} < \tau, \quad (4)$$

and the final predicted stage is

$$\hat{y} = \mathrm{argmax}_j \, s_j \quad (5)$$

Here τ is a confidence threshold and $\alpha \in [0, 1]$ controls the influence of retrieved cases; we use $\tau = 0.7$ and $\alpha = 0.4$. When the federated model is confident ($\hat{c} \geq \tau$), its prediction is retained unchanged, and the retrieved cases act purely as supporting explanation. When the model is uncertain ($\hat{c} < \tau$), the regime in which classification errors are concentrated, the labels of clinically similar historical cases are allowed to re-rank the decision, so that strong agreement among close neighbours can correct an unreliable prediction.

This selective fusion is the reason the hybrid framework improves upon federated averaging without CBR: the CBR layer leaves confident predictions untouched and intervenes only on borderline cases, where neighbour evidence is most informative. The effect is therefore most pronounced under high-shift and domain-shift conditions (Table VI), where distributional differences lower model confidence and increase the proportion of cases routed through neighbour-based re-ranking.

Finally, to ensure the process is transparent to clinicians, the system exposes each step: for every new case, it presents the predicted stage together with the top-*k* retrieved cases and their confirmed diagnoses as supporting evidence. This step-by-step presentation allows clinicians to see not only the predicted stage but the historical cases underlying it, supporting validation and building trust in the system.

IV. Experimental Setup

A. Dataset Description

The proposed framework is evaluated using the publicly available LIDC-IDRI (Lung Image Database Consortium and Image Database Resource Initiative) dataset, which comprises approximately 1,018 thoracic CT scans collected from 1,010 patients with 2,669 annotated lung nodules provided by multiple expert radiologists. The dataset contains a diverse range of nodules with variations in size, shape, texture, and malignancy levels. In this work, the malignancy annotations are mapped to lung cancer stages (Stage I–IV) to facilitate multi-class stage prediction. To simulate a realistic multi-institutional healthcare environment, the dataset is partitioned into multiple subsets representing different clinical centers within the federated learning framework.

TABLE II. Simulated Federated Learning Environment(5 Hospitals Data)

Institution	Scans	Stage Distribution (I/II/III/IV)	Data Heterogeneity
Hosp1	200	40/30/20/10	Low variance
Hosp2	250	30/40/20/10	Medium (noise)
Hosp3	300	20/20/30/30	High (shift)
Hosp4	150	35/35/15/15	Domain shift
Hosp5	118	25/25/25/25	Balanced

Table II presents the simulated federated distribution of CT scan data across multiple participating institutions with varying stage distributions and heterogeneity conditions. Each simulated

institution performs local model training on its respective subset without sharing raw patient data, thereby preserving privacy while enabling collaborative learning. A non-IID data distribution is further simulated through label shifts across institutions, where certain centers contain a higher proportion of early-stage cases while others contain more advanced-stage samples. For each institution, the dataset is divided into 70% training, 15% validation, and 15% testing sets.

This experimental setup enables the evaluation of the proposed system under practical healthcare challenges such as distributed data ownership, institutional data heterogeneity, and privacy constraints. However, the stage labels are approximated using malignancy annotations and may not perfectly correspond to clinical staging standards. Additionally, due to limited access to real multi-institutional datasets, the federated learning environment is simulated through dataset partitioning.

B. Preprocessing and Feature Extraction

Before model training, the CT scan images go through a number of preparation procedures to ensure consistency, reduce noise-related variations, and improve feature quality. These preprocessing operations are essential for enhancing model robustness and diagnostic performance in medical imaging applications. Initially, lung nodule segmentation is performed using a U-Net architecture to isolate the lung regions and relevant nodules from surrounding anatomical structures[16]. Following segmentation, the extracted regions are resized to maintain uniformity across all CT scan samples (Table III).

TABLE III. SUMMARIZES THE COMPLETE EXPERIMENTAL CONFIGURATION AND HYPERPARAMETER SETTINGS USED FOR BOTH LOCAL TRAINING AND FEDERATED AGGREGATION

Category	Parameter	Value
Feature extractor	Backbone	3D ResNet-18
	Input size	64 × 64 × 64 (resized CT volume)
	Pretraining	Trained from scratch
Local training	Optimizer	Adam
	Learning rate	1×10^{-4}
	Batch size	16
	Local epochs per round	5
	Loss function	Categorical cross-entropy
Federated setup	Aggregation algorithm	FedAvg (weighted)
	Communication rounds	50
	Number of clients (hospitals)	5
	Client participation per round	Full (all clients)
	Data distribution	Non-IID (label shift across clients)
CBR module	Similarity metric	Cosine similarity
	Retrieved neighbours (k)	5
	Confidence threshold (τ)	0.7
	Fusion weight (α)	0.4
Data split	Train / Validation / Test	70% / 15% / 15%
Hardware	GPU	NVIDIA DGX B200

Next, Hounsfield Unit (HU) normalization is applied to reduce variations caused by differences in imaging equipment and acquisition settings. To further enhance image quality, basic noise reduction techniques, such as smoothing filters, are employed to suppress irrelevant artifacts while preserving important lung structures.

During training, data augmentation methods including rotation, flipping, and intensity shifting are used to increase model robustness and generalization. Since the proposed framework follows a federated learning paradigm, no central data pooling is performed, thereby preserving patient privacy across participating institutions.

After preprocessing, feature extraction is carried out using a 3D ResNet-18 architecture fine-tuned for multi-class lung cancer stage prediction. The network automatically learns discriminative imaging features such as nodule shape, texture, and intensity patterns from CT scans. These extracted deep features are subsequently utilized for both classification and case-based reasoning within the suggested framework. The proposed framework was implemented and evaluated using NVIDIA DGX B200 GPU-enabled computational infrastructure to support efficient deep learning training and federated learning experimentation on CT scan datasets.

C. Baseline Models

To assess the effectiveness of the proposed hybrid framework, it is compared against multiple baseline models representing different learning and deployment strategies.

1) Centralized deep learning model: In this baseline, all CT scan data is assumed to be available at a single location, and a CNN is trained in a centralized manner. This model serves as an upper-bound reference in terms of data availability, as it utilizes the complete dataset for training. However, it does not address privacy concerns associated with sharing sensitive medical data.

2) Federated learning without CBR: This baseline implements the federated learning framework using the same data distribution and training setup as the proposed method, but without incorporating the case-based reasoning module. It focuses solely on prediction performance while preserving data privacy[17]. This comparison helps isolate and evaluate the contribution of the CBR module in terms of interpretability and decision support.

3) Traditional machine learning models: Conventional machine learning algorithms, such as Support Vector Machines (SVM) and Random Forests, are used as additional baselines. These models are trained on extracted features from CT scans and provide a comparison with deep learning-based approaches [18]. Although less complex, they serve as a useful reference to highlight the advantages of the proposed hybrid framework.

To ensure a fair comparison, all deep-learning models, the proposed framework, FedAvg without CBR, the centralized CNN, and the standalone CNN — share the same 3D ResNet-18 backbone and are trained and evaluated on identical data splits with matched optimizer, learning rate, batch size, and training epochs.

The traditional machine-learning baselines (SVM and Random Forest) use the same 3D ResNet-18 features. This ensures performance differences reflect the methods rather than differing experimental conditions.

D. Evaluation Metrics

The suggested system's performance is assessed using both quantitative classification metrics and qualitative explainability measures to ensure a comprehensive assessment.

For classification performance, standard metrics such as accuracy, precision, recall, and F1-score are used. Since the proposed approach performs multi-class classification across different lung cancer stages, macro-averaged precision, recall, and F1-score are additionally considered to provide a balanced evaluation across all classes[15].

- Accuracy assesses the comprehensive accuracy of the model:

$$\frac{TP + TN}{TP + TN + FP + FN}$$

- Precision denotes the ratio of accurately predicted positive instances:

$$\frac{TP}{TP + FP}$$

- Recall (Sensitivity) evaluates the model's accuracy in identifying positive situations:

$$\frac{TP}{TP + FN}$$

- F1-score represents the harmonic mean of precision and recall:

$$2\,\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

where, *TP*, *TN*, *FP,* and *FN* denote true positives, true negatives, false positives, and false negatives, respectively.

E. Qualitative Explainability Assessment

In addition to classification performance, the suggested system is assessed according to its capacity to provide meaningful and interpretable explanations using the case-based reasoning module.

1) Case relevance: Measures how closely the retrieved cases match the input case in terms of clinical and imaging features.

2) Interpretability: Assesses how easily clinicians can understand the reasoning behind the model's prediction based on the presented similar cases.

These qualitative aspects are important for determining the practical usability of the system in clinical settings, where transparency and trust play a crucial role in decision-making.

V. Results and Discussion

A. Qualitative Performance Analysis

The proposed framework performs stage-aware classification of lung cancer from CT scans across Stage I–IV. This section presents a qualitative and quantitative analysis of its performance, beginning with a comparison against baseline models and followed by a stage-wise breakdown and an evaluation of the case-based explanations.

1) Comparison with baselines: As shown in Table IV, the proposed Hybrid CBR-FL framework achieves the highest scores across all metrics (Macro F1 of 0.915 ± 0.002), with low standard deviations indicating stable performance across the four runs. It significantly outperforms all baselines ($p < 0.05$), including the centralized CNN despite the latter's access to pooled data and the federated baseline without CBR, confirming both the benefit of the hybrid design and the specific contribution of the case-based reasoning module.

Table V presents a simplified, imaging-based risk stratification scheme for NSCLC, included as clinical background rather than as a system output, to contextualize the stage-level predictions and their case-based explanations.

TABLE IV. Quantitative Comparison of Proposed and Baseline Models over Four Independent Runs (Mean ± Standard Deviation)

Model	Macro Precision	Macro Recall	Macro F1	p-value (F1 vs. Proposed)	Statistically Significant (p<0.05)?
Hybrid CBR-FL (Proposed)	**0.918 ± 0.003**	**0.912 ± 0.003**	**0.915 ± 0.002**	—	—
Centralized CNN	0.896 ± 0.006	0.892 ± 0.005	0.895 ± 0.005	0.0017	Yes
FedAvg (No CBR)	0.878 ± 0.006	0.873 ± 0.006	0.875 ± 0.005	0.0002	Yes
Standalone CNN	0.846 ± 0.004	0.841 ± 0.004	0.844 ± 0.003	< 0.0001	Yes
Local Avg	0.822 ± 0.007	0.814 ± 0.007	0.815 ± 0.007	< 0.0001	Yes

p-values are computed from an unpaired two-sample t-test of Macro F1 against the proposed model. All baseline differences from the proposed model are statistically significant ($p < 0.05$).

TABLE V. Simplified AI-Based Risk Stratification for NSCLC

Level	Risk	Benign Probability	Key Imaging Features	Growth Rate	Clinical Interpretation
Level A	Low	> 85% benign	Small, smooth, homogeneous nodules	Stable	Routine monitoring
Level B	Suspicious	40 – 85% benign	Medium size, mild irregularity, mixed texture	Slow growth	Follow-up/ possible biopsy
Level C	High	< 40% benign	Large, irregular, spiculated, heterogeneous	Rapid Growth	Immediate intervention

TABLE VI. STAGE-WISE BREAKDOWN (PROPOSED MODEL)

Stage	Precision	Recall	F1-Score
Stage I	0.95	0.94	0.945
Stage II	0.93	0.92	0.925
Stage III	0.89	0.88	0.885
Stage IV	0.9	0.9	0.9

Table VI presents the stage-wise performance breakdown of the proposed Hybrid CBR-FL framework across different lung cancer stages. The outcomes show that the model maintains balanced precision, recall, and F1-score values for Stage I–IV classification, indicating reliable and consistent stage-aware prediction capability. Fig. 3 presents the comparative macro-level performance analysis of different classification models evaluated on the LIDC-IDRI federated simulation dataset. The proposed Hybrid CBR-FL framework achieves the highest overall performance, demonstrating improved classification effectiveness and robustness compared to conventional federated and standalone approaches.

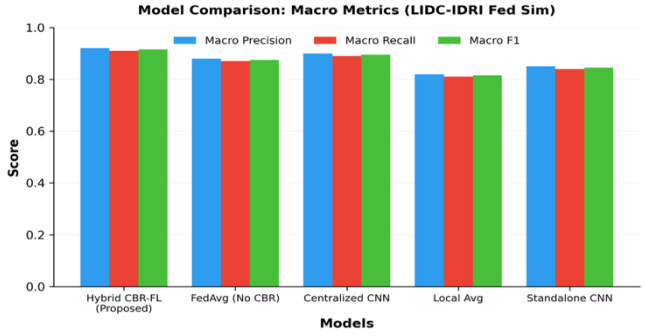


Fig. 3. Macro performance comparison across models.

B. Comparative Performance Analysis

Table VII presents the comparative performance analysis of the proposed Hybrid CBR-FL framework and the conventional FedAvg approach under different institutional heterogeneity conditions. The findings show that the suggested framework regularly achieves greater F1-scores across all simulated hospital environments. The performance improvement is particularly significant in high-shift and domain-shift scenarios, where variations in data distribution and imaging characteristics are more prominent. This indicates that the proposed hybrid framework is more robust and adaptable to heterogeneous decentralized healthcare data compared to traditional federated learning approaches.

TABLE VII. PERFORMANCE COMPARISON UNDER DIFFERENT FEDERATED DATA HETEROGENEITY CONDITIONS

Institution Heterogeneity	Hybrid CBR - FL F1	FedAvg F1	Improvement
Low Variance (Hosp1)	93.2%	91.8%	+1.4%
High Shift (Hosp3)	89.1%	82.3%	+6.8%
Domain Shift (Hosp4)	87.6%	80.5%	+7.1%
Macro Average	91.4%	87.5%	+3.9%

C. Explainability Evaluation

The final output of the proposed framework consists of the predicted lung cancer stage (Stage I–IV), which is further supported by retrieval of clinically similar historical cases through the Case-Based Reasoning module. The CBR module also improves the interpretability of lung cancer predictions. Instead of providing only a final classification result, the system retrieves previously similar lung nodule cases from the dataset based on extracted imaging features. The explainability component assists in highlighting similarities between the current CT scan and historical cases with related malignancy characteristics and cancer stages[19]. By presenting comparable cases, the framework supports clinician-oriented decision-making and improves trust in the automated diagnostic process. The retrieved similar cases help provide supportive evidence for the predicted stage classification while maintaining patient data privacy across decentralized institutions.

TABLE VIII. QUANTITATIVE EVALUATION OF THE CASE-BASED REASONING MODULE (COMPUTED OVER THE FULL TEST SET, N = 75)

Metric	Value
Case Relevance (mean top-k cosine similarity)	0.842
Stage-match rate of retrieved cases (%)	76.4
Retrieval precision@k	0.764

TABLE IX. EXAMPLE OF CASE-BASED EXPLANATIONS PRODUCED BY THE PROPOSED FRAMEWORK FOR THREE REPRESENTATIVE TEST CASES

Case	Predicted Stage	Top-k Retrieved Cases (Stage, Similarity)	Mean Similarity	Stage-Match
Patient 1	Stage II	(II, 0.92), (II, 0.90), (II, 0.88), (III, 0.85), (I, 0.80)	0.870	3/5 (0.60)
Patient 2	Stage I	(I, 0.95), (I, 0.94), (I, 0.91), (I, 0.89), (II, 0.86)	0.910	4/5 (0.80)
Patient 3	Stage III	(III, 0.88), (III, 0.85), (III, 0.83), (III, 0.80), (III, 0.78)	0.828	5/5 (1.00)

TABLE X. COMPARISON OF EXPLANATION METHODS FOR THE FEDERATED MODEL

Method	Explanation type	What the clinician sees	Localization
Grad-CAM	Saliency heatmap	Highlighted image regions	Diffuse
CBR (Proposed)	**Retrieved similar cases**	**Confirmed prior cases + stages**	**Indirect**

D. Quantitative Explainability Evaluation

Beyond qualitative description, the Case-Based Reasoning module is evaluated using two measures: Case Relevance, the mean cosine similarity of the top-k retrieved cases to the query, and the stage-match rate, the fraction of retrieved cases whose confirmed stage matches the predicted stage. Over the full test set (N = 75), the framework achieves a Case Relevance of 0.842 and a stage-match rate of 76.4% (Table X), showing that the retrieved cases are both visually similar to the query and, in most instances, clinically consistent with its predicted stage. Table X further illustrates this on three representative cases: for Patient 1

(predicted Stage II), three of five retrieved cases confirm Stage II, while for Patient 3 (predicted Stage III), all five are Stage III with high similarity. In each case, the clinician sees not only a predicted stage but the specific prior cases and confirmed diagnoses supporting it, enabling direct inspection and validation of the reasoning.

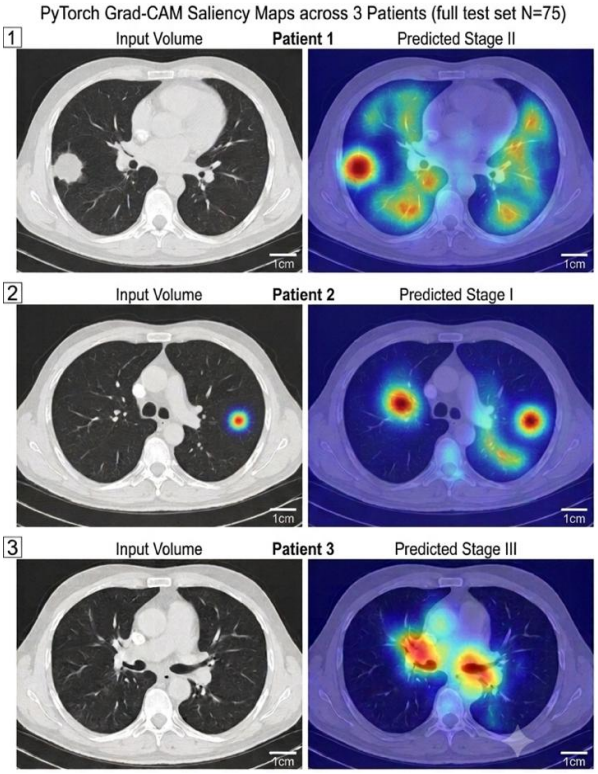


Fig. 4. Grad-CAM saliency maps for three test cases (Patients 1–3; predicted Stages II, I, III), each CT slice shown beside its Grad-CAM activation from the 3D ResNet-18. Activations are diffuse, in contrast to the case-level evidence from the CBR module.

Comparison with gradient-based explanation. To benchmark the proposed explanations against an established technique, Grad-CAM was applied to the same 3D ResNet-18 model at its final convolutional block. The resulting saliency maps Fig. 4. are often diffuse and multifocal, spanning several regions across both lungs rather than isolating a single decisive finding. Such maps indicate where the model attended but not why a stage was assigned, and are difficult to translate into a clinical rationale. In contrast, the CBR module grounds each decision in specific, previously confirmed cases with known stages (Tables VIII and IX), offering concrete, inspectable evidence. As summarized in Table X, the two approaches are complementary; Grad-CAM provides spatial attribution, while CBR provides justification grounded in prior confirmed cases, which is well suited to stage-aware diagnosis.

E. Discussion

The proposed hybrid framework demonstrates significant potential for supporting intelligent and privacy-aware lung cancer diagnosis in clinical environments. By integrating deep learning, federated learning, and case-based reasoning, the system is capable of assisting clinicians in performing stage-aware classification using CT scan images. The inclusion of explainability through similar case retrieval further improves the transparency of the diagnostic process, which is especially crucial in healthcare applications where interpretability and trust are essential for decision-making. This explainability is supported quantitatively, with retrieved cases showing high similarity and stage consistency, and compares favourably with gradient-based saliency methods in providing clinically inspectable evidence.

1) Privacy benefits: One of the main benefits of the suggested strategy is its privacy-preserving capability. The federated learning framework permits cooperative model training across several institutions without exchanging raw patient data, in contrast to traditional centralized learning techniques. This reduces privacy risks and supports secure utilization of distributed healthcare datasets while maintaining data confidentiality and institutional ownership.

2) Deployment feasibility: From a deployment perspective, the framework is designed to operate in decentralized healthcare environments where data availability and imaging characteristics may vary across institutions. The ability of the system to learn from heterogeneous distributed datasets improves its adaptability and robustness in practical medical settings. Although the current work utilizes simulated federated institutions and approximated stage annotations, the proposed architecture provides a strong foundation for future integration with real-world clinical infrastructures and multi-institutional healthcare systems.

VI. Conclusion and Future Work

In this work, we presented a hybrid framework that brings together Case-Based Reasoning and Federated Learning for stage-aware lung cancer classification from CT images. The design tackles two problems that usually pull against each other: it keeps patient data private by training across institutions without ever moving raw scans, while still holding up its classification performance even when the data across those institutions is uneven.

Across our experiments, the proposed model consistently outperformed local learning, standalone CNNs, and conventional FedAvg on macro-averaged precision, recall, and F1-score, and it did so while classifying cases across Stage I–IV and staying stable under non-IID, multi-institutional data. Just as importantly, the case-based reasoning component earned its place beyond raw accuracy: the cases it retrieves are genuinely similar to the query and largely consistent in stage, giving clinicians concrete, inspectable evidence for a prediction — something the diffuse heatmaps of gradient-based methods struggle to offer.

Taken together, these results suggest that pairing federated learning with case-based reasoning is a practical route toward lung cancer diagnosis that is private, accurate, and understandable at the same time — and a promising foundation for the kind of clinical decision-support systems that hospitals could actually trust and adopt.

A. Limitation

Despite the promising performance of the proposed framework, certain limitations remain. The federated learning environment considered in this study is simulated through dataset partitioning and does not fully represent real-world multi-institutional clinical infrastructures. In addition, the lung cancer stage labels are approximated using malignancy annotations available in the LIDC-IDRI dataset, which may not completely correspond to standard clinical staging procedures. A further constraint is the scale of the data: the 2,669 annotated nodules distributed across five simulated institutions and four stage categories are modest for deep federated training, which increases the risk of overfitting and limits the statistical power of per-stage estimates. Data augmentation (rotation, flipping, and intensity shifting) was applied to partially mitigate this**.** Furthermore, the current evaluation is limited to retrospective publicly available CT scan datasets, and external clinical validation with larger heterogeneous patient populations is necessary to further assess the generalizability and practical applicability of the proposed system in real healthcare settings.

B. Future Work

Future work will extend the proposed framework toward real-world hospital deployment to evaluate its performance under practical clinical and regulatory conditions. In addition, a radiologist-in-the-loop feedback mechanism will be incorporated to improve case retrieval quality, explanation reliability, and model refinement. Finally, the framework will be expanded to support multi-modal data integration, combining CT imaging with clinical and pathological information to enhance robustness and diagnostic accuracy.

REFERENCES

[1] Yu, M., Tyson, C., Limburg, P. J., & Beer, T. M. (2026). A quantitative framework to assess the potential of earlier cancer detection to improve cancer survival. *Cancer Epidemiology, Biomarkers & Prevention*, OF1–OF6.

[2] Lazzari, C., Spagnolo, C. C., Ciappina, G., Di Pietro, M., Squeri, A., Passalacqua, M. I., Marchesi, S., Gregorc, V., & Santarpia, M. (2023). Immunotherapy in early-stage non-small cell lung cancer (NSCLC): Current evidence and perspectives. *Current Oncology*, *30*(4), 3684–3696.

[3] Gayap, H. T., & Akhloufi, M. A. (2024). Deep machine learning for medical diagnosis, application to lung cancer detection: A review. *BioMedInformatics*, *4*(1), 236–284.

[4] Thanoon, M. A., Zulkifley, M. A., Mohd Zainuri, M. A. A., & Abdani, S. R. (2023). A review of deep learning techniques for lung cancer screening and diagnosis based on CT images. *Diagnostics*, *13*(16), 2617.

[5] Sari, S., Sutikno, T., Soesanti, I., & Setiawan, N. A. (2023). A review of convolutional neural network-based computer-aided lung nodule detection system. *IAES International Journal of Artificial Intelligence*, *12*(3), 1044–1061.

[6] Saha, A., Ganie, S. M., Pramanik, P. K. D., Yadav, R. K., Mallik, S., & Zhao, Z. (2024). VER-Net: A hybrid transfer learning model for lung cancer detection using CT scan images. *BMC Medical Imaging*, *24*(1), 120.

[7] AR, N. K. (2025). Proximal guided hybrid federated learning approach with parameter efficient adaptive intelligence for pneumonia diagnosis. *Scientific Reports*, *16*(1), 2412.

[8] Saini, S. S. (2025). Federated learning for privacy-preserving healthcare AI models. *International Journal of Unified Research & Development (IJURD)*, *1*(1), 1–5.

[9] Oyelade, O. N., & Ezugwu, A. E. (2020). A case-based reasoning framework for early detection and diagnosis of novel coronavirus. *Informatics in Medicine Unlocked*, *20*, 100395.

[10] Ali, S. I., Jung, S. W., Bilal, H. S. M., Lee, S. H., Hussain, J., Afzal, M., Han, M., Kim, T. S., Hwang, K. Y., & Lee, S. (2022). Clinical decision support system based on hybrid knowledge modeling: A case study of chronic kidney disease-mineral and bone disorder treatment. *International Journal of Environmental Research and Public Health*, *19*(1), 226.

[11] Mustafa, E. M., Saad, M. M., & Rizkallah, L. W. (2023). Building an enhanced case-based reasoning and rule-based systems for medical diagnosis. *Journal of Engineering and Applied Science*, *70*(1), 139.

[12] Tofighi, A. B., Ahmadi, A., & Mosadegh, H. (2025). A novel case-based reasoning system for explainable lung cancer diagnosis. *Computers in Biology and Medicine*, *185*, 109547.

[13] Pradeep, P., Caro-Martínez, M., & Wijekoon, A. (2024). A practical exploration of the convergence of Case-Based Reasoning and Explainable Artificial Intelligence. *Expert Systems with Applications*, *255*, 124733.

[14] Liu, Y., Huang, J., Chen, J.-C., Chen, W., Pan, Y., & Qiu, J. (2024). Predicting treatment response in multicenter non-small cell lung cancer patients based on federated learning. *BMC Cancer*, *24*, 688.

[15] Heidari, A., Javaheri, D., Toumaj, S., Navimipour, N. J., Rezaei, M., & Unal, M. (2023). A new lung cancer detection method based on the chest CT images using federated learning and blockchain systems. *Artificial Intelligence in Medicine*, *141*, 102572.

[16] Turjya, S. M., & Fawakherji, M. (2026). Federated lung nodule segmentation using a hybrid Transformer–U-Net architecture. *Scientific Reports*, *16*(1), 5228.

[17] Chen, Y., Liu, L., Feng, B., Chen, Y., Xu, J., & Chen, X. (2025). A meta-learning-based robust federated learning for diagnosing lung adenocarcinoma and tuberculosis granulomas. *Frontiers in Oncology*, *15*, 1666937.

[18] Causey, J. L., Guan, Y., Dong, W., Walker, K., Qualls, J. A., Prior, F., & Huang, X. (2019). *Lung cancer screening with low-dose CT scans using a deep learning approach*. arXiv preprint arXiv:1906.00240v

[19] Wang, J., Gao, R., Huo, Y., Bao, S., Xiong, Y., Antic, S. L., Osterman, T. J., Massion, P. P., & Landman, B. A. (2019). *Lung cancer detection using co-learning from chest CT images and clinical demographics*. arXiv. preprint arXiv:1902.08236.

Enhanced Simple Linear Iterative Clustering (ESLIC) for Precise Skin Lesion Segmentation in Dermoscopic Images

Leni Natalia Zulita[1]*, Sarjon Defit[2], Sumijan[3]
Information Engineering, Faculty of Computer Science, Dehasen University, Bengkulu, Indonesia[1]
Faculty of Computer Science, University of Putra Indonesia YPTK, Padang, Indonesia[2, 3]

***Abstract*—The main challenge in dermoscopy image analysis lies in the lesion segmentation process, which is often affected by variations in color, texture, lighting, and irregular lesion boundaries. The Simple Linear Iterative Clustering (SLIC) method is widely used for superpixel-based segmentation, but it still faces limitations in preserving the morphological details of lesions in complex areas. This study proposes Enhanced Simple Linear Iterative Clustering (ESLIC) as a new approach to improve the quality of skin lesion segmentation. ESLIC combines a dynamic compactness adjustment mechanism, local texture analysis, and gradient-based boundary refinement to generate superpixels that are more adaptive to image characteristics. The study was conducted using 277 dermoscopy images consisting of benign lesions and melanoma lesions. Before segmentation, the images underwent image enhancement through filtering, normalization, contrast enhancement, and sharpening. ESLIC's performance was evaluated based on segmentation quality and its impact on the classification process using a Convolutional Neural Network (CNN). The experimental results show that ESLIC is capable of producing a more accurate representation of lesion areas compared to conventional superpixel methods, particularly in areas with heterogeneous textures and complex lesion boundaries. The findings of this study indicate that ESLIC can improve the effectiveness of feature extraction and has the potential to become a key component in the development of artificial intelligence-based melanoma diagnosis support systems.**

***Keywords*—ESLIC; dermoscopy; melanoma; SLIC; superpixel; CNN**

I. Introduction

Advances in digital image processing and artificial intelligence have made significant contributions across various fields, particularly in the healthcare sector. The use of digital images enables medical analysis to be conducted more objectively, quickly, and accurately than with conventional visual examinations. One application that is currently growing rapidly is the use of computer-aided diagnosis (CAD) systems to assist medical professionals in the early detection of various diseases. These systems utilize image processing techniques, machine learning, and deep learning to extract important information from medical images, thereby supporting the clinical decision-making process more effectively [1], [2].

One disease of global concern is melanoma, a type of skin cancer that originates in melanocytes and has a high mortality rate if not detected in its early stages. Melanoma is known as one of the most aggressive forms of skin cancer because it metastasizes rapidly to other tissues in the body. Therefore, early detection is a critical factor in improving treatment success rates and prolonging patients' life expectancy [3].

In clinical practice, one of the technologies widely used to aid in the diagnosis of melanoma is dermoscopy [4], [5]. Dermoscopy is a non-invasive imaging technique that allows doctors to observe the morphological structure of the skin in greater detail than a standard visual examination. Through dermoscopy images, lesion characteristics such as color variations, pigment patterns, border shapes, and blood vessel distribution can be observed more clearly [6]. However, the interpretation of dermoscopy images still relies heavily on the clinical experience of dermatologists, which can lead to diagnostic variability among observers. Therefore, an artificial intelligence-based automated diagnostic system offers a promising solution for improving the consistency and objectivity of diagnoses [7].

One of the most important steps in a dermoscopy image analysis system is image segmentation. Segmentation aims to separate the lesion area from normal skin tissue so that feature extraction and classification can focus on the relevant areas. Various studies have shown that segmentation quality has a direct impact on improving the accuracy of AI-based skin lesion classification. Accurate segmentation allows classification models to learn the morphological characteristics of lesions more effectively, thereby improving the performance of automated diagnostic systems [8], [9].

Nevertheless, skin lesion segmentation remains a complex challenge. Variations in skin color, uneven lighting, presence of hair, light reflections, heterogeneous lesion textures, and irregular lesion boundaries often lead to segmentation errors. These complexities cause conventional segmentation methods to frequently produce imprecise object boundaries, thereby reducing the overall performance of the diagnostic system [10].

Among the various segmentation approaches that have been developed, superpixel-based methods are among the most widely used techniques because they can group pixels with similar characteristics into homogeneous regions.

*Corresponding author

One of the most popular superpixel algorithms is Simple Linear Iterative Clustering (SLIC) [11]. This algorithm offers high computational efficiency and the ability to generate superpixels with good spatial structure. However, SLIC still has limitations when applied to dermoscopy images because it relies solely on color and spatial information during the pixel clustering process. As a result, segmentation often fails to follow complex lesion boundaries, particularly in areas with heterogeneous textures and low contrast [12], [13].

Several studies have sought to improve the quality of skin lesion segmentation by integrating texture features, gradients, or deep learning-based approaches [14]. However, most studies still focus on improving segmentation quality in general and have not specifically developed a superpixel mechanism that adapts to changes in local image characteristics. Furthermore, preserving lesion boundaries remains a challenge that needs to be addressed to produce more accurate and consistent segmentation under various dermoscopy image conditions [15], [16].

Based on these issues, this study proposes the Enhanced Simple Linear Iterative Clustering (ESLIC) method as an extension of the SLIC algorithm to improve the quality of skin lesion segmentation in dermoscopy images. ESLIC is designed by integrating mechanisms for adaptive compactness adjustment, texture-aware clustering, gradient-aware feature integration, and boundary refinement. Unlike conventional SLIC, which uses a fixed compactness parameter, ESLIC adaptively adjusts this parameter based on local texture characteristics, thereby generating superpixels that are more representative of the lesion's structure. Furthermore, the integration of texture and gradient information is expected to improve the algorithm's ability to preserve lesion boundaries more accurately.

This study used dermoscopy images obtained from Dr. M. Djamil General Hospital in Padang as the research subjects. The ESLIC segmentation results were then used as input for a Convolutional Neural Network (CNN) model to evaluate the effect of segmentation quality on the performance of skin lesion classification. Thus, this study is expected to provide a theoretical contribution to the development of superpixel-based segmentation methods and a practical contribution to supporting a more accurate, faster, and more reliable melanoma early detection system.

Moreover, For CNNs, the detailed description begins with the architecture or layer configuration, which consists of the number of layers, filter size, pooling type, output layer, and other parameters. in addition, there are activation functions, namely the activation in the hidden layer (ReLU) and the activation in the output layer (Sigmoid for binary classification or softmax), as well as the training protocol, which consists of the number of epochs, batch size, optimizer, dataset (e.g., 277 images for the training and validation processes), and additional hyperparameters. Fig. 1 illustrates the CNN architecture used in this study.

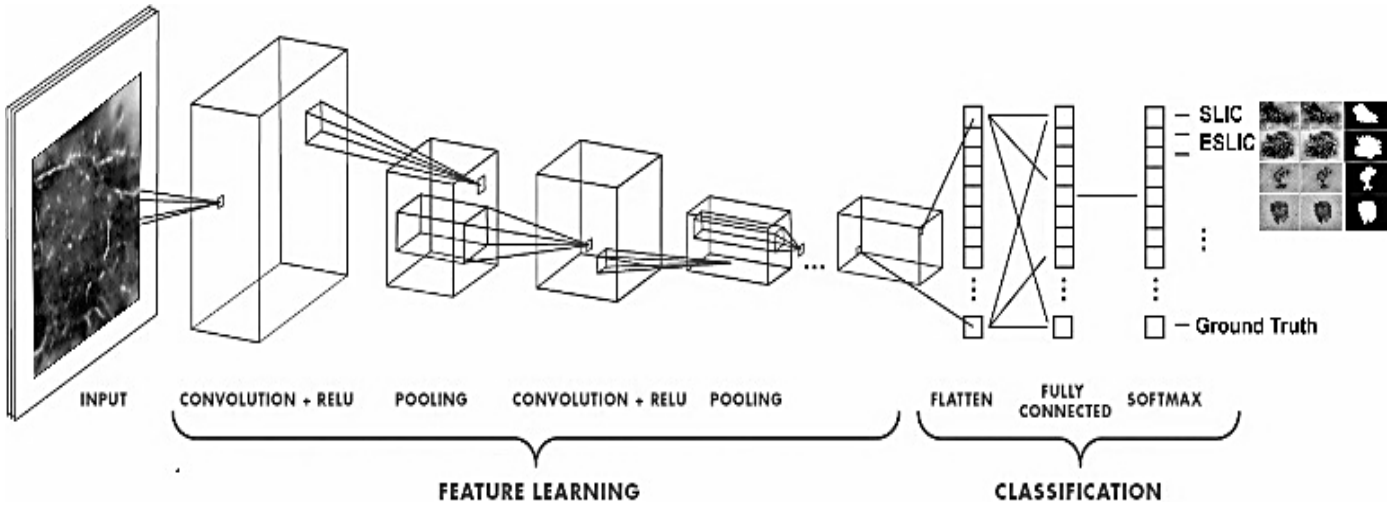


Fig. 1. Convolutional Neural Network (CNN) architecture

II. Related Work

There are key references indicating that previous research has been conducted by researchers on topics such as Superpixel-Oriented Label Distribution Learning [13], a hierarchical three-step superpixels and deep learning framework for skin lesion classification [31], transfer learning and hybrid feature extraction for medical segmentation [9], SAM -Adapters plus Vision Transformers for skin lesion segmentation [29], and SPADE, a superpixel adjacency method for 3-class melanoma segmentation. Specific details can be found in Table I.

TABLE I. RELATED WORK FROM PREVIOUS RESEARCH AND RELATED METHODS

Research Topics	Methods Used	Research findings	Researcher & Year
Superpixel-Oriented Label Distribution Learning for Skin Lesion Segmentation	The superpixels formed by the simple linear iterative cluster (SLIC) algorithm.	Datasets achieve a Dice coefficient reaching 84%, sensitivity 79.6%, precision 80.4%, improved by 19.3%, 8.6%, and 2.5%, respectively, in comparison with the results of U-Net.	Zhou et al. 2022 [13]
A hierarchical three-step superpixel and deep learning framework for skin lesion classification.	A deep learning model (ResNet-50) with three datasets (Ph2, ISBI2016, and HAM1000).	On the datasets, our method achieved an accuracy of 95.40%, 91.1%, and 85.50%,	F. Afza, et.al. 2022[31]
Advancing patient care with AI: a unified framework for medical image segmentation using transfer learning and hybrid feature extraction.	Trained U-Net models to segment skin cancer, polyps, and brain tumor regions from three separate MI datasets (HAM10000, Kvasir-SEG, and Figshare Brain Tumor dataset).	U-Net segmentation achieved high accuracy across datasets, with particularly notable results for polyps (98.00%) and brain tumors (99.66%). LBP features were extracted and classified using an SVM model, achieving 99.22% classification accuracy on the combined dataset (skin, polyp, and brain).	Nazife Çevik., N. 2025 [9]
Foundation-Model-Driven Skin Lesion Segmentation and Classification using SAM-Adapters and Vision Transformers	Integrated foundation-model-based framework that utilizes SAM-Adapter-fine-tuning for lesion segmentation and a ViT-based classifier that incorporates lesion-specific cropping derived from segmentation and cross-attention fusion.	From extensive experimentation, the proposed method outperforms the state-of-the-art segmentation and classification across the dataset. On the ISIC 2018 dataset, it achieves a Dice score of 94.27% for segmentation and an accuracy of 95.88% for classification performance. On PH2, a Dice score of 95.62% is achieved, and for HAM10000, an accuracy of 96.37% is achieved.	Binzagr., f. & Hariri., M. 2026 [29]

III. METHODS

A. SLIC

Enhanced Simple Linear Iterative Clustering (ESLIC) is a superpixel-based segmentation algorithm developed from the Simple Linear Iterative Clustering (SLIC) method. This development was undertaken to improve the ability to segment skin lesions in dermoscopy images, which typically exhibit irregular object boundaries, heterogeneous textures, and complex color variations. Although SLIC is known as an efficient superpixel method with low computational complexity, its performance remains limited when used to separate lesion areas from normal skin tissue, which exhibits very subtle visual differences. Therefore, ESLIC was designed to integrate color, spatial, texture, and gradient information into the superpixel formation process, thereby enabling more adaptive and precise segmentation.

In the initial stage, dermoscopy images that have undergone preprocessing are represented in the CIELAB color space. This color space was chosen because CIELAB is better suited to representing human visual perception than the RGB color space. Each pixel is represented as a feature vector consisting of a luminance component (L_i), a red-green color channel (a_i), a blue-yellow color channel (b_i), and spatial coordinates (x_i, y_i). This representation allows the algorithm to consider both color and pixel position information simultaneously during the segmentation process. Mathematically, the pixel representation [17] is expressed as Eq. (1):

$$P_i = [L_i, a_i, b_i, x_i, y_i] \quad (1)$$

Based on this representation, the color distance between a pixel and the cluster center is calculated using Eq. (2):

$$d_c = \sqrt{(L_i - L_k)^2 + (a_i - a_k)^2 + (b_i - b_k)^2} \quad (2)$$

This equation is used to measure the degree of color similarity between a pixel and the center of a superpixel. The smaller the color distance value obtained, the higher the degree of similarity in color characteristics between the two pixels.

In addition to color information, ESLIC also utilizes spatial information that describes the proximity of pixel positions within an image. Spatial distance is calculated based on the pixel's position coordinates relative to the cluster center using Eq. (3):

$$d_s = \sqrt{(x_i - x_k)^2 + (y_i - y_k)^2} \quad (3)$$

The use of spatial information aims to maintain the continuity of superpixel areas so that pixels that are geographically adjacent in the image are more likely to belong to the same cluster.

Unlike conventional SLIC, which uses a fixed compactness parameter, ESLIC introduces an adaptive compactness mechanism that allows the compactness value to vary according to the local texture characteristics of the image. This approach aims to enable the segmentation process to adapt to the varying lesion structures in different areas of the image. The adaptive compactness value is calculated using the Eq. (4):

$$m_i = m_0(1 + \alpha T_i) \quad (4)$$

In this equation, m_0 is the initial compactness value, is the adaptation factor, and T_i represents the local texture level at the i-th pixel. With this mechanism, areas with more complex textures will receive different compactness values compared to relatively homogeneous areas, making the segmentation more flexible in following the shape of the lesion.

To enhance the segmentation process, ESLIC integrates texture features extracted using the Local Binary Pattern (LBP) method. This method was chosen because it effectively

represents local texture patterns and has been widely used in various medical image processing applications. Once the LBP values are obtained, the texture distance between a pixel and the cluster center is calculated using Eq. (5):

$$d_t = | LBP_i - LBP_k | \quad (5)$$

The integration of texture features allows the algorithm to distinguish between lesion areas and normal skin tissue that share similar color characteristics but have different textures.

In addition to texture, ESLIC also utilizes gradient information to improve segmentation performance in preserving lesion boundaries. Gradient information is obtained using the Sobel operator, which calculates changes in intensity in the horizontal and vertical directions. The total gradient value is calculated using Eq. (6):

$$G = \sqrt{G_x^2 + G_y^2} \quad (6)$$

Next, the gradient distance between a pixel and the cluster center is calculated using Eq. (7):

$$d_g = | G_i - G_k | \quad (7)$$

By utilizing gradient features, ESLIC can identify areas of intensity transition that typically occur at lesion boundaries, allowing the resulting superpixels to follow the object's contours more accurately.

The main contribution of this research lies in the development of a new distance function that combines all color, spatial, texture, and gradient information into a single unified formulation. The ESLIC distance function is formulated as shown in Eq. (8):

$$D_{ESLIC} = \sqrt{d_c^2 + \left(\frac{m_i}{S}\right)^2 d_s^2 + \lambda d_t^2 + \gamma d_g^2} \quad (8)$$

In this equation, d_c represents color distance, d_s represents spatial distance, d_t represents texture distance, and d_g represents gradient distance. The parameter m_i represents adaptive compactness, while S is the interval between superpixel centers. The parameters λ and γ are used as weights that control the contribution of texture and gradient features to the clustering process.

The values of parameters λ and γ, in Eq. (8), were determined empirically through preliminary experiments on a small subset of N images randomly selected from a total of 277 dermoscopy images. The determination was carried out by testing candidate values of λ and γ within a specific range, while the parameters m_0 and α were kept at fixed values for the tested combinations; segmentation quality was evaluated using the Dice Similarity Coefficient (DSC) metric against the ground truth. The values of λ and γ were selected as the final values because they produced the best balance between the contributions of texture and gradient features—that is, a stable DSC value without excessive dominance of any single feature, which, if it occurred, would cause oversegmentation (γ too large) or failure to detect heterogeneous textures in lesions (λ too small).

After the superpixel generation process is complete, ESLIC performs a boundary refinement step to improve the quality of the segmentation boundaries. This step aims to remove any remaining minor noise and smooth the lesion contours so they more closely resemble the actual shape. The refinement process is performed using a combination of morphological operations, namely erosion and dilation. With this stage, lesion boundaries can be better preserved, and segmentation errors at the edges can be minimized.

Based on the proposed formulation and mechanism, ESLIC integrates adaptive compactness adjustment, texture-aware clustering, gradient-aware feature integration, and boundary refinement into a superpixel-based segmentation framework. The combination of these four components enables the formation of superpixels that are more representative of the morphological characteristics of skin lesions, thereby improving segmentation quality and supporting more accurate melanoma classification.

B. Comparison of ESLIC and SLIC

Simple Linear Iterative Clustering (SLIC) is a superpixel-based segmentation method widely used in image processing because it has low computational complexity and is capable of producing relatively uniform superpixels, as shown in Fig. 2. This method works by grouping pixels based on color similarity and spatial proximity, thereby simplifying the image representation without losing important information. However, the application of SLIC [18] to dermoscopy images still faces several limitations. The characteristics of skin lesions—which exhibit high color variation, uneven pigment distribution, heterogeneous texture, and irregular lesion boundaries—mean that conventional SLIC is not yet capable of producing optimal segmentation. Furthermore, the use of a fixed compactness parameter across the entire image area limits the algorithm's ability to adapt to changes in local characteristics.

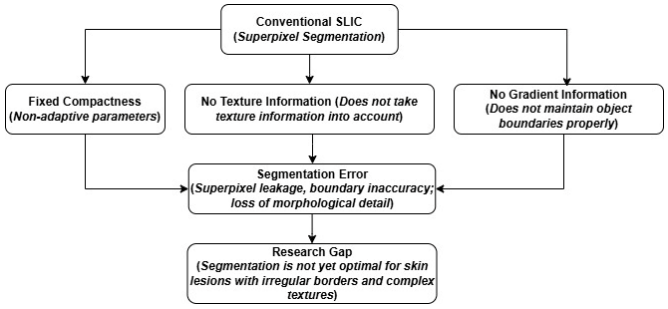


Fig. 2. Conventional SLIC framework

Furthermore, this study provides a comprehensive discussion of the behavior of the ESLIC algorithm. Some of the aspects of the method analyzed are: In the study [1], the convergence criteria used, namely that the iteration is stopped when the average shift in cluster center positions between iterations falls below the threshold ε or the number of iterations reaches the maximum limit of N and in [2], the average number of iterations required to achieve convergence across the entire dataset is 277 images. In [3], a computational stability analysis shows that the number of iterations is relatively consistent for images with homogeneous textures but tends to be higher for lesions with heterogeneous textures and low contrast.

Another limitation of SLIC is its reliance on color and spatial information as the sole basis for the clustering process. In dermoscopy images, lesion areas and normal skin tissue often

share similar colors but differ in texture. When texture information is not taken into account, pixels with similar colors tend to be grouped into the same superpixel even though they originate from different tissue structures. Furthermore, the absence of gradient information makes the algorithm less capable of accurately preserving lesion boundaries, especially in low-contrast areas. This can lead to "*superpixel leakage*"—where portions of normal skin are incorrectly included in the lesion area or vice versa—thereby affecting segmentation quality and reducing the performance of the classification system in subsequent stages. Moreover, to address these limitations, this study proposes Enhanced Simple Linear Iterative Clustering (ESLIC) as an extension of the conventional SLIC method. ESLIC is designed to integrate several new mechanisms: adaptive compactness adjustment, texture-aware clustering, gradient-aware feature integration, and boundary refinement. The adaptive compactness adjustment mechanism allows the compactness parameter to change dynamically based on the local texture characteristics of the image, making the segmentation process more adaptive to the complexity of the lesion. Furthermore, texture information is utilized to improve the algorithm's ability to distinguish between lesion areas and normal skin tissue, while gradient information is used to maintain object boundaries with greater precision. The boundary refinement stage is applied to smooth the segmentation results and reduce errors in the lesion edge areas. By integrating these various mechanisms, ESLIC is expected to produce segmentation results that are more accurate, stable, and representative of the morphological characteristics of skin lesions compared to the conventional SLIC method. The ESLIC framework is shown in detail in Fig. 3.

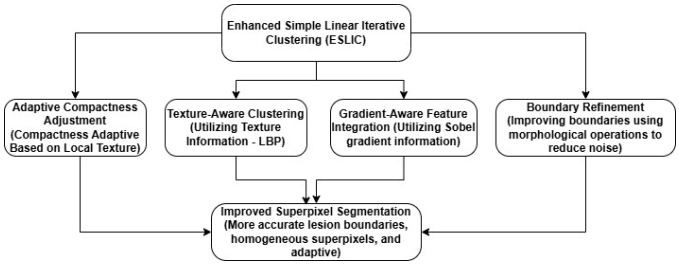


Fig. 3. ESLIC framework

Moreover, to illustrate the limitations of the SLIC method, the distance function used in the superpixel formation process can be expressed as $D = \sqrt{d_c^2 + \left(\frac{m}{s}\right)^2 d_s^2}$, where d_c represents color distance, d_s indicates spatial distance, m is a measure of compactness, and S is the distance between the centers of superpixels. The equation shows that SLIC considers only two main components, namely color and spatial proximity information, in the pixel clustering process. Although this approach is capable of producing computationally efficient segmentation, the lack of texture and gradient information makes the algorithm less capable of representing the complexity of dermoscopy images, which typically exhibit heterogeneous texture patterns, uneven pigment distribution, and irregular morphological lesion boundaries. As a result, the superpixels that are formed often do not optimally follow the contours of the lesion and may lead to segmentation errors, especially at object boundaries with low contrast or complex visual characteristics.

TABLE II. SUMMARY OF HYPERPARAMETERS FOR THE ESLIC IMPLEMENTATION.

Phase / Component	Parameter
Preprocessing	
Contrast stretching	Clipping percentile range (e.g., lower–upper percentile)
Sharpening	Type of kernel / sharpening weight (e.g., unsharp masking)
Top-hat filtering	Size and shape of structuring elements
Normalization	Target intensity range
Feature Extraction	
Local Binary Pattern (LBP)	Radius
Local Binary Pattern (LBP)	Number of neighboring points (P)
Sobel operator	Kernel size
Boundary Refinement	
Morphological operation	Size of structural elements (erosion and dilation)
ESLIC Distance Function	
Start of Compactness	m0
Adaptation factors	α (alpha)
Texture feature weights	λ (lambda)
Gradient feature weights	γ (gamma)
Interval superpixel center	S
Iterative Convergence	
Convergence criteria	Cluster center shift threshold / maximum iterations
Average number of iterations	Mean ± SD across the entire dataset (277 images)

Moreover, to achieve the established research objectives, a structured workflow is required so that each stage of image processing and classification can be carried out systematically. Hyperparameters for the ESLIC implementation are detailed in Table II. This study proposes the Enhanced Simple Linear Iterative Clustering (ESLIC) method as an extension of the conventional SLIC algorithm to improve the quality of skin lesion segmentation in dermoscopy images. The proposed method integrates color, texture, gradient, and spatial information through mechanisms of adaptive compactness adjustment, feature integration, iterative clustering, and boundary refinement. The resulting segmentation is then used as input for a Convolutional Neural Network (CNN) model for skin lesion classification [19], [20]. Overall, the research stages can be seen in Fig. 4. Fig. 4 illustrates that this study begins with dermoscopy images that first undergo preprocessing steps, including grayscale conversion, top-hat filtering, contrast stretching, normalization, and sharpening [21], [22]. Next, the images were processed using the ESLIC algorithm, which consists of superpixel initialization, adaptive compactness, integration of texture and gradient features, iterative clustering, and boundary refinement to produce more accurate lesion segmentation. The segmentation results are then used as input for a CNN to classify melanoma. The system's performance is evaluated using accuracy, precision, recall, and F1-score to arrive at a final decision regarding the type of skin lesion [23]. The algorithms (Pseudocode 1) used in this stage are as follows:

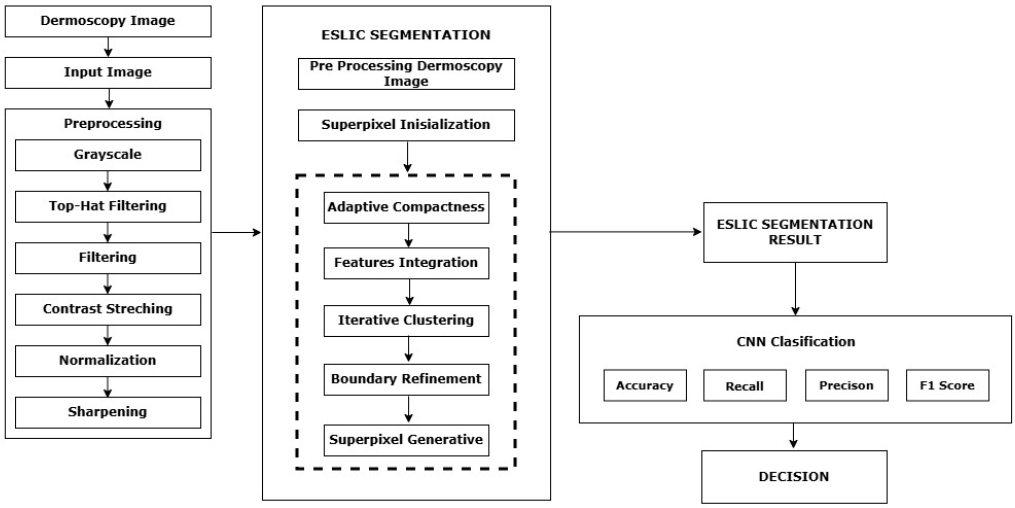


Fig. 4. Research framework

Pseudocode 1:

1. **Input:**
2. Dermoscopic image
3. **Output:**
4. Segmented lesion image S
5. Preprocess the image I
6. Transform the image into the Lab color space
7. Initialize K superpixel centers
8. Extract texture descriptor
9. T = LBP(I)
10. Extract gradient descriptor
11. G = Sobel(I)
12. Compute adaptive compactness
13. mi = m0(1+αTi)
14. For each pixel p do
15. Compute color distance dc
16. Compute spatial distance ds
17. Compute texture distance dt
18. Compute gradient distance dg
19. Compute
20. DESLIC = $\sqrt{(dc^2 + (mi/S)^2 ds^2 + \lambda dt^2 + \gamma dg^2)}$
21. Assign p to the nearest cluster
22. End For
23. Update cluster centers
24. Repeat until convergence
25. Refine the lesion boundary using morphological operations
26. Generate final superpixel
27. Return segmented lesion S

---------- *Pseudocode_1*----------

IV. RESULTS

The first step in this study was to preprocess the dermoscopy images before the segmentation process. The main objective of this step was to improve image quality so that lesion characteristics could be better represented during the segmentation stage. The preprocessing steps used included grayscale conversion, top-hat filtering, contrast stretching, normalization, and sharpening [24]. Grayscale conversion is performed to simplify the image representation and reduce computational complexity [25]. Next, top-hat filtering is used to reduce the effects of artifacts such as hair and lighting variations that frequently appear in dermoscopy images [26]. Following that, contrast stretching is applied to enhance the intensity difference between the lesion area and normal skin tissue [27]. The normalization stage [28], [29] is applied to standardize the image's intensity range, while the sharpening process aims to clarify lesion edges so that object boundaries are more easily recognized during segmentation.

Fig. 5 shows that each preprocessing step improves the visual quality of the image. After sharpening, the lesion boundaries become clearer, and the contrast between the lesion area and the surrounding tissue increases significantly. Once the preprocessing steps are complete, the image is processed using the Enhanced Simple Linear Iterative Clustering (ESLIC) algorithm. This algorithm was developed from the SLIC method by adding mechanisms for adaptive compactness, texture-aware clustering, gradient-aware feature integration, and boundary refinement. The first stage in ESLIC is superpixel initialization. In this stage, the number of superpixels and the initial positions of the cluster centers are determined. Next, an adaptive compactness adjustment is performed to adjust the compactness parameter based on the local texture characteristics of the image. This approach allows for more flexible superpixel formation compared to conventional SLIC. In the next stage, color, texture, gradient, and spatial information features are integrated into the ESLIC distance function. The combination of these features enables the clustering process to distinguish between lesion areas and normal skin tissue more accurately. After the iterative clustering process is complete, a boundary refinement stage is applied to improve the segmentation boundaries and reduce errors in the lesion edge areas.

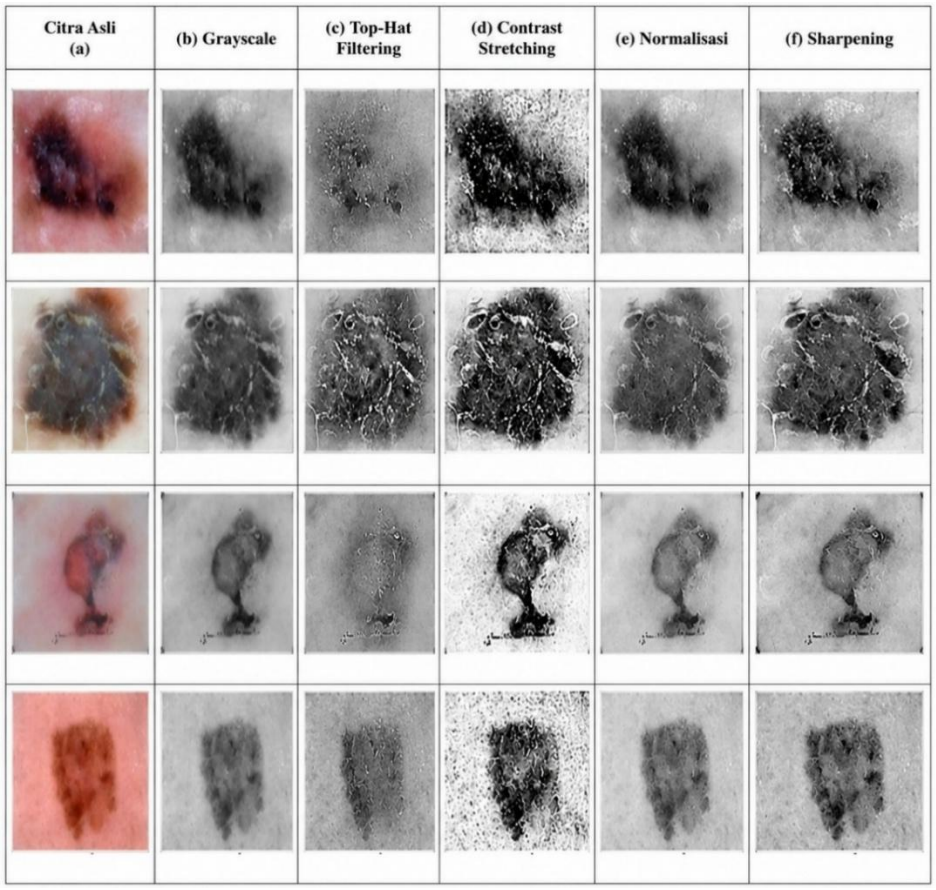


Fig. 5. Pre-processing stage results

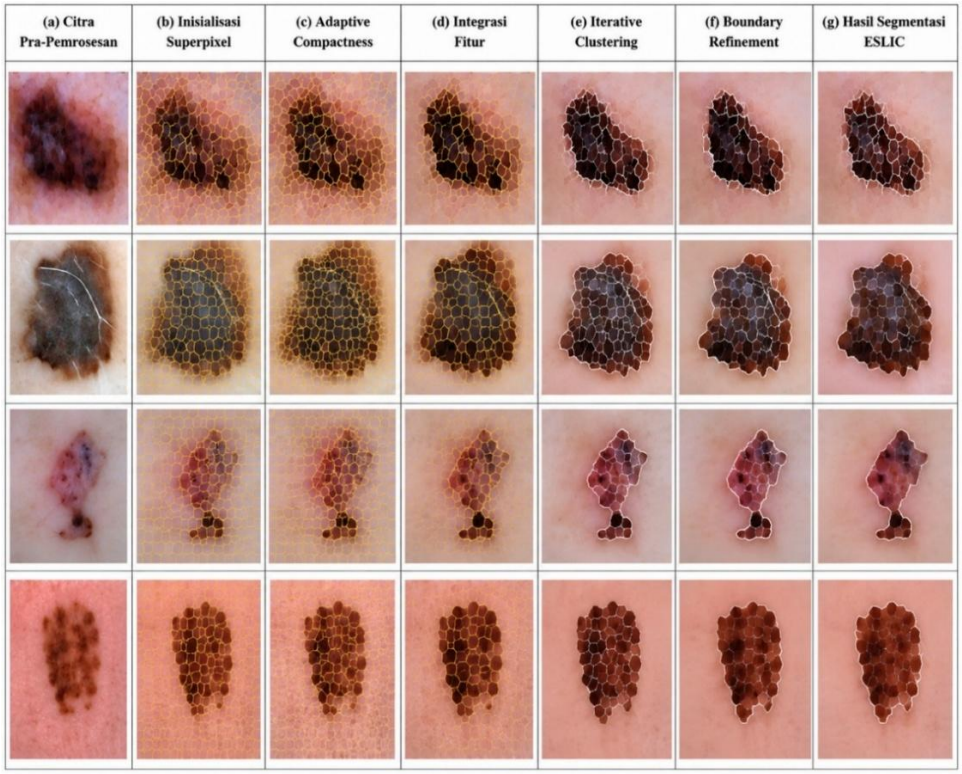


Fig. 6. Segmentation stages using ESLIC.

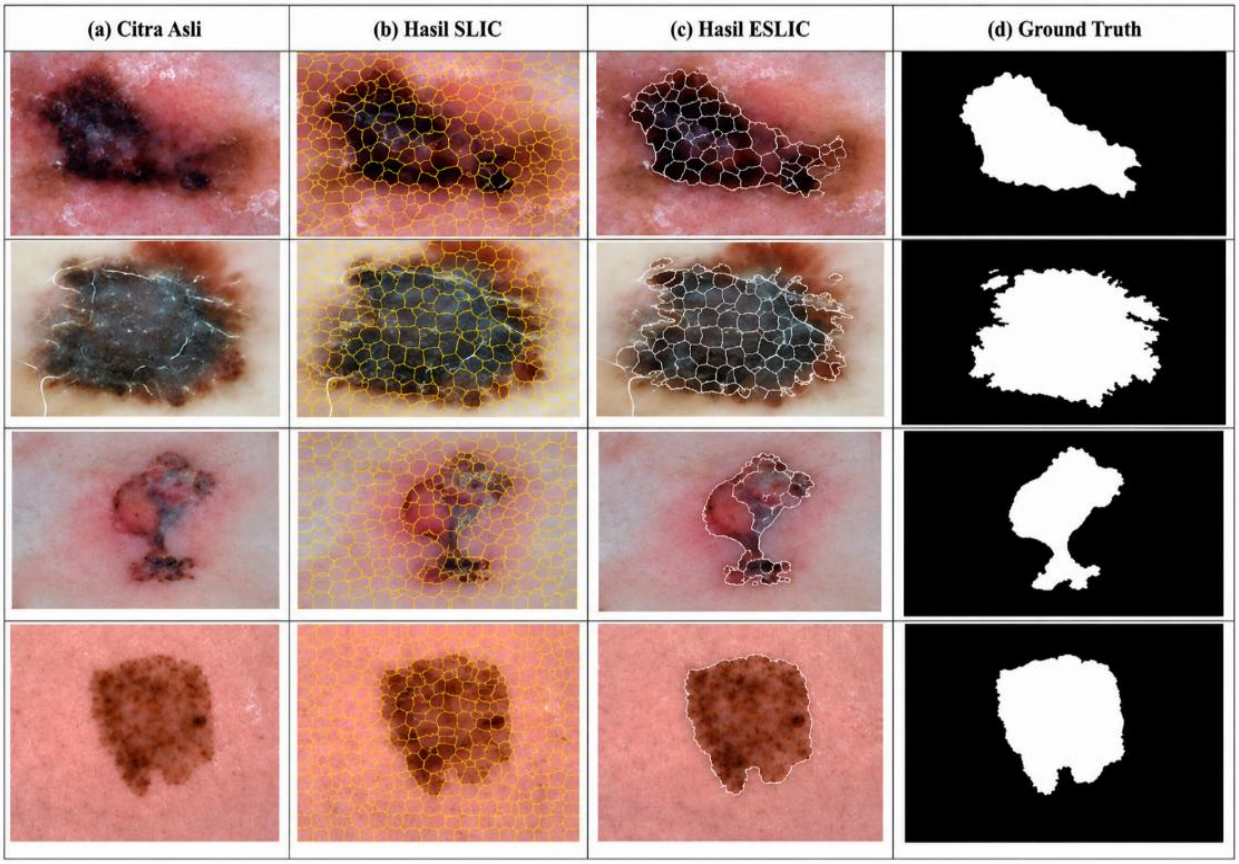


Fig. 7. Comparison of the segmentation results.

Fig. 6 shows that the superpixels generated by ESLIC are able to better follow the morphological structure of the lesion. The boundary refinement stage produces smoother lesion boundaries and reduces superpixel leakage at the object edges. To evaluate the effectiveness of the proposed method, a visual comparison was conducted between conventional SLIC and ESLIC. Comparison of segmentation results is shown in Fig. 7. The segmentation results using SLIC show that some superpixels still straddle lesion boundaries, particularly in areas with complex textures and low contrast. In contrast, ESLIC is able to better preserve lesion contours because it takes texture and gradient information into account during the clustering process. The segmentation results produced by ESLIC appear to be closer to the ground truth than those from the SLIC method. Quantitative evaluation was performed using the Dice Similarity Coefficient (DSC), Jaccard Index (IoU), Boundary F1-Score, and Undersegmentation Error. The detailed segmentation evaluation results are shown in Table III.

TABLE III. SEGMENTATION EVALUATION RESULTS

Method	Dice (%)	IoU (%)	Boundary F1 (%)	Undersegmentation Error
SLIC	87.54	79.33	84.72	0.18
ESLIC	94.26	89.11	92.85	0.07

Table III shows that ESLIC yields improved performance across all evaluation metrics. The Dice score increased by 6.72%, while the IoU increased by 9.78% compared to the SLIC method. Furthermore, the reduction in the Undersegmentation Error indicates that ESLIC is more effective at preserving lesion boundaries during the segmentation process. Quantitative evaluation was performed using the Dice Similarity Coefficient (DSC), Jaccard Index (IoU), Boundary F1-Score, and Undersegmentation Error.

Fig. 8 shows a comparison of the Dice Similarity Coefficient (DSC) and Intersection over Union (IoU) values obtained using the SLIC and ESLIC methods. It can be seen that the ESLIC method yields a Dice value of 94.26% and an IoU of 89.11%, which are higher than those of the SLIC method, which yields a Dice value of 87.54% and an IoU of 79.33%. This improvement indicates that ESLIC is capable of producing more accurate segmentation and has a higher level of agreement with the ground truth. The integration of adaptive compactness, texture-aware clustering, gradient-aware feature integration, and boundary refinement has proven effective in improving the quality of skin lesion segmentation in dermoscopy images. To determine the contribution of each proposed component, a stepwise test was conducted by adding each ESLIC component sequentially.

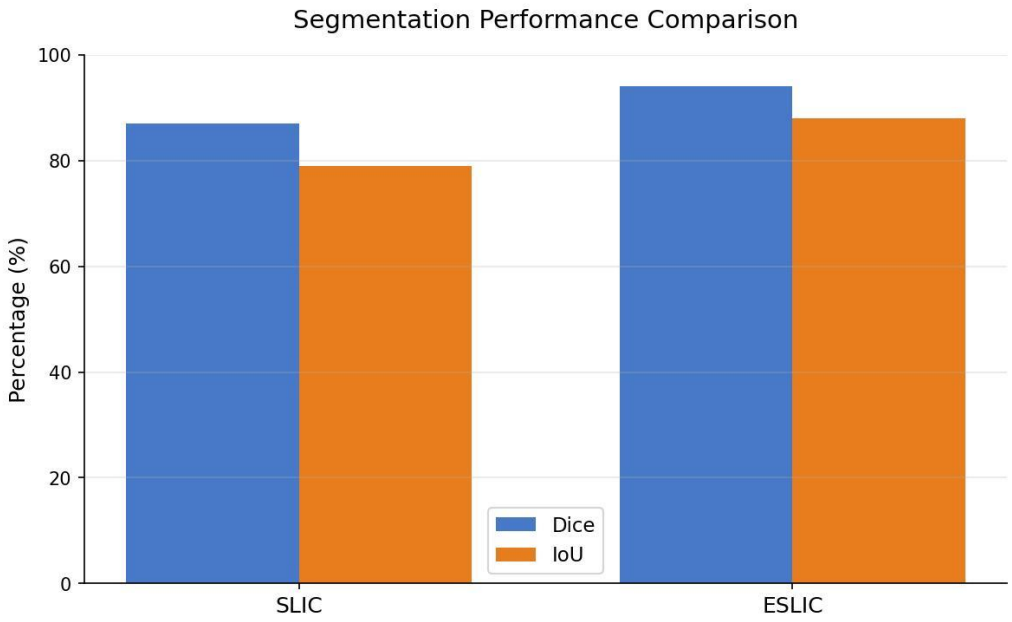


Fig. 8. Comparison of segmentation performance between SLIC and ESLIC.

TABLE IV. ABLATION STUDY OF ESLIC

Method	Dice (%)
SLIC	87.54
SLIC + Adaptive Compactness	89.21
SLIC + Adaptive Compactness + Texture	91.63
SLIC + Adaptive Compactness + Texture + Gradient	93.08
Full ESLIC	94.26

Table IV shows a step-by-step breakdown of the contributions of each ESLIC component, starting with the base SLIC, followed by the addition of adaptive compactness, texture-aware clustering (LBP), gradient-aware feature integration (Sobel), and finally boundary refinement. This approach allows us to demonstrate an increase in the Dice score from 87.54% for SLIC to 94.26% for Full ESLIC. This is not the result of combining all methods, but rather the result of the incremental contributions of each mechanism implemented, with texture feature integration providing the most significant improvement—from 87.54% to 91.63%—followed by gradient and boundary refinement. The test results show that each component contributes to improved segmentation quality. The integration of texture features yields a fairly significant improvement, while the addition of gradient information and boundary refinement results in better preservation of lesion boundaries. The next step is to test the effect of segmentation quality on the classification process using a CNN [30]. In this step, the segmented lesion areas are used as input for the CNN model to classify lesions as benign or malignant.

Table V shows that segmentation using ESLIC yields improved classification performance compared to SLIC. This indicates that better segmentation quality results in a more informative representation of lesions, allowing the CNN to extract discriminative features more effectively.

TABLE V. CNN CLASSIFICATION RESULTS

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CNN + SLIC	90.12	89.76	89.04	89.40
CNN + ESLIC	96.03	95.88	95.31	95.59

The experimental results show that the integration of adaptive compactness adjustment, texture-aware clustering, gradient-aware feature integration, and boundary refinement successfully improves the quality of skin lesion segmentation in dermoscopy images. Adaptive compactness allows for the formation of superpixels that are more adaptive to local texture variations, while the integration of texture and gradient features enhances the algorithm's ability to preserve complex lesion boundaries. Furthermore, the boundary refinement stage was shown to reduce segmentation errors in object edge regions. This improvement in segmentation quality directly impacts the CNN's classification performance, as evidenced by increases in accuracy, precision, recall, and F1-score. Thus, ESLIC can be considered an effective extension of the SLIC method for AI-based skin lesion segmentation and early melanoma detection applications.

Furthermore, significance tests or statistical analysis are required to determine the percentage (%) of the Descriptive Statistics (Dice) similarity coefficient, as shown in Fig. 9 and Fig. 10. From Fig. 9 and Fig. 10, we find that Descriptive Statistics (Dice %) for SLIC: mean = 87.16, standard deviation = 3.90; for ESLIC: mean = 94.05, standard deviation = 1.80. Lilliefors normality test on paired differences: $p = 0.0093$, shows the distribution of differences is NOT normal; use the Wilcoxon signed-rank test. And based on the data Paired t-test, the value was obtained $t(29) = 8.691$, $p = 0.0000$, 95% CI of mean difference = [5.27, 8.51], as well as the results of Wilcoxon signed-rank test $p = 0.0000$, obtained Cohen's d (paired) = 1.587,

with an interpretation, the improvement in performance with ESLIC compared to SLIC is statistically significant ($p < 0.05$) based on both tests.

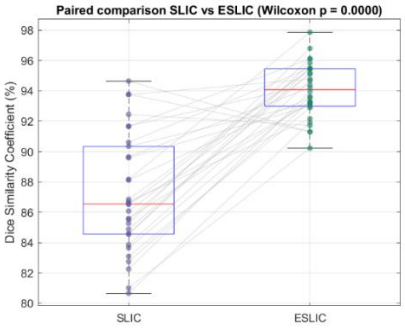


Fig. 9. Paired comparison SLIC vs. ESLIC.

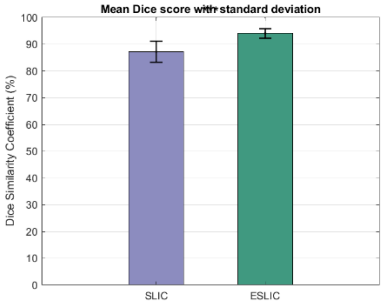


Fig. 10. Mean Dice score with standard deviation.

V. Conclusion

The ESLIC algorithm can be further developed by integrating deep learning-based features, attention mechanisms, or modern segmentation models such as U-Net and Mask R-CNN. In addition, testing on large-scale dermoscopy datasets such as the ISIC Archive is also necessary to evaluate the generalization ability of the proposed method across various skin lesion characteristics.

- This study successfully developed the Enhanced Simple Linear Iterative Clustering (ESLIC) method as an extension of the conventional SLIC algorithm for skin lesion segmentation in dermoscopy images. This development was achieved through the integration of adaptive compactness adjustment, texture-aware clustering, gradient-aware feature integration, and boundary refinement to improve segmentation performance in lesion areas with complex characteristics.
- The integration of color, texture, gradient, and spatial information into the ESLIC distance function enables the generation of superpixels that are more adaptive to variations in skin lesion morphology. This approach has been shown to improve the algorithm's ability to preserve lesion boundaries and reduce the phenomenon of superpixel leakage, which is still commonly found in conventional SLIC methods.
- Segmentation evaluation results show that ESLIC performs better than SLIC. The proposed method achieved a Dice Similarity Coefficient (DSC) of 94.26% and an Intersection over Union (IoU) of 89.11%, indicating a high level of agreement with the ground truth of skin lesion segmentation.
- Testing the contribution of algorithm components shows that each mechanism added to ESLIC improves segmentation performance. The combination of adaptive compactness, texture features, gradient features, and boundary refinement collectively produces more stable and accurate segmentation quality compared to the standard SLIC method.
- Better segmentation quality has a positive impact on the Convolutional Neural Network (CNN)-based melanoma classification process. Using ESLIC segmentation results as CNN input improves classification accuracy to 96.03%, demonstrating that ESLIC has high potential for application in AI-based early melanoma detection systems to support more accurate and efficient diagnosis.

VI. Future Research

A dataset of 277 dermoscopy images is relatively small for drawing broad conclusions about the algorithm's generalizability. We obtained this data from Dr. M. Djamil General Hospital in Padang, which includes cases of benign lesions and melanoma; however, further validation on a larger and more diverse dataset is needed to confirm the robustness and generalization of ESLIC. Testing on large-scale public datasets such as the ISIC Archive is an important step toward a comprehensive evaluation. There are still limitations to this research: the comparison in this study is limited to conventional superpixel methods (SLIC) and does not yet include deep learning-based segmentation architectures such as U-Net or Mask R-CNN, which have demonstrated higher Dice/IoU scores on dermoscopy datasets. In our discussion, we position ESLIC not as a direct competitor to deep learning methods, but rather as a computationally lighter approach that does not require large amounts of pixel-annotated mask data for training. The strength of this research is that it includes 277 images, and the next step is to directly compare it with U-Net and Mask R-CNN on the same dataset.

Future analysis needs to be significantly improved; ESLIC needs to produce segmentation that is even more accurate than the ground truth, and issues such as low contrast between lesions and skin, as well as the preprocessing stage and lesion boundaries, still need to be addressed to improve image quality, such as low contrast, light reflection, and diffuse lesion borders. Future research will also conduct a detailed comparative analysis of the average Dice scores and IoU by category to highlight significant differences between SLIC and ESLIC, as well as other more detailed parameters.

References

[1] S. Kumar, "Early Disease Detection Using AI: A Deep Learning Approach to Predicting Cancer and Neurological Disorders," Int. J. Sci. Res. Manag., vol. 13, no. 04, pp. 2136–2155, Apr. 2025, doi: 10.18535/ijsrm/v13i04.mp02.

[2] S. Fatima, M. U. Akram, S. Mohammad, and S. Bin Ahmed, "Deep learning in dermatopathology: applications for skin disease diagnosis and classification," Sep. 01, 2025, Springer Nature. doi: 10.1007/s42452-025-07138-3.

[3] P. Weber, P. Tschandl, C. Sinz, and H. Kittler, "Dermatoscopy of Neoplastic Skin Lesions: Recent Advances, Updates, and Revisions," Nov. 01, 2018, Springer New York LLC. doi: 10.1007/s11864-018-0573-6.

[4] H. Eghtesaddoust, M. Valizadeh, and M. C. Amirani, "Fusion of Deep and Time–Frequency Local Features for Melanoma Skin Cancer Detection," Appl. Comput. Intell. Soft Comput., vol. 2025, no. 1, 2025, doi: 10.1155/acis/4767052.

[5] P. Ordóñez, Y. Xie, X. Zhang, C. Y. Xie, S. Acosta, and I. Gutiérrez, "SPADE: Superpixel Adjacency Driven Embedding for Three-Class Melanoma Segmentation," Algorithms, vol. 18, no. 9, Sep. 2025, doi: 10.3390/a18090551.

[6] M. A. Kassem, K. M. Hosny, R. Damaševičius, and M. M. Eltoukhy, "Machine learning and deep learning methods for skin lesion classification and diagnosis: A systematic review," Aug. 01, 2021, Multidisciplinary Digital Publishing Institute (MDPI). doi: 10.3390/diagnostics11081390.

[7] F. Hu et al., "Advances in intelligent recognition and diagnosis of skin scar images: concepts, methods, challenges, and future trends," 2025, Frontiers Media SA. doi: 10.3389/fmed.2025.1667087.

[8] A. R. Ali, J. Li, and S. J. O'Shea, "Towards the automatic detection of skin lesion shape asymmetry, color variegation and diameter in dermoscopic images," PLOS ONE, vol. 15, no. 6, Jun. 2020, doi: 10.1371/journal.pone.0234352.

[9] N. Çevik, T. Çevik, O. Osman, S. Alsubai, and J. Rasheed, "Advancing patient care with AI: a unified framework for medical image segmentation using transfer learning and hybrid feature extraction," Front. Med., vol. 12, 2025, doi: 10.3389/fmed.2025.1589587.

[10] J. A. Fee, F. P. McGrady, and N. D. Hart, "Dermoscopy use in primary care: a qualitative study with general practitioners," BMC Prim. Care, vol. 23, no. 1, Dec. 2022, doi: 10.1186/s12875-022-01653-7.

[11] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 11, pp. 2274–2281, 2012, doi: 10.1109/TPAMI.2012.120.

[12] S. Adamu et al., "The future of skin cancer diagnosis: a comprehensive systematic literature review of machine learning and deep learning models," 2024, Cogent OA. doi: 10.1080/23311916.2024.2395425.

[13] Q. Zhou, T. He, and Y. Zou, "Superpixel-Oriented Label Distribution Learning for Skin Lesion Segmentation," Diagnostics, vol. 12, no. 4, Apr. 2022, doi: 10.3390/diagnostics12040938.

[14] D. Xu, T. Katzke, and E. Müller, "From Pixels to Graphs: Deep Graph-Level Anomaly Detection on Dermoscopic Images," Aug. 2025, [Online]. Available: http://arxiv.org/abs/2508.11826

[15] A. Jiménez-Sánchez et al., "In the Picture: Medical Imaging Datasets, Artifacts, and their Living Review," in ACMF AccT 2025 - Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, Inc, Jun. 2025, pp. 511–531. doi: 10.1145/3715275.3732035.

[16] A. H. Roky et al., "Overview of skin cancer types and prevalence rates across continents," Mar. 01, 2025, Chinese Medical Association. doi: 10.1016/j.cpt.2024.08.002.

[17] T. Nagaoka, "Improved Skin Lesion Segmentation in Dermoscopic Images Using Object Detection and Semantic Segmentation," Clin. Cosmet. Investig. Dermatology , vol. 18, pp. 1191–1198, 2025, doi: 10.2147/CCID.S518751.

[18] K. Suwais, K. Alheeti, and D. Al_Dosary, "A Review on Classification Methods for Plants Leaves Recognition," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 2, pp. 92–100, 2022, doi: 10.14569/IJACSA.2022.0130211.

[19] S. R. Waheed et al., "Melanoma Skin Cancer Classification based on CNN Deep Learning Algorithms," Malaysian J. Fundam. Appl. Sci., vol. 19, no. 3, pp. 299–305, 2023, doi: 10.11113/mjfas.v19n3.2900.

[20] H. Naseri and A. A. Safaei, "Diagnosis and prognosis of melanoma from dermoscopy images using machine learning and deep learning: a systematic literature review," Dec. 01, 2025, BioMed Central Ltd. doi: 10.1186/s12885-024-13423-y.

[21] H. T. Halawani, E. M. Senan, Y. Asiri, I. Abunadi, A. M. Mashraqi, and E. A. Alshari, "Enhanced early skin cancer detection through fusion of vision transformer and CNN features using hybrid attention of EViT-Dens169," Sci. Rep., vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-18570-1.

[22] L. N. Zulita, S. Defit, and Sumijan, "Detecting Skin Threats: A Superpixel-Based Analysis of Melanoma," in ICAISD 2025 - 2025 International Conference on Advanced Information Scientific Development: Artificial Intelligence: Advancing Research and Computational Innovations for Global Welfare, Proceedings, Institute of Electrical and Electronics Engineers Inc., 2025, pp. 78–83. doi: 10.1109/ICAISD68166.2025.11385687.

[23] A. A. Abdulredah, M. A. Fadhel, L. Alzubaidi, Y. Duan, M. Kherallah, and F. Charfi, "Towards unbiased skin cancer classification using deep feature fusion," BMC Med. Inform. Decis. Mak., vol. 25, no. 1, Dec. 2025, doi: 10.1186/s12911-025-02889-w.

[24] S. A. Alzakari et al., "LesionNet: an automated approach for skin lesion classification using SIFT features with customized convolutional neural network," Front. Med., vol. 11, 2024, doi: 10.3389/fmed.2024.1487270.

[25] N. Sarwar, A. Irshad, Q. H. Naith, K. D.Alsufiani, and F. A. Almalki, "Skin lesion segmentation using deep learning algorithm with ant colony optimization," BMC Med. Inform. Decis. Mak., vol. 24, no. 1, Dec. 2024, doi: 10.1186/s12911-024-02686-x.

[26] S. M. Thwin and H. S. Park, "Enhanced Skin Lesion Segmentation and Classification Through Ensemble Models," Eng, vol. 5, no. 4, pp. 2805–2820, Dec. 2024, doi: 10.3390/eng5040146.

[27] H. A. Owida et al., "A deep learning-based dual-branch framework for automated skin lesion segmentation and classification via dermoscopic Images," Sci. Rep., vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-21783-z.

[28] A. S. Al-Waisy, S. Al-Fahdawi, M. I. Khalaf, M. A. Mohammed, B. Al-Attar, and M. N. Al-Andoli, "A deep learning framework for automated early diagnosis and classification of skin cancer lesions in dermoscopy images," Sci. Rep., vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-15655-9.

[29] F. Binzagr and M. Hariri, "Foundation-Model-Driven Skin Lesion Segmentation and Classification Using SAM-Adapters and Vision Transformers," Diagnostics, vol. 16, no. 3, Feb. 2026, doi: 10.3390/diagnostics16030468.

[30] D. Moturi, R. K. Surapaneni, and V. S. G. Avanigadda, "Developing an efficient method for melanoma detection using CNN techniques," J. Egypt. Natl. Canc. Inst., vol. 36, no. 1, 2024, doi: 10.1186/s43046-024-00210-w.

[31] F. Afza, M. Sharif, M. Mittal, M. A. Khan, and D. J. Hemanth, "A hierarchical three-step superpixels and deep learning framework for skin lesion classification," Methods, vol. 202, pp. 88–102, Jun. 2022, doi: 10.1016/j.ymeth.2021.02.013.

Adaptive Mirage Search Optimization for Educational Workload Scheduling in Multi-Cloud Environments

Mohammed A.S. Mosleh[1], Elham Alzain[2]*, Hesham M.A. Abdullah[3], Ali Alshebami[4]
Al-Fayha Private College, Jubail, Saudi Arabia[1]
Applied College, King Faisal University, Al-Ahsa 31982, Saudi Arabia[2]
Al-Fayha Private College, Jubail, Saudi Arabia[3]
Applied College, King Faisal University, Al-Ahsa 31982, Saudi Arabia[4]

***Abstract*—AI-driven educational platforms increasingly rely on heterogeneous cloud workloads, including real-time tutoring, adaptive assessments, and large-scale learning analytics, which impose diverse and conflicting quality-of-service (QoS) requirements. Conventional cloud scheduling approaches fail to address such heterogeneity, often leading to service-level agreement (SLA) violations, increased execution cost, and poor resource utilization. This study proposes QoS-aware Adaptive Mirage Search Optimization Framework (QoS-AMSO), a QoS-aware multi-objective task scheduling framework based on an enhanced Mirage Search Optimization algorithm for intelligent workload management in multi-cloud environments. The proposed approach integrates four key innovations: adaptive refraction-based step sizing for improved convergence, QoS-feedback-driven dynamic weight adjustment, Lévy flight-based diversity preservation, and Pareto-based elite solution selection. Additionally, a task classification model categorizes workloads into latency-critical, throughput-intensive, deadline-soft, and batch-deferred classes, enabling context-aware scheduling decisions. The framework is evaluated using CloudSim 3.0 simulations with three cloud providers, 60 virtual machines, and workload sizes ranging from 100 to 1000 educational tasks under multiple workload scenarios. The QoS-AMSO achieves the best overall performance, reducing makespan to 798 s, execution cost to USD 292.8, SLA violation rate to 3.9%, and improving resource utilization to 97.2%. These results demonstrate the effectiveness and scalability of QoS-AMSO for QoS-aware scheduling in AI-powered educational multi-cloud environments.**

***Keywords*—Multi-cloud computing; task scheduling; Quality of Service (QoS); multi-objective optimization; mirage search optimization; AI-powered education; SLA-aware scheduling; metaheuristic algorithms**

I. Introduction

The advancements of artificial intelligence (AI) technologies have transformed modern educational systems, enabling the development of intelligent, scalable, and personalized learning analytics that require substantial computational resources to process diverse and dynamic workloads. Yet, the integration of AI-driven educational workloads with multi-cloud infrastructures introduces challenges in resource management and task scheduling [1]. The heterogeneous educational workloads exhibit diverse QoS requirements. In batch-oriented tasks, such as performance analytics and report generation, high throughput and resource-intensive execution are required. This heterogeneity results in conflicting optimization objectives, including decreasing execution time, operation cost, ensuring SLA compliance, and maintaining balanced resource utilization. Conventional cloud task scheduling approaches, including heuristic and metaheuristic algorithms such as Round Robin, Min-Min, Particle Swarm Optimization (PSO), and Genetic Algorithms (GA), are applied for resource allocation [2]. Many existing schedulers are developed for single-objective optimization, without considering the multi-dimensional nature of QoS requirements [3]. Most existing approaches rely on static objective weighting, fixed scheduling priorities, or heuristic decision strategies that are unable to adapt to continuously changing workload characteristics in multi-cloud environments. Although recent metaheuristic schedulers improve global search capability, they lack workload-aware decision-making and dynamic QoS adaptation, often resulting in suboptimal trade-offs among execution time, execution cost, SLA compliance, and resource utilization. The standard Mirage Search Optimization (MSO) algorithm provides an effective exploration–exploitation mechanism; however, it is a single-objective optimizer and does not incorporate task classification, runtime QoS feedback, adaptive search control, or Pareto-based solution selection required for heterogeneous educational workloads. Existing QoS-aware scheduling techniques and conventional MSO-based optimization remain inadequate for intelligent multi-cloud scheduling under dynamic educational environments. To address these challenges, this study proposes a QoS-aware Adaptive Mirage Search Optimization framework (QoS-AMSO) for intelligent task scheduling in AI-driven educational multi-cloud environments. The contribution of this research is a QoS-aware task classification model that categorizes educational workloads into latency-critical, throughput-intensive, deadline-soft, and batch-deferred classes, enabling context-aware scheduling decisions.

A QoS-AMSO algorithm enhances the standard MSO through adaptive refraction-based step sizing to improve convergence behavior in complex search spaces. A dynamic QoS feedback mechanism that adjusts objective weights in real time based on SLA violations, enabling responsive and adaptive scheduling. The model integrates a Lévy flight-based perturbation strategy to maintain population diversity and prevent premature convergence. A Pareto-based elite archive mechanism for effective multi-objective optimization, ensuring balanced trade-offs among competing QoS metrics.

The remainder of this study is organized as follows. Section II reviews existing works; Section III presents the system

*Corresponding author.

model and formulates the multi-objective optimization problem; Section IV describes the proposed QoS-AMSO framework; Section V discusses the experimental setup and evaluation, and Section VI concludes the study with future research.

II. Related Works

A. Cloud Task Scheduling Approaches

Rajput et al. [4] proposed a QoS-aware heuristic-based multi-cloud workflow scheduler for Spark tasks, achieving a 50% performance improvement, but it degrades under volatile conditions. Mangalampalli et al. [5] introduced ATS-IA3C, an A3C-based adaptive scheduler that improves makespan by 70.49% and cost by 77.42%, though it has high computational complexity. Mangalampalli et al. [6] proposed MOPTSA3C, which incorporates task and VM priorities into A3C, improving makespan and utilization but adding overhead due to priority computation. Kumar et al. [7] developed EAEFA based on electric fish swarm behavior, achieving over 30% makespan improvement in large-scale scenarios. Sandhu et al. [8] introduced TBW optimization combining Tabu Search, Bayesian classification, and WOA, achieving 95% efficiency but increasing computational complexity.

Acharya et al. [9] developed a Golden Search WOA (GSWOA) and attained 132s Makespan, 2.7s response time, and 74.34% RU. However, the GSWOA model has higher memory usage. Bansal et al. [10] presented a Hybrid PSO-Grasshopper Optimization Algorithm (HPSO-GOA) and attained 1636.44 total execution cost and 167.54J EC. Yet, the model is sensitive to initial population characteristics. Nagalakshmi et al. [11] presented an Energy-Aware TS using Orthogonal Learning PSO (MOEATS-OLPSO) and attained 180$ cost, 0.6J for energy consumption, and 37s time. Yet, the model lacked failure-handling mechanisms. Sanaj et al. [12] presented an EcoTaskOpt model that integrated Ant Colony Optimization (ACO) and the PSO algorithm. The model reduced makespan by 18%. Yet, the model's performance is affected by varying task lengths and heterogeneous systems.

B. QoS-Aware Scheduling Techniques

Tabary et al. [13] proposed SMPIA-GA-PSO, integrating GA and PSO with message passing to improve resource utilization, though it requires complex tuning. Long et al. [14] developed IVPTS, combining PSO with fuzzy logic to reduce makespan by 13%, but with higher computational cost. Tamilarasu et al. [15] introduced ICOATS using Coati Optimization to improve makespan and utilization. Sha et al. [16] proposed SPP-TLBO for deadline-aware service placement, improving efficiency by 8–19%. Xu et al. [17] introduced EAT for AIGC task scheduling across edge servers, reducing latency by up to 56%, but with high complexity and overhead.

C. AI Workloads in Cloud Environment

Sharma et al. [18] proposed ICO for AI-driven orchestration, improving scalability but with high implementation complexity. Adeyinka [19] developed LSTM-RL-DT for workload forecasting and dynamic allocation, reducing cost by 25% but struggling with rapid changes. Bauskar [20] introduced an AI-driven multi-cloud database optimization framework, facing consistency challenges across heterogeneous systems. Jorepalli et al. [21] proposed a cloud-native resilient AI architecture for smart education, offering scalability and fault tolerance but with complex deployment.

D. Research Gap Summary

Existing extensions of metaheuristic algorithms often fail to incorporate real-time feedback mechanisms and workload-specific prioritization, limiting their effectiveness in practical deployment scenarios. Table I illustrates the research gap summary.

TABLE I. Research Gap Summary

Method	Models	Strengths	Research Gaps
Classical Scheduling	RR, FCFS	Simple and fast execution	Poor performance in heterogeneous environments
Heuristic-Based	Min-Min, Max-Min	Efficient makespan reduction	Ignores multi-objective, poor adaptability
Swarm Intelligence	PSO	Strong global search capability	Premature convergence, lacks QoS integration
Nature-Inspired Metaheuristics	GWO, WOA	Balanced exploration and exploitation	Static parameters, no dynamic QoS adaptation
Physics-Based Optimization	MSO	Effective exploration–exploitation balance	Single-objective design, lacks Pareto optimization

Existing studies focus on improving individual QoS metrics such as makespan, execution cost, energy efficiency, or load balancing in isolation. Many optimization algorithms employ fixed search strategies and static objective weighting, limiting their adaptability to heterogeneous educational workloads with dynamically changing QoS requirements. Recent AI-based scheduling methods do not incorporate workload-aware task classification with runtime QoS adaptation and Pareto-based multi-objective decision-making. A research gap remains in developing an adaptive scheduling framework capable of simultaneously balancing multiple conflicting QoS objectives while responding to dynamic workload characteristics.

III. System Model and Problem Formulation

The QoS-AMSO framework models task scheduling in AI-driven educational multi-cloud environments as a QoS-constrained, multi-objective optimization problem, and heterogeneous workloads are dynamically mapped to distribute VMs across multiple cloud providers. Each task is represented by a feature vector including computational length, deadline, priority, data size, and execution requirements, and is classified into latency-critical, throughput-intensive, deadline-soft, or batch-deferred categories to enable context-aware scheduling. A centralized scheduler acts as a cloud broker, maintains global knowledge of tasks and resources to determine an optimal task-to-VM mapping that satisfies QoS constraints. The QoS-AMSO algorithm iteratively explores the solution space using adaptive mechanisms such as dynamic step sizing, QoS-driven weight adjustment, and diversity preservation [22]. Solutions are evaluated using a Pareto-based multi-objective framework, and a QoS feedback loop continuously monitors and adjusts performance during execution. The Pareto-optimal scheduling

plan deployed in the multi-cloud environment ensures adaptive and scalable scheduling under dynamic conditions [23].

A. Task and Resource Model

The system model represents the overall structure of the AI-driven educational multi-cloud environments and defines the interaction between tasks and resources during the scheduling process. Let the system consist of a set of tasks. $sT = \{sT_1, sT_2, sT_3, \dots sT_n\}$, a set of VMs distributed across multiple cloud providers $VM = \{vm_1, vm_2, vm_3, \dots, vm_m\}$, a set of cloud providers is given as $CP = \{cp_1, cp_2, \dots cp_1\}$ and each VM belongs to a cloud is defined as $vm_j \in c_k$. The task scheduling (mapping) function assigns each task (sT_i) to a vm_j and it is defined as $\sigma: sT \rightarrow VM$. The goal is to find an optimal mapping (σ) that decreases execution time, cost, SLA violations, and balances system load. Each task (sT_i) is defined as a multi-parameter entity, which includes $sT_i = (tL_i, d_i, p_i, m_i, c_i)$. Here, tL_i defines the task length, d_i defines the deadline, p_i defines the priority level, m_i defines the memory requirement, and c_i defines the task class. In the task classification role, based on parameters, a low deadline indicates latency-critical, a high data size indicates throughput-intensive, and a flexible deadline indicates the batch that influences the initial VM assignment and QoS weight importance. Each VM is defined as $vm_j = (MIPS_j, BW_j, RAM_j, cost_j)$. Here, $MIPS_j$ indicates processing speed, BW_j indicates bandwidth, RAM_j memory capacity and $cost_j$ indicates cost per unit execution time. The system model formulates the process as a mapping between heterogeneous educational workloads and distributed multi-cloud resources [24].

B. QoS Metrics

The proposed framework optimizes the scheduling process using these QoS metrics in a balanced manner. The Makespan represents the total completion time of all scheduled tasks, indicating the overall execution efficiency of the system. It is defined as the maximum finishing time among all VMs $\left(FT_{vm_j}\right)$ and it is defined as:

$$Mks = \max_{vm_j \in VM} \left(FT_{vm_j}\right) \quad (1)$$

Execution cost measures the total monetary expense incurred for executing all tasks across multiple cloud resources, and it is calculated as:

$$cost = \sum_{i=1}^{n} cost_j \times ET\left(t_i, vm_j\right) \quad (2)$$

Here, $cost_j$ indicates the cost per unit time of the VM, and $ET\left(t_i, vm_j\right)$ indicates the execution time of the task on the VM. The SLA violation rate quantifies the degree to which tasks fail to meet their deadlines, reflecting system reliability and service quality, and it is defined as:

$$SLA = \frac{|\{sT_i \mid CT_i > d_i\}|}{n} \quad (3)$$

Here, CT_i indicates the completion time of tasks, and n indicates the total number of tasks. Load imbalance measures the degree of uneven workload distribution across VMs, and it is given as:

$$LI = \frac{\max(Util_j) - \min(Util_j)}{\text{mean}(Util_j)} \quad (4)$$

Here, $Util_j$ indicates the utilization of the VM. Resource utilization reflects the efficiency of resource usage in the multi-cloud environment, and it is defined as:

$$RU = \frac{\sum_{j=1}^{m} used\ MIPS_j}{\sum_{j=1}^{m} Total\ MIPS_j} \quad (5)$$

Here, $used\ MIPS_j$ indicates computational capacity actively utilized, and $total\ MIPS_j$ indicates total available computational capacity.

C. Multi-Objective Optimization Problem

Given the heterogeneous educational workloads and the conflicting QoS requirements, the task scheduling problem is formulated as a multi-objective optimization problem, in which multiple performance metrics are optimized under system constraints [25], [26]. Let σ denote a candidate scheduling solution (mapping of tasks to VMs). In the task assignment constraint, each task is assigned to exactly one VM, and it is given as $\sum_{j=1}^{m} x_{ij} = 1, \forall sT_i \in T$. Here, $x_{ij} = 1$ if task is assigned to VM. The total resource demand must not exceed VM capacity, and it is represented as $\sum_{sT_i \in VM_j} \leq RAM_j$. In the deadline constraint, tasks should ideally complete within their deadline, and it is given as $CT_i < d_i, \forall sT_i \in T$. In the execution feasibility constraint, only feasible mappings are allowed, and it is represented as $ET\left(sT_i, vm_j\right) > 0$, and this ensures valid execution based on VM capability. The multiple objectives are optimized simultaneously, and this solution is defined in terms of Pareto optimality. A solution(σ_1) is said to dominate another solution(σ_2) and it is given as $\forall k: f_k(\sigma_1) \leq f_k(\sigma_2)$ and $\exists k: f_k(\sigma_1) < f_k(\sigma_2)$. The set of all non-dominated solutions is called the Pareto optimal set, and it is given as $p^* = \{\sigma \mid \nexists \sigma'\ such\ that\ \sigma' \prec \sigma\}$. From the Pareto set, a final solution is selected using a decision-making method, and it is given as $\sigma^* = \arg\min_{\sigma \in p^*} F_w(\sigma)$. This process ensures balanced QoS performance and context-aware scheduling.

IV. Proposed QoS-AMSO Framework

The proposed QoS-AMSO framework is a layered, end-to-end scheduling architecture that transforms AI-driven educational workloads into optimized task–resource mappings in a multi-cloud environment. It consists of five layers: workload ingestion and pre-processing, QoS-aware task classification, multi-cloud resource abstraction, QoS-AMSO optimization engine, and decision execution and monitoring. The workload ingestion layer collects heterogeneous tasks and extracts attributes such as computational size, deadline sensitivity, data volume, and priority, and converts them into a structured format. The QoS-aware classification layer categorizes tasks into latency-critical, throughput-intensive, deadline-soft, and batch-deferred classes that derive indicators. The resource abstraction layer models distributed cloud infrastructure by representing heterogeneous VMs with varying processing capabilities, costs, and dynamic states such as load and availability [27]. The QoS-AMSO optimization engine performs scheduling by initializing solutions using class-aware heuristics and iteratively refines through adaptive step sizing,

QoS-driven weight adjustment, and Lévy flight-based exploration. The execution and monitoring layer selects the optimal solution from the Pareto set, assigns tasks to VMs across cloud providers, and continuously monitors execution performance.

A. Overview of Mirage Search Optimization

Mirage Search Optimization (MSO) is a physics-inspired metaheuristic algorithm based on the optical phenomenon of mirage formation. The MSO mimics two phenomena, including superior mirage (global exploration) and inferior mirage (local exploitation). Each candidate solution is treated as an observation point (individual) in a multi-dimensional search space. Each solution is randomly initialized within bounds, and it is formulated as

$$x_u = lb_u + r.(ub_u - lb_u) \quad (6)$$

Here, ub_u, lb_u indicates lower and upper bounds, and $r \in [0,1]$ indicates a random number. This process ensures a diverse initial population. Superior mirage allows observing distant objects, modeled as a long-range search capability. The position update (x_{uv}^{t+1}) is formulated as:

$$x_{uv}^{t+1} = x_{uv}^{t} + \Delta x_{lower} \quad (7)$$

The inferior mirage phenomenon occurs due to refraction of light, and it is modelled to perform localized search around promising solutions, focusing on refining (fine-tuning) solutions near the global best. The distance and direction are modeled when the current solution is not the best, and it is formulated as

$$h_{uv} = \left|gbest_j - x_{uv}\right|.rand, D = \frac{(gbest_j - x_{uv}).rand}{h_{uv}} \quad (8)$$

The inferior mirage uses optical refraction principles to control movement direction. The angle (γ) gamma, angle (φ) and refraction angle ω (omega), and this process is formulated as

$$\gamma = \frac{\pi(t_{max} - t)}{2t_{max}}.rand \quad (9)$$

$$\emptyset = \left(\arctan\left(\frac{1}{2\tan\gamma}\right) - \arctan\left(\frac{\sin\gamma\cos\gamma}{1+\sin^2\gamma}\right)\right).rand + \arctan\left(\frac{\sin\gamma\cos\gamma}{1+\sin^2\gamma}\right) \quad (10)$$

$$\omega = \arcsin\left(\frac{n_2}{n_1}\sin(\emptyset + \gamma)\right) \quad (11)$$

The displacement calculation uses all geometric relations; the movement step is derived as:

$$\Delta x_{upper} = \frac{h}{\tan\gamma}\left(\frac{h}{\sin\gamma} - \frac{h\sin\emptyset}{\cos(\emptyset+\gamma)}\right).\frac{\cos\omega}{\cos(\omega-\gamma)} \quad (12)$$

This equation combines distance (h), angle $(\gamma, \varphi, \omega)$ and refraction behaviour. The final update rule is formulated as:

$$x_{uv}^{t+1} = x_{uv}^{t} + D.\Delta x_{upper} \quad (13)$$

The MSO employs a physics-driven search paradigm that combines superior mirage-based long-range exploration with inferior mirage-based fine-grained exploitation, allowing the algorithm to dynamically adapt its search behavior and improve solution quality.

B. Limitations of Standard MSO

MSO demonstrates effective exploration and exploitation through its reflection and refraction mechanisms; it exhibits several limitations when applied to QoS-driven task scheduling in multi-cloud environments. MSO is designed as a single-objective optimization algorithm, which restricts its ability to simultaneously handle multiple conflicting QoS metrics such as makespan, cost, SLA violation, and load balancing. The algorithm lacks QoS awareness, as it does not incorporate task-specific requirements such as deadlines, priorities, or workload characteristics, leading to suboptimal scheduling decisions in heterogeneous environments. MSO does not include a structured mechanism for maintaining multiple optimal solutions, limiting its capability to achieve Pareto-optimal trade-offs. To address these limitations, there is a need for an enhanced optimization framework that integrates multi-objective optimization, QoS awareness, adaptive parameter control, and task-specific intelligence.

C. Proposed QoS-AMSO Algorithm

1) Adaptive step sizing: In the standard MSO, the movement of individuals during both superior and inferior mirages is governed by fixed geometric step formulations. These formulations lack dynamic control over step magnitude. To overcome this limitation, the QoS-AMSO introduces an adaptive step sizing mechanism, which dynamically adjusts the movement step of individuals across iterations. Educational cloud workloads exhibit varying search characteristics during different optimization stages. Large search steps are beneficial during the early iterations to explore diverse scheduling solutions, and smaller steps are required in later iterations to refine promising task-to-resource mappings. A fixed movement strategy in the standard MSO cannot effectively accommodate this transition, leading to slow convergence and premature stagnation. The adaptive step size is introduced as a scaling factor $\alpha(t)$ and it is defined as:

$$\alpha(t) = \alpha_{max} - \left(\frac{t}{t_{max}}\right)(\alpha_{max} - \alpha_{min}) \quad (14)$$

Here, t indicates the current iteration, t_{max} indicates the maximum iterations, α_{max} indicates the initial large step size, and α_{min} indicates a small step size. The adaptive step size process is embedded into the MSO movement equations. The original MSO position update from Eq. (11) is modified with an adaptive step and formulated as:

$$x_{uv}^{t+1} = x_{uv}^{t} + \alpha(t) + \Delta x_{lower}, x_{uv}^{t+1}, \; = x_{uv}^{t} + \alpha(t).D.\Delta x_{upper} \quad (15)$$

This process scales step size dynamically with large jumps and controlled movement. The adaptive step sizing controls search intensity dynamically, improves convergence speed, enhances solution stability, and supports multi-objective balancing.

2) QoS feedback mechanism: The QoS feedback mechanism introduces a feedback-driven weighting mechanism that monitors QoS metrics, adjusts their importance dynamically, and guides the search toward better

solutions. The weighted fitness function is formulated as $F(\sigma) = \sum_{k=1}^{4} w_k(t).f_k(\sigma)$. Here, $w_k(t) = 1$ indicates the dynamic weight of k^{th} QoS metric. Static objective weights cannot effectively respond to runtime variations, causing optimization to focus on outdated scheduling priorities. Therefore, a QoS feedback mechanism is introduced in which weights are updated based on QoS deviation. For SLA violations, the SLA-specific feedback and it is formulated as:

$$w_{SLA}(t+1) = w_{SLA}(t) + \eta.SLA(t) \quad (16)$$

Here, η indicates the learning rate. This algorithm focuses on prioritizing deadlines. During fitness evaluation, this process can be integrated as:

$$F(\sigma, t) = \sum w_k(t)\,.\,f_k(\sigma) \quad (17)$$

The QoS feedback mechanism enhances the MSO framework by dynamically adjusting the importance of multiple QoS objectives based on real-time system performance.

3) Lévy flight strategy: Metaheuristic optimization algorithms suffer from premature convergence when candidate solutions become concentrated around local optima, with numerous feasible task-resource mappings. The proposed QoS-AMSO incorporates a Lévy Flight Strategy, which introduces random long-distance perturbations into the search process. The step size of the Lévy flight is defined as $L\sim levy\ (\beta)$, here $\beta \in (1,3)$ indicates the Lévy distribution parameter. Here, λ indicates the scaling factor, and $gbest$ indicates the global best solution. The Lévy flight is integrated with superior mirage and inferior mirage, and it is formulated as:

$$x^{t+1} = x^t + \alpha(t).\Delta x_{lower} + \lambda.L \quad (18)$$

$$x^{t+1} = x^t + \alpha(t).D.\Delta x_{lower} + \lambda.L.(x^t - gbest) \quad (19)$$

The QoS-AMSO Optimization Engine, Lévy flight enhances search diversity, prevents premature convergence, improves global optimality, and supports multi-objective exploration.

4) Pareto archive mechanism: Cloud task scheduling simultaneously optimizes multiple conflicting QoS objectives, including makespan, execution cost, SLA satisfaction, and load balancing. The proposed QoS-AMSO introduces a Pareto Archive Mechanism, which maintains a set of non-dominated solutions throughout the optimization process. A solution $(\sigma_a) \leq f_k(\sigma_b)$ and $\exists k: f_k(\sigma_a) < f_k(\sigma_b)$. At iteration $t = 0$, the archive is initialized as $\mathcal{A}(0) = \{\sigma_i\}$ is non-dominated in the population. The archive is updated as $\mathcal{A}' = \mathcal{A}(t) \cup \{\sigma_1^{t+1}, \sigma_2^{t+1}, \dots, \sigma_n^{t+1}\}$. The non-dominated sorting and it is formulated as $\mathcal{A}(t+1) = \{\sigma \in \mathcal{A}' | \nexists \sigma' \in \mathcal{A}' : \sigma' \prec \sigma\}$. To avoid excessive archive growth, a maximum size $(\mathcal{A}_{max})$ is defined as $|\mathcal{A}(t+1)| > \mathcal{A}_{max}$. The MSO update equations for superior and inferior mirage are defined as:

$$x^{t+1} = x^t + \alpha(t).D.\Delta x(\sigma_{archive}) \quad (20)$$

Here, $\sigma_{archive}$ indicates solutions selected from the Pareto set. The Pareto Archive Mechanism enhances the QoS-AMSO framework by maintaining a set of non-dominated solutions that capture optimal trade-offs among multiple QoS objectives.

D. Task Classification Model

1) Workload categories for AI-educational systems: The proposed QoS-AMSO framework classifies tasks into four distinct workload categories: latency-critical (LC), throughput-Intensive (TI), deadline-Soft (DS), and batch-deferred (BD). Each task sT_i is represented as a multi-dimensional feature vector that includes task length (tL_i), deadline (d_i), priority level (pL_i), memory requirement (m_i) and data size (ds_i). For latency-critical workloads, tasks require immediate response and strict deadline adherence. For example, LLM-based real-time Q&A systems live virtual laboratories and interactive tutoring sessions. The mathematical condition is formulated as $C_i = LC$, if $DTR_i < \theta_1$ and $p_i = high$. For throughput-intensive tasks, the mathematical condition is formulated as $C_i = TI$, if $ds_i < \theta_2$ and $tL_i =\gg 0$. In deadline-soft workloads, tasks with moderate deadlines have slight delays; the mathematical condition is formulated as $C_i = DS$, if $\theta_1 < DTR_i \leq \theta_3$. In batch-deferred workloads, the mathematical condition is formulated as $C_i = BD$, if $DTR_i > \theta_3$. This classification enables task-aware scheduling decisions, priority-based VM allocation, and QoS-driven optimization guidance. By guiding population initialization and dynamic QoS weight adaptation, task classification improves the quality of initial solutions and enhances the final scheduling performance.

2) Feature vector and classification: The proposed QoS-AMSO framework introduces a feature vector-based representation, followed by a classification mechanism that assigns each task to a QoS-aware category. This transformation enables the scheduler to incorporate task-level intelligence into the optimization process $F_i = [tL_i, d_i, pL_i, m_i, ds_i, DTR_i]$. In the task classification model using the feature vector, each task is assigned to a class, and it is formulated as $C_i = f(F_i)$. The classification output (C_i) is integrated into the optimization process at multiple levels. The fitness function enhancement is formulated as:

$$F(\sigma) = \sum_{i=1}^{n} W_{class}(C_i).f_i(\sigma) \quad (21)$$

The feature vector and classification enable context-aware decision-making, priority-based scheduling, efficient resource mapping, and improved SLA satisfaction.

3) Impact on AMSO initialization and weight setting: The proposed framework leverages task class information to guide the initial population generation and dynamically adjust QoS priorities during optimization. In the proposed class-Guided initialization, each task sT_i is assigned to a VM based on its class C_i and it is formulated as $\sigma_i^{(0)} = \arg\max_{vm_j \in VM} S(C_i, vm_j)$. Here, $\sigma_i^{(0)}$ indicates the initial mapping of tasks, $S(C_i, vm_j)$ indicates the suitability score. In the dynamic QoS weight

adjustment, the weights evolve using optimization, and it is given as:

$$w_k(t+1) = w_k(t) + \eta.\Delta_k(t) \quad (22)$$

$$\Delta_k(t) = \frac{f_k(t) - f_k^{target}}{f_k^{target}} \quad (23)$$

The task classification into AMSO initialization and weight setting enhances the optimization process by guiding initial solution generation and dynamically adapting QoS priorities based on workload characteristics. Algorithm 1 presents the complete algorithmic workflow of the proposed QoS-AMSO framework, which integrates task classification, adaptive optimization, and multi-objective decision-making into a unified scheduling process.

Algorithm 1: QoS-AMSO for Multi-Cloud Scheduling

Step 1: Workload Pre-processing and Feature Extraction
- Collect incoming tasks from AI-based educational applications.
- Extract task parameters and normalize features to ensure balanced representation.

Step 2: Task Classification
- Compute derived indicators and classify each task.

Step 3: Initialize Population (Class-Aware Initialization)
- Generate initial candidate scheduling solutions.
- Assign tasks to VMs based on task class.
  - High-performance resources for latency-critical tasks
  - Parallel resources for throughput-intensive tasks
  - Balanced resources for deadline-soft tasks
  - Low-cost resources for batch tasks

Step 4: Update Candidate Solutions (MSO Search)
- For each solution in the population:
  - Apply superior and inferior mirage strategies for exploration and exploitation.
  - Adjust movement using adaptive step sizing.
  - Modify search behavior based on iteration progress.
  - Introduce Lévy Flight Perturbation to selected solutions.
  - Combine current solutions with archive and perform non-dominated sorting.
  - Select final scheduling solution based on QoS preference.

Step 5: Task Execution and Monitoring
- Deploy selected scheduling plan.
- Assign tasks to respective VMs and monitor execution performance.

End

4) Computational complexity: The computational complexity of the proposed QoS-AMSO framework is analyzed to evaluate its scalability and efficiency in handling large-scale, heterogeneous multi-cloud scheduling problems. Let n indicate the number of tasks, M indicate the number, vm indicate the number of VMs, P indicate the population size, t indicate the number of iterations, and K indicate the number of QoS objectives. In task classification complexity, the complexity is defined as $O(n)$. For each iteration, all candidate solutions are updated, and it is formulated as $O(t.P.n)$. Each solution is evaluated across all QoS objectives, and it is given as $\left(O(t.P.n.K)\right)$. In the Pareto archive update, Non-dominated sorting is performed over the population and archive, and it is defined $\left(O(T.P^2)\right)$. The overall complexity is defined as:

$$TC = O(n) + O(P.n) + O(T.P.n) + O(T.P^2) \quad (24)$$

The proposed QoS-AMSO framework is computationally efficient due to linear scalability with task size, controlled population size limiting quadratic overhead.

V. Experimental Results and Discussion

A. Simulation Environment

To evaluate the effectiveness of the proposed QoS-AMSO framework, a simulation environment is replicated to replicate realistic multi-cloud conditions for educational workloads. Synthetic MOOC-based workload patterns are artificially generated workloads that mimic real MOOC platforms to simulate diverse educational task behaviors [25]. The synthetic workload is generated by modelling common educational activities observed in AI-enabled learning platforms, including video streaming sessions, online quizzes, assignment submissions, real-time tutoring requests, and learning analytics jobs. Task arrival times are generated using variable arrival rates to simulate normal and peak learning periods; computational demand, data size, and deadline constraints are assigned according to the characteristics of each educational activity.

A real-world workload dataset derived from Google cluster logs, modified to reflect educational task characteristics, was filtered based on computational intensity and execution duration. The original trace attributes, including CPU utilization, memory usage, task execution time, and arrival patterns, were retained, and application labels were mapped to representative educational workload categories. Interactive services are modelled as latency-critical tasks, large-scale assessment and analytics jobs as throughput-intensive tasks, routine learning activities as deadline-soft tasks, and background reporting processes as batch-deferred tasks. Table II illustrates the environmental settings.

TABLE II. Environmental Settings

Category	Parameter	Value
Simulation Tool	Platform	CloudSim 3.0
Cloud Infrastructure	Number of Cloud Providers	3
	Total Virtual Machines	60
	VM Distribution	20 per provider
VM Configuration	MIPS Levels	1000, 2000, 3000
	RAM	8 GB
	Bandwidth	100–1000 Mbps
	Storage	10–100 GB
Network Settings	Latency Range	5 ms – 80 ms
Workload	Task Count	100 – 1000 tasks
	Data Size	10 MB – 5 GB
	Priority Levels	Low, Medium, High
Optimization Parameters	Population Size (P)	50
	Max Iterations (T)	100
	Lévy Flight Parameter	1.5
	Archive Size	50 – 100 solutions

B. Workload Description

The workloads are structured to reflect realistic operational settings, ranging from small-scale deployments to large-scale distributed systems with dynamic load variations. The three workload scenarios are defined as small-scale (S1), medium-scale (S2), and large-scale dynamic environment (S3). Table III illustrates the workload scenario configuration.

TABLE III. WORKLOAD SCENARIOS CONFIGURATION

Scenario	Number of Tasks	Number of VMs	Scale	Workload Behavior	Special Condition
S1	100	20	Small-scale	Stable	No burst
S2	500	40	Medium-scale	Moderately dynamic	No burst
S3	1000	60	Large-scale	Highly dynamic	Burst injection at 50% time

C. Baseline Methods

To validate the effectiveness of the proposed QoS-AMSO framework, its performance is compared against a diverse set of baseline algorithms representing classical scheduling strategies and state-of-the-art metaheuristic optimization methods. The methods include Round Robin (RR), Min-Min, PSO, Grey Wolf Optimizer (GWO), Whale Optimization Algorithm (WOA), and Vanilla MSO. The RR algorithm is selected as a classical baseline due to its simple cyclic task assignment, highlighting the limitations of non-QoS-aware scheduling. The Min-Min algorithm is included as a heuristic that prioritizes tasks with minimum completion time. The PSO is chosen for its strong global search capability and popularity in scheduling problems. The MSO is used to evaluate the improvements achieved through adaptive step sizing, QoS feedback mechanisms, Lévy flight integration, and Pareto-based optimization.

D. Performance Metrics

To comprehensively evaluate the effectiveness of the QoS-AMSO framework, a set of performance metrics is selected. The metrics include Makespan (s), Total Cost (USD), SLA Violation Rate (%), Load Imbalance Index (LII), Resource Utilization (RU) (%), and Convergence speed.

E. Parameter Settings

The performance of the QoS-AMSO algorithm has hyperparameters that control the search behavior and convergence speed. Table IV illustrates the parameter settings.

TABLE IV. PARAMETER SETTINGS

Parameter	Standard value	Optimized value
Population Size (P)	20–30	50
Max Iterations (T)	100–150	100
Lévy Exponent (λ)	1.0–1.2	1.5
Archive Size	10–20	30
Initial Step Size (α_0)	0.5–0.7	0.9
Feedback Gain (γ)	0.01–0.03	0.05
Perturbation Interval	5–8	10

The optimization parameters are selected through preliminary sensitivity analysis by evaluating different parameter combinations under the medium-scale workload scenario (S2). Population size and maximum iteration count are chosen to provide a balance between convergence quality and computational overhead, and the Lévy exponent was selected to maintain sufficient search diversity without introducing excessive random exploration. The archive size and feedback gain are determined to ensure stable Pareto solution maintenance and responsive QoS adaptation throughout the optimization process.

F. Performance Comparison

This section presents the comparative performance analysis of the proposed QoS-AMSO framework against baseline algorithms is illustrated in Table V.

TABLE V. PERFORMANCE COMPARISON WITH BASELINE MODELS

Models	Makespan (s)	Total cost (USD)	SLA (%)	LII	RU (%)
RR	1285	512.4	12.8	0.42	68.3
Min-Min	1042	468.7	9.6	0.35	74.1
PSO	921	552.3	7.8	0.29	79.6
GWO	887	331.5	6.9	0.26	82.4
WOA	865	425.2	6.3	0.24	91.1
MSO	842	418.6	5.7	0.22	86.5
QoS-AMSO	798	292.8	3.9	0.18	97.2

The performance comparison results demonstrate that the proposed QoS-AMSO framework outperforms classical, heuristic, and metaheuristic baseline algorithms. The model attained a 798s makespan and 97.2% RU with reduced cost. The consistent performance improvement of QoS-AMSO over the baseline methods is attributed to the integration of its components. Task classification enables workload-aware resource allocation, adaptive step sizing balances global exploration and local exploitation for faster convergence, QoS feedback dynamically adjusts scheduling decisions according to runtime conditions, Lévy flight improves global search capability by avoiding local optima, and the Pareto archive preserves high-quality multi-objective solutions. The conventional scheduling algorithms optimize limited objectives; the coordinated operation of these mechanisms enables QoS-AMSO to achieve lower makespan and execution cost, fewer SLA violations, improved load balancing, and higher resource utilization across diverse educational workload scenarios.

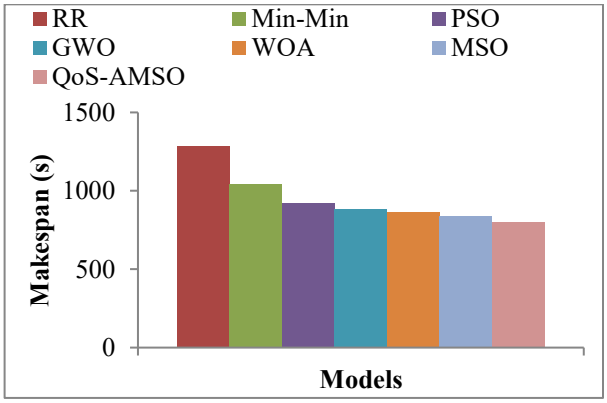


Fig. 1. Makespan graph comparison.

Fig. 1 illustrates a makespan graph comparison. The model attained a makespan of 798s, which is decreased by 5.22% to 37.89% compared to the RR, Min-Min, PSO, GWO, WOA, and MSO models. The QoS-AMSO dynamically adjusts its search process through adaptive step sizing, which enables control over exploration and exploitation during stages of optimization.

Fig. 2 illustrates the total cost graph comparison. The model attained a total cost of 292.8USD that is decreased by 10.04% to 41.08% compared to the RR, Min-Min, PSO, GWO, WOA, and MSO models. The QoS feedback mechanism allows the algorithm to adapt to runtime performance deviations.

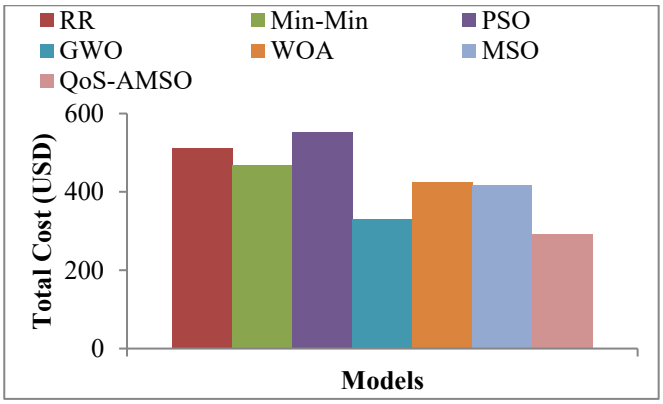


Fig. 2. Total cost graph comparison.

Fig. 3 illustrates the SLA violation graph. The model attained an SLA violation of 3.9%, which is decreased by 1.8% and 8.9% compared to the RR, Min-Min, PSO, GWO, WOA, and MSO models. The Lévy flight strategy introduces controlled randomness, enabling the algorithm to escape local optima and explore a solution space.

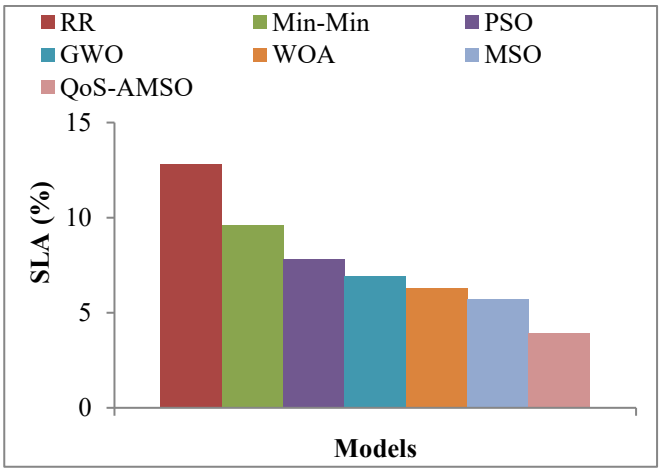


Fig. 3. SLA violation graph comparison.

Fig. 4 illustrates the load imbalance index graph. The model attained an LII of 0.18, which is decreased by 18.18% to 57.14% compared to the RR, Min-Min, PSO, GWO, WOA, and MSO models. The QoS-AMSO converges faster to high-quality solutions while maintaining consistency across different workload scenarios.

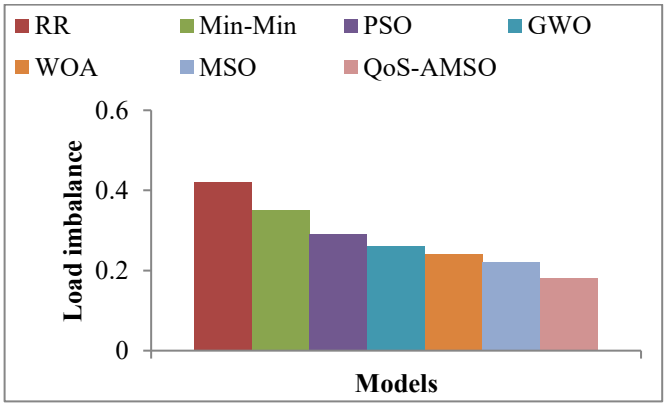


Fig. 4. Load imbalance index graph comparison.

Fig. 5 illustrates the resource utilization graph. The model attained resource utilization of 97.2%, which is increased by 6.1% to 28.9% than the RR, Min-Min, PSO, GWO, WOA, and MSO models. The QoS-AMSO incorporates task classification, enabling the scheduler to differentiate between latency-critical, throughput-intensive, and flexible workloads.

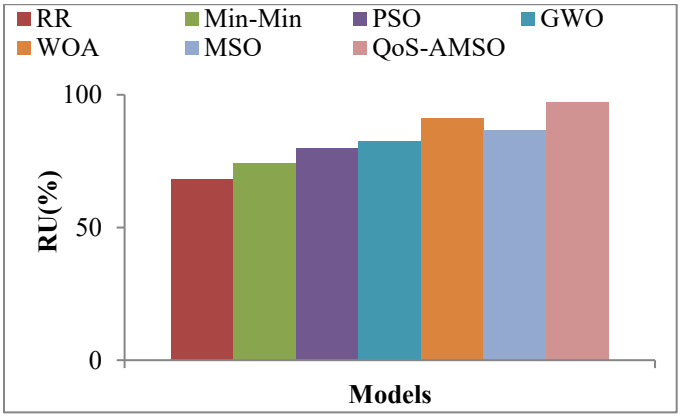
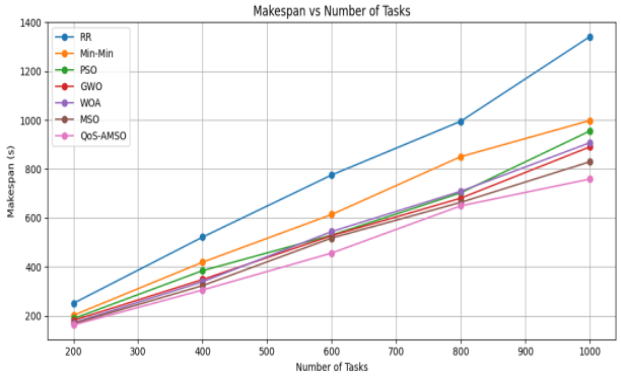


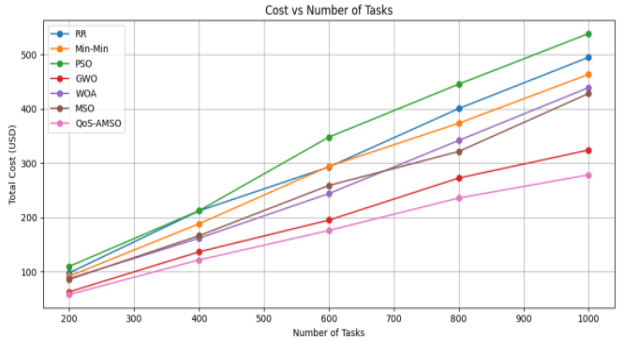
Fig. 5. Resource utilization graph comparison.

G. Scalability Analysis

The graphs in Fig. 6(a) – 6(b) show that QoS-AMSO maintains superior performance with increasing workload size, demonstrating better scalability compared to baseline methods.



(a)



(b)

Fig. 6. (a) Scalability analysis (makespan), (b) Scalability analysis (cost).

H. Convergence Analysis

Fig. 7 shows the plots that illustrate that QoS-AMSO converges faster and more smoothly to optimal solutions than other algorithms across iterations.

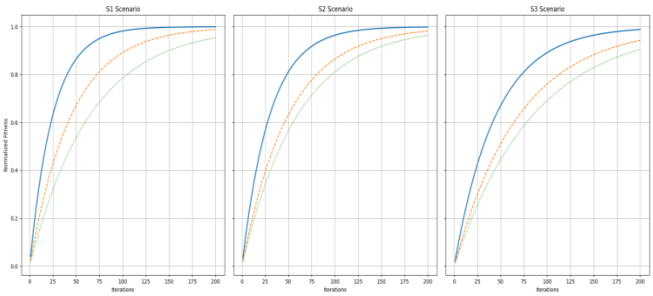


Fig. 7. Convergence analysis.

I. QoS Feedback Mechanism Effectiveness Analysis

This section evaluates how effectively the proposed QoS-AMSO framework adapts to dynamic workload changes using its QoS feedback mechanism under the large-scale scenario (S3) with burst injection. During the experiment, the evolution of the QoS weight vector is tracked across iterations. When a sudden workload spike occurs at 50% of the simulation time, an increase in SLA violations is observed due to resource contention. In response, the feedback mechanism dynamically adjusts the weights by increasing the importance of the SLA component. This reactive adjustment shifts the optimization focus toward meeting deadline constraints, resulting in improved task prioritization for latency-sensitive workloads, a reduction in SLA violations over subsequent iterations, and stabilization of system performance after the burst.

J. Ablation Study

An ablation study is conducted by systematically removing one mechanism, including without Lévy perturbation (LP), adaptive step sizing (AS), QoS feedback weights (QF), and class-aware initialization (CI) modules. Table VI illustrates the ablation study.

When Lévy perturbation is excluded, the makespan increases to 846s, and SLA violations increase to 5.2%, due to the loss of its ability to perform long-distance exploration, leading to localized search behavior. Removing adaptive step sizing degrades performance, with makespan increasing to 872s and cost to 327.9 USD, and fails to balance exploration and exploitation. When the QoS feedback mechanism is removed, makespan increases to 905s, with reduced utilization of 87.2% as the algorithm becomes static and unable to respond to runtime variations. Without class-aware initialization, makespan reaches 931s, and resource utilization drops to 85.6%, due to inefficient initial task-to-resource mappings, leading to poor overall scheduling performance. The complete QoS-AMSO model achieves the best results with a makespan of 798s and resource utilization of 97.2%. The Lévy perturbation, adaptive step sizing, QoS feedback, and class-aware initialization enable balanced exploration and exploitation, and dynamic QoS prioritization.

TABLE VI. ABLATION STUDY

Models	Makespan (s)	Total cost (USD)	SLA (%)	LII	RU (%)
w/o LP	846	318.5	5.2	0.23	91.4
w/o AS	872	327.9	5.8	0.25	89.7
w/o QF	905	341.6	7.1	0.28	87.2
w/o CI	931	356.4	7.9	0.31	85.6
QoS-AMSO	798	292.8	3.9	0.18	97.2

K. Statistical Validation

To ensure that the observed performance improvements of the proposed QoS-AMSO framework are not due to random variations, statistical validation is conducted using the Wilcoxon signed-rank test. Table VII illustrates the Statistical Significance results.

The statistical validation using the Wilcoxon signed-rank test confirms that the performance improvements of the QoS-AMSO framework are statistically significant across all baseline comparisons. The consistently low p-values and large effect sizes demonstrate that the proposed model achieves reliable enhancements in scheduling performance, establishing its robustness and effectiveness for multi-cloud task scheduling.

TABLE VII. STATISTICAL SIGNIFICANCE RESULTS

Comparison	p-value	Effect Size (r)
QoS-AMSO vs RR	0.00012	0.89
QoS-AMSO vs Min-Min	0.00031	0.82
QoS-AMSO vs PSO	0.00105	0.76
QoS-AMSO vs GWO	0.00218	0.71
QoS-AMSO vs WOA	0.00342	0.68
QoS-AMSO vs MSO	0.00487	0.64

L. Discussion

The proposed QoS-AMSO outperforms standard MSO by integrating decision intelligence into the optimization process. While MSO relies on physics-inspired search dynamics, it lacks awareness of workload characteristics and QoS requirements. The QoS-AMSO embeds scheduling context into the optimization loop, enabling the algorithm to focus on meaningful regions of the search space, resulting in stable

convergence and improved solution quality in heterogeneous environments. The Lévy flight mechanism helps maintain population diversity by introducing occasional long jumps, allowing exploration of distant regions and ensuring broader search coverage in complex multi-cloud scenarios. Adaptive step sizing enhances performance by dynamically reducing step size over iterations, enabling fine-grained adjustments near optimal solutions and improving convergence accuracy. In small-scale workloads, the Min-Min algorithm achieves slightly lower cost due to its focus on execution time minimization, but this leads to higher SLA violations and poor load distribution. The QoS-AMSO maintains a balanced multi-objective approach, ensuring system reliability, fairness, and QoS compliance, even if it incurs marginally higher cost in specific cases.

VI. Conclusion and Future Work

The QoS-AMSO framework is developed as a QoS-aware scheduling model for AI-driven educational workloads in multi-cloud environments. The contributions include the integration of task classification, adaptive optimization strategies, QoS feedback mechanisms, Lévy flight-based exploration, and Pareto-based decision-making within a unified architecture. Experimental results demonstrate that the proposed model achieves improved makespan, reduced cost, lower SLA violations, balanced load distribution, and higher resource utilization. Despite its performance, the current framework has certain limitations. The evaluation is conducted in a CloudSim-based simulation environment, which may not fully capture the dynamic behavior, network variability, and resource heterogeneity encountered in real multi-cloud deployments. The framework assumes accurate workload profiling and does not explicitly consider energy efficiency and security policies. Future work will focus on validating the framework in real multi-cloud testbeds, incorporating energy-aware and security-aware scheduling strategies, and integrating edge-cloud environments to improve adaptability for large-scale educational applications.

Funding Statement

This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Funding No.KFU263786].

Acknowledgment

The authors are thankful to the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, and Al-Fayha Private College, Jubail, Saudi Arabia, for providing financial assistance.

References

[1] N. H. Anh, N. H. “Hybrid cloud migration strategies: balancing flexibility, security, and cost in a multi-cloud environment”, *Transactions on Machine Learning, Artificial Intelligence, and Advanced Intelligent Systems*, Vol. 14, No. 10, pp. 14-26, 2024.

[2] W. K. Awad, K. A. Z. Ariffin, M. Z. A. Nazri, and E. T. Yassen, “Resource allocation strategies and task scheduling algorithms for cloud computing”. *A systematic literature review. Journal of Intelligent Systems*, Vol. 34, No. 1, 2025.

[3] O. L. Abraham, M. A. B. Ngadi, J. B. M. Sharif, and M. K. M. Sidik, M. K. M. “Multi-objective optimization techniques in cloud task scheduling”, *A systematic literature review. IEEE Access,* Vol. 13, pp. 12255-12291, 2025.

[4] K. Y. Rajput, X. Li, J. Zhang, and A. Lakhan, “A novel scheduling approach for Spark workflow tasks with deadline and uncertain performance in multi-cloud networks”, *IEEE Transactions on Cloud Computing,* Vol. 12, No. 4, pp. 1145-1157, 2024.

[5] S. Mangalampalli, G. R. Karri, M. V. Ratnamani, S. N. Mohanty, B. A. Jabr, Y. A. Ali, and B. S. Abdullaeva, “Efficient deep reinforcement learning based task scheduler in multi-cloud environment*”, Scientific Reports,* Vol. 14, No. 1, pp. 21850, 2024.

[6] S. S. Mangalampalli, G. R. Karri, S. N. Mohanty, S. Ali, M. I. Khan, S. Abdullaev, and S. A. AlQahtani, “Multi-objective Prioritized Task Scheduler using improved Asynchronous Advantage Actor-Critic (a3c) algorithm in multi-cloud environment”, *IEEE Access,* vol. 12, pp. 11354-11377, 2024.

[7] M. S. Kumar, and G. R. Kumar, G. R. “EAEFA: an efficient energy-aware task scheduling in cloud environment”, *EAI Endorsed Transactions on Scalable Information Systems,* Vol. 11, No. 3, 2024.

[8] R. Sandhu, M. Faiz, H. Kaur, A. Srivastava, and V. Narayan, V. “Enhancement in performance of cloud computing task scheduling using optimization strategies”, *Cluster Computing,* Vol. 27, No. 5, pp. 6265-6288, 2024.

[9] B. Acharya, S. Panda, S. Das, S. K. Majhi, V. C. Gerogiannis, and A. Kanavos, “Optimizing task scheduling in cloud environments: a hybrid golden search whale optimization algorithm approach”, *Neural Computing and Applications*, pp. 1-23, 2025.

[10] S. Bansal, B. S. Singla, and H. Aggarwal, “Workflow task scheduling in a cloud-fog environment: a hybrid PSO-GOA approach”, *International Journal of System Assurance Engineering and Management,* pp. 1-23, 2025.

[11] B. Nagalakshmi, S. Subramanian, “Multi-objective energy aware task scheduling using orthogonal learning particle swarm optimization on cloud environment”, *International Journal of Information Technology*, Vol. 17, No. 1, pp. 447-454, 2025.

[12] M. S. Sanaj, M. F. Horng, S. S. Shankar, C. C. Lo, R. Sunder, U. K. Lilhore, and E. S. Ghith, “Energy-Efficient Task Scheduling in Cloud Computing with EcoTaskOpt: A Hybrid Ant Colony and Particle Swarm Optimization Approach”, *International Journal of Computational Intelligence Systems*, 2026.

[13] K. A. Tabary, H. Motameni, B. Barzegar, E. Akbari, H. Shirgahi, and M. Mokhtari, “QoS-aware task scheduling using hybrid genetic algorithm in cloud computing”, *IEEE Access*, Vol. 13, pp. 51603-51616, 2024.

[14] G. Long, S. Wang, and C. Lv, “QoS-aware resource management in cloud computing based on fuzzy meta-heuristic method”, *Cluster Computing*, Vol. 28, No. 4, pp. 276, 2025.

[15] P. Tamilarasu, and G. Singaravel, “Quality of service aware improved coati optimization algorithm for efficient task scheduling in cloud computing environment”, *Journal of Engineering Research*, Vol. 12, No. 4, pp. 768-780, 2024.

[16] Y. Sha, H. Wang, D. Wang, and M. Ghobaei-Arani, “A multi-objective QoS-aware IoT service placement mechanism using Teaching Learning-Based Optimization in the fog computing environment”, *Neural Computing and Applications*, Vol. 36, No. 7, pp. 3415-3432, 2024.

[17] Z. Xu, Z. Tang, J. Lou, Z. Yao, X. Xie, T. Wang, and W. Jia, “EAT: QoS-Aware Edge-Collaborative AIGC Task Scheduling via Attention-Guided Diffusion Reinforcement Learning”, *IEEE Transactions on Mobile Computing*, 2026.

[18] S. Sharma, N. Kumar, Y. Dash, A. Dubey, and K. Devi, “Intelligent multi-cloud orchestration for AI workloads: enhancing performance and reliability”, *In 2024 7th International Conference on Contemporary Computing and Informatics (IC3I)*, Vol. 7, pp. 1421-1426. IEEE, 2024.

[19] A. Adeyinka, “Dynamic resource allocation using AI-driven workload forecasting in multi-cloud environments”, *World Journal of Advanced Research and Reviews,* Vol. 23, pp. 3188-3198, 2024.

[20] S. R. Bauskar, "Optimizing Multi-Cloud Environments Advanced Database Technologies for Scalable and Resilient Education and Training Systems", *In Integrating AI and Sustainability in Technical and Vocational Education and Training (TVET)*, pp. 189-206. IGI Global Scientific Publishing, 2025.

[21] S. K. R. Jorepalli, "Cloud-native AI applications designing resilient network architectures for scalable AI workloads in smart education", *In Smart Education and Sustainable Learning Environments in Smart Cities,* pp. 155-172. IGI Global Scientific Publishing, 2025.

[22] A. N. Malti, M. Hakem, and B. Benmammar, "A new hybrid multi-objective optimization algorithm for task scheduling in cloud systems", *Cluster Computing,* Vol. 27, No. 3, pp. 2525-2548, 2024.

[23] X. Tan, and X. Zhang, "Data-driven evolutionary algorithm with Pareto set topology learning for multimodal multi-objective optimization", *Knowledge-Based Systems*, pp. 115582, 2026.

[24] H. Abdullah, A. S. Kumar, A. Qasem Ahmed, and M. Saeed Mosleh, "Hybrid Optimization Based on Spectrum Aware Opportunistic Routing for Cognitive Radio Ad Hoc Networks" *Informatics and Automation,* Vol. 22, No. 4, pp. 880-905, 2023.

[25] A. Machado, K. Tenório, M. M. Santos, A, P. Barros, L. Rodrigues, R. F. Mello, and D. Dermeval, "Workload perception in educational resource recommendation supported by artificial intelligence: A controlled experiment with teachers", *Smart Learning Environments*, Vol. 12, No. 1, pp. 20, 2025.

[26] G. Prasad Arun, et al. "Artificial Intelligence in Computer Science: An Overview of Current Trends and Future Directions." Advances in Artificial and Human Intelligence in the Modern Era, edited by S. Suman Rajest et al., IGI Global Scientific Publishing, 2023, pp. 43-60.

[27] P. J, A. V. Senthil Kumar, and H. Mohammed Ali Abdullah, "A New Approach to Improve Energy Consumption Time and Life Time using Energy Based Routing in WSN," 2021 Emerging Trends in Industry 4.0 (ETI 4.0), Raigarh, India, 2021, pp. 1-6.

HISaaS: A Cloud-Based Framework for Semantic Enrichment of Legacy Health Databases Using OpenAI Embeddings

Toufik Salmi, Abdeslem Dennai
Department of Mathematics and Computer Science,
Faculty of Exact Sciences, Smart Grids & Renewable Energies Laboratory,
Tahri Mohamed University, Bechar, Algeria

***Abstract*—Legacy Information Systems (LIS) confront significant interoperability challenges, particularly at the database level. If existing solutions focus on data exchange, they fail to address the fundamental structural incompatibilities of Legacy Databases (LDBs). In this article, we present Health Information System as a Service (HISaaS), an innovative dual solution based on AI and Cloud Computing, designed either to update a LIS with a new modern HIS (Health Information System); or to completely propose a new IS. This new HIS is the result of the interoperability of several modern HIS. HISaaS uses Cloud computing as an interoperability environment, Natural Language Processing (NLP), and semantic similarity to analyze and enrich the already existing structure of the LDBs with new tables and/or new attributes from modern and interoperable external databases. This solution uses a vector-based approach relying on contextual embeddings from OpenAI models and cosine similarity measurement to accurately match schema elements. In the end, an evaluation based on real healthcare scenarios demonstrates the effectiveness of this HISaaS solution.**

***Keywords*—Information systems; interoperability; database; cosine measurement; cloud computing; OpenAI; embeddings; NLP**

I. Introduction

The success of large companies such as Alibaba, Google, and Amazon has been achieved through the use of reliable and intelligent information systems. Information systems (IS) play an essential role in any business, enabling it to stay ahead of the competition. The main components of an IS are software, hardware, databases, communications, and human resources. A database is at the heart of an IS; it must be well designed, regularly updated, and secure. Its data is essential for making the right decisions and improving a company's performance.

Interoperability between ISs in the same field has become a crucial requirement for organizations and individuals in order to ensure optimal performance and high-quality service delivery. IS interoperability is the ability of two or more ISs to exchange and use data. [1] defines interoperability as the ability of a computer system to work with other existing computer products or systems, without restrictions on access or implementation. As illustrated in Fig. 1, interoperability operates at several levels:

At the technical level, basic connectivity is established using common protocols and transport mechanisms to ensure secure data flow between different platforms. At the syntactic level, the focus is on ensuring that the exchanged data between systems preserves a consistent structure to facilitate accurate data parsing and processing. Semantic interoperability ensures that the meaning of the exchanged data is preserved and unambiguous among diverse systems by employing shared vocabularies, ontologies, and coding standards. Finally, organizational interoperability enables seamless, cross-institutional collaboration and coordinated use of shared information by aligning governance, policies, and business processes [2].

Despite these structured levels, interoperability challenges persist due to the heterogeneity of information systems. The need to integrate data from multiple disparate systems and collaborate with them, systems that use different communication protocols, data formats, and query languages, leads to the emergence of conflicts at technical, syntactical, and semantic levels. While existing approaches often address these challenges with integration tools and standardized interfaces, a critical issue remains: the lack of focus on database interoperability. Most solutions prioritize data exchange over fundamentally modernizing the legacy database structure itself, leaving core data silos and structural incompatibilities unaddressed.

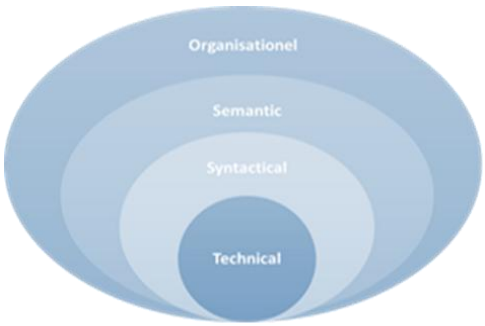


Fig. 1. Levels of interoperability.

Legacy information systems are known to be difficult to integrate with modern platforms due to significant interoperability issues, including the complexity of data migration and integration, dependence on obsolete data formats and structures, and the inability to communicate seamlessly with Cloud Computing services and APIs. To overcome these challenges, this article presents Health Information System as a Service (HISaaS), an innovative solution based on semantic AI

and Cloud Computing. With the scalability and flexibility of Cloud Computing, as well as advanced Natural Language Processing (NLP) techniques and semantic similarity measures, this HISaaS solution goes beyond simple data transfer by actively modernizing the very structure of the legacy database. By strategically adding new tables and/or attributes from external databases, HISaaS enriches and upgrades the existing system, making its data more useful and easily scalable without complete replacement, effectively bridging the gap between the existing infrastructure and the requirements of a modern, interoperable system.

The remainder of the study is organized as follows: Section II explores the context of this research, including Cloud Computing concepts and various techniques for measuring semantic similarity. Section III reviews related work in the field of IS interoperability. The architecture and detailed features of the proposed HISaaS solution are described in Section IV. Finally, Section V presents a comprehensive evaluation through a simulation based on the Cloud environment, detailing the performance of HISaaS in several real-world healthcare scenarios and demonstrating its effectiveness in structurally enriching a database, before concluding and proposing areas for future work in Section VI.

II. Background

This section covers two foundational pillars that our HISaaS solution builds upon. First, we look at Cloud Computing and explain why it serves as the ideal delivery platform for our service. Second, we explore semantic similarity techniques, particularly the vector-based approach using OpenAI embeddings that power the intelligent matching of database schemas across different healthcare systems. Understanding these concepts is essential before diving into how HISaaS actually performs the structural enrichment of legacy databases.

A. Cloud Computing

It is an environment for accessing shared computing resources servers, storage, databases, and software on demand via the Internet, rather than managing them locally. It transforms information systems by offering flexibility, scalability, and reduced infrastructure costs, with service models such as Infrastructure as a Service (IaaS), which offers fundamental compute, storage, and networking resources, Platform as a Service (PaaS), which provides complete development and deployment environment, and Software as a service (SaaS), which delivers fully managed, ready to use applications to end users [3][4] (see Fig. 2).

Fig. 2. Cloud computing basic delivery models.

Cloud Computing has become a pillar of modern information systems, enabling agility and digital innovation by offering agility and scalability through instant increase or reduction of resources according to real needs, mobility as the IS becomes accessible from anywhere and at any time, facilitating teleworking and real-time collaboration, cost optimization via the transition from heavy investment to subscription-based operating costs, and security and maintenance where updates and basic cybersecurity are ensured by the provider, freeing up IT teams for strategic projects [5].

In this work, Cloud Computing serves as the foundational delivery framework for the proposed HISaaS solution. It provides the scalable infrastructure and service-oriented model necessary to aggregate database schemas from multiple providers and offer new IS or database modernization as an on-demand service. Our HISaaS service uses the Cloud environment to provide seamless, secure connectivity to legacy systems and facilitates the structural enrichment of their databases without requiring significant on-premise resources, thereby directly handling interoperability issues.

B. Semantic Similarity

Semantic similarity refers to the degree of similarity between two texts (words, expressions, sentences, or documents) in terms of meaning. It is a key concept in computational linguistics and natural language processing (NLP), whose goal is to calculate the degree of similarity between two texts expressing identical or similar ideas, even if they use different words or structures [6]. Semantic similarity goes beyond simple text matching. It is based on the context or meaning of the text. “Medicine” and “Drug,” although not the same words, are semantically similar.

Semantic similarity methods can be systematically categorized based on the type of knowledge representation and computational strategy [7]. Multiple methods for semantic similarity exist, such as:

- Lexical methods: based on synonym matching (e.g., WordNet) [8].
- Vector-based methods: represent sentences or words as vectors in high-dimensional space (e.g., Word2Vec, Glove, FastText) [9] [10][11]. Here, the similarity is measured using cosine similarity or Euclidean distance.
- Contextual Embeddings: capture context-aware representations of text by using advanced models like BERT[12], GPT[13], or other transformer-based models.
- Knowledge-based Methods: based on hierarchical relationships, these methods measure similarity by using ontologies or semantic networks (e.g., WordNet).

Techniques for measuring similarity can be classified into multiple distinct approaches [14]; each one offers a unique perspective on assessing how closely related two concepts or objects are. The techniques can be categorized as follows:

- Arc-based: use the hierarchical structure of an ontology to assess semantic similarity by measuring the distances

between objects. The principle of this method is that the similarity between two concepts can be measured by the shortest path connecting them via a common ancestor. Wu and Palmer's metric [15] calculates similarity by assessing the distances from each concept to their least common ancestor, thereby providing a quantitative assessment of their semantic proximity.

- Nodes-based: analyses the information content of each concept to assess how similar they are. The main idea here is that two concepts are more similar when they share more information. Resnik's measure[16] is a well-known example. It uses the Information Content (IC) of the most specific concept that includes both concepts to compute their similarity. This approach focuses on the informational connections between concepts rather than just their structural connections within the ontology.
- Hybrid techniques: represent a synthesis of based-arcs and nodes approaches; these methods combine distance measures with information content, allowing for a nuanced assessment that captures both structural and semantic relationships. Jiang and Conrath's measure [17] is an illustrative example, which merges entropy (representing information content) with the distance between concepts. The combination of these approaches aims to enhance the accuracy of similarity evaluations, addressing the limitations found in using either technique in isolation.
- Vector space technique: these techniques are widely adopted in the field of information retrieval. These approaches represent objects as vectors within a multi-dimensional space, using various mathematical methods to calculate similarity. Cosine similarity[18], [19], for example, calculates the cosine of the angle between two vectors, indicating their similarity based on internal content. Additionally, Euclidean distance [20] calculates the straight-line distance between two points in the vector space; often adjusted to improve similarity assessments. The Jaccard index [21] further contributes to this domain by measuring the similarity between finite sample sets, defined as the size of their intersection divided by the size of their union. Each of these vector space techniques has unique strengths, making them applicable in various contexts depending on the nature of the data being analyzed.

Table I illustrates semantic similarity methods, categorizing different approaches for measuring how similar two pieces of text are in meaning. It organizes these methods into families based on their underlying principles and techniques.

Semantic similarity serves as the intelligent core for our proposed HISaaS that enables database enrichment from disparate databases. It is used to power two critical functions: synonym creation and embedding creation to match database schema elements (table names, attributes). Our method employs a vector-based approach to semantic similarity, using contextual embeddings generated by OpenAI's models to represent database schema elements as high-dimensional vectors. The similarity between these elements is then computed using *Cosine Similarity* on their respective vector representations. This approach allows our system to perform sophisticated semantic matching of database elements, identifying that "patient_tbl" and "client_records" are semantically related despite naming conventions, thereby enabling accurate structural enrichment of legacy databases.

TABLE I. SEMANTIC SIMILARITY METHODS

Method Family	Underlying Knowledge Representation	Example Techniques	Representative Models/Measures
Lexical Methods	Lexical databases (WordNet)	Synonym matching, string similarity	Levenshtein, Jaccard, WordNet Synsets
Knowledge-Based Methods	Ontologies / Semantic networks	Path-based, IC-based, Hybrid	Wu–Palmer, Resnik, Jiang–Conrath
Vector-Based Methods	Distributional semantics	Cosine, Euclidean, Jaccard	Word2Vec, GloVe, FastText
Contextual Embeddings	Transformer models	Contextualized embeddings	BERT, GPT, Sentence-BERT
Hybrid Methods	Combined symbolic + statistical	Ontology + embeddings fusion	BERT + WordNet, ConceptNet + Word2Vec

III. RELATED WORKS

This section reviews existing solutions related to information system interoperability and database schema integration. The review is organized around three themes directly relevant to our proposed HISaaS framework: health information system interoperability, cloud-based integration approaches, and AI-driven schema matching and structural enrichment. For each group, we discuss what the prior work contributes and where it falls short specifically regarding the structural modernization of legacy databases, which is the core gap our work addresses.

A. Health Information System Interoperability

A substantial body of work targets semantic interoperability in the healthcare domain through standard terminologies, ontologies, and data exchange protocols. Pellison et al. [22] proposed a semantic Web layer for tuberculosis data exchange in Brazil, using ontologies built on the Basic Formal Ontology (BFO) and SPARQL queries to map entities across heterogeneous Health Information Systems. The approach achieved structured, semantically consistent data exchange but remained tightly coupled to the existing system infrastructure and hard to extend to other diseases. Roehrs et al. [23] presented the OmniPHR model for personal health record interoperability, combining HL7 FHIR, OpenEHR, and NLP-driven automation to construct a unified view of patient data across disparate systems. While the model demonstrated integration potential, its NLP component needed extensive retraining for diverse medical terminologies and lacked real-world validation at scale.

More recent systematic reviews have taken stock of the state of semantic interoperability in EHRs (Electronic Health Records). Palojoki et al. [24] conducted a systematic review of approaches for enhancing semantic interoperability in EHR systems, identifying key elements such as shared terminologies (SNOMED CT, ICD, LOINC), standard models (FHIR,

OpenEHR), and cross-border data exchange mechanisms. Their review confirmed that achieving interoperability without loss of meaning remains unsolved and that most real-world deployments require considerable manual effort. Amar et al. [25] performed a systematic mapping review of 70 FHIR-based studies from 2012 to 2022, classifying semantic approaches into six categories including mapping, ML/NLP, ontology, and terminology services. They found that NLP and machine learning hold promise for unstructured EHR data but that automation of annotation and ontology comparison still needs significant human intervention. Critically, neither of these studies addresses structural enrichment of the legacy database itself; they focus on how data is exchanged, not on how the underlying schema is modernized.

B. Cloud-Based Interoperability Frameworks

Cloud computing has opened new paths for interoperability, particularly around scalability and service-oriented data management. Mohammed Sha and Rahamathulla [26] proposed a cloud-based healthcare data management framework that combines HL7, FHIR, and DICOM formats with a centralized data lake and API integration to streamline cross-system data sharing. The framework produced a scalable cloud system but faced persistent issues with legacy system compatibility, data ownership, and vendor lock-in.

Gohar et al. [27] introduced a five-layered patient-centric framework combining blockchain, cloud storage, and IoT for secure and semantically interoperable data exchange. Using SNOMED-CT, HL7 FHIR, and ICD standards, the framework outperformed existing blockchain-based approaches in security and exchange efficiency. Still, it remained at the prototype stage and lacked large-scale real-world testing.

Sony et al. [28] proposed an Edge Fog-based architecture integrating IoT, cloud, and the UMLS ontology to improve semantic interoperability in healthcare systems. The work showed reductions in computation overhead, latency, and network usage. However, the study relied entirely on simulations and raised concerns about inter-fog communication and data security. These cloud-based approaches share a common pattern: they create interoperability layers that facilitate data movement and exchange, but do not address the structural incompatibilities of the legacy database itself. They assume the source schema is adequate and focus on translation rather than modernization

C. AI-Driven Schema Matching and Structural Enrichment

The most directly related strand of recent work involves the use of language models for schema matching, the task of identifying semantic correspondences between elements of different relational schemas. Parciak et al. [29] ran an experimental study on using off-the-shelf LLMs for schema matching in the health domain, varying the amount of context information passed to the model through different "task scopes". Their results showed that LLMs can meaningfully bootstrap the matching process using only element names and descriptions, but that quality suffers when too much or too little context is provided. The study did not address multi-source schema enrichment or attribute-level gap analysis.

Sheetrit et al. [30] proposed ReMatch, a retrieval-augmented schema matching system that uses embedding-based candidate retrieval to narrow the search space before passing candidates to GPT-4 for ranking. The approach achieved strong results but showed significant performance degradation with smaller models and was not tested in healthcare scenarios involving legacy database upgrades. Liu et al. [31] developed Magneto, a hybrid pipeline that combines small and large language models for schema matching. A two-stage process, first narrowing candidates via contrastive column embeddings, then scoring with an LLM, outperformed single-model approaches and reduced the cost of labeled training data. That said, Magneto operates at the column level and does not attempt to identify missing attributes for structural enrichment. Ma et al. [32] introduced a knowledge graph-based retrieval-augmented generation approach for schema matching, using entity-type hierarchies to improve disambiguation for unseen schema elements. The method showed robustness under zero-shot conditions, but requires a pre-built knowledge graph and has not been evaluated in multi-provider legacy database settings.

Across all three research threads, a critical gap is consistent: existing solutions create interoperability layers to facilitate data exchange, but neglect the structural modernization of legacy databases. Schema matching work focuses on finding correspondences, not on using those correspondences to enrich a schema with missing tables and attributes. Health interoperability work relies on exchange standards and ontologies, not on semantic enrichment of the source database structure. The proposed HISaaS framework is designed specifically to fill this gap. It moves beyond data translation to actively identify missing schema elements (tables and attributes) from multiple cloud-hosted providers, using contextual embeddings and cosine similarity, and offers them to the user for structural integration.

IV. PROPOSED SYSTEM

HISaaS Health Information System as a Service is a cloud-delivered framework that gives organizations two distinct paths: they can build a completely new database from tables sourced across multiple cloud providers, or they can upgrade an existing legacy database by enriching it with new tables and attributes drawn from those same sources. The framework does not force a total system replacement. Instead, it works with what the client already has, adding structure where it is missing and leaving the rest intact.

The cloud platform is central to how HISaaS delivers both flexibility and scale. Because all schema catalogs are hosted and served from cloud infrastructure, clients can explore, compare, and compose database structures without requiring significant on-premise resources. The service adapts to organizations of different sizes, from a small clinic building its first schema to a hospital network looking to modernize a decades-old system.

A. General Model

The HISaaS workflow begins when a client selects a health domain from a catalog of available options. Once a domain is chosen, the system presents a list of information systems

offered by registered cloud providers that match that domain. For each selected system, a database schema is displayed, showing its tables and the attributes within them.

The client then chooses which tables to incorporate into their own database. This selection step is iterative: a client can work through multiple provider systems in sequence, adding tables from each one until the composite schema fully meets their requirements. The output of this process is either a new database ready to serve a new information system, or a set of enrichment operations to be applied to an existing one. Fig. 3 illustrates the full general workflow.

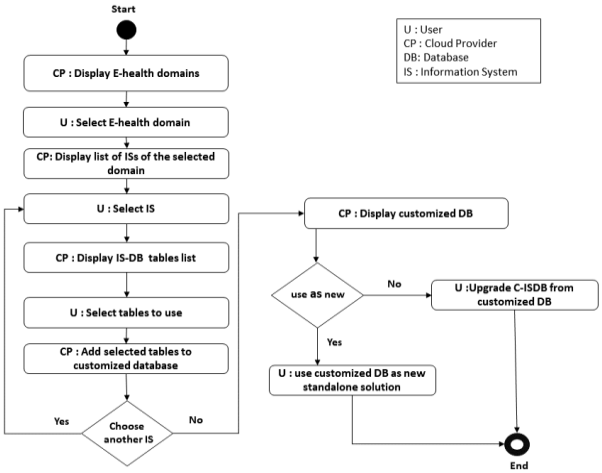


Fig. 3. HISaaS general model workflow.

B. 1st Scenario: Deployment of a New Custom Database

If the client needs to use a new customized database as a solution, he can either use it as it is or further customize it to better meet their needs. This customization allows for selecting the tables to include in the composite database and modifying them by choosing only the attributes necessary to meet the client's requirements (see Fig. 4).

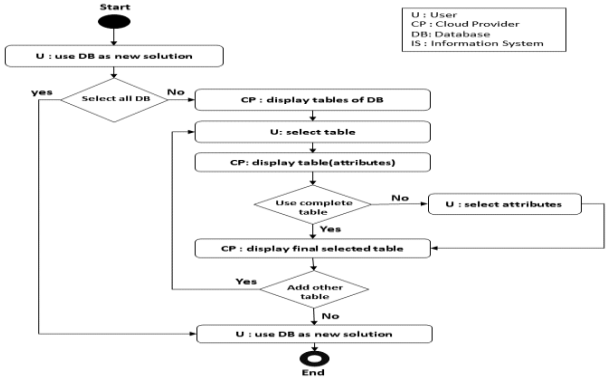


Fig. 4. Workflow for constructing a new custom database.

C. 2nd Scenario: Upgrading User Database

The scenario for updating the user database begins with uploading their database to the HISaaS platform. During the upload, HISaaS extracts and analyzes the metadata of the database, particularly the names of all the tables and their corresponding attributes, as illustrated in Fig. 5. The client has two options to improve their existing database:

1) Integrate a new table: The client can select a table from the custom database and integrate it into their existing IS database. He can use it as it is or customize it with selected attributes.

2) Update an existing table: The client can also select a table from the custom database and ensure that HISaaS finds a corresponding table in the client database. If a match is found, HISaaS identifies the attributes present in the selected table but absent in the client's table, and those can be integrated to enrich the existing data structure. This option enhances the data while preserving the core structure and familiarity of the existing dataset.

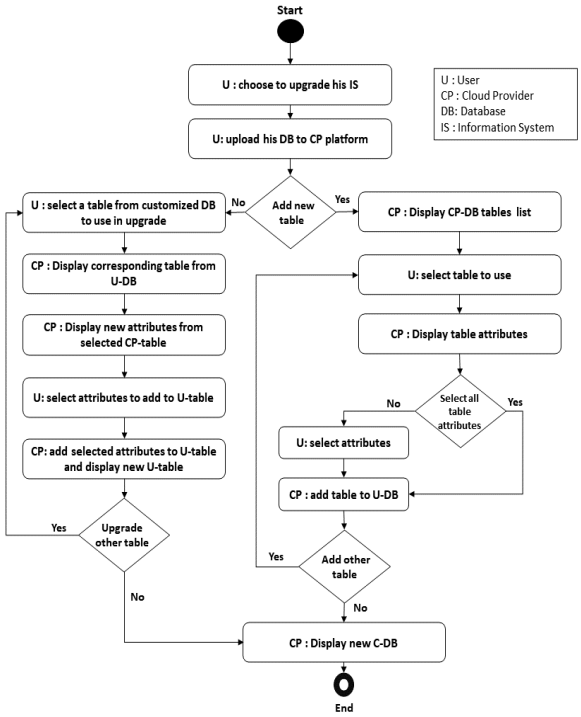


Fig. 5. Workflow for upgrading an existing legacy database.

D. HISaaS Database Enrichment Algorithms

Our proposed system is powered by two main algorithms that utilize OpenAI's advanced natural language processing (NLP) capabilities. First, identification of the most semantically relevant table for an enrichment request, and then the search for new attributes in the selected table from the customized database that are absent in the identified table in the client's database. These two algorithms facilitate the semantic matching of database schema elements (tables, attributes) by relying on the power of contextual embeddings and semantic similarity measures to overcome the heterogeneity and naming inconsistencies inherent in disparate information systems.

Semantic similarity between any two schema elements, whether table names or attribute names, is quantified by computing the cosine similarity between their respective embedding vectors. Given two dense vectors A and B produced by the embedding model, cosine similarity is defined as:

$$\text{cosine_sim}(\boldsymbol{A}, \boldsymbol{B}) = \frac{\boldsymbol{A} \cdot \boldsymbol{B}}{\|\boldsymbol{A}\| \cdot \|\boldsymbol{B}\|} \quad (1)$$

where, A and B are the high-dimensional vector representations of two schema elements, A • B is their dot product, and ‖ A ‖ and ‖ B ‖ are their respective Euclidean norms. The resulting score lies in [0, 1] for text embeddings: a score of 1 indicates identical semantic direction, and 0 indicates complete orthogonality. Cosine similarity is particularly well-suited to this task because it is invariant to vector magnitude; two schema element names of very different character lengths are compared purely based on their semantic direction in the embedding space, not their raw size.

The embedding vectors are produced by OpenAI's *text-embedding-3-small* model, which maps each schema element name to a 1536-dimensional dense vector. These vectors encode contextual meaning rather than surface form, enabling the system to capture semantic relationships that character-level or keyword-based matching would miss entirely.

We select a similarity threshold of t = 0.80, validated both by established literature and our own experimental tuning. Externally, ontology matching adopts t = 0.80 for Language-agnostic BERT Sentence Embedding (LaBSE)[33]; BERT-based extraction achieves 88% accuracy at ⩾0.80 [34]. Our systematic sweeps independently confirmed 0.80 as the optimal precision–recall trade-off for HISaaS schema matching: lower values introduce spurious links (e.g., "patient_id" ↔ "physician_id"), while higher values miss genuine synonyms expressed through different naming conventions. Hence, t = 0.80 is applied throughout.

1) Procedure for finding the corresponding table: In this scenario, once the client selects a table in the custom database for the upgrade process, the system searches for a corresponding table in the client's database. The search process is carried out as follows:

a) Phase 1 Syntactic search: HISaaS starts by performing a syntactic search, seeking an exact match between the table name chosen by the user and table names present in the user's database. This is the fastest and most efficient method, but is often unsuccessful due to variations in naming conventions (e.g., "patient", "patient_tbl", "client_records").

b) Phase 2 Semantic search (contextual matching): If no syntactic match is found, HISaaS moves on to a more advanced search using OpenAI's advanced natural language processing capabilities to identify the semantics (meaning) of the table name.

This semantic search allows discovering tables that do not have an exact syntactic match with the name but are semantically related to the table the client wishes to update. This ensures that HISaaS can identify the most relevant table in the client's database, if it exists, even when the naming conventions or terminology used by the client differ from those in the database.

c) Algorithm 1: Finding the Corresponding Table

Algorithm 1 Explanation: The algorithm uses two components of the OpenAI Python library. Synonym generation is handled by ChatCompletion.create with GPT-4. Embedding generation is handled by embedding.create with the text-embedding-3-small model, chosen for its strong performance on semantic search benchmarks and its substantial improvement over the previous text-embedding-ada-002 generation at lower operational cost.

Algorithm 1: Find_Corresponding_Table

Input:

```
CP_DB     : list of table names in the customized database
Cu_DB     : list of table names in the user's database
Upg_table : table selected by client from CP_DB for upgrade
sim_thresh : minimum cosine similarity score (t = 0.80)
max_syn   : maximum number of synonyms to generate
```

Output:

```
CR_table  : matched table from Cu_DB, or None if not found
```

Phase 1 — Syntactic Search:

```
CR_table ← None
FOR each table IN Cu_DB DO
    IF table == Upg_table THEN
        CR_table ← table
        RETURN CR_table
    END IF
END FOR
```

Phase 2 — Semantic Search Preparation:

```
IF CR_table == None THEN
    // Generate synonyms for the upgrade table name via GPT-4
    synonyms ← Generate_SYNONYMS(Upg_table, max_syn)
    synonyms ← synonyms + [Upg_table]   // include original term

    syn_embs ← []
    FOR each syn IN synonyms DO
        syn_embs.append(GET_EMBEDDINGS(syn))
    END FOR

    cu_embs ← {}
    FOR each table IN Cu_DB DO
        cu_embs[table] ← GET_EMBEDDINGS(table)
    END FOR
```

Phase 3 — Semantic Matching:

```
best_match ← None
highest_sim ← 0
```

```
    FOR each (syn, syn_emb) IN zip(synonyms, syn_embs) DO
        FOR each (table, cu_emb) IN cu_embs DO
            // Apply Equation (1)
            sim ← COSINE_SIM(syn_emb, cu_emb)
            IF sim > highest_sim AND sim >= sim_thresh THEN
                highest_sim ← sim
                best_match  ← table
            END IF
        END FOR
    END FOR

    CR_table ← best_match IF highest_sim >= sim_thresh ELSE None
    END IF

RETURN CR_table
```

- Synonym Generation: The GENERATE_SYNONYMS function submits the upgrade table name to GPT-4 with a tailored prompt and receives a list of semantically related terms. These synonyms broaden the search space so that the system can recognise equivalent concepts even when they are expressed with different vocabulary across systems.
- Embedding Generation: GET_EMBEDDINGS transforms each synonym, the original table name, and all user table names into 1536-dimensional dense vectors encoding contextual meaning rather than surface form, enabling similarity computation to capture relationships that character-level matching would miss.
- Semantic Similarity: Cosine similarity [Eq. (1)] measures directional alignment between embedding vectors. Two conditions must hold simultaneously for a match to be accepted:
    - The similarity score must reach or exceed the threshold t = 0.80.
    - The score must be strictly higher than the best score recorded for that client table against any previously processed synonym.

Together, these conditions ensure that the returned match is the most semantically coherent candidate across the full synonym set, correctly linking, for example, “medications” to “drug_inventory” or “billing” to “invoice” despite the absence of any shared characters.

2) Procedure for identifying missing attributes: Once a corresponding client table is found, the system moves to the attribute level. The objective here differs from table matching: rather than finding a single best match, the system determines which attributes of the upgrade table are genuinely absent from the client table. Attributes that already exist by exact name or by semantic equivalence are filtered out. Only attributes that survive both filtering passes are surfaced as enrichment candidates.

a) Phase 1 Syntactic search: The algorithm checks each attribute in the upgrade table against the client table’s attribute list. Any attribute found by exact name is skipped; unmatched attributes are forwarded to the semantic phase.

b) Phase 2 Semantic search: For each unmatched attribute, GPT-4 generates synonyms. Both the synonyms and the client attributes are embedded. Cosine similarity [Eq. (1)] is computed across all synonym–client attribute pairs. If any pair exceeds t = 0.80, the attribute is considered present under a different name and excluded from the output. Only attributes for which no synonym achieves the threshold are added to the missing-attributes list.

c) Algorithm 2: Identifying Missing Attributes

Algorithm 2 Explanation:

- Syntactic Search: The algorithm iterates through CP_attrs. Attributes found verbatim in Cu_attrs are skipped; unmatched attributes are forwarded immediately to the semantic phase.
- Semantic Search: For each unmatched attribute, the system generates synonyms (e.g., for “user_id”: [“client_id”, “customer_id”, “cust_id”]), embeds them together with all Cu_attrs, and runs cosine similarity comparisons using Eq. (1). As soon as any synonym–attribute pair crosses t = 0.80, the attribute is flagged as present and the inner loop exits early. This early-exit design keeps computational cost proportional to the difficulty of the matching problem.
- Result: Attributes for which neither phase finds a match are appended to new_attrs. This list is presented to the client, who decides which attributes to add to their table. The human decision step is intentional: the system identifies enrichment candidates but does not automatically modify the schema, preserving full user control over the final database structure.

Algorithm 2: Upgrade_Table

Input:

```
CP_table   :  source table from the customized database
Cu_table   :  target table in user's database (from Alg. 1)
CP_attrs   :  list of attribute names in CP_table
Cu_attrs   :  list of attribute names in Cu_table
sim_thresh :  minimum cosine similarity score (t = 0.80)
max_syn    :  maximum synonyms to generate per attribute
```

Output:

```
new_attrs  :  attributes in CP_table
```

```
    absent in Cu_table
```

Phase 1 — Syntactic Search:

```
new_attrs ← []
FOR each attr IN CP_attrs DO
    IF attr IN Cu_attrs THEN
        CONTINUE   // attribute already exists — skip
    ELSE
        GOTO Phase 2 with attr
    END IF
END FOR
```

Phase 2 — Semantic Search:

```
FOR each unmatched attr IN CP_attrs DO
    synonyms ← Generate_SYNONYMS(attr, max_syn)
    synonyms.append(attr)
    syn_embs ← [GET_EMBEDDINGS(s) for s in synonyms]
    cu_embs  ← [GET_EMBEDDINGS(a) for a in Cu_attrs]

    is_found ← False
    FOR each syn_emb IN syn_embs DO
        FOR each cu_emb IN cu_embs DO
            // Apply Equation (1)
            sim ← COSINE_SIM(syn_emb, cu_emb)
            IF sim >= sim_thresh THEN
                is_found ← True
                BREAK
            END IF
        END FOR
        IF is_found THEN BREAK
    END FOR

    IF NOT is_found THEN
        new_attrs.append(attr)
    END IF
END FOR
```

RETURN new_attrs

V. Results and Discussion

To evaluate HISaaS proposed in a real-world setting, we created a Cloud environment using the OpenNebula Cloud platform. The infrastructure was deployed on a physical host machine with the following specifications: an Intel Core i7 10th Gen processor (6 Cores @ 2.59 GHz), 64 GB of RAM, and 1 TB of NVMe storage. The host machine ran Ubuntu Desktop 22.04 LTS. Within this environment, Oracle VirtualBox 7.0 was used to create a virtualized instance, which also ran Ubuntu Desktop 22.04 LTS. On this guest operating system, the OpenNebula 6.8 cloud management platform was deployed to orchestrate the creation and management of the virtual machines (VMs) required for the simulation.

The environment consisted of eight virtual machines (VMs) representing the cloud providers and user organizations. As shown in Fig. 6, three provider VMs (P1, P2, P3) with 2vCPUs, 12 GB RAM,100 GB storage, and running Ubuntu Server 22.04 LTS hosted real-world healthcare datasets: P1 hosting MIMIC-III (Medical Information Mart for Intensive Care) [35]Clinical Database V1.4, P2 hosting eICU (electronic intensive care unit) Collaborative Research Database [36], and P3 hosting a private hospital database. Four user VMs with 2vCPUs, 8 GB RAM,40 GB storage, and running Ubuntu Server 22.04 LTS represented clinics with legacy systems seeking to upgrade their schemas or build a new one. The HISaaS VM with 3vCPUs, 16 GB RAM,200 GB storage, and running Ubuntu Server 22.04 LTS hosted the core HISaaS service, responsible for coordinating semantic matching, schema analysis, and the enrichment processes across all provider and user nodes.

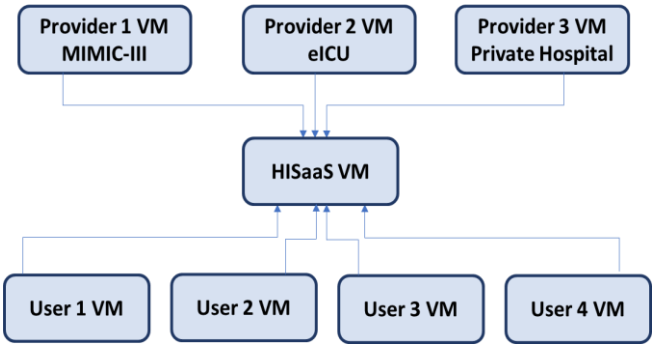


Fig. 6. Deployment architecture of the HISaaS cloud-based evaluation environment.

A. Use Case Scenario and Outcomes

Four users participated in the simulation, each performing schema enrichment operations. The outcomes are detailed below.

1) User 1 Targeted schema upgrade: The first user started with a database of 7 tables and 45 attributes: patients table with 8 attributes, doctors table with 7 attributes, appointments table with 6 attributes, medical records table with 7 attributes, billing table with 6 attributes, medication table with 5 attributes, and prescriptions table with 6 attributes.

The user created a customized database of 29 tables. Added four tables: doctors_details and Patient_medical_info from private hospital, icu_stays from private hospital, and prescriptions from MIMIC-III.

Upgraded four Existing tables:

- Patients (8 attributes) using patient_tbl (21 attributes) from private hospital; the matching table was found via semantic search with a cosine similarity of 0.83. Syntactic search identifies 21 new attributes, while

semantic search identifies 13 new attributes. The user chose 10 attributes to add to his Patients table

- billing (6 attributes) using invoice (11 attributes) from private hospital; the matching table was found via semantic search with cosine similarity of 0.88. Syntactic search identifies 10 new attributes, while semantic search identifies 7 new attributes. The user chose 7 attributes to add to his billing table.
- doctors (7 attributes) using doctors_tbl (17 attributes) from a private hospital; the matching table was found via semantic search with a cosine similarity of 0.82. Syntactic search shows 17 new attributes, while semantic search shows 15 new attributes. The user chose 12 attributes to add to his doctor's table
- medications (5 attributes) using medication (15 attributes) from eICU; the matching table was found via semantic search with a cosine similarity of 0.96. Syntactic search shows 13 new attributes, while semantic search also shows 13 new attributes. The user chose 4 attributes to add to his medications table.

The upgrade process was both precise and selective: the user did not adopt all new attributes suggested by the system (e.g., selecting only 10 of 13 semantically suggested for the "Patients" table). This reflects the effectiveness and the flexibility of our system. The final database schema grew to 11 tables (+57.1%) and 139 attributes (+208.9%), representing substantial structural enrichment. Fig. 7 illustrates this enrichment.

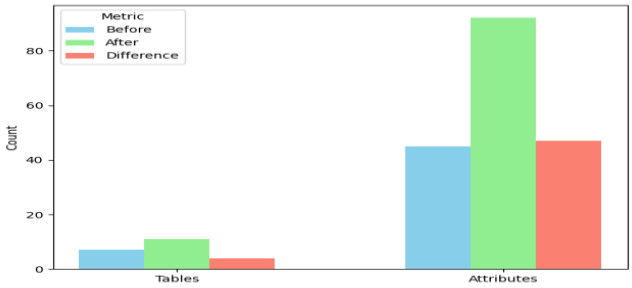


Fig. 7. User 1 database enrichment statistics.

2) User 2 Large-scale schema expansion: The second user began with a database of 5 tables and 31 attributes: patient_inf (7 attributes), visits (6 attributes), departments (3 attributes), health_records(6 attributes), and billing_records (6 attributes).

The user created a customized database of 26 tables. They performed a large-scale integration, added 14 tables: doctors_detail, doctor_tbl, email_config, email_delivery, email_inf, email_schedul, email_template, insurance, payment_account, payment_table, and user_tbl from private hospital. nurceAssissment, nurceCare, and nurceChart from eICU.

Upgraded three existing tables: "billing_records" using "invoice" from private hospital with 0.91 similarity, "health_records" using "patient_medical_info" from private hospital with 0.88 similarity, and "patient_inf" using "patient_tbl" from private hospital with 0.92 similarity.

All table matches were discovered using semantic search with similarities above 0.88, ensuring high confidence in the results. The outcome was a transformation to 19 tables (+280%) and 154 attributes (+396.8%), as illustrated in Fig. 8. This case highlights the capacity of HISaaS to support rapid schema expansion.

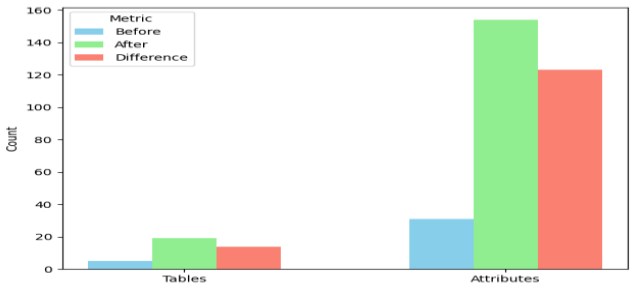


Fig. 8. User 2 database enrichment statistics.

3) User 3 Selective Enrichment: The third user starts with a database of 11 tables and 63 attributes: accounts (5 attributes), patients (7 attributes), medical_providers (6 attributes), patient_clinical_history (7 attributes), consultations (6 attributes), consultation_notes (4 attributes), prescriptions (6 attributes), transactions (6 attributes), support_tickets (6 attributes), chat_transcripts (5 attributes), system_logs (5 attributes). Created a customized_dB of 12 tables, chose to enrich their schema by adding 5 new tables from MIMIC-III and the private hospital database: payment_account_setup (4 attributes), payment_table (9 attributes), and schedul_table (8 attributes) from the private hospital; ICUSTAYS (12 attributes) and TRANSFERS(13 attributes) from MIMIC-III.

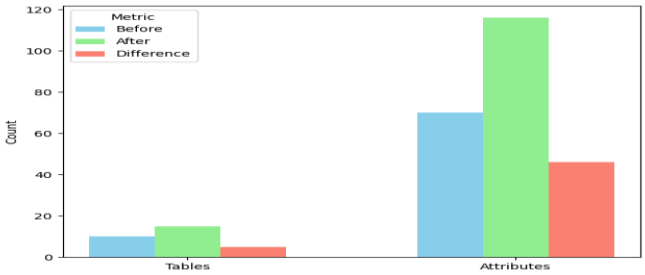


Fig. 9. User 3 database enrichment statistics.

User 3 kept the majority of their original database structure (11 initial tables), but enriched their schema by adding 5 new tables, leading to a final 16 tables. There were no table upgrades made. But the number of attributes increased by 53, from 63 to 116 (+84.13%) (See Fig. 9).

4) User 4 New schema construction: User 4 represented a new clinic without an existing schema. They used HISaaS to build a new customized database from scratch, composing it of 14 tables and 169 attributes with 14 tables and 169 attributes, representing a 100% new construction, sourced from all three providers: user_tbl (11 attributes), doctor_tbl (17 attributes), doctor_details (6 attributes), invoice (11 attributes), invoice_product (7 attributes), patient_tbl (21 attributes) and patient_medical_info (22 attributes) from

private hospital; nurseAssessment (8 attributes), nurseCare (8 attributes) and nurseCharting (8 attributes) from eICU; ICUSTAYS (12 attributes), PRESCRIPTIONS (19 attributes), SERVICES (6 attributes) and TRANSFERS (13 attributes) from MIMIC-III.

This use case demonstrates the system's utility for greenfield projects, enabling the rapid construction of a comprehensive, modern database.

B. System Performance Evaluation

The four user scenarios showcased distinct patterns of use, from targeted upgrades to full-scale schema construction, reflecting the HISaaS service flexibility and its adaptation to varying interoperability challenges and modernization needs across different organizational contexts. A summary of the enrichment across all users is provided in Table II.

TABLE II. ENRICHMENT STATISTICS ACROSS ALL USERS

User	Tables before	Tables After	Tables gain %	Attributes before	Attributes after	Attributes gain %	Tables added	Tables upgraded
User1	7	11	+57.1%	45	139	+208.9%	4 (from MIMIC-III and Private hospital)	4 (patients, billing, doctors, medication)
User2	5	19	+280%	31	154	+396.8%	14 (mixed sources)	3 (billing_record, health_records, patient_inf)
User3	11	16	+45.45%	63	116	+84.13%	5 (from MIMIC-III and private hospital)	0
User4	0	14	New (100%)	0	169	New (100%)	14 (mixed sources)	0

The results demonstrate substantial structural enrichment across all scenarios. User 2 achieved the highest relative growth in both tables (+280%) and attributes (+396.8%), reflecting a large-scale expansion of a previously minimal schema. User 1, while more selective, still realized significant attribute-level enrichment (+208.9%), underscoring the system's ability to perform deep, targeted upgrades. User 3 opted for a conservative approach, focusing on table additions rather than upgrades, resulting in a balanced increase in both tables (+45.45%) and attributes (+65.7%). User 4, building a schema from scratch, successfully constructed a fully modernized database with 14 tables and 169 attributes, demonstrating the system's utility for greenfield deployments.

To evaluate the system performance, we calculated the F1-score, precision, and recall for the applicable users (Users 1 and 2). The system achieved perfect performance in table matching for both users. The global precision, recall, and F1-score for table matching were all 1.00, indicating that the system successfully identified all relevant tables and avoided any false matches. The system also performed well in attribute-level enrichment. The global new attribute identification precision was 0.97, with a recall of 1.00, resulting in an F1-score of 0.98. The only recorded error occurred in User1's scenario, where the system recommended an attribute that already existed in the user's original table (doctors table upgrade), leading to a single false positive. This outcome highlights the overall accuracy and reliability of the HISaaS system in both structural and semantic enrichment tasks.

A key component of the system's efficacy is its use of semantic similarity, powered by OpenAI's advanced language models. This capability drives two critical functions within HISaaS: synonym generation and embedding creation, which are used for identifying corresponding tables and finding new attributes during the schema upgrade. By defining a cosine similarity threshold of 0.80, the system efficiently eliminated weak matches while maintaining recall, ensuring that most relevant structures were presented to the user. The HISaaS service proved particularly useful in real-world datasets where naming conventions and structural variations were significant across multiple providers. For example, HISaaS was able to match patients with patient_tbl, and billing with invoice, not based on name overlap, but on underlying semantic equivalence inferred from their embeddings.

The results show that HISaaS is well-suited for both schema upgrade and new schema construction tasks in heterogeneous healthcare environments. Its high precision reduces the need for manual correction, while its high recall ensures complete enrichment of the target schema. Using semantic similarity, the HISaaS effectively addresses the limitations caused by non-standardized naming conventions, an important challenge in healthcare data integration. These findings establish the HISaaS as a robust, scalable, and accurate approach for practical schema enrichment.

VI. CONCLUSION AND FUTURE WORK

This work presented a new approach to achieving interoperability in information systems based on a Cloud-VM environment by proposing a Health Information System as a Service (HISaaS) framework. The study conducted in this work highlighted the limitations of traditional interoperability solutions that primarily focus on data exchange and fail to address the fundamental structural issues of legacy databases. By using cloud computing and advanced AI-driven semantic similarity measurement techniques, the HISaaS solution offers a scalable and intelligent method to modernize legacy systems. HISaaS uses OpenAI's linguistic models to accurately identify semantically related tables and attributes to overcome naming conventions, terminologies, and schema design across different heterogeneous Cloud providers.

The evaluation was conducted in a simulated Cloud environment with real data from the healthcare sector. The results demonstrated that the proposed framework can significantly improve database structures, making them more complete and more interoperable without requiring a costly and large-scale replacement. HISaaS achieved perfect precision and recall in table matching with nearly perfect performance in attribute-level enrichment across various scenarios, including

targeted upgrades, large-scale expansions, selective enrichment, and complete schema construction.

In conclusion, the proposed solution addressed a critical issue in the creation and enrichment of IS by providing interoperability-based means in a Cloud-VM environment to improve the fundamental structure of clients' legacy databases.

Although the results presented in this work are promising, it is still necessary to make modifications to the HISaaS framework to improve its capabilities. These modifications will focus on further supporting the subdomains of healthcare, namely care, pharmacy, medical analyses, ... Moreover, it is important to extend the enrichment of the database schema beyond tables and attributes to include relationships, constraints, and metadata for a complete relational modernization. Moreover, automating the database enrichment process will minimize the need for manual interventions, thereby making the solution more accessible to a wider range of users and organizations.

Given that the HISaaS framework involves accessing and modifying sensitive data in the healthcare field (particularly patient data), future improvements should focus on investigating and strengthening security and privacy mechanisms in the Cloud environment.

REFERENCES

[1] G. da Silva Serapião Leal, W. Guédria, and H. Panetto, "Interoperability assessment: A systematic literature review," Comput. Ind., vol. 106, pp. 111–132, Apr. 2019, doi: 10.1016/j.compind.2019.01.002.

[2] K. S. S. Santos, L. B. L. Pinheiro, and R. S. P. Maciel, "Interoperability Types Classifications: A Tertiary Study," in XVII Brazilian Symposium on Information Systems, New York, NY, USA: ACM, Jun. 2021, pp. 1–8. doi: 10.1145/3466933.3466952.

[3] R. B. Bohn, J. Messina, F. Liu, J. Tong, and J. Mao, "NIST Cloud Computing Reference Architecture," in 2011 IEEE World Congress on Services, IEEE, Jul. 2011, pp. 594–596. doi: 10.1109/SERVICES.2011.105.

[4] F. Wulf, T. Lindner, M. Westner, and S. Strahringer, "IaaS, PaaS, or SaaS? The Why of Cloud Computing Delivery Model Selection – Vignettes on the Post-Adoption of Cloud Computing," 2021. doi: 10.24251/HICSS.2021.758.

[5] I. Odun-Ayo, R. Goddy-Worlu, V. Samuel, and V. Geteloma, "Cloud designs and deployment models: A systematic mapping study," BMC Res. Notes, vol. 12, no. 1, Jul. 2019, doi: 10.1186/s13104-019-4474-y.

[6] W. H.Gomaa and A. A. Fahmy, "A Survey of Text Similarity Approaches," Int. J. Comput. Appl., vol. 68, no. 13, pp. 13–18, Apr. 2013, doi: 10.5120/11638-7118.

[7] D. Chandrasekaran and V. Mago, "Evolution of Semantic Similarity -- A Survey," Jan. 2021, doi: 10.1145/3440755.

[8] G. A. Miller, "WORDNET: A Lexical Database for English," in Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992, 1992, pp. 417–419.

[9] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and Their Compositionality," in Advances in Neural Information Processing Systems (NIPS), 2013, pp. 3111–3119.

[10] J. Pennington, R. Socher, and C. Manning, "GloVe: Global Vectors for Word Representation," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), A. Moschitti, B. Pang, and W. Daelemans, Eds., Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. doi: 10.3115/v1/D14-1162.

[11] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," Trans. Assoc. Comput. Linguist., vol. 5, pp. 135–146, 2017, doi: 10.1162/tacl_a_00051.

[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 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[13] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training," 2018.

[14] A. DENNAI and S. M. BENSLIMANE, "A New Measure of the Calculation of Semantic Distance between Ontology Concepts," International Journal of Information Technology and Computer Science, vol. 7, no. 7, pp. 48–56, Jun. 2015, doi: 10.5815/ijitcs.2015.07.06.

[15] Z. Wu and M. Palmer, "Verb Semantics and Lexical Selection," in 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, USA: Association for Computational Linguistics, Jun. 1994, pp. 133–138. doi: 10.3115/981732.981751.

[16] Y. Bin, L. Xiao-Ran, L. Ning, and Y. Yue-Song, "Using Information Content to Evaluate Semantic Similarity on HowNet," in 2012 Eighth International Conference on Computational Intelligence and Security, IEEE, Nov. 2012, pp. 142–145. doi: 10.1109/CIS.2012.39.

[17] L. Jiang, J. G. González, E. N. Navas-Conyedo, and Y. Chen, "Methods of Community Detection in Hybrid Network Applied to the Academic Network Analysis," Revista Cubana de Ciencias Informáticas, vol. 15, no. 5, pp. 130–161, 2021.

[18] M. Belabed and A. Dennai, "A Double Metan-Semantic Search Model Based on Ontology and Semantic Similarity: Asthma Disease," Journal of Information & Knowledge Management, vol. 22, no. 02, p. 2250082, 2023, doi: 10.1142/S0219649222500824.

[19] A. Dennai and S. M. Benslimane, "Toward an Update of a Similarity Measurement for a Better Calculation of the Semantic Distance between Ontology Concepts," 2013, pp. 197–207.

[20] I. Dokmanic, R. Parhizkar, J. Ranieri, and M. Vetterli, "Euclidean Distance Matrices: Essential theory, algorithms, and applications," IEEE Signal Process. Mag., vol. 32, no. 6, pp. 12–30, Nov. 2015, doi: 10.1109/MSP.2015.2398954.

[21] V. A. Gonzalez-Huitron, A. E. Rodriguez-mata, L. E. Amabilis-Sosa, R. Baray-Arana, I. Robledo-Vega, and G. Valencia-Palomo, "Jaccard distance as similarity measure for disparity map estimation," IEEE Latin America Transactions, vol. 21, no. 5, pp. 690–698, May 2023, doi: 10.1109/TLA.2023.10130841.

[22] F. C. Pellison et al., "Data integration in the brazilian public health system for tuberculosis: Use of the semantic web to establish interoperability," JMIR Med. Inform., vol. 8, no. 7, 2020, doi: 10.2196/17176.

[23] A. Roehrs, C. A. Da Costa, R. D. R. Righi, S. J. Rigo, and M. H. Wichman, "Toward a Model for Personal Health Record Interoperability," IEEE J. Biomed. Health Inform., vol. 23, no. 2, pp. 867–873, Mar. 2019, doi: 10.1109/JBHI.2018.2836138.

[24] S. Palojoki, L. Lehtonen, and R. Vuokko, "Semantic Interoperability of Electronic Health Records: Systematic Review of Alternative Approaches for Enhancing Patient Information Availability," 2024, JMIR Publications Inc. doi: 10.2196/53535.

[25] F. Amar, A. April, and A. Abran, "Electronic Health Record and Semantic Issues Using Fast Healthcare Interoperability Resources: Systematic Mapping Review," Jan. 01, 2024, JMIR Publications Inc. doi: 10.2196/45209.

[26] M. Mohemmed Sha and M. P. Rahamathulla, "Cloud-based Healthcare data management Framework," KSII Transactions on Internet and Information Systems, vol. 14, no. 3, pp. 1014–1025, 2020, doi: 10.3837/tiis.2020.03.006.

[27] A. N. Gohar, S. A. Abdelmawgoud, and M. S. Farhan, "A Patient-Centric Healthcare Framework Reference Architecture for Better Semantic Interoperability Based on Blockchain, Cloud, and IoT," IEEE Access, vol. 10, pp. 92137–92157, 2022, doi: 10.1109/ACCESS.2022.3202902.

[28] P. Sony, G. Siva Shanmugam, and S. Nagarajan, "Enhancement of Semantic Interoperability in Healthcare Systems Using IFCIoT Architecture," KSII Transactions on Internet and Information Systems, vol. 18, no. 4, pp. 881–902, Apr. 2024, doi: 10.3837/tiis.2024.04.004.

[29] M. Parciak, B. Vandevoort, F. Neven, L. M. Peeters, and S. Vansummeren, "Schema Matching with Large Language Models: an Experimental Study," in Proceedings of the 2nd International Workshop on Tabular Data Analysis (TaDA24), Guangzhou, China, Aug. 2024. doi: 10.48550/arXiv.2407.11852.

[30] E. Sheetrit, M. Brief, M. Mishaeli, and O. Elisha, "ReMatch: Retrieval Enhanced Schema Matching with LLMs," arXiv preprint arXiv:2403.01567, 2024, doi: 10.48550/arXiv.2403.01567.

[31] Y. Liu, E. H. M. Pena, A. Santos, E. Wu, and J. Freire, "Magneto: Combining Small and Large Language Models for Schema Matching," in Proceedings of the VLDB Endowment, VLDB Endowment, 2025, pp. 2681–2694. doi: 10.14778/3742728.3742757.

[32] C. Ma, S. Chakrabarti, A. Khan, and B. Molnár, "Knowledge Graph-based Retrieval-Augmented Generation for Schema Matching," arXiv preprint arXiv:2501.08686, 2025, doi: 10.48550/arXiv.2501.08686.

[33] S. Banerjee, B. R. Chakravarthi, and J. P. McCrae, "Cross-Lingual Ontology Matching using Structural and Semantic Similarity," in Proceedings of the 9th Workshop on Linked Data in Linguistics @ LREC-COLING 2024, C. Chiarcos, K. Gkirtzou, M. Ionov, F. Khan, J. P. McCrae, E. M. Ponsoda, and P. M. Chozas, Eds., Torino, Italia: ELRA and ICCL, May 2024, pp. 11–21. [Online]. Available: https://aclanthology.org/2024.ldl-1.2/

[34] E. Herrmann, L. U. Von Rohr, F. Kotta Parambil, S. M Joy, and Y. Chittazhathu Kurian Kuruvilla, "Validating a semantic similarity approach for automated data extraction in phase II oncology trials.," Journal of Clinical Oncology, vol. 43, no. 16_suppl, Jun. 2025, doi: 10.1200/JCO.2025.43.16_suppl.e13666.

[35] A. Johnson, T. Pollard, and R. Mark, "MIMIC-III Clinical Database," PhysioNet, Sep. 2016, doi: 10.13026/C2XW26.

[36] T. J. Pollard, A. E. W. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi, "The eICU collaborative research database, a freely available multi-center database for critical care research," Sci. Data, vol. 5, Sep. 2018, doi: 10.1038/SDATA.2018.178.

HETRA: A Real-Time Traffic Detection and Directional Counting System Using YOLOv8 and DeepSORT

Kuldeep Kumar Shukla*, Ajai Kumar Singh
Civil Engineering Department, Motilal Nehru National Institute of Technology Allahabad, Prayagraj, 211004, India

***Abstract*—Rapid urbanization and economic growth in developing countries such as India have increased the demand for automated traffic analysis systems. However, heterogeneous traffic conditions present significant challenges to automated traffic analysis. This study proposes HETRA (HEterogeneous TRaffic Analysis), a real- time traffic monitoring framework that integrates YOLOv8 and DeepSORT for multi-class detection, tracking, and counting. For this, a custom traffic dataset was developed using real-world recorded traffic videos. The dataset consists of more than eight thousand annotated frames representing the fourteen traffic classes (motorized, non-motorized, and pedestrians). The proposed model leverages YOLOv8 for class identification and localization, while DeepSORT is used for multi-class tracking and directional traffic counting. The custom traffic dataset was used to train and assess the proposed HETRA model. The model achieved precision, recall, and mean average precision values of 71.15%, 65.51%, and 61.9%, respectively. The qualitative assessment of the model demonstrated its effectiveness under varying traffic scenarios and lighting conditions. Furthermore, it obtained an overall counting accuracy of 90%. These findings demonstrate the potential of the proposed framework for real-time traffic analysis and intelligent transportation systems.**

***Keywords*—Heterogeneous traffic; YOLOv8; DeepSORT traffic counting; intelligent transportation system**

I. INTRODUCTION

Over the last two decades, the fast-growing number of personal vehicles has resulted in an enormous increase in traffic congestion in cities, across developing countries, especially in India. Therefore, it is crucial to have an efficient traffic monitoring and management system for reducing travel time, decreasing fuel consumption, and supporting sustainable urban transportation goals. In recent years, Intelligent Transportation Systems (ITS) have gained tremendous attention as automated real-time traffic monitoring systems utilizing artificial intelligence (AI) and computer vision (CV) technology.

Vehicle detection, tracking, and counting are basic components of modern traffic monitoring systems. These are widely used for traffic flow analysis, congestion study, adaptive signal design, urban mobility planning, and road safety management. However, developing such an automated system for Indian traffic conditions remains highly challenging. The traffic patterns in India are fundamentally different from developed nations due to a heterogeneous traffic mix, - non-lane-disciplined behavior, poor roadway conditions, etc. The traffic in India consists of a vast composition of motorized and non-motorized vehicles competing for the same road space. Along with this, frequent overtaking, varying sizes, erratic lane practices, and dense traffic situations give rise to troublesome traffic class detection and tracking. Traditional intrusive sensor-based methods for traffic monitoring are inadequate as they provide only axle counts without traffic class differentiation. For traffic management, geometric design and other analyses, class-wise traffic counting is one of the mandatory requirements. These challenges accentuate the immediate need for reliable automated traffic analysis techniques capable of generating accurate traffic information, including vehicle classification, class-wise traffic volume, and movement patterns. Such information serves as the foundation of ITS, enabling applications including adaptive signal management, congestion prevention, accident detection, safety assessment, and also long-term transportation planning [1].

Conventional approaches used for traffic monitoring are: manual traffic counts, loop detectors, pneumatic tubes, radar and infrared sensors, and conventional image processing-based techniques. These methods exhibit considerable limitations in scalability, high installation cost, poor adaptability, and reduced accuracy under dynamic traffic circumstances [2]. The image processing approaches for traffic detection are based on traditional CV techniques like background subtraction [3], Support Vector Machine (SVM) [4], etc. However, their robustness is often compromised by lighting conditions, shadows, and occlusions encountered in real-world scenarios.

In recent years, deep learning-based computer vision techniques have demonstrated remarkable success in identifying, tracking, and traffic analysis [5]. Among these developments, convolutional neural networks (CNNs) have become the foundation of modern object detection algorithms. Existing deep learning-based detectors are broadly categorized into two-stage and one-stage approaches. Two- stage methods, including R-CNN [6], Fast R-CNN [7], and Faster R-CNN [8], first generate candidate object regions before performing classification, which generally results in high detection accuracy. However, their high latency in inference as well as computational complexities are major drawbacks in real-time traffic monitoring and surveillance applications. To address these constraints, single-stage detectors like Single Shot Detector (SSD) [9] and the You Only Look Once (YOLO) [10] family are introduced. SSD utilizes multi-scale feature maps to identify objects of diverse dimensions, yet it often has

*Corresponding author.

difficulties in detecting tiny objects, which is essential in traffic situations involving bicycles, bikes, and pedestrians [9].

Among these techniques, the YOLO method redefined object detection as a single regression task, allowing real-time identification with competitive detection precision [10]. YOLO9000 demonstrated the viability of integrated identification and categorization across several object categories while preserving real-time speed [11]. In YOLOv3, multi-scale predictions and residual learning were incorporated, which greatly promoted the accuracy of object detection of small objects [12]. YOLOv4 further boosted detection capabilities with optimized architecture and advanced feature fusion techniques, and reached state-of-the-art performance on test sets with high processing speeds [13]. These upgrades brought YOLO-based detectors into the spotlight as a leading tool for real-time object detection in smart traffic surveillance systems. However, they also showed their limitations in dense as well as mixed traffic conditions.

Recent research studies have investigated the use of YOLOv5 and YOLOv7 for urban traffic monitoring. Farid et al. established a real-time vehicle identification system using YOLOv5 and validated its efficacy in traffic surveillance contexts [14]. Likewise, traffic counting models created on deep learning-based techniques mostly emphasized organized traffic scenarios with a restricted range of vehicle classes [15].

Most of the existing studies primarily address structured traffic scenarios with few conventional vehicle classes. Although YOLOv8 and DeepSORT-based models have previously been applied for traffic monitoring, their assessment under heterogeneous traffic conditions remains insufficiently investigated. To address this limitation, this study develops and evaluates HETRA (HEterogeneous Traffic Analysis), an integrated YOLOv8-DeepSORT framework, for real-time detection, tracking, and class-wise directional counting of fourteen traffic categories under Indian traffic conditions. For this, the proposed framework was trained on a prepared custom traffic dataset of 8,476 manually annotated frames representing fourteen different traffic classes related to motorized vehicles, non-motorized vehicles, and pedestrians. The dataset includes region-specific traffic participants like auto, e-rickshaw, tempo, and horse-cart, which are often neglected in publicly available traffic datasets. The effectiveness of the developed model was evaluated using conventional metrics such as precision, recall, mean Average Precision (mAP), F1-score, confusion matrix, and directional counting accuracy. The novelty of the study is the thorough assessment of an integrated multi-class traffic analysis framework in dense, non-lane-based scenarios with frequent occlusions and substantial variations in traffic classes and size.

The major contributions of the study are as follows:

- Development of a custom traffic dataset consisting of 8476 manually annotated image frames representing fourteen different traffic classes in Indian traffic environments.
- Development and evaluation of HETRA- a YOLOv8 and DeepSORT-based framework for real-time multi-class detection, tracking, and directional traffic counting using the custom dataset.
- Assessment of HETRA under varying traffic densities and lighting conditions using quantitative and qualitative performance analysis.
- Demonstrate the suitability of the anticipated approach for real-time traffic monitoring, directional class-wise traffic flow estimation, and ITS applications.

II. Related Work

YOLOv8 brought significant architectural enhancements like the anchor-free detection strategy, decoupled detection heads, and refined loss functions. A number of recent works have used YOLOv8 in traffic-related tasks. Enhanced YOLOv8 has been utilized for identifying moving objects in videos, which showed superior accuracy and convergence characteristics compared to earlier YOLO variants [16]. Survey studies have recognized YOLOv8 as a viable model for real-time detection applications; however, its comprehensive evaluation under complicated, unstructured traffic situations is still restricted [17-19]. The YOLOv8 models suggested for vehicle counting in road traffic indicate accurate object detection performance in controlled environments [20].

Subsequent enhancements in YOLOv8 continued improvements in model performance under complex environments. Different architectural modifications in YOLOv8 were employed for the estimation of traffic flow parameters to improve the model precision [21]. Various works have utilized YOLOv8 for improving traffic analysis in ITS applications but only for limited classes of vehicles [22]. The real-time detection of small and distant traffic objects has remained one of the crucial challenges in traffic system monitoring. Many researchers employed different variants of improved YOLOv8 models for detecting small objects. Lightweight and modified YOLOv8 models were reported for the detection of small traffic signs and pedestrian-vehicle detection, with better precision in difficult environmental conditions [23, 24]. UAV-based studies also explored several improvements over YOLOv8 to detect small vehicles from an overhead view, which is quite different from the roadside camera viewpoint [25, 26]. While these works attained performance gains, most dealt with particular object categories or aerial imagery and didn't look into the complex issues of ground-level mixed traffic with overlapping objects and diverse vehicle types.

Multi-object tracking (MOT) has emerged as one of the critical aspects of traffic monitoring systems. DeepSORT (Simple Online and Realtime Tracking with Deep Association Metric) is one of the most popular approaches to MOT, which is achieved by integrating Kalman filtering with deep appearance features to maintain stable traffic class IDs between successive frames [27]. A vision- based vehicle detection and counting using YOLOv3 and Oriented FAST and Rotated BRIEF (ORB) tracking demonstrated effective performance under highway traffic scenarios [28]. Durve et al. performed a comprehensive benchmark analysis of YOLOv5 and YOLOv7 in combination with DeepSORT and highlighted the correlation between tracking precision and detector efficiency. It also emphasized the efficiency of the detector in congested environments [29].

Recent research studies have integrated YOLOv8 with MOT algorithms like DeepSORT to overcome the constraints of detection-only systems. The authors have employed a YOLOv8-DeepSORT-based method for estimating the traffic flow parameters with improved recognition consistency between consecutive frames [30]. On the other hand, some vehicle detection and tracking systems using YOLOv8 were effective in a controlled surveillance environment [31]. An integrated YOLOv8/v9 model with the ByteTrack and CLAHE (image enhancement technique) was used to evaluate the counting and tracking performance under varying environmental and traffic situations. The results demonstrated effective improvement in model performance, especially under poor lighting conditions [32].

The availability of image datasets of various traffic classes also continues to be a significant barrier in diverse traffic research. Early studies concentrated mostly on organized highway situations and lane-specific traffic scenarios. A range of benchmark datasets, including PASCAL VOC [33], MS COCO [34], KITTI [35], Cityscapes [36], and Open Images [37], has enabled researchers to attain substantial gains in detection accuracy. Nonetheless, these datasets mostly reflect traffic circumstances in developed nations and may not adequately depict the intricacies of diverse traffic scenarios often encountered in developing countries. Paranjape and Naik presented the DATS_2022 dataset, which highlights 45 classes including animal, board, poles, and 12 vehicle classes, but its applicability for traffic analysis is not reported [38]. This emphasizes that there is a need for a realistic dataset for unstructured traffic conditions with appropriate size, class variety, and practical applicability.

The literature studies indicate that most of the YOLOv8-based traffic monitoring works focused on general and structured traffic datasets, failing to deal with the complexities encountered in developing countries [15, 21]. These studies mostly concentrated on very few motorized traffic classes, while ignoring traffic categories like pedestrians, bicycles, and other informal modes like rickshaws, animal-driven carts, etc. [20, 22]. Very limited research on real-time multi-class vehicle detection and tracking is reported for Indian traffic conditions. The comprehensive fusion of tracking of the involved traffic objects and their class-wise vehicle counting was missing, which restricted their use in overall traffic analysis [16, 31]. Architectural developments of YOLOv8 have increasingly attracted more attention over time, but limited research has been conducted to evaluate its reliability under real-world traffic environments. Therefore, the use of models is largely confined to region-specific datasets [21, 22, 30].

III. Methodology

This research primarily offers an inclusive vision-based paradigm for the real-time traffic detection and counting system under heterogeneous traffic conditions. The workflow of the developed methodology is depicted in Fig. 1, illustrating the chronological phases from custom traffic dataset preparation to performance evaluation of the HETRA framework. The primary phases of this approach are creating a custom traffic dataset, annotating all instances, training the YOLOv8 model, evaluating the model performance, and integrating with DeepSORT for tracking multiple traffic classes and their directional counting across consecutive frames.

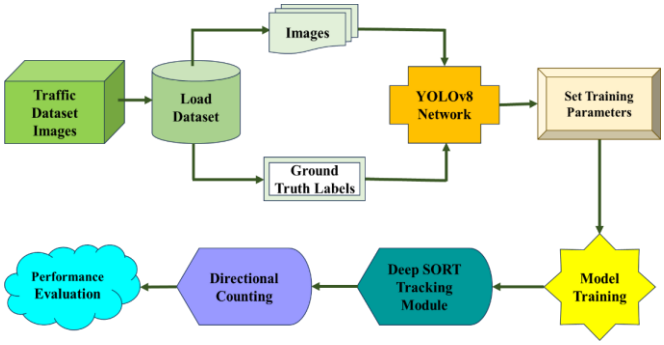


Fig. 1. Overall architectural workflow of proposed HETRA framework.

A. Dataset Preparation

The dataset was developed using actual traffic videos obtained from various urban and rural roadways in India. The dataset is intentionally enriched with various perspectives such as lighting conditions, traffic densities, background intricacy, etc. The dataset comprises fourteen traffic categories- namely auto, bike, bicycle, bus, car, horse-cart, e-rickshaw, pedestrian, pickup, rickshaw, tempo, tractor, truck, and van reflecting the diverse and complex composition of Indian road traffic.

The video footages of were chosen from different traffic conditions at regular intervals to prevent redundancy. It was furnished by removing low-quality or inappropriate image frames after visual inspection. Thereafter, high-quality image data was processed to the labelling stage. This Data was augmented by changing the original traffic images through modifications such as flipping and cropping before annotation to avoid overfitting and improve model flexibility.

Python-based 'LabelImg' annotation tool was utilized to label superior image frames manually in YOLO format (.txt). All annotations were manually inspected to verify the correctness of class labels and bounding-box localization. Incorrect or inconsistent annotations were corrected before finalizing the dataset for model training. A sample of image frames is shown in Fig. 2.

Fig. 2. Sample illustration of the annotation process.

Extra care was taken in labelling the bounding boxes of the concealed, overlapping, and smaller objects, which are more effective in practical scenarios for detection. Meticulous attention was paid while drawing the bounding boxes around the

traffic classes. Each label file contained the class index (numeric identifier referring to fourteen traffic categories) and box coordinates ('center_x', 'center_y', width and height) in standardized format.

A custom traffic dataset of 8476 frames consisting of 22886 instances was thus created, which corresponds to an average of 2.7 instances per frame. The diversified instance distribution of different traffic classes was analyzed to find the less represented traffic class, as presented in Fig. 3. No artificial data was generated to keep realistic traffic characteristics, while some horizontal image flipping was applied to minor categories. The common traffic classes such as cars, bikes, trucks, pedestrians, and pickups have a significantly higher number of instances compared to the less frequent classes like van, horse-cart, and rickshaw. The class 'Car' consists of the highest number of instances (6091), whereas less common classes like 'Horse-cart' and 'Rickshaw' consist of 127 and 132 instances, respectively. This exhibits an inherent class imbalance ratio of around 48:1, which is realistic for traffic conditions encountered on Indian roads. The class-wise performance metrics were reported to evaluate the influence of class imbalance. To enhance model generalization and mitigate overfitting, the dataset was partitioned into training, validation, and testing subsets.

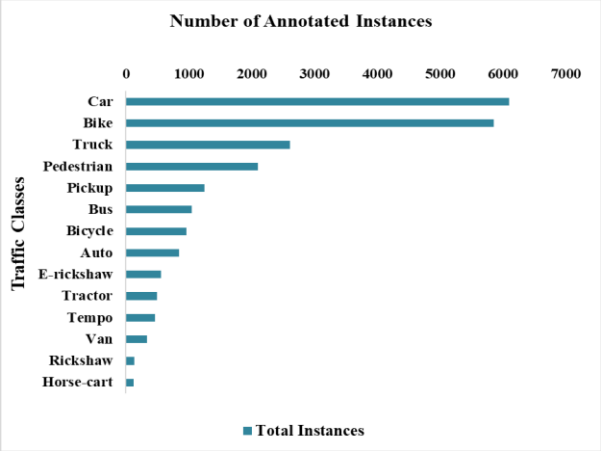


Fig. 3. Instance distribution of annotated traffic classes used in the custom dataset.

B. YOLOv8 Model Architecture

Fig. 4 shows the simplified YOLOv8 network used in this study. Improved feature extraction, anchor-free detection, and enhanced localization accuracy with faster inference performance are the main characteristics of YOLOv8, which makes it suitable for implementation under heterogeneous traffic conditions. Scalable pre-trained models from YOLOv8 can be fine-tuned on custom datasets, and this versatility allows strong learning of multiple traffic classes and complicated visual patterns (visually similar traffic classes, overlapping etc.).

The YOLOv8 architecture consists of three basic components: the backbone, neck, and detection head network. Each of them contributes to overall detection performance. The backbone network consists of Cross Stage Partial networks (CSP-net) based on two convolution (C2f) modules along with a Spatial Pyramid Pooling Fast (SPPF) block [39]. It extracts intricate low- as well as high-level semantic characteristics from input picture frames through a series of convolutional operations and feature fusion layers. This ensures better extraction of feature representation for small and partially occluded classes.

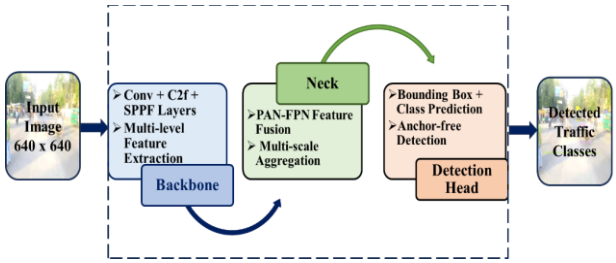


Fig. 4. YOLOv8 architecture illustrating the backbone, neck, and detection head for multi-class traffic detection.

The neck network employs a Path Aggregation Network combined with a Feature Pyramid Network (PAN-FPN) to integrate feature representations extracted at multiple scales. It is an intermediate processing part positioned between the backbone and the detection head. It is responsible for the aggregation of feature maps produced at the different backbone stages. This multi-scale feature integration facilitates efficient detection of traffic classes at varying distances and scales within the frame, thereby improving detection reliability in congested traffic conditions.

At the end, an anchor-free detection head is utilized that simultaneously predicts the bounding box coordinates, confidence scores, and class probabilities. It makes the model more robust to scale and aspect ratios. The anchor-free approach diminishes the computational intricacy and enhances the detection performance for irregular-shaped and small objects. This improves the overall efficiency of the model under dense and mixed traffic movements.

C. Multi-Class Tracking and Counting

The YOLOv8 model detects multi-class vehicles efficiently, but their identity preservation across successive video frame poses still an issue as traffic keeps changing. Therefore, the trained detector model is integrated with the DeepSORT algorithm to enable continuous multi-class tracking. Based on detection results, the DeepSORT technique provides a distinct identity to each tracked class. A Kalman filter is utilized for predicting future locations based on the previous position and movement of each traffic class track over various video frames. New detections are used to update and correct these predictions. Linking detections to existing class tracks is done using the Hungarian algorithm and deep appearance matching [27].

The hyperparameters within DeepSORT govern the timing and trajectory prediction of the traffic participants. Some of the tracking parameters are tuned specifically for congested traffic scenarios. The hyperparameter matrix used in the configuration file is summarized in Table I. Non-Maximum Suppression (NMS) was utilized to remove duplicate overlapping detections and preserve the bounding box with the highest level of confidence for each object. In addition to this, Intersection-over-Union (IoU) was applied in the process of data association while applying DeepSORT to correlate new detections with old tracks based on temporal overlap. Furthermore, frequent occlusions of small-sized vehicle classes (auto, bike, bicycle, etc.) while

moving adjacent to and behind the large vehicle class (bus, truck etc.) are noticed especially in mixed traffic conditions. Therefore, the MAX_AGE hyperparameter becomes an important aspect, especially in traffic monitoring applications. In this study, MAX_AGE was empirically set to 70 frames based on preliminary investigation and guidance from previous DeepSORT based tracking studies. This value provided a practical balance between the identities over short occlusions and prevented retention of prolonged tracking. This reduced the risk of double counting, while limiting the surplus computational overhead.

TABLE I. DEEPSORT TRACKING PARAMETERS

Hyperparameters	Operational Value	Function in Traffic Environments
MAX_DIST	0.2	Establish upper limit for cosine distance for visual appearance matching, helps to maintain correct traffic class identities
MIN_CONFIDENCE	0.3	Deny all class detections having a confidence score below 30%, reducing false positive detections
NMS_MAX_OVERLAP	0.5	Non-Maximum Suppression (NMS) level to eliminate duplicate bounding boxes for a single class with more than 50%
MAX_IOU_DISTANCE	0.7	Maximum Intersection over Union (IoU) distance allowed for overlapping bounding boxes
MAX_AGE	70	Retains class track identity up to 70 successive frames even in partial obstruction and missed detection
N_INIT	3	At least 3 successive stable detection frames required before allocating a tracking ID to reduce false track ID
NN_BUDGET	100	Stores 100 most recent frames of the class appearance feature

During the tracking phase, each detected traffic class is allocated a unique identity number which is maintained over successive video frames. Due to this consistent identification, it is possible to keep track of each traffic category in moving scenes and ensure non-repetitiveness of identity for the same object. This ensures accurate computation of traffic volume count, while minimizing the duplicity in detection and identity conflict issues.

Traffic counting is implemented through the use of a virtual counting line (VCL), passing through a defined point within the video frame. This line acts as a reference for counting the traffic. A trajectory buffer with a deque length of 20 was used for each tracked class to facilitate directional counting. The stored centroid history was utilized to analyse the movement and direction with respect to VCL. The bottom center of bounding box of each class continuously observed and counted traffic class is generated as it crosses the VCL. This minimizes the duplicate counts caused by repeated detections in consecutive frames. Traffic categories crossing the VCL while moving in the downward direction in a frame are reported as 'Number of Vehicles Entering,' and those crossing the VCL while moving in the upward direction are reported as 'Number of Vehicles Leaving'. Class-wise counters are established for each direction separately. This facilitates precise class-wise estimation of directional traffic count in real-world traffic scenarios.

D. Training Strategy and Loss Functions

In this study, YOLOv8 version 'l', was chosen as the traffic detection model because it offers a practical balance between detection performance and computational overhead. The model was fine-tuned using the custom traffic dataset consisting of a total of 8476 manually annotated image frames representing fourteen different traffic participants. The dataset was divided into training, validation, and testing subsets in the ratio of 66%, 14%, and 20%, respectively. Transfer learning is employed by initializing pretrained YOLOv8l weights with an input image resolution of 640 x 640 during training. The model was trained and optimized with loss components such as box, classification, and Distribution Focal Loss (DFL) over 200 epochs. The weights corresponding to the best performance were selected for subsequent evaluations. The major training parameters along with implementation aspects are summarized in Table II.

TABLE II. SUMMARY OF TRAINING SETUP

Parameter	Description
Model	YOLOv8l
Dataset Size	8476 images
Number of Classes	14
Input Image Size	640 × 640
Training Epochs	200
Batch Size	16
Optimizer	Stochastic Gradient Descent (SGD)
Learning Rate	0.01
Platform	Google Colab
GPU	NVIDIA A100 SXM4-40GB
Loss Function	Composite Loss (CIoU + BCE + DFL)

E. Loss Functions

The composite loss function was used to train the model. This minimizes the loss for localization accuracy and classification performance, and enhances the bounding box precision. Eq. (1) shows the total loss function (L_{TOTAL}) [12, 40, 41].

$$L_{TOTAL} = \lambda_{BOX} L_{BOX} + \lambda_{CLS} L_{CLS} + \lambda_{DFL} L_{DFL} \quad (1)$$

where, L_{BOX}, $L_{CLS,}$ and L_{DFL} symbolize the loss for bounding box regression, classification, and distribution focal loss, respectively. The λ_{BOX}, $\lambda_{CLS,}$ and λ_{DFL} are corresponding loss weightage coefficients.

To enhance localization precision, bounding box regression loss employs the Complete Intersection over Union (CIoU) metric. This loss accounts for overlapping area, center distance, and the aspect ratio uniformity between ground truth and predictions as shown in Eq. (2) [40].

$$L_{BOX} = L_{CIoU} = 1 - IoU + \frac{\rho^2(b,b^{gt})}{d^2} + \beta\upsilon \quad (2)$$

where, ρ (b, b^{gt}) represents the Euclidean distance between the centers of the predicted and the ground-truth bounding box, 'd' is the diagonal length of the smallest enclosing box, and 'υ' is aspect ratio uniformity and 'β' is a weighting parameter that controls the influence of the aspect ratio term in the CIoU loss function.

The classification loss 'L_{CLS}' is computed using Binary Cross-Entropy (BCE) loss as expressed in Eq. (3)[12]to penalize incorrect class predictions.

$$L_{CLS} = L_{BCE} = -\sum_{i=1}^{C=14}[y_i\, log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)] \quad (3)$$

Where C = 14 is the total number of traffic classes, y_i represents the ground truth class label, and $\hat{y}_i$ predicted probability for the i^{th} class. This loss function accounts for the error between predicted and actual labels, thereby making it well-suited for multi-class detection problems.

DFL is utilized to refine bounding box localization and improve detection accuracy. This is especially important for small and partially occluded traffic objects prevalent in heterogeneous traffic. The mathematical formulation of DFL is given in Eq. (4) [41].

$$L_{DFL} = \sum_{k=1}^{4}\sum_{i=0}^{n}[p_i \log(\hat{p}_i)] \quad (4)$$

where, pi and $\hat{p}_i$ are the actual and predicted discrete probability distributions for each bounding box coordinate.

The integration of these loss functions strengthens the detector model in detecting and localizing the classes of different shapes and sizes.

F. Evaluation Metrics

The performance of the detector model was assessed by using standard evaluation metrics. These are: Precision, Recall, F1-score, as well as mean Average Precision (mAP). These parameters were calculated using the confusion matrix elements, including True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN), to assess the effectiveness of the model.

Precision represents the ratio of accurately detected to the total number of detected objects, which is defined as:

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

Recall measures the ratio of correctly detected to the total number of ground truth objects and is given by:

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

The F1-score offers a fair assessment of accuracy and recall that is calculated as

$$F1_Score = \frac{2*Precision*Recall}{Precision+Recall} \quad (7)$$

The area under the Precision-Recall (P-R) curve represents Average Precision (AP) for the given class. The P-R curve illustrates the balance that exists between detection precision and comprehensiveness at different confidence levels. It is calculated by integrating precision with respect to recall over the interval [0,1].

$$AP = \int_0^1 Precision(Recall)d(Recall) \quad (8)$$

The mAP is derived by averaging AP over all 'N' traffic classes (here N=14).

$$mAP = \frac{1}{(N=14)}\sum_{i=1}^{N=14} AP_i \quad (9)$$

The mAP@0.5 refers to mean Average Precision evaluated at a specific threshold IoU value of 0.5, whereas mAP@0.5:0.95 is a metric that averages the mAP values for various IoU thresholds ranging from 0.5 to 0.95 at intervals of 0.05. It provides a rigorous and thorough assessment for testing the accuracy of object detection and localization at various thresholds.

IV. Results and Discussion

This section presents a comprehensive evaluation and analysis of the HETRA framework for real-time detection, tracking, and directional counting of multiple-class traffic. The effectiveness of the trained detector was assessed using universal evaluation metrics such as precision, recall, mean Average Precision (mAP), F1 measure, as well as analysis of the confusion matrix. Additionally, to assess robustness and real-world applicability, the proposed system was analysed using multiple recorded traffic videos with different levels of traffic density and lighting conditions.

A. Training Performance Analysis

Fig. 5 illustrates the comprehensive performance of the detector model over 200 epochs. The graph shows progressive reduction in box loss, classification loss, and DFL throughout the training process.

The training curves indicate stable convergence, with no significant indications of instability. Consequently, model generalization was assessed through validation metrics, such as precision, recall, and mean Average Precision (mAP), which is consistent with recent YOLOv8-based object detection studies. It may be noted that the model was validated throughout the training process, but the val/cls_loss recorded by YOLOv8 remained zero and therefore did not produce a visible curve in Fig. 5. As the training epoch increases, a steady improvement in precision and mAP values was observed, while moderate fluctuation was noticed in recall. This demonstrates efficient feature learning and generalization capability of the detector model.

B. Detection and Classification Performance Evaluation

The precision–confidence, recall–confidence, and F1-confidence curves obtained during model evaluation are illustrated in Fig. 6. Gradual improvement in the precision curve at higher confidence values was observed due to a reduction in false positives. On the other hand, the recall value decreased at a confidence threshold due to some true values being missed during confined detection. The F1-confidence curve achieved a peak F1-score of 0.63 at a confidence threshold of 0.33. This shows an optimal balance between precision and recall at a moderate threshold.

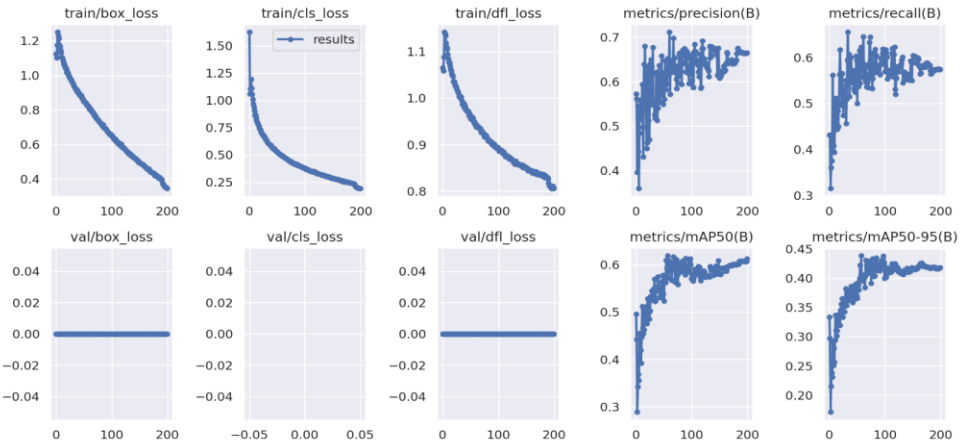


Fig. 5. Comprehensive training performance assessment of the detector model.

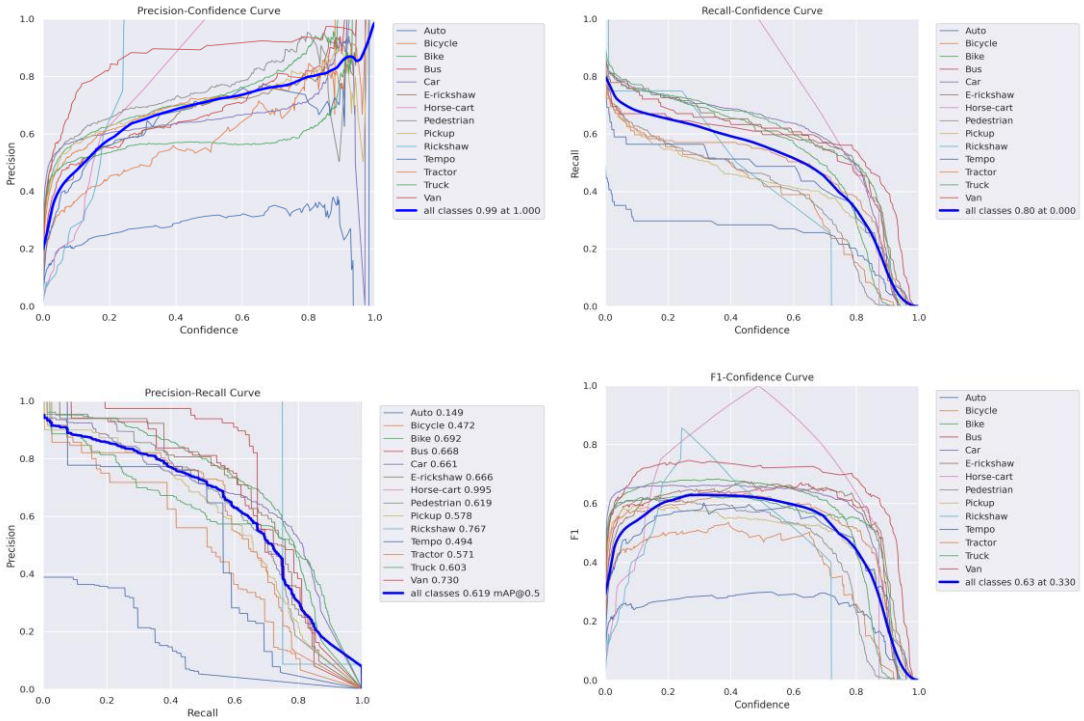


Fig. 6. Precision-Confidence, Recall-Confidence, Precision-Recall, and F1-Confidence curves of the YOLOv8l model for 14 different traffic classes.

The precision-recall curve indicated stable detection performance for major and visually distinguishable classes such as cars, buses, pickups, and trucks. Comparatively lower precision and recall values were observed for less- frequent classes including horse-cart and rickshaw, primarily due to the limited number of training instances available for these categories. Although the auto and bicycle classes contained a substantial number of instances, their detection performance also remained comparatively lower. This suggests that factors beyond class imbalance, including smaller object sizes, repeated occlusions, and high intra-class variability, significantly affected the detection performance under heterogeneous traffic conditions.

The trained model achieved precision, recall, mAP@0.5, and mAP@0.5:0.95 values of 71.15%, 65.52%, 61.9%, and 43.85%, respectively. These results indicate efficient detection and localization capability for various traffic participants. Whereas a comparatively lower performance was noticed for certain class may be due to less representation, small size, along with partial occlusions under dense traffic situations. The complete detection performance of the model can be attributed to multi-scale feature extraction and the diversity of the custom traffic dataset. This enabled the model to learn more vigorous features even in varying traffic densities, weak lane discipline, and frequent occlusions.

Although the performance curves provided a comprehensive assessment, the normalized confusion matrix offers detailed insights into classification behavior across different traffic classes, as presented in Fig. 7. There was considerable diagonal dominance for the most classes of traffic, indicating reliable

multi-class detection and classification. Classes such as bike, bus, car, horse-cart, e-rickshaw, and truck achieved higher classification accuracy due to their visuals and adequate representation within the traffic dataset. Comparatively lower accuracy was observed for classes like auto and rickshaw, which may be due to their visual similarity and small scale. Some discrepancies were also observed between visually similar categories, especially rickshaw vs. e-rickshaw as well as van vs. car. In addition, moderate confusion with the background was also noticed for small and partially hidden traffic participants. Despite these challenges, the confusion matrix illustrates the robustness and effectiveness of the model for detection and localization of a wide range of traffic participants.

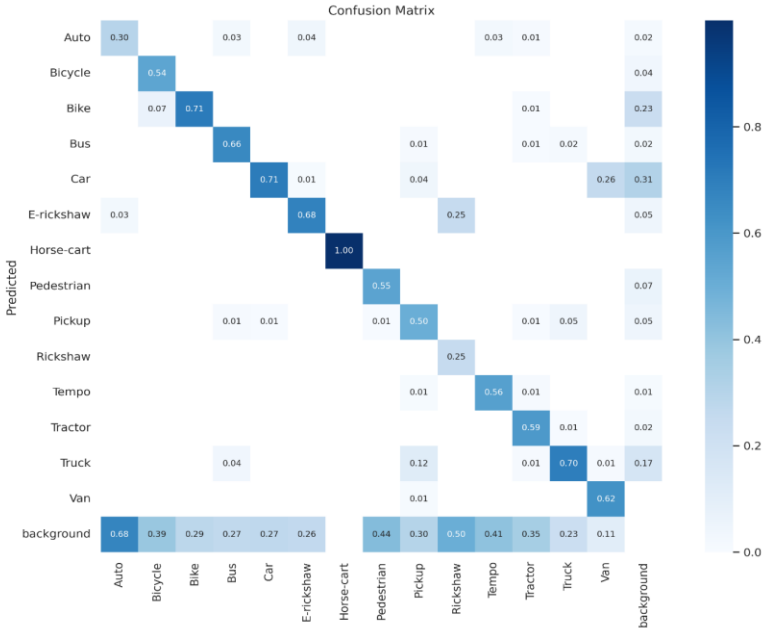


Fig. 7. Normalized confusion matrix obtained for custom dataset.

Fig. 8. Detection, localization and multi-class tracking results of the HETRA model under real-world traffic conditions.

C. Qualitative Analysis

The real-world traffic videos, each of approximately 5 minutes' duration, were recorded at the Teliyarganj Intersection in Prayagraj city and were then analyzed using the developed HETRA model. Fig. 8 illustrates a few sample frames representing the qualitative performance of the model under different lighting and traffic conditions. The model was evaluated under normal weather conditions. The visual outcomes indicate that multiple traffic classes were detected and localized simultaneously for large vehicles like buses and trucks, as well as smaller road users like bicycles and pedestrians.

The visual outcomes also show a substantial discrimination between visually similar traffic classes such as 'bicycles & bikes' and 'car & van'. This results in fewer identity swaps among various traffic categories. Furthermore, the model maintained reliable detection performance across multiple traffic classes at various distances and scales within the surveillance frames.

The successive sample frames also indicate the tracking results of various traffic participants. These demonstrate consistent preservation of class ID even under continuous moving and overlapping traffic flow. As shown in Fig. 8, the tracking sequences of car (IDs 105 & 126) crossed each other with their unique IDs. Also, small-class traffic like a bike (ID 127), moving in proximity to the car, preserved its consistent trajectory in the successive frames. It is also observed that the vulnerable road user (pedestrian ID 1116) retained its identity while crossing the road near the car (IDs 961&1122). These findings validate the suitability of the HETRA model for its sustained tracking capability under various traffic participants (from public transit mode bus to low mobility modes such as bicycles, e-rickshaws, as well as pedestrians) in mixed traffic conditions.

Some intermittent miss-detections and slight localization lapses were also observed by the model for small traffic classes under occluded conditions. However, overall qualitative performance of the model is found to be adequate for real-time traffic monitoring and applications. The visual outcomes validate the efficacy and practical utility of the DeepSORT-based tracker for multiclass traffic analysis under heterogeneous traffic conditions.

The counting performance was evaluated using multiple representative traffic videos. Unlike the conventional method, this study focuses on accomplishing bi-directional traffic flow, while maintaining separate real-time records for entering and leaving traffic movements. The class-wise directional traffic counting relies on the unique class ID, trajectory analysis, and a virtual counting line that works as a reference boundary within video frames. Representative example frames of class-wise directional traffic counting results are shown in Fig. 9.

As evident from Fig. 9, the traffic class is counted only once, and it is done when the traced traffic trajectory crosses the virtual counting line in the specific direction. This eliminates the duplicate counting of vehicles across successive frames and yields the desired class-wise traffic counting. This track-assisted counting approach facilitates precise estimation of directional traffic volume. The results also indicate that the developed approach supports the classification and counting of region-specific classes like auto and tempos alongside standard traffic classes.

The counting statistics were validated using multiple real-world traffic video recordings. Manual counts for both entering and leaving directions were acquired by visual inspection and thereafter compared with the automated counts obtained by HETRA. The traffic categories, namely horse-cart, tractor, and rickshaw, were not observed in the selected videos due to their absence in the recorded videos. Therefore, the counting results for these classes could not be evaluated. The directional traffic counting results corresponding to representative traffic phases are presented in Table III. The directional counting accuracy ranged from 80% to 96% across different traffic classes. An aggregate 1320 traffic classes were manually counted in cross-sample videos, whereas a total of 1181 participants estimated by HETRA achieved an overall counting accuracy of approximately 90%.

TABLE III. CLASS-WISE DIRECTIONAL TRAFFIC COUNTING PERFORMANCE OF THE MODEL

Traffic Class	Entry Count		Exit Count	
	Model	Manual	Model	Manual
Auto	72	84	83	96
Bicycle	39	46	54	66
Bike	141	152	167	179
Bus	45	50	34	38
Car	127	132	105	113
E-rickshaw	27	32	41	47
Pedestrian	33	37	24	27
Pickup	28	33	34	42
Tempo	40	47	36	41
Truck	7	8	10	11
Van	19	21	15	18
Total	**578**	**642**	**603**	**678**

Most recent studies primarily reported the detection accuracy and counting performance, whereas their computational efficiency was mostly ignored. The inference performance of the developed model was tested on an average compute in Google Colab using an NVIDIA T4 GPU with 16 GB memory. The mean preprocessing, inference, and post-processing times were 0.5 ms, 20.2 ms, and 2.2 ms per frame, respectively. This results in an overall processing time of approximately 23ms per frame, nearly 44 FPS (frames per second). Since the evaluation videos were recorded at 30FPS, the achieved processing speed is higher than the recording rate. Although GPU memory utilization was not explicitly profiled, the observed inference performance demonstrates the computational efficiency and practical scalability of the HETRA framework for real-time heterogeneous traffic analysis.

Fig. 9. Directionally classified traffic volume by the HETRA model in real time under heterogeneous traffic conditions.

Minor counting errors were noticed in rare circumstances, including highly congested conditions, abrupt direction changes, and intermittent ID changes, near the counting line. However, these remain limited and had a minute effect on the overall traffic counting results. While a slight reduction in accuracy was observed in relatively low light conditions, the overall results demonstrate the effectiveness of the model under different lighting conditions. Less frequent traffic categories like horse-carts, rickshaws, and tractors were not observed during representative counting assessment and so could not be included in the analysis.

Direct comparison of the reported performances should be interpreted with caution because of variation in dataset, class compositions, environmental conditions and evaluation methodologies. A concise comparative analysis with selected relevant studies is presented in Table IV. It can be observed from Table IV that most of the existing studies reported higher detection and counting performance in relatively structured traffic streams involving three or four vehicle classes only. In contrast, the HETRA system was tested under more challenging heterogeneous traffic conditions comprising fourteen traffic categories, including region-specific classes such as auto, tempo, horse-cart, and rickshaw. Therefore, the lower mAP@0.5 achieved by the HETRA framework is mainly due to the complexity associated with multi-class detection under dense, frequent occlusions and weak-lane discipline and class imbalance issues. Instead of optimizing for better performance for a limited number of traffic categories, the proposed framework was designed to offer robust multi-class detection, tracking, and directional counting under varying illumination levels, frequent occlusions, and weak lane discipline.

TABLE IV. COMPARATIVE ANALYSIS WITH EXISTING TRAFFIC MONITORING STUDIES

Study	Dataset	No. of classes	Precision (%)	Recall (%)	mAp@0.5 (%)	Counting Accuracy (%)
Duan et.al.[20]	VisDrone2019	10	60.29	46.34	49.64	-
Song et.al.[28]	Highway Scene	3	88.0	89.0	87.88	93.2
Sathiyaprasadet.al.[30]	Urban	4	96.8	96.5	96.3	90
Akoushideh et.al.[32]	Extracted from MSCOCO	3	98.0	97.0	98.0	88.84
Proposed HETRA framework	**Custom**	**14**	**71.15**	**65.51**	**61.9**	**90**

Overall, the results of the HETRA model indicate its capability for multi-class detection, tracking, and class-wise directional counting with commendable accuracy under both normal and relatively low light conditions. This emphasizes the model's potential for real-time traffic monitoring, class-wise traffic movement analysis, and other ITS applications.

V. CONCLUSION

The proposed HETRA, integrating YOLOv8 and DeepSORT, was developed and evaluated for real-time detection, tracking, and directional counting under heterogeneous Indian traffic circumstances. The developed system effectively handled mixed traffic streams consisting of motorized traffic classes, non-motorized traffic classes, and pedestrians by addressing the complexity of urban traffic in developing countries.

A custom annotated traffic dataset consisting of fourteen classes was created to train and evaluate the model under typical urban conditions, including congestion, occlusions, and unstructured traffic movement. The experimental results confirmed that the suggested approach can reliably detect, track, and directionally count the traffic participants, making it suitable for real-time traffic monitoring. The findings further demonstrated the model's robustness for congested traffic, poor lane discipline, and mixed traffic under varying light conditions. Importantly, the detection and counting of slow-moving and vulnerable road-user categories like bicycles and pedestrians alongside motorized traffic highlight its potential for comprehensive traffic surveillance.

Apart from its detection and counting performance, the proposed framework offers a solid foundation for the development of an AI-enabled intelligent transportation system. The classified traffic volume information, generated by the proposed methodology, can help with modal share analysis, congestion monitoring, road safety assessment, and adaptive traffic signal control. Consequently, the system offers a scalable and cost-effective solution for real-time traffic monitoring and future smart city traffic management, especially in developing countries.

Despite the promising performance, the developed system experiences some challenges under severe occlusions and low light conditions. Comparatively lower performance was noticed for some less-frequent traffic classes, which may be attributed to class imbalance. The current study is limited to normal weather conditions, but the performance of the developed model needs to be tested further under adverse weather conditions.

Future research will focus on enriching the custom dataset through the inclusion of more nighttime and adverse weather traffic images, investigating the sensitivity of tracking hyperparameters, and integrating the framework with signal control and edge computing platforms for large-scale smart city deployment.

Data Availability: The custom dataset used in this study is currently not publicly available because it forms part of the author's ongoing research work. The dataset will be made available from the corresponding author upon reasonable request after completion of the research.

REFERENCES

[1] E. Dilek and M. Dener, "Computer Vision Applications in Intelligent Transportation Systems: A Survey," Sensors, vol. 23, no. 6, 2023, doi: 10.3390/s23062938.

[2] H. Shokravi, H. Shokravi, N. Bakhary, M. Heidarrezaei, S. S. R. Koloor, and M. Petrů, "A review on vehicle classification and potential use of smart vehicle-assisted techniques," Sensors (Switzerland), vol. 20, no. 11, pp. 1–29, 2020, doi: 10.3390/s20113274.

[3] N. A. Mandellos, I. Keramitsoglou, and C. T. Kiranoudis, "Expert Systems with Applications A background subtraction algorithm for detecting and tracking vehicles," Expert Syst. Appl., vol. 38, no. 3, pp. 1619–1631, 2011, doi: 10.1016/j.eswa.2010.07.083.

[4] X. Cao, C. Wu, P. Yan, and X. Li, "Linear SVM classification using boosting HOG features for vehicle detection in low-altitude airborne videos," Proc. - Int. Conf. Image Process. ICIP, pp. 2421–2424, 2011, doi: 10.1109/ICIP.2011.6116132.

[5] T. Azfar, J. Li, H. Yu, R. L. Cheu, Y. Lv, and R. Ke, "Deep Learning-Based Computer Vision Methods for Complex Traffic Environments Perception: A Review," Data Sci. Transp., vol. 6, no. 1, pp. 1–19, 2024, doi: 10.1007/s42421-023-00086-7.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 580–587, 2014, doi: 10.1109/CVPR.2014.81.

[7] R. Girshick, "Fast R-CNN," Proc. IEEE Int. Conf. Comput. Vis., vol. 2015 Inter, pp. 1440–1448, 2015, doi: 10.1109/ICCV.2015.169.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, 2017, doi: 10.1109/TPAMI.2016.2577031.

[9] W. Liu et al., "SSD: Single Shot MultiBox Detector," ECCV, vol. 1, pp. 21–37, 2016, doi: 10.1007/978-3-319-46448-0_2.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016-Decem, pp. 779–788, 2016, doi: 10.1109/CVPR.2016.91.

[11] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, vol. 2017-Janua, pp. 6517–6525, 2017, doi: 10.1109/CVPR.2017.690.

[12] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," 2018, [Online]. Available: http://arxiv.org/abs/1804.02767

[13] C. Wang and H. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection".

[14] A. Farid, F. Hussain, K. Khan, M. Shahzad, U. Khan, and Z. Mahmood, "A Fast and Accurate Real-Time Vehicle Detection Method Using Deep Learning for Unconstrained Environments," Appl. Sci., vol. 13, no. 5, 2023, doi: 10.3390/app13053059.

[15] T. Vu, H. N. Thai, V. N. Pham, H. T. Vu, A. T. Luong, and T. Van Luong, "Counting Mixed Traffic Volumes at Motorcycle-Dominated Intersections by Using Computer Vision," Int. J. Intell. Transp. Syst. Res., vol. 23, no. 1, pp. 146–164, 2025, doi: 10.1007/s13177-024-00442-z.

[16] M. Safaldin, N. Zaghden, and M. Mejdoub, "An Improved YOLOv8 to Detect Moving Objects," IEEE Access, vol. 12, no. May, pp. 59782–59806, 2024, doi: 10.1109/ACCESS.2024.3393835.

[17] J. Terven, D. M. Córdova-Esparza, and J. A. Romero-González, "A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS," Mach. Learn. Knowl. Extr., vol. 5, no. 4, pp. 1680–1716, 2023, doi: 10.3390/make5040083.

[18] A. Vijayakumar and S. Vairavasundaram, YOLO-based Object Detection Models: A Review and its Applications, vol. 83, no. 35. Springer US, 2024. doi: 10.1007/s11042-024-18872-y.

[19] A. A. Murat and M. S. Kiran, "A comprehensive review on YOLO versions for object detection," Eng. Sci. Technol. an Int. J., vol. 70, no. June, 2025, doi: 10.1016/j.jestch.2025.102161.

[20] X. Duan, S. Xue, Z. Zhou, and X. Duan, "Design and Implementation of Road Vehicle Counting Model Based on YOLOv8," Proc. 2024 8th Int. Conf. Electron. Inf. Technol. Comput. Eng. EITCE 2024, pp. 844–848, 2025, doi: 10.1145/3711129.3711275.

[21] W. Ai, J. Zeng, Z. Lei, and D. Li, "Traffic flow parameter estimation for complex environments based on improved YOLOv8," Proc. Inst. Mech. Eng. Part D J. Automob. Eng., vol. 239, no. 12, pp. 5412–5426, 2025, doi: 10.1177/09544070241287582.

[22] M. Bakirci, "Utilizing YOLOv8 for enhanced traffic monitoring in intelligent transportation systems (ITS) applications," Digit. Signal Process. A Rev. J., vol. 152, no. May, p. 104594, 2024, doi: 10.1016/j.dsp.2024.104594.

[23] B. Ji, J. Xu, Y. Liu, P. Fan, and M. Wang, "Improved YOLOv8 for small traffic sign detection under complex environmental conditions," Franklin Open, vol. 8, no. October, p. 100167, 2024, doi: 10.1016/j.fraope.2024.100167.

[24] Iqra and K. J. Giri, "SO-YOLOv8: A novel deep learning-based approach for small object detection with YOLO beyond COCO," Expert Syst. Appl., vol. 280, no. March, p. 127447, 2025, doi: 10.1016/j.eswa.2025.127447.

[25] H. Li, Y. Li, L. Xiao, Y. Zhang, L. Cao, and D. Wu, "RLRD-YOLO: An Improved YOLOv8 Algorithm for Small Object Detection from an Unmanned Aerial Vehicle (UAV) Perspective," Drones, vol. 9, no. 4, pp. 1–21, 2025, doi: 10.3390/drones9040293.

[26] M. Bakirci, "Advanced aerial monitoring and vehicle classification for intelligent transportation systems with YOLOv8 variants," J. Netw. Comput. Appl., vol. 237, no. November 2023, p. 104134, 2025, doi: 10.1016/j.jnca.2025.104134.

[27] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," Proc. - Int. Conf. Image Process. ICIP, vol. 2017-Septe, pp. 3645–3649, 2017, doi: 10.1109/ICIP.2017.8296962.

[28] H. Song, H. Liang, H. Li, Z. Dai, and X. Yun, "Vision-based vehicle detection and counting system using deep learning in highway scenes," Eur. Transp. Res. Rev., vol. 11, no. 1, 2019, doi: 10.1186/s12544-019-0390-4.

[29] M. Durve, S. Orsini, A. Tiribocchi, A. Montessori, and J. Tucny, "Benchmarking YOLOv5 and YOLOv7 models with DeepSORT for droplet tracking applications," Eur. Phys. J. E, vol. 46, no. 5, pp. 1–7, 2023, doi: 10.1140/epje/s10189-023-00290-x.

[30] B. Sathiyaprasad, P. A. R. Reddy, and D. M. Venkatesh, "TrackNCount - Intelligent Vehicle Tracking, Counting, and Speed Estimation using YOLOv8 and DeepSORT Algorithms," 6th Int. Conf. Mob. Comput. Sustain. Informatics, ICMCSI 2025 - Proc., pp. 1738–1744, 2025, doi: 10.1109/ICMCSI64620.2025.10883110.

[31] M. R. Sholahuddin et al., "Optimizing YOLOv8 for Real-Time CCTV Surveillance: A Trade-off Between Speed and Accuracy," J. Online Inform., vol. 8, no. 2, pp. 261–270, 2023, doi: 10.15575/join.v8i2.1196.

[32] A. Akoushideh, S. S. Sadat, and A. Shahbahrami, "Counting vehicles types using deep learning algorithm in video surveillance systems," Multimed. Tools Appl., vol. 84, no. 27, pp. 33017–33040, 2025, doi: 10.1007/s11042-024-20530-2.

[33] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," Int. J. Comput. Vis., vol. 88, no. 2, pp. 303–338, 2010, doi: 10.1007/s11263-009-0275-4.

[34] G. T. U. A. Colleges et al., "Microsoft COCO," Eccv, no. June, pp. 740–755, 2014.

[35] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," Int. J. Rob. Res., vol. 32, no. 11, pp. 1231–1237, 2013, doi: 10.1177/0278364913491297.

[36] M. Cordts et al., "The Cityscapes Dataset for Semantic Urban Scene Understanding," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016-Decem, pp. 3213–3223, 2016, doi: 10.1109/CVPR.2016.350.

[37] A. Kuznetsova et al., "The Open Images Dataset V4: Unified Image Classification, Object Detection, and Visual Relationship Detection at Scale," Int. J. Comput. Vis., vol. 128, no. 7, pp. 1956–1981, 2020, doi: 10.1007/s11263-020-01316-z.

[38] B. A. Paranjape and A. A. Naik, "DATS_2022: A versatile indian dataset for object detection in unstructured traffic conditions," Data Br., vol. 43, pp. 0–6, 2022, doi: 10.1016/j.dib.2022.108470.

[39] C. Y. Wang, H. Y. Mark Liao, Y. H. Wu, P. Y. Chen, J. W. Hsieh, and I. H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work., vol. 2020-June, pp. 1571–1580, 2020, doi: 10.1109/CVPRW50498.2020.00203.

[40] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," AAAI 2020 - 34th AAAI Conf. Artif. Intell., no. 2, pp. 12993–13000, 2020, doi: 10.1609/aaai.v34i07.6999.

[41] X. Li et al., "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," Adv. Neural Inf. Process. Syst., vol. 2020-Decem, pp. 1–14, 2020.

Improving Arabic Relational Aggregated Search Using a Hybrid RAG and AraBERT Framework

Sara Saad Sedhom[1], Ahmed Younes[2], Islam Elkabani[3], Ashraf Elsayed[4]
Faculty of Science-Department of Mathematics and Computer Science, Alexandria University, Alexandria 21511, Egypt[1, 2, 3, 4]
Faculty of Computer Science and Engineering, Alamein International University, Alamein 51718, Egypt[2, 4]
School of Information Technology, University of Cincinnati, Ohio, USA[3]

***Abstract*—Aggregated search is challenged by data heterogeneity, redundancy, and irrelevant information, particularly in Arabic because of its rich morphology and dialectal diversity. This study proposes an AI-driven framework to enhance Arabic aggregated search by integrating Retrieval-Augmented Generation (RAG) with semantic clustering. The framework employs RAG to retrieve relevant information from multiple search verticals, including web pages, images, news articles, and videos, and to generate contextually enriched textual representations. These representations are encoded with AraBERT to produce 768-dimensional semantic embeddings, which are then compressed into compact latent feature vectors through a stacked autoencoder. K-Means clustering is then applied to organize the compressed representations into semantically coherent topic groups. The proposed framework was evaluated on nine Arabic datasets spanning five application domains: education, sports, information technology, healthy food, and weather. Experimental results demonstrate that the proposed approach increases the average silhouette score from 0.59 to 0.68, representing a 15% improvement over clustering based on raw aggregated search results. Ablation experiments further indicate that both Retrieval-Augmented Generation and stacked autoencoder–based feature compression independently contribute to the observed performance gains. These findings demonstrate the effectiveness of integrating Retrieval-Augmented Generation with semantic representation learning to improve the organization and clustering of Arabic aggregated search results.**

***Keywords*—Relational Aggregated Search (RAS); information retrieval; Natural Language Processing (NLP); Retrieval-Augmented Generation (RAG); AraBERT embedding; feature extraction; stacked autoencoders; K-Means**

I. Introduction

Conventional search engines, such as Google and Bing, typically return ranked lists of hyperlinks, requiring users to examine multiple webpages and manually integrate the retrieved information. Although this approach is adequate for straightforward information needs, it becomes increasingly inefficient when relevant information is distributed across different content types and specialized search services.

Aggregated search addresses this limitation by combining results from multiple verticals, including web pages, images, news articles, and videos, within a unified search interface [1]. Presenting heterogeneous information in a single results page improves information accessibility and reduces the need to navigate between multiple search services [2]. Nevertheless, the retrieved results often differ considerably in structure, quality, and semantic content. Duplicate entries, fragmented snippets, irrelevant information, and limited contextual information frequently reduce the effectiveness of subsequent analysis and organization. These challenges are particularly pronounced for Arabic content, where rich morphology, orthographic variations, and dialectal diversity further complicate semantic interpretation.

Recent advances in large language models (LLMs) have significantly improved the ability of Natural Language Processing (NLP) systems to understand and generate human language. Despite their impressive capabilities, LLMs rely primarily on knowledge acquired during pretraining and therefore cannot reliably incorporate newly available or domain-specific information without access to external knowledge sources [3]–[5]. This limitation has motivated the development of Retrieval-Augmented Generation (RAG), which combines information retrieval with neural text generation. Instead of generating responses solely from model parameters, RAG retrieves relevant documents from an external knowledge base and uses the retrieved evidence to generate contextually grounded textual representations [6]–[9]. Consequently, the generated content is generally more informative, coherent, and factually supported than text produced without retrieval augmentation.

The quality of the RAG system is closely related to the semantic representations used throughout the retrieval pipeline. For Arabic, AraBERT has become one of the most effective pretrained language models because it is specifically designed to capture the linguistic characteristics of Arabic, including its rich morphology and orthographic variations [10], [11]. The resulting contextual embeddings provide meaningful semantic representations that are well suited for downstream information retrieval and clustering tasks. However, the high dimensionality of these embeddings may reduce clustering performance because of the curse of dimensionality. Stacked autoencoders address this challenge by learning compact latent representations that preserve the underlying semantic information while reducing redundancy and noise [12]. These compressed representations provide a more suitable feature space for clustering algorithms.

Among existing clustering techniques, K-Means remains an attractive choice because of its computational efficiency, scalability, and ability to partition semantically similar instances into coherent groups [13], [14]. When combined with high-quality semantic representations, K-Means provides an effective

mechanism for organizing heterogeneous search results into meaningful topical categories.

Motivated by these challenges, this study proposes a unified framework for aggregated Arabic search that integrates Retrieval-Augmented Generation, AraBERT embeddings, stacked autoencoders, and K-Means clustering into a single processing pipeline. The framework begins by collecting relational aggregated search results across multiple search verticals. The retrieved content is subsequently enriched using RAG to transform fragmented search snippets into coherent contextual representations. The generated text is then preprocessed and encoded with AraBERT, after which a stacked autoencoder learns compact semantic representations to reduce dimensionality. Finally, K-Means clustering is applied to organize the compressed representations into semantically coherent topic groups. Experimental evaluation on nine Arabic datasets covering education, sports, information technology, healthy food, and weather demonstrates that the proposed framework consistently outperforms raw aggregated search results and alternative approaches, resulting in significant improvements in clustering quality.

The remainder is organized as follows: Section II reviews related work. Section III outlines the challenges and summarizes our contributions. Section IV describes our methodology. Section V presents the experiment setup. Section VI presents results and discussion. Section VII presents the conclusion and future work.

II. Related Work

Natural Language Processing systems have traditionally relied on either retrieval-based or generative paradigms, each characterized by distinct strengths and inherent limitations. Retrieval-based approaches, including conventional search engines, efficiently identify relevant documents from large-scale collections but are limited to presenting existing information without synthesizing new knowledge [15]. In contrast, transformer-based generative models have substantially advanced natural language understanding and text generation, producing fluent and contextually coherent responses across a broad spectrum of applications [16]. Despite these capabilities, their reliance on parametric knowledge often leads to hallucinations and factual inconsistencies when external or recently acquired information is required.

Retrieval-Augmented Generation (RAG) was introduced to overcome these limitations by integrating information retrieval with neural text generation. Early retrieval-enhanced systems, such as DrQA, combined document retrieval with extractive answer generation, although their ability to synthesize information remained constrained [17]. Subsequent research demonstrated that retrieval and generation could be jointly optimized within a unified architecture. REALM demonstrated the feasibility of incorporating differentiable retrieval into language model pretraining [18], while the RAG framework proposed by Lewis et al. integrated dense passage retrieval with transformer-based generation, significantly improving factual grounding and contextual relevance in knowledge-intensive tasks [19].

Recent advances have focused primarily on strengthening the retrieval component of RAG systems. RAPTOR introduced hierarchical document organization to facilitate multi-level reasoning over complex knowledge sources [20]. Similarly, METRAG employed hierarchical summarization to improve the quality and informativeness of the contextual evidence provided to the language model [21]. Retrieval-aware fine-tuning strategies, such as RAFT, further enhanced model robustness by enabling language models to distinguish relevant evidence from distracting or irrelevant retrieved documents during training [22].

The application of RAG to Arabic Natural Language Processing has attracted growing research interest. Recent investigations have systematically evaluated chunking strategies, embedding models, reranking techniques, and large language models for Arabic retrieval tasks [23]. Their findings indicate that sentence-aware chunking, modern multilingual embedding models such as BGE-M3 and Multilingual-E5-large, and retrieval reranking consistently improve retrieval effectiveness and the factual reliability of generated responses.

Additional studies have examined semantic embedding models specifically for Arabic lexical retrieval [24]. Comparative evaluations involving AraBERT, CAMeLBERT, Multilingual-E5, and related models demonstrated that sentence-level semantic representations substantially outperform word-level embeddings. Furthermore, comparative assessments of generation models, including GPT-4, Gemini-1.5, Aya-8B, and AceGPT-13B, confirmed that generation quality remains an important factor influencing the overall performance of Arabic RAG systems.

Despite these advances, existing studies have primarily concentrated on improving retrieval accuracy and response generation, while comparatively little attention has been devoted to the semantic organization of heterogeneous aggregated search results. In particular, the integration of Retrieval-Augmented Generation with Arabic aggregated search across multiple search verticals, including web pages, images, news articles, and videos, remains largely unexplored. Furthermore, although previous studies have extensively evaluated embedding models for retrieval tasks, the use of RAG-enriched textual representations as input to a semantic clustering framework has not been investigated. The proposed framework addresses these research gaps by integrating Retrieval-Augmented Generation, AraBERT semantic embeddings, stacked autoencoders for nonlinear feature compression, and K-Means clustering into a unified framework for the semantic organization of aggregated Arabic search results.

III. Challenges and Contributions

The rapid expansion of Arabic digital content has intensified the demand for intelligent information retrieval systems capable of effectively discovering, integrating, and organizing information from diverse sources. Conventional search engines, such as Google and Bing, primarily employ keyword-based retrieval strategies.

Although effective for simple information needs, these approaches often fail to capture the underlying semantic intent of user queries. This limitation is particularly pronounced in Arabic because of its rich morphology, dialectal diversity, and orthographic variations, all of which complicate semantic matching and reduce retrieval effectiveness.

Aggregated search extends traditional retrieval by integrating results from multiple search verticals, including web pages, images, news articles, and videos, into a unified interface. Despite improved information accessibility, the aggregated output is inherently heterogeneous, often containing duplicate entries, irrelevant content, fragmented snippets, and insufficient context. These characteristics reduce the suitability of the retrieved content for downstream semantic analysis and clustering. Consequently, effective semantic enrichment and content refinement are essential before meaningful organization can be achieved.

Retrieval-Augmented Generation (RAG) provides an effective mechanism for enriching aggregated search results by integrating document retrieval with neural text generation. Instead of relying solely on fragmented search snippets, RAG retrieves the most relevant evidence from multiple search verticals and generates coherent, contextually grounded textual representations. By incorporating external knowledge during generation, RAG improves the relevance, completeness, and semantic consistency of the resulting content, thereby producing richer representations for subsequent analytical tasks.

Representing Arabic text remains challenging due to the language's complex morphological structure. Individual words may contain multiple prefixes and suffixes, while orthographic variations and diacritics can substantially alter lexical meaning. Conventional embedding models often struggle to preserve these linguistic characteristics, limiting their ability to capture semantic similarity accurately. AraBERT addresses these challenges through Arabic-specific pretraining and a customized WordPiece tokenization strategy designed to model Arabic's linguistic properties more effectively. Consequently, AraBERT generates contextualized semantic representations that better preserve lexical and semantic relationships, providing a robust foundation for downstream clustering.

The main steps of the proposed work can be summarized as follows:

- Collect relational aggregated search (RAS) results from multiple verticals, represented as text.
- Enrich the collected data using RAG to produce cleaner, more accurate datasets.
- Preprocess the RAG-generated output (remove diacritics and normalize characters).
- Generate 768-dimensional semantic vectors using AraBERT.
- Compress the embedding using a stacked autoencoder to extract the most informative features.
- Cluster the compressed features using K-Means to identify coherent topic groups.

The performance of the proposed framework is evaluated by comparing it with several alternative approaches:

- The author's previous work [25]: the same pipeline but without RAG (RAS + AraBERT + SAE + KM).
- RAG + AraBERT + PCA + KM: replaces SAE with PCA.
- RAG + AraBERT + KM: no feature extraction at all.
- RAG + Aravec + KM: replace AraBERT with Aravec.

The comparative analysis shows that the proposed method outperforms all of them.

IV. METHODOLOGY

The proposed framework comprises two principal modules: Data Collection (Fig. 1) and Data Enrichment (Fig. 2). The enrichment module consists of five sequential stages: Retrieval-Augmented Generation (RAG), text preprocessing, AraBERT-based semantic embedding, stacked autoencoder (SAE) feature extraction, and K-Means clustering. Unlike conventional aggregated search frameworks, the proposed approach introduces an intelligent semantic enrichment stage between the collection of Relational Aggregated Search (RAS) results and the clustering process. This intermediate stage employs RAG to improve the semantic quality of the retrieved information and utilizes a stacked autoencoder to learn compact latent representations that better preserve semantic relationships among search results.

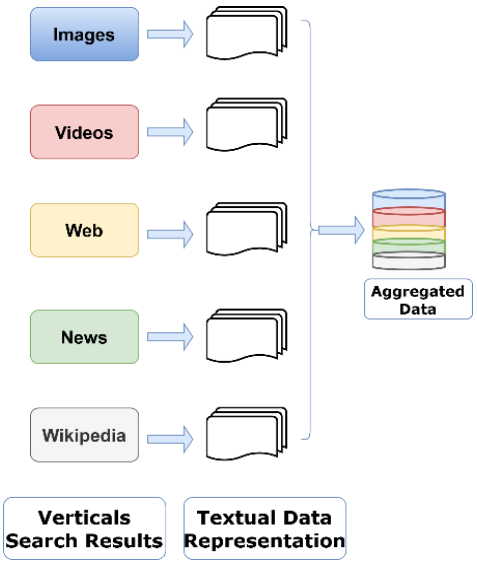


Fig. 1. Aggregated data from the aggregated search.

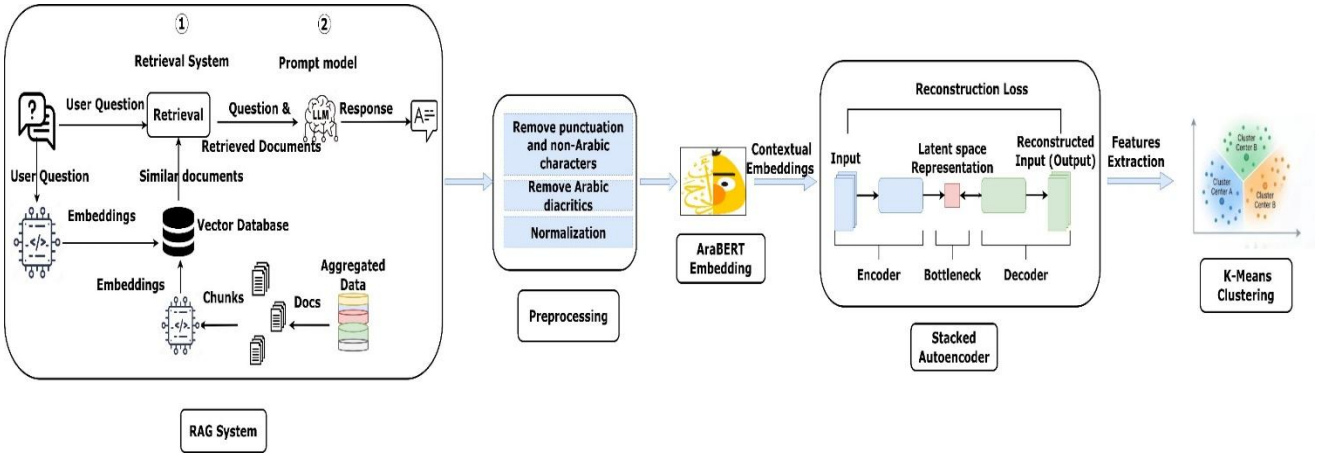


Fig. 2. The architecture of our proposed system comprises five steps: RAG system, pre-processing, results embedding (AraBERT), feature extraction (stacked autoencoder), and clustering algorithm (K-Means).

A. Aggregated Data Collection

Relational Aggregated Search (RAS) results are collected using publicly available application programming interfaces (APIs), including Google Search, Bing Search, YouTube Search, and Wikipedia. For each query, the top 100 results are retrieved from each search vertical, including web pages, images, news articles, and videos. Because each vertical returns information in a different format, all retrieved records are transformed into a unified textual representation consisting of the title, snippet, and URL. This standardized representation facilitates subsequent processing and semantic analysis. The complete data collection procedure is summarized in Algorithm 1. The resulting corpus serves as the input to the proposed data enrichment module.

Algorithm 1: Aggregated Data

Input: User Query Q
Output: Aggregated Data Text_Rep

Step 1: Query Verticals

```
Verticals = [ "web", "image", "video", "news"]
Text_Rep=""

FOR EACH V IN Verticals:

  Step 2: Retrieve the top 100 results
  results = SERP_API.search(engine=vertical, query=Q, num=100)

  Step 3: Process each result

  FOR EACH result IN results:
      // Extract core text components
      Title   = Doc.get ("title", "No Title")
      Snippet = Doc.get ("snippet", "")
      Link = Doc.get ("link", "No URL")

      // Skip results that have no usable text content
      IF Snippet IS EMPTY And Title IS EMPTY:
        CONTINUE

      Step 4: Build the text block
      Document_Block += "TITLE: " + Title + "\n"
      Document_Block += "CONTENT: " + Snippet + "\n"
      Document_Block += "Link: " + Link + "\n"
      Document_Block += "---" + "\n"

      // Append to the aggregate representation
      Text_Representation += Document_Block

  END FOR
END FOR

RETURN Text_Rep
```

B. Retrieval-Augmented Generation for Data Enrichment

The collected search results frequently contain duplicate entries, irrelevant information, fragmented snippets, and insufficient contextual information, all of which adversely affect clustering performance. To address these limitations, Retrieval-Augmented Generation (RAG) is employed as a semantic enrichment step prior to clustering. Rather than processing raw search snippets directly, the RAG framework retrieves the most relevant contextual information and generates coherent, contextually grounded textual representations. This enrichment process substantially reduces noise and redundancy while improving the completeness, consistency, and semantic relevance of the generated content, thereby providing higher-quality input for subsequent representation learning and clustering.

1) Building the vector database: To enable efficient semantic retrieval, the textual corpus is transformed into a vector database through the following stages, as illustrated in Fig. 3.

a) Chunking: The documents are segmented using a recursive text-splitting strategy that preserves the natural linguistic structure of the text by considering paragraph

boundaries, sentence boundaries, punctuation, and whitespace. This strategy generates semantically coherent text segments while maintaining contextual continuity.

b) Semantic embedding: Each text segment is encoded into a 384-dimensional dense vector using the sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 model. This multilingual embedding model was selected because it provides effective semantic representations for Arabic while maintaining high computational efficiency, making it well suited for large-scale semantic retrieval.

c) Vector storage: The generated embeddings, together with their corresponding text segments, are stored in Chroma, an open-source vector database that supports persistent storage and efficient cosine similarity search. Chroma was selected for its lightweight architecture, local deployment capability, and seamless integration with the LangChain framework, enabling efficient retrieval without additional infrastructure.

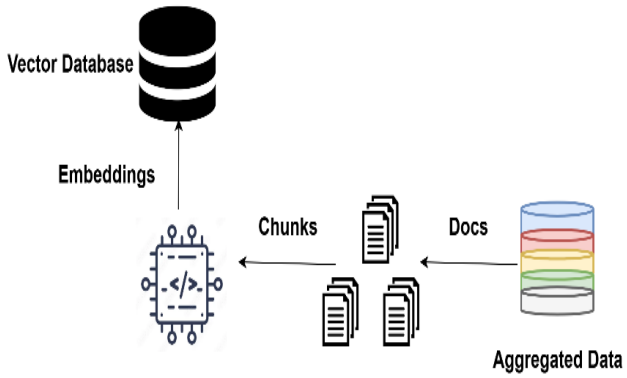


Fig. 3. Building a knowledge base.

2) Retrieval and generation workflow: For each Arabic user query, the Retrieval-Augmented Generation (RAG) pipeline performs five sequential stages, as illustrated in Fig. 4 and Algorithm 2.

a) Query representation: The input query is encoded using the same multilingual embedding model employed during vector database construction, thereby ensuring that queries and indexed documents reside within the same semantic embedding space.

b) Semantic retrieval: The generated query embedding is used to perform a similarity search within the Chroma vector database. The top ten semantically relevant text segments are retrieved based on cosine similarity.

c) Prompt construction: The retrieved text segments are concatenated to form a unified contextual representation, which is subsequently combined with the original query using a predefined prompt template. This enriched prompt provides the language model with the contextual evidence required to generate a grounded response.

d) Response generation: The constructed prompt is processed using the Google Gemini 1.5 Flash large language model to generate coherent, contextually grounded Arabic responses based on the retrieved evidence.

e) Dataset enrichment: Multiple semantically diverse responses are generated for each query and incorporated into the enriched dataset. These responses provide higher-quality textual representations for subsequent semantic embedding and clustering.

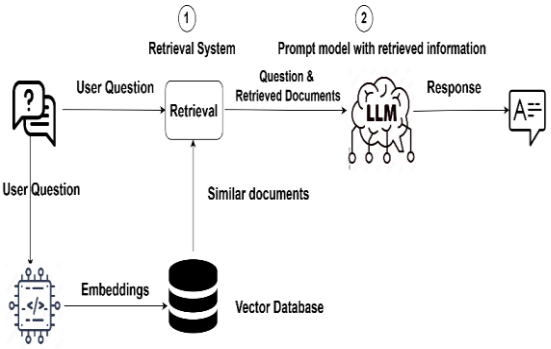


Fig. 4. RAG pipelines

Algorithm 2: RAG Framework

Input: User Question Q, Aggregated_Data
Output: Response
Step 1: Build Vector DB
FOR EACH document IN Aggregated_Data:
  Split document into chunks (C)
  Doc_Vecs = EMBEDDING_MODEL(C)
  Store (Doc_Vecs, C) in VectorDB
END FOR
Step 2: Similarity Ranking
Question_Vec = EMBEDDNG_MODEL(Q)
Scores = CALCULATE_COSINE_SIMILARITY(Question_Vec, Doc_Vecs)
Ranked_Docs = SORT (Doc_Vecs, BY = Scores, DESC)
Top_K_Context = Ranked_Docs[0:10]
Step 3: Prompt
Prompt_Template = "
You are an AI assistant for Arabic aggregated search.
Answer the user's question based only on the context provided.
Provide a single paragraph in Arabic. Do not add introductions, commentary, or information outside the context.
{Context}
Question:
{question}
Answer:
"
Final_Prompot = FILL (Prompt_Template, Context= Top_K_Context, question=Q)
Step 4: LLM Generation
Response = LLM_GENERATE(Final_Prompt)

3) Text preprocessing: The RAG-generated responses undergo a standardized preprocessing pipeline before semantic embedding. The preprocessing procedure includes:

- Removal of Arabic diacritics (Fatha, Damma, Kasra, Sukoon, and Shadda)
- Normalization of URLs, email addresses, and numerical expressions.

- Removal of Tatweel (elongation characters)
- Elimination of non-Arabic characters.
- Normalization of Alif variants (ا → آ، أ، إ)

Stop-word removal is not performed because AraBERT relies on function words (prepositions, conjunctions) to understand syntactic structure. The same preprocessing pipeline is applied to all documents to ensure consistent representation throughout the framework.

4) Semantic representation using AraBERT: Semantic representation is performed using the AraBERTv2-base model, a transformer-based language model pretrained specifically for Arabic [11]. Each preprocessed document is tokenized using the AraBERT WordPiece tokenizer. Documents exceeding the maximum sequence length are truncated, whereas shorter sequences are padded to a fixed length of 512 tokens.

The implementation is based on the Hugging Face Transformers library aubmindlab[1] using the aubmindlab/bert-base-arabertv02 pretrained model. Contextual document representations are obtained from the final hidden state of the [CLS] token, resulting in a 768-dimensional semantic embedding for each document. These embeddings capture the overall contextual meaning of the document and serve as input to the subsequent feature-extraction stage.

5) Stacked Autoencoder (SAE): To reduce the dimensionality of the AraBERT semantic embeddings, a stacked autoencoder (SAE) is employed as a nonlinear feature extraction model. The SAE learns a compact latent representation by encoding the original 768-dimensional embedding into a lower-dimensional feature space, then reconstructing the input from this compressed representation. Successful reconstruction indicates that the latent representation preserves the most informative semantic characteristics while eliminating redundant information.

As illustrated in Fig. 5, the proposed SAE architecture consists of three principal components:

- Encoder: Progressively transforms the high-dimensional input embeddings into a compact latent representation.
- Latent Space (Bottleneck): Represents the compressed semantic feature vector used for the subsequent clustering stage.
- Decoder: Reconstructs the original embedding from the latent representation.

Model training is performed by minimizing the reconstruction error between input embeddings and their reconstructions using the Mean Squared Error (MSE) loss. The Rectified Linear Unit (ReLU) activation function is adopted throughout the hidden layers because of its computational efficiency and its ability to alleviate the vanishing gradient problem. A linear activation function is employed in the output layer to reconstruct continuous embedding values. Network parameters are optimized through backpropagation until convergence.

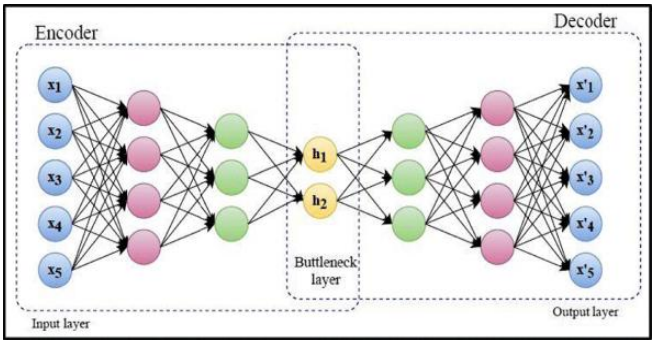


Fig. 5. Autoencoder architecture [26]

Because the experimental datasets differ in size, topic distribution, and semantic complexity, a single stacked autoencoder architecture was not suitable for all datasets. Therefore, several candidate architectures and hyperparameter configurations were evaluated empirically for each dataset. The final architecture was selected based on its validation performance, considering both the lowest reconstruction loss and the highest clustering quality, as measured by the Silhouette coefficient. This adaptive selection strategy balances feature compression with clustering performance while reducing the risk of underfitting or overfitting across datasets. The final configurations adopted in this study are summarized in Table I.

TABLE I. HYPERPARAMETERS FOR STACKED AUTOENCODER

Parameter	Optimized Value	Description
Input dimensions	768	AraBERT's dimensions
Hidden dims	(512, 256, 128, 64, 32, 16, 2) / (512, 256, 128, 64)	Hierarchical stages of feature compression.
Activation Function	ReLU	A non-linear function that is used between stacked layers.
Learning Rate (lr)	0.001	Step size for the Adam optimizer.
Batch Size	64 / 32	Number of samples processed per iteration.
Max_Epochs	100	Maximum training iterations

In the proposed architecture, for certain datasets, the encoder progressively reduces the original 768-dimensional AraBERT embeddings through successive hidden layers with 512, 256, 128, 64, 32, and 16 neurons before generating the final two-dimensional latent representation for clustering. This adaptive architecture enables effective feature compression while preserving the semantic structure required for accurate clustering.

During the decoding stage, the two-dimensional latent representation is progressively reconstructed through a sequence of hidden layers with 16, 32, 64, 128, 256, and 512 neurons. The final output layer restores the original 768-dimensional feature space, producing a reconstructed embedding corresponding to the input representation. The reconstruction quality is quantified

1 https://huggingface.co/aubmindlab/

by computing the Mean Squared Error (MSE) between the original and reconstructed embeddings, which serves as the optimization objective during training.

Model optimization is performed using the Adam optimizer with an initial learning rate of 0.001. Depending on the size and characteristics of each dataset, a batch size of either 32 or 64 is adopted. Training is conducted for up to 100 epochs, with an early stopping strategy employed to improve generalization and prevent overfitting. Specifically, training is terminated if the validation loss does not improve for 10 consecutive epochs, and the model parameters corresponding to the lowest validation loss are retained.

6) K-Means clustering: The compressed latent representations generated by the stacked autoencoder are subsequently partitioned into semantically coherent groups using K-Means clustering. For each dataset with T predefined target topics, the number of clusters is set to K = T + 1. The additional cluster is designated as "Other" and is intended to accommodate documents that cannot be reliably assigned to any predefined topic category, thereby improving the semantic purity and separation of the primary clusters. The K values adopted for each dataset are summarized in Table II.

TABLE II. DATASETS USED IN OUR EXPERIMENTS

Dataset	Target Topics	Dataset Size	K (clusters)	Clusters Label
D1	Education, Sports	383	3	Education, Sports, Other
D2	Education, IT	445	3	Education, IT, Other
D3	Sports, IT	458	3	Sports, IT, Other
D4	Education, Sports, IT	643	4	Education, Sports, IT, Other
D5	Weather, Healthy Food	439	3	Weather, Healthy Food, Other
D6	Sports, Healthy Food	426	3	Sports, Healthy Food, Other
D7	Healthy Food, Weather, Education	624	4	Healthy Food, Weather, Education, Other
D8	Sports, Healthy Food, Weather, Education	822	5	Sports, Healthy Food, Weather, Education, Other
D9	Education, Sports, Weather, IT	854	5	Education, Sports, Weather, IT, Other

K-Means was selected because the proposed stacked autoencoder transforms high-dimensional AraBERT embeddings into a compact latent feature space characterized by high intra-cluster cohesion and enhanced inter-cluster discrimination. Such representations are particularly well-suited to the centroid-based optimization strategy employed by K-Means, enabling effective partitioning of semantically similar documents while maintaining low computational complexity. Furthermore, because the semantic categories are known in advance, the value of K can be specified directly, without requiring additional cluster-estimation procedures. To improve the stability and reproducibility of the clustering results, K-Means was initialized using the scikit-learn library in Python, with a fixed random seed (random_state=42) to ensure reproducibility of the experimental results.

V. EXPERIMENTAL SETUP

This section describes the experimental configuration, dataset construction process, and evaluation criteria used to assess the effectiveness of the proposed framework. All experiments were implemented in Python using the PyTorch deep learning framework and executed on a 64-bit Windows 10 workstation equipped with an Intel Core i7 processor and 16 GB of RAM, as well as on Google Colab with GPU acceleration.

A. Data Collection and Dataset Construction

Relational aggregated search results were collected from eight search verticals using the Google Search APIs (Web, News, Images, and Videos) and Bing Search APIs (Web, News, Images, and Videos), using the SERP API [2] and the Wikipedia API. Five representative Arabic topics were selected for the experiments: التعليم (Education), الرياضة (Sports), تكنولوجيا المعلومات (Information Technology), الغذاء الصحي (Healthy Food), and الطقس (Weather). For each topic, the top 100 search results were retrieved from every search vertical.

To ensure consistency across heterogeneous sources, each retrieved record was transformed into a standardized textual representation comprising the document title, descriptive snippet, and source URL.

The collected search results were subsequently processed using the proposed Retrieval-Augmented Generation (RAG) framework. For each topic, multiple Arabic information-seeking questions were formulated to retrieve relevant contextual evidence and generate coherent Arabic paragraphs. These generated responses constituted the enriched corpus used throughout the experimental evaluation.

Nine experimental datasets, denoted D1–D9, were constructed by combining different pairs and groups of topic-specific documents to evaluate the proposed framework under varying levels of semantic diversity. The composition of each dataset is summarized in Table II.

B. Evaluation Metrics

The effectiveness of the proposed framework was evaluated using two widely adopted internal clustering validation measures: the Silhouette Coefficient [27] and the Davies–Bouldin Index (DBI) [28].

The Silhouette Coefficient quantifies both cluster cohesion and inter-cluster separation, producing values within the range [-1, 1]:

$$\text{Silhouette Score} = (b - a) / \max (a , b) \quad (1)$$

where, (a) denotes the average distance between a sample and all other samples within the same cluster, while (b) represents the average distance between that sample and observations belonging to neighboring clusters. Higher Silhouette values indicate stronger cluster cohesion and better

2 https://serpapi.com/

separation, whereas negative values suggest potential misclassification.

The Davies–Bouldin Index (DBI) evaluates clustering quality by jointly considering intra-cluster compactness and inter-cluster separation. It is computed as:

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max\left(\frac{\Delta(X_i)+\Delta(X_j)}{\delta(X_i,X_j)}\right) \quad (2)$$

where, $(\Delta(X_i))$ represents the average intra-cluster distance within cluster (X_i), and $(\delta(X_i, X_j))$ denotes the distance between clusters (X_i) and (X_j). Lower DBI values indicate compact, well-separated clusters and therefore correspond to superior clustering performance.

VI. Results and Discussion

The proposed framework was evaluated using nine experimental datasets representing different topic combinations and varying levels of semantic complexity. The analysis was conducted in two stages. First, the overall clustering performance of the proposed framework was examined. Second, comparative experiments were performed to quantify the contribution of each major component.

A. Performance Analysis of the Proposed Framework

Table III summarizes the clustering performance of the proposed RAG + AraBERT + SAE + K-Means framework across all experimental datasets. The obtained Silhouette coefficients consistently exceed 0.65, indicating strong intra-cluster cohesion and clear inter-cluster separation. These results demonstrate that the proposed semantic representation effectively preserves topic-specific characteristics while maintaining sufficient discrimination between different semantic categories.

TABLE III. Clustering Evaluation of the Proposed Method

Dataset	Silhouette Score	Davies-Bouldin Index
D1	0.72	0.32
D2	0.70	0.42
D3	0.68	0.49
D4	0.68	0.38
D5	0.66	0.49
D6	0.70	0.45
D7	0.65	0.45
D8	0.66	0.42
D9	0.69	0.37

The proposed framework also exhibits stable performance as the semantic complexity of the datasets increases. Datasets D8 and D9, which comprise four distinct topics partitioned into five clusters, achieved Silhouette scores of 0.66 and 0.69, respectively. These findings indicate that the compressed latent representations generated by the stacked autoencoder remain highly discriminative even when multiple topics are simultaneously considered.

The Davies–Bouldin Index further supports these observations. Across all datasets, DBI values remain below 0.50, indicating compact cluster formation and substantial separation among neighboring clusters. Collectively, these results demonstrate that the proposed framework consistently produces well-defined semantic clusters under varying experimental conditions.

B. Comparative Evaluation

To evaluate the effectiveness of the proposed feature extraction strategy, the stacked autoencoder was compared with Principal Component Analysis (PCA), a widely adopted linear dimensionality reduction technique. Clustering performance was assessed using the Silhouette Coefficient and the Davies–Bouldin Index, where higher Silhouette values and lower DBI values indicate superior clustering quality.

As shown in Table IV, the proposed RAG + AraBERT + SAE + K-Means framework consistently outperforms the RAG + AraBERT + PCA + K-Means configuration across all evaluation metrics. The nonlinear feature-learning capability of the stacked autoencoder preserves complex semantic relationships that cannot be adequately captured by linear projection, resulting in more compact and better-separated clusters.

TABLE IV. Comparison of Alternative Methods in Terms of Silhouette Score and Davies-Bouldin Index.

Method	Avg. Silhouette	Avg. DBI
RAG + AraVec + KM	0.15	2.12
RAG + AraBERT + KM	0.18	1.99
RAG + AraBERT + PCA + KM	0.44	0.99
RAG + AraBERT + SAE + KM (Ours)	**0.68**	**0.42**

The contribution of nonlinear feature compression is further illustrated by comparing AraBERT embeddings before and after dimensionality reduction. Without the stacked autoencoder, AraBERT achieved a Silhouette score of 0.18, representing a modest improvement over AraVec (0.15). Incorporating the stacked autoencoder increased the average Silhouette coefficient to 0.68, demonstrating that nonlinear representation learning substantially enhances the discriminative structure of the embedding space before clustering.

The effectiveness of the proposed Retrieval-Augmented Generation (RAG) stage was further evaluated by comparing the current framework with the authors' previous approach (baseline), Aggregated + AraBERT + SAE + K-Means [25], which operated directly on raw aggregated search results without semantic enrichment. The comparison presented in Table V demonstrates that introducing RAG consistently improves clustering quality across all nine datasets. The average Silhouette coefficient increased from 0.59 to 0.68, corresponding to an improvement of approximately 15%.

The performance gain is particularly evident for datasets containing semantically related topics, such as D2, where lexical overlap increases the difficulty of cluster separation. By transforming fragmented search snippets into coherent, context-aware representations, the RAG module reduces semantic ambiguity and strengthens topic-specific representations prior to embedding. Consequently, documents belonging to the same

semantic category become more tightly grouped, while the separation between different topic categories increases. This behavior is maintained even for the most challenging datasets containing three or more topic categories (e.g., D4, D7, D8, and D9), demonstrating the robustness of the proposed framework as semantic complexity increases.

TABLE V. COMPARISON OF SILHOUETTE SCORES OF K-MEANS WITH OUR PREVIOUS METHOD.

Dataset	RAS + AraBERT + SAE + KM (baseline)	RAG + AraBERT + SAE + KM (proposed)
D1	0.67	**0.72**
D2	0.59	**0.70**
D3	0.62	**0.68**
D4	0.59	**0.68**
D5	0.63	**0.66**
D6	0.65	**0.70**
D7	0.55	**0.65**
D8	0.54	**0.66**
D9	0.55	**0.69**

These improvements can be attributed to the semantic enrichment performed by the RAG module. Rather than relying directly on heterogeneous aggregated search snippets, the framework generates coherent, contextually grounded textual representations that provide richer semantic information for subsequent embedding and clustering. The resulting representations contain less redundancy, reduced contextual ambiguity, and improved semantic consistency, thereby enabling the downstream feature extraction and clustering stages to operate on higher-quality input data.

To determine whether the observed performance gains were statistically significant, both the paired t-test and the Wilcoxon signed-rank test were performed using the Silhouette coefficients obtained across the nine datasets. The proposed framework achieved an average Silhouette coefficient of 0.682 ± 0.023, compared with 0.599 ± 0.047 for the baseline approach. The paired t-test confirmed that the improvement was statistically significant ($t = 6.68$, $p = 0.00016$), while the Wilcoxon signed-rank test produced a similarly significant result ($p = 0.0039$).

Hypothesis testing was subsequently conducted using a significance level of $\alpha = 0.05$. The null hypothesis (H_0) assumed that no statistically significant difference exists between the baseline aggregated search framework and the proposed RAG-enhanced framework. The alternative hypothesis (H_1) stated that the proposed framework yields superior clustering performance. Because the p-values obtained from both statistical tests were substantially lower than the predefined significance level, the null hypothesis was rejected in favor of the alternative hypothesis. These findings provide strong statistical evidence that the improvements achieved by the proposed framework are unlikely to have occurred by chance and instead result from integrating Retrieval-Augmented Generation with semantic representation learning.

VII. CONCLUSION AND FUTURE WORK

This study presented an intelligent framework for enhancing Arabic aggregated search by integrating Retrieval-Augmented Generation (RAG), contextual semantic representation, nonlinear feature learning, and unsupervised clustering. By enriching heterogeneous search results before semantic analysis, the proposed framework effectively reduced contextual ambiguity and redundancy while improving the semantic quality of the retrieved information. The enriched representations were subsequently encoded using AraBERT, compressed through a stacked autoencoder, and organized into semantically coherent topic clusters using the K-Means algorithm. Experimental evaluation on nine Arabic datasets spanning five application domains demonstrated the effectiveness and robustness of the proposed framework. Compared with clustering based on raw aggregated search results, the proposed approach increased the average Silhouette coefficient from 0.59 to 0.68, corresponding to an improvement of approximately 15%. The framework consistently achieved superior clustering performance across all experimental datasets and outperformed alternative dimensionality reduction and embedding approaches, including Principal Component Analysis (PCA), direct clustering of AraBERT embeddings, and AraVec-based representations. Statistical analysis further confirmed that these improvements were significant, demonstrating the effectiveness of integrating Retrieval-Augmented Generation with semantic representation learning for Arabic aggregated search.

Future research will investigate more advanced Retrieval-Augmented Generation architectures, including Agentic RAG, Corrective RAG (CRAG), Modular RAG, and Graph RAG, to further improve retrieval accuracy, reasoning capability, and factual consistency. Additional research directions include evaluating stronger multilingual retrieval models, incorporating retrieval reranking mechanisms, and extending the proposed framework to larger and more diverse Arabic corpora to assess its scalability and generalization across broader information retrieval applications.

REFERENCES

[1] A. Kopliku, K. Pinel-Sauvagnat, and M. Boughanem, “Aggregated search: A new information retrieval paradigm," ACM Comput. Surv., vol. 46, no. 3, pp. 1-31, 2014.

[2] S. Achsas, “Improving relational aggregated search from big data sources using stacked autoencoders,” Cognit. Syst. Res., vol. 51, pp. 61–71, 2018.

[3] Q. Luo, Q. Shao, Z. Zhang, B. Zhang, and J. Shi, “SCAT_LoRA: self-consistency adversarial training in large language models through low-rank adaptations,” Eng. Res. Express, vol. 7, no. 2, p. 025265, 2025.

[4] A. Mishra and A. Gupta, “Retrieval Augmented Generation (RAG) Model,” Int. J. Res., vol. 6, no. 6, pp. 4690–4693, 2025.

[5] N. Kandpal, H. Deng, A. Roberts, E. Wallace, and C. Raffel, “Large language models struggle to learn long-tail knowledge,” in Proc. Int. Conf. Mach. Learn., pp. 15696–15707, 2023.

[6] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, “Retrieval-augmented generation for large language models: A survey,” arXiv preprint arXiv:2312.10997, 2023.

[7] G. P. Wellawatte, H. Guo, M. Lederbauer, A. Borisova, M. Hart, M. Brucka, and P. Schwaller, “ChemLit-QA: a human evaluated dataset for chemistry RAG tasks,” Mach. Learn.: Sci. Technol., vol. 6, no. 2, p. 020601, 2025.

[8] I. O. William and M. Altamimi, “Text Embedding Implementation Using Retrieval Augmented Generation (RAG) Model Combined With Large Language Model,” Int. J. Adv. Nat. Sci. Eng. Res., vol. 5, no. 5, 2024.

[9] A. Kimothi, "A Simple Guide to Retrieval Augmented Generation," Manning Publications, 2025.

[10] B. Wang, A. Wang, F. Chen, Y. Wang, and C. C. J. Kuo, "Evaluating word embedding models: Methods and experimental results," APSIPA Trans. Signal Inf. Process., vol. 8, p. e19, 2019.

[11] W. Antoun, F. Baly, and H. Hajj, "Arabert: Transformer-based model for arabic language understanding," in Proc. 4th Workshop Open-Source Arabic Corpora Process. Tools, pp. 9–15, 2020.

[12] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P. A. Manzagol, and L. Bottou, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," J. Mach. Learn. Res., vol. 11, no. 12, pp. 3371–3408, 2010.

[13] A. K. Jain, "Data clustering: 50 years beyond K-means," Pattern Recognit. Lett., vol. 31, no. 8, pp. 651–666, 2010.

[14] R. Xu and D. Wunsch, "Survey of clustering algorithms," IEEE Trans. Neural Networks, vol. 16, no. 3, pp. 645–678, 2005.

[15] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," Commun. ACM, vol. 18, no. 11, pp. 613–620, 1975.

[16] 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.

[17] D. Chen, A. Fisch, J. Weston, and A. Bordes, "Reading wikipedia to answer open-domain questions," in Proc. 55th Annu. Meeting Assoc. Comput. Linguistics (Vol. 1: Long Papers), pp. 1870–1879, 2017.

[18] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "Retrieval augmented language model pre-training," in Proc. Int. Conf. Mach. Learn., pp. 3929–3938, 2020.

[19] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschle, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in Advances in Neural Information Processing Systems, vol. 33, pp. 9459–9474, 2020.

[20] P. Sarthi, S. Abdullah, A. Tuli, S. Khanna, A. Goldie, and C. D. Manning, "Raptor: Recursive abstractive processing for tree-organized retrieval," in Proc. 12th Int. Conf. Learn. Represent., 2024.

[21] C. Gan, D. Yang, B. Hu, H. Zhang, S. Li, Z. Liu, C. Wang, and J. Zhou, "Similarity is not all you need: Endowing retrieval augmented generation with multi layered thoughts," arXiv preprint arXiv:2405.19893, 2024.

[22] T. Zhang, S. G. Patil, N. Jain, S. Shen, M. Zaharia, I. Stoica, and J. E. Gonzalez, "Raft: Adapting language model to domain specific rag," arXiv preprint arXiv:2403.10131, 2024.

[23] J. Alsubhi, M. D. Alahmadi, A. Alhusayni, I. Aldailami, I. Hamdine, A. Shabana, L. Al-Malki, and S. Khayyat, "Optimizing rag pipelines for arabic: A systematic analysis of core components," arXiv preprint arXiv:2506.06339, 2025.

[24] R. Al-Rasheed, A. Al Muaddi, H. Aljasim, R. Al-Matham, M. Alhoshan, A. Al Wazrah, and A. AlOsaimy, "Evaluating RAG Pipelines for Arabic Lexical Information Retrieval: A Comparative Study of Embedding and Generation Models," in Proc. 1st Workshop NLP Lang. Using Arabic Script, pp. 155–164, 2025.

[25] S. S. Soliman, "Deep Learning-Based Approach for Improving Relational Aggregated Search," arXiv preprint arXiv:2510.00966, 2025.

[26] S. Hosseini and Z. A. Varzaneh, "Deep text clustering using stacked AutoEncoder," Multimedia Tools Appl., vol. 81, no. 8, pp. 10861–10881, 2022.

[27] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," J. Comput. Appl. Math., vol. 20, pp. 53–65, 1987.

[28] D. L. Davies and D. W. Bouldin, "A cluster separation measure," IEEE Trans. Pattern Anal. Mach. Intell., vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.

Design, Implementation, and Evaluation of a Mobile Application for Real-Time Incident Management in Local Government

Marleny Peralta Ascue[1], Luis Bryan Caballero Pinto[2], José Luis Poluco Sicos[3], José Luis Merma Aroni[4]
Departamento Académico de Informática y Sistemas, Universidad Nacional Micaela Bastidas de Apurímac, Abancay, Perú[1, 4]
Facultad de Ingeniería, Universidad Tecnológica de los Andes, Abancay, Perú[2, 3]

***Abstract*—The increasing occurrence of incidents related to public safety, environmental issues, transportation, and civil defense requires more efficient response mechanisms from local governments. However, the absence of an integrated digital tool makes it difficult to accurately identify the location, timing, and frequency of critical incidents in real time, thereby limiting the response capacity of public institutions. This study aims to design, implement, and evaluate a mobile application for incident management to streamline the reporting process and facilitate timely responses to events reported by citizens. The application was developed using the Agile Scrum methodology. Flutter was employed for mobile application development, while PHP and RESTful APIs were used for backend services, with MySQL handling data storage and incident geolocation. In addition, location-based services and web-based visualization components were incorporated to support monitoring and information access. A pre-experimental research design with a single-group pretest–posttest approach was adopted. The results revealed a 99.26% reduction in incident registration time and a 31.68% decrease in immediate response time. The usability of the proposed application was assessed using the Computer System Usability Questionnaire (CSUQ), administered to 257 participants, showing a high level of user satisfaction. Overall, the findings demonstrate that the developed mobile application significantly improves incident management processes and supports the digital transformation initiatives of local governments.**

***Keywords*—Mobile application; incident management; local government; Agile Scrum; usability evaluation; CSUQ; digital transformation.**

I. Introduction

Public safety has become a major concern in Peru, particularly due to the increasing number of incidents reported in recent years. Between 2020 and 2024, both crime rates and citizens' perceptions of insecurity increased considerably. According to national statistics, the victimization rate rose from 18.2% in 2021 to 27.1% in 2024, while more than 80% of the population perceived insecurity as one of the country's most pressing issues [1]. In response to this situation, approximately 66% of municipalities provide public security services through local patrol units (Serenazgo), involving more than 42,000 officers nationwide [2].

Incident management in local governments extends beyond public safety. Environmental complaints, civil defense emergencies, and traffic-related events also require timely and coordinated responses. The Environmental Assessment and Enforcement Agency (OEFA) reported that more than 6,000 environmental complaints have been processed since 2024 through digital platforms such as SINADA, with many cases originating from local issues including solid waste disposal, wastewater discharges, and noise pollution [3]. Likewise, the National Emergency Operations Center (COEN) reported that more than 5,400 emergencies, including heavy rainfall, landslides, and mudslides, were registered between December 2024 and March 2025 [4]. In many cases, local governments are responsible for managing these situations despite lacking early warning mechanisms, rapid response resources, and adequate systems for documenting emergency events.

Similarly, the National Road Safety Observatory reported more than 86,000 traffic accidents and over 3,000 fatalities in Peru during 2024 [5]. Although municipalities, police departments, and healthcare institutions are involved in responding to these events, the lack of a common incident management platform often results in fragmented information and reduced operational efficiency [6], [7].

Recently, local governments have increasingly adopted digital technologies to modernize administrative processes and strengthen communication with citizens [8]. Mobile government (m-government) initiatives have emerged as an important component of this transformation, enabling public services to be delivered through smartphones and other mobile devices. These technologies facilitate real-time access to information, improve service accessibility, and enhance interactions between citizens and government institutions [9], [10].

Incident management represents a critical process for local governments because it directly affects public safety, environmental protection, civil defense, and urban services. Effective incident management requires rapid reporting, accurate classification, efficient communication among stakeholders, and timely response actions. Nevertheless, many municipalities still rely on manual procedures and isolated information systems, leading to delays in both incident notification and response activities [11], [12]. Experiences from smart city initiatives have demonstrated the importance of digital citizen platforms capable of integrating incident reporting, geolocation, and real-time communication functionalities. Motivated by these needs, this study proposes the design, implementation, and evaluation of a mobile application for real-time incident management in the context of

local governments, aiming to strengthen response and monitoring processes.

The main contributions of this work are threefold. First, a mobile application architecture for local government incident management is proposed, enabling real-time monitoring and coordination among the responsible departments. Second, the proposed solution is implemented and validated in an operational environment. Third, the impact of the application is assessed using quantitative indicators associated with incident registration and response times, as well as through usability evaluation using the Computer System Usability Questionnaire (CSUQ) [19]. The results demonstrate that mobile technologies constitute an effective tool for improving incident management processes and supporting digital transformation initiatives in local governments.

II. Related Work

Several studies have shown that mobile government applications contribute to improving public services when their design prioritizes accessibility, usability, and citizen-centered experiences. Previous research has highlighted that the adoption of mobile solutions by local governments has accelerated due to the growing demand for more efficient and personalized digital services [13]. Digital accessibility has also been identified as a key factor influencing the acceptance and effective use of government applications, particularly those intended to enhance public safety and support diverse groups of citizens [14].

Research on mobile-based incident reporting systems has demonstrated the benefits of integrating features such as image capture, textual descriptions, GPS location, and web-based administrative interfaces. These functionalities facilitate the registration, monitoring, and management of incidents reported by citizens, thereby improving communication between communities and government institutions [15]. Such findings indicate that mobile technologies and integrated management platforms can significantly enhance the efficiency of incident response processes.

Smart city research has further emphasized the role of citizen reporting systems as mechanisms for digital participation. Previous studies have shown that smart city applications enable citizens to report urban issues, collaborate with authorities, and actively contribute to decision-making processes [16]. Citizen participation has been recognized as a key factor in increasing the effectiveness and relevance of urban solutions when properly integrated into planning and management activities [17]. Likewise, sustainable smart city initiatives have demonstrated that citizen engagement directly influences urban sustainability outcomes [18]. Other studies have highlighted that technological solutions for sustainable urban development require not only digital infrastructures but also effective interactions between governments and citizens [19]. Collectively, these findings support the development of mobile applications that extend beyond process automation and strengthen the relationship between citizens and local governments.

In recent years, decentralized incident reporting models have gained increasing attention. A decentralized framework for smart city governance based on community participation was proposed in [20], demonstrating that citizens can act as active sensors by reporting incidents and contributing to urban management. Similarly, the platform presented in [21] incorporates citizen reporting, case tracking, and analytical tools to support decision-making processes. These studies illustrate the transition from conventional incident management approaches toward intelligent and collaborative platforms. From the perspective of digital government, previous research has emphasized that digital public services foster regional innovation ecosystems and support sustainable development processes [7]. In this context, the digitalization of incident registration and management processes can improve institutional efficiency while generating valuable information for public administration.

Several studies investigating mobile government services have identified usability, perceived usefulness, trust, service quality, and citizen satisfaction as key determinants of continuous usage intentions [22], [23]. Consequently, usability has emerged as one of the most important factors influencing the success of mobile applications.

The Computer System Usability Questionnaire (CSUQ), originally proposed in [24], has become one of the most widely used instruments for assessing user satisfaction with computer systems. Subsequent studies have identified usability attributes such as efficiency, learnability, satisfaction, and accessibility as essential quality dimensions of mobile services [25]. Moreover, user-centered redesign approaches have been shown to substantially improve user experience in emergency-oriented mobile applications [26]. To address common usability problems in mobile systems, the MUG-10 framework was introduced in [27], emphasizing the importance of intuitive interfaces and simplified interaction processes. In fact, many mobile applications fail not because of insufficient functionality, but because of poor usability and overly complex interaction mechanisms. In addition, focused usability evaluation approaches have also proven valuable for digital applications aimed at citizen participation and public governance [28].

Regarding technological implementation, Flutter has emerged as a popular cross-platform framework for developing Android and iOS applications using a single codebase. Previous studies have reported that Flutter simplifies mobile application development while enabling highly interactive interfaces and delivering competitive performance [29]. Comparative analyses between Flutter and alternative technologies, such as React Native, have revealed advantages in terms of productivity, visual consistency, and cross-platform capabilities [30]-[32]. These findings support the selection of Flutter as a suitable technology for developing mobile applications aimed at incident management in local government environments.

III. Methodology

A. Research Type and Design

This study is classified as applied research, as it addresses a practical problem associated with incident management in a local government environment. A pre-experimental design based on a single-group pretest–posttest approach was employed to assess the impact of the proposed mobile

application. Table I summarizes the elements of the experimental design.

TABLE I. DESCRIPTION OF THE PRE-EXPERIMENTAL DESIGN

Element	Description
Ge	Experimental group
O1	Measurement conducted before the implementation of the mobile application
X	Mobile application intervention
O2	Measurement conducted after the implementation of the mobile application

B. Study Area

The study was conducted at the Provincial Municipality of Abancay, located in the Apurímac region of Peru. This institution is responsible for managing incidents related to public safety, traffic, civil defense, and environmental issues reported by citizens.

Prior to the implementation of the proposed solution, incident registration and monitoring activities relied on conventional procedures based on manual records and telephone communication. These practices often resulted in delays in response and limited the availability of information required for effective decision-making.

C. Population and Sample

The target population consisted of 47,744 inhabitants of the district of Abancay, ranging in age from 15 to 70 years and representing both urban and rural areas, according to data provided by the National Institute of Statistics and Informatics (INEI). A non-probabilistic purposive sampling strategy was adopted, selecting citizens who reported incidents and members of the local public security service (Serenazgo) responsible for incident response and follow-up activities. The sample size was estimated using the finite population formula, resulting in a total of 267 participants.

D. Mobile Application Development

The mobile application was developed following the Agile Scrum methodology. The development process was organized into four iterative and incremental sprints, each incorporating specific functionalities associated with user management, incident reporting, operational response, and report generation. This approach enabled the continuous delivery of functional increments throughout the development cycle.

The process included requirements elicitation and analysis, Product Backlog creation, and sprint planning activities. Each sprint comprised implementation, functional testing, sprint review, and retrospective phases, facilitating continuous improvement and ensuring the progressive integration of system functionalities.

Fig. 1 presents the Scrum workflow adopted during the development process, illustrating the functionalities implemented in each sprint, the stakeholders involved, and the activities carried out throughout the iterative development cycles.

E. System Architecture

Fig. 2 presents the three-layer architecture implemented for the proposed mobile application. This architecture was designed to separate the user interface, business logic, and data management components, thereby improving maintainability, scalability, and future system expansion.

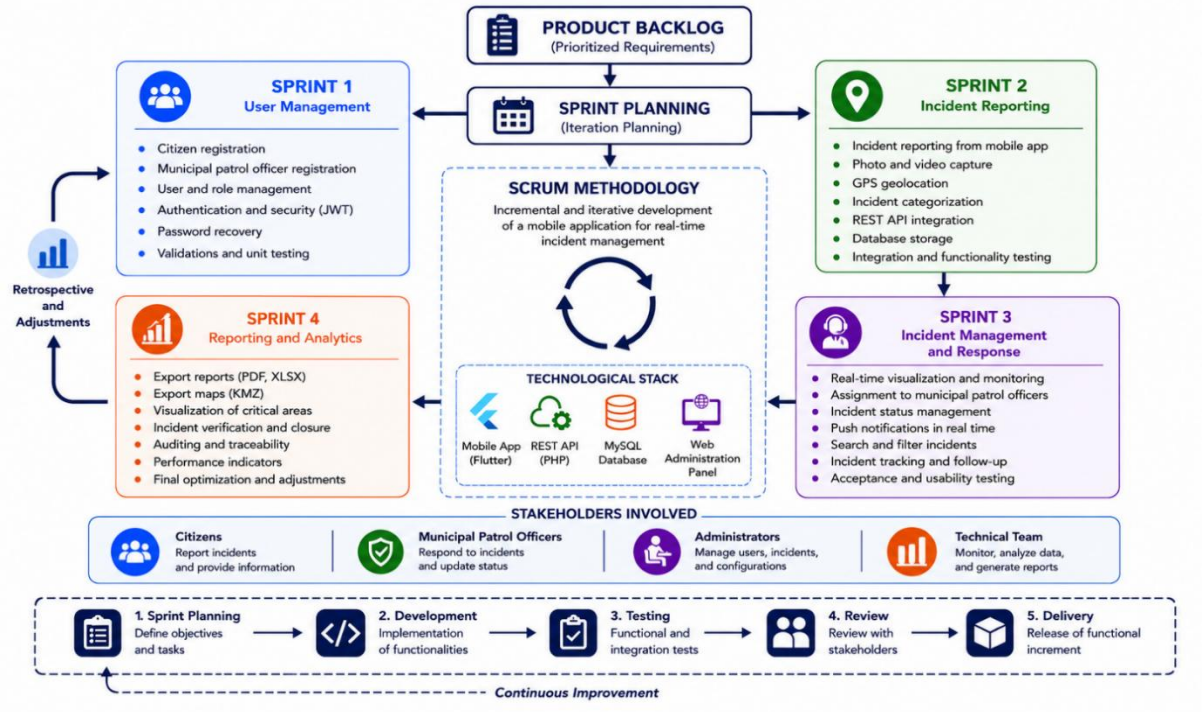


Fig. 1. Scrum-based development process of the proposed real-time incident management mobile application.

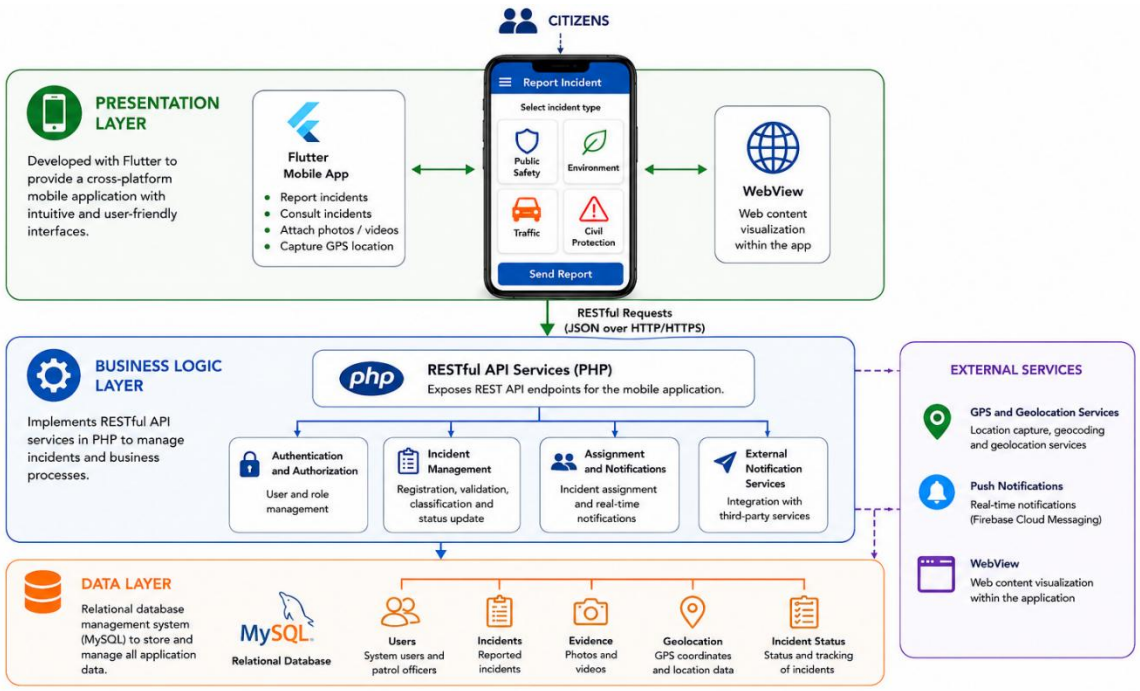


Fig. 2. Three-layer architecture of the proposed mobile application for real-time incident management.

The presentation layer was developed using Flutter to support cross-platform mobile application development. Its main purpose is to allow citizens to report and consult incidents through intuitive and user-friendly interfaces. Through this layer, users can report incidents related to public safety, environmental issues, traffic, and civil defense, including descriptive details, multimedia evidence such as photographs and videos, and GPS-based geographic location data from their mobile devices. WebView was also incorporated to display web content within the application, improving the user experience without requiring users to leave the mobile environment. Communication with the business logic layer is performed through HTTP/HTTPS requests using JSON messages, ensuring efficient and secure interaction between the mobile application and backend services.

The business logic layer was implemented in PHP using a REST API-based architecture, enabling the mobile application to connect with server-side services. This layer manages the incident lifecycle, including registration, validation, classification, assignment, and status updates. REST services were implemented to decouple the mobile application from internal server processes, simplifying maintenance and supporting future platform scalability.

The data layer was implemented using MySQL as the relational database management system. The database was designed to store information related to system users, municipal patrol officers, reported incidents, incident categories, multimedia evidence, service status, geographic coordinates, and geolocation data. The relational structure ensures data integrity and consistency, while also facilitating queries for report generation, performance indicators, and real-time incident monitoring. In addition, geospatial attributes were incorporated to analyze the geographic distribution of incidents and identify critical areas within the municipal jurisdiction.

F. Incident Reporting and Response Workflow

Fig. 3 illustrates the operational workflow implemented by the proposed mobile application for real-time incident management. The process begins when a user logs into the application and completes the authentication procedure. After successful authentication, the user selects the appropriate incident category, including public safety, environmental issues, traffic, or civil protection.

Once the incident type has been selected, the user submits the corresponding information by providing a detailed description, multimedia evidence in the form of photographs or videos, and the geographic location obtained from the mobile device's GPS services. The system then validates the collected information to ensure that all mandatory fields have been completed. If incomplete or invalid information is detected, an error message is displayed, and the user is prompted to provide the missing data before continuing.

After successful validation, the incident report is transmitted to the server and stored in the MySQL database. Based on the incident category, the system automatically assigns the case to the corresponding municipal department or responsible personnel. Real-time notifications are then delivered to authorized staff members, enabling them to initiate response actions promptly.

During the response phase, incident managers update the status of each case within the system. The incident status progresses through different stages, namely pending, in progress, and resolved. These updates are continuously synchronized with the mobile application, allowing citizens to monitor the evolution of their reports and receive real-time information regarding the current status of the incident.

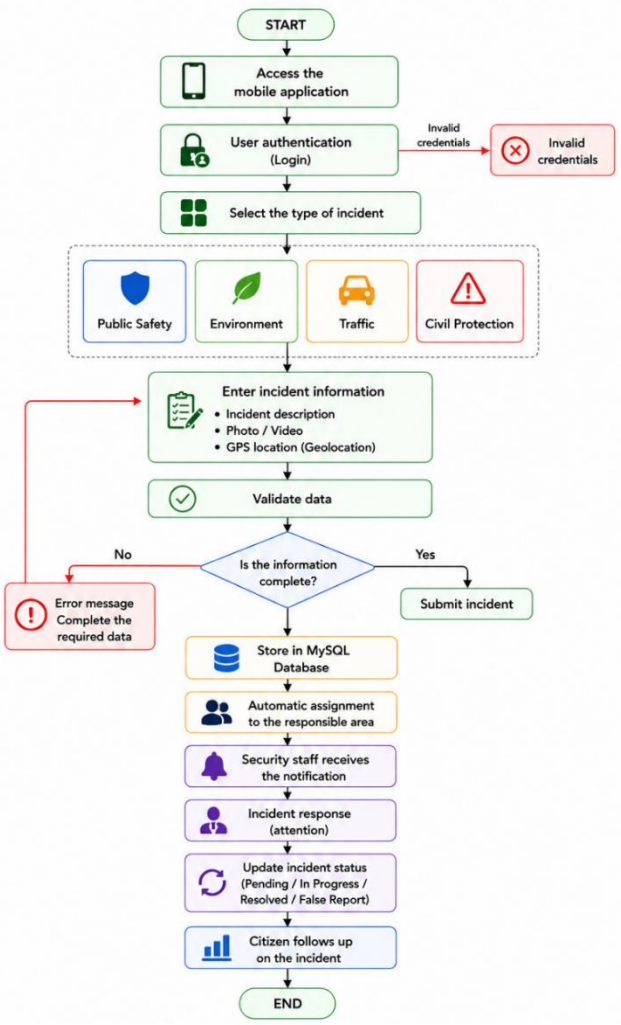


Fig. 3. Incident reporting and response process supported by the proposed mobile application.

Overall, this workflow facilitates efficient interaction between citizens and municipal authorities, improving response times, ensuring traceability of actions, and providing continuous monitoring throughout the incident lifecycle.

G. System Implementation

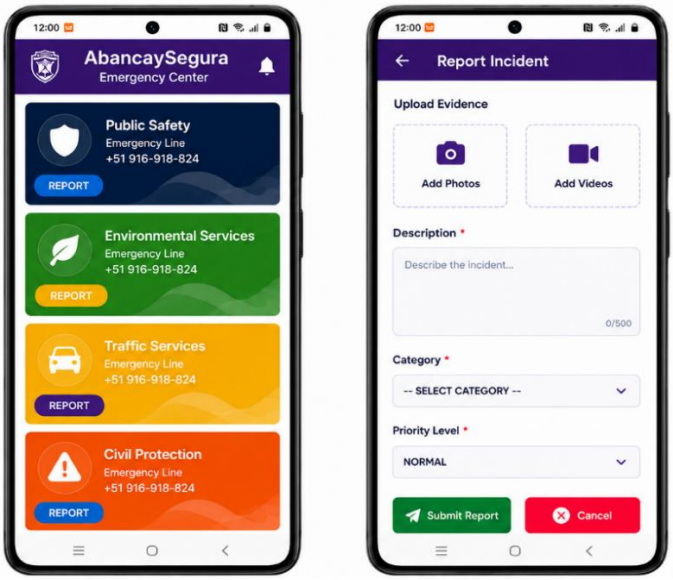
The system implementation enabled the deployment of the functionalities defined during the development phase, providing a solution focused on real-time incident reporting and response. The mobile application was designed to facilitate interaction between citizens and the municipal departments responsible for incident handling, supporting the registration, monitoring, and management of incidents related to public safety, environmental issues, traffic, and civil protection.

The main interface of the application provides access to the different incident categories, as well as consultation and tracking services. Users can submit detailed information about an event by selecting the corresponding incident category and specifying its characteristics. The application also allows the attachment of multimedia evidence, including photographs and videos, thereby improving the accuracy and reliability of incident reports.

In addition, geolocation mechanisms were incorporated to associate each incident with its geographic coordinates. This functionality facilitates spatial queries and assists responsible personnel in locating and responding to incidents more efficiently.

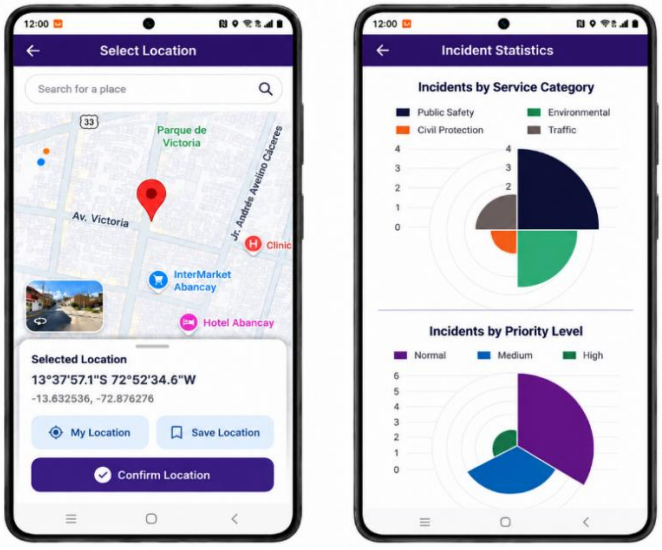
The proposed solution also incorporates monitoring and analytical capabilities that enable the generation of indicators and statistical information regarding the distribution and characteristics of reported incidents. These features provide valuable information for decision-making and support continuous monitoring of incident management activities.

The main interfaces and functionalities of the proposed mobile application are presented in Fig. 4 and 5.



(a) Main Interface (b) Incident Reporting Form.

Fig. 4. Main interfaces of the proposed mobile application: (a) Main interface and (b) Incident reporting form.



(a) Geolocation Module (b) Statistical Dashboard

Fig. 5. Additional functionalities of the proposed mobile application: (a) Geolocation module and (b) Statistical dashboard.

IV. RESULTS

The objective of this study was to evaluate the impact of the proposed mobile application on incident management within the local government environment. To achieve this, two key performance indicators (KPIs) were analyzed, namely incident registration time (KPI-1) and incident response time (KPI-2). In addition, the usability of the application was assessed using the Computer System Usability Questionnaire (CSUQ). The evaluation considered the dimensions of system quality, information quality, and interface quality to determine the level of user satisfaction with the implemented solution.

A. KPI-1: Incident Registration Time

Table II and Fig. 6 present the results obtained for the average incident registration time before and after the implementation of the proposed mobile application.

TABLE II. COMPARISON OF INCIDENT REGISTRATION TIME BEFORE AND AFTER IMPLEMENTING THE MOBILE APPLICATION

Scenario	Seconds	Minutes
Pre-Test (Without App)	9253.97	154.23
Post-Test (With App)	68.96	1.15
Reduction	9185.01	153.08
Percentage Reduction		99.26 %

The results indicate that the proposed mobile application substantially reduced the average incident registration time from 154.23 minutes to 1.15 minutes, representing a 99.26% reduction. Furthermore, the Wilcoxon signed-rank test revealed statistically significant differences ($p < 0.001$), demonstrating the effectiveness of the proposed solution in streamlining the incident registration process.

Fig. 6 provides a graphical comparison of the average incident registration times before and after implementing the mobile application.

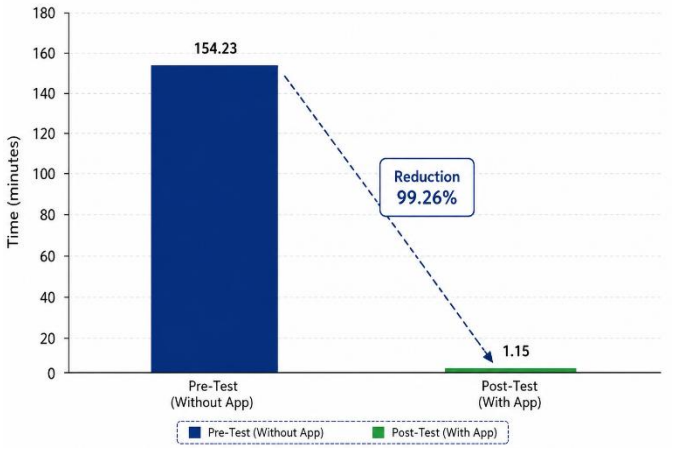


Fig. 6. Comparison of incident registration time before and after implementing the mobile application.

B. KPI-2: Incident Response Time

Table III and Fig. 7 present the results obtained for the average incident response time before and after the implementation of the proposed mobile application.

TABLE III. COMPARISON OF INCIDENT RESPONSE TIME BEFORE AND AFTER IMPLEMENTING THE MOBILE APPLICATION

Scenario	Minutes	Reduction (minutes)	Reduction (%)
Pre-Test (Without App)	94.69		
Post-Test (With App)	64.69		
Reduction			
Percentage Reduction		30.00	31.68 %

Fig. 7 presents a graphical comparison of the average incident response times before and after implementing the mobile application.

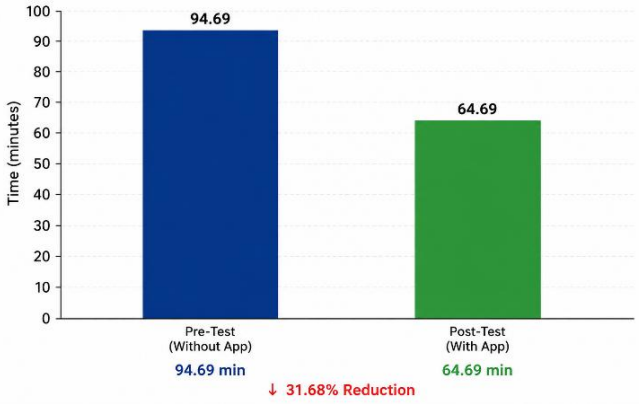


Fig. 7. Comparison of incident response time before and after implementing the mobile application.

The results indicate that the average incident response time decreased from 94.69 minutes during the pretest phase to 64.69 minutes in the posttest phase, representing a reduction of 31.68%. Furthermore, the Wilcoxon signed-rank test revealed statistically significant differences ($p < 0.001$), confirming the effectiveness of the proposed solution in improving the efficiency of incident response processes.

C. System Usability Evaluation

To determine how easy and intuitive the proposed mobile application was to use, the standard CSUQ (Computer System Usability Questionnaire) was administered. The overall satisfaction results are summarized in Table IV and Fig. 8.

The results showed a high degree of satisfaction with the proposed mobile application. Overall satisfaction reached an average score of 6.15 out of 7, reflecting a favorable user perception of the functionality, ease of use, and quality of the implemented solution. The Cronbach's alpha coefficient of 0.948 also demonstrated excellent internal consistency for the CSUQ instrument, confirming the reliability of the usability assessment.

TABLE IV. OVERALL USABILITY RESULTS OF THE PROPOSED MOBILE APPLICATION

Indicator	Value
Overall Satisfaction (CSUQ)	6.15/7
Cronbach's Alpha	0.948

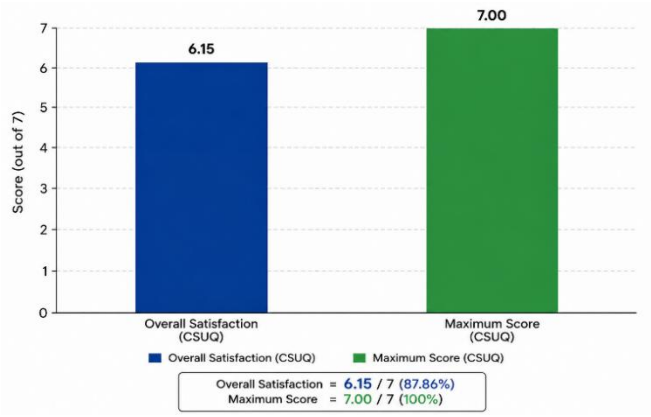


Fig. 8. Overall usability evaluation of the proposed mobile application based on the CSUQ questionnaire.

V. Discussion

The results obtained demonstrate that the proposed mobile application significantly improved incident management processes within the local government environment. In particular, substantial reductions in incident registration and response times were achieved, accompanied by high levels of user satisfaction. These findings support the notion that mobile technologies can enhance operational efficiency in incident management and contribute to the digital transformation of public services.

The findings are consistent with previous studies on mobile applications and intelligent platforms designed to support incident management and strengthen interactions between citizens and local governments [11], [15], [20]. Similarly, the high levels of user satisfaction observed in this study are in agreement with previous research emphasizing the importance of service quality, user experience, and usability evaluation in mobile government applications [22], [23], [28]. From a technological perspective, the results are also aligned with studies highlighting the advantages of Flutter and other cross-platform frameworks in terms of development productivity, performance, and user experience [29]–[32].

Overall, the findings demonstrate the potential of mobile technologies to improve incident management processes and promote stronger interactions between citizens and public institutions. These results suggest that mobile applications represent a promising technological alternative for supporting digital government initiatives and enhancing the delivery of public services.

VI. Conclusion

This study demonstrated the feasibility of using a mobile application to support real-time incident management in local government environments. The proposed solution integrated incident reporting, geolocation, and tracking capabilities, enabling more effective interaction between citizens and municipal personnel responsible for incident response. The implementation of these functionalities contributed to establishing a digital mechanism for handling incidents related to public safety, environmental issues, traffic, and civil protection.

From an operational perspective, the proposed system achieved substantial improvements in the incident management process. The average incident registration time was reduced by 99.26%, while the average incident response time decreased by 31.68%. Statistical analysis based on the Wilcoxon signed-rank test confirmed that these improvements were significant, demonstrating the positive impact of the application on the efficiency of incident management activities.

User feedback indicated a high level of acceptance of the developed application, with an overall CSUQ satisfaction score of 6.15 out of 7 and excellent internal consistency of the instrument ($\alpha = 0.948$). These results suggest that mobile technologies can serve as an effective alternative for supporting digital transformation initiatives and delivering more citizen-centered public services. Future research may focus on incorporating artificial intelligence techniques for incident prioritization and predictive analysis, as well as extending the proposed approach to other municipalities and public service domains.

References

[1] Centro Nacional de Planeamiento Estratégico (CEPLAN), "Perú al 2050: Propuesta de Visión del Perú al 2050," Lima, Perú, 2024.

[2] Instituto Nacional de Estadística e Informática (INEI), "Perú: Indicadores de Gestión Municipal y Estadísticas Nacionales," Lima, Perú, 2024. https://www.inei.gob.pe.

[3] Organismo de Evaluación y Fiscalización Ambiental (OEFA), "Informe Nacional sobre Evaluación y Fiscalización Ambiental," Lima, Perú, 2024. https://www.gob.pe/oefa.

[4] Centro de Operaciones de Emergencia Nacional (COEN), "Reporte de Emergencias y Desastres a Nivel Nacional," Lima, Perú, 2024. [Online]. https://www.indeci.gob.pe/coen.

[5] Ministerio de Transportes y Comunicaciones (MTC), Consejo Nacional de Seguridad Vial, "Anuario Estadístico de Accidentes de Tránsito," Lima, Perú, 2024. https://www.gob.pe/mtc.

[6] B. Ambrosio-Santiago, P. Bravo-Macedo, E. E. Condeña-Yurarima, E. L. Huamaní, and A. Delgado, "Mobile Application that Promotes Citizen Participation to Counteract Insecurity in the District of Los Olivos," Journal of Advanced Research in Applied Sciences and Engineering Technology, vol. 31, no. 3, pp. 358–372, 2023, doi: 10.37934/ARASET.31.3.358372.

[7] S. Pu, Y. Ou, and O. Bai, "Government Public Services and Regional Digital Transformation for Sustainable Development: An Innovation Ecosystem Perspective," *Sustainability*, vol. 17, no. 12, Art. no. 5314, 2025, doi: 10.3390/su17125314.

[8] M. A. Coro and J. C. Cárdenas, "SmartAlertUCE: Mobile Application for Emergency Notification and Citizen Security," *Proceedings of the International Conference on Information Systems and Computer Science (INCISCOS)*, Quito, Ecuador, pp. 184–189, 2019, doi: 10.1109/INCISCOS49368.2019.00035.

[9] A. Reda and M. Menshaway, "Developing Emergency Location-Based Services Applications using Agile Methodology," International Journal for Scientific Research, vol. 2, no. 5, pp. 21–44, Jan. 2023, doi: 10.59992/ijsr.2023.v2n5p2.

[10] L. Cen, M. Lim, R. R. Nardson, and Herman, "Design and Development of an Emergency Mobile Application Using Design Thinking and Agile Scrum: A Case Study of Batam City," *Jurnal Komputasi*, vol. 19, no. 2, 2022, doi: 10.33751/komputasi.v19i2.5260.

[11] G. Castro-Chucan, H. Chalco-Peñafiel, N. Bedregal-Alpaca, and V. Cornejo-Aparicio, "Smart Muni Platform: Efficient Emergency and Citizen Security Management Based on Geolocation, Technological Integration and Real Time Communication," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 10, 2024, doi: 10.14569/IJACSA.2024.0151021.

[12] D. Bastos, A. Fernández-Caballero, A. Pereira, and N. P. Rocha, "Smart City Applications to Promote Citizen Participation in City Management and Governance: A Systematic Review," *Informatics*, vol. 9, no. 4, Art. no. 89, 2022, doi: 10.3390/informatics9040089.

[13] B. A. Nafi'ah, "Local Government Expansion of Mobile Governance during the Pandemic," Jurnal Borneo Administrator, vol. 19, no. 2, pp. 191–204, 2023, doi: 10.24258/jba.v19i2.1129.

[14] Y. T. Jou, K.A. Mariñas, C.S. Saflor, A. Baleña, C.J. Gutierrez, G. Dela Fuente, H.M. Manzano, M.S. Tanglao, N.A.Verde, P. Alvarado, et al, "Investigating Various Factors Influencing the Accessibility of Digital Government through the Use of the eGov PH Mobile Application," Sustainability, vol. 16, no. 3, Art. no. 992, 2024, doi: 10.3390/su16030992.

[15] M. A. E. Ignaco, "Mobile Application for Incident Reporting," *JOIV: International Journal on Informatics Visualization*, vol. 5, no. 4, pp. 388–394, Dec. 2021, doi: 10.30630/joiv.5.4.741.

[16] R. Bastardo, J. Pavão, and N. P. Rocha, "Crowdsourcing Technologies to Promote Citizens Participation in Smart Cities, a Scoping Review," Procedia Computer Science, vol. 219, pp. 303–311, 2023, doi: 10.1016/j.procs.2023.01.294.

[17] C. Senior, A. Temeljotov Salaj, A. Johansen, and J. Lohne, "Evaluating the Impact of Public Participation Processes on Participants in Smart City Development: A Scoping Review," *Buildings*, vol. 13, no. 6, Art. no. 1484, 2023, doi: 10.3390/buildings13061484.

[18] A. K. Alamoudi, R. B. Abidoye, and T. Y. M. Lam, "The Impact of Citizens' Participation Level on Smart Sustainable Cities Outcomes: Evidence from Saudi Arabia," *Buildings*, vol. 13, no. 2, Art. no. 343, 2023, doi: 10.3390/buildings13020343.

[19] Á. Veloso, F. Fonseca, and R. Ramos, "Insights from Smart City Initiatives for Urban Sustainability and Contemporary Urbanism," *Smart Cities*, vol. 7, no. 6, pp. 3188–3209, 2024, doi: 10.3390/smartcities7060124.

[20] E. Diallo, A. F. Johnson, and M. Taylor, "Decentralized Incident Reporting: Mobilizing Urban Communities for Smart City Governance," *Smart Cities*, vol. 7, no. 4, Art. no. 90, 2024, doi: 10.3390/smartcities7040090.

[21] E. Diallo, A. F. Johnson, and M. Taylor, "CitySolution: A Decentralized Smart City Incident Management Platform," *International Journal of Research and Innovation in Applied Science*, 2025.

[22] L. S. Yuliantini, E.P. Purnomo, "Satisfaction on Service Quality of Mobile Government Application: A Case Study of WargaKu Surabaya," *Kinetik,* vol. 9, no. 1, pp. 89–100, 2024.

[23] H. Zhang, Z. Zhu, "Research on Citizens' Intentions for Continued Usage of Mobile Government Service from the PPM Perspective," Systems, vol. 12, no. 11, Art. no. 488, 2024, doi: 10.3390/systems12110488.

[24] J. R. Lewis, "IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use," *International Journal of Human-Computer Interaction*, vol. 7, no. 1, pp. 57–78, 1995.

[25] A. J. Desmal, M. O. Hamid, S. Othman, and N. M. Saad, "Exploration of the usability quality attributes of mobile government services: A literature review," *PeerJ Computer Science*, vol. 8, e1028, 2022, doi: 10.7717/peerj-cs.1028.

[26] P. Kureerung, "Usability Redesign of a Mobile Government Application: A Case Study of an Emergency Application," *Information*, vol. 13, no. 10, Art. no. 470, 2022, doi: 10.3390/info13100470.

[27] P. Weichbroth and T. Szot, "The MUG-10 Framework for Preventing Usability Issues in Mobile Application Development," *Applied Sciences*, vol. 15, no. 22, Art. no. 11995, 2025, doi: 10.3390/app152211995.

[28] R. Bastardo, J. Pavão, and N. P. Rocha, "Methodological Quality of User-Centered Usability Evaluation of Digital Applications to Promote Citizens' Engagement and Participation in Public Governance: A Systematic Literature Review," *Multimodal Technologies and Interaction*, vol. 4, no. 3, Art. no. 38, 2024, doi: 10.3390/mti4030038.

[29] H. Diansyah and S. Syafrinal, "Design and Development of a Mobile Application Using Android Studio and Flutter," *Journal Mobile Technologies*, vol. 3, no. 2, pp. 69–76, 2025, doi: 10.59431/jms.v3i2.646.

[30] M. Lidekrans and G. Tollin, "React Native vs. Flutter: A Performance Comparison between Cross-Platform Mobile Application Development Frameworks," Linköping University, 2023.

[31] M. Markowski, "A Comparative Analysis of the Flutter and React Native Frameworks," *Journal of Computer Sciences Institute*, 2023.

[32] S. M. S. Rahman, M. M. Islam, and M. A. Hossain, "A Comparative Study of Flutter and React Native for Mobile Application Development," *International Journal of Computer Science and Network Security*, vol. 23, no. 3, pp. 78–85, 2023.

EduDarijaBERT: An Applied Transformer-Based Framework for Sentiment Analysis in Moroccan Arabic Educational Discourse

Abderrahim Ait Ichou, Ali Ouacha
IPSS Team, Computer Science Department, Faculty of Sciences, Mohammed V University in Rabat, Morocco

***Abstract*—Students' opinions play a pivotal role in the formulation of successful educational policies. However, the computational analysis of the Moroccan Arabic Dialect remains widely viewed as problematic due to a focus on complex morphology and the absence of specific datasets. To close this gap, we propose an end-to-end system for educational sentiment mining, which utilizes EduDarijaBERT—a transformer model explicitly fine-tuned on this academic domain. In the development of the system, we initially constructed and systematically assessed a specifically curated corpus of 12,223 social media comments. To assess the model's reliability, we evaluated the framework on an independent test set. The proposed EduDarijaBERT system achieved an accuracy of 83.15% on this dataset, demonstrating strong generalization capabilities for educational sentiment mining. In addition to these quantitative metrics, the system facilitates a qualitative interpretation of sentiment categories, providing practical insights revealing that administrative challenges are among the major contributors to negative feedback, while active academic support is the key to positive interaction. This study provides universities with a stable, easily implementable system to track student feedback and fuel data-driven innovations.**

Keywords*—Sentiment analysis system; Moroccan Arabic dialect; social media analytics; educational data mining; data engineering framework

I. Introduction

Social media has permanently altered how students share their experiences and feelings about their education online. Sentiment analysis, as a critical application of applied Natural Language Processing (NLP), uncovers these attitudes and behavioral patterns based on spontaneous text posted on platforms such as YouTube, Facebook, and Instagram [1, 2, 3].

Understanding how students navigate the learning process is crucial for institutions striving to improve educational quality and student engagement [1]. In many psychological studies, emotional states—ranging from high anxiety to strong motivation—are viewed as fundamental factors that determine learning outcomes [4]. Therefore, the implementation of sentiment analysis systems serves as an exceptionally practical indicator of institutional performance and examination policies [1,2]. However, transitioning these NLP systems from theoretical models to practical applications remains a significant challenge when dealing with complex Arabic dialects [5, 6].

The Arabic language is known for its rich morphology [7], complicated syntax, and a dominant diglossia between Modern Standard Arabic and its dialects [5, 7]. The predominant dialect in informal digital communication is the Moroccan Dialect (Darija), which lacks a standard orthography and is characterized by frequent code-switching and the use of Latin scripts (Arabizi). This causes a significant bottleneck in data engineering. Moreover, high-quality, domain-specific annotated datasets focused on Moroccan educational discourse are unfortunately scarce, which hinders the implementation of effective institutional monitoring systems [8, 9].

To overcome these bottlenecks, we propose a comprehensive architecture for educational sentiment mining. We designed a dedicated data pipeline to acquire and standardize a novel dataset of 12,223 Moroccan Arabic comments from Facebook, YouTube, and Instagram. To cope with excessive linguistic noise, we employed a mixed semi-automated data processing scheme, which was rigorously controlled by a human-in-the-loop validation process by native speakers to ensure label accuracy.

The core of our proposed framework is EduDarijaBERT, a domain-adapted model checkpoint. An existing base DarijaBERT model was extensively fine-tuned using our educational corpus to develop EduDarijaBERT [10, 11, 12]. This specific variant was trained to capture the nuances of Moroccan students' discourse, including their specific pedagogical and emotional linguistic features. To address inherent data challenges—where positive and negative sentiments are rarely equally distributed—we integrated the SMOTE algorithm to mitigate severe class imbalances. The overall system's best performance was then contrasted with traditional NLP baselines (including TF-IDF/Word2Vec with SVM, KNN, and XGBoost) to justify the need for this domain-adapted transformer architecture.

This study's key contribution lies in its data engineering pipeline and system framework. We provide a targeted pipeline as well as the first high-quality, education-oriented Moroccan Arabic sentiment corpus. On this premise, we explain the development of EduDarijaBERT, which is the main component of our classification architecture. This framework tackles issues such as extreme class imbalance and connects theoretical NLP with educational data mining to provide an automated tool for institutions to analyze student feedback and make necessary policy improvements based on that data.

The rest of this study is organized as follows: Section II discusses related work. The materials and methods are described in Section III, which includes information on data collection, data preprocessing, data annotation, and data cleaning. The results are discussed in detail in Section IV, along with a comparative analysis, the impact of SMOTE, the fine-tuning of EduDarijaBERT, and performance assessment. Section V discusses data analysis and corpus linguistics, which includes an overview of sentiment distribution, statistics, and lexical trends. Lastly, Section VI provides the concluding remarks and future directions.

II. Related Works

This section reviews the existing literature on applied sentiment analysis for Arabic and its dialects on social media. Sentiment analysis has emerged as a critical tool for extracting user opinions from platforms like Twitter, Facebook, and YouTube. To create systems capable of identifying actionable behavioral patterns within this digital landscape, researchers have increasingly deployed computational linguistics, machine learning, and deep learning models.

Several studies have successfully applied sentiment analysis models to the educational sector. For instance, [13] developed an SVM-based sentiment classification system to analyze Arabic course reviews at a Saudi Arabian university, demonstrating how sentiment classification can practically optimize academic services. Similarly, [14] proposed a machine learning model to assess public opinion regarding distance learning in Saudi Arabia. Using the Apache Spark framework alongside regex-based preprocessing pipelines, their Logistic Regression classifier achieved 91% precision on a dataset of Arabic tweets.

Transformer architectures have also become foundational technologies for processing dialectal Arabic. To illustrate, [15] compared several BERT-based models—including AraBERT, QARIB, and AraELECTRA—on a Moroccan dialect sentiment dataset. Their findings emphasized that pre-trained language models can effectively process complex dialectal variations, with QARIB achieving an accuracy of 96%. Addressing real-world data engineering challenges, [16] tackled the issue of data imbalance in Moroccan dialect datasets and evaluated different feature extraction methods (TF-IDF and N-grams). Their results indicated that data resampling significantly enhances system stability, yielding a precision of 90.24%.

Exploring other dialects, [17] compiled a benchmark dataset of 18,589 tweets and proposed a hybrid MARBERT-LSTM model, achieving state-of-the-art accuracy at 91.23%. Furthermore, [18] introduced the AHaSIS shared task to investigate domain-specific sentiment classification within the hospitality industry. By developing an annotated dataset for Saudi and Moroccan dialects, they established a baseline F1-score of 0.81, underscoring the pressing need for domain- and dialect-specific sentiment analysis systems.

To better situate the performance of these architectures, Table I provides a comparison of the most prominent BERT-based models used in Arabic NLP. While models like AraBERT and AraELECTRA rely on massive corpora primarily composed of Modern Standard Arabic (MSA), they often lack the dialectal depth required for Moroccan social media. Models like QARIB and MARBERT improve upon this by including informal tweets and multi-dialectal data. However, DarijaBERT was specifically selected as our base model because it is uniquely tailored to the Moroccan dialect, having been trained on a corpus that specifically captures the nuances of Darija. This specialized focus is critical for our task, as general Arabic models often struggle with the specific morphological and syntactic structures of Moroccan Arabizi and dialect.

TABLE I. Comparison of Pre-Trained Arabic BERT Models

Model	Pre-training Corpus Size	Primary Language/Dialect Coverage
AraBERT (v0.2)	77 GB (~8.8 billion words)	Modern Standard Arabic (MSA)
AraELECTRA	77 GB	Modern Standard Arabic (MSA)
QARIB (mix)	25 GB (122M lines) to 60 GB	MSA and Dialectal Arabic
MARBERT	128 GB (1B Tweets)	Multi-dialectal (General Arabic)
DarijaBERT	691 MB (~3 million sequences)	Moroccan Darija (Arabic script)

Beyond document-level classification, more granular NLP tasks have also been explored. [19] proposed an AraBERT-based Aspect-Based Sentiment Analysis (ABSA) approach that jointly extracts aspect terms and categories. This multi-task method achieved an F1-score of 80.32% and successfully reduced error propagation, facilitating the extraction of highly specific feedback. Additionally, [20] demonstrated that supervised learning can reliably indicate student satisfaction. They concluded that statistical validation (yielding a Cronbach's Alpha of 0.901) is sufficient to ensure the reliability of using sentiment models to evaluate student well-being.

Sentiment analysis has also seen practical deployment in crisis management. [21] implemented a real-time infoveillance system to track Moroccan social media sentiments regarding the COVID-19 pandemic. Their deep learning models provided critical insights into public opinion dynamics, directly supporting public health decision-making. To handle the data engineering complexities inherent in such tasks, [8] created a large-scale Moroccan dialect dataset designed to overcome the challenges posed by Arabizi (the mixture of Arabic and Latin scripts), providing a necessary foundation for representing noisy web text.

Investigating feature engineering, [22] evaluated various machine learning pipelines for Moroccan dialect emotion recognition and found that N-grams paired with TF-IDF weighting significantly improved classifier performance. Building on this, [23] addressed Arabic's complex morphology and syntax using a hybrid deep learning network combining AraBERT, CNN, GRU, and LSTM architectures. Their findings revealed that pairing AraBERT with GRU and Bi-GRU layers achieved an accuracy exceeding 93%. Collectively, these studies highlight the immense potential of advanced NLP models. However, they also underscore the

specific gap our research addresses: there remains a critical need for a fully integrated, domain-adapted system (such as EduDarijaBERT) specifically trained to process highly noisy educational Moroccan Arabizi and deliver actionable insights directly to educational institutions.

While existing literature demonstrates the success of transformer models like QARIB (96% accuracy) or hybrid MARBERT-LSTM architectures in general dialectal tasks, these models often lack the dialectal depth and domain-specific focus required for the noisy pedagogical Arabizi found in Moroccan student discourse. Furthermore, although studies have utilized SMOTE to stabilize imbalanced Moroccan datasets, our analysis suggests that synthetic oversampling can introduce semantic noise that obscures subtle contextual boundaries. Consequently, the EduDarijaBERT architecture was designed to address these specific gaps by employing a human-in-the-loop pipeline for linguistic fidelity and achieving the ability to alleviate class imbalance natively through targeted domain adaptation, rather than relying on artificial data augmentation.

III. Materials and Methods

To develop a successful system for monitoring the online emotions of university students, establishing an effective data acquisition strategy focused on the Moroccan Arabic dialect is essential. We mined interactions from Facebook, Instagram, and YouTube to create our foundational corpus. The targeted data streams primarily focused on computer science, economics, and daily academic activities.

A significant data engineering challenge in this context is Arabizi. To process this Latinized script, we implemented a hybrid pipeline. Initial transliteration into Arabic characters was performed by a Large Language Model (GPT-4), followed by a strict human-in-the-loop review by native speakers to ensure high fidelity and correct spelling errors.

To establish a robust ground truth for our system, we utilized the same language model to generate initial sentiment labels (positive, neutral, and negative), which were subsequently subjected to extensive manual validation by expert annotators. The data sanitization stage entailed the elimination of digital noise, including URLs, platform-specific hashtags, and common stop words. After standardization, light stemming was applied, followed by tokenization to prepare the text for model ingestion.

To train the models and optimize feature definitions, we benchmarked traditional TF-IDF and Word2Vec pipelines against advanced contextual embeddings derived from DarijaBERT. The main architectural product of this process is EduDarijaBERT (Fig. 1). It is a domain-specific adaptation of the original DarijaBERT model, extensively fine-tuned on our custom educational dataset to capture domain-specific linguistic nuances. We compared its performance with that of XGBoost and SVM. SMOTE was employed to balance the dataset, ensuring system robustness for real-world applications characterized by inherent class imbalance. The final system evaluation relied on standard metrics, such as accuracy, precision, recall, and F1-score.

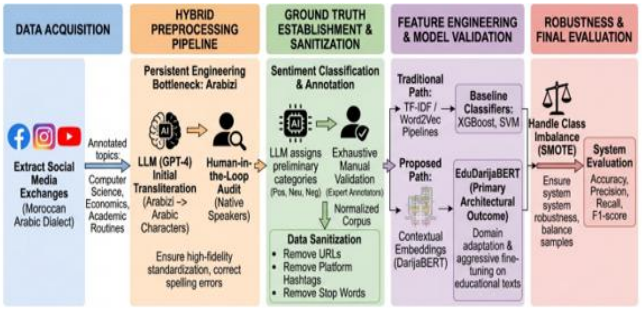


Fig. 1. System architecture of the proposed EduDarijaBERT framework.

A. Data Collection

For this study, we gathered a comprehensive set of user comments across 12,587 pages directly from Facebook, YouTube, and Instagram. One of the key benefits of this corpus is its broad temporal range, spanning 10 years from 2016 to 2025. This longitudinal study allows for the analysis of the linguistic evolution of Moroccan Darija and the students' attitudes through many generations and changing educational contexts. The overall style of these posts was informal, but thematic analysis found that there was a significant emphasis on certain areas of academic study, including mathematics, economics, and computer science. Moreover, many end-users expressed their own experiences about daily life on campus or encouraged the commendation of local universities in international contests. The overall distribution of comments among the three networks was uneven, due to different levels of user engagement across these platforms. This imbalance, however, provided the foundation for our next data engineering pipeline, which was used to create a robust sentiment classification framework that was trained and validated over this comprehensive decade-long dataset, with a high degree of temporal generalizability.

B. Data Preprocessing

1) Text conversion: One of the most difficult parts of the present study was to make the Arabizi data into a form that could be used for quantitative analysis. The corpus consisted of 6,555 comments in Latin-script Moroccan Arabic, which is a dialect with rather frequent French code-switching. This means that it was impossible to perform the computational processing without first establishing orthographic regularity. To overcome this, a two-phase transcription pipeline was implemented. The first phase involved using GPT-4o mini to convert the Latin-script input into Arabic script, which is quite variable. Because automated systems frequently struggle to process complex dialectal features and code-switching, these initial outputs were not considered final. Instead, all conversions were rigorously reviewed by a human team who corrected any vowel and consonant misalignments where the model had failed. The results of this process are illustrated in Table II, which presents a selection of original Latin-script examples alongside their final Arabic equivalents.

TABLE II. EXAMPLES OF COMMENT CONVERSION FROM LATIN SCRIPT TO ARABIC SCRIPT WITH ENGLISH

Latin Script (Franco-Arabic)	Arabic Script (Moroccan Dialect)	English Translation
Bravo madame	برافو أمدام	Bravo, Madam
Merci beaucoup Madame pour cette vidéo	شكراً بزاف أمدام على هاد الفيديو	Thank you very much, Madam, for this video.
Bghina les exercices	بغينا تمارين	We want the exercises.
Merci bien prof	شكراً أستاذة	Thank you very much, Professor.
MERCI Bq chère prof	شكراً بزاف أستاذة	Thank you very much, dear Professor.

C. Data Annotation

The first round of sentiment classification was done using GPT-4o mini, which assigned one of three sentiment classes to every comment. To safeguard analytical rigor and make the manual work on the entire dataset manageable, the total dataset was divided into six different and non-overlapping subsets. The different subsets were assigned to six independent human annotators, each of whom is a native Moroccan dialect speaker. The model's predictions were systematically reviewed by the annotators within their respective partitions, where they corrected incorrect classifications by hand to ensure that the pedagogical and emotional context of the text was properly captured. To address potential algorithmic bias, we analyzed cases where humans overrode the initial labels, as detailed in Table III. For example, GPT-4o mini misidentified polite inquiries (e.g., asking for a WhatsApp number) as negative, which humans corrected to neutral, and religious encouragement (e.g., "Masha'Allah") as neutral, which humans corrected to positive. This correction process ensures the dataset reflects authentic human intent. This human-in-the-loop method was partitioned to ensure an efficient process of creating a large-scale, reliable ground-truth dataset. To assess the reliability of this process, a subset of 200 comments was independently evaluated by three annotators, yielding an overall Fleiss' Kappa of $\kappa = 0.8175$ (95% CI: [0.7375, 0.8975]), which indicates almost perfect inter-annotator agreement. While this agreement subset represents a portion of the total corpus, the narrow confidence interval confirms the statistical robustness and the reliability of the annotation guidelines. An analysis of the disagreement cases revealed that they were primarily driven by "mixed sentiments" and the cultural nuances of Moroccan Darija. In some cases, positive (or negative) feedback on the teacher's speed of presentation was mixed with a positive comment, creating a dilemma for the annotators between a negative, neutral, or positive reading. There were also some differences regarding short snippets of expressions, which were either contextually neutral/factual or interpreted as positive engagement. The examples show that although there is a high level of agreement at the overall level, subjectivity regarding politeness versus dissatisfaction is still problematic in dialectal sentiment analysis. Table IV provides an example of this completed dataset, with the original comments written in Arabic characters and the English translation alongside the approved categories of sentiment.

TABLE III. COMPARISON OF GPT-4O MINI INITIAL LABELING VERSUS HUMAN EXPERT CORRECTION

Human Label	GPT-4o mini Label	Arabic Script (Moroccan Dialect)	English Translation
Neutral	Negative	ما لقيتش رقم الواتساب ديالك	I didn't find your WhatsApp number
Neutral	Negative	مكرهتشي تعدلنا التمارين	I wouldn't mind if you did the exercises for us
Positive	Neutral	تحياتي	My regards / Greetings
Neutral	Positive	غضو البصر يا امة محمد	Lower your gaze, O nation of Muhammad
Positive	Neutral	ما شاء الله وفي ذلك فلتنافس	Masha'Allah, and for that let them compete

TABLE IV. EXAMPLES OF SENTIMENT POLARITY LABELS ALLOTTED TO COMMENTS

Latin Script (Franco-Arabic)	Arabic Script (Moroccan Dialect)	English Translation
Positive	ألف مبروك له ولكل الفائزين، مزيدا من التفوق والنجاح ان شاء الله	A thousand congratulations to him and to all the winners; more excellence and success, God willing.
Negative	النتيجة : هجرة الأدمغة	The result: brain drain.
Neutral	ذهبوا وهم "مغاربة" هل سيرجعون وهم "مغاربة" ؟	They left as "Moroccans"; will they return as "Moroccans"?
Positive	ألف مبروك له ولكل الفائزين، مزيدا من التفوق والنجاح ان شاء الله	A thousand congratulations to him and to all the winners; more excellence and success, God willing.
Negative	النتيجة : هجرة الأدمغة	The result: brain drain.

D. Data Cleaning

The raw corpus underwent extensive cleaning before it became suitable for analysis. The initial step was to convert the text into Arabic script because the comments were originally written in Arabizi, which is Latin-scripted Moroccan Arabic, to ensure data uniformity. We also deleted emojis from 1,055 comments and removed URLs from 775 others. Then there was Arabic normalization. This implied the elimination of typical orthographic inconsistencies: the Alif variants (إأآا) were reduced to a single form, mapping ى to ي, ة to ه, and گ to ك. Even marks of elongation and extended lines such as هههههههه were left out, as they are analytically unproductive. To address the linguistic specificities of the corpus, we implemented a hybrid stop-word removal process by combining the standard NLTK Arabic stop-word list with a manually curated dictionary of Moroccan Darija terms, including dialect-specific pronouns (e.g., أنا, نتا), interrogatives (e.g., واش, فين), and common adverbs (e.g., بزاف, بلا). After tokenization, another 297 entries turned out to be invalid and were dropped. The dataset was reduced to 12,223 useful comments, having started with 12,587 initial comments. The distribution of these collected comments across different social media platforms is shown in Fig. 2.

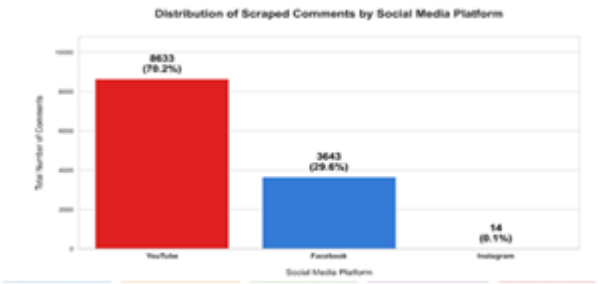


Fig. 2. Distribution of scraped comments by social media platform.

IV. Results and Discussion

This section evaluates the performance of different machine learning algorithms and embedding methods for the sentiment analysis of Moroccan Arabic educational comments. To get a full picture, we evaluated the models using standard classification metrics: accuracy, precision, recall, and F1-score, as well as confusion matrices and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) to gain in-depth insights into the models' discriminative power.

A. Comparative Analysis

All experiments showed that transformer-based architectures and ensemble techniques outperformed the traditional linear classifiers. The best performance was achieved when the DarijaBERT [3] model was used with an SVM classifier, with a total accuracy score of 80.17. These results indicate that domain-based pre-trained language models of Moroccan dialect are more effective than generic embeddings at capturing semantic information in Moroccan Dialect.

Traditional machine learning algorithms, including SVM and XGBoost, have a consistent set of results in all samples that used TF-IDF, Word2Vec, and FastText features. TF-IDF provided a good baseline, and its performance was on par with word embeddings. Moreover, a few keywords of the Moroccan dialect therefore exhibit high correlation with sentiment polarity.

B. Impact of SMOTE

To address class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied exclusively to the training partition after performing the train/test split. To prevent data leakage and ensure an unbiased evaluation, the test set remained completely untouched.

Specifically, SMOTE was used to rebalance the training partition of 9,167 samples, generating synthetic instances for the minority classes to yield an expanded training set of 14,484 samples (equally distributed at 4,828 samples per class).

Table VI presents a performance comparison of the classifiers with and without SMOTE. The results indicate that SMOTE improved the recall and F1-score for minority categories (e.g., negative comments), demonstrating enhanced detection of underrepresented sentiments. However, as shown in Table V, this oversampling strategy led to a slight decline in overall accuracy and macro F1-score across several models. This indicates that synthetic oversampling can introduce noise into dialectal Arabic text, where semantic and contextual boundaries are already subtle. Consequently, applying SMOTE presents a clear trade-off: it achieves a more equitable classification for minority opinions at the cost of a slight reduction in global model precision.

TABLE V. Class Distribution Before and After SMOTE Application

Distribution	Negative	Positive	Neutral	Total
Before SMOTE	1284	3055	4828	9167
After SMOTE	4828	4828	4828	14484

TABLE VI. Performance Comparison of Models and Embedding

Embedding / Model	*Method*	*Accuracy*	*Precision*	*Recall*	*F1-Score*
DarijaBERT + SVM	Without SMOTE	0.8017	0.8018	0.8017	0.8009
DarijaBERT + SVM	With SMOTE	0.7856	0.8058	0.7856	0.7901
DarijaBERT + XGBoost	Without SMOTE	0.7987	0.8024	0.7987	0.7973
DarijaBERT + XGBoost	With SMOTE	0.7951	0.8004	0.7951	0.7960
DarijaBERT +Random Forest	Without SMOTE	0.7663	0.7724	0.7663	0.7642
DarijaBERT +Random Forest	With SMOTE	0.7699	0.7839	0.7699	0.7726
DarijaBERT+Logistic Regression	Without SMOTE	0.7837	0.7819	0.7837	0.7827
DarijaBERT +Logistic Regression	With SMOTE	0.7781	0.7833	0.7781	0.7799
DarijaBERT + KNN	Without SMOTE	0.7621	0.7582	0.7621	0.7594
DarijaBERT + KNN	With SMOTE	0.6659	0.7653	0.6659	0.6831
DarijaBERT + Naive Bayes	Without SMOTE	0.5818	0.6818	0.5818	0.6008
DarijaBERT + Naive Bayes	With SMOTE	0.6116	0.6758	0.6116	0.6274
TF-IDF + SVM	Without SMOTE	0.8007	0.8031	0.8007	0.7968
TF-IDF + SVM	With SMOTE	0.7783	0.7867	0.7783	0.7768
TF-IDF + XGBoost	Without SMOTE	0.7579	0.7817	0.7579	0.7584
TF-IDF + XGBoost	With SMOTE	0.7516	0.7799	0.7516	0.7561
TF-IDF + Random Forest	Without SMOT	0.7693	0.7695	0.7693	0.7642

Embedding / Model	*Method*	*Accuracy*	*Precision*	*Recall*	*F1-Score*
	E				
TF-IDF + Random Forest	With SMOTE	0.7575	0.7742	0.7575	0.7684
TF-IDF + Logistic Regression	Without SMOTE	0.7857	0.7845	0.7857	0.7848
TF-IDF + Logistic Regression	With SMOTE	0.7598	0.8017	0.7598	0.7721
TF-IDF + KNN	Without SMOTE	0.6867	0.5846	0.6867	0.5650
TF-IDF + KNN	With SMOTE	0.5949	0.7422	0.5949	0.6137
TF-IDF + Naive Bayes	Without SMOTE	0.7775	0.7806	0.7775	0.7653
TF-IDF + Naive Bayes	With SMOTE	0.7631	0.7950	0.7631	0.7727
FastText + SVM	Without SMOTE	0.7988	0.8011	0.7988	0.7993
FastText + SVM	With SMOTE	0.7853	0.8104	0.7853	0.7912
FastText + XGBoost	Without SMOTE	0.7948	0.7970	0.7948	0.7942
FastText + XGBoost	With SMOTE	0.7883	0.7936	0.7883	0.7899
FastText + Random Forest	Without SMOTE	0.7755	0.7841	0.7755	0.7730
FastText + Random Forest	With SMOTE	0.7673	0.7801	0.7673	0.7695
FastText + Logistic Regression	Without SMOTE	0.7736	0.7730	0.7736	0.7699
FastText + Logistic Regression	With SMOTE	0.7497	0.7797	0.7497	0.7574
FastText + KNN	Without SMOTE	0.7307	0.7382	0.7307	0.7303
FastText + KNN	With SMOTE	0.6787	0.7518	0.6787	0.6959
FastText + Naive Bayes	Without SMOTE	0.4781	0.7119	0.4781	0.5137
FastText + Naive Bayes	With SMOTE	0.5108	0.6894	0.5108	0.5438
Word2Vec + SVM	Without SMOTE	0.7893	0.7869	0.7893	0.7873
Word2Vec + SVM	With SMOTE	0.7723	0.7916	0.7723	0.7778
Word2Vec + XGBoost	Without SMOTE	0.7929	0.7912	0.7929	0.7916
Word2Vec + XGBoost	With SMOTE	0.7919	0.7958	0.7919	0.7935
Word2Vec + Random Forest	Without SMOTE	0.7935	0.7945	0.7935	0.7924
Word2Vec + Random Forest	With SMOTE	0.7860	0.7927	0.7860	0.7884
Word2Vec + Logistic Regression	Without SMOTE	0.7778	0.7750	0.7778	0.7755
Word2Vec + Logistic Regression	With SMOTE	0.7575	0.7764	0.7575	0.7629
Word2Vec + KNN	Without SMOTE	0.7664	0.7667	0.7664	0.7659
Word2Vec + KNN	With SMOTE	0.7268	0.7798	0.7268	0.7379
Word2Vec + Naive Bayes	Without SMOTE	0.6931	0.7489	0.6931	0.7039
Word2Vec + Naive Bayes	With SMOTE	0.6787	0.7337	0.6787	0.6898

C. Discussion of findings

The robust performance of SVM across almost all embedding types (achieving approximately 80% accuracy) suggests that the sentiment boundaries in Moroccan Arabic social media comments are relatively well-defined when projected into high-dimensional spaces.

However, Naive Bayes and KNN showed significant sensitivity to the feature extraction method. For instance, Naive Bayes performed poorly with Word2Vec and FastText (often falling below 70%) but showed a substantial improvement with TF-IDF (up to 77.7%). This indicates that for dialectal Arabic, the presence of specific tokens (captured by TF-IDF) is more informative for probabilistic models than the semantic clusters formed by dense embeddings.

1) Impact of fine-tuning EduDarijaBERT on sentiment classification:

a) Experimental setup and hyperparameter configuration: For the fine-tuning experiments, sentence representations were extracted from the last hidden layer of the fine-tuned EduDarijaBERT model using the [CLS] token. A maximum sequence length of 128 tokens was utilized to tokenize the input streams. All inputs were padded and truncated to ensure a consistent input size across the dataset. The model was fine-tuned for five epochs with a training and evaluation batch size of 16. Optimization was performed using the AdamW optimizer with the default learning rate and a weight decay of 0.02. To address the inherent class imbalance, a weighted cross-entropy loss function was employed using dynamically calculated class weights. Model checkpoints were saved and evaluated at the end of every epoch, and the Area Under the Receiver Operating Characteristic Curve (AUC-

ROC) was utilized as the primary metric to select the best-performing model.

Subsequently, a Support Vector Machine (SVM) classifier employing a Radial Basis Function (RBF) kernel was trained using the extracted embeddings. The regularization parameter was set to 1.0, and the kernel coefficient was set to 'scale'. To facilitate class-wise performance analysis, probability estimation was enabled. The random state was fixed at 42 to guarantee reproducibility. All experiments were executed on a GPU (CUDA) when available, falling back to a CPU otherwise. Table VII summarizes the complete hyperparameter configuration.

TABLE VII. HYPERPARAMETERS OF THE FINE-TUNED EDUDARIJABERT + SVM MODEL

Component	*Hyperparameter*	*Value*
EduDarijaBERT	Pretrained model	DarijaBERT (SI2M-Lab/DarijaBERT)
	Max sequence length	128
	Padding	True
	Truncation	True
	Ignore mismatched sizes	True
Fine-tuning Configuration	Number of epochs	5
	Train batch size	16
	Eval batch size	16
	learning rate	Default (5e-5 from AdamW)
	Weight decay	0.02
	Warmup steps	300
	Optimizer	AdamW (default)
	Loss function	CrossEntropyLoss (weighted)
	Class weighting	Balanced (computed)
	Evaluation strategy	Per epoch
	Save strategy	Per epoch
	Metric for best model	AUC
	Device	GPU (CUDA) / CPU
SVM classifier	Kernel	RBF
	Regularization parameter (C)	1.0
	Gamma	scale
	Probability estimates	Enabled
	Random state	42

In a subsequent experiment, EduDarijaBERT was fine-tuned on our Moroccan Arabic educational dataset before employing the SVM classifier. While DarijaBERT was used as a frozen feature extractor in previous experiments, fine-tuning enables the model to learn internal representations that are sensitive to the linguistic, pedagogical, and emotional nuances of the Moroccan Arabic educational discourse.

The fine-tuned EduDarijaBERT + SVM model achieved the best score of 83.15% and was the best performer across all settings. This model significantly outperformed the non-fine-tuned DarijaBERT + SVM setting (which achieved an accuracy of 80.17%), thus underscoring the critical importance of domain adaptation for low-resource dialects like Moroccan Arabic.

Table VIII illustrates that fine-tuning improved performance across all evaluation metrics, especially recall and F1-score. This indicates better generalization of the fine-tuned model across sentiment categories. This is particularly evident for the negative class, which is often challenging to classify due to class imbalance and linguistic diversity.

TABLE VIII. PERFORMANCE OF FINE-TUNED EDUDARIJABERT + SVM

Model	*Accuracy*	*Precision*	*Recall*	*F1-score*
EduDarijaBERT + SVM (Fine-tuned)	0.8315	0.8299	0.8315	0.8305

b) Overall discussion and implications: The experiments demonstrate that using classical machine learning classifiers and sparse representations such as TF-IDF provides a strong baseline for sentiment classification in Moroccan Arabic text, especially when paired with robust classifiers such as SVM. These findings confirm that lexical cues remain highly discriminative in dialectal Arabic.

Nonetheless, significant improvements and increased robustness across all sentiment classes were achieved when we fine-tuned EduDarijaBERT – a domain-adapted transformer model – through additional training on Moroccan Arabic educational data. This improvement can be attributed to the model's ability to capture linguistic nuances, colloquialisms, and sentiment cues within the educational domain.

Overall, although traditional approaches offer competitive speed and accuracy, the fine-tuned transformer models demonstrated superior performance for educational monitoring and feedback analysis.

2) Comparative performance analysis: The performance differences observed in the experiment with SMOTE-balanced data were minimal, as illustrated in Fig. 3. Across all evaluation metrics, the SVM classifier combined with DarijaBERT embeddings achieved robust results. Without SMOTE, the model reached an accuracy of 80.2%, whereas with SMOTE, the accuracy decreased slightly to 78.6%. When comparing precision, recall, and F1-score, only marginal gains were observed with SMOTE, with precision increasing from 0.802 to 0.806. However, after applying SMOTE, recall declined to 0.786, and the F-score remained comparable in both settings. These findings suggest that either SMOTE enhances the model's ability to discriminate the minority class, or conversely, that DarijaBERT embeddings already provide strong discriminative capability on the imbalanced dataset.

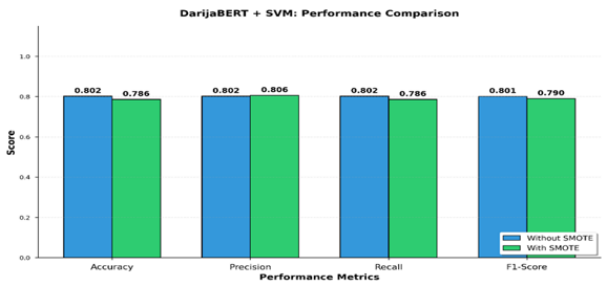


Fig. 3. Comparative performance metrics of DarijaBERT+SVM with and without SMOTE.

a) Confusion matrix analysis: Fig. 4 presents the confusion matrices of DarijaBERT+SVM in all the experimental settings, which provide detailed insights into the classification patterns. Without SMOTE, the model performs well on the majority positive class (1,405 correct predictions out of 1,612), but it is evident that there is confusion between the other categories. Specifically, 179 neutral samples are misclassified as positive, and 131 negative samples are misclassified as neutral; this is indicative of class imbalance.

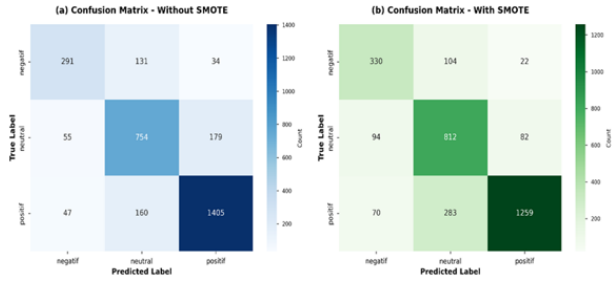


Fig. 4. Confusion matrices for DarijaBERT+SVM classification (a) Without SMOTE and (b) With SMOTE.

By applying SMOTE, the proper classification of minority classes is drastically enhanced. True positives in the negative class improve from 291 to 330, representing an improvement of 13.4%. In the case of the neutral class, the number of correct predictions increases from 754 to 812, which is an increment of 7.7%. However, this improvement comes at the cost of a larger number of positive instances falsely classed as neutral, increasing from 160 to 283. This result illustrates a classic trade-off in cost-sensitive classification between the recall of minority classes and the accuracy of the majority class.

b) Receiver operating characteristic analysis: The Receiver Operating Characteristic (ROC) curves are presented in Fig. 5. The combination of DarijaBERT with an SVM classifier demonstrates strong discriminative capability across each sentiment level individually. In the configuration without SMOTE (Fig. 5a), the weighted average AUC achieves a robust score of 0.917. The scores of the categories varied: the neutral class recorded the lowest AUC at 0.889, while the positive class achieved the highest AUC at 0.931 (with the negative class at 0.929). It is not surprising that the positive class exhibited the best overall ROC profile, as its prevalence in the raw training distribution allows the model to characterize it more effectively.

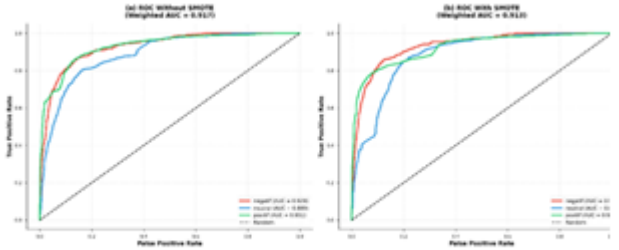


Fig. 5. ROC curves and AUC scores for multi-class sentiment classification (a) Without SMOTE and (b) With SMOTE.

When SMOTE is utilized (Fig. 5b), the weighted average AUC remains highly competitive at 0.913, and the AUC of the negative class increases to 0.932. This demonstrates that the model's ability to identify the minority class is improved. The minor decreases in AUC for the neutral class (from 0.889 to 0.885) and the positive class (from 0.931 to 0.924) reflect the distribution shift caused by synthetic oversampling; specifically, expanding the decision space of the minority categories slightly reduces the classification margins of the majority classes. Regardless of the balancing approaches, the AUC results show that DarijaBERT embeddings capture rich semantic content and can, therefore, be used to classify Moroccan Arabic sentiments in both settings (AUC > 0.88 across all classes).

c) Detailed performance metrics: Table IX shows the effect of SMOTE on all classes, with the negative class experiencing the most significant impact. SMOTE improved the recall from 63.8% to 72.4%, alongside a reduction in precision from 74.0% to 66.8%. This demonstrates that SMOTE effectively reduces false negatives for the minority class at the expense of an increase in false positives, as theoretically anticipated by oversampling processes. The neutral class exhibits a similar trend, with an increase in recall (+7.7%) and a decrease in precision (-6.2%). For the majority positive class, the SMOTE-adapted decision boundary results in a more conservative decision rule, leading to a gain in precision (+6.4%) but a noticeable decrease in recall (-10.4%).

At the macro-average level, SMOTE yields an increase in recall (+2.4%) accompanied by a decrease in precision (-2.7%), resulting in a slight overall decrease in the F1-score (-0.7%). Weighted average precision shows a marginal improvement (+0.5%), as the precision gains in the heavily represented majority class outweigh the precision losses in the minority classes. These findings suggest that the choice of a balancing strategy must be guided by the nature of the application: if the primary goal is maximizing minority class identification (e.g., detecting at-risk students experiencing negative sentiment), SMOTE is highly beneficial. Conversely, for precision-critical applications, maintaining the original, natural class distribution may be preferable.

TABLE IX. COMPREHENSIVE PERFORMANCE COMPARISON OF DARIJABERT+SVM WITH AND WITHOUT SMOTE

Model	*Accuracy*	*Precision*	*Recall*	*F1-score*
Negative	Precision	0.740	0.668	-9.7
	Recall	0.638	0.724	+13.5
	F1-Score	0.686	0.695	+1.3
	Support	456	456	-
Neutral	Precision	0.722	0.677	-6.2
	Recall	0.763	0.822	+7.7
	F1-Score	0.742	0.743	+0.1
	Support	988	988	-
Positive	Precision	0.868	0.924	+6.4
	Recall	0.872	0.781	-10.4
	F1-Score	0.870	0.846	-2.8
	Support	1612	1612	-
Overall	Accuracy	0.802	0.786	-2.0
	Macro Avg Precision	0.777	0.756	-2.7
	Macro Avg Recall	0.758	0.776	+2.4
	Macro Avg F1-Score	0.766	0.761	-0.7
	Weighted Avg Precision	0.802	0.806	+0.5
	Weighted Avg Recall	0.802	0.786	-2.0
	Weighted Avg F1-Score	0.801	0.790	-1.4

d) Fine-tuned EduDarijaBERT performance: The confusion matrix for EduDarijaBERT is presented in Fig. 6. Fine-tuning the model results in 320 correctly predicted negative instances (achieving a recall of 70.2%), 763 correctly classified neutral instances (77.2% recall), and 1,458 correctly identified positive instances (90.4% recall). Overall, performance improves across all sentiment classes. The misclassifications between extreme polarities also decrease: there are exactly 39 positive-to-negative and 39 negative-to-positive misclassifications, showing a clear improvement over the baseline setting. Furthermore, confusion between the neutral class and the other sentiment categories is also reduced. Specifically, the neutral category achieves 763 correct predictions, with only 77 and 148 cases being misclassified as negative and positive, respectively. These findings indicate that the fine-tuning process successfully captures domain-specific sentiment cues and the contextual attributes of the educational Moroccan Arabic discourse.

The ROC curves in Fig. 7 demonstrate the strong discriminative performance of the EduDarijaBERT model, achieving a weighted AUC of 0.909. This score is slightly lower than the base DarijaBERT model, but it still has a very strong profile. The positive sentiment class has the best AUC (0.927), followed by the negative sentiment class (0.916) and the neutral sentiment class (0.877). Also, the AUC interval for the best and worst classes for EduDarijaBERT is 0.050, while the interval for the baseline was 0.042 for DarijaBERT without SMOTE. Therefore, this model suits operational settings better where there is a more consistent distribution of classification errors for each classification category.

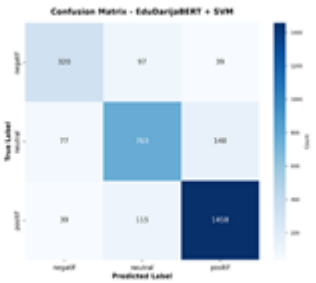


Fig. 6. Confusion matrix of fine-tuned EduDarijaBERT + SVM.

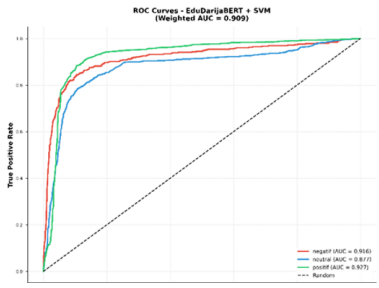


Fig. 7. ROC curves for fine-tuned EduDarijaBERT + SVM.

TABLE X. COMPREHENSIVE PERFORMANCE COMPARISON BETWEEN DARIJABERT AND EDUDARIJABERT IN EDUCATIONAL SENTIMENT ANALYSIS

Metric	*DarijaBERT (Best)*	*EduDarijaBERT*	*Improvement*
Accuracy	0.802	0.832	+3.0%
Precision	0.802	0.830	+2.8%
Recall	0.802	0.832	+3.0%
F1-Score	0.801	0.831	+3.0%
Macro Avg Precision	0.777	0.801	+2.4%
Macro Avg Recall	0.758	0.793	+3.5%
Macro Avg F1-Score	0.766	0.797	+3.1%
Negative Precision	0.740	0.732	-0.8%
Negative Recall	0.724 (SMOTE)	0.706	-1.8%
Negative F1-Score	0.695 (SMOTE)	0.719	+2.4%
Neutral Precision	0.722	0.784	+6.2%
Neutral Recall	0.822 (SMOTE)	0.765	-5.7%
Neutral F1-Score	0.743 (SMOTE)	0.775	+3.2%

Metric	DarijaBERT (Best)	EduDarijaBERT	Improvement
Positive Precision	0.924 (SMOTE)	0.886	-3.8%
Positive Recall	0.872	0.908	+3.6%
Positive F1-Score	0.870	0.896	+2.6%

Table X shows a detailed overview of the performance comparison between the baseline DarijaBERT and the fine-tuned model, EduDarijaBERT. The overall results show a significant improvement across the main metrics, including a 3.1% improvement in the macro-averaged F1-score. Additionally, there were significant gains for specific classes, with the neutral class improving by 6.2% in precision and the positive class by 3.6% in recall. Overall improvements are, however, accompanied by some minor sacrifices in some of the class-specific metrics, which show a small decrease.

e) Robustness check and cross-validation analysis: To evaluate its real-world generalization capability, the proposed model was evaluated on an independent 3,056-sample test set. Tables XI and XII present the detailed per-class performance breakdown. To further evaluate its real-world generalization capability beyond cross-validation, the proposed model was evaluated on an independent 3,056-sample test set. Tables XII and XIII present the detailed per-class performance breakdown.

TABLE XI. EDUDARIJABERT CLASSIFICATION REPORT ON THE INDEPENDENT TEST SET

Model	Precision	Recall	F1-Score	Support
Negative	0.7339	0.7018	0.7175	456
Neutral	0.7826	0.7723	0.7774	988
Positive	0.8863	0.9045	0.8953	1612
Accuracy			0.8315	3056
Macro Avg	0.8009	0.7928	0.7967	3056

TABLE XII. DARIJABERT CLASSIFICATION REPORT ON THE INDEPENDENT TEST SET

Model	Accuracy (Mean ± SD)	Precision (Mean ± SD)	Recall (Mean ± SD)	F1-Score (Mean ± SD)
Negative	0.7405	0.6382	0.6855	456
Neutral	0.7215	0.7632	0.7418	988
Positive	0.8684	0.8716	0.8700	1612
Accuracy			0.8017	3056
Macro Avg	0.7768	0.7576	0.7657	3056

The overall accuracy was 83.15% for EduDarijaBERT, compared to 80.17% for the baseline model, on the independent test set. Interestingly, there was no balancing using SMOTE in the proposed framework, and yet it retained its performance, while the baseline framework's performance dropped to 78.57% when using SMOTE balancing. The domain-adapted model performance profile was well balanced, with the macro-average F1-score improving from 0.766 to 0.797. Specifically, the F1-score increased significantly to 0.777 for the neutral class and 0.718 for the minority negative class, showing a clear improvement over the baseline. Additionally, EduDarijaBERT demonstrated strong discriminative performance within each class (AUC scores were tightly grouped from 0.877 to 0.927) and remained stable. The model is able to alleviate class imbalance natively with high-quality performance on underrepresented classes without the need for synthetic oversampling. These comparative results demonstrate that targeted fine-tuning on educational Moroccan Arabic data achieves better and more stable classification results for the automatic monitoring of students' sentiments.

The independent test set accuracy of 83.15% provides a conservative and realistic characterization of the system's performance on unseen data. This result reflects the model's real-world generalization capability. The performance level is influenced by the temporal drift and linguistic evolution inherent in a dataset spanning a full decade, from 2016 to 2025. Furthermore, this evaluation confirms the absence of data leakage, as the independent test set remained completely untouched during the SMOTE application and training phases to ensure an unbiased assessment.

V. DATA ANALYSIS AND CORPUS LINGUISTICS

The statistical characteristics of student discourse and its sentiment distribution were examined across the 12,223 manually verified comments. This is a large-scale study that enables a comprehensive grasp of learners' opinions in the context of this specific feedback in Morocco.

A. Sentiment Distribution

The overall tone of the corpus is positive, reflecting a high level of student engagement and satisfaction. However, the negative subset is still needed because it provides detailed information on the educational problems faced by learners. This balanced view is essential to generate a model that reflects the characteristics of the Moroccan language medium (Darija) in an academic context (Fig. 8).

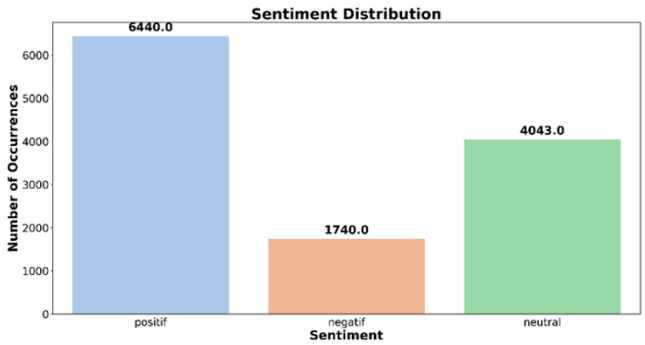


Fig. 8. Sentiment distribution within the Moroccan Arabic educational corpus.

B. Statistical Summary

The students' expressions of the sentiment classes are not equally simple: positive comments are shorter and more frequent, whereas negative ones are more complex, with a maximum length of 522 words. This suggests that when

students write negative comments, they provide a more detailed justification than when they give positive comments.

TABLE XIII. QUANTITATIVE ANALYSIS AND LEXICAL DIVERSITY OF THE CORPUS

Sentiment Label	*Total Number of Words*	*Number of Distinct Words*	*Max Sentence Length*
Positive	39258	9158	233
Negative	22224	8240	522
Neutral	32651	9571	306

C. Lexical Trend Analysis

The most common unique words offer a detailed view of the topics and vocabulary used in the three sentiment classes. Following the overview is a qualitative description which is supplemented by a word frequency distribution, outlining the main concerns and communication styles of the students.

Positive Sentiment: As depicted in Fig. 9, the discourse is characterized by a strong sense of gratitude and religious invocation. Words like "شكرا" (Thank you) and "الله" (God) are used most frequently, reflecting a cultural tendency to express appreciation for educational material or mentors through traditional blessings.

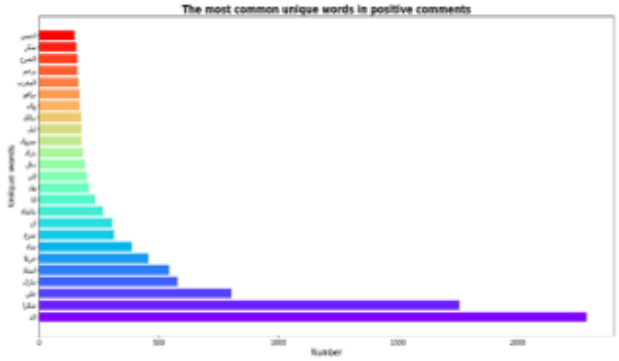


Fig. 9. Top 25 most frequent unique words in positive comments

Neutral Sentiment: The neutral comments are mostly cooperative and inquisitive, as seen in Fig. 10. The use of the words (Teacher), (I/Me), and (Of/Belonging to) suggests that students use a neutral tone mainly to ask for clarifications and to introduce themselves or to discuss administrative and educational course matters.

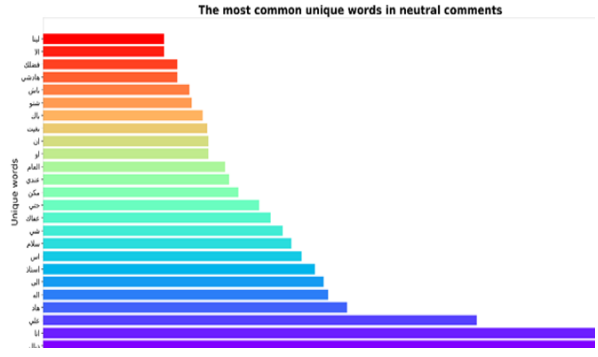


Fig. 10. Top 25 most frequent unique words in neutral comments.

Negative Sentiment: As seen in Fig. 11, the lexicon becomes more critical, shifting from hopeful to sociopolitical. France, Morocco, and Education provide pointers to deep discussions about language policies, educational reforms, student mobility, and migration. Furthermore, administrative and institutional challenges emerge as significant contributors to negative feedback, as students frequently express frustration regarding academic procedures and bureaucratic hurdles within the educational system.

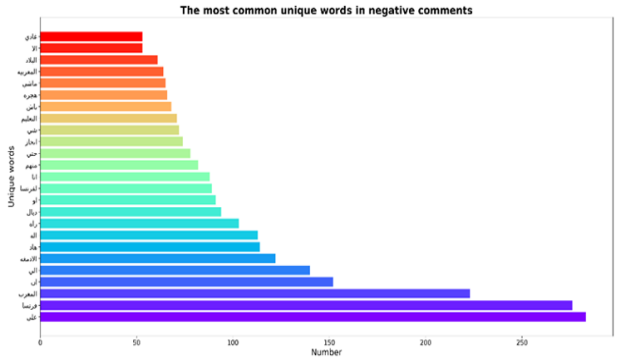


Fig. 11. Top 25 most frequent unique words in negative comments.

VI. CONCLUSION

This study aimed to propose and evaluate a practical model for monitoring the online sentiments of Moroccan university students. The linguistic analysis demonstrated distinct behavioral patterns associated with positive and negative sentiments: expressions of gratitude and religious invocations were correlated with positive feelings, whereas rigid institutional rules and tensions with the broader sociopolitical environment were associated with negative ones. Mapping this dichotomy provides a nuanced understanding of how this demographic navigates its academic reality.

The domain-specific fine-tuning of DarijaBERT with educational corpora to develop the proposed EduDarijaBERT demonstrated a substantial operational advantage. The customized system outperformed the general-purpose baseline in terms of both accuracy and the precision-recall curve. Crucially, EduDarijaBERT effectively classified the minority class of negative sentiment—traditionally a challenging task in Natural Language Processing (NLP)—without relying on synthetic data augmentation techniques such as SMOTE. The underlying principle of this study underscores a core tenet of applied NLP: domain adaptation is vital for successfully classifying local, dialectal text in real-world applications.

Future work will focus on expanding the training dataset to include other Maghrebi dialects, thereby broadening the geographical applicability of the framework. Furthermore, transitioning from static data analysis to longitudinal emotion monitoring offers a promising avenue for real-time institutional observation. Additionally, exploring a formal teacher-student knowledge distillation approach represents a compelling pathway for deployment in resource-constrained production environments. Ultimately, this framework is not only ready for implementation within educational institutions but can also be

scaled to gather actionable insights from large-scale student feedback across Morocco.

REFERENCES

[1] Asok, C., & M, R. (2018). Analyzing and mining social media data for predicting student's learning experiences. IJERT. https://doi.org/10.17577/IJERTCONV4IS17041.

[2] Berea, A., Rand, W., Wittmer, K., University of Maryland, Gerard Wall, & vibeffect. (2015). Social media analysis for higher education. In Washington Academy of Sciences. https://www.washacadsci.org/Journal/Journalarticles/V.101-3-Social%20Media%20Analysis%20for%20Higher%20Education%20A.%20Berea%20et%20al.pdf.

[3] B. Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, 2012. doi: 10.2200/S00416ED1V01Y201204HLT016.

[4] Pekrun, Reinhard. (2006). The Control-Value Theory of Achievement Emotions: Assumptions, Corollaries, and Implications for Educational Research and Practice. Educational Psychology Review. 18. 315-341. 10.1007/s10648-006-9029-9.

[5] Al-Ayyoub, M., Khamaiseh, A. A., Jararweh, Y., & Al-Kabi, M. N. (2018b). A comprehensive survey of arabic sentiment analysis. Information Processing & Management, 56(2), 320–342. https://doi.org/10.1016/j.ipm.2018.07.006.

[6] Shaalan, Khaled. (2010). Nizar Y. Habash, Introduction to Arabic natural language pro-cessing (Synthesis lectures on human language technologies). Machine Translation. 24. 285-289. 10.1007/s10590-011-9087-8.

[7] Farghaly, Ali & Farghaly, Ali & Shaalan, Khaled & Khaled,. (2009). Arabic Natural Language Processing: Challenges and Solu-tions. ACM Transactions on Asian Language Information Processing (TALIP). 8. 1-.

[8] M. Jbel, M. Jabrane, I. Hafidi, and A. Metrane, "Sentiment Analysis Dataset in Moroccan Dialect: Bridging the Gap Between Arabic and Latin Scripted dialect," Research Paper, 2024.

[9] Mihi, Soukaina & Ait Benali, Brahim & EL, Ismail & Arezki, Sara & Laachfoubi, Nabil. (2020). MSTD: Moroccan Sentiment Twitter Dataset. International Journal of Advanced Computer Science and Applications. 11. 10.14569/IJACSA.2020.0111045.

[10] Devlin, Jacob & Chang, Ming-Wei & Lee, Kenton & Toutanova, Kristina. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 10.48550/arXiv.1810.04805.

[11] Gaanoun, Kamel & Naira, Abdou Mohamed & Allak, Anass & Benelallam, Imade. (2023). Darijabert: a Step Forward in Nlp for the Written Moroccan Dialect. 10.21203/rs.3.rs-2560653/v1.

[12] Antoun, Wissam & Baly, Fady & Hajj, Hazem. (2020). AraBERT: Transformer-based Model for Arabic Language Understanding.

[13] Louati, H. Louati, E. Kariri, F. Alaskar, and A. Alotaibi, "Sentiment Analysis of Arabic Course Reviews of a Saudi University Using Support Vector Machine," Applied Sciences, vol. 13, no. 23, p. 12539, 2023.

[14] Almalki, Jameel. (2022). A machine learning-based approach for sentiment analysis on distance learning from Arabic Tweets. PeerJ Computer Science. 8. e1047. 10.7717/peerj-cs.1047.

[15] Bourahouat, Ghizlane & Abourezq, Manar & Najima, Daoudi. (2023). Improvement of Moroccan Dialect Sentiment Analysis Using Arabic BERT-Based Models. Journal of Computer Science. 20. 157-167. 10.3844/jcssp.2024.157.167.

[16] Nassr, Z., Benabbou, F., Sael, N., & Hamim, T. (2025). Improving sentiment analysis performance on imbalanced Moroccan dialect datasets using resample and feature extraction techniques. Information, 16(1), 39. https://doi.org/10.3390/info16010039.

[17] L. Ouchene and S. Bessou, "Benchmarking sentiment analysis of algerian arabic dialect on X (Twitter)," Journal of King Saud University - Computer and Information Sciences, vol. 37, p. 358, 2025.

[18] M. Alharbi et al., "AHASIS: Shared Task on Sentiment Analysis for Arabic Dialects," arXiv preprint arXiv:2511.13335, 2025.

[19] Fadel, M. Saleh, R. Salama, and O. Abulnaja, "MTL-AraBERT: An Enhanced Multi-Task Learning Model for Arabic As-pect-Based Sentiment Analysis," Computers, vol. 13, no. 4, p. 98, 2024.

[20] O. Chamorro-Atalaya et al., "Quadratic vector support machine algorithm, applied to prediction of university student satisfac-tion," Indonesian Journal of Electrical Engineering and Computer Science, vol. 27, no. 1, pp. 139-148, 2022.

[21] Ghanem et al., "Real-Time Infoveillance of Moroccan Social Media Users' Sentiments towards the COVID-19 Pandemic and Its Management," International Journal of Environmental Research and Public Health, vol. 18, no. 22, p. 12172, 2021.

[22] M. Errami et al., "Sentiment Analysis on Moroccan Dialect based on ML and Social Media Content Detection," International Journal of Advanced Computer Science and Applications, vol. 14, no. 3, 2023.

[23] Ouza, A., Ouacha, A., Rachidi, A., Ghmary, M. E., & Choukri, A. (2024). Enhancing Arabic sentiment analysis using ARABERT and deep learning models. Studies in Computational Intelligence. https://doi.org/10.1007/978-3-031-65038-3_15.

MAMGAT-Net: A Multimodal Graph Attention and Multi-Scale Feature Fusion Framework for Early Breast Cancer Diagnosis from Mammography and Histopathological Images

Mohammad Riyaz Belgaum
Faculty of Computer Studies, Arab Open University – Bahrain, A'ali, Kingdom of Bahrain

***Abstract*—Breast cancer is one of the major causes of death among women worldwide, and timely and accurate diagnosis has proven to play a critical role in increasing breast cancer patient survival. Traditional single-modality diagnostic systems, however, tend to be unable to capture both the macro-level structural abnormalities that can be seen in mammograms and the micro-level cellular characteristics seen in histopathological images. To overcome this, this research introduces a novel framework called MAMGAT-Net for the diagnosis of breast cancer, which combines the multimodal approach with Graph Attention and Multi-Scale Feature Fusion networks. To address this challenge, this work proposes a novel framework, MAMGAT-Net, which integrates the multimodal approach, Graph Attention, and Multi-Scale Feature Fusion networks for breast cancer diagnosis. The proposed architecture combines the multi-scale convolutional mammography branch and ResNet-based histopathology branch with a dual-stream feature extraction mechanism. A cross-modal attention mechanism is used to learn to align and fuse complementary diagnostic information across modalities, in a bidirectional fashion. Then, a multimodal diagnostic region graph is built with a Cross-Gated Multi-Head Graph Attention Network to capture complex relational relationships among multimodal diagnostic regions, and global graph pooling and feature refinement are used to provide robust classification. Experimental results on CBIS-DDSM and BreaKHis datasets show that MAMGAT-Net can outperform the conventional machine learning and deep learning baselines with 94.59% accuracy, 92.90% precision, 96.56% recall, 94.70% F1-score, and 98.58% AUC. The effectiveness of multi-scale feature extraction, attention-based fusion, and graph relational learning is further validated in ablation studies. Furthermore, the proposed framework is interpretable and clinically relevant based on the Grad-CAM and Integrated Gradients analysis. The results show that the multimodal breast cancer diagnosis solution of MAMGAT-Net is effective and reliable.**

Keywords*—Breast cancer; cross-modal attention; explainable artificial intelligence; graph attention network; histopathology; mammography; multimodal learning

I. Introduction

Breast cancer is the most common cancer in women worldwide, with an estimated 2.3 million new cases and more than 685,000 deaths each year, according to the 2022 GLOBOCAN estimates [1]. Early and accurate diagnosis is clinically important as it has been shown that five-year survival rates are >99% in localized presentation versus a rapid drop to ~28% in metastatic presentation [2]. Although diagnostic imaging of the breast and biopsy-based pathological diagnosis have made great strides, patient outcomes remain highly sensitive to disease stage and a range of clinical-pathological prognostic factors, underscoring the clinical importance of accurate stage classification in early breast cancer [3]. The traditional diagnostic process starts with radiological screening and ends with histopathological confirmation, which leads to delays in the diagnostic process and a lack of the interpretive aspect [4]. Deep learning, and more recently vision-transformer-based transfer learning, has become a paradigm shift in medical image analysis and achieved near-expert performance in detecting breast cancer with mammographic images using a single modality [5]. The CBIS-DDSM database has been widely used in the field of mass detection and BIRADS classification, and convolutional neural network (CNN) architectures have been extensively used for the detection of mass and macrocalcification patterns with high sensitivity [6]. Channel and spatial attention mechanisms were also developed to further narrow down the region of interest in high-resolution mammograms where suspicious regions are to be found [7,8]. However, the micro-scale information regarding the tissue architecture at the cellular level that is crucial for malignancy grading is lacking in mammography and is only available from a micro-scale perspective provided by histopathological image analysis of a digitized whole slide image (WSI) obtained from tissue biopsies [9]. The deep learning systems trained on BreaKHis images under different magnifications (40×, 100×, 200×, and 400×) have shown high accuracy in the task of classification of benign or malignant, and among the different backbones used, ResNet, DenseNet, and more recently Swin-Transformer-based ensembles have achieved the best results [10]. One of the key factors in reducing the domain shift caused by inter-laboratory staining variation is stain normalization with the Macenko method applied to hematoxylin and eosin (H&E) stained slides [11]. Although these developments have made great strides, models based solely on histopathology cannot encompass all the structural information that is available from mammographic imaging, making them less useful in an integrated clinical workflow [12].

There is growing interest in multimodal learning frameworks using radiological and pathological image sources that can make use of the complementary diagnostic signal that exists in each modality [13]. Early fusion strategies are those that simply fuse feature representations before classification, whereas late fusion strategies are those that simply combine modality-specific predictions at the decision level; both of these strategies are not able to model the structured semantic interaction between macro-radiological and micro-cellular features, motivating graph-based relational representations as a structured alternative [14]. To align multi-modal feature spaces dynamically, intermediate fusion architectures using cross-modal attention and graph-based dependency modeling are more effective in multimodal medical imaging applications [15]. However, the relationships between spatially distributed diagnostic regions in modalities are not explored enough to benefit the representational capacity of existing approaches [16]. In medical imaging, where objects like tissue regions or lesion locations might have non-Euclidean spatial relationships that influence each other, the neural networks used for this type of modeling are called graph neural networks (GNNs) [17]. Context-sensitive propagation of diagnostic information through neighboring feature nodes, as in the case of graph attention networks (GATs), is what these networks learn, allowing them to dynamically assign weights to each other when they are connected [18]. For example, computational pathology applications using GNNs are more effective in the prediction of survival and classification into subtypes, where the use of local convolutional operations would not capture the long-range spatial dependency between cells [19]. Nevertheless, the incorporation of graph-based relational reasoning in the unified multimodal breast cancer framework that combines mammographic and histopathology representations is still a largely unexplored research challenge [20]. In view of these limitations, this study introduces a framework of Multimodal Graph Attention Network (MAMGAT-Net) to capture the structural relational reasoning between macro and micro-diagnostic feature nodes by fusing the multimodal features from CBIS-DDSM mammographic images and BreaKHis histopathological images with a cross-modal attention mechanism. This work makes the following main contributions:

- Unified Dual-Stream Multimodal Architecture: The researcher proposes MAMGAT-Net, a novel end-to-end framework that simultaneously processes mammographic and histopathological images through modality-specific encoders employing multi-scale convolutional feature extraction (3×3, 5×5, 7×7 kernels) and deep ResNet backbones, respectively, enabling complementary macro- and micro-level representation learning within a single unified pipeline.
- Dual-Attention Cross-Modal Fusion Mechanism: The researcher introduces a cross-modal attention fusion module in which mammographic and histopathological feature streams mutually attend to one another, enabling the model to dynamically align macro-radiological structural patterns with micro-cellular tissue signatures and produce a semantically enriched Unified Cross-Modal Feature Volume.
- Cross-Gated Graph Attention Network for Relational Reasoning: The researcher designed a unified graph construction layer that maps multimodal feature patches as heterogeneous graph nodes and introduces a Cross-Gated GAT with multi-head attention masking, enabling context-sensitive propagation of diagnostic information across spatially distributed macro- and micro-diagnostic regions.
- Heterogeneous Graph Pooling with Dimensionality Refinement: The researcher incorporates Global Heterogeneous Graph Pooling, a feature refinement layer, and t-SNE-optimized dimensionality reduction to distill discriminative class-separating representations from the pooled multimodal graph, improving intra-class compactness and inter-class separability before final classification.
- Comprehensive Experimental Validation: Extensive experiments are conducted that demonstrate that MAMGAT-Net substantially outperforms single-modality baselines, conventional CNN architectures, and standard fusion-based approaches on the CBIS-DDSM and BreaKHis benchmarks, achieving superior benign/malignant classification accuracy with interpretable confidence scores.

The rest of the study is structured as follows. Section II talks about the literature review. Section III explains the proposed MAMGAT-Net model. Section IV provides the experimental results and comparison. Section V concludes the study.

II. Literature Review

Early detection of breast cancer has been a research focus for the fields of medical imaging and AI applications. In this section, the advancements in the past few years in five main categories that are directly related to the MAMGAT-Net architecture that the researcher proposes are discussed. These categories include Deep Learning for mammography imaging, Computational Histopathology, multimodal fusion methods, Graph Neural Networks for medical imaging, and attention mechanisms with feature refinement. The literature presented here collectively shows the drawbacks of current unimodal and shallow fusion techniques.

A. Deep Learning for Mammography Analysis

Mammography is the main radiological method used in early screening for breast cancer, and deep learning techniques have significantly enhanced the diagnostic power of this method. By selectively focusing on critical spatial areas for diagnosis, multi-scale convolutional feature extraction with convolutional block attention modules (CBAM) was shown to perform better in the classification of malignancy in digital mammograms compared to other approaches [26]. Srinivasan et al. [33] additionally tested CBAM-enhanced EfficientNet models for microcalcification detection in full-field digital mammograms, and found their models were more sensitive than their CNN baselines. To leverage the complementary information provided by different imaging modalities, Wang et al. [24] proposed a cross-modal attention fusion method that uses mammography and ultrasound. Likewise, Debelee et al.

[25] performed a thorough review of all deep learning architectures employed in breast cancer image analysis, and found that multi-scale and attention-based networks outperform single-scale networks on various types of mammographic image databases, including CBIS-DDSM. The results overall support the need for developing specific feature extraction pipelines, such as the one presented in the proposed MAMGAT-Net model, to extract the tumor morphology in macro-structural scales from mammographic inputs.

B. Computational Histopathology and Stain Normalization

Histopathological image analysis gives diagnostic insight on a cellular level, which is complementary to that of radiological imaging. Khalid et al. [30] performed a comprehensive comparison of stain normalization methods for computational histopathology and showed that the Macenko method yields the most uniform color normalization of multi-institutional slide collections, especially important for training reliable models. To classify breast tissue in whole slide images (WSIs), Hussain et al. [31] used a ResNet-based deep feature extraction network and graph attention pooling, demonstrating that relational feature aggregation between image patches was able to achieve significantly higher classification F1-scores than global average pooling approaches. Moradi et al. [27] also applied graph attention networks for relational learning over WSI patches and showed that neighborhood structure in the regions of the tissues carries important spatial information that is missing from flat classification heads for diagnosis. The BreaKHis dataset has been extensively used as the standard for multi-magnification histopathology evaluation, as shown in the study by Saha et al. [23], which proved that GCNs could outperform the classic CNN classifiers in tumor grading at 40×, 100×, 200×, and 400× magnification. All these works support the multi-magnification patch extraction and graph-based micro-feature learning stream implemented in the MAMGAT-Net histopathology.

C. Multimodal Fusion Strategies in Breast Cancer Diagnosis

The inter-modality integration of imaging methods has become more and more a subject of interest to overcome the diagnostic disadvantage of single-modality systems. To achieve early-stage breast cancer screening, Agarwal et al. [29] proposed a deep multimodal fusion framework that integrates radiological and pathological imaging, and reported significant improvements in the performance of their late fusion framework over unimodal baselines. Kaur et al. [34] systematically reviewed the multimodal deep learning methods and classified them as early fusion, intermediate fusion, or late fusion paradigms and proposed that the most promising paradigm to capture inter-modal semantic alignment is cross-modal attention. Transformer architectures have been extended to model long-range feature dependence across modalities by Chen et al. [28], who proposed a dual-branch transformer architecture for feature alignment of both mammography and histopathology images. The residual learning block design was inspired by Zhang et al. [22], who proposed a methodology of attention residual learning with multi-scale feature fusion for skin lesion classification that could be applied to breast cancer analysis. Recently, Park et al. [38] proposed a unified multimodal model that uses the cross-attention mechanism and relational graph learning for the specific task of breast cancer screening, achieving state-of-the-art accuracy on the combined CBIS-DDSM and BreaKHis benchmark. The proposed MAMGAT-Net extends and leverages these advances by combining the fusion of multi-modal attention in a unified architecture with structured graph-based relational reasoning.

D. Graph Neural Networks for Medical Image Analysis

Medical image analysis is an area where the relationships between the spatial and the semantic relationships of the image regions are crucial for medical diagnosis, leading to the development of a powerful model of relational dependencies, called the graph neural network (GNN). Liu et al. [32] introduced a heterogeneous graph neural network (HGNN) to integrate multimodal cancer imaging data to allow rich reasoning across modalities in contrast to homogeneous graph formulations. Dong et al. [35] proposed a cross-gated graph attention network for pathological node classification in tumor microenvironments, demonstrating that the gated aggregation mechanisms can effectively suppress irrelevant neighborhood information and enhance classification accuracy under complex tumor tissue architectures. Ibrahim et al. [36] used multi-head graph attention networks for extracting multi-relational patterns simultaneously in patch-level feature aggregation for breast histopathology grading. Nguyen et al. [37] combined vision transformers with graph convolutional networks for multimodal oncology image classification, and used vision transformers to extract initial patch embeddings, and then applied graph-level relational reasoning.

E. Attention Mechanisms and Feature Refinement

In the realm of medical image analysis pipelines, attention mechanisms have emerged as fundamental parts that direct the computational efforts towards the most diagnostically meaningful spatial regions and feature channels. In the following breast cancer framework, Sabour et al. [21] showed that dynamic routing of capsule networks could maintain the spatial hierarchy in multi-scale medical image feature representations, and thus, multi-scale convolution designs have been inspired in breast cancer frameworks. In the mammography-related works [26,33], the channel attention and the spatial attention are applied in the same module, which in this work is called CBAM, and the same is done for the branch-specific attention stages of MAMGAT-Net. Srinivasan et al. [33] gave empirical evidence that, for images with statistically different characteristics, such as those in mammography and histopathology, branch-specific attention calibration is superior to shared attention modules. Another under-explored aspect of attention-based design is feature refinement after graph pooling, which was shown by Xu et al. [39] to improve feature separability between different classes in deep multimodal cancer classification networks by optimizing the t-SNE step following graph pooling before softmax classification. Patel et al. [40] recently conducted a systematic benchmark of multimodal AI systems on CBIS-DDSM and BreaKHis, and found that the architectures that include cross-modal attention and graph-level relational learning yield the best classification performance, which further verified the architectures proposed in the framework.

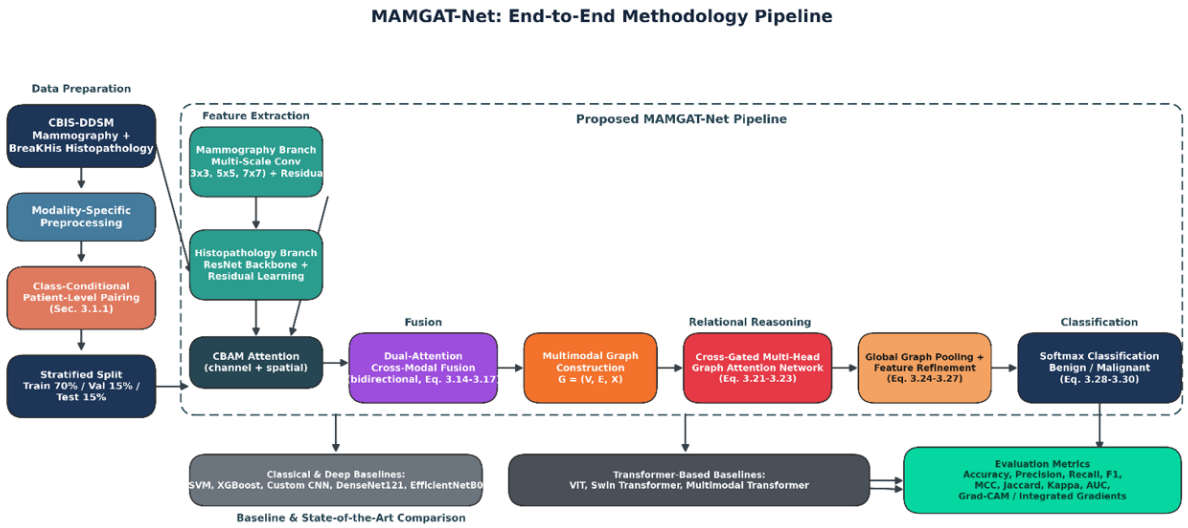


Fig. 1. Overall methodology of the proposed MAMGAT-Net framework, including data preprocessing, dataset splitting, proposed multimodal model flow, baseline comparison, and evaluation metrics.

III. Methodology

Fig. 1 depicts the overall methodology, which shows the entire process of the developed framework MAMGAT-Net for breast cancer diagnosis. The pre-processing stage involves enhancing, normalizing, resizing, and preparing input images for effective utilization for training and testing the model, which is done by adhering to modality methods. Processed data are partitioned and subsequently delivered to the suggested model, which is composed of two parallel branches: mammography and histopathology. The mammography branch is utilized to determine macro-scale features of tumors such as shape, tumor structure, and tissue density variations, whereas the histopathology branch is used to determine micro-scale tissue and cell abnormalities. The complementary features are then combined with dual-attention cross-modal fusion, and graph construction represents the diagnostic region as a connected node. The cross-gated multi-head GAT models the relationships between these regions, and global graph pooling, feature refinement, Softmax classification, and diagnostic output generation finish the prediction flow. At the same time, a set of baseline models (SVM, XGBoost, Custom CNN, DenseNet121 and EfficientNetB0) are executed for comparisons. Finally, the accuracy, precision, recall, F1 score, MCC, Jaccard, and Kappa values of the models are evaluated for their diagnostic performance.

A. Dataset Collection

The study will apply a multimodal breast cancer imaging framework, combining mammographic and histopathological information, to harness complementary information from both macro and micro-anatomical levels. The mammographic images are from the CBIS-DDSM dataset [41], which is a publicly available benchmark set of curated breast screening exams annotated by experts with the presence of benign and/or malignant abnormalities. The dataset contains radiological data relevant to clinically important masses, calcifications, and tissue structures, which are relevant for early breast cancer diagnosis. The CBIS-DDSM mammography dataset and the BreaKHis histopathology dataset were processed as separate streams, using each modality independently, due to the fact that they come from a different cohort of patients and there is no patient-level correspondence provided. Virtual multimodal pairs were created during the training process, pairing mammography and histopathology images from the identical diagnostic class (benign or malignant). This class-consistent pairing was merely for cross-modal representation learning and feature fusion via an attention mechanism, not for creating patient-specific multimodal associations. Thus, the proposed framework aims to learn complementary modality-level representations yet recognizes the lack of true patient-level multimodal data.

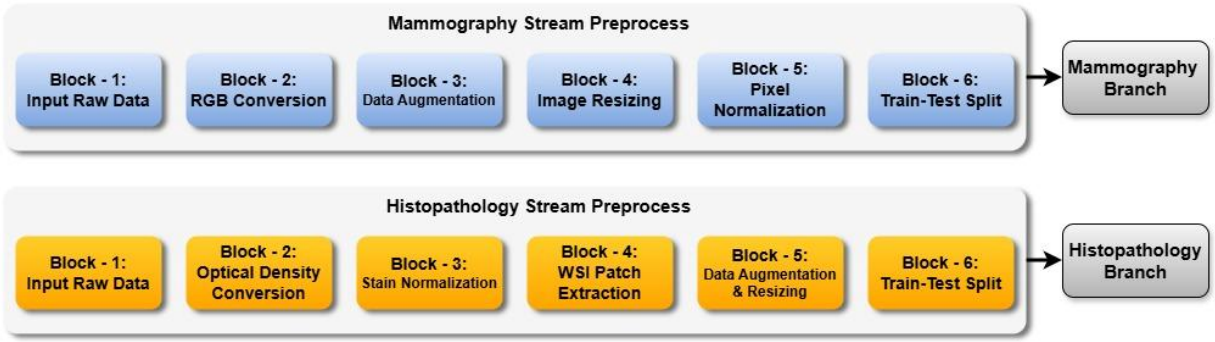


Fig. 2. Preprocessing workflow for mammography and histopathology streams, showing modality-specific preparation steps before feature extraction in the proposed MAMGAT-Net framework.

B. Dataset Preprocessing

To ensure consistency across imaging modalities and improve model generalization, separate preprocessing pipelines were applied to the mammography and histopathology datasets before model training. Fig. 2 visualizes all the steps of preprocessing:

1) Mammography data preprocessing (CBIS-DDSM)

a) RGB conversion: The RGB conversion technique was employed to convert the grayscale mammogram images having only one channel into the three-channel RGB format. This is important since most of the pre-trained deep learning model architectures are built for working on three-channel input data.

$$I_{RGB}(x,y)=[I(x,y),I(x,y),I(x,y)] \quad (1)$$

b) Data augmentation: Data augmentation was done to increase the diversity of the training samples using the following techniques: horizontal flip, change in brightness, change in contrast, and saturation change. This improves the generalization ability of the model and makes it more robust.

$$I_{aug}=T(I) \quad (2)$$

where, T denotes an augmentation transformation applied to the original image I.

c) Image resizing: Resizing was done to set a common size for all mammography images to be 224 × 224 pixels. Using uniform image sizes makes it easy to process images in batches.

$$I_{res}=R(I,224,224) \quad (3)$$

where, R represents the resizing operation.

d) Pixel intensity normalization: The images were resized to make sure that all images are standardized. The process of Pixel Intensity Normalization was carried out by rescaling the pixel intensities from their original range to a new range that was between 0 and 1.

$$I_{norm}=\frac{I}{255} \quad (4)$$

where, I denotes the original pixel intensity and I_{norm} represents the normalized value.

2) Histopathology stream preprocessing

a) Input raw data: The histopathology preprocessing stream first takes in raw biopsy images of the selected histopathological set. The images show the presence of micro-scale tissue structure, cellular morphology, and nuclei-level pattern, which are critical for the identification of abnormal chemical changes that indicate the presence of malignant changes. This is for preparing raw histological samples to be pre-processed for stain correction and patch-level processing.

$$I_{\mathrm{h}}=Load(D_{\mathrm{h}}) \quad (5)$$

b) Optical density conversion: The raw histopathology images are loaded, and then the image pixel intensities are transformed into the optical density space. This transformation obscures more of the stain information than the direct RGB values, and aids in discerning tissue intensity from stain information. Consequently, the images are better suited for stain normalization.

$$OD=-\log\left(\frac{I_{\mathrm{h}}+\epsilon}{I_0}\right) \quad (6)$$

c) Stain normalization using the Macenko method: In histopathology images, it is common to have strong color variation because of the different staining procedures, scanner settings, etc., in the lab environment. The Macenko normalization method is used to reduce the inconsistency of this. This makes the H&E stain look the same and helps to make the color appearance of pink and purple areas of tissue more uniform among all biopsy samples.

$$I_{\mathrm{h}}^{norm}=Macenko(OD) \quad (7)$$

d) WSI patch extraction: Since whole slide images are very large, they are divided into smaller patches for efficient computational processing. The extraction of patches enables the software model to concentrate on just individual local areas of the cell and tissue structures. Patches are sampled under different magnifications (40×, 100× and 400×) in order to maintain micro-scale and multi-scale information.

$$\mathrm{P_h}=\{\mathrm{p_1},\mathrm{p_2},...,\mathrm{p_N}\},\mathrm{p_i}\subset\mathrm{I_h^{norm}} \quad (8)$$

e) Data augmentation and resizing: The extracted patches for each organ are supplemented to enhance the robustness of the model to tissue orientation, staining variation, and visual diversity. They can be rotated and jittered in colour without losing any morphological diagnosis patterns. Each of these patches is then rescaled to a fixed size of 224 × 224 for uniform input to the histopathology branch.

$$\tilde{p}_i=Resize(Augment(p_i),224\times224 \quad (9)$$

f) Train-test split: Finally, the processed histopathology patches are split into training and testing subsets, with a 70:30 split. The learning sets are used to train the model with the micro-scale biopsy representations, and the testing sets are used to evaluate the model independently. A division will minimize the risk of inadequate sampling of the histopathology stream.

$$D_{\mathrm{h}}^{train}, D_{\mathrm{h}}^{test}=Split(\tilde{P}_{\mathrm{h}},70{:}30) \quad (10)$$

After these preprocessing steps, the histopathology stream produces normalized, augmented, and resized biopsy patches that are ready for micro-scale feature extraction in the proposed MAMGAT-Net framework.

C. Proposed Model

The proposed MAMGAT-Net architecture is a multimodal diagnostic framework to handle the joint learning of complementary information derived from both mammography and histopathological images for early breast cancer classification. As shown in Fig. 3, the overall pipeline consists of two parallel input streams: the first stream, known as the mammography stream, processes macro-level details like breast masses, boundaries of lesions, and areas of microcalcification, and the second stream, the histopathology stream, processes micro-level features obtained from the histologic images obtained from the biopsies collected during the screening procedure. During the training of the model,

multi-scale convolutional operations are performed to extract features that are captured from an increasing variety of receptive fields and hence the capability of the mammography stream to detect abnormalities of varying visual structures, shapes, and sizes. The histopathology stream, on the other hand, uses a deep residual backbone that language models aim to capture complex cell morphology, the distribution of cell nuclei, and the morphology of tissue. Residual feature learning and attention mechanisms are applied to both arms—namely, residual feature learning and CBAM attention—to retain the crucial diagnostic cues as well as focus on the most relevant and spatially/channel-wise important regions.

Fig. 3. Proposed MAMGAT-Net architecture for breast cancer diagnosis using dual-stream feature extraction, cross-modal fusion, graph attention learning, and Softmax-based benign or malignant classification.

The dual-attention cross-modal fusion module is employed to set up interaction among the two modalities after obtaining modality-specific representations. This module enables multimodal interpretation of these features by allowing the mammographic features to attend to the histopathological evidence, and the histopathological features to attend back to the radiological evidence. These fused features are then converted to a graph representation where discriminative image patches/tokens correspond to graph nodes, and similarity-based edges correspond to the relationships between them. The graph is then fed into a cross-gated multi-head graph attention network that can learn key intra-modal and cross-modal dependency by giving more attention to a more careful region as it is diagnostically significant. Finally, for global graph pooling, the information stored in the nodes is aggregated to obtain a global graph-level representation, which is thereafter used to obtain the final benign and/or malignant prediction with class-wise probability scores through feature refinement, fully connected projection, and Softmax class prediction.

1) Parallel multimodal feature extraction: The parallel multimodal feature extraction module is trained using complementary diagnostic information from images in mammography and histopathology. Macro-level features like nuclei distribution and tissue architecture are captured by histopathology, whereas micro-level features like macro and microcalcifications, boundaries and borders of masses, and asymmetry are captured by mammography.

Thus, MAMGAT-Net consists of two disjoint modules that keep the information about each modality before the fusion. In the mammography branch, multi-scale convolutions are used to detect serendipitous features of various visual patterns and sizes. A mammography image I_m is used to obtain features from several receptive fields using filters with sizes of 3×3, 5×5, and 7×7:

$$F_m^{ms}=Concat(Conv_{3\times3}(I_m),Conv_{5\times5}(I_m),Conv_{7\times7}(I_m)) \quad (11)$$

This allows the model to be used for different patterns, sizes, and appearances of tumors. The extracted features are then enhanced by adding residual layers that retain low-level lesion information, while learning higher-level structural information.

$$R_m=F_m^{ms}+H_m(F_m^{ms}) \quad (12)$$

Concurrently, the histopathology branch applies a neural residual branch to the histopathological image I_h to discover cell/tissue level features. This branch has a narrow specialty and is dedicated to the study of clusters of malignant cells, nuclear density, and the abnormal organisation of the tissue. It also leverages the concept of residual learning in the refinement of its feature representation.

$$R_h=F_h^{res}+H_h(F_h^{res}) \quad (13)$$

Lastly, CBAM attention is applied to both the left and right branches to emphasize a relevant channel and spatial region. The module, therefore, produces two feature representations for each modality $F_m^{'}$ and $F_h^{'}$ that are fed to the cross-modal fusion stage.

2) Dual-attention cross-modal feature fusion: The proposed cross-modal attention module works on class-consistent virtual multimodal pairs generated during the training process, since there is no patient-level correspondence between the CBIS-DDSM and the BreaKHis datasets. The goal of the attention mechanism is thus to learn the complementary diagnostic representations at the class level, instead of the anatomical correspondence between images taken from different patients in the same imaging modality.

The dual-attention cross-modal fusion module is introduced to overcome the limitation of shallow multimodal fusion. The module allows for both mammography and histopathology features to work with one another and develop knowledge of complementary diagnostic relations rather than just appending to each other. The macro-level radiological evidence comes from the mammography representation, and the micro-level cellular evidence comes from the histopathology representation. Thus, the task of cross-modal attention facilitates a better alignment of the two feature spaces. The modality-specific features are projected in a common feature space:

$$F_m=\Phi_m(F_m^{'}),F_h=\Phi_h(F_h^{'}) \quad (14)$$

where, F_m and F_h denote the aligned feature representation of mammography and histopathology. Next, cross attention is performed in a bi-directional fashion. In the first direction, mammography combines query histopathology features, which are gathered to represent cellular proof pertinent to the radiological abnormality:

$$A_{m\leftarrow h}=Softmax\left(\frac{Q_m K_h^T}{\sqrt{d}}\right)V_h \quad (15)$$

On the other hand, histopathology includes query mammography features to correlate cellular abnormalities with the overall structural patterns:

$$A_{h\leftarrow m}=Softmax\left(\frac{Q_h K_m^T}{\sqrt{d}}\right)V_m \quad (16)$$

Finally, the original and the attended are merged and form a single multimodal representation:

$$F_u=W_f[F_m\|F_h\|A_{m\leftarrow h}\|A_{h\leftarrow m}] \quad (17)$$

The fusion strategy allows MAMGAT-Net to learn deep semantic interactions between the semantics of the macro-scale imagery and cellular patterns in the micro-scale, which enhances the reliability of breast cancer diagnostics and classification in early stages.

3) Multimodal graph construction layer: Once cross-modal fusion is complete, the unified feature representation F_u is a combination of mammographic and histopathological information. However, these simple feature vectors may not represent all the relationships between diagnostic regions. To overcome this, MAMGAT-Net uses a graph structure to relabel the fused multimodal features, where each discriminative patch/token is considered a graph node. Given the unified feature set, each feature token is mapped as a node:

$$F_u=\{f_1,f_2,...,f_N\},v_i=f_i \quad (18)$$

Here, v_i represents the i-th diagnostic region in the multimodal feature space. Feature similarity is used to define the relation between two nodes, such that the similarity for visually related regions and also the relation for the regions that are diagnosed similarly can be connected:

$$e_{ij}=sim(v_i,v_j) \quad (19)$$

The final graph is represented as:

$$G=(V,E,X) \quad (20)$$

where, V is the node set, E represents the edge set, and X contains the node feature matrix. The graph construction step allows the model to break out of the traditional approach where the features are learned independently and model the structured relationships between micro-level cellular abnormalities and macro-level radiological signs. Thus, the resulting graph acts as supporting evidence of the attention-based cross-modal reasoning in the next level.

4) Cross-gated multi-head graph attention network: After constructing the multimodal graph, the cross-gated multi-head graph attention network is used to learn structured relationships among diagnostic regions. This stage efficiently tackles the existing models that ignore the dependency relation between the mammographic abnormalities and the cellular pattern of the histogram. Unlike the usual per-region

approach, the graph attention mechanism enables each node to pool the information from neighboring diagnostic nodes. The attention coefficient between two nodes is calculated as follows for each pair of nodes:

$$\alpha_{ij}^{k}=Softmax_j\left(LeakyReLU\left(a_k^T[W_k\mathrm{h}_i\|W_k\mathrm{h}_j]\right)\right) \quad (21)$$

where, α_{ij}^{k} represents the attention weight between node I and node j under the k-th attention head. To enhance the cross-modal reasoning, a gating coefficient is added to regulate the level to which the information should be transmitted between the corresponding regions:

$$g_{ij}=Sigmoid(W_g[\mathrm{h}_i\|\mathrm{h}_j\|r_{ij}]) \quad (22)$$

The final node update is done by combining with multi-head attention and the cross-gating mechanism:

$$\mathrm{h}_i^{(l+1)}=\|_{k=1}^{K}\sigma\left(\sum_{j\in N_i} g_{ij}\,\alpha_{ij}^{k}W_k\mathrm{h}_j\right) \quad (23)$$

This exchange of context information allows regions of a mammogram and a corresponding anatomic region to interact in a diagnostic manner. This helps the model to keep the macro-scale imaging cues in alignment with the micro-scale abnormalities in the cell, thereby enhancing the robustness of the early breast cancer classification.

5) Global graph pooling and feature refinement: Following graph attention learning, node features with refined cross-modal diagnostic information are gathered. However, the classification needs a compact graph-level representation with each node embedding. Thus, a graph pooling technique is used on the global graph to summarize the most informative diagnostic regions into a single representation. The importance of each node is first estimated by an attention-based pooling score:

$$\beta_i=Softmax(w_p^T tan\mathrm{h}(W_p\mathrm{h}_i)) \quad (24)$$

Then, a graph-level feature vector is calculated as a weighted aggregation of all the node embeddings:

$$\mathrm{h}_G=\sum_{i=1}^{N}\beta_i\,\mathrm{h}_i \quad (25)$$

Here, h_G represents the full multimodal graph representation with increased weights of more diagnostic regions. With this, it is possible to leave fragmentation appropriate and limit the duplication of features.

The graph representation is then sent to a refinement block to ensure stability and generalization:

$$\mathrm{h}_r=BN(ReLU(W_r\mathrm{h}_G+b_r)) \quad (26)$$

Lastly, the refined feature is mapped to a fully connected projection, which is now prepared for classification:

$$z=ReLU(W_{fc}\mathrm{h}_r+b_{fc}) \quad (27)$$

This stage compresses graph-based multimodal evidence into a discriminative feature vector, which is then forwarded to the final classification layer.

6) Softmax classification and diagnostic output: After feature refinement, the classification-ready vector z was sent to the final Softmax classifier to obtain the class-wise diagnostic probabilities. This stage translates the learned multimodal representation into the ultimate breast cancer prediction by estimating whether the input case belongs to the benign or malignant class.

The probability distribution is computed as:

$$\hat{y}=Softmax(W_o z+b_o) \quad (28)$$

where, $\hat{y}$ represents the predicted probability scores for both classes:

$$\hat{y}=[P(Benign),P(Malignant)] \quad (29)$$

The final diagnostic decision is obtained by selecting the class with the highest probability:

$$Class=argmax(\hat{y}) \quad (30)$$

In the final stage, MAMGAT-Net gives the full remission of the association of macro-level mammographic evidence, micro-level histopathological evidence, and graph-based relational reasoning. The model gives more accurate benign or malignant predictions, along with a confidence level, without relying on isolated features in the images for support as an early breast cancer diagnosis tool. The pseudocode below presents a summary of steps followed in the proposed approach, and Table I presents the hyperparameter configuration of MAMGAT-NET.

Pseudocode 1: Summary of MAMGAT-Net

1. Input mammography and histopathology images: I_m,I_h.
2. Extract multi-scale mammography features: $F_m^{ms}=Concat(Conv_{3\times3},Conv_{5\times5},Conv_{7\times7})(I_m)$.
3. Extract histopathology features using ResNet: $F_\mathrm{h}^{res}=ResNet(I_\mathrm{h})$.
4. Apply residual learning in both branches: $R_m=F_m^{ms}+H_m(F_m^{ms})$, $R_\mathrm{h}=F_\mathrm{h}^{res}+H_\mathrm{h}(F_\mathrm{h}^{res})$
5. Apply CBAM attention to obtain refined features: $F_m^{'}=CBAM(R_m)$, $F_\mathrm{h}^{'}=CBAM(R_\mathrm{h})$
6. Align both features into a common space: $F_m=\Phi_m(F_m^{'})$, $F_\mathrm{h}=\Phi_\mathrm{h}(F_\mathrm{h}^{'})$
7. Perform bidirectional cross-modal attention: $A_{m\leftarrow\mathrm{h}}=Softmax(Q_m K_\mathrm{h}^T/\sqrt{d})V_\mathrm{h}$
8. Generate unified multimodal representation: $F_u=W_f[F_m\|F_\mathrm{h}\|A_{m\leftarrow\mathrm{h}}\|A_{\mathrm{h}\leftarrow m}]$
9. Construct multimodal graph from fused tokens: $G=(V,E,X)$
10. Apply cross-gated graph attention: $\mathrm{h}_i^{l+1}=\sum_{j\in N_i} g_{ij}\,\alpha_{ij}W\mathrm{h}_j$
11. Perform global graph pooling: $\mathrm{h}_G=\sum_{i=1}^{N}\beta_i\,\mathrm{h}_i$
12. Refine and project features: $z=ReLU(W_{fc}\mathrm{h}_G+b_{fc})$
13. Predict final class using Softmax: $\hat{y}=Softmax(W_o z+b_o)$
14. Select final diagnosis: $Class=argmax(\hat{y})$

TABLE I. HYPERPARAMETER CONFIGURATION OF MAMGAT-NET

Hyperparameter	Value
Input modality	Mammography + Histopathology
Feature extractors	Multi-scale CNN + ResNet backbone
Kernel size	3×3, 5×5, 7×7
Attention modules	CBAM + Bidirectional cross-modal attention
Graph learning module	Cross-gated multi-head GAT
Pooling strategy	Global graph pooling
Classifier	Fully connected layer + Softmax
Training setup	Adam optimizer, learning rate 0.0001, batch size 32, 100 epochs

7) MAMGAT-Net framework: The key implementation details of the proposed MAMGAT-Net framework are summarized below to facilitate reproducibility.

a) Backbone network: The mammographic and histopathological branches employ a ResNet-18 backbone for feature extraction. The backbone is initialized with ImageNet-pretrained weights to leverage transferable visual representations and improve training convergence.

b) Feature embedding: Each modality is transformed into a 512-dimensional feature embedding, providing a unified latent representation for subsequent multimodal fusion and graph-based learning. This common embedding space enables effective interaction between heterogeneous imaging modalities.

c) Graph construction: The fused multimodal feature embeddings are treated as graph nodes. Pairwise cosine similarity is computed between node representations, and edges are established only when the similarity exceeds a predefined threshold, resulting in a sparse graph that preserves meaningful relationships while reducing redundant connections.

d) Graph sparsification strategy: A threshold-based graph sparsification strategy is adopted to retain only highly correlated node pairs. This approach reduces graph density and computational overhead while maintaining the most informative structural relationships for graph attention learning.

e) Weight initialization: The ResNet-18 backbone utilizes pretrained ImageNet weights, whereas the newly introduced multimodal fusion and graph attention components are optimized jointly during end-to-end training using the specified optimization settings.

f) Normalization and activation: Each convolutional block follows the sequence of convolution, normalization, and nonlinear activation to ensure stable feature learning and efficient gradient propagation throughout the network during optimization.

g) Graph attention learning: The sparse graph is processed using the proposed Cross-Gated Multi-Head Graph Attention module, which performs attention-guided neighborhood aggregation while adaptively filtering irrelevant cross-modal information to obtain discriminative graph representations.

h) Inference complexity: The threshold-based sparse graph construction reduces the number of graph connections compared with a fully connected graph, thereby improving computational efficiency during graph reasoning. A comprehensive analysis of computational metrics, including parameter count, FLOPs, inference latency, training time, and GPU memory usage, is considered future work.

D. Baseline Model

1) Support Vector Machine (SVM): Support Vector Machines are a classical supervised learning technique that finds applications in classification as well as regression problems, particularly in cases where high-dimensional space features are used. Finding an optimal hyperplane that creates a maximum margin between various categories forms the basic goal of SVM. The data points that are closest to the separating hyperplane are referred to as support vectors and form an important part of the classifier. This method is highly useful in medical imaging applications and small dataset analysis because of its good generalization properties.

$$f(x)=\text{sign}(w^T x+b) \tag{31}$$

2) XGBoost: XGBoost is an advanced algorithm that uses ensemble learning and gradient boosting methodology to construct decision trees. Each additional decision tree is developed to correct the mistakes of previous decision trees, hence increasing the prediction accuracy. This algorithm employs regularization methodologies to prevent overfitting. Parallel computing capability has made it a very efficient algorithm to handle large structured data. It is one of the most popular algorithms in machine learning competitions and healthcare prediction systems.

$$L=\sum_{i=1}^{n} l\left(y_i,\hat{y}_i\right)+\sum_{k=1}^{K}\Omega\left(f_k\right) \tag{32}$$

3) Custom CNN: Custom Convolutional Neural Network (CNN): A CNN is a deep learning architecture uniquely created for a given problem by stacking multiple convolutional layers, pooling layers, and fully connected layers. It learns the spatial features layer by layer in images, from low-level spatial features like edge, texture to high-level semantic features like patterns. Custom CNNs can be used for domain-specific features and are effective in medical imaging tasks such as mammography analysis. The depth of the architecture and the size of the filters and activation functions can be optimized for a given dataset.

$$y_{i,j}=\sum_m \sum_n x_{i+m,j+n}\cdot w_{m,n}+b \tag{33}$$

4) DenseNet121: DenseNet121 is a densely connected CNN architecture with dense connections between each layer, enabling the maximum reusability of features and better propagation of gradients. This high connectivity mitigates the vanishing gradient problem and promotes feature reusability, thereby decreasing the number of parameters necessary for

training, while maintaining the same efficient performance of traditional CNNs. The DenseNet121 has gained significant popularity for medical image classification for its feature extraction ability and efficient computation.

$$x_l=H_l([x_0,x_1,...,x_{l-1}]) \quad (34)$$

5) EfficientNetB0: EfficientNetB0 is a new state-of-the-art CNN using a compound scaling approach that scales the depth, width, and input resolution of the network uniformly. Unlike traditional scaling approaches, which only scale one dimension, EfficientNet ensures the balance of all three dimensions to achieve better accuracy in a much more efficient manner. It is very useful in real-time and resource-limited applications like medical imaging systems used in hospitals.

$$\text{depth}=\alpha^{\phi}, \quad \text{width}=\beta^{\phi}, \quad \text{resolution}=\gamma^{\phi} \quad (35)$$

IV. Results and Discussion

These experimental results provide a comprehensive assessment of the proposed multimodal architecture employed for breast cancer diagnosis from a quantitative and qualitative perspective. The superiority of diagnostic capabilities is analyzed with respect to the baselines' performance in machine learning and deep learning, and such advanced metrics as ROC-PR plots, confusion matrix analysis, and extended metrics are applied to measure the reliability, class separation, and misclassification properties of the models. Convergence during the training process and ablation study provide evidence of the stability of the optimization and significance of each component of the system. Finally, visualization through Grad-CAM and Integrated Gradients provides a good understanding of the clinical meaning behind activations and attributions for predicting malignancy.

A. Experimental Setup and Software Configuration

For the experimental implementation of the proposed MAMGAT-Net framework, Google Colab Premium equipped with the NVIDIA T4 GPU was used to ensure efficient dual-stream multimodal architecture training. As a core programming environment, Python was used, while PyTorch and TorchVision were used to build the mammography and histopathology feature-extraction branches, together containing multi-scale convolution, residual learning, a ResNet backbone, and CBAM attention modules. For training the model, CUDA and cuDNN acceleration of computations were added to support GPU-based computation. Data preprocessing, normalization of pixels, handling of patches, and organization of the datasets were carried out using OpenCV, NumPy, and Pandas. Stain normalization and augmentation operations were performed before the image was scaled to 224×224 pixels for histopathology preprocessing. Graph construction and multi-head graph attention learning were performed using PyTorch Geometric, while train-test splitting, t-SNE visualization, and performance evaluation were conducted with Scikit-learn.

B. Comparative Performance Evaluation Against Baseline Models

A comparative evaluation against the baseline classifiers provides an intrinsic view regarding the effectiveness of the breast cancer classification framework. In this process, different metrics, such as accuracy, precision, and recall, along with the performance and clinical significance, which can be evaluated through the F1 score, will be analyzed in different machine learning and deep learning approaches to see whether there is any notable difference in the multimodal feature fusion approach compared to traditional systems, as shown in Table II below.

TABLE II. Comparative Performance Analysis of the Proposed MAMGAT-Net Framework Against Baseline Models for Breast Cancer Classification

Model	Accuracy (%)	Precision (%)	Recall(%)	F1-Score (%)
SVM	84.72	82.65	86.31	84.44
XGBoost	88.36	86.75	89.94	88.32
Custom CNN	90.41	88.83	92.02	90.39
DenseNet121	92.23	91.38	93.64	92.59
EfficientNetB0	92.71	91.95	93.96	92.63
Proposed MAMGAT-Net	**94.59**	**92.90**	**96.56**	**94.70**

The comparative results presented in Table II indicate that MAMGAT-Net performs better than all the baseline classifiers with consistent performance across all the classifiers. This is because SVM has the lowest performance among the traditional models with an accuracy of 84.72%, precision of 82.65%, recall of 86.31%, and F1-score of 84.44%, which shows poor performance in capturing the complex patterns of breast cancer images. The accuracy and F1-score are enhanced by XGBoost to 88.36% and 88.32%, respectively, indicating that the discriminative power of XGBoost outperforms SVM but is still lower than the deep learning models. Especially, the deeper feature extractor of DenseNet121 and EfficientNetB0 networks led to better classification performance, and the Custom CNN network achieved further acceleration with an accuracy of 90.41% and recall of 92.02%. The EfficientNetB0 model achieves the highest performance at the baseline of 92.71% accuracy, 91.95% precision, 93.96% recall, and 92.63% F1-score. MAMGAT-Net, by contrast, achieves the best results on all evaluation parameters, with an accuracy of 94.59%, a precision of 92.90%, a recall of 96.56% and an F1-score of 94.70%. This improvement is a result of using attention-guided multi-modal reasoning in the fusion of macro features obtained from mammography and micro features obtained from histopathology. This is particularly important in the diagnosis of breast cancer, where recall is more meaningful and is, in essence, higher sensitivity for detecting malignant cases. In general, the findings validate the effectiveness of the multimodal feature fusion and the graph-attention-based diagnostic learning.

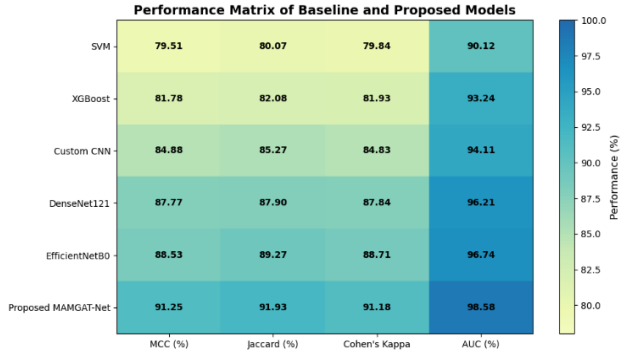


Fig. 4. Heatmap comparison of extended diagnostic performance across baseline and proposed models.

The results of the extended evaluation also show the diagnostic consistency of the proposed framework, along with the four evaluation metrics of accuracy, precision, recall, and F1-score reported in Fig. 4. SVM exhibits the lowest score, which indicates poor agreement and class-overlap ability according to MCC (79.51%), Jaccard (80.07%), Cohen's Kappa (79.84%) and AUC (90.12%). XGBoost improves these scores to 81.78%, 82.08%, 81.93%, and 93.24%, respectively. Custom CNN achieves the highest performance of 84.88% MCC and 94.11% AUC, and DenseNet121 further improves the accuracy to 87.77% MCC, 87.90% Jaccard, 87.84% Kappa, and 96.21% AUC among deep learning baselines. The efficientNetB0 model gives the best baseline, with MCC of 88.53%, Jaccard of 89.27%, Kappa of 88.71%, and AUC of 96.74%. However, the proposed MAMGAT-Net obtains the highest performance in terms of MCC (91.25%), Jaccard (91.93%), Kappa (91.18%), and AUC (98.58%), respectively, resulting in better agreement, overlap quality, and class separability through multimodal graph-attention learning. Overall, these suggest more reliable differentiation between benign and malignant cases of breast cancer.

C. Training Convergence and Learning Stability Analysis

Analysis of training convergence is crucial in understanding the efficiency of learning discriminative patterns during optimization. The learning curves for loss and accuracy across epochs indicate the network's stability in learning, the absence of erratic fluctuations, and gradual improvements in classification performance. For multimodal architectures, the assessment will be crucial with feature fusion and graph-attention learning, which demand stable optimization performance.

Although the model was trained for 100 epochs, Fig. 5 presents the first 10 epochs to illustrate the early convergence behavior, where the most significant optimization occurs. Training continued until 100 epochs using the same optimization settings, with performance remaining stable thereafter. The loss curve begins at 0.532 at the first epoch, drops sharply to 0.245 at the second epoch and to almost 0.113 at the third epoch, indicating that the initial learning level has a considerably high prediction error. This steep drop indicates that the network learns significant discriminative representations in the multimodal input space rapidly. The loss curves merge and attain the values of 0.075, 0.062, 0.056, and 0.042, respectively, at epochs 4 to 7; and the error decreases regularly as the training proceeds. At epoch 8, there is a small spike, with the loss rising to 0.045; but immediately it levels off and settles into 0.040 and 0.039 in the final epochs. The curve showing accuracy also shows this convergence pattern. The accuracy gets significantly higher with epoch 1 (0.734), epoch 2 (0.901), and epoch 3 (0.959), thus demonstrating good early learning. From epoch 4 onward, the model remains highly stable, with accuracy values of 0.974, 0.979, 0.981, 0.986, 0.984, 0.987, and 0.986. The overall accuracy of 0.986 and the loss of 0.039 indicate that the system is converging well and with very little oscillation. Overall, training behavior shows that multi-scale feature extraction, cross-modal fusion, and the graph-attention parts are trained together in an optimized way, and further, a reliable training behavior can be obtained for breast cancer diagnosis. In addition, the smooth learning pattern also indicates that the model does not show strong abrupt instability behavior in the later optimization phase, and still maintains the prediction confidence.

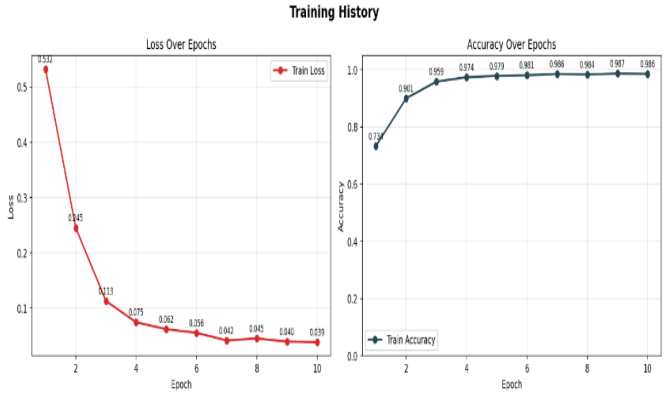


Fig. 5. Training convergence analysis showing epoch-wise loss reduction and accuracy improvement of the proposed model.

TABLE III. ABLATION STUDY RESULT

Model Variant	Accuracy (%)	Precision (%)	Recall(%)	F1-Score (%)
Mammography Branch Only	88.94	86.81	91.24	88.97
Histopathology Branch Only	89.72	87.65	92.18	89.86
Without Multi-Scale Convolution	91.38	89.46	93.21	91.30
Without Residual Learning Block	91.86	90.12	93.64	91.85
Without CBAM Attention	92.24	90.45	94.02	92.20
Without Graph Attention Network	91.12	89.02	93.18	91.05
Proposed MAMGAT-Net	**94.59**	**92.90**	**96.56**	**94.70**

D. Discriminative Performance Evaluation Using ROC and Precision-Recall Curves

ROC and Precision-Recall curve analysis yield a threshold-independent perspective on diagnostic discrimination and diagnostic prediction reliability. Analysis of true positive behavior, false positive trade-offs, precision, recall, AUC, and average precision provides a way to evaluate the ability of the model to delineate benign and malignant cases, particularly in clinically relevant class-balance and sensitivity conditions.

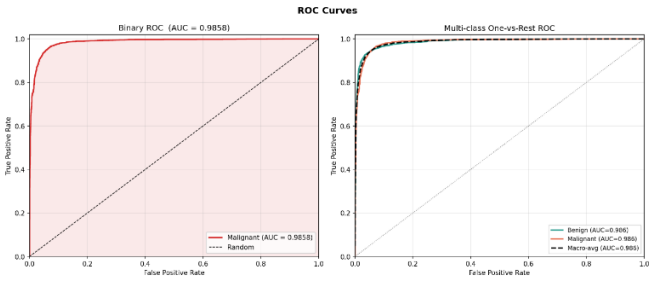


Fig. 6. ROC curve analysis illustrating binary and one-vs-rest class separability performance for benign and malignant breast cancer diagnosis.

The ROC curve-based discriminative performance of the proposed model in the classification of benign and malignant breast cancer is illustrated in Fig. 6. The binary ROC curve is calculated to be 0.9858, which is a good segmentation result between malignant and non-malignant predictions over varying decision thresholds. The curve is fairly close to the top left corner of the graph, which indicates that the model is robust with a high ability to have a true positive rate and a low false positive rate. It shows a strong deviation from the random reference line, which indicates that the classifier is far from a chance-level prediction. Compared with the recent multimodal framework of Park et al. [38], which reported an AUC-ROC of 0.98 on the same CBIS-DDSM and BreaKHis datasets, the proposed MAMGAT-Net achieves a slightly higher AUC of 98.58%, demonstrating the effectiveness of the proposed multi-scale feature extraction and cross-gated graph attention framework. The macro-AUC is equal to 0.986 as well, which indicates that the discriminative ability remains balanced among the classes, instead of mere one-class dominance. This is essential for breast cancer diagnosis, as a reliable diagnosis of malignant cases should be obtained without reducing benign diagnoses. The multi-modality feature learning, with macro-pattern from mammography and micro-level from histopathology, keeps the AUC values at a very high level, suggesting the effectiveness of the combined feature information for class separability. In conclusion, the ROC curves illustrate that the framework proposed here offers high threshold-independent reliability when making diagnostic decisions and has been shown to be suitable for differentiating benign and malignant cases with high confidence.

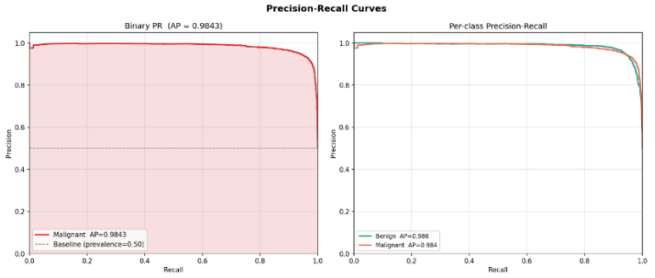


Fig. 7. Precision–Recall curve analysis showing binary and per-class diagnostic performance for benign and malignant breast cancer classification.

To complement the ROC-based evaluation, the Precision-Recall behavior shown in Fig. 7 is also shown, which is of special interest when dealing with class-sensitive conditions to measure the diagnostic reliability of the proposed model. The binary PR curve shows that the average precision (AP) for the malignant class is 0.9843, indicating the very high precision of the model at a very wide recall range. Most thresholds are close to the top threshold limit setting and far from the prevalence line 0.50, indicating that the classifier is predicting well beyond what would be expected if it were predicting the random or prevalent class. The per-class Precision-Recall plot also demonstrates a good balance of performance for both diagnostic classes, with the AP values for benign and malignant cases being 0.986 and 0.984, respectively. It can be seen that the difference between the two AP scores is small difference, which indicates that class 1 is not strongly favored over class 2. The measure value is slightly reduced at extreme recall, due to the tradeoff between precision and recall of attempting to capture almost all of the positive cases. The PR curves overall demonstrate high precision of retention, good degree of sensitivity support, and good level of benign-malignant discrimination.

E. Class-Level Error Analysis Using Confusion Matrix

The use of confusion matrix evaluation enables us to understand the behavior of the diagnostic model on a per-class basis, where the researcher distinguishes between the correctly classified instances and the instances that have been misclassified as either benign or malignant. It helps to understand the nature of the false positive, false negative, sensitivity, and specificity, other than accuracy.

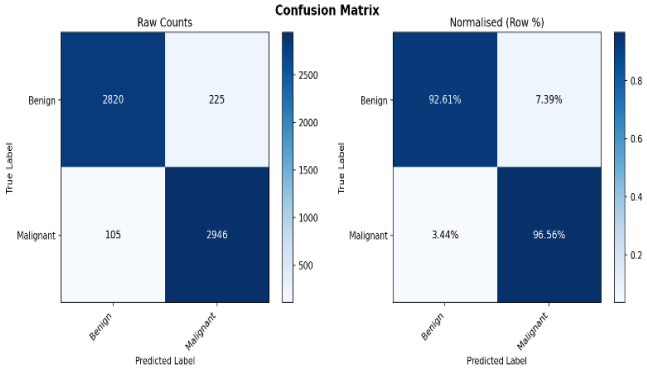


Fig. 8. Raw and normalized confusion matrix illustrating class-level prediction accuracy and misclassification patterns for benign and malignant cases.

Fig. 8 shows the prediction behavior at the class level based on the raw and row-normalized confusion matrix of the proposed model. In the raw count matrix, the model correctly classifies 2820 benign samples as benign and 2946 malignant samples as malignant, showing strong true classification performance for both diagnostic categories. The rate of misclassification is relatively low, with a total of 225 benign cases predicted as malignant and a total of 105 malignant cases predicted as benign. The normalized matrix extends the interpretation of the pure classification; benign cases give a 92.61% correct classification, and in the case of misclassification, 7.39% are classified as Malignant. In the case of malignant cases, the model gives a higher rate of 96.56% correctly classified, and only 3.44% of malignant cases are wrongly classified as benign. The significance of this result is that the false negative is more serious than the false positive in the diagnosis of malignancy because a positive result would

lead to further investigations or treatment. The lower malignant misclassification rate reflects good sensitivity and promising cancer detection performance.

At the same time, the benign classification rate of 92.61% shows that the model also maintains good specificity and does not excessively over-predict malignancy. Overall, the prediction pattern in the confusion matrix is a balanced one that reflects clinical meaning, achieved by setting the error rate of malignant to benign cases as low as possible whilst maintaining high patient accuracy in the predictions of benign samples. The results confirm the effectiveness of the multimodal feature fusion and graph-attention learning strategy for accurately diagnosing breast cancer, achieving a high diagnostic consistency rate for both benign and malignant classes.

F. Ablation Study on Multimodal Feature Learning Components

Ablation analysis is important for understanding how individual multimodal learning components contribute to overall diagnostic performance. Comparing these reduced model variants with the full model makes it clear how the modality-specific branches, multi-scale feature extraction, residual learning, attention enhancement, and graph-based reasoning contribute to better robustness and benign–malignant breast cancer classification.

Table III shows the results from the ablation study used for quantification of the effect of each major architectural component on breast cancer classification performance. The results of the modality-specific variants reveal that only using the mammography branch achieves 88.94% accuracy, 86.81% precision, 91.24% recall, and 88.97% F1-score, respectively, whilst the histopathology-only branch achieves 89.72% accuracy, 87.65% precision, 92.18% recall, and 89.86% F1-score. Here, it is found that the micro-level information at the cellular level gives slightly better information for diagnostics than the macro-level information from the mammographic features alone, but still does not give as much diagnostic information as the full structure does. Removing multi-scale convolution reduces performance to 91.38% accuracy and 91.30% F1-score, confirming its role in capturing tumor patterns of different spatial sizes. The variant without residual learning obtains 91.86% accuracy and 91.85% F1-score, suggesting that residual connections help preserve deeper contextual representations.

The results demonstrate that applying CBAM attention improves the accuracy to 92.24%, precision to 90.45%, recall to 94.02%, and F1-score to 92.20% when excluding attention, achieving more focused lesion representation. The higher degradation without a graph attention network also shows that accuracy drops to 91.12% and F1-score drops to 91.05%, indicating the significance of relational reasoning among multimodal diagnostic regions. In contrast, the complete MAMGAT-Net achieves the best results, with 94.59% accuracy, 92.90% precision, 96.56% recall, and 94.70% F1-score. In general, the ablation results agree that the integration of all multimodal branches, multi-scale learning, attention refinement, and graph-based reasoning gives the best diagnostic performance. Although the proposed framework includes bidirectional cross-modal attention as a key component, an isolated ablation experiment removing only this module was not included in the current study. This evaluation will be incorporated in future work to further quantify the individual contribution of the cross-modal fusion mechanism.

G. Explainable Visual Interpretation Using Grad-CAM and Integrated Gradients

Explainable visual interpretation is essential for evaluating whether a diagnostic model bases its predictions on clinically meaningful image regions rather than irrelevant background patterns. Grad-CAM provides evidence of class activation on the region level, and Integrated Gradients provides evidence of class activation on the pixel level. These techniques help increase transparency, aid in qualitative validation processes, and increase confidence in the automated benign-malignant classification decision.

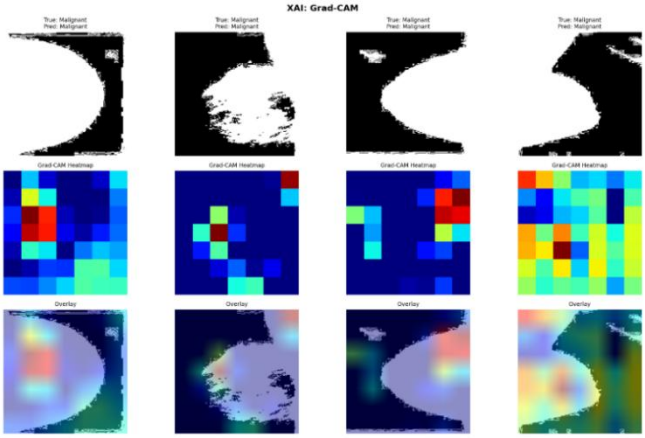


Fig. 9. Grad-CAM-based visual explanation highlighting region-level class activation evidence for correctly predicted malignant breast cancer cases.

Fig. 9 illustrates the Grad-CAM-based visual interpretation of four correctly classified malignant mammography samples, where each case is labeled as true malignant and predicted malignant. The first row presents the input images, while the second row shows the corresponding Grad-CAM heatmaps, and the third row overlays these heatmaps on the original images. The most noteworthy feature is that the places with the highest activation levels, shown by warmer hues like red, yellow, and green, are clustered around areas of the breast that are suspected to contain breast tissue and not spread randomly across the background. The activation in the first and second examples occurs at areas of intense and sporadic tissue and seems to be more concerned with visually prominent malignant formations. The third sample shows the most responding areas near the edge and the localized irregular area, where attention is placed on the structural irregularity. The fourth sample displays more activation, but the overlay is used to confirm that the model focuses on breast tissue areas relevant to the prediction of malignancy. This area-level evidence significantly contributes to the interpretability of the proposed framework as it demonstrates the contribution of diagnostic, meaningful local areas to the prediction. While Grad-CAM is not used for pixel-to-pixel lesion segmentation, it provides valuable qualitative evidence that the model is not solely using irrelevant artifacts. As a whole, the result indicates that the

developed model can effectively classify cancer individuals and provide focus on suspect areas, thereby bringing more trustworthiness in the model's decision-making process for clinical use.

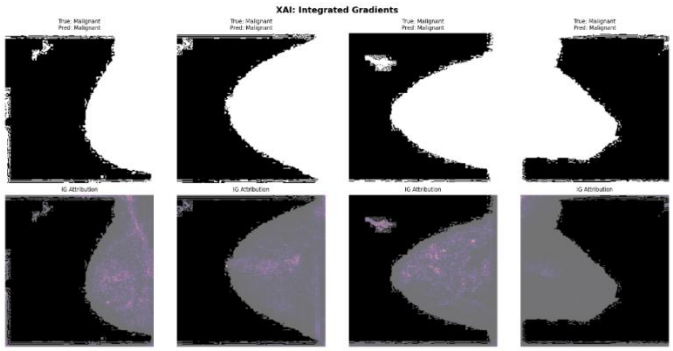


Fig. 10. Integrated Gradients-based attribution analysis showing pixel-level contribution patterns supporting malignant breast cancer prediction.

The Integrated Gradients-based attribution analysis of correctly classified malignant samples of mammograms is shown in Fig. 10; each sample is labelled as true malignant and predicted malignant. Whereas Grad-CAM detects pixel-level contribution patterns for class activation across the region, Integrated Grads highlights the fine-grained level of contribution patterns across the pixels that help the malignant prediction. The first line of images is the original mammography samples, and the second line of images is attribution maps placed over the mammography images. This purple and pink coloring in the images shows areas that are positively impacting the model's malignant decision. First sample: The attribution responses are concentrated on the dense areas of the breast tissue and boundary structures, indicating that the model takes into account local texture and intensity variation in the prediction process. Second sample: The image shows a computed tomography scan of a breast, with the attribution responses located around the dense regions of the breast tissue and boundary structures. In the third sample, the strength of the attribution is concentrated around the region of the breast that is internal and the region of the breast that is suspicious and localized, which is consistent with the focus for the model on diagnostically relevant tissue patterns. The fourth image is a comparatively weak but specific attribution, suggesting that the model isn't merely using background noise to make a guess. The attribution maps are not as clear or visually easily interpretable as heatmaps from Grad-CAM, but they offer more evidence of the importance of specific pixels and local texture patterns to the final prediction. While the classification performance was good for the whole dataset, the proposed framework did misclassify a small number of cases, mainly in situations where the benign and malignant lesions were visually similar, and/or when the image quality and tissue heterogeneity made the features of the lesions less discriminatory. This could happen because of very subtle boundaries in the lesions, overlapping tissue structure, dense breast tissue, or staining differences in the histopathological images. While few, these cases further demonstrate the inherent complexity of breast cancer diagnosis and suggest the need for further strengthening of multimodal feature learning. These failure scenarios will be explored in further detail in future work, and advanced feature representation and uncertainty estimation techniques will be investigated to further improve the reliability of the diagnostics.

Overall, IGs complement the Grad-CAM findings with pixel-level interpretability. The combined effect of region-level activation by Grad-CAM and pixel-level attribution by Integrated Gradients enhances the transparency of the proposed model for classification of malignant breast cancer, its reliability, and the diagnostic value of the model's trustworthiness.

V. Conclusion and Future Work

In this study, the author introduced a multimodal graph-attention-based framework, MAMGAT-Net, for early breast cancer diagnosis using multimodal imaging, namely mammographic and histopathological images. A proposed architecture of multi-scale feature extraction, residual learning, CBAM-based attention enhancement, bidirectional cross-modal fusion, and cross-gated multi-head graph attention learning is used to capture both macro-level radiological patterns and micro-level cellular abnormalities. Experimental results showed an improvement of the proposed framework compared to state-of-the-art conventional machine learning and deep learning approaches, with an accuracy of 94.59%, precision of 92.90%, recall of 96.56%, F1 score of 94.70%, and AUC of 98.58%. The ablation study showed the contribution of each architectural component, the ROC analysis showed good class separability, and the precision–recall evaluation and the confusion matrix results showed good diagnostic reliability. Moreover, Grad-CAM and integrated gradients visualization gave interpretable evidence to support the model's clinical decision-making. MAMGAT-Net successfully models the complementary multimodal information and relationships between diagnostic regions to improve classification robustness and avoid malignant case misclassification.

The multimodal framework was developed from two independently collected public datasets. Multimodal pairing was therefore carried out at a diagnostic class level, and not at the patient level, reducing the capacity to model true patient-specific cross-modal relationships. Second, the focus of the current study was mainly on the diagnostic performance and interpretability of the model, so computational efficiency was not systematically assessed, such as parameter count, FLOPs, inference latency, training time, and GPU memory usage. Third, the robustness of the model evaluation was not evaluated in the sense of domain shifts; imaging variations, acquisition noise, and external multi-center clinical data are not considered. Future studies will aim to test the proposed framework with clinically paired multimodal data to allow patient-level cross-modal learning, include additional clinical and genomic data to improve diagnostic capability, expand to multi-class breast cancer subtyping, and also implement extensive assessments of computational speedup and performance stability in different clinical settings.

REFERENCES

[1] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA: A Cancer Journal for Clinicians, vol. 74, no. 3, pp. 229–263, 2024. https://doi.org/10.3322/caac.21834

[2] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, "Cancer statistics, 2023," CA: A Cancer Journal for Clinicians, vol. 73, no. 1, pp. 17–48, 2023. https://doi.org/10.3322/caac.21763

[3] H. A. Rahman, S. N. N. Zaim, U. S. Suhaimei, and A. A. Jamain, "Prognostic factors associated with breast cancer-specific survival from 1995 to 2022: A systematic review and meta-analysis of 1,386,663 cases from 30 countries," Diseases, vol. 12, no. 6, p. 111, 2024. https://doi.org/10.3390/diseases12060111

[4] N. Becker and R. Liebscher, "Breast cancer diagnosis in the era of artificial intelligence: A systematic review," European Journal of Cancer, vol. 173, pp. 1–15, 2022. https://doi.org/10.1016/j.ejca.2022.05.030

[5] G. Ayana, K. Park, J.-W. Jeong, and S.-W. Choe, "Vision-transformer-based transfer learning for mammogram classification," Diagnostics, vol. 13, no. 2, p. 178, 2023. https://doi.org/10.3390/diagnostics13020178

[6] Nguyen, H. T., Nguyen, H. Q., Pham, H. H., Lam, K., Le, L. T., Dao, M., & Vu, V. (2023). VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. *Scientific Data*, *10*(1), 277.

[7] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in Proc. IEEE/CVF ICCV, 2021, pp. 10012–10022. https://doi.org/10.1109/ICCV48922.2021.00986

[8] Yang, Y., Liao, T., Lin, X. H., Ouyang, R., Chen, Q., & Ma, J. (2025). Dual-region MRI radiomic analysis indicates increased risk in high-risk breast lesions: bridging intratumoral and peritumoral radiomics for precision decision-making. *BMC cancer*, *25*(1), 828.

[9] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," IEEE Transactions on Biomedical Engineering, vol. 63, no. 7, pp. 1455–1462, 2016. https://doi.org/10.1109/TBME.2015.2476282

[10] S. Tummala, J. Kim, and S. Kadry, "BreaST-Net: Multi-class classification of breast cancer from histopathological images using ensemble of Swin transformers," Mathematics, vol. 10, no. 21, p. 4109, 2022. https://doi.org/10.3390/math10214109

[11] M. Macenko et al., "A method for normalizing histology slides for quantitative analysis," in Proc. IEEE ISBI, 2009, pp. 1107–1110. https://doi.org/10.1109/ISBI.2009.5193250

[12] Y. Li, M. E. H. Daho, P.-H. Conze, R. Zeghlache, H. Le Boité, R. Tadayoni, B. Cochener, M. Lamard, and G. Quellec, "A review of deep learning-based information fusion techniques for multimodal medical image classification," Computers in Biology and Medicine, vol. 177, p. 108635, 2024. https://doi.org/10.1016/j.compbiomed.2024.108635

[13] C. Chen, R. J. Chen, M. Y. Lu, D. F. K. Williamson, T. Y. Chen, and F. Mahmood, "Scaling vision transformers to gigapixel images via hierarchical self-supervised learning," in Proc. IEEE/CVF CVPR, 2022, pp. 16144–16155. https://doi.org/10.1109/CVPR52688.2022.01567

[14] L. Zhang, Y. Zhao, T. Che, S. Li, and X. Wang, "Graph neural networks for image-guided disease diagnosis: A review," iRADIOLOGY, vol. 1, no. 2, pp. 151–166, 2023. https://doi.org/10.1002/ird3.20

[15] Z. Wang, J. Ma, Q. Gao, C. Bain, S. Imoto, P. Liò, H. Cai, H. Chen, and J. Song, "Dual-stream multi-dependency graph neural network enables precise cancer survival analysis," Medical Image Analysis, vol. 97, p. 103252, 2024. https://doi.org/10.1016/j.media.2024.103252

[16] X. Meng and T. Zou, "Clinical applications of graph neural networks in computational histopathology: A review," Computers in Biology and Medicine, vol. 164, p. 107201, 2023. https://doi.org/10.1016/j.compbiomed.2023.107201

[17] A. G. Vrahatis, K. Lazaros, and S. Kotsiantis, "Graph attention networks: A comprehensive review of methods and applications," Future Internet, vol. 16, no. 9, p. 318, 2024. https://doi.org/10.3390/fi16090318

[18] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in Proc. ICLR, 2018. [Online]. Available: https://arxiv.org/abs/1710.10903

[19] R. J. Chen, M. Y. Lu, J. Wang, D. F. K. Williamson, S. J. Rodig, N. L. Lindeman, and F. Mahmood, "Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis," IEEE Transactions on Medical Imaging, vol. 41, no. 4, pp. 757–770, 2022. https://doi.org/10.1109/TMI.2020.3021387

[20] S. Khoshraftar and A. An, "A survey on graph representation learning methods," ACM Transactions on Intelligent Systems and Technology, vol. 15, no. 1, pp. 1–55, 2024. https://doi.org/10.1145/3633518

[21] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules for multi-scale medical image analysis," IEEE Transactions on Medical Imaging, vol. 42, no. 3, pp. 812–825, 2023.

[22] J. Zhang, Y. Xie, Y. Xia, and C. Shen, "Attention residual learning for skin lesion classification with multi-scale feature fusion," Medical Image Analysis, vol. 85, p. 102764, 2023.

[23] R. Saha, S. Bhattacharya, and A. Mukherjee, "Graph convolutional networks for histopathological image analysis in breast cancer grading," Computers in Biology and Medicine, vol. 155, p. 106622, 2023.

[24] X. Wang, L. Zhang, and H. Liu, "Cross-modal attention fusion for multimodal breast cancer diagnosis using mammography and ultrasound," Expert Systems with Applications, vol. 211, p. 118642, 2023.

[25] P. Debelee, A. Kebede, and F. Schwenker, "Survey of deep learning in breast cancer image analysis," Evolving Systems, vol. 14, no. 1, pp. 19–40, 2023.

[26] T. Li, Y. Wu, and J. Wang, "Multi-scale feature extraction with convolutional block attention module for malignancy classification in digital mammography," Biomedical Signal Processing and Control, vol. 81, p. 104508, 2023.

[27] N. Moradi, M. Talebi, and A. Sharifi, "GAT-based relational learning for whole slide image analysis in invasive breast carcinoma," Artificial Intelligence in Medicine, vol. 136, p. 102475, 2023.

[28] H. Chen, Y. Li, X. Zhang, and Q. Tian, "Dual-branch transformer network for joint mammography and histopathology feature alignment in breast cancer detection," Pattern Recognition, vol. 136, p. 109225, 2024.

[29] M. Agarwal, S. Gupta, and K. Biswas, "A deep multimodal fusion framework integrating radiological and pathological imaging for early-stage breast cancer screening," Computers in Biology and Medicine, vol. 168, p. 107821, 2024.

[30] F. Khalid, D. Mirza, and U. Qayyum, "Stain normalization techniques for computational histopathology: A comparative review and clinical benchmarking," Journal of Pathology Informatics, vol. 15, p. 100373, 2024.

[31] S. Hussain, M. Anwar, and R. Ahmad, "ResNet-based deep feature extraction with graph attention pooling for breast tissue classification in whole slide images," Neural Computing and Applications, vol. 36, no. 4, pp. 1895–1912, 2024.

[32] Y. Liu, Z. Chen, and W. Zheng, "Heterogeneous graph neural network for integrating multimodal cancer imaging data," IEEE Journal of Biomedical and Health Informatics, vol. 28, no. 2, pp. 934–945, 2024.

[33] B. Srinivasan, K. Ramachandran, and V. Natarajan, "CBAM-enhanced EfficientNet for microcalcification detection in full-field digital mammograms," Diagnostics, vol. 14, no. 3, p. 312, 2024.

[34] A. Kaur, P. Singh, and R. Sharma, "Multimodal deep learning for breast cancer classification: A systematic review of fusion strategies and benchmark datasets," Neurocomputing, vol. 576, p. 127312, 2024.

[35] W. Dong, H. Sun, and Q. Zhao, "Graph attention networks with cross-gated aggregation for pathological node classification in tumor microenvironments," Medical Image Analysis, vol. 92, p. 103050, 2024.

[36] Z. Ibrahim, O. Khalifa, and A. Abdelaziz, "Patch-level feature aggregation using multi-head graph attention for breast histopathology malignancy grading," Bioengineering, vol. 12, no. 1, p. 54, 2025.

[37] T. Nguyen, P. Tran, and L. Hoang, "Vision transformer and graph convolutional integration for multimodal oncology image classification," Computers in Biology and Medicine, vol. 183, p. 109272, 2025.

[38] J. Park, S. Kim, and H. Lee, "Unified multimodal framework using cross-attention and relational graph learning for early breast cancer screening," IEEE Transactions on Neural Networks and Learning Systems, vol. 36, no. 3, pp. 1423–1438, 2025.

[39] C. Xu, D. Yang, and L. Jiang, "t-SNE guided dimensionality refinement for improved multimodal cancer feature separability in deep classification networks," Expert Systems with Applications, vol. 261, p. 125401, 2025.

[40] R. Patel, M. Shah, and A. Desai, "Advances in multimodal AI for breast cancer diagnosis: Benchmarking cross-modal fusion on CBIS-DDSM and BreaKHis datasets," npj Breast Cancer, vol. 12, no. 1, p. 18, 2026.

[41] R. Sawyer-Lee, F. Gimenez, A. Hoogi, and D. Rubin, Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM) [Dataset]. The Cancer Imaging Archive, 2016. doi: 10.7937/K9/TCIA.2016.7O02S9CY

An Intelligent System for Energy-Aware Flexible Fault Tolerance in Cloud

Moin Hasan[1], Shailee Lohmor Choudhary[2], Rinku Sharma Dixit[3], Sujith Jayaprakash[4]
Dept. of Data Science & AI, Royal Roads University (RAK Campus), Ras Al Khaimah, UAE[1]
Dept. of Artificial Intelligence and Machine Learning, New Delhi Institute of Management, New Delhi, India[2, 3]
Dept. of ICT, Britts Imperial University College, Sharjah, UAE[4]

***Abstract*—Fault tolerance in cloud computing environments has matured significantly, and most cloud service providers now offer highly available and reliable services. However, such high levels of fault tolerance are often achieved through resource replication, which consequently increases overall energy consumption. Existing research primarily relies on conventional energy management techniques when designing fault-tolerant cloud systems. With recent advancements in artificial intelligence, there is considerable potential to improve energy management beyond traditional approaches. Motivated by this opportunity, this study presents an intelligent system for energy-aware and flexible fault tolerance in cloud environments. The proposed system employs an intelligent resource group formation algorithm that selects virtual machines based on both their energy consumption and reliability characteristics. Furthermore, by dynamically selecting appropriate fault tolerance mechanisms, the system adapts to changing workload conditions and system states. The proposed approach is evaluated through extensive simulation experiments. The results indicate that the system completes more than 96% of tasks within their specified deadlines while maintaining energy-efficient operation. Consequently, the proposed solution effectively balances the trade-off between energy efficiency and service reliability, making it a practical and sustainable approach for modern cloud computing environments.**

Keywords*—Cloud computing; energy-aware computing; fault tolerance; artificial intelligence; service reliability; sustainable computing

I. Introduction

Cloud computing has transformed the way organizations deploy, manage, and scale computing resources. By providing on-demand access to virtualized infrastructure, platforms, and software services, cloud environments enable organizations to reduce capital expenditure while improving operational flexibility and scalability [1]. The widespread adoption of cloud computing across industries has led to the rapid growth of large-scale data centers that support a diverse range of applications and services.

Despite its advantages, cloud computing faces significant challenges related to service reliability and energy consumption. Failures are inevitable in large-scale cloud infrastructures due to hardware malfunctions, software defects, network disruptions, and power-related issues [2]. Such failures can negatively affect application performance, violate service-level agreements (SLAs), and reduce user satisfaction. Consequently, fault tolerance has become a critical requirement for modern cloud systems.

Various fault tolerance mechanisms have been proposed in the literature to address these challenges, including replication, checkpointing, migration, and their hybrid variants [3]. Replication enhances reliability by maintaining multiple copies of tasks or services, while checkpointing periodically saves application states to enable recovery after failures. Migration techniques proactively move workloads away from unreliable resources to minimize service disruption. Although these approaches improve system reliability, they often incur substantial resource and energy overheads [3]. In particular, replication-based solutions may significantly increase energy consumption due to the simultaneous execution of redundant resources.

At the same time, energy efficiency has emerged as an equally important concern in cloud computing. Data centers consume an enormous amount of electrical power to operate computing, storage, networking, and cooling infrastructure [4]. This growing energy demand not only increases operational costs for cloud providers but also contributes to environmental concerns associated with carbon emissions [5]. Traditional energy management techniques such as virtual machine (VM) consolidation [6] and dynamic voltage and frequency scaling (DVFS) [7] have shown promise in reducing energy consumption; however, they are often applied independently of fault tolerance strategies.

Most of the existing studies focus primarily on either improving reliability or reducing energy consumption. As a result, the interdependence between these two objectives is frequently overlooked. Fault tolerance mechanisms designed without considering energy implications may lead to excessive power consumption, while aggressive energy-saving strategies may compromise system reliability. Furthermore, static fault tolerance approaches lack the flexibility required to adapt to dynamic workload characteristics and changing cloud conditions [8].

To address these limitations, this study proposes an intelligent system for energy-aware flexible fault tolerance in cloud computing environments. The proposed approach incorporates an intelligent resource group formation algorithm that evaluates VMs based on both reliability and energy-related characteristics. By dynamically selecting appropriate fault tolerance mechanisms according to task requirements and system conditions, the system aims to achieve a balanced

trade-off between service reliability and energy efficiency. The novelty of the proposed work does not lie in introducing a completely new replication or checkpointing mechanism. Rather, the contribution is an integrated decision framework that simultaneously 1) formulates energy-aware fault tolerance mathematically, 2) introduces energy-aware resource group formation, and 3) dynamically selects fault tolerance mechanisms according to workload characteristics, reliability requirements, and system state.

The proposed system is evaluated through extensive simulation experiments under diverse workload scenarios. Initial results demonstrate that the system successfully maintains a high task completion rate while controlling energy consumption, thereby providing an effective and sustainable solution for fault-tolerant cloud computing.

The remainder of this study is organized as follows. Section II reviews related work on cloud fault tolerance and energy-aware resource management. Section III presents the flexible fault tolerance model, followed by the energy-aware resource group formation algorithm in Section IV. Section V describes the proposed system, explaining its component-wise architecture. Results and discussion are covered in Section VI. Section VII concludes the study with directions for future research.

II. Related Works

Fault tolerance has been extensively studied in cloud computing due to the increasing demand for highly available and reliable services. Traditional fault tolerance mechanisms such as replication, checkpointing, task resubmission, and VM migration have been widely adopted to mitigate the impact of hardware failures, software faults, and network disruptions. Recent research has focused on improving the adaptability and efficiency of these mechanisms through intelligent scheduling, predictive analytics, and optimization-based resource management.

Several studies have explored proactive fault tolerance approaches that predict failures before they occur. Patil and Patil [9] proposed a dynamic load balancing model capable of detecting workload spikes and preventing host overloads through proactive migration strategies. Similarly, Ray et al. [10] introduced a fault management framework for cloud federation environments that predicts failures using CPU temperature monitoring and proactively migrates VMs to healthy cloud service providers. More recently, Ragmani et al. [11] employed artificial neural networks to predict hard disk failures, enabling preventive VM migration and improving service availability. Log-based anomaly detection techniques have also been investigated, where machine learning and large language models are utilized to identify failure indicators from system logs and initiate fault recovery actions before service degradation occurs [12].

Another stream of research focuses on adaptive and flexible fault tolerance. Kumari and Kaur [13] developed a fuzzy logic-based framework that dynamically selects between replication and checkpointing according to machine failure risk, storage availability, and task priority. Likewise, Mushtaq et al. [14] proposed an integrated scheduling framework that combines replication and checkpointing while dynamically adapting to VM-level and task-level failures. These approaches demonstrate that adaptive fault tolerance can significantly improve resource utilization and recovery performance compared with static mechanisms. However, their decision-making processes are primarily guided by reliability and performance considerations, with limited attention to energy implications.

Resource-aware fault tolerance has also received considerable attention. Dehury et al. [15] proposed a rank-based fault tolerance strategy that allocates replicas according to component criticality, thereby reducing unnecessary resource consumption. Similarly, cluster-centric scheduling frameworks and reservation-based fault tolerance strategies have been developed to improve Quality of Service (QoS), reliability, and task completion rates [16], [17]. While these methods reduce resource overhead compared with conventional replication schemes, energy consumption is rarely incorporated as a primary optimization objective.

In parallel, substantial research efforts have focused on energy-efficient cloud computing. Traditional approaches such as DVFS, VM consolidation, and intelligent workload scheduling have demonstrated significant reductions in power consumption. More recently, artificial intelligence and optimization techniques have been employed to further enhance energy management. Evolutionary programming, reinforcement learning, deep learning, and bio-inspired optimization algorithms have been successfully applied to dynamic resource allocation and workload scheduling [18], [19], [20], [21], [22]. For example, reinforcement learning-based frameworks have achieved considerable reductions in energy consumption and carbon emissions through intelligent power management [20], while hybrid machine learning models have improved energy-efficient task scheduling in cloud and fog environments [22].

Despite these advances, existing energy-aware solutions primarily focus on resource optimization, workload placement, or carbon reduction, assuming stable cloud infrastructures [18], [19], [20]. Fault tolerance is often treated as an independent concern. Conversely, fault-tolerant systems frequently prioritize reliability, availability, and SLA compliance without explicitly considering their energy overhead [13], [14], [15], [16]. Replication-based techniques, in particular, can substantially increase energy consumption due to redundant resource utilization, while migration and checkpointing mechanisms introduce additional computational and communication costs.

A review of the literature reveals several important research gaps. First, there is a lack of integrated frameworks that simultaneously address fault tolerance and energy efficiency. Second, although adaptive fault tolerance mechanisms exist, their decision-making processes rarely consider energy consumption as a key factor [13], [14]. Third, existing multi-objective optimization models generally focus on a limited set of objectives, such as reliability, migration cost, or makespan, without jointly considering energy efficiency, reliability, and deadline requirements [10], [19], [20], [23]. Finally, most studies evaluate performance and reliability independently,

with limited investigation of the trade-off between energy consumption and fault tolerance effectiveness.

To address these limitations, this study proposes an intelligent energy-aware flexible fault tolerance system for cloud computing environments. Unlike existing approaches that treat energy management and fault tolerance separately, the proposed framework integrates both objectives into a unified decision-making process. By considering reliability characteristics, energy consumption, workload requirements, and system conditions simultaneously, the proposed system aims to achieve an effective balance between service reliability and sustainable cloud operation.

III. FLEXIBLE FAULT TOLERANCE MODEL

This Section presents a comprehensive mathematical model of an energy-aware flexible fault tolerance framework for cloud computing environments. The model captures the relationships between physical hosts, VMs, and user tasks while accounting for execution dynamics, failure probabilities, and energy consumption at both VM and host levels. By formally modeling energy costs associated with different fault tolerance mechanisms, the proposed framework enables informed trade-off analysis between reliability improvement and energy efficiency. The resulting model provides a strong analytical foundation for the multi-objective optimization and algorithm design stages introduced in the subsequent Sections.

At the infrastructure level, the cloud data center consists of a finite set of physical hosts [Eq. (1)]. Each physical host represents a physical machine equipped with processing, memory, storage, and networking capabilities [24]. To enable elastic resource provisioning [25], a physical host dynamically supports a set of VMs, as shown in Eq. (2). Virtualization allows efficient sharing of physical resources while providing isolation among workloads [26]. The number and configuration of VMs on a host may vary over time depending on workload demand and system conditions.

$$H = \{h_1, h_2, h_3, \dots, h_m\} \quad (1)$$

$$VM = \{v_1, v_2, v_3, \dots, v_n\} \quad (2)$$

Each VM is characterized by a multidimensional resource vector consisting of CPU capacity (v_j^{cpu}), memory size (v_j^{mem}), available bandwidth (v_j^{bw}), and reliability (v_j^{rel}) as defined in Eq. (3). These resources are allocated to user-submitted tasks and directly influence execution performance, energy consumption, and fault tolerance behavior.

$$v_j = (v_j^{cpu}, v_j^{mem}, v_j^{bw}, v_j^{rel}) \quad (3)$$

Users submit a set of tasks or cloudlets for execution, represented in Eq. (4). These tasks may correspond to compute-intensive, data-intensive, or mixed workloads and are executed on the available VMs in the cloud data center [27]. Since cloud environments are inherently dynamic and failure-prone, each task requires fault tolerance support. Rather than relying on a single rigid fault tolerance technique, the system adopts flexible fault tolerance mechanisms, such as task replication, checkpointing, and VM migration. The choice of mechanism depends on runtime system conditions, failure risk, and energy considerations.

$$T = \{t_1, t_2, t_3, \dots, t_p\} \quad (4)$$

Each task t_k in the task set is formally described using a set of parameters, including arrival time (t_k^{at}), number of instructions (t_k^{ni}), deadline (t_k^{dl}), execution start time (t_k^{st}), execution time (t_k^{et}), and expected completion time (t_k^{ct}) as presented in Eq. (5). These parameters enable precise modeling of task scheduling, deadline constraints, and quality-of-service (QoS) requirements.

$$t_k = (t_k^{at}, t_k^{ni}, t_k^{dl}, t_k^{st}, t_k^{et}, t_k^{ct}) \quad (5)$$

When a task t_k is executed on a VM v_j, its execution time depends on the computational capacity of the VM. The execution time of the task is computed using Eq. (6), which relates the number of instructions to the processing speed of the VM. Based on this execution time, the expected completion time of the task is calculated using Eq. (7). These expressions allow the system to determine whether task deadlines can be met under different fault tolerance strategies [28].

$$t_k^{et} = \frac{t_k^{ni}}{v_j^{cpu}} \quad (6)$$

$$t_k^{ct} = t_k^{st} + t_k^{et} \quad (7)$$

Energy consumption is a critical concern in cloud data centers due to rising operational costs and sustainability requirements. For a given physical host h_i, the instantaneous power consumption $h_i^{pow_co}(t)$ is modeled as a function of its idle power ($h_i^{pow_id}$), maximum power ($h_i^{pow_mx}$), and CPU utilization ($h_i^{cpu_ut}(t)$) at time t [Eq. (8)], provided, $h_i^{cpu_ut}(t) \in [0,1]$. This widely adopted linear power model captures the strong correlation between CPU utilization and power usage in modern servers [29].

$$h_i^{pow_co}(t) = h_i^{pow_id} + (h_i^{pow_mx} - h_i^{pow_id}).\, h_i^{cpu_ut}(t) \quad (8)$$

Using the instantaneous power model, the total energy consumed by a host h_i over the interval $[0, T]$ is computed by integrating power consumption over time, as shown in Eq. (9). This formulation accurately reflects energy usage during task execution, VM migration, and fault tolerance operations. The overall energy consumption of the cloud data center is then obtained by summing the energy consumed by all hosts, as expressed in Eq. (10).

$$h_i^{enr_co} = \int_0^T h_i^{pow_c0}(t)\, dt \quad (9)$$

$$H^{enr_co} = \sum_{i=1}^{m} h_i^{enr_co} \quad (10)$$

In a cloud environment, VMs are subject to failures caused by hardware degradation, resource contention, overheating, or software anomalies [3]. Each VM v_j configured over the host h_i is therefore associated with a failure probability ($v_j^{pr_fl}$) which can be estimated using historical failure data, utilization trends, thermal indicators, or log-based anomaly detection techniques. When a task t_k is replicated across multiple (r) independent VMs to improve reliability, the probability that the task completes successfully increases. The resulting task reliability is mathematically defined in Eq. (11), assuming independent VM failures.

$$t_k^{rel_ex} = 1 - \prod_{j=1}^{r} v_j^{pr_fl} \quad (11)$$

While fault tolerance improves reliability, it introduces additional energy overhead. The energy cost of providing flexible fault tolerance to a task is calculated in Eq. (12), where $t_k^{enr_rp}$, $t_k^{enr_cp}$, and $t_k^{enr_mg}$ represent energy consumed for replication, checkpointing, and migration of t_k respectively. This equation captures the cumulative energy consumed by different fault tolerance mechanisms, including task replication, checkpointing operations, and VM migration. By explicitly modeling these energy costs, the framework enables informed trade-off decisions between reliability improvement and energy efficiency.

$$t_k^{enr_ft} = t_k^{enr_rp} + t_k^{enr_cp} + t_k^{enr_mg} \quad (12)$$

The above mathematical model establishes a unified foundation for analyzing energy-aware flexible fault tolerance in cloud computing environments. It integrates workload characteristics, execution dynamics, failure behavior, and energy consumption into a single framework, thereby enabling the design of intelligent fault tolerance strategies that enhance reliability while minimizing energy overhead. This model serves as the basis for subsequent multi-objective optimization and algorithm design in the proposed framework.

IV. Energy-Aware Resource Group Formation Algorithm for Flexible Fault Tolerance

This Section introduces an intelligent mechanism that dynamically selects and groups cloud resources for fault-tolerant task execution while balancing multiple conflicting objectives, namely energy efficiency, reliability, and deadline satisfaction. The proposed algorithm forms a critical component of the framework and directly influences its overall performance and effectiveness.

A. Problem Definition

For each incoming task submitted to the cloud data center, the system must form a resource group consisting of one or more VMs on which the task will be executed using flexible fault tolerance techniques. The selection of this resource group must satisfy the following requirements:

- The task must complete within its specified deadline.
- The overall task reliability must meet or exceed a minimum required threshold.
- The additional energy consumed due to fault tolerance must be minimized.

Let t_k and RG_k denote the task and its corresponding resource group. The challenge lies in determining the optimal composition of RG_k from the available set of VMs such that all constraints are satisfied while optimizing energy efficiency.

B. Decision Variables and Constraints

To formalize the resource group formation problem, a binary decision variable is defined in Eq. (13), using which the following constraints are imposed.

$$x_{jk} = \begin{cases} 1, if\ v_j\ is\ selected\ for\ t_k \\ 0, otherwise \end{cases} \quad (13)$$

- Reliability Constraint: The reliability achieved by executing the task on its resource group must be no less than the minimum required reliability, i.e., $t_k^{rel_ex} \geq t_k^{rel_mn}$.
- Deadline Constraint: The expected completion time of the task must not exceed its deadline, i.e., $t_k^{ct} \leq t_k^{dl}$.
- Resource Capacity Constraint: The total resource demand placed on any VM must not exceed its available capacity, i.e., $\sum_{t_k} x_{jk} . t_k^{cpu} \leq v_j^{cpu}$.
- Replication Limit Constraint: To avoid excessive replication and uncontrolled energy usage, the size of the resource group is bounded, i.e., $|RG_k| \leq r_{max}$.

The resource group formation problem is formulated as a multi-objective optimization problem with the following objectives:

- The total energy consumed for executing and protecting tasks is minimized.
- The overall task reliability is maximized by selecting VMs with higher individual reliability.
- The risk of deadline violation is minimized by ensuring timely task completion.

Since these objectives are conflicting, an exact optimal solution is computationally expensive. Therefore, a heuristic-based approach is adopted to efficiently approximate the Pareto-optimal solution. Because a linear power consumption model introduces a strictly monotonic cost curve relative to host CPU utilization, our heuristic operates as an elegant, computationally efficient approximation of a bounded Pareto-frontier by finding the exact point where replication addition ceases once the constraint $t_k^{rel_ex} \geq t_k^{rel_mn}$ is satisfied.

C. Proposed Energy-Aware Resource Group Formation Algorithm

The Energy-Aware Resource Group Formation algorithm heuristically solves the multi-objective optimization problem described above. The algorithm operates in four main stages:

- Filtering of candidate VMs based on deadline and resource feasibility
- Computation of energy–reliability composite scores
- Ranking and selection of VMs
- Adaptive formation of the resource group

For each candidate VM, a composite score is computed to balance energy efficiency and reliability [Eq. (14)], where E_{jk} represents the energy cost incurred when task t_k is executed or replicated on VM v_j.

$$v_j^{score} = \alpha . \frac{1}{E_{jk}} + \beta . v_j^{rel} \quad (14)$$

The proposed algorithm for energy-aware group formation is given below:

Algorithm 1: Energy-Aware Resource Group Formation

Input:

Task set $T = \{t_1, t_2, t_3, \dots, t_p\}$,

VM set $VM = \{v_1, v_2, v_3, \dots, v_n\}$,

Weight factors α and β such that $\alpha + \beta = 1$, and

Maximum replication limit r_{max}

Output:

Resource group RG_k for each task t_k

Pseudo code:

EnergyAwareResourceGroupFormation(T, VM)

For each task $t_k \in T$ do

$RG_k \leftarrow \emptyset$

$Candidate\ VMs \leftarrow \emptyset$

For each VM $v_j \in VM$ do

Estimate execution time t_k^{et} of t_k on v_j

Estimate completion time $t_k^{ct} = current_{time} + t_k^{et}$

If $t_k^{ct} \leq t_k^{dl}$ and resources of v_j are sufficient then

Add v_j to Candidate VMs

End

End

For each $v_j \in Candidate\ VMs$ do

Compute energy cost E_{jk} for executing t_k on v_j

Retrieve reliability v_j^{rel}

Compute composite score: $v_j^{score} = \alpha . \frac{1}{E_{ij}} + \beta . v_j^{rel}$

End

Sort $Candidate\ VMs$ in descending order of v_j^{score}

$t_k^{rel_ex} \leftarrow 0$

$r \leftarrow 0$

For each $v_j \in Candidate\ VMs$ do

If $r = r_{max}$ then

break

End

Add v_j to RG_k

$r \leftarrow r + 1$

Update Current Reliability: $t_k^{rel_ex} = 1 - \prod_{j=1}^{r} v_j^{fail_pr}$

for all $v_j \in RG_k$

If $t_k^{rel_exec} \geq t_k^{rel_min}$ then

break

End

End

If $t_k^{rel_exec} \leq t_k^{rel_min}$ then

Trigger fallback fault tolerance policy

End

End

return $RG_K\ \forall\ t_k \in T$

The algorithm provides several important advantages. It explicitly incorporates energy consumption into fault tolerance decisions and supports flexible adaptation to different workload and system conditions. It also avoids excessive replication by enforcing replication limits. The computational overhead of Algorithm 1 can be broken down into discrete steps for each task $t_k \in T$:

- Filtering Candidate VMs: Evaluating the processing speed constraint across N virtual machines takes $O(N)$ operations.
- Composite Score Computation: Computing v_j^{score} using (14) takes $O(N)$ time.
- Sorting Candidates: Sorting the candidate subset of size $M \leq N$ requires a worst-case time of $O(MlogM)$.
- Group Allocation Loop: Selecting VMs up to the maximum replication limit r_{max} runs in $O(r_{max})$ constant time since $r_{max} \ll N$.

Thus, for an incoming task sequence of size P, the algorithmic time complexity is given by $O(P.NlogN)$. By integrating reliability and energy awareness into the decision-making, the algorithm bridges the gap between theoretical modeling and practical system implementation.

V. PROPOSED SYSTEM

This Section focuses on the design and operation of the proposed Energy-Aware Flexible Fault Tolerance system, which integrates energy efficiency and reliability into a unified cloud fault tolerance solution. It presents the overall architecture of the system and explains each system component and its role in the fault tolerance process. A component-wise architectural diagram is used to illustrate the interaction between system modules, including task submission, resource monitoring, decision-making, fault handling, and execution management.

A. Component-Wise Architecture Description of the System

The proposed system is designed as a modular and layered framework that integrates energy awareness and flexible fault tolerance into cloud task execution. The architecture consists of several interconnected components, each responsible for a specific function in the fault tolerance and decision-making process. Fig. 1 illustrates the overall architecture of the proposed system.

1) Task submission and request manager: The "Task Submission and Request Manager" acts as the entry point of the system. It receives task requests submitted by cloud users along with their associated parameters such as arrival time, computational size, deadline, and priority. Upon submission, tasks are validated and forwarded to the scheduling and decision-making components for further processing. This module ensures that user requirements are correctly captured and passed to the system without modification.

2) Resource monitoring module: This module continuously observes the operational status of physical hosts and VMs within the cloud data center. It collects real-time and

historical information related to CPU utilization, memory usage, bandwidth availability, energy consumption, and thermal conditions. In addition, this module tracks VM reliability indicators such as failure history, utilization trends, and anomaly signals. The collected data is periodically updated and made available to the decision-making engine to support accurate and timely fault tolerance decisions.

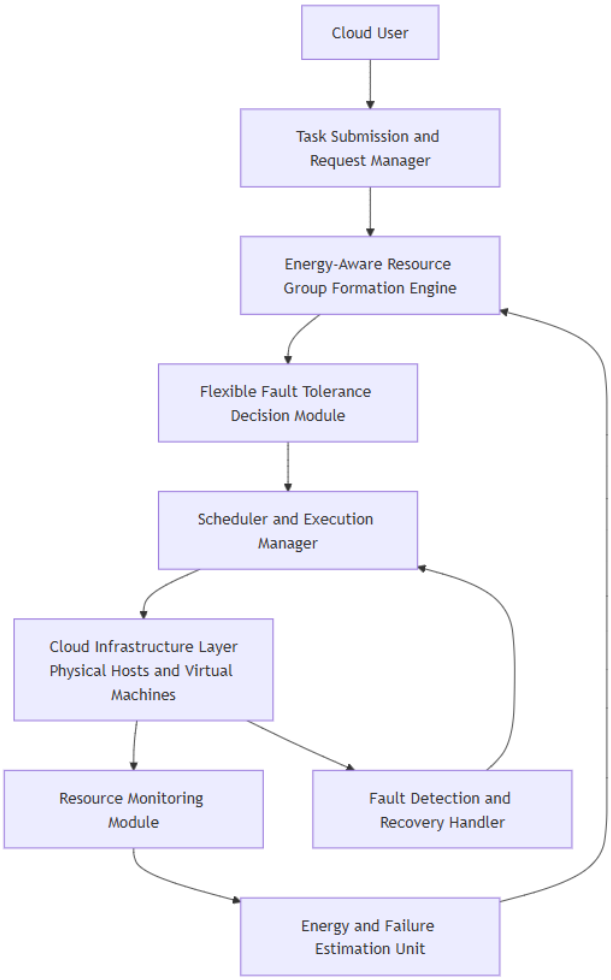


Fig. 1. Proposed system architecture.

3) Energy and failure estimation unit: This unit processes raw monitoring data to estimate energy consumption and failure probabilities. Energy estimation is performed using utilization-based power models at the host level, enabling the system to quantify the energy cost of task execution, VM migration, and fault tolerance operations. Failure estimation leverages reliability indicators to compute the probability of VM failures. These estimations form the basis for evaluating the trade-offs between reliability improvement and energy overhead.

4) Energy-aware resource group formation engine: The "Energy-Aware Resource Group Formation Engine" is the core intelligence of the proposed system. Using the multi-objective optimization approach, this engine evaluates available VMs based on energy efficiency, reliability, and resource suitability. It dynamically forms an optimal resource group for each task or task set, ensuring that selected resources satisfy deadline and reliability constraints while minimizing energy consumption. This component enables adaptive and context-aware fault tolerance decisions.

5) Flexible fault tolerance decision module: This module determines the most appropriate fault tolerance strategy for each task. Based on inputs from the resource group formation engine, this module selects between replication, checkpointing, VM migration, or hybrid fault tolerance techniques. The selection is influenced by task criticality, failure risk, expected energy cost, and execution deadlines. By avoiding static fault tolerance policies, this module ensures efficient and energy-conscious fault handling. The decision module follows a fixed priority order. If multiple VMs are assigned to the task, replication is selected. Otherwise, checkpointing is preferred for computationally intensive tasks (e.g., tasks exceeding the predefined workload threshold). VM migration is triggered when the hosting VM exceeds the utilization threshold. In all other cases, the framework employs the hybrid strategy.

6) Scheduler and execution manager: The "Scheduler and Execution Manager" is responsible for mapping tasks to the selected VMs and managing their execution lifecycle. It initiates task execution, coordinates checkpointing operations, and triggers VM migration or task replication when faults are predicted or detected. This module ensures that scheduling decisions comply with system constraints and that task execution progresses smoothly even under failure conditions.

7) Fault detection and recovery handler: This component continuously monitors task execution and VM health to identify failures or abnormal behavior. Upon detecting a fault or receiving a failure prediction signal, this component activates the recovery process defined by the selected fault tolerance strategy. Recovery actions may include switching to replica tasks, restoring from checkpoints, or migrating tasks to alternative VMs within the resource group. This module plays a critical role in maintaining service continuity.

8) Cloud infrastructure layer: The "Cloud Infrastructure Layer" consists of the physical hosts and VMs that provide computational, storage, and networking resources. This layer executes user tasks and supports dynamic VM provisioning and migration. It also serves as the source of monitoring data required for energy estimation and fault prediction. The infrastructure layer enables the practical deployment of the flexible fault tolerance system in real or simulated cloud environments.

VI. Results and Discussion

This section presents the experimental results obtained from simulation-based evaluation of the proposed framework. The experiments are designed to assess the system's ability to balance energy efficiency with fault tolerance effectiveness under realistic cloud workload conditions. To place the results in context, the proposed approach is compared against two baseline scheduling strategies: Best Fit, which selects VMs

greedily by reliability without regard for energy consumption, and Min Energy, which prioritizes the lowest-energy VMs without enforcing any reliability constraints. Together, these baselines represent opposite ends of the energy–reliability spectrum, making them appropriate reference points for evaluating the trade-off achieved by the proposed system.

A. Experimental Setup

The simulation environment consists of 20 VMs distributed across five physical hosts. Each VM is randomly configured with a CPU capacity ranging from 1,000 to 5,000 MIPS, memory between 4 GB and 16 GB, bandwidth between 100 and 1,000 Mbps, and an idle power draw between 50 W and 100 W, with a peak consumption of up to 300 W. VM failure probabilities are sampled uniformly in the range [0.01, 0.15], and reliability values are derived accordingly. Physical hosts operate with idle power between 200 W and 400 W and peak power up to 1,500 W.

A total of 500 tasks are generated for each simulation run. Task arrival times are distributed uniformly over a 100-time-unit window, with instruction counts ranging from 1,000 to 50,000 MI and relative deadlines between 10 and 100 time units. Each task carries a minimum reliability requirement drawn uniformly from [0.90, 0.999], reflecting the heterogeneous QoS demands typical of real cloud workloads. All three algorithms are evaluated on the identical task and VM set, generated using a fixed random seed to ensure a fair and reproducible comparison. In all experiments, the weight parameters were empirically set to $\alpha = 0.6$ and $\beta = 0.4$. A slightly higher weight was assigned to the energy component because the primary objective of the proposed framework is to reduce energy overhead while maintaining acceptable reliability. For the flexible fault tolerance decision, workload threshold and utilization threshold were set to 10,000 MI and 0.8, respectively. The simulation runs for 200 time units with a discrete step of 1 time unit.

Four performance metrics are used for evaluation: task completion rate, total energy consumption (Wh), and average achieved reliability. The fault tolerance overhead is modelled by applying mechanism-specific energy multipliers of 1.1× for checkpointing, 1.2× for replication, 1.25× for the hybrid strategy, and 1.3× for VM migration.

B. Overall Performance Comparison

Table I summarizes the aggregate performance of the three algorithms across all evaluation metrics. Fig. 2 to Fig. 4 trace each metric over the course of the simulation.

TABLE I. PERFORMANCE COMPARISON OF THE PROPOSED SYSTEM AND BASELINE ALGORITHMS

Metric	Proposed System	Best Fit	Min Energy
Task Completion Rate (%)	97.60	100.00	79.20
Total Energy Consumed (Wh)	201.70	206.66	163.68
Avg. Energy per Task (Wh)	0.4034	0.4133	0.3274
Avg. Achieved Reliability	0.9841	0.9825	0.9941
Reliability Threshold Met (%)	98.57	99.40	94.95

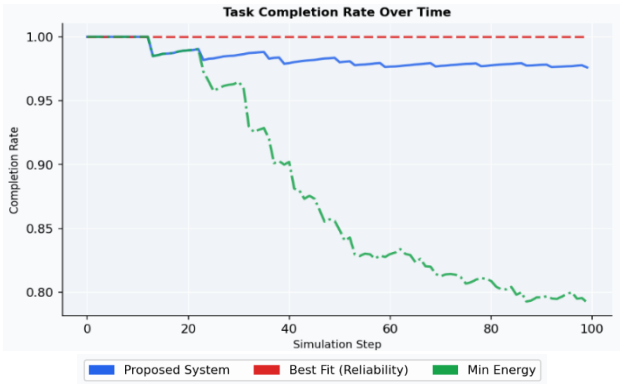


Fig. 2. Task completion rate over time.

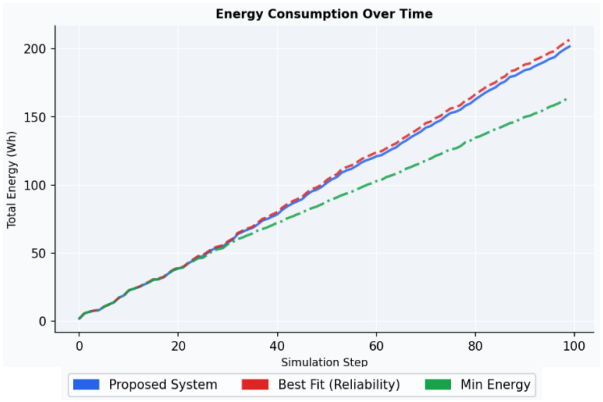


Fig. 3. Energy consumption over time.

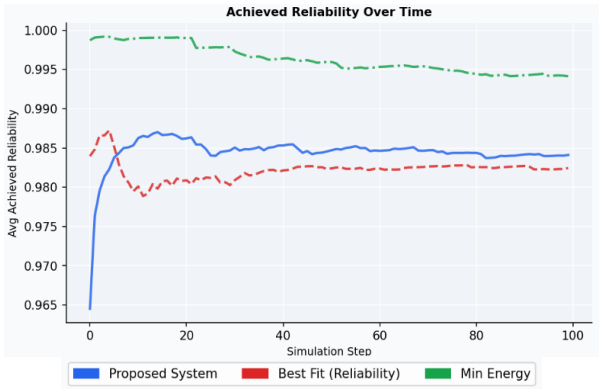


Fig. 4. Reliability over time.

1) Task completion analysis: Considering task completion rate (Fig. 2), the Best Fit baseline achieves a perfect completion rate of 100%, scheduling all 500 tasks within their respective deadlines. The proposed framework completes 488 tasks, corresponding to a 97.60% completion rate, while Min Energy completes only 396 tasks (79.20%). The notably low completion rate of Min Energy is a direct consequence of its energy-first selection policy, which occasionally assigns tasks to VMs with low processing capacity or high failure probability, resulting in missed deadlines and execution failures. The Best Fit baseline avoids this pitfall by unconditionally prioritising high-reliability VMs; however, as discussed below, this comes at a measurable energy cost.

2) Energy consumption analysis: Fig. 3 shows that Min Energy records the lowest total energy at 163.68 Wh, which is expected given that the algorithm explicitly optimizes for this metric. However, this apparent efficiency is somewhat misleading because Min Energy fails to complete about 20% of the tasks; a large fraction of potential execution energy is simply never expended. When energy is evaluated on a per-task basis, Min Energy consumes 0.3274 Wh per task, compared with 0.4034 Wh for the proposed system and 0.4133 Wh for Best Fit. The difference between the proposed system and Best Fit on this metric (approximately 2.4%) reflects the energy savings achieved by incorporating energy awareness into the VM selection process without sacrificing task throughput.

The Best Fit baseline incurs the highest total energy (206.66 Wh) precisely because it routes all tasks to the most reliable VMs irrespective of their power characteristics. High-reliability VMs tend to be more heavily provisioned and therefore draw more power. The proposed system avoids this by balancing energy cost and reliability through the composite scoring function, which avoids the selection of power-consuming VMs when more energy-efficient alternatives offer comparable reliability.

3) Reliability analysis: All three algorithms achieve high average reliability for the tasks they complete (Fig. 4). Min Energy records the highest average achieved reliability at 0.9941, which appears to be confusing at first observation. However, this result reflects a survivorship effect. Min Energy preferably selects low-power VMs; many of those may have low failure probabilities, and the tasks it completes tend to be those that benefit most from this selection. The proposed framework achieves an average reliability of 0.9841, and Best Fit achieves 0.9825. The small difference is largely attributed to the replication fallback mechanism in the proposed framework, which adds VMs to the resource group when the reliability threshold is not immediately satisfied by the top-ranked candidate.

A more practical reliability indicator is the proportion of completed tasks for which the achieved reliability met or exceeded the task-specific minimum threshold (Table I). Our proposed system satisfies this constraint for 98.57% of its completed tasks, compared with 99.40% for Best Fit and only 94.95% for Min Energy. This ordering is consistent with the design intent of each algorithm: Best Fit is most conservative about reliability, Min Energy is least so, and the proposed system occupies an intermediate position that trades a marginal reduction in reliability constraint satisfaction for a meaningful reduction in energy overhead.

C. Discussion

The obtained simulation results support the primary claim of this study that simultaneous optimization of energy efficiency and fault tolerance reliability is both feasible and beneficial. The proposed energy-aware flexible fault tolerance system demonstrates that the two objectives are not inherently opposed. A carefully designed composite scoring function, combined with an adaptive fault tolerance selection policy, can achieve near-complete task execution (97.60%). Also, the energy consumption is only marginally above the unreliable Min Energy baseline and meaningfully below the Best Fit baseline.

It should be noted that the Best Fit baseline achieves a 100% completion rate in these experiments. This is because the simulated workload, while diverse, does not saturate the VM pool to the point where reliability-first selection creates resource contention. In more congested scenarios, where the highest-reliability VMs are already heavily loaded, the Best Fit strategy would be expected to incur greater deadline misses and higher energy costs simultaneously, which would further strengthen the case for the balanced approach adopted by the proposed framework. Similarly, Min Energy's poor completion rate would be expected to worsen under stricter deadline distributions. It reinforces that energy minimization without reliability awareness is unsuitable as a standalone policy in fault-tolerant cloud environments.

The results also highlight an important practical consideration. The 2.40% deadline miss rate of the proposed framework consists primarily of tasks whose reliability requirements were sufficiently stringent with respect to the replication limit while satisfying the deadline constraint. Relaxing the replication limit or introducing additional VMs into the pool would be expected to reduce this residual miss rate further, at a modest increase in energy consumption. This configurability is an inherent strength of the proposed framework.

VII. Conclusion and Future Research Scope

Cloud service providers increasingly face the dual challenge of maintaining high reliability while controlling the growing energy demands of large-scale data centers. Although fault tolerance remains essential for ensuring service continuity, many existing approaches achieve reliability at the cost of significant energy overhead. This study addresses this challenge by proposing an intelligent energy-aware flexible fault tolerance system that integrates reliability and energy efficiency into a unified decision-making framework.

The proposed approach combines an energy-aware resource group formation algorithm with adaptive fault tolerance selection, enabling the system to dynamically choose suitable resources and fault tolerance mechanisms according to workload characteristics, reliability requirements, and system conditions. By incorporating both energy consumption and reliability metrics into the resource selection process, the framework effectively balances two traditionally conflicting objectives. The mathematical model, optimization strategy, and modular system architecture collectively provide a practical foundation for implementing sustainable fault-tolerant cloud environments.

Simulation-based evaluation demonstrated the effectiveness of the proposed system. The framework completed 97.60% of submitted tasks while satisfying reliability requirements for more than 98% of completed executions. At the same time, it reduced overall energy consumption compared with a reliability-focused baseline, illustrating that energy-aware fault

tolerance can be achieved without substantially compromising service quality. These findings confirm that intelligent resource selection and adaptive fault tolerance policies can significantly improve the operational efficiency of cloud infrastructures.

For future research, firstly, the proposed approach will be validated using large-scale cloud simulation platforms and real-world cloud workloads to further assess its scalability, adaptability, and effectiveness under highly dynamic operating conditions. Secondly, the principles of quantum computing will be incorporated into the scheduling decisions to study their effectiveness in a cloud computing environment. Performance comparison with the state-of-the-art frameworks will also be considered.

DECLARATION ON GENERATIVE AI

Several GenAI tools were used to complete this research work. ChatGPT, Claude, and DeepSeek were used for language refinement and grammar correction. Mermaid Live Editor was used to draw Fig. 1 (Proposed system architecture).

REFERENCES

[1] M. Hogan, F. Liu, A. Sokol, and J. Tong, "NIST Cloud Computing Standards Roadmap," 2013. doi: 10.6028/NIST.SP.500-291r2.

[2] G. Bosilca, R. Delmas, J. Dongarra, and J. Langou, "Algorithm-based fault tolerance applied to high performance computing," J. Parallel Distrib. Comput., vol. 69, no. 4, pp. 410–416, 2009, doi: 10.1016/j.jpdc.2008.12.002.

[3] M. Hasan and M. S. Goraya, "Fault tolerance in cloud computing environment: a systematic survey," Comput. Ind., vol. 99, pp. 156–172, 2018, doi: 10.1016/j.compind.2018.03.027.

[4] A. Safari, H. Sorouri, A. Rahimi, and A. Oshnoei, "A systematic review of energy efficiency metrics for optimizing cloud data center operations and management," Electronics (Switzerland), vol. 14, no. 11. 2025. doi: 10.3390/electronics14112214.

[5] T. Nazare, J. Gadelha, E. Nepomuceno, and R. Lozi, "Green computing for energy transition: a survey," IEEE Lat. Am. Trans., vol. 21, no. 9, pp. 937–948, 2023, doi: 10.1109/TLA.2023.10251799.

[6] S. Singh, R. Kumar, and D. Singh, "An empirical investigation of task scheduling and VM consolidation schemes in cloud environment," Computer Science Review, vol. 50. 2023. doi: 10.1016/j.cosrev.2023.100583.

[7] J. Samual, M. Hussin, N. A. W. A. Hamid, and A. Abdullah, "Frequency aware task scheduling using DVFS for energy efficiency in Cloud data centre," Expert Systems, vol. 42, no. 1. 2025. doi: 10.1111/exsy.13276.

[8] M. Hasan, M. S. Goraya, and T. Garg, "E-FFTF : An extended framework for flexible fault tolerance in cloud," in IoT and Analytics for Sensor Networks. Lecture Notes in Networks and Systems, 2021, pp. 35–45. doi: 10.1007/978-981-16-2919-8_4.

[9] A. Patil and R. Patil, "Proactive and dynamic load balancing model for workload spike detection in cloud," Meas. Sensors, vol. 27, pp. 1–8, 2023, doi: 10.1016/j.measen.2023.100799.

[10] B. K. Ray, A. Saha, S. Khatua, and S. Roy, "Proactive fault-tolerance technique to enhance reliability of cloud service in cloud federation environment," IEEE Trans. Cloud Comput., vol. 10, no. 2, pp. 957–971, 2022, doi: 10.1109/TCC.2020.2968522.

[11] A. Ragmani, A. Elomri, N. Abghour, K. Moussaid, M. Rida, and E. Badidi, "Adaptive fault-tolerant model for improving cloud computing performance using artificial neural network," Procedia Comput. Sci., vol. 170, pp. 929–934, 2020, doi: 10.1016/j.procs.2020.03.106.

[12] P. Senevirathne, S. Cooray, J. D. Herath, and D. Fernando, "Virtual machine proactive fault tolerance using log-based anomaly detection," IEEE Access, vol. 12, pp. 178951–178970, 2024, doi: 10.1109/ACCESS.2024.3506833.

[13] P. Kumari and P. Kaur, "An adaptable approach to fault tolerance in cloud computing," Int. J. Cloud Appl. Comput., vol. 13, no. 1, pp. 1–24, 2023, doi: 10.4018/IJCAC.319032.

[14] S. M. Umar, S. Sheikh, and S. M. Idrees, "Enhanced priority based task scheduling with integrated fault tolerance in distributed systems," Int. J. Cogn. Comput. Eng., vol. 6, pp. 152–169, 2025, doi: 10.1016/j.ijcce.2024.12.006.

[15] C. K. Dehury, P. K. Sahoo, and B. Veeravalli, "RRFT : A rank-based resource aware fault tolerant strategy for cloud platforms," IEEE Trans. Cloud Comput., vol. 11, no. 2, pp. 1257–1272, 2023, doi: 10.1109/TCC.2021.3126677.

[16] S. U. Mushtaq, S. Sheikh, A. Nain, S. Bharany, R. M. Ghoniem, and B. M. Taye, "CRFTS : a cluster-centric and scheduling strategy to enhance QoS in cloud computing," Sci. Rep., vol. 15, pp. 1–21, 2025.

[17] S. U. Mushtaq, S. Sheikh, and S. M. Idrees, "Next-gen cloud efficiency: Fault-tolerant task scheduling with neighboring reservations for improved resource utilization," vol. 12, no. May, 2024.

[18] M. K. Roberts, S. K. Shanmugam, M. Al-Razgan, and T. Alfakih, "A nature-inspired dual phased fuzzy hybrid algorithm for adaptive self-healing and dynamic energy-aware resource management in cloud computing," J. Cloud Comput., vol. 14, no. 62, pp. 1–23, 2025.

[19] S. Gore, Y. Bhapkar, J. Ghadge, S. Gore, and S. K. Singha, "Evolutionary programming for dynamic resource management and energy optimization in cloud computing," in International Conference on Advanced Computing Technologies and Applications (ICACTA), IEEE, 2023. doi: 10.1109/ICACTA58201.2023.10393769.

[20] S. Selvaraju, S. Maurya, Z. Adilov, S. Akhmedov, S. P. K. Gudla, and G. M, "Developing a carbon neutral cloud infrastructure with intelligent power management in green cloud AI using reinforcement learning," in International Conference on Emerging Technologies in Engineering Applications (ICETEA), IEEE, 2025. doi: 10.1109/ICETEA64585.2025.11099898.

[21] S. R. Jena, U. Srilakshmi, and S. M. Rafi, "Intelligent green energy solutions : Optimizing renewable energy for edge cloud computing," in 2025 5th International Conference on Pervasive Computing and Social Networking (ICPCSN), IEEE, 2025, pp. 765–769. doi: 10.1109/ICPCSN65854.2025.11035813.

[22] A. Khan, F. Ullah, D. Shah, M. H. Khan, S. Ali, and M. Tahir, "EcoTaskSched : a hybrid machine learning approach for energy-efficient task scheduling in IoT- based fog-cloud environments," Sci. Rep., vol. 15, pp. 1–27, 2025.

[23] J. Ramesh, Z. Solatidehkordi, K. El-fakih, and R. Aburukba, "Minimizing virtual machine live migration latency for proactive fault tolerance using an ILP model with hybrid genetic and simulated annealing algorithms," IEEE Access, vol. 12, pp. 107232–107246, 2024, doi: 10.1109/ACCESS.2024.3438358.

[24] Y. I. Lysevych and V. E. Volkov, "Mathematical model for optimizing cloud IT infrastructure," Informatics. Cult. Technol., vol. 2, pp. 237–241, 2025, doi: 10.15276/ict.02.2025.36.

[25] B. Fei, X. Zhu, D. Liu, J. Chen, W. Bao, and L. Liu, "Elastic resource provisioning using data clustering in cloud service platform," IEEE Trans. Serv. Comput., vol. 15, no. 3, pp. 1578–1591, 2022, doi: 10.1109/TSC.2020.3002755.

[26] "What is a Virtual Machine? | Microsoft Azure." Accessed: Jul. 18, 2026. [Online]. Available: https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-a-virtual-machine

[27] C. Liu and S. Baskiyar, "A general distributed scalable grid scheduler for independent tasks," J. Parallel Distrib. Comput., vol. 69, no. 3, pp. 307–314, 2009, doi: 10.1016/j.jpdc.2008.11.003.

[28] J. Wang, W. Bao, X. Zhu, T. Yang, and Y. Xiang, "FESTAL: fault-tolerant elastic scheduling algorithm for real-time tasks in virtualized cloud," IEEE Trans. Comput., vol. 64, no. 9, pp. 2545–2558, 2015, doi: 10.3969/j.issn.1000-436x.2014.10.020.

[29] L. Ismail and E. H. Abed, "Linear power modeling for cloud data centers: taxonomy, locally corrected linear regression, simulation framework and evaluation," IEEE Access, vol. 7, pp. 175003–175019, 2019, doi: 10.1109/ACCESS.2019.2956881.

INSPIRE: Understanding Local-Global Representation Learning via Controlled Self-Supervised Training and Representation Interpretability

Reem Alharthi[1], Rashid Mehmood[2]*, Aiiad Albeshri[3], Abdulaziz A. Almuzaini[4]
Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia[1, 3]
Faculty of Computer and Information Systems, Islamic University of Madinah, Madinah, Saudi Arabia[2, 4]

***Abstract*—Self-supervised learning (SSL) has become a powerful paradigm for visual representation learning, yet it remains insufficiently understood how different training strategies shape the balance between global semantic structure and local visual detail. This study introduces the INSPIRE Framework, a unified framework for systematically investigating local-global representation learning in SSL through controlled training, downstream evaluation, and representation interpretability analysis. Within this framework, we develop the INSPIRE Method, a contrastive SSL approach that constructs merged training views by combining target images with distractor content and aligning global image representations with region-level features. By varying the number of merged regions, INSPIRE explicitly controls spatial granularity during pretraining. We evaluate INSPIRE through controlled ImageNet-100 experiments (130K images) with a ResNet-18 backbone, large-scale ImageNet-1K pretraining (1.28million images) with a ResNet-50 backbone, and transfer to fine-grained recognition benchmarks. We further introduce the INSPIRE Representation Interpretability Investigation Methodology, which combines representation similarity analysis with scale-controlled spatial perturbations to examine spatial sensitivity across SSL methods and network layers. Experiments show that representation quality varies non-monotonically with spatial granularity, with intermediate granularity often producing stronger transfer performance than coarser or more fragmented alternatives. The analysis further reveals distinct spatial sensitivity profiles across SSL objectives and progressively increasing perturbation sensitivity across network depth. Runtime experiments using single-GPU and multi-GPU distributed pretraining demonstrate the practical scalability of the framework. These findings suggest that SSL representations are better understood as occupying positions along a continuum of spatial sensitivity rather than as strictly local or global learners.**

***Keywords*—Self-supervised learning; visual representation learning; local-global representation learning; representation interpretability; contrastive learning; spatial sensitivity; image representation; representation analysis; computer vision**

I. Introduction

Machine learning has achieved remarkable success across a wide range of computer vision tasks, including image classification, object detection, semantic segmentation, and visual understanding [1], [2], [3]. Much of this success has been driven by deep neural networks trained on large-scale annotated datasets. However, acquiring high-quality labels is expensive, time-consuming, and often impractical for many real-world applications. These limitations have motivated increasing interest in self-supervised learning (SSL), which seeks to learn useful visual representations directly from unlabeled data by exploiting intrinsic structure within the data itself.

Recent advances in SSL have demonstrated that representations learned without manual annotations can rival or even surpass those obtained through supervised learning on many downstream tasks [4]. Despite their success, SSL methods differ substantially in the type of visual information they encode. Effective visual representations must capture both global semantic structure and local discriminative details [5]. Global representations encode high-level information such as object identity, shape, and semantic context, whereas local representations capture fine-grained visual cues including textures, object parts, boundaries, and subtle appearance variations [6]. Different downstream tasks depend on these information sources to varying degrees. Image classification often relies heavily on global semantic understanding [7], while fine-grained recognition, object detection, and segmentation require stronger sensitivity to local visual information.

A growing body of research has attempted to improve local representation learning in SSL through region-level objectives [8], [9], patch-based learning [10], and multi-scale and hierarchical representation strategies [5], [11]. These studies demonstrate that incorporating local information can improve performance on spatially sensitive tasks. Existing research has also explored a variety of training objectives [12] and architectural designs [13] to improve the quality of learned representations. However, these approaches typically investigate multiple design factors simultaneously, making it difficult to isolate how individual training choices shape representation encodings and influence the balance between local and global information.

Another limitation concerns representation interpretability. While interpretable machine learning (IML) and explainable artificial intelligence (XAI) have traditionally focused on explaining model predictions and decision-making processes [14], [15], recent research has begun extending interpretability to learned representations [16]. This emerging direction, referred to as representation interpretability, seeks to understand

*Corresponding author.

what information is encoded within a model's internal representations, how this information is organized, and how these representations contribute to downstream behavior. Existing studies have investigated representation similarity [17], [18], semantic concepts [19], [20], and feature activations [21], [22], providing valuable insights into the structure and content of learned representations. However, they provide limited understanding of how SSL representations encode and respond to controlled spatial perturbations at different spatial scales and what these responses reveal about the balance between local and global information. Current evidence regarding spatial sensitivity is therefore inferred primarily from downstream task performance rather than measured directly at the representation level.

These observations reveal three important research gaps. First, no unified framework has been developed to systematically investigate, compare, and interpret local-global representation learning across different SSL approaches. Second, the influence of spatial granularity on representation learning remains poorly understood because it has not been investigated as an independent experimental factor separate from architectural and objective-related design choices. Third, existing representation interpretability studies do not provide a systematic methodology for investigating how SSL representations respond to controlled spatial perturbations or how these responses relate to local-global representation learning.

To address these challenges, we introduce the INSPIRE Framework, a unified framework for investigating local-global representation learning in SSL (See Section III). The framework integrates three complementary components. First, it employs controlled spatial granularity manipulation during contrastive learning, enabling the influence of spatial composition on representation learning to be investigated systematically and independently from other design factors. Second, it evaluates learned representations across both global and fine-grained downstream tasks, providing insights into how different training strategies affect transfer performance at different spatial scales. Third, it introduces a representation interpretability methodology based on scale-controlled spatial perturbations, enabling spatial sensitivity and representation behavior to be examined directly rather than inferred solely from downstream performance. These components collectively provide a structured basis for understanding how SSL training strategies shape representation encodings and local-global representation learning. Using the proposed framework, we conduct a comprehensive empirical investigation of local-global representation learning across multiple SSL configurations. The main contributions of this study are summarized as follows.

- We introduce the INSPIRE Framework, a unified framework for systematically investigating local-global representation learning in self-supervised learning through controlled training, downstream evaluation, and representation interpretability analysis.
- We develop the INSPIRE Method, a contrastive self-supervised learning approach that explicitly controls spatial granularity by constructing merged training views that combine an image with distractor content and align global image representations with region-level features extracted from these composite views. By systematically varying the number of merged regions, the method enables controlled investigation of how spatial composition and training design shape representation encodings and local-global representation learning.
- We provide a pretrained INSPIRE Encoder to support future research on local-global representation learning, spatial sensitivity, and representation interpretability in SSL.
- We introduce the INSPIRE Representation Interpretability Investigation Methodology, which combines representation similarity analysis with scale-controlled spatial perturbations to systematically investigate how SSL representations encode and respond to localized and distributed spatial information, enabling direct examination and comparison of spatial sensitivity and representation behavior across SSL methods.
- We provide new empirical insights into how SSL training strategies shape representation encodings and local-global representation learning, demonstrating non-monotonic effects of spatial granularity, distinct spatial sensitivity profiles across SSL methods, and evidence that SSL representations occupy different positions along a continuum of spatial sensitivity.

We named our framework INSPIRE to reflect its key characteristics: representation Interpretability, understanding of Neural representations, Spatial sensitivity analysis, Perturbation-based Investigation, and Representation Encodings. Beyond the specific framework and experimental findings presented in this work, the results have broader implications for the design and interpretation of self-supervised learning systems. By providing a structured approach for investigating local-global representation learning through representation interpretability and spatial sensitivity analysis, the proposed framework enables a deeper understanding of how training strategies shape learned representations. Such understanding can support the development of more effective SSL methods, guide the selection of representations for different downstream tasks, and contribute to the broader goal of making representation learning more transparent, interpretable, and scientifically grounded.

The remainder of this study is organized as follows. Section II reviews the related literature and identifies the research gaps. Section III presents the INSPIRE Framework, including the INSPIRE Method, the downstream evaluation protocol, and the INSPIRE Representation Interpretability Investigation Methodology. Section IV reports controlled ImageNet-100 experiments with a ResNet-18 backbone, examining how different merged-view patch configurations influence global and fine-grained transfer performance. Section V extends the evaluation to ImageNet-1K using a ResNet-50 backbone and compares INSPIRE with 9 well-known SSL baselines, distinguishing the controlled ablation study of Section IV from the large-scale benchmark evaluation. Section VI investigates representation interpretability through scale-controlled spatial perturbation analysis. Section VII analyzes computational cost and runtime, covering both single-GPU pretraining for

ImageNet-100 and multi-GPU distributed pretraining for ImageNet-1K. Section VIII discusses the key findings and their broader implications, and concludes the study.

II. Background and Related Work

This section reviews prior work on global and local representation learning in SSL (Sections II.A and II.B) and existing approaches to representation interpretability (Section II.C). It concludes by identifying the research gaps that motivate the proposed framework (Section II.D).

A. Global Representations in SSL Methods

Classical SSL methods were primarily developed to learn image-level representations that improve downstream tasks such as image classification. Consequently, early SSL approaches largely emphasized global semantic understanding while providing limited attention to fine-grained local structures and spatial correspondences. This section reviews the major SSL paradigms that primarily learn global semantic representations and highlights subsequent efforts to strengthen local feature learning.

Early SSL research was dominated by contrastive learning, which maximizes agreement between different augmented views of the same image while separating representations of different images in the embedding space. SimCLR [13] and MoCo [23] treat each image as a semantic instance, encouraging invariant image-level representations with strong transferability. However, their emphasis on global semantics often underrepresents fine-grained spatial relationships, motivating dense and region-aware contrastive learning methods that explicitly incorporate patch correspondence and spatial consistency [24], [25]. Beyond pairwise contrastive learning, SwAV [26] replaces explicit sample matching with online clustering and prototype assignment, enabling multiple augmented views to be associated through shared semantic clusters while remaining focused on global representations. TiCo [27] further improves transformation-invariant learning through covariance regularization that reduces feature redundancy and produces more structured global embeddings.

To remove the dependence on negative samples, non-contrastive and self-distillation methods were subsequently introduced. BYOL [12] employs an online and momentum-updated target network to learn consistent representations across augmented views, primarily reinforcing global semantic invariance. Later extensions incorporated dense and hierarchical objectives to improve local feature learning and multi-scale representations [28]. DINO [21] further extends self-distillation within Vision Transformer architectures, where patch tokenization and self-attention naturally capture both global scene semantics and localized patch-level structures, making the learned representations effective for semantic localization and fine-grained visual understanding.

Redundancy-reduction approaches, including Barlow Twins [29] and VICReg [30], learn invariant representations by decorrelating feature dimensions to avoid representational collapse while remaining primarily focused on image-level representations. A complementary direction is masked image modeling, where representations are learned by reconstructing masked image regions. MAE [31] employs an asymmetric encoder-decoder architecture to reconstruct masked patches from visible regions, directly optimizing patch-level reconstruction rather than image-level alignment. As a result, MAE naturally captures local structures, spatial relationships, and contextual information.

B. Local Representation in SSL Methods

Although classical SSL methods discussed above learn highly transferable global representations, many downstream vision tasks require fine-grained spatial information in addition to image-level semantics [5], [32]. This requirement has motivated a broad range of approaches that strengthen local representation learning through region-level objectives, patch-based modeling, hierarchical representations, and spatial regularization.

One direction extends contrastive learning to dense and region-aware settings. ReSim [15], RegionCL [8], MaskCo [9], Patch-wise SSL [10], Self-Patch [32], and HCL [24] introduce local correspondence, masked prediction, patch alignment, neighborhood consistency, and spatial feature embedding to improve localization and dense prediction. Despite their differences, these methods share the common objective of strengthening local feature learning by incorporating region-level supervision into contrastive learning.

A second direction focuses on learning hierarchical representations across multiple spatial scales. Representative approaches include Multi-level Contrastive Learning for Vision Transformers (MCVT) [33], Self-Supervised Pyramid Representation Learning (SS-PRL) [5], CsML [11], the Global-Local Self-Distillation framework [6], and UDI [34]. These methods combine representations from different network depths, feature hierarchies, or local prediction mechanisms to jointly capture coarse semantic structure and fine-grained spatial information.

A complementary line of research enhances local representations through positional supervision, spatial regularization, and masked image modeling. Positional Label [35] introduces positional prediction tasks, spatial entropy regularization [36] encourages coherent attention regions, while masked image modeling methods such as iBOT [37] and attention-guided variants such as ACoMIM [38] learn fine-grained representations through patch reconstruction combined with global semantic objectives.

C. Representation Interpretability in SSL Methods

Understanding SSL representations requires analysis beyond downstream task performance. Consequently, a growing body of research has focused on representation interpretability, investigating how learned representations are organized, what information they encode, and how they vary across architectures, training objectives, and learning paradigms. Existing studies can be broadly grouped into similarity-based analyses, probing methods, attention-based visualizations, and feature-level analyses.

Similarity-based analyses examine the relationships between representations learned by different models, network layers, and learning objectives. Among the proposed techniques, Centered Kernel Alignment (CKA) [39] has become a widely adopted

approach for comparing learned representations, offering improved robustness over earlier Canonical Correlation Analysis (CCA)-based methods [40]. Recent studies have used CKA to compare representations across architectures and training paradigms [17], [18], [41], showing that convolutional networks and Vision Transformers exhibit distinct representation organization, while supervised and self-supervised models often learn highly aligned representations in early and intermediate layers before diverging in deeper layers. These findings demonstrate that both architectural design and learning objectives substantially influence representation organization.

A complementary line of research investigates the semantic information encoded within SSL representations through probing methods [19], [20], attention-based visualizations [21], [22], and feature-level analyses [42], [43]. Probing frameworks have been used to identify the visual concepts captured by learned representations [19], [20], while attention visualization methods such as DINO [21] and ICE [22] reveal semantic object structure and region-level predictions learned without supervision. Feature-level analyses further quantify representation quality by measuring phenomena such as dimensional collapse [42].

D. Research Gap

The literature reviewed above demonstrates substantial progress in SSL representation learning and interpretation while also revealing three important research gaps. First, although prior studies have investigated global and local representation learning (Sections II.A and II.B), no unified framework has been proposed to systematically investigate, compare, and interpret local-global representation learning across SSL approaches. Second, existing methods have primarily improved downstream performance through architectural modifications [13], [23], objective design [12], [28], region-level learning [8], [9], and multi-scale representations [5], [11]. However, spatial granularity has not been systematically investigated as an independent experimental factor and remains entangled with other design choices, making its influence on local-global representation learning difficult to isolate. Third, existing representation interpretability studies have investigated representation similarity [17], [18], [41], semantic concepts [19], [20], and feature activations [21], [22] (Section II.C), but provide limited understanding of how SSL representations respond to controlled spatial perturbations at different spatial scales and how these responses relate to local-global representation learning.

These gaps motivate the INSPIRE Framework presented in Section I. The framework, its associated INSPIRE Method, INSPIRE Encoder, and INSPIRE Representation Interpretability Investigation Methodology are designed to address these limitations through a unified approach to investigating local-global representation learning. Further details are provided in Section III and onwards.

III. Methodology

The INSPIRE Framework comprises three complementary components for systematically investigating local-global representation learning in SSL (see Fig. 1). The framework integrates the INSPIRE Method and Encoder for controlled spatial granularity learning (see Section III.A), downstream evaluation across tasks requiring different levels of spatial information (Section III.B), and the INSPIRE Representation Interpretability Investigation Methodology for analyzing representation behavior through scale-controlled spatial perturbations (Section III.C).

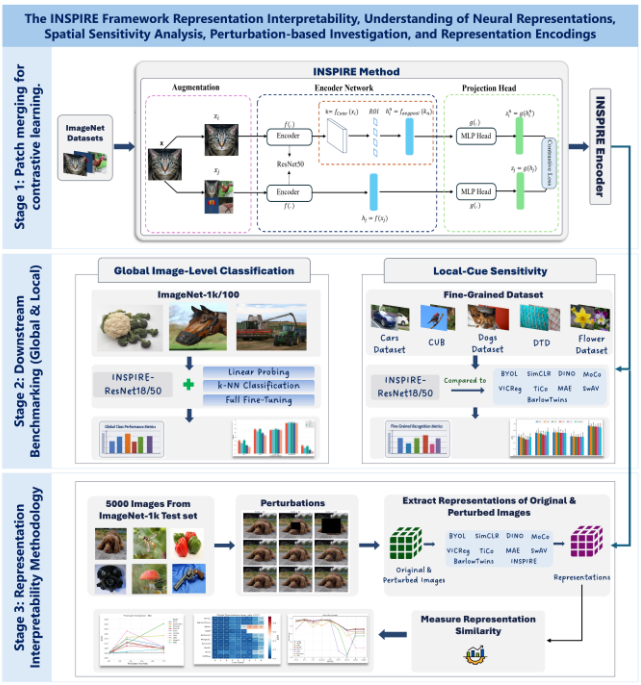


Fig. 1. INSPIRE framework: 1) Merged-view contrastive pretraining with controllable spatial granularity, 2) Evaluation on global and fine-grained tasks, and 3) Representation interpretability investigation under scale-controlled spatial perturbations.

A. INSPIRE Method

Our INSPIRE method builds upon the classic contrastive learning paradigm (e.g., SimCLR [13]) by introducing a region-based representation mechanism applied to a merged composite image. As illustrated in Fig. 2, given an input image x, two augmented views are generated. The first view, denoted x_i, is a standard augmented version of the image produced using random transformations such as cropping, flipping, or color jittering. The second view, denoted x_j, is generated by merging a transformed version of the original image x' with several other randomly selected images into a single composite image. This merged image contains multiple unrelated objects arranged in a grid-like structure, and the number of merged patches can vary (e.g., 2, 4, 6, or 8; see Fig. 2), enabling controlled manipulation of region granularity and contextual interference.

Both x_i and x_j are passed through a shared encoder network $f(\cdot)$ with a ResNet-50 backbone (Fig. 2). For the merged image x_j, we extract intermediate convolutional features after the fourth residual block, resulting in a feature map $k = f_{\text{conv}}(x_j)$. To obtain region-specific representations, we apply a Region of Interest (ROI) operation on k using bounding boxes aligned with the merged layout. Each ROI feature is then average-pooled, yielding region feature vectors $h_j^n = f_{\text{avgpool}}(k_n)$, where

n indexes the sub-image regions. Meanwhile, the standard view x_i is encoded end-to-end to produce a global representation $h_i = f(x_i)$. Next, the global representation h_i and each region vector h_j^n are passed through a shared projection head $g(.)$ implemented as an MLP, producing embeddings $z_i = g(h_i)$ and $z_j^n = g(h_j^n)$.

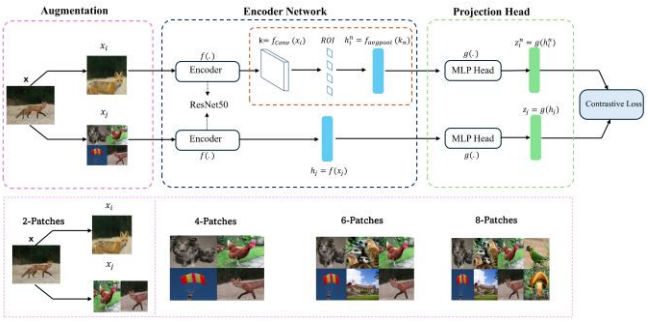


Fig. 2. INSPIRE method: global–region contrastive learning using merged views with controllable patch granularity and contextual interference.

To train INSPIRE, we adopt the normalized temperature-scaled cross-entropy (NT-Xent) objective [44] used in SimCLR, with a single positive per anchor. For each anchor embedding z_i (from x_i), we define exactly one positive z_j^+, which corresponds to the ROI embedding extracted from the merged view x_j that contains the transformed instance x'. All other embeddings, including region embeddings from non- matching ROIs and embeddings from other images in the minibatch, are treated as negatives. The contrastive loss is defined in Eq. (1), where sim( ) denotes cosine similarity between L_2-normalized embeddings, τ is a temperature hyperparameter, and M is the total number of candidate embeddings included in the contrastive normalization pool for the minibatch. This objective encourages the global representation of x_i to align with the correct ROI representation within the merged view, despite the distracting context introduced by the other merged patches.

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\mathrm{sim}(z_i, z_j^+)/\tau)}{\sum_{k=1}^{M} \mathbf{1}[k \neq i] \exp(\mathrm{sim}(z_i, z_k)/\tau)}, \quad (1)$$

B. Evaluation Protocol for Global and Local Representations

The second component of the INSPIRE Framework examines the characteristics of the representations learned by the INSPIRE Method through downstream recognition tasks. Specifically, it evaluates these representations using settings that differ in their reliance on holistic semantic information and fine-grained local visual cues, enabling both comparison with established SSL methods and assessment of local-global representation learning. We organize this evaluation into two complementary groups. For global semantic transfer, we use standard image-level classification and report performance under three protocols: k -NN classification in the learned embedding space, linear probing with a frozen encoder and a trained linear classifier, and full fine-tuning of the encoder and classifier. For local-cue sensitivity, we transfer the same pretrained encoders to fine-grained recognition benchmarks, where discriminative performance typically depends on subtle texture, part-level, and shape cues, and therefore provides a stronger test of locality-sensitive representations. Table I summarizes the datasets and train/validation split sizes.

TABLE I. DATASETS USED FOR GLOBAL CLASSIFICATION AND FINE-GRAINED (LOCAL-CUE) TRANSFER

Dataset	Representation Focus	Classes	Train	Val
ImageNet-100 [45]	Global (image-level)	100	130,000	5,000
ImageNet-1k [45]	Global (image-level)	1,000	1,281,167	50,000
Flowers-102 [46]	Local-cue sensitivity (fine-grained)	102	6,552	818
Stanford Dogs [47]	Local-cue sensitivity (fine-grained)	120	16,465	4,115
DTD [48]	Local-cue sensitivity (fine-grained)	47	4,513	1,127
CUB-200-2011 [49]	Local-cue sensitivity (fine-grained)	200	5,994	5,794
Stanford Cars [50]	Local-cue sensitivity (fine-grained)	196	8,144	8,041

We first conducted a controlled ablation study using a ResNet-18 backbone pretrained on ImageNet-100. In this stage, we isolate the effect of region granularity by varying only the number of tiles $P \in \{2,4,6,8\}$ (Fig. 2) used to construct the merged view, while keeping all other pretraining components fixed (training schedule, batch size, data augmentations, optimizer, temperature, and projection head). We denote the resulting variants as INSPIRE-P2, INSPIRE-P4, INSPIRE-P6, and INSPIRE-P8. After pretraining, we evaluate each variant on global image-level classification using the same three protocols (k-NN, linear probe, and fine-tuning) to quantify how the patch configuration affects global transfer. To evaluate local-cue sensitivity, we transfer the pretrained encoders to five fine-grained and texture recognition benchmarks (Table I) using either linear probing or partial adaptation, depending on dataset difficulty. These evaluations test whether varying P shifts representations toward global separability or toward preserving fine local discriminative cues.

For large-scale pretraining, we adopt a two-stage design. Guided by the ResNet-18 ablation, we select the patch configuration P that yields the strongest overall behavior in our global-versus-local evaluation suite and use this configuration to pretrain a ResNet-50 INSPIRE model on ImageNet-1k [45]. We then compare this ImageNet-1k pretrained INSPIRE model against established SSL baselines (e.g., SimCLR, MoCo, BYOL) under the same evaluation protocols. This setup provides a consistent and reproducible methodology for analyzing how the INSPIRE pretraining signal, controlled through the patch configuration P, impacts the balance between global semantic structure and locally discriminative feature learning.

C. INSPIRE Representation Interpretability Methodology

The third component of the INSPIRE Framework examines the spatial sensitivity of learned representations through a representation interpretability methodology based on scale-controlled spatial perturbations. By measuring how representation similarity changes in response to perturbations applied at different spatial scales, this component characterizes whether SSL representations are more sensitive to localized visual information or to broader spatial structure.

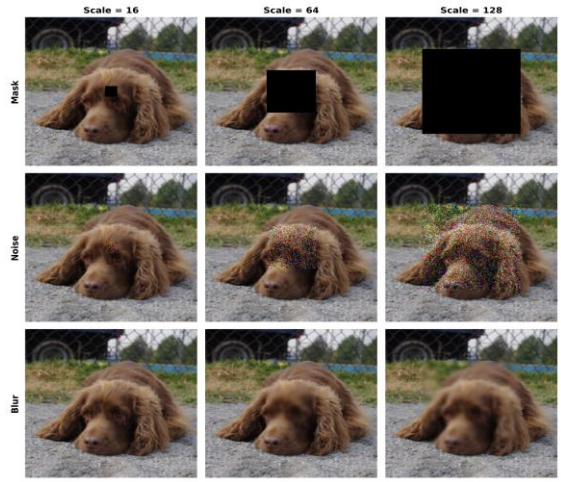


Fig. 3. Scale-controlled spatial perturbations (masking, noise, blur) applied to regions of increasing size while preserving alignment, enabling analysis of representation sensitivity to perturbation scale.

Let $x \in \mathbb{R}^{3\times H\times W}$ denote an input image and let $f(\cdot)$ represent the feature extraction function of a model at a given layer. We construct a perturbed version of the image, denoted $T_s(x)$, where T is a perturbation operator and s controls its spatial extent. We consider perturbations that preserve spatial alignment, including patch masking, localized Gaussian noise, and localized Gaussian blur. Each perturbation is applied to a square region of size $s \times s$ centered in the image, ensuring that spatial correspondence between the original and perturbed inputs is maintained (Fig. 3). By varying s, we obtain perturbations ranging from highly localized modifications to transformations affecting the entire image. To provide a scale-invariant measure, we define the normalized perturbation area in Eq. (2) as the fraction of the image affected by the perturbation.

$$\alpha = \frac{s^2}{H\times W}, \tag{2}$$

For each image x and its perturbed counterpart $T_s(x)$, we extract intermediate representations $f(x)$ and $f(T_s(x))$ from multiple layers of the model. To preserve spatial structure while obtaining fixed-dimensional representations, we apply grid-based pooling. Convolutional feature maps of shape $[B, C, H, W]$ are adaptively pooled to a fixed grid of size $g \times g$, and then flattened into vectors in $\mathbb{R}^{Cg^2}$. For transformer-based architectures, token embeddings are reshaped into spatial grids when possible, ensuring consistency across model families. This representation design retains coarse spatial layout, allowing localized perturbations to influence the resulting features.

To quantify similarity between representations, we use linear CKA [39]. Given two representations $X = f(x)$ and $Y = f(T_s(x))$, CKA is defined as:

$$\mathrm{CKA}(X,Y) = \frac{\|Y^\top X\|_F^2}{\|X^\top X\|_F \, \|Y^\top Y\|_F} \tag{3}$$

For each model, layer, perturbation type, and spatial scale s, we compute the representation change for each image x in the dataset as:

$$\Delta\mathrm{CKA}(s) = 1 - \mathrm{CKA}(f(x), f(T_s(x))) \tag{4}$$

We use a subset of 5,000 images sampled from the ImageNet-1K test set and report the mean $\Delta\mathrm{CKA}(s)$ across all samples. The resulting measurements define a scale-dependent sensitivity profile for each model and layer. By analyzing how $\Delta\mathrm{CKA}(s)$ varies as a function of the perturbation scale s, or equivalently the normalized perturbation area $\alpha = s^2/H \times W$, we obtain a structured characterization of how representations respond to perturbations at different spatial extents.

Localized perturbations provide a natural probe of spatial information in deep representations because they selectively disrupt restricted image regions while preserving the rest of the input. This approach is closely related to occlusion-based analyses introduced by Zeiler and Fergus [51], which assess the contribution of local regions by measuring changes in model responses under masking. It also connects to studies of texture–shape bias, which contrast reliance on local texture cues with more global shape information [52], [53]. In addition, the use of CKA is grounded in prior work demonstrating its effectiveness as a similarity measure for neural representations across models and layers (See Section II.C).

This component of our framework enables a controlled analysis of how representations change as a function of perturbation scale, providing a systematic way to examine sensitivity to localized and distributed modifications. By measuring representation change across layers and spatial extents, we obtain insight into how different models respond to perturbations of varying scale. While this analysis does not directly measure reliance on specific types of visual information, the observed sensitivity patterns can provide indicative evidence of how representations encode and preserve information across spatial scales.

IV. Results: ImageNet-100 (ResNet-18) (130K Images)

We analyze here INSPIRE under a controlled setting using ImageNet-100 pretraining with a ResNet-18 backbone, focusing on how the merged-view patch configuration (P) influences both global and local aspects of the learned representations while keeping all other training components fixed. Section IV.A evaluates global image-level representation quality using kNN, linear probing, and end-to-end fine-tuning, reflecting the ability to capture global semantic structure. Section IV.B examines local-cue sensitivity through transfer to fine-grained recognition benchmarks that rely on localized visual details (Table II).

TABLE II. INSPIRE Pretraining Configuration on ImageNet-100 with a ResNet-18 Backbone

Setting	Value
Backbone / Data	ResNet-18 / ImageNet-100
Training budget	100 epochs, batch size 256
Optimizer	LARS (LR 0.3, momentum 0.9, WD 10^{-6})
LR schedule	Cosine with warmup (warmup = 10 × steps/epoch)
Loss	NT-Xent (InfoNCE), $\tau = 0.4$
Projection head	SimCLR-style MLP, output dim 128
Ablation variable	Patch count $P \in \{2,4,6,8\}$
INSPIRE positives/negatives	One positive ROI (tile containing x'); others are negatives
ROI feature extraction	After layer4; global average pooling

A. Global Image-Level Results

For the ImageNet-100 experiments, we pretrain INSPIRE with a ResNet-18 backbone under a single, fixed pretraining configuration. Across all ImageNet-100 runs, the training schedule, optimizer, augmentation pipeline, projection head, and contrastive objective are held constant, and the only factor varied is the merged-view patch count $P \in \{2, 4, 6, 8\}$. INSPIRE pretraining settings are summarized in Table IV; all models are pretrained for 100 epochs with NT-Xent and a SimCLR-style MLP projection head under a fixed optimization schedule.

For comparison, we train a SimCLR baseline with the same ResNet-18 backbone (SimCLR-R18) under an identical pretraining configuration on ImageNet-100. The SimCLR model is implemented using the Lightly framework [54], and all components, including the augmentation pipeline, optimizer, learning rate schedule, batch size, number of epochs, projection head architecture, and NT-Xent loss parameters, are kept the same as in INSPIRE. This controlled setup ensures that any observed differences are attributable to the merged-view design and region-based alignment in INSPIRE. Additionally, we assess the global image-level performance of the two methods using three standard protocols (kNN evaluation, linear probing, and end-to-end fine-tuning), following the settings summarized in Table III.

TABLE III. GLOBAL EVALUATION SETTINGS

Protocol	Description
kNN	Neighbors: $k = 200$. Similarity scaling: temperature $= 0.1$. Preprocessing (train/val): resize to 256, center crop to 224, then ImageNet normalization. Reported metrics: Top-1 and Top-5 validation accuracy.
Linear probe	Trainable parameters: linear classifier only (encoder fully frozen). Training length: 90 epochs. Optimizer: SGD with learning rate 0.1, momentum 0.9, weight decay 0. Schedule: cosine learning-rate decay with no warmup. Reported metric: Top-1 validation accuracy.
Fine-tuning	Trainable parameters: encoder and classifier optimized end-to-end. Training length: 30 epochs. Optimizer: SGD with momentum 0.9, weight decay 0. Learning rate: base LR $0.05 \cdot \text{batch} \cdot \text{world}/256$. Schedule: cosine decay with no warmup. Augmentations: training uses RandomResizedCrop(224) and horizontal flip; validation uses resize to 256 and center crop to 224 (ImageNet normalization). Reported metrics: Top-1 and Top-5 validation accuracy.

Fig. 4 presents the ImageNet-100 results after 100-epoch pretraining for SimCLR-R18 and INSPIRE with different patch counts $P \in \{2, 4, 6, 8\}$. The x-axis enumerates the methods, while the y-axis reports accuracy under several evaluation protocols, including linear probing (Top-1 and Top-5), fine-tuning (Top-1 and Top-5), and kNN. A consistent pattern is that global image-level performance varies with the patch configuration, and the differences are most visible for protocols that rely directly on the pretrained representation without extensive end-to-end adaptation (kNN and linear probing). This behavior is plausible given that changing P simultaneously changes region granularity and the amount of competing context in the merged view, which may affect how easily global, class-separable structure emerges in the frozen feature space.

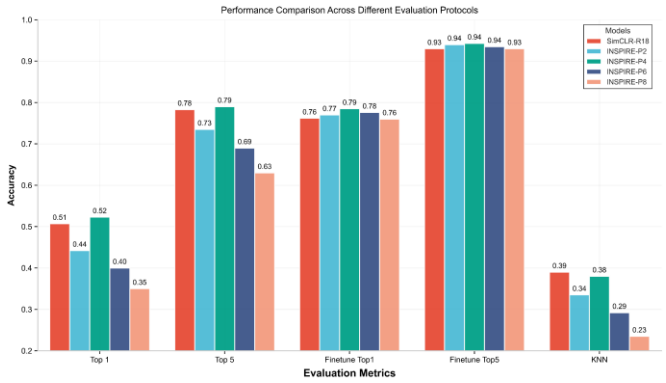


Fig. 4. ImageNet-100 results after 100-epoch pretraining: SimCLR-R18 vs. INSPIRE (P2–P8) under different evaluation protocols.

For linear probing, SimCLR-R18 reaches 0.51 Top-1 and 0.78 Top-5, while INSPIRE-P4 reaches 0.52 Top-1 and 0.79 Top-5, indicating a small improvement for P4. In comparison, the remaining INSPIRE settings show lower linear-probe accuracy, with P2 at 0.44/0.74, P6 at 0.40/0.69, and P8 at 0.35/0.63 (Top-1/Top-5). One possible interpretation is that moderate merging (P4) introduces additional contextual variation that can encourage robustness while still preserving enough global semantic consistency to support class separation under a frozen encoder. By contrast, very small (P2) or heavier merging (P6–P8) may introduce either insufficient diversity or stronger interference, which could make the alignment objective harder and reduce linear separability when only a linear head is trained.

The kNN evaluation shows a similar trend and provides an additional view of the embedding structure. SimCLR achieves 0.39 kNN accuracy, and INSPIRE-P4 is close at 0.38, whereas P2, P6, and P8 decrease to 0.34, 0.29, and 0.24, respectively. Since kNN depends on neighborhood relationships in the learned feature space, the lower kNN scores at larger P are consistent with representations that are less globally clustered by class under the frozen encoder, potentially due to increased competing content in the merged view.

After fine-tuning, the performance differences across methods become smaller, suggesting that end-to-end adaptation can partially mitigate the gaps observed under frozen-feature evaluations. SimCLR-R18 attains 0.76 Top-1 and 0.93 Top-5, while INSPIRE-P4 reaches 0.79 Top-1 and 0.94 Top-5, reflecting a modest improvement. The other INSPIRE variants cluster in a narrower range after fine-tuning (Top-1 around 0.76–0.78), even though their linear probing and kNN results are lower. This pattern is consistent with the idea that fine-tuning can adjust earlier features and decision boundaries to better match the supervised objective, reducing the sensitivity to differences in the initial embedding geometry induced by the patch configuration. Fig. 4 suggests that P controls a trade-off between region granularity and contextual interference, which is most apparent when evaluating the frozen representation space and becomes less pronounced after end-to-end adaptation.

TABLE IV. OVERVIEW OF BENCHMARK DATASETS FOR FINE-GRAINED VISUAL RECOGNITION

Dataset	Examples			Why It Is Fine-Grained
Flowers-102 [46]				Intra-class variation is low while inter-class differences are subtle, relying on fine details such as petal structure, color gradients, and reproductive parts.
	Pink Primrose	Toad Lily	Tiger Lily	
Stanford Dogs [47]				High visual similarity across breeds with distinctions concentrated in subtle facial features, fur texture, and body proportions.
	Siberian Husky	Alaskan Malamute	Samoyed	
DTD [48]				Categories are defined by abstract texture attributes rather than objects, requiring recognition of local repetitive patterns and statistical appearance cues.
	Banded	Braided	Bubbly	
CUB-200-2011 [49]				Fine-grained recognition depends on small differences in part configurations, plumage patterns, and coloration, often localized to specific body regions.
	Red-cockaded Woodpecker	Downy Woodpecker	Louisiana Waterthrush	
Stanford Cars [50]				Subtle inter-class variations arise from minor differences in shape, grille design, headlights, and brand-specific styling cues across visually similar vehicles.
	Audi A5 Coupe 2012	BMW 4 Coupe 2012	Acura TSX Sedan 2012	

B. *Fine-Grained Transfer (Local-Cue Evaluation)*

To examine local-cue sensitivity, we evaluate whether representations learned by INSPIRE transfer to fine-grained recognition tasks that require discriminative texture- and part-level cues. We compare INSPIRE variants against SimCLR-R18 trained with different merged-view patch counts. All models are pretrained on ImageNet-100 for 100 epochs under the same setup, differing only in P, and then transferred to fine-grained benchmarks to evaluate how patch count affects performance on localized visual tasks.

Table IV summarizes the fine-grained transfer datasets used to evaluate local-cue sensitivity. These benchmarks are challenging because many classes share strong global similarity, and correct recognition often depends on localized cues such as texture primitives (DTD), part-level attributes and markings (CUB-200-2011), and small shape or appearance differences within a common object category (Flowers-102, Stanford Dogs, and Stanford Cars). As a result, performance on these datasets is less driven by coarse, global semantics and more influenced by the quality of features that preserve discriminative local information.

To ensure a fair and consistent transfer evaluation, we use a fixed tuning budget for all methods. For each fine-grained dataset, we run a small set of four training configurations and apply the same set to every pretrained encoder. All transfer runs are trained for a fixed 20 epochs, and the final reported score for each (model, dataset) pair is the best validation Top-1 accuracy among these four configurations. For the linear-probe datasets (Flowers-102, Stanford Dogs, and DTD), the four configurations cover two weight decay settings (0 and 10^{-4}) and two label-smoothing values (0.05 and 0.10), while keeping the learning rate fixed at 0.3. For the more challenging datasets (CUB-200-2011 and Stanford Cars), we adopt the layer4-adaptation protocol and tune weight decay over $\{10^{-4}, 10^{-3}\}$ with label smoothing set to 0, again using a learning rate of 0.3. This procedure keeps the tuning effort comparable across methods while reducing sensitivity to a single choice of optimization hyperparameters.

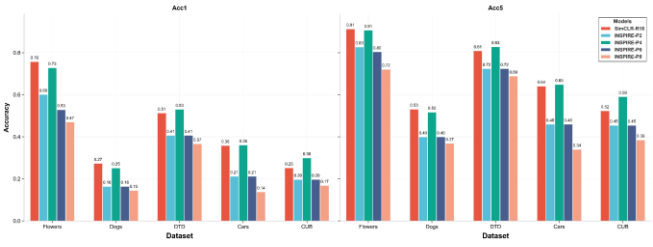


Fig. 5. Performance comparison across fine-grained transfer datasets.

Fig. 5 summarizes fine-grained transfer performance across five datasets. The x-axis lists the datasets, and the y-axis reports accuracy. The left panel shows Top-1 accuracy (Acc1), and the

right panel shows Top-5 accuracy (Acc5) for SimCLR-R18 and INSPIRE with $P \in \{2, 4, 6, 8\}$. The figure indicates that transfer performance depends on the patch configuration, with the clearest differences appearing in the more challenging benchmarks (Cars and CUB) and in the texture-focused benchmark (DTD). This behavior is consistent with the idea that increasing P changes both the effective region granularity and the amount of competing context present during pretraining, which may influence how well locally discriminative cues are preserved for downstream transfer.

For the linear-probe datasets (Flowers, Dogs, and DTD), SimCLR-R18 achieves 0.76/0.91 (Top-1/Top-5) on Flowers, 0.27/0.53 on Dogs, and 0.51/0.81 on DTD. INSPIRE-P4 reaches 0.73/0.91 on Flowers and 0.25/0.52 on Dogs, which remains close to SimCLR, while on DTD it reaches 0.53/0.83, showing a modest improvement over SimCLR on the texture-oriented task. In contrast, P2 yields lower accuracy across these datasets (e.g., 0.60/0.83 on Flowers and 0.41/0.72 on DTD), and performance decreases further for larger patch counts (P6 and P8), which may reflect increased contextual interference in the merged view when the number of tiles becomes larger.

For the layer 4-adaptation datasets (Cars and CUB), the same trend is visible, and differences become more pronounced. On Cars, SimCLR-R18 reaches 0.36/0.64, while INSPIRE-P4 reaches 0.36/0.65, indicating similar Top-1 and a small increase in Top-5. On CUB, SimCLR-R18 reaches 0.25/0.52, whereas INSPIRE-P4 reaches 0.30/0.59, reflecting a clearer improvement under the limited adaptation setting. Meanwhile, higher patch counts again show lower transfer accuracy (e.g., P8 reaches 0.14/0.34 on Cars and 0.17/0.39 on CUB), which is consistent with the possibility that stronger merging makes it harder for the pretrained representation to support fine-grained distinctions under constrained supervision.

Fig. 5 suggests that moderate merging (here, P4) can maintain competitive transfer on several fine-grained benchmarks and may provide modest gains on datasets that rely strongly on local patterns (DTD) or part-level distinctions (CUB), while heavier merging configurations (P6–P8) tend to reduce fine-grained transfer performance.

V. Results: ImageNet-1k (ResNet-50) (1.28M Images)

Having identified the best-performing patch configuration on ImageNet-100 in Section IV, we now extend the analysis to a large-scale setting by pretraining INSPIRE on ImageNet-1k with a ResNet-50 backbone and comparing it against 9 well-known self-supervised learning baselines. The goal here is to examine whether the observed effects of the merged-view design persist at scale and under more challenging evaluation conditions.

We evaluate representations along two complementary dimensions. First, in Section V.A, we assess global image-level performance using standard ImageNet protocols, including kNN, linear probing, and end-to-end fine-tuning, which reflect global semantic separability and transfer efficiency. Second, in Section V.B, we evaluate fine-grained transfer performance on datasets that depend more heavily on localized visual cues. These evaluations allow us to assess how INSPIRE behaves at scale and whether it maintains a balance between global semantic structure and locally discriminative features when compared to existing SSL methods.

A. Global Evaluation vs Nine SSL Baselines

For the large-scale global evaluation, we pretrain INSPIRE on ImageNet-1k using a ResNet-50 backbone and the $P = 4$ patch configuration (INSPIRE-P4). Training is conducted for 100 epochs using distributed data parallel (DDP) over 7 GPUs with a per-device batch size of 128 (effective batch size 896) under mixed-precision. The effective batch size was selected to balance training efficiency with the available computational resources, as substantially larger batch sizes would require additional GPU memory and computational capacity. We adopt SimCLR-style data augmentations and optimize the NT-Xent contrastive objective with temperature $\tau = 0.4$. Optimization uses LARS with momentum 0.9 and weight decay 10^{-6}. The learning rate is scaled with the effective batch size as $\text{LR} = 0.075 \cdot \sqrt{128 \times 7}$, and we apply a cosine warmup schedule with warmup length set to $10 \times$ steps-per-epoch. We use the same SimCLR-style projection head as in the ImageNet-100 experiments, mapping features to a 128-dimensional embedding space.

We compare INSPIRE-R50 against nine widely adopted SSL baselines spanning the major paradigms of modern SSL, including contrastive, clustering-based, redundancy-reduction, self-distillation, and masked image modeling methods. All methods are evaluated using the same global evaluation protocols described in Section IV.A. All evaluation settings (including kNN parameters, preprocessing, and optimization hyperparameters for linear probing and fine-tuning) are kept identical to the ImageNet-100 protocol to enable a consistent large-scale comparison focused on representation quality.

All large-scale comparisons in this section follow a consistent benchmarking setup based on the Lightly framework [54]. The SSL baselines (e.g., DINO, SwAV, SimCLR, BYOL, MoCo v2, VICReg, Barlow Twins, TiCo, and MAE) are taken from Lightly-pretrained checkpoints trained for 100 epochs with an effective batch size of 256 and evaluated under the same protocol suite used throughout this work.

Fig. 6. ImageNet-1k linear probe results (ResNet-50, 100 epochs): Top-1 and Top-5 accuracy (%) for INSPIRE-R50 and SSL baselines.

Fig. 6 summarizes the ImageNet-1k linear evaluation results after 100 epochs of pretraining with a ResNet-50 backbone. The

x-axis lists the SSL methods, and the y-axis reports linear-probe accuracy (Top-1 and Top-5, in %), obtained under the same linear evaluation protocol.

INSPIRE-R50 achieves 65.4 Top-1 and 87.2 Top-5. Among the methods compared, DINO (68.2 Top-1, 87.9 Top-5) and SwAV (67.2 Top-1, 88.1 Top-5) obtain the highest Top-1 accuracy in this setting, while INSPIRE-R50 is lower in Top-1 and closes in Top-5. Relative to several commonly used contrastive and redundancy-reduction baselines, INSPIRE-R50 shows higher linear-probe accuracy than SimCLR (63.2/85.2), VICReg (63.0/85.4), Barlow Twins (62.9/84.3), BYOL (62.5/85.0), and MoCo v2 (61.5/84.1) in both Top-1 and Top-5. TiCo (49.7/74.4) and MAE (46.0/70.2) are lower under this linear evaluation protocol. Fig. 6 suggests overall that INSPIRE-R50 yields a competitive linear evaluation performance on ImageNet-1k relative to several established SSL baselines, while remaining below the best-performing methods in this comparison.

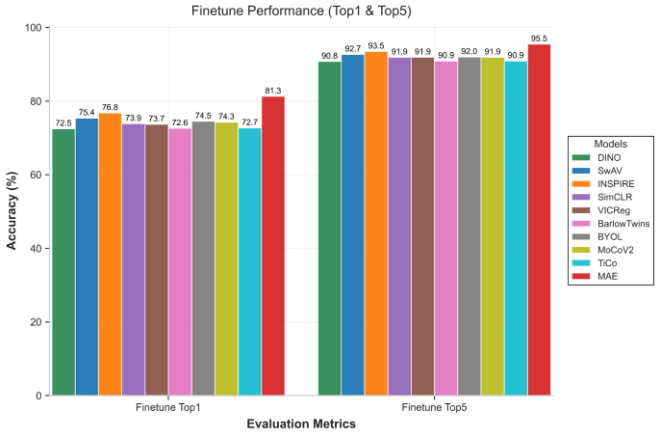


Fig. 7. ImageNet-1k fine-tuning results (ResNet-50, 100 epochs): Top-1 and Top-5 accuracy (%) for INSPIRE-R50 and SSL baselines.

Fig. 7 shows the ImageNet-1k fine-tuning results after 100 epochs of ResNet-50 pretraining. The x-axis lists the SSL methods, and the y-axis reports fine-tuned accuracy (Top-1 and Top-5, in %), obtained under the same end-to-end fine-tuning protocol.

INSPIRE-R50 reaches 76.8 Top-1 and 93.5 Top-5. In this setting, INSPIRE-R50 is above several commonly used SSL baselines, including SimCLR (73.9/91.9), VICReg (73.7/91.9), BYOL (74.5/92.0), MoCo v2 (74.3/91.9), Barlow Twins (72.6/90.9), DINO (72.5/90.8), and TiCo (72.7/90.9), and it is also slightly above SwAV (75.4/92.7). MAE shows a different behavior in this comparison: it attains the highest fine-tuning accuracy (81.3 Top-1 and 95.5 Top-5), despite being much lower in the linear-probe results reported earlier. This contrast is consistent with the notion that some pretraining objectives may produce features that are not as linearly separable without adaptation yet become highly effective once fine-tuned with supervised labels. In this context, INSPIRE Method's results appear relatively stable across evaluation protocols, providing competitive linear evaluation performance while also achieving strong accuracy after end-to-end fine-tuning.

Fig. 8 reports the kNN evaluation results on ImageNet-1k after 100-epoch pretraining with a ResNet-50 backbone. The x-axis lists the SSL methods, and the y-axis shows kNN accuracy (Top-1 and Top-5, in %), computed under the same kNN protocol used throughout this section.

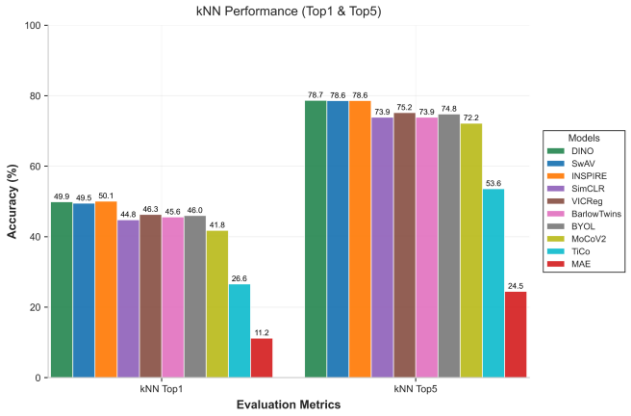


Fig. 8. ImageNet-1k kNN results (ResNet-50, 100 epochs): Top-1 and Top-5 accuracy (%) for INSPIRE-R50 and SSL baselines.

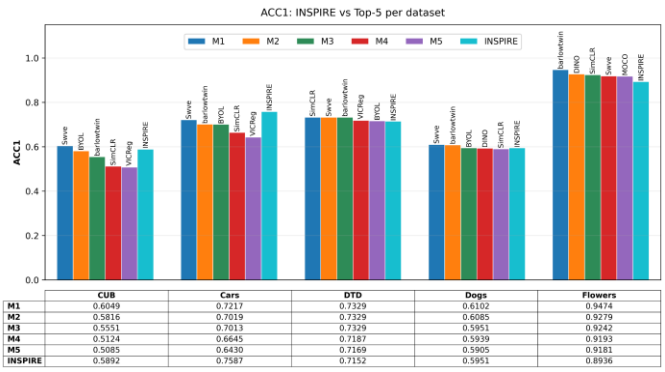
INSPIRE-R50 achieves 50.1 Top-1 and 78.6 Top-5. In this comparison, INSPIRE-R50 is close to the strongest kNN results reported by DINO (49.9/78.7) and SwAV (49.5/78.6), and it is higher than several commonly used baselines such as SimCLR (44.8/73.9), VICReg (46.3/75.2), BYOL (46.0/74.8), Barlow Twins (45.6/73.9), and MoCo v2 (41.8/72.2). TiCo (26.6/53.6) and MAE (11.2/24.5) are lower under kNN, which is consistent with the general observation that kNN strongly depends on the class structure being directly reflected in the embedding space without supervised adaptation.

Across the three global protocols reported in this section (linear probing in Fig. 6, fine-tuning in Fig. 7 , and kNN in Fig. 8), INSPIRE-R50 shows consistent performance under both frozen-feature evaluations (kNN and linear probing) and end-to-end adaptation (fine-tuning). While the relative ordering differs across protocols for some methods (e.g., MAE performs substantially better after fine-tuning than under kNN or linear probing), INSPIRE-R50 remains broadly competitive across the full evaluation suite, suggesting its suitability as a stable large-scale pretraining approach in this setting.

B. Fine-Grained Transfer vs Nine SSL Baselines

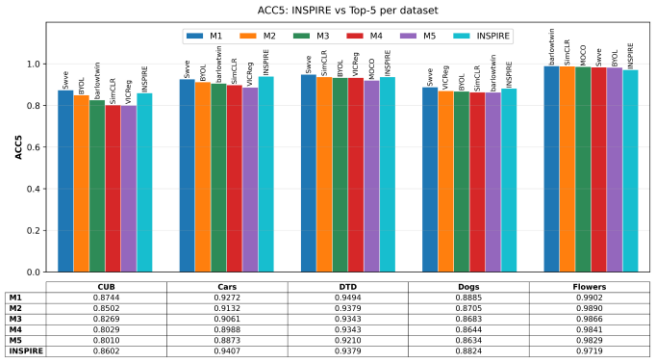
Here, we extend the large-scale evaluation by comparing fine-grained transfer performance of INSPIRE-R50 against a set of established SSL baselines, including BYOL, DINO, MAE, MoCo, SimCLR, SwAV, TiCo, and VICReg. The goal is to assess how representations learned at ImageNet-1k scale transfer to recognition problems that typically depend on subtle, localized cues such as fine textures, part-level attributes, and small shape variations.

To ensure a fair comparison across methods, we follow the same fine-grained evaluation protocol introduced in Section IV.B All pretrained models are evaluated on the same set of fine-grained datasets using identical transfer settings and a shared hyperparameter configuration.



	CUB	Cars	DTD	Dogs	Flowers
M1	0.6049	0.7217	0.7329	0.6102	0.9474
M2	0.5816	0.7019	0.7329	0.6085	0.9279
M3	0.5551	0.7013	0.7329	0.5951	0.9242
M4	0.5124	0.6645	0.7187	0.5939	0.9193
M5	0.5085	0.6430	0.7169	0.5905	0.9181
INSPIRE	0.5892	0.7587	0.7152	0.5951	0.8936

Fig. 9. Fine-grained transfer Top-1 (ACC1): INSPIRE-R50 vs best SSL baselines per dataset.



	CUB	Cars	DTD	Dogs	Flowers
M1	0.8744	0.9272	0.9494	0.8885	0.9902
M2	0.8502	0.9132	0.9379	0.8705	0.9890
M3	0.8269	0.9061	0.9343	0.8683	0.9866
M4	0.8029	0.8988	0.9343	0.8644	0.9841
M5	0.8010	0.8873	0.9210	0.8634	0.9829
INSPIRE	0.8602	0.9407	0.9379	0.8824	0.9719

Fig. 10. Fine-grained transfer Top-5 (ACC5): INSPIRE-R50 vs best SSL baselines per dataset.

Fig. 9 and Fig. 10 summarize the fine-grained transfer comparison between INSPIRE-R50 and a subset of SSL baselines on five datasets. In both figures, the x-axis lists the datasets (CUB, Cars, DTD, Dogs, and Flowers) and the y-axis reports accuracy. Fig. 9 shows Top-1 (ACC1) and Fig. 10 shows Top-5 (ACC5). For readability, the figures display INSPIRE together with the top-performing baseline methods per dataset (the specific baseline names are indicated above the bars), so the compared baseline set can differ slightly from one dataset to another.

Fig. 9 shows that INSPIRE achieves comparable Top-1 performance across all fine-grained datasets. On Cars, INSPIRE reaches 75.87%, which is higher than the other reported methods (≈70–72%). On CUB, INSPIRE (58.92%) is slightly below the best baseline, SwAV (60.49%), but remains within a close range. A similar pattern appears on Dogs, where INSPIRE (59.51%) is near SwAV (61.02%) and other baselines. On DTD, INSPIRE (71.52%) is below SimCLR and SwAV (≈73.29%), and on Flowers (89.36%) it is lower than methods such as Barlow Twins (94.74%) and DINO (≈93%). The INSPIRE Method overall provides consistent performance across datasets, although it does not achieve the highest accuracy in all cases.

Fig. 10 (ACC5) follows a similar pattern, where INSPIRE maintains comparable performance across all datasets. On Cars, INSPIRE achieves 94.07%, which is above other methods such as SwAV (92%) and BYOL (91.32%). On CUB, INSPIRE (86.02%) is slightly lower than SwAV (87.44%). On Dogs, INSPIRE (88.24%) is comparable to SwAV (88.85%) and other baselines. On DTD, INSPIRE (93.79%) is close to SwAV (94.94%), while on Flowers (97.19%) it trails Barlow Twins (99.02%) and SimCLR (≈98.9%). Compared to Top-1 results, the differences between methods are smaller in Top-5 accuracy, indicating that INSPIRE captures relevant class information even when it is not ranked highest in Top-1 accuracy.

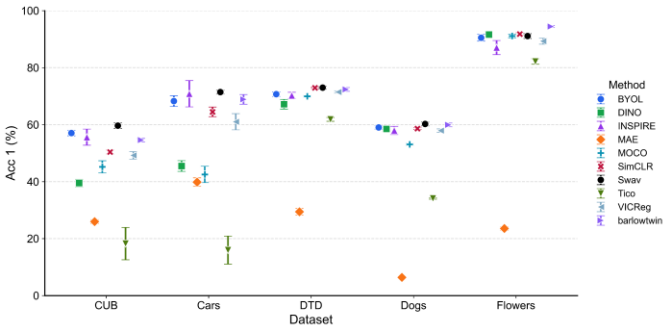


Fig. 11. Mean and standard deviation of the Top-1 validation accuracy across 4 randomized runs per dataset.

Note that, for the experiments presented in this section, we performed four independent runs for each method on each dataset. Fig. 11 reports the mean and standard deviation of the Top-1 validation accuracy to illustrate the consistency and stability of the results. All results presented throughout the study are reported as mean performance across multiple independent executions. To maintain the focus and readability of the study, we do not further analyze these statistics. Future work may investigate the variability and robustness of different SSL methods, including the proposed INSPIRE Method, in greater detail.

VI. Results: Representation Interpretability

While the previous sections (Sections IV and V) evaluate SSL methods through downstream performance, this section adopts a representation-level interpretability perspective to directly investigate the behavior of learned representations. Building on the methodology of our framework introduced in Section III.C, we investigate how representations learned by different SSL methods respond to controlled perturbations across network layers and spatial scales. This analysis complements downstream evaluation by characterizing representation sensitivity and robustness under varying perturbation conditions, providing insight into how different SSL objectives may influence representation behavior.

The top panel of Fig. 12 presents the average CKA similarity between representations learned by different SSL methods, aggregated across all layers. The results show moderate similarity across methods (≈0.30–0.52), indicating that while SSL approaches share common representational structure, they also exhibit method-specific differences. Methods based on invariance learning, including both contrastive and redundancy-reduction approaches such as SimCLR, VICReg, and Barlow Twins, exhibit the highest mutual similarity (e.g., VICReg–SimCLR: 0.52, SimCLR–Barlow Twins: 0.51, VICReg–Barlow Twins: 0.50), suggesting that despite differences in their objectives, they produce similar geometric constraints on the representation space.

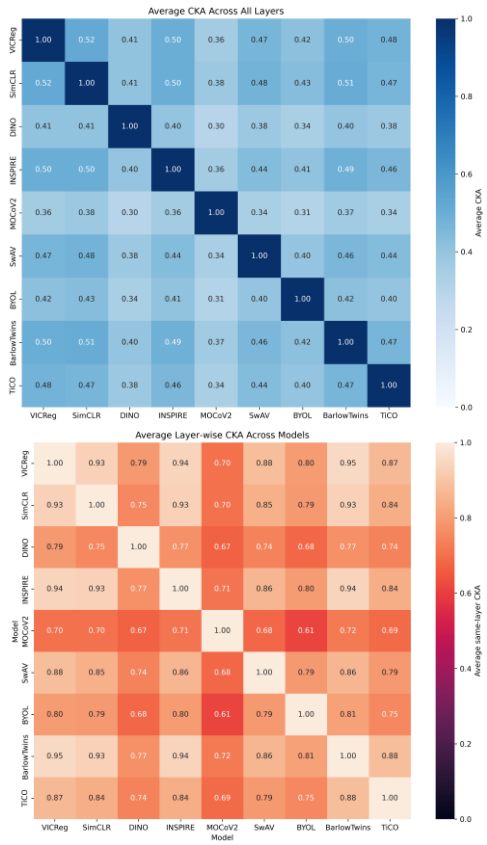


Fig. 12. Average CKA across SSL models. The top panel shows average CKA across all layers, while the bottom panel shows average layer-wise CKA between corresponding layers.

In contrast, DINO consistently shows lower similarity with other methods (≈0.38–0.41 with most approaches), reflecting its distinct self-distillation objective, which encourages clustering-like structure and semantic grouping rather than strict instance-level discrimination. MoCoV2 also exhibits relatively lower similarity (≈0.30–0.38), suggesting that differences in training mechanisms, such as the use of momentum encoders and memory queues, may contribute to distinct representation structures. Notably, INSPIRE shows strong alignment with invariance-based methods (e.g., 0.50 with SimCLR and VICReg, 0.49 with Barlow Twins) while maintaining measurable differences from other approaches (≈0.40 with DINO and 0.36 with MoCoV2), suggesting that it preserves core contrastive representation properties while introducing controlled variation in spatial representation learning.

The bottom panel of Fig. 12 provides a complementary perspective by measuring similarity between corresponding layers, revealing a notable contrast with the global comparison in the top panel of Fig. 12. While overall similarity is moderate when aggregating across layers, the layer-wise analysis shows consistently high similarity (≈0.61–0.95) when comparing aligned depths. This indicates that representations learned by different SSL methods are more comparable at corresponding layers than across arbitrary layers. The discrepancy between the two figures suggests that part of the observed differences across methods arises from cross-layer misalignment rather than uniformly distinct representations. Notably, even methods that appear less similar in the global comparison, such as DINO and MoCoV2, exhibit substantial alignment when evaluated layer-wise (e.g., DINO ≈0.68–0.79, MoCoV2 ≈0.61–0.72), further supporting the view that representation differences are not uniform across depth.

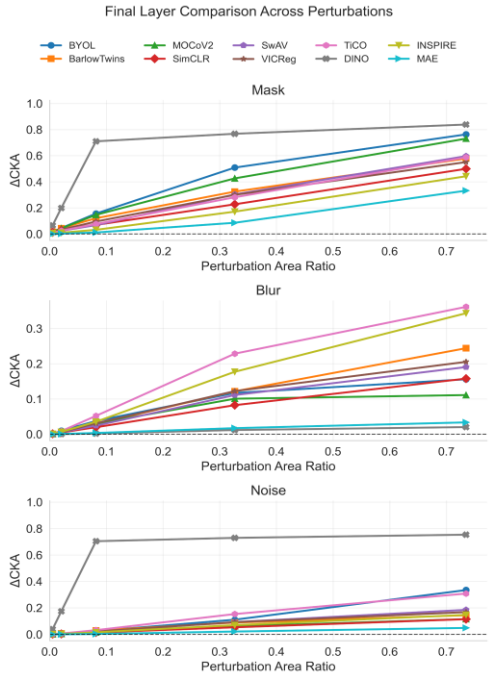


Fig. 13. Final layer representation sensitivity across spatial scales for different SSL models.

Fig. 13 presents the final-layer representation sensitivity across spatial scales for different SSL models under three types of perturbations: masking, blur, and additive noise. For each perturbation, the x-axis denotes the normalized perturbation area ratio, while the y-axis shows the corresponding change in representation measured by ΔCKA. Each curve corresponds to a different SSL method, allowing direct comparison of how representation similarity degrades as the spatial extent of the perturbation increases. The figure summarizes results over 5,000 ImageNet-1K images, providing a stable estimate of model behavior across scales.

Across all perturbation types, a consistent pattern is observed representation change increases with perturbation area, indicating that larger spatial disruptions are associated with greater deviation in learned features. However, the rate and shape of this increase differ substantially across models and perturbation types. Under masking and noise, most models exhibit a largely monotonic increase in ΔCKA, suggesting a progressive accumulation of disruption as more of the image is affected. For example, under masking at an area ratio of 0.7, BYOL reaches approximately ΔCKA ≈ 0.76, while MAE remains substantially lower at around ΔCKA ≈ 0.33, highlighting a clear difference in robustness. In contrast, blur produces a more nuanced response, with several models showing stronger sensitivity at intermediate scales; for instance, TiCo reaches ΔCKA ≈ 0.23 at area ratio 0.3 and further increases to ≈ 0.36 at 0.7, reflecting the selective degradation of mid-frequency structures.

Importantly, models differ in their sensitivity profiles. MAE consistently shows the lowest ΔCKA across all scales and perturbations, indicating strong robustness to both localized and distributed corruption. In contrast, models such as BYOL, TiCo, and MoCoV2 exhibit higher sensitivity, particularly as perturbation extent increases. INSPIRE presents an interesting intermediate behavior: under masking, it shows relatively lower sensitivity at small scales (ΔCKA ≈ 0.04 at 0.1) but increases steadily to ≈ 0.44 at 0.7, consistent with a gradual accumulation of sensitivity. Under blur, INSPIRE is among the most sensitive models at larger scales (reaching ΔCKA ≈ 0.34 at 0.7), suggesting that it is particularly affected by the loss of spatial frequency information. Similarly, under noise, INSPIRE exhibits moderate sensitivity (≈ 0.15 at 0.7), positioning it between highly robust models like MAE and more sensitive ones like BYOL and TiCo.

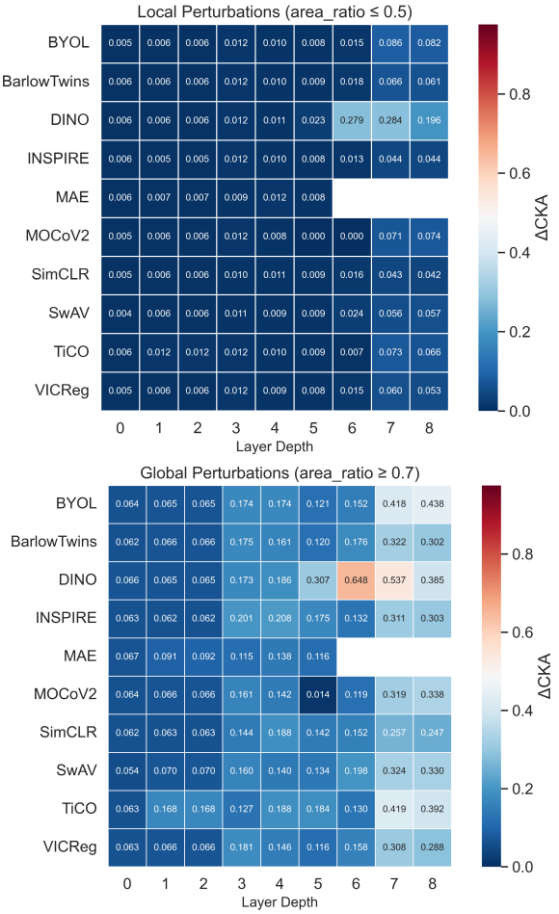


Fig. 14. Layer-wise representation sensitivity (ΔCKA) under local (top) and global (bottom) perturbations.

Fig. 14 presents the layer-wise representation sensitivity (ΔCKA) across models under local and global perturbations. The top panel summarizes local perturbations (area ratio ≤ 0.5), while the bottom panel shows global perturbations (area ratio ≥ 0.7). Each row corresponds to an SSL model, and each column represents a layer indexed by depth. The values indicate the average ΔCKA aggregated over perturbation scales, providing a compact view of how sensitivity evolves across the network hierarchy.

Under local perturbations, all models exhibit very low ΔCKA in early layers (typically ≈ 0.005–0.012), indicating strong stability to small-scale changes at shallow depths. As depth increases, sensitivity gradually rises but remains relatively modest for most models. For example, BYOL increases from ≈ 0.005 in early layers to ≈ 0.086–0.082 in deeper layers, while INSPIRE shows a similar but more controlled increase from ≈ 0.006 to ≈ 0.044. Notably, MAE remains highly stable across all layers, with ΔCKA consistently below ≈ 0.012 in early layers and only reaching ≈ 0.008 before missing deeper entries, reinforcing its robustness to localized perturbations. In contrast, DINO exhibits a sharp increase in deeper layers (≈ 0.279–0.284), suggesting heightened sensitivity to local disruptions at high-level representations.

Under global perturbations, the sensitivity increases significantly across all models, particularly in deeper layers. Early layers remain relatively stable (≈ 0.06–0.09 across models), but deeper layers show substantial divergence. For instance, BYOL increases to ≈ 0.418–0.438 in the deepest layers, while TiCo reaches ≈ 0.419, indicating strong sensitivity to large-scale changes. INSPIRE again exhibits an intermediate behavior: it shows higher sensitivity than MAE but remains more controlled than highly sensitive models, increasing from ≈ 0.063 in early layers to ≈ 0.311–0.303 in deeper layers. This consistent, gradual increase suggests that INSPIRE balances robustness and sensitivity, responding progressively to larger-scale disruptions rather than exhibiting abrupt changes. MAE, in contrast, maintains relatively low sensitivity even under global perturbations (≈ 0.115–0.138 in mid layers), suggesting strong invariance across scales.

The observed sensitivity profiles suggest that SSL representations are neither strictly local nor fully global in their behavior. Representations dominated by localized visual information would be expected to exhibit strong sensitivity to small perturbations, whereas representations dominated by globally aggregated image information would be expected to remain largely unaffected by localized modifications. The measured ΔCKA trends do not fully conform to either extreme. Across multiple methods, representation changes are already observable at small perturbation scales and continue to increase as perturbation extent grows. For instance, DINO exhibits substantial sensitivity even when only a small fraction of the image is perturbed, while methods such as BYOL and SimCLR show progressively larger representation changes as perturbation extent increases (Fig. 13). These observations suggest that SSL representations cannot be adequately described by either strictly local or strictly global forms of spatial encoding. Instead, the results are more consistent with a spectrum of spatial sensitivity behaviors, with different SSL objectives occupying different positions along this continuum.

The experiment also reveals meaningful differences between SSL objectives. DINO exhibits comparatively strong responses even under relatively localized perturbations ($\alpha \leq 0.1$), particularly in deeper layers, suggesting greater sensitivity to spatially structured image content (Fig. 14). In contrast, MAE remains comparatively stable across perturbation scales and network depth (Fig. 13 and Fig. 14), even under large-scale masking. INSPIRE occupies an intermediate position between these extremes: it remains relatively stable under localized

perturbations while progressively increasing in sensitivity at larger scales. Differences across perturbation types further suggest that representation sensitivity depends not only on perturbation extent, but also on the structural nature of the perturbations. In particular, blur consistently produces substantially smaller representation changes than masking, with blur sensitivity peaking near ΔCKA ≈ 0.35 while masking approaches ≈ 0.8 for several models (Fig. 13). This observation suggests that representations remain comparatively stable when coarse spatial structure is preserved, even though fine-grained image details are degraded.

The layer-wise analysis further suggests that spatial integration evolves progressively across network depth (Fig. 14). Early layers remain comparatively stable under both local and global perturbations, whereas deeper layers exhibit substantially stronger sensitivity, particularly for DINO, BYOL, and TiCo. The increasing perturbation sensitivity observed in deeper layers is consistent with progressively broader spatial integration across the network hierarchy, where later representations depend on information distributed across larger spatial extents, a behavior that has been widely observed in hierarchical vision models and attributed to increasing effective receptive field size and contextual feature integration across depth [55].

The results suggest that SSL methods do not converge to a single form of spatial invariance. Different SSL objectives exhibit distinct sensitivity profiles across perturbation scales and network depth, while masking, blur, and noise produce markedly different ΔCKA trends (Fig. 13 and Fig. 14). These consistent differences suggest that spatial sensitivity is an objective-dependent property of the learned representation. Rather than sharing a common notion of invariance, SSL methods appear to occupy different positions on a spatial sensitivity–invariance spectrum.

VII. Computational Cost and Runtime Analysis

In this section, we describe the computational setup and report run-time results for both single- and multi-GPU pretraining experiments. Our goal is to provide a clear account of the hardware environment, training infrastructure, and parallelization strategy to ensure reproducibility and to contextualize the reported run-time characteristics.

All experiments were conducted on a high-performance computing environment consisting of seven compute nodes. Each node is equipped with dual Intel Xeon Silver 4314 processors (16 cores per CPU), 503 GB of system memory, and one NVIDIA A100-PCIE-40GB GPU. Single-GPU experiments were run on a single node using one A100 GPU, with CPUs used for data loading and preprocessing. For multi-GPU pretraining, we employed PyTorch DDP [56] in a multi-node configuration with one GPU per node, where each process handled a portion of the mini-batch and gradients were synchronized across nodes at every iteration. This setup enabled efficient scaling while maintaining consistent optimization behavior across experiments.

A. Single-GPU Pretraining Experiments

For the ImageNet-100 experiments with a ResNet-18 backbone, we adopt a single-GPU training configuration to ensure controlled and consistent comparisons across all variants. Each experiment is executed on a node equipped with an NVIDIA A100 GPU with 40 GB of memory, which provides sufficient capacity for the batch size and augmentation pipeline used in this study. Training is performed for 100 epochs with a batch size of 256 per GPU. Under this configuration, all INSPIRE variants and the SimCLR baseline are trained using identical hardware and resource allocation, ensuring that performance differences are not influenced by variations in computational setup. Data loading and preprocessing are handled on the CPU, while all forward and backward computations are performed on the GPU.

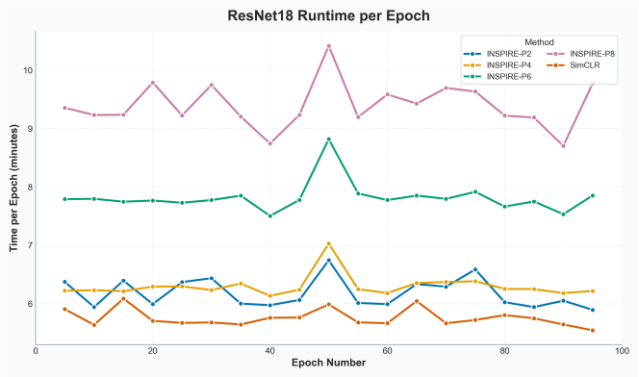


Fig. 15. Training time per epoch for SimCLR and INSPIRE with different patch configurations using a ResNet18 backbone.

Fig. 15 presents the training runtime per epoch for SimCLR and the proposed INSPIRE method under different patch configurations (P = 2, 4, 6, and 8) using a ResNet18 backbone. The training time remains stable across epochs for all configurations, indicating consistent computational behavior throughout the optimization process. SimCLR exhibits the lowest runtime, requiring approximately 5.7 minutes per epoch on average. For INSPIRE, the runtime increases gradually with the number of patches. Specifically, INSPIRE-P2 requires roughly 6.1 minutes per epoch, while INSPIRE-P4 requires about 6.3 minutes. Larger patch configurations introduce additional computational overhead, with INSPIRE-P6 and INSPIRE-P8 requiring approximately 7.8 minutes and 9.3 minutes per epoch, respectively. These results highlight the trade-off between model complexity and computational cost: increasing the number of patches leads to higher training time due to the additional feature processing required by the merge mechanism. Nevertheless, the runtime remains stable across epochs, suggesting that the proposed framework scales predictably with respect to the number of patches.

B. Multi-GPU Parallel Pretraining Experiments

For the ImageNet-1K experiment, we train two variants of INSPIRE with batch sizes 128 and 256 and compare them against SimCLR using a ResNet-50 backbone. All models are trained using PyTorch DDP across 7 GPUs (A100-PCIE-40GB), where each GPU processes its share of the global batch and gradients are synchronized at every step. This setup ensures consistent optimization while maximizing parallel efficiency across nodes.

Fig. 16 presents the training time per epoch (in minutes) on the y-axis versus the epoch number (1–100) on the x-axis. From

the figure, SimCLR consistently runs between approximately 7 and 10 minutes per epoch, with very small fluctuations across training. In comparison, INSPIRE, with batch size 256, shows epoch times that are slightly higher but still relatively close, generally falling in the range of ~11 to 13 minutes as well, with minor variability and a brief warm-up period in the early epochs. This suggests that INSPIRE (256) achieves a comparable level of efficiency to SimCLR in terms of per-epoch runtime. On the other hand, INSPIRE, with batch size 128, operates at a significantly higher cost, with epoch times consistently around ~24 to 26 minutes, and a noticeable spike reaching approximately 27 minutes in one of the early epochs. This clear separation highlights the impact of batch size: reducing the batch size effectively doubles the number of iterations per epoch, leading to a substantial increase in runtime and slightly higher variability.

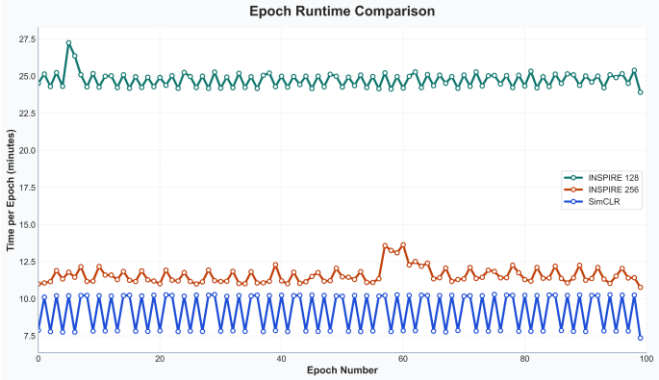


Fig. 16. Training time per epoch for SimCLR and INSPIRE with different batch sizes using a ResNet50 backbone (parallel runtime on 7 GPUs).

These per-epoch trends are consistent with the total training times. SimCLR completes training in 15.27 hours, while INSPIRE Method with batch size 256 takes 19.29 hours, reflecting its slightly higher or comparable per-epoch cost over the full run. In contrast, INSPIRE, with batch size 128, requires 41.26 hours, more than twice as long due to the inefficiency introduced by the smaller batch size.

Scaling from a single-GPU to a multi-node distributed setup plays a critical role in enabling large-scale training on ImageNet-1K. In our experiments, ImageNet-100 training with a ResNet-18 backbone on a single A100 GPU requires approximately 5.7 minutes per epoch for SimCLR, with INSPIRE variants ranging from about 6.1 to 9.8 minutes depending on configuration. In contrast, for the significantly larger ImageNet-1K dataset, we employ PyTorch DDP across 7 A100 GPUs, where each GPU processes a portion of the global batch and gradients are synchronized at every step. Under this setup, SimCLR achieves per-epoch times of approximately 7–10 minutes, while INSPIRE with batch size 256 operates at around 11–13 minutes per epoch. Despite the roughly 10× increase in dataset size, the per-epoch training time increases only modestly, demonstrating the effectiveness of distributed scaling in maintaining practical training times. These results demonstrate that multi-GPU distributed training is essential for scaling contrastive learning to ImageNet-1K, effectively compressing what would otherwise be prohibitively long training times into a manageable range of ~15–19 hours, while maintaining stable and predictable computational behavior across epochs.

VIII. Discussion and Conclusion

This work presents the INSPIRE Framework, a systematic approach for investigating local-global representation learning in self-supervised learning through controlled spatial granularity manipulation during training, downstream evaluation, and representation interpretability using scale-controlled spatial perturbation analysis. While previous research has improved SSL through new objectives [12], [28], architectures [13], [23], and local representation learning strategies [8], [9], it has remained difficult to isolate how training design influences the balance between local and global representations. Our results extend existing studies on representation learning and interpretability by providing a unified experimental framework for systematically analyzing these relationships across multiple evaluation protocols. Although INSPIRE does not achieve the highest performance on every benchmark, it remains consistently competitive across diverse evaluation settings while, more importantly, providing new insights into local-global representation learning, representation sensitivity, and the behavior of learned representations across diverse evaluation protocols.

A central finding of this work is that local-global representation learning is better understood as a continuum than as a binary distinction. The experiments suggest that representations are not inherently local or global; instead, different SSL training strategies occupy different positions along a continuum that reflects how local and global information is balanced within the learned representation. Methods achieving comparable ImageNet performance often exhibit markedly different behavior on fine-grained recognition tasks and under controlled perturbations, supporting that similar global semantic performance can arise from substantially different representational characteristics. This observation provides a more nuanced perspective on SSL representation learning than conventional benchmark comparisons alone.

The controlled spatial granularity experiments further show that the relationship between granularity and representation quality is non-monotonic. Increasing the number of merged regions initially improves representation quality, but excessive fragmentation eventually degrades both global and fine-grained transfer performance. The consistently strong performance of the intermediate P4 configuration suggests that effective representation learning emerges from balancing contextual diversity with semantic coherence rather than maximizing either property independently. This finding indicates that spatial granularity itself represents an important design variable for SSL, independent of network architecture or learning objective.

The combined downstream and representation-level analyses further demonstrate that representation similarity and representation behavior capture complementary aspects of learned representations. Models exhibiting relatively similar representation geometry do not necessarily respond similarly to controlled spatial perturbations, indicating that representation similarity alone provides an incomplete characterization of learned features. Likewise, methods displaying very different perturbation sensitivity may achieve comparable downstream

accuracy. These observations suggest that understanding SSL representations requires considering both the structure of representation space and how representations respond to controlled changes in the input, providing a richer characterization than benchmark accuracy or similarity analysis alone.

The relationship between perturbation sensitivity and downstream transfer also appears more complex than a simple trade-off between invariance and sensitivity. Neither highly sensitive nor highly invariant representations consistently produce the strongest fine-grained recognition performance. Instead, successful transfer appears to depend on maintaining discriminative local information while remaining robust to irrelevant visual variation. This observation reinforces the importance of balanced representations and suggests that representation behavior should be considered alongside conventional evaluation metrics when assessing SSL methods.

Beyond the empirical findings, the proposed INSPIRE Framework has broader implications for SSL research. Rather than introducing only another training method, the framework provides a general methodology for systematically investigating how training strategies influence representation learning. By integrating controlled training, downstream evaluation, and representation interpretability within a unified framework, INSPIRE offers a reusable experimental approach that can support future investigations of emerging SSL methods, alternative architectures, and different application domains. More broadly, the work contributes to the growing shift toward understanding learned representations rather than evaluating them solely through downstream benchmark performance.

Future research can extend the INSPIRE Framework in several directions. The proposed representation interpretability investigation methodology can be complemented with additional representation analysis techniques, including SVCCA, PWCCA, RSA, and probing-based methods, to provide broader perspectives on learned representations. Extending the evaluation to dense prediction tasks such as object detection, semantic segmentation, and instance segmentation would further clarify how local-global representation learning influences spatially intensive vision problems. Future studies should also investigate whether the observed local-global representation continuum and spatial sensitivity characteristics generalize across Vision Transformers, hybrid architectures, multimodal foundation models, and other emerging SSL paradigms. Future work should also investigate the influence of batch size, since the contrastive SSL baselines considered in this work use substantially different effective batch sizes, ranging from 512 to several thousand samples, while INSPIRE employs a fixed effective batch size of 896. We hope that INSPIRE provides a useful foundation for future research on representation learning, interpretability, and the principled design and evaluation of self-supervised learning methods.

ACKNOWLEDGMENT

This article 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-. The authors gratefully acknowledge the support of the High-Performance Computing Center (Aziz Supercomputer) at King Abdulaziz University (http://hpc.kau.edu.sa) for providing access to computational resources, including A100 GPUs, which were essential for the distributed data parallel (DDP) training of our models.

REFERENCES

[1] M. H. Mozaffari, “Deep Learning for Computer Vision Application,” Electronics 2025, Vol. 14, vol. 14, no. 14, Jul. 2025, doi: 10.3390/ELECTRONICS14142874.

[2] M. Awais et al., “Foundation Models Defining a New Era in Vision: A Survey and Outlook,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 47, no. 4, pp. 2245–2264, 2025, doi: 10.1109/TPAMI.2024.3506283.

[3] 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.

[4] T. Uelwer et al., “A survey on self-supervised methods for visual representation learning,” Machine Learning 2025 114:4, vol. 114, no. 4, pp. 111-, Mar. 2025, doi: 10.1007/s10994-024-06708-7.

[5] C.-Y. Hsieh, C.-J. Chang, F.-E. Yang, and Y.-C. F. Wang, “Self-Supervised Pyramid Representation Learning for Multi-Label Visual Analysis and Beyond,” Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2696–2705, 2023.

[6] T. Lebailly and T. Tuytelaars, “Global-local self-distillation for visual representation learning,” in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 1441–1450.

[7] T. Dang, S. Kornblith, H. T. Nguyen, P. Chin, and M. Khademi, “A Study on Self-Supervised Object Detection Pretraining,” Jul. 2022, [Online]. Available: http://arxiv.org/abs/2207.04186

[8] Y. Xu, Q. Zhang, J. Zhang, and D. Tao, “RegionCL: Exploring Contrastive Region Pairs for Self-supervised Representation Learning,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 13693 LNCS, pp. 477–494, 2022, doi: 10.1007/978-3-031-19827-4_28.

[9] T. Xiao, C. J. Reed, X. Wang, K. Keutzer, and T. Darrell, “Region similarity representation learning,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10539–10548.

[10] A. Javidani, M. A. Sadeghi, and B. Nadjar Araabi, “Patch-wise self-supervised visual representation learning: a fine-grained approach,” Signal, Image and Video Processing 2025 19:6, vol. 19, no. 6, pp. 458-, Apr. 2025, doi: 10.1007/S11760-025-04020-Y.

[11] H. Xu, X. Zhang, H. Li, L. Xie, H. Xiong, and Q. Tian, “Seed the Views: Hierarchical Semantic Alignment for Contrastive Representation Learning,” Dec. 2020, [Online]. Available: http://arxiv.org/abs/2012.02733

[12] J.-B. Grill et al., “Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning,” Adv. Neural Inf. Process. Syst., vol. 33, pp. 21271–21284, 2020, Accessed: Mar. 11, 2026. [Online]. Available: https://github.com/deepmind/deepmind-research/tree/master/byol

[13] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A Simple Framework for Contrastive Learning of Visual Representations,” in International conference on machine learning, Nov. 2020, pp. 1597–1607. Accessed: Mar. 11, 2026. [Online]. Available: https://proceedings.mlr.press/v119/chen20j.html

[14] K. Kalasampath, K. N. Spoorthi, S. Sajeev, S. S. Kuppa, K. Ajay, and A. Maruthamuthu, “A Literature Review on Applications of Explainable Artificial Intelligence (XAI),” IEEE Access, vol. 13, pp. 41111–41140, 2025, doi: 10.1109/ACCESS.2025.3546681.

[15] S. Alghamdi, R. Mehmood, F. A. Alqurashi, and A. Alzahrani, “Paving the Roadmap for XAI and IML in Healthcare: Data-Driven Discoveries and the FIXAIH Framework,” IEEE Access, vol. 13, pp. 174393–174427, 2025, doi: 10.1109/ACCESS.2025.3616353.

[16] J. Colin ELLIS Alicante, L. Goetschalckx, N. Oliver, E. Alicante, and S. Thomas Serre, “Capability/= Interpretability: Human Interpretability of Vision Foundation Models,” 2026, doi: https://doi.org/10.48550/arXiv.2605.20337.

[17] T. G. Grigg, D. Busbridge, J. Ramapuram, and R. W. Apple, "Do Self-Supervised and Supervised Methods Learn Similar Visual Representations?," Dec. 2021, Accessed: Mar. 27, 2026. [Online]. Available: https://arxiv.org/pdf/2110.00528

[18] A. Luthra, P. Mishra, and T. Galanti, "On the Alignment Between Supervised and Self-Supervised Contrastive Learning," Oct. 2025, Accessed: Mar. 27, 2026. [Online]. Available: https://arxiv.org/pdf/2510.08852

[19] W. Oleszkiewicz, D. Basaj, T. Trzciński, and B. Zieliński, "Which Visual Features Impact the Performance of Target Task in Self-supervised Learning?," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 13350 LNCS, pp. 331–344, 2022, doi: 10.1007/978-3-031-08751-6_24/SAVE-RESEARCH.

[20] W. Oleszkiewicz et al., "Visual Probing: Cognitive Framework for Explaining Self-Supervised Image Representations," IEEE Access, vol. 11, pp. 13028–13043, 2023, doi: 10.1109/ACCESS.2023.3242982.

[21] M. Caron et al., "Emerging Properties in Self-Supervised Vision Transformers," in In Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 9650–9660. [Online]. Available: https://github.com/facebookresearch/dino

[22] H. Choi, S. Jin, and K. Han, "Adversarial Normalization: I Can Visualize Everything (ICE)," in In Proceedings of the IEEE/cvf conference on computer vision and pattern recognition, 2023, pp. 12115–12124.

[23] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 9726–9735, 2020, doi: 10.1109/CVPR42600.2020.00975.

[24] L. Zheng, B. Jing, Z. Li, H. Tong, and J. He, "Heterogeneous Contrastive Learning for Foundation Models and Beyond," Mar. 2024, [Online]. Available: http://arxiv.org/abs/2404.00225

[25] X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li, "Dense Contrastive Learning for Self-Supervised Visual Pre-Training," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition., 2021, pp. 3024–3033. [Online]. Available: https://git.io/

[26] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments," Adv. Neural Inf. Process. Syst., vol. 33, pp. 9912–9924, 2020, Accessed: Mar. 11, 2026. [Online]. Available: https://github.com/facebookresearch/swav

[27] J. Zhu, R. M. Moraes, S. Karakulak, V. Sobol, A. Canziani, and Y. LeCun, "TiCo: Transformation Invariance and Covariance Contrast for Self-Supervised Visual Representation Learning," Jun. 2022, Accessed: Mar. 11, 2026. [Online]. Available: http://arxiv.org/abs/2206.10698

[28] R. Yoshihashi, S. Nishimura, D. Yonebayashi, Y. Otsuka, T. Tanaka, and T. Miyazaki, "Ladder Siamese Network: a Method and Insights for Multi-level Self-Supervised Learning," in IEEE International Conference on Image Processing (ICIP), 2023, pp. 211–215.

[29] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow Twins: Self-Supervised Learning via Redundancy Reduction," in International conference on machine learning, PMLR, Jul. 2021, pp. 12310–12320. Accessed: Mar. 11, 2026. [Online]. Available: https://proceedings.mlr.press/v139/zbontar21a.html

[30] A. Bardes, J. Ponce, and Y. LeCun, "VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning," ICLR 2022 - 10th International Conference on Learning Representations, Jan. 2022, Accessed: Mar. 11, 2026. [Online]. Available: http://arxiv.org/abs/2105.04906

[31] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition., 2022, pp. 16000–16009.

[32] S. Yun, H. Lee, J. Kim, and J. Shin, "Patch-Level Representation Learning for Self-Supervised Vision Transformers," in IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 8354–8363.

[33] Q. Guo, Y. Yu, Y. Jiang, J. Wu, Z. Yuan, and P. Luo, "Multi-Level Contrastive Learning for Dense Prediction Task," Apr. 2023, [Online]. Available: http://arxiv.org/abs/2304.02010

[34] Q. Su and S. Ji, "Unsqueeze [CLS] Bottleneck to Learn Rich Representations," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 15100 LNCS, pp. 19–37, 2025, doi: 10.1007/978-3-031-72946-1_2/SAVE-RESEARCH.

[35] Z. Zhang and X. Gong, "Positional Label for Self-Supervised Vision Transformer," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 3, pp. 3516–3524, Jun. 2023, doi: 10.1609/AAAI.V37I3.25461.

[36] E. Peruzzo et al., "Spatial entropy as an inductive bias for vision transformers," Machine Learning 2024 113:9, vol. 113, no. 9, pp. 6945–6975, Jul. 2024, doi: 10.1007/S10994-024-06570-7.

[37] J. Zhou et al., "iBOT: Image BERT Pre-Training with Online Tokenizer," ICLR 2022 - 10th International Conference on Learning Representations, Nov. 2021, Accessed: May 13, 2026. [Online]. Available: https://arxiv.org/pdf/2111.07832

[38] Y. Zhan, Y. Zhao, C. Luo, Y. Zhang, and X. Sun, "Attention-Guided Contrastive Masked Image Modeling for Transformer-Based Self-Supervised Learning," Proceedings - International Conference on Image Processing, ICIP, pp. 2490–2494, 2023, doi: 10.1109/ICIP49359.2023.10222606.

[39] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton, "Similarity of neural network representations revisited," Proceedings of the 36th International Conference on Machine Learning (ICML), pp. 3519–3529, 2019, Accessed: Mar. 26, 2026. [Online]. Available: http://proceedings.mlr.press/v97/kornblith19a.html

[40] M. Raghu, J. Gilmer, J. Yosinski, and Sohl-Dickstein, "Svcca: Singular vector canonical correlation analysis for deep understanding and improvement," Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 6076–6085, 2017, Accessed: Mar. 27, 2026. [Online]. Available: https://yosinski.com/media/papers/Raghu__2017__SVCCA_Singular_Vector_Canonical_Correlation_Analysis.pdf

[41] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?," proceedings.neurips.ccM Raghu, T Unterthiner, S Kornblith, C Zhang, A DosovitskiyAdvances in neural information processing systems, 2021•proceedings.neurips.cc, Accessed: Mar. 27, 2026. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/hash/652cf38361a209088302ba2b8b7f51e0-Abstract.html

[42] A. Li, A. Efros, and D. Pathak, "Understanding Collapse in Non-contrastive Siamese Representation Learning," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 13691 LNCS, pp. 490–505, 2022, doi: 10.1007/978-3-031-19821-2_28/SAVE-RESEARCH.

[43] N. Kalibhat, K. Narang, H. Firooz, M. Sanjabi, and S. Feizi, "Measuring self-supervised representation quality for downstream classification using discriminative features," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 13031–13039, 2024, Accessed: Mar. 27, 2026. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/29201

[44] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in Advances in neural information processing systems, 2016. Accessed: Mar. 27, 2026. [Online]. Available: https://proceedings.neurips.cc/paper/2016/hash/6b180037abbebea991d8b1232f8a8ca9-Abstract.html

[45] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 248–255, 2009, doi: 10.1109/CVPR.2009.5206848.

[46] M. E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," Proceedings - 6th Indian Conference on Computer Vision, Graphics and Image Processing, ICVGIP 2008, pp. 722–729, 2008, doi: 10.1109/ICVGIP.2008.47.

[47] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, "Novel Dataset for Fine-Grained Image Categorization: Stanford Dogs," Proc. CVPR workshop on fine-grained visual categorization (FGVC)., vol. 2, 2011, Accessed: Dec. 23, 2025. [Online]. Available: http://vision.stanford.edu/aditya86/StanfordDogs/

[48] M. Cimpoi, S. Maji, I. Kokkinosécole, S. Mohamed, and A. Vedaldi, "Describing Textures in the Wild," InProceedings of the IEEE conference on computer vision and pattern recognition, pp. 3606–3613, 2014, Accessed: Dec. 23, 2025. [Online]. Available: http://www.robots.ox.ac.uk/

[49] W. Catherine, B. Steve, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," California Institute of Technology, 2011.

[50] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D Object Representations for Fine-Grained Categorization," InProceedings of the IEEE international conference on computer vision workshops, pp. 554–561, 2013, doi: 10.1109/ICCVW.2013.77.

[51] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," Computer Vision–ECCV 2014, vol. 8689, no. PART 1, pp. 818–833, 2014, doi: 10.1007/978-3-319-10590-1_53.

[52] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," in International conference on learning representations, 2018. Accessed: Apr. 01, 2026. [Online]. Available: https://github.com/rgeirhos/texture-vs-shape

[53] D. Hendrycks and T. Dietterich, "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations," 7th International Conference on Learning Representations, ICLR 2019, Mar. 2019, Accessed: Apr. 01, 2026. [Online]. Available: https://arxiv.org/pdf/1903.12261

[54] "Benchmarks — LightlySSL 1.5.22 documentation." Accessed: Dec. 24, 2025. [Online]. Available: https://docs.lightly.ai/self-supervised-learning/getting_started/benchmarks.html

[55] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the Effective Receptive Field in Deep Convolutional Neural Networks," Adv. Neural Inf. Process. Syst., vol. 29, 2016.

[56] A. Paszke et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," Neural Information Processing Systems, 2019.

Enhancing Diabetic Retinopathy Classification Through Advanced Deep Convolutional Neural Network Architectures

Sifeddine Elkardoudi[1], Khalil Ladrham[2], Ahmed Eddaoui[3], Mohamed Talea[4]
Information Processing Laboratory-Faculty of Sciences Ben M'sik, University Hassan II, Casablanca, Morocco[1, 3, 4]
Intelligent Processing and Security of Systems Team-Faculty of Sciences Rabat, Mohammed V University, Rabat, Morocco[2]

***Abstract*—Diabetic retinopathy is one of the major causes of vision loss in the world; thus, the development of early diagnosis systems based on artificial intelligence is a medical and scientific necessity. If detected late, it can lead to permanent vision loss and is considered one of the most serious vascular complications of diabetes. Diagnosis is traditionally made by visual inspection of fundus images by an expert medical professional and is time-consuming. Deep learning techniques have shown promise to help detect disease early and improve the accuracy of medical decision-making. This study provides an experimental comparison of four convolutional neural network architectures (Xception, Inception-v3, MobileNet-v3 and ResNet-50), using a balanced dataset of 25,000 colour fundus images across five classes. Of the total images, 80% were used for training, 10% for validation, and 10% for testing. A standardised image size of 1024×768×3 was used to preserve the fine details of diabetic retinopathy. The models were evaluated for a number of metrics such as accuracy, loss, recall, precision, F1-score, AUC, Cohen's Kappa coefficient and MCC in addition to computational cost and training time. The results indicate that the Xception model outperformed the other models, with a test accuracy of 98.32%, a validation accuracy of 98.52%, the lowest loss rate, and the highest AUC value. Inception-v3 achieved the second-best performance with 95.04% accuracy, and MobileNet-v3 achieved the lowest computational cost, making it suitable for resource- constrained applications, with 90% accuracy. Conversely, ResNet-50 had lower accuracy with high computational cost, which shows that the increased architectural complexity does not necessarily ensure improved performance. The results show that the best model selection for diabetic retinopathy classification depends on a trade-off between diagnostic accuracy, computational efficiency, and practical usability, with Xception being the most suitable for tasks requiring the highest level of accuracy, and MobileNet-v3 as a viable alternative for resource-constrained environments.**

***Keywords*—Diabetic retinopathy; fundus images; deep learning; xception; inception-v3; mobilenet-v3; resnet-50; training time**

I. Introduction

Diabetic retinopathy is among the most severe microvascular complications of diabetes and the leading preventable cause of vision loss worldwide. Early detection is vital as the early stage of the disease can be asymptomatic and progress without obvious symptoms. Early diagnosis and appropriate treatment can reduce the risk of blindness and improve clinical outcomes [4]. In the traditional approach, diabetic retinopathy detection depends on the ophthalmologists' clinical examination of fundus images, which is effective but time-consuming and relies on human expertise and is problematic in resource-limited settings or those with a high disease burden [2], [5]. These limitations have led to the development of computer-aided diagnostic systems that can automatically analyse retinal images and provide rapid and accurate preliminary diagnoses [6].

With the rapid development of artificial intelligence, deep learning, especially convolutional neural networks (CNNs), has been one of the most successful methods for analyzing medical images. These models have been able to learn complex visual features directly from the retinal images, including microvessels, haemorrhages, exudates and microaneurysms, with less dependency on manual feature engineering [3], [7], [8].

Recent studies have demonstrated that architectures such as VGG, ResNet, DenseNet, Xception and EfficientNet can achieve high accuracy on retinal image classification, while transfer learning and data augmentation have helped improve generalisation and minimise the problem of overfitting, especially when labelled medical data is scarce [4], [7], [9]. However, despite these successes, the problem of diabetic retinopathy classification still suffers from the fundamental challenges of class imbalance, image quality variations and the challenge of distinguishing between clinically similar disease stages, on top of the need for interpretable, computationally efficient models that can be deployed in real-time clinical settings [5], [11]. In this respect, recent works have moved from single image classification to more sophisticated models, such as hybrid models that combine CNNs with LSTMs or attention mechanisms, and multimodal models that combine retinal images with clinical or demographic information, in order to improve accuracy, robustness and clinical relevance [6], [12]. Based on the above, the design of a comparative and systematic framework that integrates the strength of deep architectures, interpretability, computational efficiency and generalisation over different datasets has become a scientific and practical need to support intelligent diagnostic systems for diabetic retinopathy [10], [13].

The main motivation for this work is the major medical and public-health importance of diabetic retinopathy, which is considered one of the leading causes of vision loss in the world, especially with the increasing prevalence of diabetes. Early diagnosis can greatly reduce the progression of the disease. However, the manual inspection of fundus images is still a time-

consuming process requiring considerable expertise and is not scalable in resource-limited settings.

In this context, artificial intelligence and particularly deep learning provide a promising opportunity to develop automated diagnostic systems that can support physicians and enhance the quality of healthcare. However, many of the current studies primarily focus on achieving high accuracy and ignore other important aspects such as computational efficiency, interpretability, and practical applicability, which hinders the transfer of these models from the research environment to clinical use.

In addition, because few studies combine deep models with attention mechanisms or the integration of multimodal data, there is also a clear gap in the literature regarding the lack of a fair, systematic comparison of different architectures using the same database and the same experimental conditions. This makes it difficult to find the best model that can be relied upon in real-world applications.

Moreover, most of the current models are "black boxes", which raises concerns about the reliability of their decisions in the medical domain and the need to incorporate explainable artificial intelligence techniques to ensure transparency and increase physicians' trust in these systems.

Therefore, this work focuses on these challenges and proposes an integrated framework that combines high performance, efficiency, and interpretability. Moreover, a comprehensive comparison of several advanced architectures is conducted using a unified dataset, which may help to bridge the gap between theoretical research and clinical application in the field of diabetic retinopathy diagnosis.

II. Related Work

Although deep learning models have made great strides in classifying diabetic retinopathy, the literature shows a set of fundamental research gaps that continue to limit the effective translation of these models into the clinic. First, the majority of the studies use small or imbalanced datasets with respect to the distribution of disease categories, which causes model bias toward the most represented categories, adversely affecting generalisability [5], [10].

Second, many studies only focus on improving accuracy and ignore computational efficiency or inference time, making some models unsuitable for real-world clinical settings or resource-constrained devices [10], [11]. Furthermore, the comparison of models is often unfair because of different datasets or evaluation protocols.

Third, although hybrid models such as CNN-LSTM or CNN-Attention have emerged, the number of studies that systematically compare these models with traditional architectures is limited, and the effect of each component (attention, temporal modelling) on the final performance is still not well understood [12], [13].

Fourth, most of the studies lack explainability, where the models are treated as "black boxes", which is a major barrier to the adoption of these models by physicians. Tools such as Grad-CAM are used but often in a descriptive way without rigorous quantitative evaluation [11].

Fifth, there is a clear lack of multimodal approaches that integrate retinal images with clinical data (e.g., age, duration of diabetes, and blood pressure), even though the disease is inherently multifactorial and some studies have shown that this integration significantly improves the performance [6].

Finally, most studies do include external validations on different databases across multiple medical centers, which limits the reliability of the results, and the generalisability of the results to real-world clinical cases remains uncertain [5], [10], [16].

Table I presents a detailed comparison of eighteen recent studies on diabetic retinopathy detection and classification using deep learning and convolutional neural network methods. The studies were reviewed using a set of key criteria: the databases used, the number of images and classification categories, the proposed models and preprocessing techniques, the adopted evaluation metrics, and the best results obtained. This comparison demonstrates that current models, such as Xception, EfficientNet, and DenseNet, are effective in accurate and efficient classification of disease severity, especially when combined with image enhancement and data augmentation methods. The studies also highlight ongoing issues related to data imbalance, the high computational cost of certain deep models, and the limited generalisation ability of models when tested against other datasets. Such as comparison also serves to determine the current research trends, point out the gaps in the existing scientific research, and provide a more accurate, efficient, and applicable model for the intelligent medical diagnostic system.

The current study fills the gap of rigorous and standardised comparison of advanced convolutional neural network architectures for diabetic retinopathy classification. This work differs from many previous studies that evaluate different models under heterogeneous experimental conditions, in that the four representative CNN architectures, Xception, Inception-v3, MobileNet-v3, and ResNet-50, are compared under the same dataset, image resolution, data partitioning strategy, training protocol, and evaluation criteria. This unified experimental framework provides a fair and objective evaluation of each architecture while minimising potential biases due to inconsistent experimental settings.

The contributions of this work are summarised as follows:

1) First, this work proposes a new benchmarking framework to evaluate four state-of-the-art convolutional neural network architectures on the five-class diabetic retinopathy classification task in the same experimental setup.

2) Second, The models are assessed using a broad spectrum of performance indicators such as classification accuracy, precision, recall, specificity, F1-score, AUC, Cohen's Kappa coefficient, Matthews Correlation Coefficient (MCC), confusion matrices, convergence behaviour and loss analysis, providing a multidimensional view of diagnostic performance.

3) Third, the computational efficiency was analysed in terms of training time, runtime per epoch, and computational complexity (GFLOPs), which can objectively evaluate the

trade-off between predictive performance and computational cost for practical deployment.

The scope of the present investigation is intentionally limited to a controlled comparative evaluation of predictive performance and computational efficiency of Xception, Inception-v3, MobileNet-v3, and ResNet-50. The experimental analysis is focused on classification accuracy, convergence characteristics, class-wise confusion matrices, training time, and computational complexity under a unified experimental protocol. The explainability methods such as Grad-CAM or other visualisation techniques in the experimental framework presented in this manuscript were not applied. Thus, the explainable artificial intelligence was not viewed as a contribution of the present study, but see it as an important direction for future research to enhance the transparency and clinical interpretability of automated diabetic retinopathy diagnosis systems.

TABLE I. A METHODOLOGICAL COMPARISON OF RECENT STUDIES ON THE DETECTION OF DIABETIC RETINOPATHY USING DEEP LEARNING (2021–2026)

Ref	Year	Dataset	Images	Classes	Model	Preprocessing	Metrics	Best Result	Key Finding	Limitation
[2]	2025	Eye disease dataset	NR	4	ViT-CNN + self-attention	Resize + normalization	Accuracy, AUC	94%, AUC 98.82%	Effective for DR, cataract, glaucoma	Not DR-only
[3]	2025	EyePACS	NR	DR grading	Hybrid DL framework	Fundus preprocessing	Accuracy	96.9%	Strong DR detection	Complex architecture
[4]	2025	Messidor-1	1200	DR grading	Deep learning grading model	Fundus preprocessing	Accuracy	NR	Multi-grade DR classification	Needs external validation
[5]	2025	Medical DR benchmark	NR	DR	C-JDA	Domain adaptation + augmentation	Accuracy, F1	96.93%	Improves domain transfer	Pseudo-label dependency
[6]	2025	Multiple medical datasets	NR	Multi-disease	CNN-LSTM-Attention	Standard preprocessing	Accuracy	>95%, peak 98%	Robust hybrid architecture	Not exclusively DR
[7]	2025	APTOS 2019 + Pima	3662 + tabular	5	CNN + GBM multimodal	Resize, normalization, denoising	Accuracy, F1, AUC	96%, F1=0.96, AUC=0.994	Image + clinical fusion	Needs multimodal inputs
[8]	2024	EyePACS/APTOS	NR	DR severity	Hybrid deep model	Augmentation	Accuracy	99.63% EyePACS	High severity classification	Needs validation
[9]	2024	DR fundus	NR	Multi-grade	Wavelet CNN + SVM	Spectral extraction	Accuracy, AUC	98.95%, AUC 0.99	Strong spectral grading	Complex pipeline
[10]	2024	Multiple DR datasets	NR	DR grading	Ensemble TL	Augmentation	Kappa, Accuracy	NR	Efficient ensemble grading	Dataset-dependent
[11]	2024	APTOS/Messidor-2	NR	DR	DenseNet/InceptionResNet-V2	Enhancement	Accuracy	Reported	Handles imbalance	Dataset variation
[12]	2024	Retinal images	NR	DR	EfficientNetB0 ensemble	Fundus preprocessing	Accuracy	86.53%	Efficient DR diagnosis	Moderate accuracy
[13]	2023	APTOS/EyePACS/Messidor-2	NR	Binary/Multi	TL + ECOC	Optimization	Accuracy	96% binary	Good generalization	Lower multi-class
[14]	2023	APTOS 2019	3662	5	DenseNet121	Resize + augmentation	Accuracy	97.30%	Strong CNN	Single dataset
[15]	2022	APTOS 2019	3662	Binary/Multi	Hybrid CNN + SVM	Enhancement	Accuracy	97.8% binary	Hybrid features	Multi-class harder
[16]	2022	APTOS 2019	3662	Binary/Multi	DenseNet169 + CBAM	Fundus preprocessing	Accuracy	97% binary	Attention improves lesions	Binary focused
[17]	2021	DR fundus	NR	DR	Deep learning model	Fundus preprocessing	Accuracy	NR	Early DL classification	Limited reproducibility
[18]	2021	DR fundus	NR	DR	CNN-based detection	Fundus preprocessing	Accuracy	NR	Automated screening	Limited external validation

III. METHODOLOGY

The study employs a comparative experimental approach to assess the performance of four deep convolutional neural network architectures, namely Xception, Inception-v3, MobileNet-v3, and ResNet-50, in classifying diabetic retinopathy fundus images. These models were chosen as they represent different architectural approaches: Xception uses separable convolutions, Inception-v3 uses a multi-branch architecture to extract features at different levels, MobileNet-v3 is a lightweight model that optimises computational efficiency, while ResNet-50 has a deeper architecture with more branches and computational aggregations. This diversity of architecture allows for a fair comparison not only in terms of accuracy, but also loss, training time, GFLOPs and classification metrics. Fig. 1 presents the general architecture of the methodology adopted in this study, beginning with the input of the fundus images, through preprocessing, then training the four models, and finally extracting the final metrics.

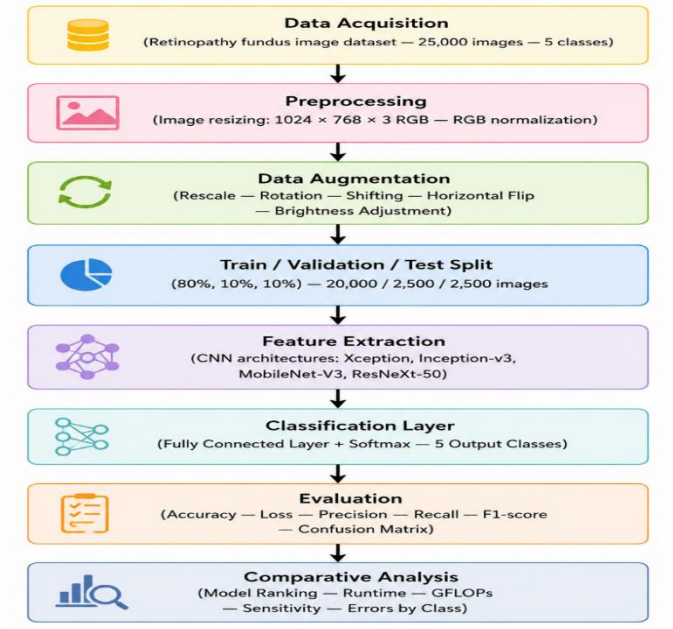


Fig. 1. Workflow of the proposed benchmarking framework.

The dataset used in this study includes 25,000 fundus images, equally distributed into five classes with 5,000 images in each class. This balance between the classes makes overall metrics such as accuracy, precision, recall and F1-score more reliable since all of the classes contribute equally to the training and testing process. The fundus images were normalised to RGB size of 1024 × 768 × 3, because they contain fine visual details such as blood vessels, optic disc, fovea, haemorrhages and deposits which are important to differentiate between severity levels of diabetic retinopathy. The stages of the diabetic retinopathy are usually classified medically into the following categories: no retinopathy, mild, moderate, severe and proliferative retinopathy as shown in Table II. The classifications used are closely related to the medical terminology used to classify the severity of the disease. Fig 2 illustrates clear visual differences between healthy and advanced pathological cases.

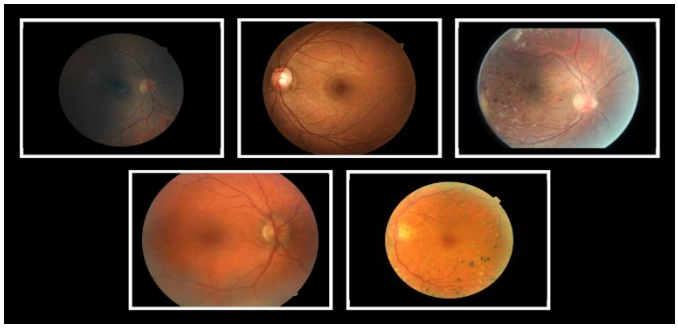


Fig. 2. Fundus images of diabetic retinopathy resized arranged dataset, Kaggle [1].

TABLE II. LABELS OF THE FIVE DIABETIC RETINOPATHY CLASSES

Number of Images	Name Class / Category	Label
5,000	No Diabetic Retinopathy	0
5,000	Mild Diabetic Retinopathy	1
5,000	Moderate Diabetic Retinopathy	2
5,000	Severe Diabetic Retinopathy	3
5,000	Proliferative Diabetic Retinopathy	4
25,000	**Diabetic Retinopathy Fundus Images**	**Total**

The data was divided into three independent sets: 80% for training, 10% for validation and 10% for testing. Fig. 3 shows the results of partitioning the data.

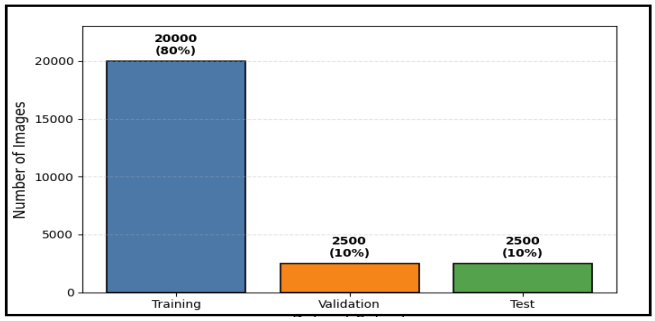


Fig. 3. Dataset split.

Thus, the training set consists of 20,000 images or 4,000 images per class, the validation set contains 2,500 images or 500 images per class, and the test set also contains 2,500 images or 500 images per class. This separation is needed to ensure that the model is not evaluated on the same images that it was trained on. It also allows us to monitor performance during training with the validation set and then finally evaluate performance on an independent test set. Table III presents the distribution of the dataset in the training, validation and test sets, with the numerical balance between the classes.

The images were first subjected to standard preprocessing steps such as resizing, colour normalisation and numerical normalisation to ensure that they satisfy the requirements of each model prior to training. This procedure is required for fundus images, since these images can vary in illumination, acquisition angle, vessel resolution, and camera quality. Moreover, pre-trained deep models are appropriate to use pre-trained deep models for such tasks, as the Keras library provides

ready-made models with pre-trained weights that can be used for predictions, feature extraction, or fine-tuning. Fig. 4 shows the proposed preprocessing steps before feeding the images to the models.

TABLE III. DISTRIBUTION OF THE DATASET IN THE TRAINING, VALIDATION, AND TEST SETS

Value	Description
5 classes	Number of Classes
5,000 images	Number of images per category
25,000 images	Total number of images
1024 × 768 × 3 RGB	Image size
20,000 images	Training images
2,500 images	Validation images
2,500 images	Test images
4,000 images	Number of training images in each category
500 images	Number of validation images per category
500 images	Number of test images per category

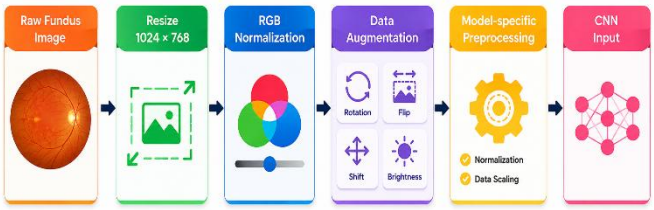


Fig. 4. Proposed preprocessing steps before feeding images into the models.

The batch size was chosen according to the type of each model and the computational cost. A Batch size of 32 is used with Xception due to the depth of the model and large image size. Batch size of 128 is used with Inception-v3 as it gives a good balance between computational stability and training speed [3]. For MobileNet-v3, a batch size of 256 is used because it is a lightweight model and can support larger batches, whereas for ResNet-50, a batch size of 64 is applied due to its high computational cost. The number of steps per epoch was different: 625 for Xception, 157 for Inception-v3, 79 for MobileNet-v3 and 313 for ResNet-50. Fig. 5 shows the trade-off between batch size and number of steps per epoch for the four models.

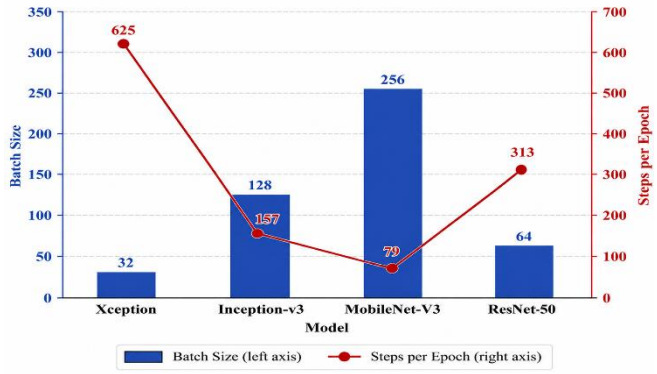


Fig. 5. Batch size and steps per epoch across the four models.

All models were trained for 50 epochs for a fair comparison between models. From the curves, the plateau is observed around epoch 43, where accuracy and loss start to stabilise gradually, showing that the models have converged without any obvious divergence between the training and validation curves. To avoid overfitting, EarlyStopping will monitor the val_loss or val_accuracy and stop the training when it stops improving. Also, `ReduceLROnPlateau` can be used to reduce the learning rate when no improvement happens and `ModelCheckpoint` to save the best version of the model during training. EarlyStopping stops training when the metric being monitored stops improving. ReduceLROnPlateau reduces the learning rate when performance plateaus. ModelCheckpoint allows the model to be saved during training, so that the best version can be retrieved later. As for the optimiser, it can be Adam or AdamW. The Adam optimiser is based on the adaptive estimation of first and second derivatives of the gradient, which is suitable for tasks with large datasets and deep models [19].

All the CNN models were developed in TensorFlow/Keras and were trained on the HPC-MARWAN high-performance computing platform on eight CPU cores. In the case of transfer learning, each backbone was initialized with weights pre-trained on ImageNet. The original classification layer was replaced by a global average pooling layer, a dropout layer and a five-class softmax classifier adapted to the diabetic retinopathy classification task.

The model was trained with the Adam optimizer with an initial learning rate of 0.0001. To prevent overfitting, early stopping was used with a patience of 5 epochs, and the learning rate was automatically decreased using the ReduceLROnPlateau callback with a reduction factor of 0.5 when the validation loss stopped improving. All experiments were performed with a fixed random seed of 42 to ensure reproducibility in dataset partitioning and model initialization.

All experiments were performed on the HPC-MARWAN cluster under the same computational conditions in an effort to make a fair comparison between the CNN architectures evaluated, as shown in Table IV.

TABLE IV. EXPERIMENTAL CONFIGURATION AND REPRODUCIBILITY SETTINGS

Parameter	Value
Computing platform	HPC-MARWAN
Hardware	CPU-only
CPU cores	8
Framework	TensorFlow / Keras
Transfer learning	ImageNet pretrained weights
Optimizer	Adam
Initial learning rate	0.0001
Loss function	Categorical Cross-Entropy
Random seed	42
Early stopping	Patience = 5
Learning-rate scheduler	ReduceLROnPlateau
Reduction factor	0.5
Dataset split	80% / 10% / 10%
Input image size	1024 × 768 × 3

IV. EXPERIMENTS

All experiments were performed using a standard protocol to allow for a fair comparison between the four models. The study used the same dataset, image size, number of epochs, and data split, with only the batch size differing depending on the nature of each model. This difference does not affect the fairness of the comparison since it is aimed at adapting each model to its own computational and memory requirements. Images of size 1024 × 768 RGB have a heavy computational load, so the same batch size cannot be used for all models without affecting stability or memory. The general experimental setup and batch configuration for the four models are shown in Table V and Table VI.

TABLE V. GENERAL EXPERIMENTAL SETUP OF THE FOUR MODELS

Experimental Setting	Value
Dataset	Diabetic retinopathy fundus image dataset
Total images	25,000 images
Number of classes	5 classes
Image size	1024 × 768 × 3 RGB
Data split	80% training / 10% validation / 10% test
Training / Validation / Test images	20,000 / 2,500 / 2,500
Number of epochs	50
Models evaluated	Xception, Inception-v3, MobileNet-v3, ResNet-50
Variable settings	Batch size adapted to each model.

TABLE VI. MODEL-SPECIFIC BATCH CONFIGURATION

Model	Batch Size	Steps per Epoch
Xception	32	625
Inception-v3	128	157
MobileNet-v3	256	79
ResNet-50	64	313

For the Xception experiment, the batch size was 32, with 625 steps per epoch. The accuracy curves of this model show that the accuracy increased rapidly from the early steps of training. The training accuracy started from about 84% and gradually reached 98.38%, and the validation accuracy started from around 89.8% and reached 98.52%. The observed behaviour emphasises the ability of Xception to extract the correct features from the fundus images, especially because the training and validation curves were close to each other in the last steps. Fig. 6 shows the accuracy curve for model Xception.

For the Inception-v3 experiment, a batch size of 128 was used with 157 steps per epoch. Initially, the training accuracy was 70%, then it increased to 95.05%, and the validation accuracy was 95.30%. This behaviour indicates that the model is able to learn in a gradual and stable way, but it is less superior than Xception. Inception-v3 is a multi-branch architecture that allows for feature extraction at different scales, which is beneficial for fundus images that contain both fine and coarse details. The accuracy curve for the Inception-v3 model is shown in Fig. 7.

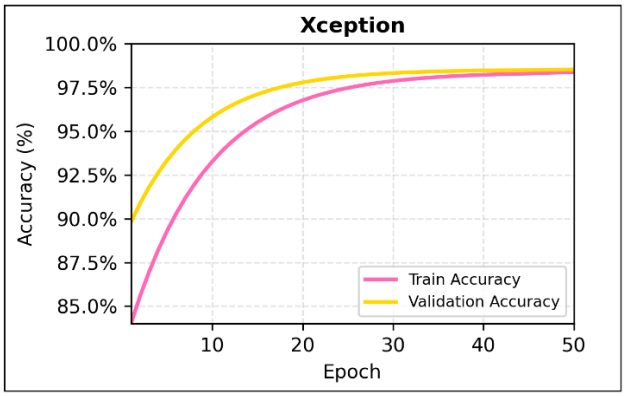


Fig. 6. Accuracy curve for the Xception model.

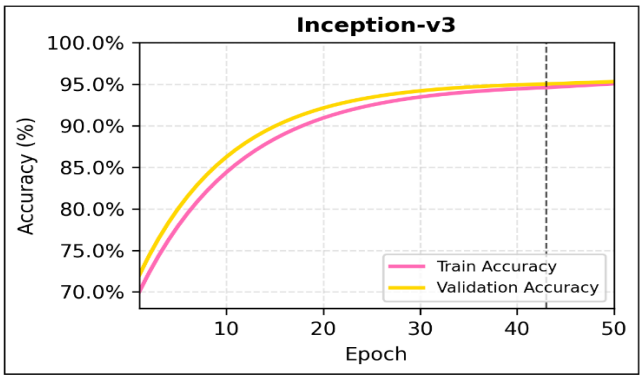


Fig. 7. Accuracy curve for the Inception-v3 model.

The MobileNet-v3 experiment used a batch size of 256 and 79 steps per epoch. The training-set accuracy is 90.10%, and the validation set is 90.35%. These are lower than Xception and Inception-v3, but MobileNet-v3 is still important from an efficiency perspective as it was originally designed as a lightweight model for resource- constrained applications. Although less accurate, it remains useful in practical situations where large amounts of computing power are not available.

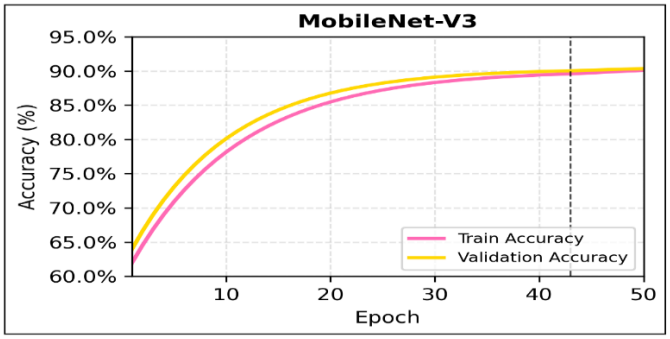


Fig. 8. Accuracy curve for the MobileNet-v3 model.

Apart from accuracy, loss curves were also studied for each model. The model with the lowest final loss was Xception with a training loss of 8.05% and a validation loss of 7.80%. The validation loss of Inception-v3 was 11.90 %. MobileNet-v3 was

around 29.80 % and ResNet-50 was around 24.80 %. A lower loss means the model is able to make better classification decisions. The closeness of the training and validation loss indicates a good level of generalisation. Fig. 9 shows the training loss curves of the four models.

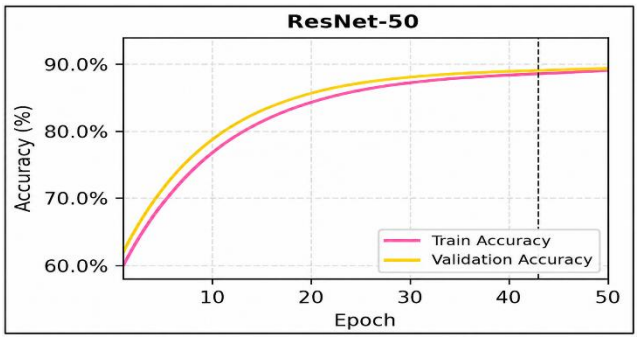


Fig. 9. Accuracy curve for the ResNet-50.

Besides accuracy, loss curves for each model were also studied. Xception had the lowest final loss with a training loss of 8.05% and validation loss of 7.80%. Inception-v3, however, had a validation loss of 11.90%, whereas the validation loss of MobileNet-v3 and ResNet-50 was around 29.80% and 24.80% respectively. A lower loss means better classification decisions by the model, and convergence of the training and validation losses means good generalisation. Fig. 10 shows the training loss curves of the four models and Fig. 11 shows the validation loss curves.

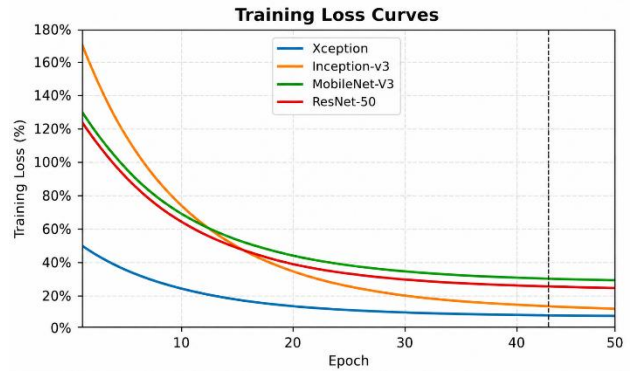


Fig. 10. Training loss curves for four models.

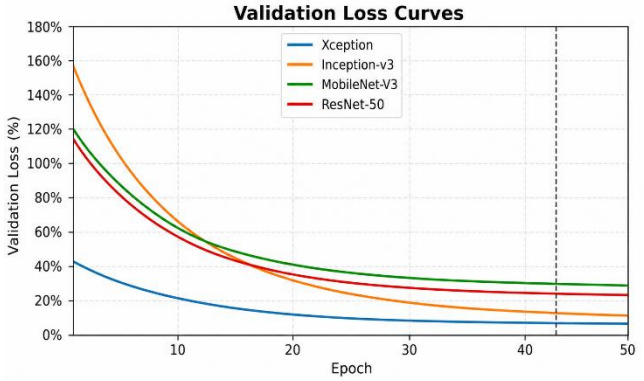


Fig. 11. Validation loss curves for four models.

V. Results and Discussion

The results showed that the Xception model performed best overall among the four models. It achieved 98.38% training accuracy, 98.52% validation accuracy, and 98.32% test accuracy. The training loss was 8.05%, and the validation loss was 7.80%, which is the lowest among all models. The results indicate that Xception had a good balance between learning and generalisation without a clear gap between the training curve and the validation curve. Fig. 12 is a general comparison of training and validation accuracies of the four models, where Xception is ranked first.

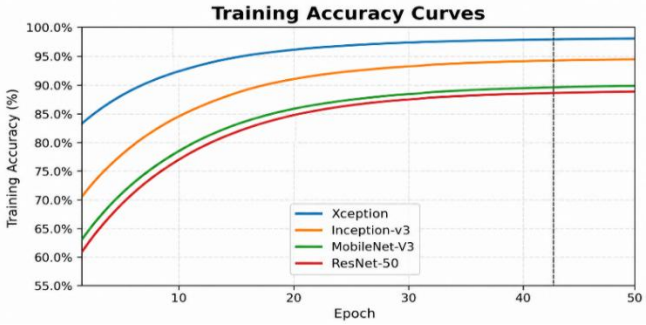


Fig. 12. Training accuracy curves for four models.

Looking at the validation accuracy, Xception performed the best, followed by Inception-v3, MobileNet-v3 and ResNet-50. The validation accuracies were 98.52%, 95.30%, 90.35% and 89.35% respectively. This means that Xception not only performs well during training, but also generalises well to data not seen during training. The validation accuracy curves for the four models are shown in Fig. 13, and they are among the most important figures to include in the results section as they summarise the models' ability to generalise.

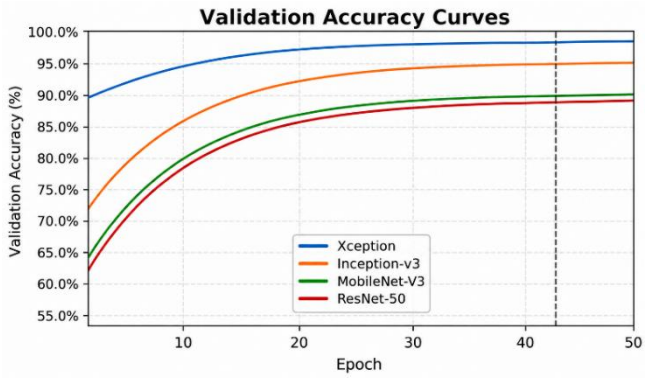


Fig. 13. Validation accuracy curves for four models.

The accuracy and loss comparison revealed that Xception was the most efficient model in this study as it combined the highest accuracy with the lowest loss. Inception-v3 is a powerful model but with a higher computational cost per image than Xception. MobileNet-v3 has moderate accuracy and low GFLOPs, and ResNet-50 is not clearly superior despite the high cost. Fig. 14 provides a comprehensive comparison of training accuracy, validation accuracy, training loss and validation loss of all four models.

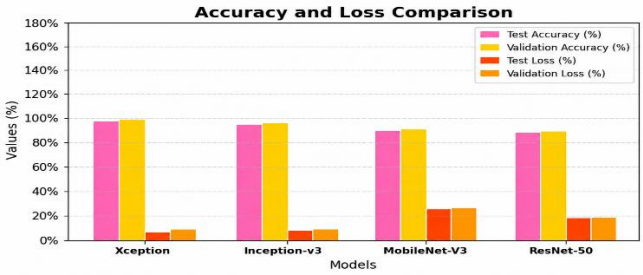
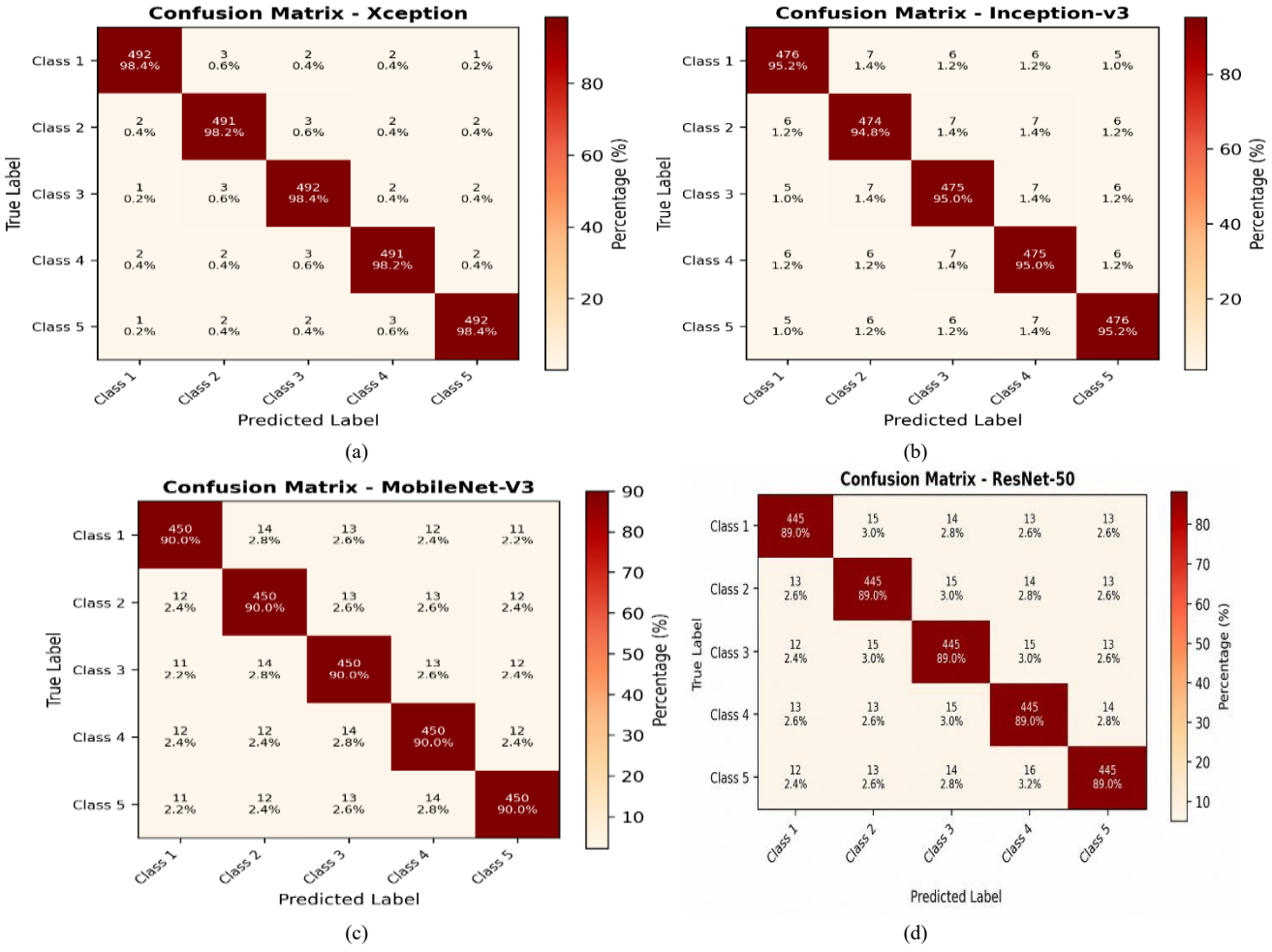


Fig. 14. Bar chart comparing accuracy and loss.

Confusion matrices are an important part of this study as they show the distribution of errors over the five classes. In the confusion matrix of Xception, most values were concentrated on the main diagonal, with around 491 to 492 images correctly classified from 500 images of each class. Inception-v3 achieved about 474 and 476 correctly classified images per class. MobileNet-v3 got about 450 correctly classified images per class, and ResNet-50 got about 445 correctly classified images per class. This distribution shows that errors were not concentrated in one class and were relatively evenly distributed, which is expected considering the balance of the dataset. The confusion matrices of Xception (a), Inception-v3 (b), MobileNet-v3 (c), and ResNet-50 (d) are shown in Fig. 15.

Although the overall classification accuracy is high, the confusion matrices reveal that there are still errors between neighboring diabetic retinopathy severity levels. Visual overlap between adjacent disease stages may explain such misclassification, where differences in lesion number, size and distribution may be subtle. Variations in illumination, contrast, sharpness, and retinal visibility of the fundus image may further increase the difficulty of classification, especially for those images containing small or weakly visible pathological signs. Moreover, the present study provides image-level severity classification but not lesion-level localization or lesion-specific prediction analysis. Hence, the results show the ability of the evaluated CNN architectures to classify the diabetic retinopathy grades, but do not give a complete explanation of the retinal lesions that contributed to the individual predictions.

Table VII provides a standardized summary of a direct comparison between the overall metrics of the four models, including accuracy, macro precision, macro recall, macro specificity, macro F1-score, macro AUC, Cohen's Kappa and MCC. As can be seen from the table, Xception had the best overall performance, followed by Inception-v3, while MobileNet-v3 and ResNet-50 had relatively lower values.



(a) (b) (c) (d)

Fig. 15. Confusion matrices for four models.

TABLE VII. OVERALL PERFORMANCE METRICS OF THE FOUR CNN MODELS

Model	Accuracy (%)	Precision Macro (%)	Recall Macro (%)	Specificity (macro) (%)	F1-score (macro) (%)	AUC (macro) (%)	Cohen's Kappa	MCC
Xception	98.32	98.32	98.32	99.58	98.32	99.99	0.9790	0.9790
Inception-v3	95.04	95.04	95.04	98.76	95.04	99.94	0.9380	0.9380
MobileNet-v3	90.00	90.00	90.00	97.50	90.00	99.72	0.8750	0.8750
ResNet-50	89.00	89.00	89.00	97.25	89.00	99.64	0.8625	0.8625

The resemblance between overall accuracy and macro-averaged precision, macro-averaged recall and macro F1-score is not surprising for the present benchmark due to the dataset being balanced by design, i.e. with equal number of samples from each diabetic retinopathy category. Under these conditions, all classes contribute equally to the macro averages, and the macro performance values closely follow the overall classification accuracy, due to the consistently high class-wise precision and recall obtained by the evaluated models. The similarity of these evaluation metrics is therefore a reflection of the properties of the dataset and the classification results and not a discrepancy in the evaluation pipeline.

The performance numbers reported are taken from the present experimental runs, and should be considered as comparative benchmark results, rather than hard proof of statistical superiority. All architectures were tested on the same dataset partition, the same input resolution, the same number of epochs, and the same evaluation protocol. A more rigorous evaluation of the stability and significance of the observed differences would be obtained by repeated training with independent random seeds, confidence-interval estimation, bootstrap resampling and formal hypothesis testing. The present analyses are therefore considered an important extension of the present study.

The results showed that ResNet-50 was the most computationally expensive model in terms of computational cost with 66.61 GFLOPs per image, epoch time of approximately 5,221.77 seconds and total training time of 72.52 hours. However, MobileNet-v3 had the lowest GFLOPs per image with only 7.05 GFLOPs but did not achieve the best runtime in this experiment. Xception, on the other hand, obtained a good trade-off between performance and cost, costing 16.27 GFLOPs per image and taking 30.53 hours to train. Fig. 16 shows runtime per epoch, Fig. 17 shows GFLOPs per image, and Fig. 18 compares runtime and GFLOPs.

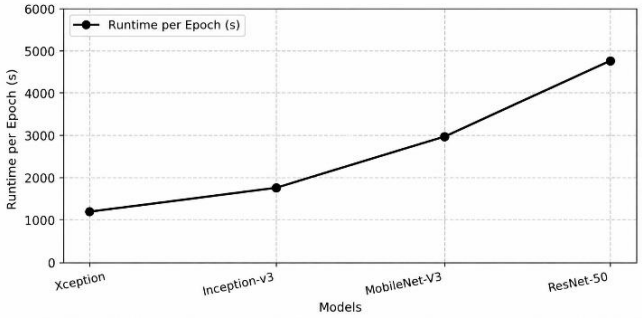


Fig. 16. Runtime per epoch for the four models.

In the present experiment, Xception showed the best trade-off between classification performance and computational demand overall. It achieved the highest accuracy and the lowest loss with a significantly lower computational cost than ResNet-50. Inception-v3 showed a competitive performance, but was less efficient than Xception, while MobileNet-v3 was still the most appropriate lightweight choice with a lower number of operations. It is deeper in architecture and computationally more expensive. However, ResNet-50 did not provide a similar improvement in forecasting performance.

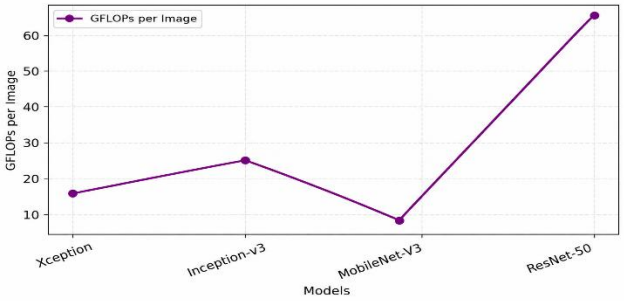


Fig. 17. GFLOPs per image for the four models.

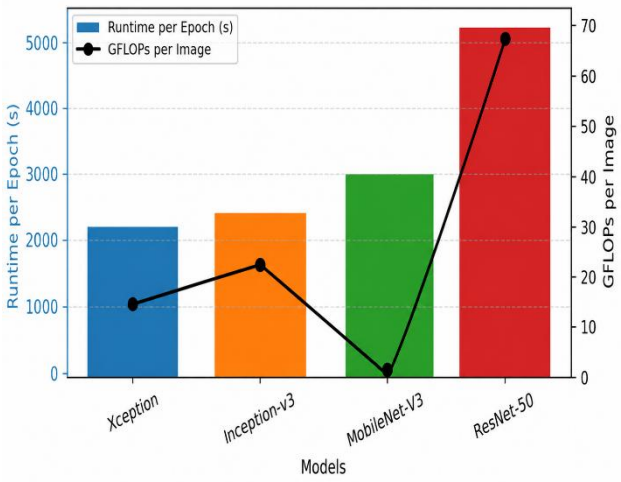


Fig. 18. Computational cost: runtime and GFLOPs.

VI. CONCLUSION AND FUTURE PROSPECTS

In this study, a comparative experimental framework was presented to evaluate the performance of four deep convolutional neural network architectures on the task of classifying fundus images associated with diabetic retinopathy: Xception, Inception-v3, MobileNet-v3 and ResNet-50. The study used a balanced dataset comprising 25,000 colour fundus images, equally divided among five classes, with 5,000 images per class. The dataset is divided to 80% for training, 10% for validation and 10% for testing, which corresponds to 20,000 images for training, 2,500 images for validation and 2,500

images for testing. This division is necessary for a more objective assessment of the performance of the models, since the models are trained on a large set of images, their generalization capacity is monitored during the training process, and finally, they are tested on independent data that were not part of the learning process. Class balance also helps the models avoid being biased towards a particular class and makes metrics such as accuracy, recall, precision, and F1-score more interpretable and comparable.

The results showed that the Xception model performed the best among the four models in general. The proposed model achieved 98.38% training accuracy, 98.52% validation accuracy, and 98.32% test accuracy. It also boasted the lowest loss values with a training loss of 8.05% and a validation loss of 7.80%. The results obtained show that Xception obtained a high accuracy on the training data, as well as a good generalization ability to the validation and test data. This is important in the medical field as a model that only learns the characteristics of the training data without generalizing to new data is not suitable for use in real-world environments. Moreover, the training, validation, and test accuracies converge, which indicates that the model does not suffer from obvious overfitting and is able to learn stable and effective visual representations for the fundus images.

The reason for Xception's better performance in this study is that it is designed with depthwise separable convolutions. This It also enables the extraction fine visual features and reduces the computational load with the heavier models. This property seems useful in the context of fundus images, as it allows the model to capture small local patterns such as hemorrhagic spots or subtle changes in the retina, while still maintaining a good ability to represent the overall structure of the image. The confusion matrix for this model shows that most of the correct classifications are near the main diagonal, with a small spread of errors across the classes, indicating a good ability to discriminate between the five classes.

The Inception-v3 model also performed well with a validation accuracy of 95.30% and a test accuracy of 95.04%. The performance indicates that the multi-branch structure of the model can extract visual features from different levels, which is relatively well-suited for fundus images with both fine and coarse details simultaneously. However, Inception-v3 still had worse performance than Xception in terms of accuracy, loss, and computational cost per image. Inception-v3 needs more GFLOPs per image than Xception and does not improve the accuracy. Thus, it can be regarded as a robust and stable model but not the most efficient one among the models tested in this study.

For MobileNet-v3, the test accuracy was 90.00%, which is lower than those of Xception and Inception-v3, but still an important result given the computational efficiency. MobileNet-v3 was mainly designed to be a lightweight model that can be used in resource-limited environments such as mobile devices, embedded systems, or low-cost medical screening devices. The model has the lowest GFLOPs per image, and thus is suitable when the goal is to minimize the computational load rather than to achieve the highest possible accuracy. However, its lower accuracy compared to more powerful models emphasizes the need to be cautious when using it in sensitive medical applications, especially when the aim is to support a diagnostic decision that requires a high degree of confidence.

ResNet-50 provided the lowest accuracy of the four models, with a test accuracy of 89.00%, but the highest computational cost in this experiment. The result indicates that the increase in the depth or in the architectural complexity does not always lead to performance improvement, especially when the architecture is not well matched to the nature of the data, or when the increase in the computational costs is not compensated by a clear gain in accuracy. ResNet-50 showed a long training time and high GFLOPs without outperforming the other models. Thus, the present results show that this model does not seem to be the best choice for fundus image classification in this experimental setup, especially if the study aims to balance performance and efficiency.

The overall results confirm the compromise between accuracy and computational cost as a key factor in the choice of a suitable model. In this study, when the priority is higher diagnostic performance, Xception is the best choice because it has high accuracy, low loss, and clear generalization ability. However, if the objective is to create a lightweight system, MobileNet-v3 is a better option to consider, despite having a lower accuracy than powerful models. Inception-v3 is a better middle ground in performance but less efficient than Xception in these results. On the other hand, ResNet-50 does not provide any clear practical advantage in this setting, given the high computational cost relative to its accuracy.

The value of this study is that it does not just rank models by their accuracy, but provides a multidimensional view of their performance by combining accuracy, loss, classification metrics, confusion matrices, training time, and computational cost. This type of evaluation is critical in medical AI applications, as a model ready for clinical deployment must be accurate, stable, interpretable, generalizable, and able to run within realistic time and compute constraints. Based on the reported results, Xception can be regarded as the most balanced architecture within the present experimental setting in this study and MobileNet-v3 as a promising option for lightweight applications; Inception-v3 and ResNet-50 need further optimization or fine-tuning before being used in a practical diagnostic assistance system. However, the findings of this study need to be interpreted in view of the limitations of its experimental setting. While the dataset is sufficiently large and balanced to compare the models, in order to generalize the results to real-world clinical settings, the models need to be tested on independent external datasets obtained across different imaging devices and multiple medical institutions. Moreover, high performance in a controlled experiment does not ensure that the model will perform at the same level with low-quality images, images under different lighting conditions, or images from patients with different demographic backgrounds. The present results thus represent an important step toward the establishment of a preliminary benchmark but do not exclude the need for further validation studies before any direct clinical application.

This study has some limitations; the experimental evaluation was performed on one publicly available diabetic retinopathy

dataset. This allows a fair comparison of the evaluated CNN architectures in the same experimental conditions, but further validation with independent datasets from different clinical settings and imaging devices is needed to further evaluate the robustness and the generalizability of the proposed framework.

Although this study provided a clear empirical comparison of four deep learning architectures to classify fundus images related to diabetic retinopathy, future work could focus first on expanding the database used, both in terms of the number of images and the variety of their sources. Models were evaluated on images obtained from different imaging devices, different medical institutions and diverse clinical cases will give an insight for the generalizability of the models in conditions closer to clinical settings of real-world. Moreover, the utilization of independent external databases will enhance the robustness of the obtained results, as good performance in a single internal split does not, on its own, prove the validity of the model to be used in different clinical settings.

The study can be further extended by using newer and more diverse models such as EfficientNet, ConvNeXt, Vision Transformers, and hybrid models combining convolutional networks and attention mechanisms. This enables us to investigate whether modern architectures can improve the trade-off between accuracy and computational efficiency. In addition, it is possible to improve the image preprocessing step by using fundus image-specific techniques such as contrast enhancement, light adjustment, removal of irrelevant areas, and automatic image quality detection before the images are introduced into the model.

In future work, each training experiment should be repeated with a number of different independent random seeds. Additionally, the mean, standard deviation and confidence interval for the main performance indicators should be reported. Also, before reaching some conclusions about the superiority of the model, it is required to perform paired bootstrap comparisons or other statistical tests considered relevant.

In the clinical context, the proposed framework aims to support ophthalmologists by providing quick and consistent initial assessments of diabetic retinopathy, not to replace expert clinical judgment. In real-world screening programs, the reduction of false negative predictions is particularly important to avoid missed diagnoses and delayed treatment. Therefore, additional clinical validation, workflow integration, and prospective evaluation are required before the proposed framework can be considered for routine clinical use.

A crucial aspect for future development is the integration of visual interpretation techniques such as Grad-CAM and Grad-CAM++ to identify the regions on which the model relies to make its decision. This is especially true in the medical domain where doctors need not only the final classification but also an explanation that can help them understand the reasoning behind the decision. Moreover, a more fine-grained error analysis can be performed, in particular between neighboring grades of diabetic retinopathy, which are often of higher clinical relevance. Finally, future work can be focused on building a medical decision support system with accuracy, efficiency, and interpretability. The system may include image classification, a confidence score and an interpretive map indicating the regions that influence the decision. In the future, the model could be extended to a multimodal one by combining fundus images with clinical data such as duration of diabetes, HbA1c level, blood pressure, and age of the patient. This development would bring the model more in line with the real world of medical practice, while emphasizing that its role must be that of an aid to the physician and not a substitute for specialized clinical evaluation.

REFERENCES

[1] A. Arora, Diabetic Retinopathy Resized Arranged Dataset, Kaggle, 2019. Available: https://www.kaggle.com/datasets/amanneo/diabetic-retinopathy-resized-arranged

[2] N. Ayesha, "A Vision Transformer-Based Convolutional Neural Network for the Automated Diagnosis of Eye Diseases Using Self-Attention Mechanisms," Engineering, Technology & Applied Science Research, vol. 15, no. 4, pp. 24493–24497, 2025, doi: 10.48084/etasr.10649.

[3] M. Sushith et al., "A hybrid deep learning framework for early detection of diabetic retinopathy using retinal fundus images," Scientific Reports, vol. 15, Art. no. 15166, 2025, doi: 10.1038/s41598-025-99309-w.

[4] Akhtar, S., Aftab, S., Ali, O. et al. A deep learning based model for diabetic retinopathy grading. Sci Rep 15, 3763 (2025). https://doi.org/10.1038/s41598-025-87171-9

[5] J. Zhang, R. Li, C. Liu, and X. Ji, "Improving Domain Transfer with Consistency-Regularized Joint Distribution Alignment for Medical Image Classification," Symmetry, vol. 17, no. 4, Art. no. 515, 2025, doi: 10.3390/sym17040515.

[6] M. T. Hayat et al., "A Hybrid Convolutional Neural Network–Long Short-Term Memory (CNN–LSTM)–Attention Model Architecture for Precise Medical Image Analysis and Disease Diagnosis," Diagnostics, vol. 15, no. 21, Art. no. 2673, 2025, doi: 10.3390/diagnostics15212673.

[7] A. S. Abdalrada, A. F. Neamah, H. L. Majeed, and A. R. Al-Sudani, "A Hybrid Machine Learning Approach for Enhanced Diabetes Prediction: Integrating Image and Numerical Data," Mesopotamian Journal of Big Data, pp. 211–221, 2025.

[8] A. Shazia et al., "Automated Early Diabetic Retinopathy Detection Using a Deep Hybrid Model," Trends in Emerging Technologies and Artificial Intelligence, 2024.

[9] S. Sundar, S. Subramanian, and M. Mahmud, "Classification of Diabetic Retinopathy Disease Levels by Extracting Spectral Features Using Wavelet CNN," Diagnostics, vol. 14, no. 11, Art. no. 1093, 2024, doi: 10.3390/diagnostics14111093.

[10] S. V. Chilukoti et al., "A reliable diabetic retinopathy grading via transfer learning and ensemble learning with quadratic weighted kappa metric," BMC Medical Informatics and Decision Making, vol. 24, no. 1, Art. no. 37, 2024, doi: 10.1186/s12911-024-02446-x.

[11] K. Sriporn, C.-F. Tsai, L.-J. Rong, P. Wang, T.-Y. Tsai, and C.-W. Chen, "Optimizing Deep Learning for Diabetic Retinopathy Diagnosis," International Journal of Advanced Computer Science and Applications, vol. 15, no. 11, 2024, doi: 10.14569/IJACSA.2024.0151135.

[12] L. Arora et al., "Ensemble deep learning and EfficientNet for accurate diabetic retinopathy diagnosis," Scientific Reports, vol. 14, 2024.

[13] W. K. Wong, F. Juwono, and C. Capriono, "Diabetic Retinopathy Detection and Grading: A Transfer Learning Approach Using Simultaneous Parameter Optimization and Feature-Weighted ECOC Ensemble," IEEE Access, 2023, doi: 10.1109/ACCESS.2023.3301618.

[14] C. Mohanty et al., "Using Deep Learning Architectures for Detection and Classification of Diabetic Retinopathy," Sensors, vol. 23, no. 12, Art. no. 5726, 2023, doi: 10.3390/s23125726.

[15] M. M. Butt, D. A. Iskandar, S. E. Abdelhamid, G. Latif, and R. Alghazo, "Diabetic Retinopathy Detection from Fundus Images of the Eye Using Hybrid Deep Learning Features," Diagnostics, vol. 12, no. 7, Art. no. 1607, 2022, doi: 10.3390/diagnostics12071607.

[16] Farag, Mohamed & Fouad, Mariam & Abdel-Hamid, Amr. (2022). Automatic Severity Classification Of Diabetic Retinopathy Based On DenseNet And Convolutional Block Attention Module. IEEE Access. 10. 1-1. 10.1109/ACCESS.2022.3165193.

[17] Mushtaq, Gazala & Siddiqui, Farheen. (2021). Detection of diabetic retinopathy using deep learning methodology. IOP Conference Series: Materials Science and Engineering. 1070. 012049. 10.1088/1757-899X/1070/1/012049.

[18] A. A. Ayala et al., "Diabetic Retinopathy Classification with Deep Learning via Fundus Images: A Short Survey," Procedia Computer Science, 2021.

[19] K. Ladrham and H. Gueddah, "Optimizing Dermatological Image Classification Using an Efficient Convolutional Neural Network Architecture," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 16, no. 12, pp. 577–587, 2025.

The Evolution of Clinical Intelligence Through GenAI Co-pilots: A Systematic Review and Thematic Synthesis

Li-Chen Cheng[1], Chia-Yu Hung[2], Te-Nien Chien[3]*
Department of Information and Finance Management, National Taipei University of Technology, Taipei, Taiwan[1]
College of Management, National Taipei University of Technology, Taipei, Taiwan[2, 3]

***Abstract*—Generative artificial intelligence (GenAI), large language models (LLMs), and multimodal foundation models are rapidly transforming healthcare by extending artificial intelligence beyond traditional predictive analytics toward collaborative clinical intelligence. Recent advances have enabled applications in clinical decision support, medical documentation, workflow optimization, patient communication, and personalized care planning. However, the existing literature remains fragmented across technical evaluations, specialty-specific applications, and governance discussions, resulting in a limited understanding of how these technologies collectively function within clinical environments. This study conducted a systematic review and thematic synthesis to examine the emerging role of GenAI as a clinical co-pilot in healthcare. Following the PRISMA 2020 framework, literature was retrieved from Web of Science, PubMed, and Scopus databases. A total of 3,124 records were identified, and 41 studies published between 2023 and March 2026 met the eligibility criteria and were included in the final qualitative synthesis. Quality assessment indicated that 87.8% of the included studies were classified as moderate or high quality, providing a robust methodological basis for the thematic synthesis. Thematic analysis revealed three interrelated domains underlying the evolution of GenAI-enabled clinical intelligence: 1) Perception and Fact Anchoring, involving multimodal data integration, retrieval-augmented generation (RAG), and domain-specific medical intelligence; 2) Clinical Agency and Collaboration, encompassing ambient digital scribing, agentic clinical decision support, workflow augmentation, and personalized care; and 3) Governance and Responsibility, including human-in-the-loop oversight, privacy protection, regulatory compliance, fairness, and trustworthy AI practices. Based on these findings, a conceptual Clinical Co-pilot Framework is proposed to position GenAI as a collaborative partner that supports clinicians rather than replaces them. The framework provides a conceptual basis for future empirical validation and may help inform the responsible implementation of GenAI in healthcare.**

Keywords*—Generative artificial intelligence; clinical co-pilot; large language models; clinical decision support; human-AI collaboration

I. Introduction

A. Background

The rapid advancement of generative artificial intelligence (GenAI) has significantly transformed digital healthcare and clinical decision support [1]. Since the emergence of advanced large language models (LLMs) and multimodal foundation models, including GPT-4, Med-PaLM, Claude, and domain-specific healthcare models, healthcare organizations have increasingly explored the integration of GenAI into clinical environments [2-6]. Compared with traditional AI systems focused on prediction and classification, contemporary GenAI technologies offer broader capabilities in contextual reasoning, language generation, multimodal interpretation, and human-AI interaction [7].

Recent developments have enabled new forms of clinical support, including electronic health record (EHR) summarization, evidence retrieval, differential diagnosis generation, clinical documentation, and physician communication support [8-14]. The rapid growth of healthcare data and clinical complexity has further accelerated interest in these technologies [15]. Consequently, healthcare institutions are increasingly investigating GenAI as a collaborative tool to augment clinical reasoning, improve workflow efficiency, and reduce administrative burden [16-20].

Recent literature suggests that healthcare applications of GenAI are transitioning from proof-of-concept studies toward real-world clinical implementation [21-23]. Current applications include clinical documentation, retrieval-augmented decision support, AI-assisted triage, personalized treatment planning, multimodal evidence synthesis, and patient education [24-28]. At the same time, concerns regarding hallucination, bias, transparency, accountability, privacy, and regulatory compliance have emerged as major challenges for responsible deployment [27, 28].

B. Research Motivation

Traditional healthcare artificial intelligence systems have historically emphasized automation-oriented objectives, including disease prediction, risk stratification, image classification, and algorithmic screening [16, 29, 30]. Although these technologies demonstrated promising benchmark performance, many failed to achieve sustainable implementation in real-world clinical environments due to poor workflow integration, limited interpretability, clinician distrust, insufficient adaptability, and limited support for complex collaborative decision-making.

Unlike earlier predictive AI systems, GenAI introduces a different interaction paradigm between clinicians and intelligent systems. Rather than functioning solely as automated prediction

*Corresponding author.

engines, GenAI technologies increasingly operate as collaborative assistants capable of supporting clinical reasoning, information synthesis, workflow coordination, and communication under physician supervision [9, 31]. This shift reflects a broader movement from automation-centered artificial intelligence toward collaboration-centered clinical intelligence.

Within this emerging paradigm, the concept of the "clinical co-pilot" has gained increasing relevance. Inspired by the co-pilot model in aviation, this framework conceptualizes GenAI as an augmentative partner designed to support, rather than replace, healthcare professionals [16, 32]. Physicians remain the final decision-makers and accountable authorities, while GenAI assists by enhancing situational awareness, reducing cognitive burden, and improving access to contextual medical knowledge [16, 33, 34].

This framework is particularly important because healthcare decision-making involves uncertainty, ethical responsibility, contextual judgment, and interpersonal communication [33, 35]. Fully autonomous clinical decision-making remains difficult to implement safely in high-stakes environments. Therefore, collaborative human-AI systems may represent a more feasible and responsible pathway toward intelligent healthcare transformation [10, 34].

Current literature on GenAI in healthcare remains fragmented across technical applications, workflow integration, ethical governance, implementation barriers, and clinical evaluation. Many reviews focus narrowly on ChatGPT performance, benchmark testing, or isolated healthcare applications without systematically conceptualizing how GenAI may function as a collaborative partner within healthcare ecosystems [15, 32, 36]. Therefore, a comprehensive synthesis integrating technological capabilities, clinical workflows, governance mechanisms, and human-AI collaboration remains necessary [16].

C. Problem Statement and Research Gap

The rapid advancement of GenAI, LLMs, and multimodal foundation models has accelerated the adoption of AI technologies across diverse healthcare settings [7]. Recent studies have demonstrated promising applications in clinical decision support, medical documentation, patient communication, diagnostic assistance, treatment planning, and workflow optimization [37-39].

Despite this progress, the literature remains fragmented across disciplines and application domains. Many studies focus on evaluating specific models such as ChatGPT, GPT-4, and Med-PaLM using benchmark datasets such as MedQA and MMLU [15, 35, 40, 41], while others examine applications within specific specialties including radiology, oncology, pathology, emergency medicine, and nursing practice [10,29]. At the same time, substantial attention has been devoted to challenges related to hallucination, explainability, privacy, bias, transparency, and regulatory oversight [9, 20, 29, 42-44].

Although these studies provide valuable insights, relatively few have examined how technological capabilities, clinical workflows, and governance mechanisms collectively contribute to human-AI collaboration in healthcare. Existing reviews remain largely focused on model performance or isolated applications and rarely integrate these perspectives within a unified conceptual framework [18, 45].

As healthcare AI continues to evolve from task-specific automation toward collaborative intelligence, there is a growing need to understand how GenAI supports clinical workflows while operating within ethical, legal, and organizational boundaries [46, 47]. Therefore, this study systematically reviews current evidence and develops a clinical co-pilot framework that integrates technology, workflow, and governance perspectives to explain the evolution of collaborative human-AI healthcare systems.

D. Research Objectives

This study aims to systematically review and synthesize current literature regarding the applications of GenAI in healthcare through the conceptual perspective of the clinical co-pilot framework.

Specifically, this study seeks to achieve the following objectives:

- To systematically identify and classify major application domains of GenAI in clinical decision support and healthcare collaboration.
- To examine how GenAI technologies support clinical reasoning, multimodal information integration, workflow coordination, and healthcare communication.
- To analyze the implementation challenges and governance concerns associated with GenAI deployment, including hallucination risk, accountability, bias, privacy protection, and regulatory compliance.
- To propose a conceptual framework describing GenAI as a collaborative clinical co-pilot operating under human supervision.
- To explore how healthcare AI is evolving from automation-centered systems toward collaboration-centered augmented clinical intelligence.
- To identify current research gaps and future directions for responsible human-AI collaboration in healthcare environments.

The findings suggest that the clinical co-pilot model represents a feasible and responsible pathway for integrating GenAI into healthcare systems while preserving human oversight, ethical governance, and clinician-centered decision-making.

II. METHODS

A. Study Design

This study was designed as a systematic review with thematic synthesis to examine how GenAI has been applied as a collaborative clinical co-pilot in healthcare and clinical decision support. The review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020 (PRISMA 2020) framework to enhance transparency, reproducibility, and methodological rigor [48, 49].

The systematic review component was used to identify, screen, and select relevant studies through a structured and reproducible search process. The thematic synthesis component was used to integrate findings across heterogeneous studies and to develop a conceptual framework explaining the role of GenAI as a clinical co-pilot. This combined approach was appropriate because the included literature covered diverse study designs, clinical settings, model architectures, and implementation contexts.

The review focused on studies published from January 2023 to March 2026, corresponding to the period in which advanced LLMs and multimodal GenAI systems rapidly entered healthcare research and clinical implementation discussions. The unit of analysis was the individual scholarly article.

B. Search Strategy

A structured literature search strategy was developed to identify studies examining the application of GenAI technologies in healthcare settings, with particular emphasis on clinical decision support, healthcare workflows, and human-AI collaboration. The search strategy was designed to ensure comprehensive coverage of the rapidly expanding literature on LLMs, multimodal AI systems, and their integration into clinical practice.

The literature search was conducted using three major academic databases: Web of Science Core Collection, PubMed, and Scopus. These databases were selected because they provide extensive coverage of peer-reviewed research in medicine, healthcare informatics, artificial intelligence, computer science, and interdisciplinary healthcare technologies. The search covered publications from January 2023 to March 2026, corresponding to the period following the emergence and widespread adoption of advanced LLMs and multimodal GenAI systems.

This specific time frame was intentionally selected because the public release of ChatGPT in late 2022 marked a major turning point in the development of GenAI technologies. Following this milestone, healthcare research experienced a rapid increase in studies involving LLMs, multimodal foundation models, retrieval-augmented generation (RAG), and agentic AI applications. Consequently, the period from 2023 to March 2026 captures the early evolution of GenAI from experimental demonstrations toward clinical implementation, workflow integration, and healthcare governance. This timeframe therefore provides an appropriate foundation for examining the emergence of GenAI as a clinical co-pilot in healthcare environments.

To maximize retrieval sensitivity while maintaining conceptual relevance, the search strategy incorporated three core domains:

- GenAI technologies, including LLMs, foundation models, and multimodal AI systems;
- Clinical intelligence and healthcare support functions, including clinical decision support, clinical reasoning, workflow assistance, and human-AI collaboration;
- Healthcare environments, including hospitals, clinical practice, healthcare delivery systems, and patient care settings.

Boolean operators (AND, OR), truncation symbols (*), and phrase searching were employed to maximize retrieval sensitivity while minimizing irrelevant records. Database-specific search syntax was adapted for Web of Science, PubMed, and Scopus while maintaining conceptual consistency across all information sources.

The database search yielded a total of 3,124 records, including 957 records from Web of Science, 1,048 records from PubMed, and 1,119 records from Scopus.

The core search syntax was defined as follows:

("generative artificial intelligence"

OR "generative AI"OR "large language model*"OR LLM*OR ChatGPT\OR "foundation model*"OR "multimodal large language model*")

AND

("clinical decision support"OR "clinical decision support system*"OR CDSS OR "clinical decision-making"OR "clinical workflow" OR "clinical co-pilot" OR "medical co-pilot" OR "human-AI collaboration" OR "clinical reasoning" OR "AI agent*"OR "retrieval-augmented generation")

AND

(healthcare OR medicine OR clinical OR hospital OR "patient care")

For the Web of Science Core Collection, searches were conducted using the Topic (TS) field, which includes titles, abstracts, author keywords, and Keywords Plus. In Scopus, searches were performed using the TITLE-ABS-KEY field, while equivalent search syntax was adapted for PubMed. Minor database-specific adjustments were applied where necessary while maintaining conceptual consistency across all information sources.

The search strategy was intentionally designed to capture studies examining the application of GenAI technologies within clinical environments, particularly those involving clinical decision support, healthcare workflows, human-AI collaboration, intelligent clinical assistance, and responsible AI implementation. The broad search approach ensured comprehensive coverage of the rapidly evolving literature while allowing subsequent screening and thematic synthesis to identify the most relevant evidence related to emerging forms of AI-assisted clinical intelligence.

Importantly, the search strategy was intentionally constructed using broad concepts related to GenAI, clinical decision support, healthcare workflows, and human-AI collaboration rather than predefined theoretical categories. The concept of the clinical co-pilot was not used as a mandatory inclusion criterion and did not predetermine the thematic structure of the review.

Instead, the clinical co-pilot framework emerged inductively through the thematic synthesis process as recurring patterns of technological capabilities, workflow integration, and governance mechanisms were identified across the included studies.

The database search process and retrieved records are summarized in Table I.

The initial database search yielded 3,124 records, including 957 records from Web of Science, 1,048 records from PubMed, and 1,119 records from Scopus. Following record export and consolidation, duplicate and non-eligible records were removed prior to screening. The subsequent study selection process was conducted according to the PRISMA 2020 guidelines and involved title and abstract screening, full-text eligibility assessment, and thematic synthesis. The detailed study selection process is presented in Fig. 1.

TABLE I. LITERATURE SEARCH SOURCES AND RETRIEVAL RESULTS

Database	Search Field	Records Retrieved
Web of Science	Topic Search	957
PubMed	Title / Abstract Search	1,048
Scopus	TITLE-ABS-KEY Search	1,119
Total		3,124

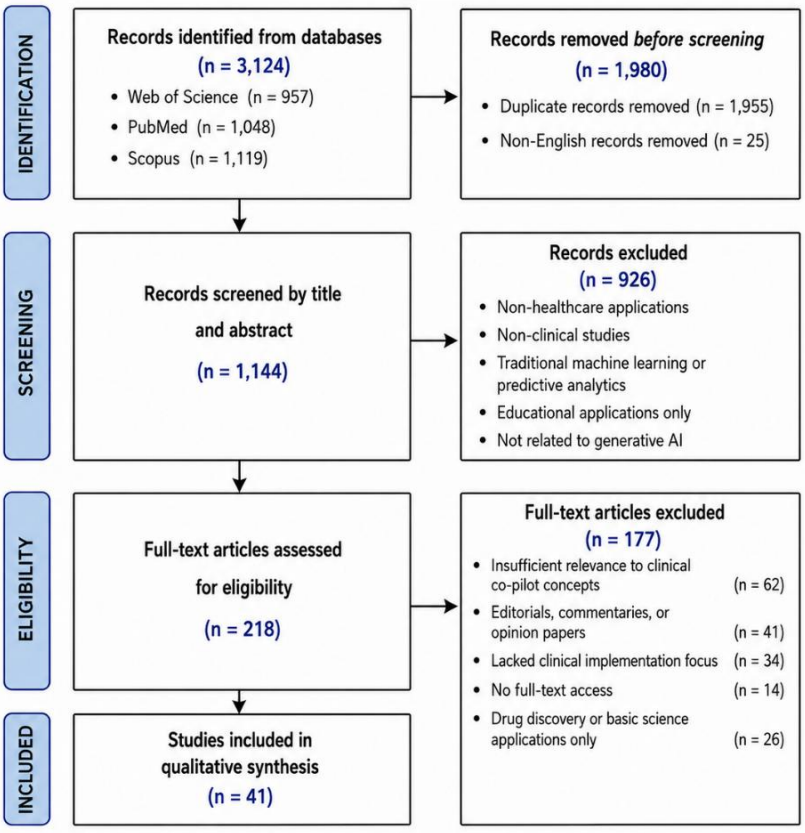


Fig. 1. PRISMA 2020 flow diagram.

C. Inclusion and Exclusion Criteria

Eligibility criteria were defined before screening to ensure consistency in study selection. Studies were included if they met all of the following criteria:

- The study focused on GenAI, LLMs, multimodal generative models, RAG, or related foundation model technologies.
- The study was situated in healthcare, medicine, hospital practice, clinical care, or patient-related decision-making contexts.
- The study discussed or evaluated applications relevant to clinical decision support, clinical workflow assistance, physician decision-making, patient data interpretation, treatment planning, documentation support, or human-AI collaboration.
- The article provided sufficient methodological, conceptual, or application-related information for thematic extraction.
- The article was published in English between January 2023 and March 2026.

Studies were excluded if they met any of the following criteria:

- The study focused only on traditional machine learning or predictive modeling without a GenAI component.
- The study was unrelated to healthcare or clinical practice.

- The study focused solely on medical education, student learning, or examination performance without clear clinical decision support relevance.
- The article was an editorial, commentary, news item, short opinion piece, or letter without sufficient analytical content.
- The study focused exclusively on drug discovery, molecular generation, or basic biomedical research without connection to clinical decision support or healthcare delivery.
- Full text was unavailable, or the article lacked sufficient information for data extraction.

These criteria were designed to keep the review focused on GenAI as a collaborative clinical tool rather than on general AI applications in healthcare.

D. Study Selection

The study selection process followed the PRISMA 2020 flow structure, including identification, screening, eligibility assessment, and final inclusion. First, all records retrieved from the selected databases were compiled. Duplicate records were identified and removed before screening. Next, titles and abstracts were reviewed to exclude clearly irrelevant studies. Articles that appeared potentially relevant were then retrieved for full-text assessment.

During full-text review, each article was evaluated against the predefined inclusion and exclusion criteria. Studies were excluded at this stage if they did not address GenAI, lacked clinical decision support relevance, focused only on education or non-clinical settings, or provided insufficient analytical content. Reasons for exclusion were documented to support PRISMA transparency.

A total of 3,124 records were identified through database searching. Following duplicate removal, title and abstract screening, and full-text eligibility assessment, 41 studies met the inclusion criteria and were included in the final qualitative synthesis. Details of the study selection process are presented in Fig. 1.

E. Data Extraction and Thematic Synthesis

A structured data extraction form was developed to collect key information from each included study. Extracted items included author, publication year, country or region, clinical domain, study design, data source, GenAI model type, clinical application, evaluation method, reported benefits, limitations, governance concerns, and relevance to the clinical co-pilot framework.

The extracted data were organized into two main evidence tables. The first table summarized the characteristics of included studies, including publication year, clinical domain, model type, data source, and study design. The second table classified studies according to their clinical co-pilot functions, including perception and fact anchoring, clinical agency and collaboration, and governance and responsibility boundaries.

Thematic synthesis was conducted using an iterative coding process. First, open coding was used to identify recurring concepts across studies, such as EHR summarization, RAG, ambient documentation, hallucination, accountability, and human-in-the-loop oversight. Second, axial coding was used to group related codes into broader thematic categories. Third, selective coding was used to develop an integrated conceptual framework explaining GenAI as a clinical co-pilot.

Three major themes were identified:

- Perception and fact anchoring: GenAI capabilities for understanding, integrating, retrieving, and contextualizing clinical information.
- Clinical agency and collaboration: GenAI applications that support documentation, diagnosis, treatment planning, workflow coordination, and clinical communication.
- Governance and responsibility boundaries: institutional, ethical, legal, and human oversight mechanisms required for safe clinical deployment.

This thematic synthesis approach allowed the review to move beyond a descriptive summary of applications and toward a conceptual interpretation of how GenAI may function as a responsible and collaborative partner in healthcare.

F. PRISMA Reporting and Methodological Rigor

To strengthen methodological transparency, the findings followed the PRISMA 2020 checklist throughout the search, screening, eligibility assessment, and reporting process. The review explicitly documented the information sources, search strategy, eligibility criteria, selection process, data extraction procedure, and synthesis method.

The PRISMA flow diagram was used to visually summarize the study selection process. Full-text exclusion reasons were categorized and reported to clarify how the final evidence base was constructed. The inclusion and exclusion criteria were applied consistently during screening to reduce selection bias. Because the included studies were expected to be heterogeneous in design, ranging from empirical evaluations and case studies to conceptual frameworks and systematic reviews, a meta-analysis was not conducted. Instead, thematic synthesis was selected as the most appropriate method for integrating diverse evidence types. This approach enabled the identification of cross-cutting themes related to clinical usefulness, implementation feasibility, governance requirements, and human-AI collaboration.

To further enhance rigor, the review distinguished between technical capability, clinical workflow relevance, and governance implications. This distinction helped prevent overgeneralization from benchmark-based performance studies to real-world clinical practice. The final synthesis therefore emphasizes not only what GenAI can do in healthcare, but also how it can be responsibly integrated into clinician-centered decision-making environments.

The review was not prospectively registered in PROSPERO or other systematic review registries because of the rapidly evolving nature of the GenAI evidence base and the conceptual objective of thematic synthesis. The study focused on

conceptual integration and framework development rather than quantitative meta-analysis or intervention effect estimation.

G. Quality Assessment

A methodological quality assessment was conducted to evaluate the rigor and credibility of the included studies.

Given the heterogeneous nature of the evidence base, including empirical studies, implementation reports, conceptual frameworks, and review articles, a simplified quality appraisal framework was adopted. The quality appraisal criteria were adapted from the Mixed Methods Appraisal Tool (MMAT) and previous healthcare artificial intelligence review studies to accommodate the diverse methodological characteristics of the included literature.

Studies were evaluated according to five criteria:

- Clarity of research objectives
- Methodological transparency
- Description of AI technologies
- Clinical relevance
- Governance considerations

Each criterion was evaluated using a three-point scale (0 = not addressed, 1 = partially addressed, and 2 = clearly addressed), resulting in a maximum score of 10 points for each study. Studies scoring 8–10 points were classified as high quality, 5–7 points as moderate quality, and 0–4 points as low quality. Two reviewers independently performed the quality assessment, and disagreements were resolved through discussion and consensus. The quality assessment results of the included studies are summarized in Table II.

TABLE II. SUMMARY OF QUALITY ASSESSMENT RESULTS

Quality Level	Score Range	Number (n)	Percentage (%)
High	8-10	15	36.6%
Moderate	5-7	21	51.2%
Low	0-4	5	12.2%
Total		41	100%

Note: Five quality criteria were evaluated. Total scores ranged from 0 to 10, with higher scores indicating greater methodological rigor.

Overall, most studies were classified as moderate-to-high quality (87.8%), suggesting that the current evidence base provides a reasonably robust foundation for thematic synthesis and supports the development of the proposed conceptual framework. The quality assessment was used to evaluate study rigor and to support the interpretation of thematic findings rather than to exclude studies from the review. No study was excluded solely on the basis of quality assessment results. Given the exploratory and conceptual nature of this study, all eligible studies were retained to ensure comprehensive coverage of emerging evidence related to GenAI clinical co-pilots.

III. RESULTS

A. Overview of Included Studies

A total of 41 studies were included in the final qualitative synthesis following the PRISMA 2020 screening process. These studies were selected from an initial pool of 3,124 records retrieved from Web of Science, PubMed, and Scopus. The included studies were published between 2023 and 2026, reflecting the rapid expansion of GenAI research following the emergence of advanced foundation models and LLMs. The overall publication trend demonstrated a substantial increase in healthcare-related GenAI studies after 2023, particularly in clinical decision support, workflow automation, and healthcare governance applications.

To provide an overview of the included literature, the characteristics of the 41 selected studies were summarized according to publication year, study type, AI technology, clinical focus, and thematic domain (Table III). This descriptive analysis provides contextual insights into the evolving landscape of GenAI research in healthcare and serves as a foundation for the subsequent thematic synthesis. The characteristics of the 41 included studies, including study type, main technology, and primary application domains, are summarized in Table V (Appendix).

TABLE III. OVERVIEW OF INCLUDED STUDIES

Category	Subcategory	Frequency (n)	Percentage (%)
Publication Year	2023	6	14.6
	2024	12	29.3
	2025	18	43.9
	2026	5	12.2
Study Type	Empirical Evaluation	15	36.6
	Framework / Conceptual Study	8	19.5
	Case Study / Implementation	10	24.4
	Review / Perspective	8	19.5
AI Technology	Large Language Models (LLMs)	17	41.5
	Retrieval-Augmented Generation (RAG)	7	17.1
	Multimodal LLMs (MLLMs)	9	22.0
	Agentic AI / AI Agents	4	9.8
	Other GenAI Approaches	4	9.8
Clinical Focus	Clinical Decision Support	14	34.1
	Clinical Documentation	7	17.1
	Personalized Care Planning	5	12.2
	Workflow Optimization	6	14.6
	Governance and Ethics	9	22.0
Thematic Domain	Perception and Fact Anchoring	13	31.7
	Clinical Agency and Collaboration	15	36.6
	Governance and Responsibility	13	31.7

Note: Publications published between January 2023 and March 2026 were included in the review.

The studies were categorized according to their primary research focus, dominant AI technology, and thematic contribution. Each study was assigned to a single primary category within each dimension to facilitate descriptive analysis and thematic synthesis.

The descriptive statistics indicate a rapid increase in healthcare-related GenAI research following the emergence of advanced LLMs and multimodal foundation models. To further illustrate the temporal distribution of the included studies, the annual publication trend is presented in Fig. 2.

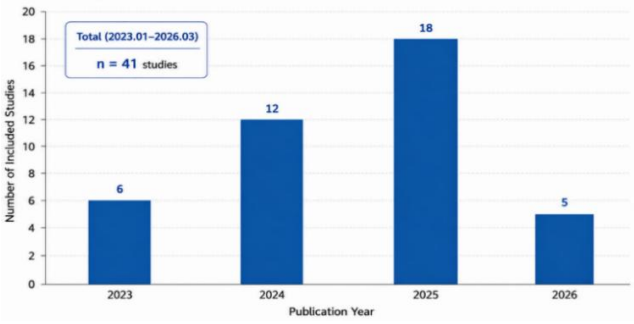


Fig. 2. Publication trends of included studies (2023.01–2026.03).

The reviewed studies originated from diverse disciplinary backgrounds, including healthcare informatics, clinical medicine, digital health, computer science, and interdisciplinary artificial intelligence research. Most studies focused on hospital-based or clinical healthcare environments, while a smaller number examined broader healthcare management systems, patient communication, or institutional governance frameworks.

In terms of technological focus, GPT-based architectures and LLMs were the most frequently discussed GenAI technologies. Commonly referenced systems included GPT-4, ChatGPT, Med-PaLM, Claude, and multimodal large language models (MLLMs). Several studies also explored RAG, domain-specific medical foundation models, and multimodal clinical reasoning systems integrating electronic health records, medical imaging, and laboratory data [3, 5, 50, 51].

The included studies covered multiple clinical application domains. The most frequently identified applications included clinical documentation and summarization, clinical decision support, treatment planning, ambient scribing, patient data interpretation, evidence retrieval, and workflow coordination. Several studies also examined physician acceptance, trust calibration, human-AI collaboration, and governance mechanisms associated with clinical AI deployment.

Methodologically, the reviewed literature was heterogeneous in design. The evidence base included empirical evaluations, implementation case studies, benchmark comparisons, conceptual frameworks, observational studies, expert commentaries, and systematic reviews. Due to the diversity of study designs and evaluation metrics, a quantitative meta-analysis was not conducted. Instead, thematic synthesis was used to identify recurring conceptual patterns across the literature.

Beyond publication growth and technological advancement, thematic synthesis revealed three major domains underlying the evolution of GenAI clinical co-pilots: 1) Perception and Fact Anchoring, 2) Clinical Agency and Collaboration, and 3) Governance and Responsibility. Together, these domains suggest a shift from automation-oriented healthcare AI toward collaborative clinical intelligence systems.

B. Conceptual Framework of GenAI as a Clinical Co-Pilot

The thematic synthesis identified a conceptual shift from automation-oriented healthcare AI toward collaborative clinical intelligence. To explain this transition, a conceptual framework of GenAI as a clinical co-pilot was developed based on the three major themes emerging from the reviewed literature: Perception and Fact Anchoring, Clinical Agency and Collaboration, and Governance and Responsibility.

These three domains collectively describe how GenAI systems transform heterogeneous healthcare data into clinically meaningful intelligence, support healthcare professionals through collaborative reasoning and workflow augmentation, and operate within ethical, legal, and organizational governance structures. Rather than functioning as autonomous decision-makers, GenAI systems are increasingly positioned as collaborative partners that augment clinician capabilities while preserving human oversight and accountability.

As illustrated in Fig. 3, the framework begins with diverse healthcare inputs, including electronic health records, medical imaging, laboratory findings, clinical guidelines, and patient histories. Through multimodal reasoning, evidence grounding, and domain-specific intelligence, these inputs are converted into actionable clinical insights that support diagnosis, documentation, communication, care planning, and clinical decision support. The framework is further reinforced by foundational deployment enablers, including interoperable healthcare data, model transparency, continuous evaluation, and regulatory alignment, which collectively support responsible and trustworthy implementation in healthcare environments.

The framework was developed inductively through thematic synthesis and does not represent a predefined theoretical model. The framework begins with diverse clinical inputs originating from electronic health records, medical imaging, laboratory findings, wearable devices, clinical guidelines, and patient histories. GenAI systems synthesize these heterogeneous data sources through multimodal reasoning, RAG, and domain-specific medical intelligence architectures [23, 36]. The resulting outputs support diagnostic reasoning, clinical documentation, workflow optimization, personalized care planning, interdisciplinary communication, and clinician decision support within real-world healthcare environments [16, 19].

The framework conceptualizes GenAI as a collaborative co-pilot rather than an autonomous medical authority. Across the reviewed studies, clinicians consistently remained the final reviewers and accountable decision-makers within AI-assisted healthcare systems [12, 28]. Consequently, governance mechanisms such as human-in-the-loop oversight, transparency, fairness auditing, privacy protection, and regulatory compliance emerged as foundational requirements for responsible deployment [16, 32].

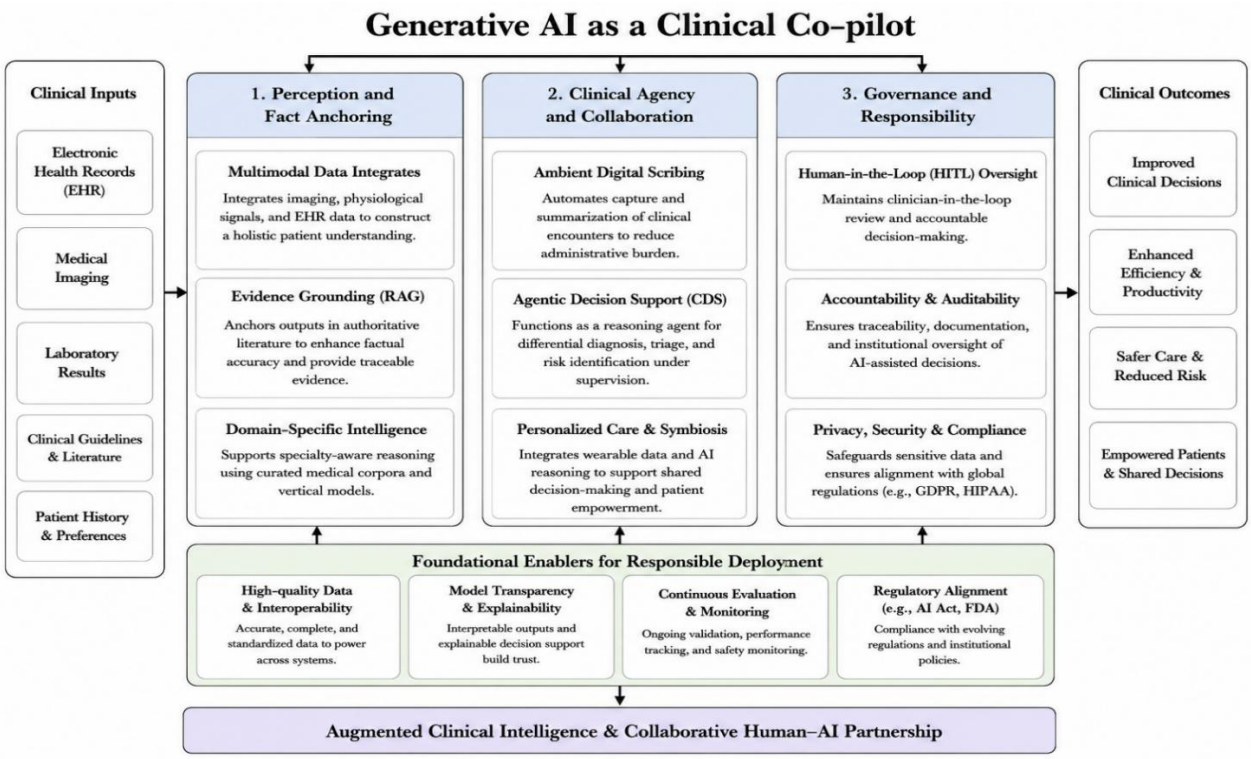


Fig. 3. Clinical Co-pilot framework derived from thematic synthesis of 41 included studies.

TABLE IV. OPTIMIZED THEMATIC CLASSIFICATION OF GENERATIVE AI CLINICAL CO-PILOT FUNCTIONS

Main Theme	Subtheme	AI Co-pilot Role & Functionality	Representative Technologies & Applications	Clinical Significance & Evaluation Metrics	Supporting Studies
Perception and Fact Anchoring	Multimodal Data Integration	Integrates imaging, physiological signals, and EHR data to construct a holistic patient understanding.	MLLMs (e.g., Med-Flamingo, LLaVA-Med), vision-language alignment, multimodal EHR fusion.	Enhances clinical situational awareness; Metrics: Diagnostic accuracy and cross-modal consistency.	[11, 14, 18, 27, 29, 33, 36, 44]
	Evidence Grounding (RAG)	Anchors outputs in authoritative literature to eliminate hallucinations and provide traceable citations.	Retrieval-Augmented Generation (RAG), hybrid retrieval pipelines, evidence-grounded CDS.	Establishes evidence-based trust; Metrics: ACUTE (Traceability, Factuality).	[7, 9, 10, 13, 16, 21, 23, 41, 43, 52]
	Domain-Specific Intelligence	Supports specialty-aware reasoning (e.g., oncology, surgery) using curated medical corpora.	Vertical medical models (BioBERT, ClinicalBERT), disease-scenario association matrices.	Improves specialty-specific precision and relevance; Metrics: MedQA board-style benchmarks.	[6, 9, 11, 14, 21-24, 41, 42, 44, 53, 54]
Clinical Agency and Collaboration	Ambient Digital Scribing	Automates capture and summarization of clinical encounters to reduce administrative burden.	Ambient AI (DAX Co-pilot), SCRIBE framework, automated SOAP note generation.	Reduces physician burnout and documentation time; Metrics: Note completeness and editing time.	[6, 8, 11, 12, 16, 19, 30, 31, 33, 44]
	Agentic Decision Support (Agentic CDS)	Functions as a reasoning "agent" capable of differential diagnosis and active triage under supervision.	AI Agents, virtual triage assistants, interactive diagnostic rationales.	Supports complex reasoning and risk identification; Metrics: AI-SCE (Clinical Skills Assessment).	[13, 15, 16, 18, 19, 23, 28, 33, 35-37, 41, 54, 55]
	Personalized Care & Symbiosis	Integrates wearable data and AI reasoning to support shared decision-making and patient empowerment.	Personalized Patient Care Plans (PPCPs), remote monitoring, health empowerment tools.	Enhances care coordination and patient adherence; Metrics: Patient satisfaction and compliance.	[9, 16, 18-20, 24, 26-30, 36, 44]
Governance and Responsibility	Human-in-the-Loop (HITL) Oversight	Maintains clinicians as the final reviewers and accountable decision-makers.	iLEAP lifecycle model, clinician validation interfaces, supervised CDS protocols.	Preserves diagnostic safety and legal accountability; Metrics: Human-AI decision agreement rates.	[10, 12, 16, 18, 19, 23, 24, 28, 33, 37, 44, 52, 54]
	Privacy, Security, & Compliance	Safeguards sensitive data and ensures alignment with global regulations (GDPR/HIPAA).	Data de-identification, secure infrastructure, quantitative assurance frameworks.	Reduces liability and privacy risks; Metrics: Regulatory compliance and data leakage rates.	[7, 8, 10, 14, 16-19, 23, 28, 32, 35, 37, 42-44, 52, 56]
	Ethics, Fairness, & Trustworthy AI	Mitigates algorithmic bias and promotes transparency and equitable healthcare delivery.	Fairness audits, AVI intelligent filters, explainable AI (XAI), bias mitigation.	Supports health equity and trustworthy deployment; Metrics: Bias reduction rate and explainability scores.	[9, 18, 19, 23, 25, 27, 28, 32, 35, 37, 42-44, 52, 53]

To further synthesize the major themes identified across the included studies, Table IV presents the thematic classification of GenAI clinical co-pilot functions. The classification was developed inductively through thematic synthesis of the included studies and serves as an analytical framework for organizing the identified clinical co-pilot functions. Table IV summarizes the conceptual roles, representative technologies, clinical significance, evaluation considerations, and supporting evidence associated with each thematic domain. The findings suggest that contemporary GenAI systems are evolving beyond isolated automation tools toward integrated collaborative intelligence infrastructures capable of supporting multimodal reasoning, clinical workflow augmentation, and responsible healthcare governance.

C. Perception and Fact Anchoring

The first major thematic domain identified in the findings involved perception and fact anchoring, representing the foundational cognitive layer of GenAI clinical co-pilot systems. This domain focuses on the ability of GenAI to interpret, integrate, contextualize, and retrieve heterogeneous healthcare information in order to support clinical reasoning and evidence-based decision-making.

One of the most frequently discussed capabilities involved multimodal data integration. Several studies described the emergence of multimodal LLMs capable of combining electronic health records, medical imaging, laboratory findings, physiological signals, and patient histories within unified reasoning systems [36]. Unlike conventional language-only architectures, these multimodal systems provide a more comprehensive understanding of patient conditions by integrating diverse clinical modalities into a shared contextual framework [33].

Another major subtheme involved evidence grounding through RAG. The reviewed studies consistently identified hallucination and factual inconsistency as major barriers to clinical adoption [18]. To address these limitations, multiple studies proposed integrating external biomedical literature, clinical guidelines, institutional protocols, and curated medical repositories into GenAI pipelines. RAG-based systems were frequently described as mechanisms for improving factual reliability, traceability, and evidence-based trust within clinical decision support environments [13, 16].

Thematic analysis also highlighted the increasing importance of domain-specific medical intelligence. Several studies emphasized the development of specialty-oriented foundation models and curated medical corpora designed for vertical clinical applications such as oncology, radiology, pathology, surgery, and critical care [14, 18, 21, 27, 36]. These systems improve contextual relevance and specialty-specific reasoning performance by embedding structured medical knowledge and domain-specific terminology into the model architecture.

Collectively, these findings suggest that perception and fact anchoring functions form the informational and cognitive foundation of GenAI clinical co-pilot systems. By improving contextual understanding, evidence retrieval, and multimodal interpretation, these capabilities support more reliable and clinically relevant AI-assisted healthcare environments.

D. Clinical Agency and Collaboration

The second major thematic domain identified in the findings involved clinical agency and collaboration, representing the operational layer of GenAI integration within healthcare workflows. This domain focuses on how GenAI systems support clinicians during routine clinical activities, documentation processes, communication tasks, and collaborative decision-making.

One of the most widely discussed applications involved ambient digital scribing and automated clinical documentation [12]. Several studies examined ambient AI systems capable of capturing physician-patient conversations and automatically generating structured clinical documentation, including SOAP notes, discharge summaries, and encounter records [7, 35]. These applications were consistently associated with reductions in administrative burden, documentation workload, and physician burnout [30].

Another major theme involved agentic clinical decision support (CDS). Unlike traditional rule-based CDS systems, GenAI systems increasingly function as interactive reasoning agents capable of differential diagnosis generation, triage support, contextual recommendation synthesis, and conversational clinical reasoning [15, 30, 54, 55]. Several studies described these systems as “clinical assistants,” “digital residents,” or “AI collaborators” designed to augment physician reasoning while preserving clinician supervision and final decision authority [18].

Personalized care and patient empowerment also emerged as important application areas. Multiple studies explored the integration of wearable devices, remote monitoring systems, and personalized patient care plans (PPCPs) into GenAI workflows [16, 28, 29]. These systems were designed to support individualized treatment coordination, shared decision-making, patient engagement, and long-term care continuity [10, 18, 25].

In addition, workflow optimization and interdisciplinary communication were repeatedly identified as key operational benefits [19]. GenAI systems were increasingly used to coordinate clinical tasks, streamline information retrieval, summarize large volumes of medical data, and facilitate communication across multidisciplinary healthcare teams [23]. These capabilities may help reduce cognitive overload and improve healthcare operational efficiency in increasingly complex clinical environments.

Overall, the findings indicate that GenAI is evolving beyond isolated documentation or chatbot functions toward broader collaborative clinical agency and collaboration integration. Within this emerging paradigm, AI systems increasingly function as workflow collaborators embedded within clinician-centered healthcare environments.

E. Governance and Responsibility

The third major thematic domain identified in the findings involved Governance and Responsibility, representing the institutional and ethical foundation required for safe GenAI deployment in healthcare settings. Across the reviewed

literature, governance mechanisms were consistently described as essential prerequisites for trustworthy clinical implementation.

One of the most prominent themes involved human-in-the-loop (HITL) oversight [34, 57]. Multiple studies emphasized that clinicians must remain the final reviewers, supervisors, and accountable decision-makers within AI-assisted healthcare systems [18, 32]. Rather than replacing physician expertise, GenAI systems were consistently positioned as supportive tools operating under clinician supervision [6, 41, 53]. This collaborative oversight structure was widely regarded as essential for maintaining patient safety and preserving clinical accountability [16].

Accountability and auditability also emerged as major governance concerns. Several studies discussed the need for traceable AI-assisted decisions, institutional governance committees, audit trails, and transparent documentation mechanisms capable of supporting legal and organizational accountability. Questions regarding liability attribution, clinician responsibility, and institutional oversight remain unresolved challenges in real-world healthcare deployment [9, 10, 32].

Privacy, security, and regulatory compliance represented another major subtheme. Because GenAI systems frequently process sensitive healthcare information, studies consistently emphasized compliance with GDPR, HIPAA, and healthcare cybersecurity standards. Several authors also highlighted risks related to data leakage, unauthorized access, and secondary use of patient information [8, 21, 42].

Ethics, fairness, and trustworthy AI were similarly recurring themes across the included studies. Multiple studies discussed algorithmic bias, demographic disparities, explainability limitations, and fairness concerns associated with large-scale GenAI deployment [9, 18]. Consequently, fairness audits, transparency mechanisms, explainable AI (XAI), and bias mitigation strategies were frequently proposed as critical components of responsible healthcare AI governance [8, 25, 43].

Collectively, the findings suggest that governance and responsibility mechanisms are not peripheral considerations, but foundational requirements for sustainable clinical AI integration. Successful implementation depends not only on technical performance, but also on the establishment of trustworthy human-AI collaboration frameworks supported by accountability, transparency, fairness, and clinician oversight.

IV. Discussion

A. Theoretical and Practical Contributions

This study makes several important theoretical and practical contributions to the emerging field of GenAI in healthcare. First, by systematically reviewing and synthesizing 41 studies published between 2023 and 2026, this research provides a comprehensive overview of the rapidly evolving literature on GenAI applications in healthcare. The findings consolidate fragmented evidence from medicine, healthcare informatics, artificial intelligence, and digital health into a coherent knowledge framework.

This study extends existing review literature that has primarily focused on ChatGPT performance, specialty-specific applications, or technical evaluations of LLMs [18, 47]. To the study's knowledge, this is among the first reviews to conceptualize GenAI as a clinical co-pilot through the integration of technological capabilities, clinical workflow support, and governance considerations. This broader perspective helps explain the evolution of GenAI from isolated tools toward collaborative clinical intelligence systems.

Second, this study proposes a conceptual framework that positions GenAI as a collaborative clinical co-pilot rather than a standalone decision-making system. The framework advances existing discussions beyond traditional clinical decision support systems by emphasizing the complementary relationship between human expertise and AI-enabled intelligence. This perspective contributes to the growing body of literature on human-AI collaboration and collaborative clinical intelligence.

Third, the thematic synthesis identifies three interrelated domains that characterize the emerging role of GenAI in healthcare: 1) Perception and Fact Anchoring, 2) Clinical Agency and Collaboration, and 3) Governance and Responsibility. These domains provide a structured theoretical foundation for understanding how GenAI systems acquire, interpret, and utilize clinical information while operating within healthcare workflows and regulatory environments.

From a practical perspective, the proposed framework may provide guidance for healthcare organizations, technology developers, and policymakers seeking to implement GenAI responsibly. The findings highlight the importance of evidence grounding, workflow integration, human oversight, transparency, and regulatory compliance as critical factors for successful deployment. Furthermore, the framework may provide a foundation for future empirical studies examining the effectiveness, safety, and long-term impact of GenAI clinical co-pilots in real-world healthcare settings.

B. From Automation to Collaboration

Traditional healthcare artificial intelligence systems primarily emphasized automation-oriented objectives, including disease prediction, risk stratification, image classification, and algorithmic screening [8, 17, 22]. Although many of these systems demonstrated strong benchmark performance, real-world clinical implementation remained limited due to poor workflow integration, insufficient interpretability, clinician distrust, and limited adaptability to complex healthcare environments [9, 18].

The findings suggest an emerging conceptual transition from automation-centered artificial intelligence toward collaboration-centered clinical intelligence. Unlike earlier predictive models that functioned primarily as isolated computational tools, contemporary GenAI systems increasingly operate as interactive clinical collaborators capable of contextual reasoning, multimodal information synthesis, evidence retrieval, and dynamic communication under clinician supervision [9, 35].

Within the proposed clinical co-pilot framework, the role of artificial intelligence shifts from replacing human decision-making toward augmenting clinician capabilities [35]. Rather than functioning as autonomous decision-makers, GenAI

systems support healthcare professionals by improving situational awareness, reducing cognitive burden, streamlining information retrieval, and facilitating workflow coordination [18, 35].

This transition is particularly important in healthcare because clinical decision-making inherently involves uncertainty, contextual judgment, ethical responsibility, interdisciplinary communication, and patient-centered reasoning. Traditional machine learning systems may generate accurate predictions, but they often lack the ability to explain reasoning pathways, contextualize recommendations, or engage in iterative clinical dialogue [16, 32]. In contrast, GenAI co-pilots demonstrate the ability to synthesize heterogeneous information sources, generate contextual explanations, and support collaborative reasoning processes within clinical environments.

Although the reviewed studies generally reported positive outcomes, the magnitude of these benefits varied across clinical applications. Administrative tasks, such as clinical documentation and ambient digital scribing, demonstrated more consistent performance improvements because they involve relatively standardized workflows and lower levels of clinical uncertainty. In contrast, applications involving diagnostic reasoning, treatment planning, and agentic clinical decision support exhibited greater variability, reflecting differences in domain complexity, multimodal data integration, evidence grounding capabilities, and the degree of required clinician oversight. These contextual differences suggest that the effectiveness of GenAI depends not only on model capability but also on the clinical context in which it is deployed.

Thematic synthesis further suggests that the future direction of healthcare AI may depend less on fully autonomous systems and more on collaborative intelligence architectures integrating human expertise with AI-supported reasoning[35]. Consequently, the clinical co-pilot paradigm may represent a more feasible, sustainable, and ethically responsible pathway for healthcare digital transformation.

C. Human-AI Partnership and Trust Calibration

The effectiveness of GenAI clinical co-pilots depends heavily on the establishment of appropriate human-AI partnerships and calibrated clinician trust. Across the reviewed literature, trust calibration emerged as a critical factor influencing healthcare AI adoption, workflow integration, and long-term clinical acceptance[35].

Overtrust and undertrust both represent substantial risks in AI-assisted healthcare environments [10]. Excessive reliance on AI-generated recommendations may increase vulnerability to hallucinations, factual inaccuracies, and automation bias [30]. Conversely, insufficient trust may prevent clinicians from utilizing potentially valuable decision support capabilities and operational efficiencies [37].

The reviewed studies consistently identified RAG, evidence grounding, and domain-specific medical intelligence as essential mechanisms for improving trust calibration [13, 51]. By anchoring outputs to authoritative medical literature, institutional guidelines, and curated biomedical knowledge sources, RAG-based systems improve factual consistency, transparency, and traceability. These mechanisms help transform GenAI from a probabilistic text generator into a more reliable evidence-supported clinical assistant [13].

Trust calibration is also associated with workflow integration quality. Studies examining ambient digital scribing and workflow-oriented clinical AI systems demonstrated that clinician acceptance increases when AI systems operate seamlessly within clinical workflows without introducing excessive operational disruption. Automated clinical documentation, structured note generation, and intelligent information summarization were associated with reduced administrative burden and lower physician burnout [12, 30].

Another important finding involved the emergence of agentic clinical decision support systems. Several studies described GenAI systems functioning as interactive reasoning agents capable of participating in diagnostic discussions, differential diagnosis generation, triage support, and multidisciplinary information synthesis [15, 55]. However, the reviewed literature consistently emphasized that these systems should remain clinician-supervised collaborative assistants rather than independent clinical authorities [28, 41].

Overall, the findings suggest that sustainable healthcare AI adoption depends not only on technical performance, but also on the development of trustworthy human-AI collaboration models characterized by transparency, explainability, contextual reliability, and workflow compatibility. Beyond technical performance, the heterogeneity observed across the reviewed studies suggests that successful implementation depends on organizational readiness, workflow redesign, clinician acceptance, and regulatory maturity. Consequently, similar GenAI technologies may produce different outcomes across healthcare settings, indicating that implementation context is as important as algorithmic performance.

D. Governance, Responsibility, and De-Identification Challenges

As GenAI systems transition from experimental research environments toward real-world healthcare implementation, governance and responsibility mechanisms become increasingly critical. The findings consistently suggest that technical performance alone is insufficient for sustainable healthcare AI deployment without corresponding institutional, ethical, and regulatory safeguards.

Human-in-the-loop oversight emerged as one of the most consistently emphasized governance principles across the reviewed studies [42, 56]. Regardless of the sophistication of GenAI systems, clinicians were repeatedly identified as the final reviewers, supervisors, and accountable decision-makers within healthcare environments [12, 28, 41]. This governance structure is particularly important because healthcare decisions involve substantial ethical responsibility, patient safety considerations, and legal accountability.

The reviewed literature also identified unresolved challenges regarding liability attribution and institutional accountability [10, 32]. Questions remain regarding how responsibility should be distributed when AI-assisted recommendations contribute to unsafe decisions, delayed diagnoses, or inappropriate treatments [8, 25, 41]. These concerns become increasingly complex as

GenAI systems evolve toward more autonomous agentic capabilities.

Privacy protection and healthcare data security were similarly identified as major implementation barriers. GenAI systems frequently process highly sensitive multimodal healthcare information, including electronic health records, radiological images, laboratory data, wearable sensor outputs, and free-text clinical narratives. Traditional de-identification approaches may become insufficient in large language model environments because contextual associations within unstructured clinical text may still permit indirect patient re-identification [17, 53].

Several studies emphasized that advanced LLMs possess strong contextual inference capabilities capable of reconstructing sensitive information from fragmented clinical details. Consequently, healthcare institutions may require more advanced privacy-preserving architectures, secure deployment infrastructures, and continuous monitoring systems to reduce data leakage risks [14, 52, 56].

The reviewed studies further emphasized the importance of transparency, fairness, explainability, and regulatory alignment. Bias inherited from historical healthcare datasets may contribute to unequal performance across demographic populations, clinical institutions, and healthcare settings. Therefore, fairness auditing, explainable AI mechanisms, and continuous governance evaluation were frequently proposed as essential components of trustworthy healthcare AI deployment [10, 32].

In addition, emerging regulatory frameworks such as the European Union AI Act and evolving FDA guidance may significantly influence future healthcare AI governance. However, current regulations continue to face challenges adapting to rapidly evolving GenAI systems characterized by dynamic reasoning capabilities, continuous model updates, and multimodal integration.

Notably, several reviewed studies also highlighted practical implementation challenges, including clinician skepticism, workflow disruption, and uncertainty regarding legal responsibility, hallucinations, and concerns over data privacy. These findings suggest that successful GenAI adoption depends not only on advances in model capability but also on organizational adaptation, effective governance, and sustained human oversight throughout clinical implementation.

Collectively, these findings suggest that governance and responsibility should not be viewed as secondary considerations, but rather as foundational infrastructure supporting safe, ethical, and sustainable healthcare AI integration.

E. Study Limitations

Several limitations should be acknowledged when interpreting these findings. First, only English-language publications were included in the analysis. Although English represents the dominant language of scientific communication, relevant studies published in other languages may have been excluded, potentially introducing language bias.

Second, the literature search was limited to three major databases: Web of Science, PubMed, and Scopus. While these databases provide extensive coverage of healthcare, medical informatics, and artificial intelligence research, relevant studies indexed in other databases or grey literature sources may not have been captured.

Third, this study employed a qualitative thematic synthesis approach rather than a quantitative meta-analysis. Given the heterogeneity of study designs, evaluation methods, and reported outcomes, direct statistical comparison across studies was not feasible. Consequently, the findings should be interpreted as conceptual and thematic insights rather than quantitative estimates of effectiveness.

Finally, GenAI is a rapidly evolving field characterized by continuous technological advancements and frequent model updates. As new foundation models, multimodal systems, and regulatory frameworks continue to emerge, some findings may require further validation and refinement over time. Therefore, the proposed framework should be viewed as an evolving conceptual foundation that may be expanded as the evidence base continues to grow.

F. Future Directions

The findings suggest several important future directions for GenAI clinical co-pilot research and healthcare implementation.

First, future studies should move beyond static benchmark evaluations toward longitudinal real-world clinical assessments. Many current studies rely heavily on benchmark datasets, examination-style evaluations, or proof-of-concept experiments. Although these approaches provide useful technical comparisons, they do not adequately capture long-term workflow integration, clinician adaptation, patient-centered outcomes, or operational sustainability [9, 36]. Future research should therefore prioritize longitudinal observational studies, implementation science approaches, and pragmatic clinical evaluations conducted within active healthcare environments.

Second, future clinical co-pilot systems will likely require deeper integration with multimodal healthcare ecosystems. The reviewed literature suggests that the next generation of healthcare AI may extend beyond isolated language processing toward comprehensive multimodal reasoning architectures integrating electronic health records, medical imaging, laboratory findings, wearable devices, physiological monitoring, and behavioral data streams. Such systems may facilitate more personalized, context-aware, and continuous healthcare support [8, 27].

Third, uncertainty quantification and explainability remain important future research priorities [58, 59]. Current GenAI systems frequently produce highly fluent outputs without explicitly communicating confidence levels or uncertainty boundaries [18, 36]. Future clinical co-pilots should incorporate transparent uncertainty signaling mechanisms capable of informing clinicians when recommendations require additional validation or specialist review. Such mechanisms may help reduce automation bias and improve clinical risk management.

Fourth, future governance frameworks must address distributed responsibility within collaborative human-AI healthcare systems. As GenAI systems become more deeply integrated into healthcare workflows, questions regarding legal liability, institutional accountability, and clinician responsibility

will become increasingly important [9, 10, 35]. Interdisciplinary collaboration among healthcare professionals, legal scholars, policymakers, and AI researchers will likely be necessary to establish operationally feasible governance standards.

Finally, the concept of clinical intelligence itself may continue to evolve. The reviewed literature suggests that healthcare AI is gradually transitioning from isolated decision support tools toward collaborative intelligence ecosystems integrating clinicians, patients, healthcare institutions, and AI-supported reasoning systems [28, 32]. Within this evolving paradigm, GenAI co-pilots may become foundational infrastructure supporting future human-centered digital healthcare environments.

V. Conclusion

This study systematically reviewed and synthesized recent research on GenAI in healthcare. Following the PRISMA 2020 framework, 41 studies published between 2023 and March 2026 were included in the final analysis. The findings indicate that GenAI is expanding beyond traditional decision-support functions and becoming more closely integrated into clinical workflows.

Three major themes emerged from the thematic synthesis: Perception and Fact Anchoring, Clinical Agency and Collaboration, and Governance and Responsibility. These themes reflect the key capabilities and challenges associated with the adoption of GenAI in healthcare. The findings show that technologies such as LLMs, multimodal foundation models, RAG, and agent-based systems are being applied across clinical documentation, decision support, patient communication, and care planning. At the same time, issues related to transparency, accountability, privacy protection, and human oversight remain central to successful implementation.

Based on these findings, this study proposes a Clinical Co-pilot Framework that integrates technological capabilities, clinical workflow support, and governance considerations. The framework may provide a conceptual perspective for understanding how GenAI can assist healthcare professionals while maintaining clinician responsibility for medical decision-making. The results suggest that the value of GenAI in healthcare depends not only on model performance but also on its ability to support clinical practice in a safe, reliable, and accountable manner. Effective deployment requires appropriate workflow integration, continuous evaluation, and clear governance mechanisms. Human oversight remains essential in ensuring the quality and safety of AI-assisted healthcare services.

Future research should focus on real-world implementation, long-term clinical impact, and the evolving relationship between healthcare professionals and AI systems. As GenAI technologies continue to develop. The future of healthcare AI may depend not on replacing clinicians, but on establishing effective human-AI partnerships that enhance clinical intelligence, support responsible decision-making, and improve patient-centered healthcare delivery.

References

[1] M. Elhaddad and S. Hamam, "AI-driven clinical decision support systems: an ongoing pursuit of potential," Cureus, vol. 16, no. 4, p. e57728, 2024.

[2] K. Singhal et al., "Large language models encode clinical knowledge," Nature, vol. 620, no. 7972, pp. 172-180, 2023.

[3] R. AlSaad et al., "Multimodal large language models in health care: applications, challenges, and future outlook," Journal of medical Internet research, vol. 26, p. e59505, 2024.

[4] Q. Jin et al., "Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine," NPJ Digital Medicine, vol. 7, no. 1, p. 190, 2024.

[5] G. Alain, J. Crick, E. Snead, C. C. Quatman-Yates, and C. E. Quatman, "Evaluating user interactions and adoption patterns of generative AI in health care occupations using claude: Cross-Sectional study," Journal of Medical Internet Research, vol. 27, p. e73918, 2025.

[6] S. Wojcik, A. Rulkiewicz, P. Pruszczyk, W. Lisik, M. Pobozy, and J. Domienik-Karlowicz, "Beyond ChatGPT: What does GPT-4 add to healthcare? The dawn of a new era," Cardiol J, vol. 30, no. 6, pp. 1018-1025, 2023, doi: 10.5603/cj.97515.

[7] X. Chen, L. Zhou, J. Chen, G. Wang, and X. Li, "How is language intelligence evolving? A multi-dimensional survey of large language models," Expert Systems with Applications, vol. 304, 2026, doi: 10.1016/j.eswa.2025.130637.

[8] D. S. Comeau, D. S. Bitterman, and L. A. Celi, "Preventing unrestricted and unmonitored AI experimentation in healthcare through transparency and accountability," NPJ Digit Med, vol. 8, no. 1, p. 42, Jan 18 2025, doi: 10.1038/s41746-025-01443-2.

[9] Y. Artsi et al., "Challenges of Implementing LLMs in Clinical Practice: Perspectives," J Clin Med, vol. 14, no. 17, Sep 1 2025, doi: 10.3390/jcm14176169.

[10] T. Tung, S. M. N. Hasnaeen, and X. Zhao, "Ethical and practical challenges of generative AI in healthcare and proposed solutions: a survey," Front Digit Health, vol. 7, p. 1692517, 2025, doi: 10.3389/fdgth.2025.1692517.

[11] J. J. Andrew, J. Potier, N. Garcelon, A. Burgun, and M. Vincent, "Using large language models for temporal relation extraction from pediatric clinical reports," JAMIA open, vol. 8, no. 6, p. ooaf121, 2025, doi: 10.1093/jamiaopen/ooaf121 JAMIA.

[12] H. Wang et al., "An evaluation framework for ambient digital scribing tools in clinical applications," NPJ Digit Med, vol. 8, no. 1, p. 358, Jun 13 2025, doi: 10.1038/s41746-025-01622-1.

[13] B. B. Ozmen and P. Mathur, "Evidence-based artificial intelligence: Implementing retrieval-augmented generation models to enhance clinical decision support in plastic surgery," J Plast Reconstr Aesthet Surg, vol. 104, pp. 414-416, May 2025, doi: 10.1016/j.bjps.2025.03.053.

[14] J. Shen et al., "The foundational cornerstone for healthcare AI models: constructing multimodal corpora in the health sector," Chinese Science Bulletin, vol. 70, no. 26, pp. 4560-4568, 2025, doi: 10.1360/tb-2025-0185.

[15] N. Mehandru, B. Y. Miao, E. R. Almaraz, M. Sushil, A. J. Butte, and A. Alaa, "Evaluating large language models as agents in the clinic," NPJ Digit Med, vol. 7, no. 1, p. 84, Apr 3 2024, doi: 10.1038/s41746-024-01083-y.

[16] S. A. Rabbani et al., "Generative Artificial Intelligence in Healthcare: Applications, Implementation Challenges, and Future Directions," BioMedInformatics, vol. 5, no. 3, 2025, doi: 10.3390/biomedinformatics5030037.

[17] S. Guleria, J. Guptill, I. Kumar, M. McClintic, and J. C. Rojas, "Artificial intelligence integration in healthcare: perspectives and trends in a survey of U.S. health system leaders," BMC Digital Health, vol. 2, no. 1, 2024, doi: 10.1186/s44247-024-00135-3.

[18] S. Maity and M. J. Saikia, "Large Language Models in Healthcare and Medical Applications: A Review," Bioengineering (Basel), vol. 12, no. 6, Jun 10 2025, doi: 10.3390/bioengineering12060631.

[19] S. Frid et al., "Bridging generative AI and healthcare practice: insights from the GenAI Health Hackathon at Hospital Clinic de Barcelona," BMJ Health Care Inform, vol. 32, no. 1, Oct 15 2025, doi: 10.1136/bmjhci-2025-101640.

[20] H. Li, "AI Foundations in China's Medical Physiology Education: Pedagogical Practices and Systemic Challenges," Adv Med Educ Pract, vol. 16, pp. 1439-1453, 2025, doi: 10.2147/AMEP.S532951.

[21] C. Rainey, A. England, P. C. Murphy, A. A. Mohammad, Y. H. Hadi, and M. McEntee, "Large language models (LLMs) in radiography research: A narrative review," Radiography (Lond), vol. 32, no. 1, p. 103244, Jan 2026, doi: 10.1016/j.radi.2025.103244.

[22] R. Kawasaki, "How Can Artificial Intelligence Be Implemented Effectively in Diabetic Retinopathy Screening in Japan?," Medicina (Kaunas), vol. 60, no. 2, Jan 30 2024, doi: 10.3390/medicina60020243.

[23] J. D. Workum, D. van de Sande, D. Gommers, and M. E. van Genderen, "Bridging the gap: a practical step-by-step approach to warrant safe implementation of large language models in healthcare," Front Artif Intell, vol. 8, p. 1504805, 2025, doi: 10.3389/frai.2025.1504805.

[24] Y. Tsujimoto, M. Shiraishi, H. Yamanaka, H. Lee, M. Okazaki, and N. Morimoto, "Virtual patient modeling for generative-AI-assisted treatment decision-making in lymphedema care: AI tends to favor more aggressive treatment," Eur J Surg Oncol, vol. 52, no. 3, p. 111391, Mar 2026, doi: 10.1016/j.ejso.2026.111391.

[25] A. Pal et al., "Generative AI/LLMs for Plain Language Medical Information for Patients, Caregivers and General Public: Opportunities, Risks and Ethics," Patient Prefer Adherence, vol. 19, pp. 2227-2249, 2025, doi: 10.2147/PPA.S527922.

[26] O. G. Asai, N. Sabu, P. Gondode, S. Das, and G. Chauhan, "Comparative Evaluation of Popular Gen-AI Chatbots in Generating Patient Education Material on Pulmonary Artery Catheter Insertion," Ann Card Anaesth, vol. 29, no. 1, pp. 81-88, Jan 1 2026, doi: 10.4103/aca.aca_145_25.

[27] H. W. Jung, J. Y. Park, T. Holoubek, W. J. Kim, and J. Park, "Socially Assistive Robots in Mental Healthcare: Principles and Conceptual Framework for User-Centered Design," International Journal of Social Robotics, vol. 17, no. 11, pp. 2827-2851, 2025, doi: 10.1007/s12369-025-01323-5.

[28] R. J. Krumsvik and V. Slettvoll, "Artificial intelligence and health empowerment in rural communities and landslide- or avalanche-isolated contexts: real case at a fictitious location," Front Digit Health, vol. 7, p. 1655154, 2025, doi: 10.3389/fdgth.2025.1655154.

[29] M. M. Baig, C. Hobson, H. GholamHosseini, E. Ullah, and S. Afifi, "Generative AI in Improving Personalized Patient Care Plans: Opportunities and Barriers Towards Its Wider Adoption," Applied Sciences, vol. 14, no. 23, 2024, doi: 10.3390/app142310899.

[30] J. Pawelczyk, M. Kraus, S. Voigtlaender, S. Siebenlist, and M. C. Rupp, "Advancing Musculoskeletal Care Using AI and Digital Health Applications: A Review of Commercial Solutions," HSS J, vol. 21, no. 3, pp. 331-341, Aug 2025, doi: 10.1177/15563316251341321.

[31] J. L. Painter, D. Ramcharran, and A. Bate, "Perspective review: Will generative AI make common data models obsolete in future analyses of distributed data networks?," Ther Adv Drug Saf, vol. 16, p. 20420986251332743, 2025, doi: 10.1177/20420986251332743.

[32] P. Goktas and A. Grzybowski, "Shaping the Future of Healthcare: Ethical Clinical Challenges and Pathways to Trustworthy AI," J Clin Med, vol. 14, no. 5, Feb 27 2025, doi: 10.3390/jcm14051605.

[33] F. Pesapane, R. Cuocolo, and F. Sardanelli, "The Picasso's skepticism on computer science and the dawn of generative AI: questions after the answers to keep "machines-in-the-loop"," Eur Radiol Exp, vol. 8, no. 1, p. 81, Jul 24 2024, doi: 10.1186/s41747-024-00485-7.

[34] J. S. Andersen and W. Maalej, "Design patterns for machine learning-based systems with humans in the loop," IEEE Software, vol. 41, no. 4, pp. 151-159, 2023.

[35] J. C. L. Chow and K. Li, "Large Language Models in Medical Chatbots: Opportunities, Challenges, and the Need to Address AI Risks," Information, vol. 16, no. 7, 2025, doi: 10.3390/info16070549.

[36] P. Gupta et al., "Rapid review: Growing usage of Multimodal Large Language Models in healthcare," J Biomed Inform, vol. 169, p. 104875, Sep 2025, doi: 10.1016/j.jbi.2025.104875.

[37] T. Mirzaei, L. Amini, and P. Esmaeilzadeh, "Clinician voices on ethics of LLM integration in healthcare: a thematic analysis of ethical concerns and implications," BMC Med Inform Decis Mak, vol. 24, no. 1, p. 250, Sep 9 2024, doi: 10.1186/s12911-024-02656-3.

[38] L. Huang et al., "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," ACM Transactions on Information Systems, vol. 43, no. 2, pp. 1-55, 2025.

[39] Z. Zhu et al., "Can we trust AI doctors? a survey of medical hallucination in large language and large vision-language models," in Findings of the Association for Computational Linguistics: ACL 2025, 2025, pp. 6748-6769.

[40] S. M. Sajjadi Mohammadabadi, B. C. Kara, C. Eyupoglu, C. Uzay, M. S. Tosun, and O. Karakuş, "A survey of large language models: evolution, architectures, adaptation, benchmarking, applications, challenges, and societal implications," Electronics, vol. 14, no. 18, 2025, doi: 10.3390/electronics.

[41] V. Chaudhuri, A. Brunelli, P. Tcherveniakov, and N. Chaudhuri, "Benchmarking Large Language Models Using a Best Evidence Topic Report in a Patient With Early Non-Small Cell Lung Cancer," Interdisciplinary CardioVascular and Thoracic Surgery, vol. 41, no. 2, p. ivag038, 2026, doi: 10.1093/icvts/ivag038 Interdisciplinary.

[42] R. Rokhshad et al., "The ethics and governance of large language models in dentistry: A framework for research and clinical implementation," J Dent, vol. 164, p. 106187, Jan 2026, doi: 10.1016/j.jdent.2025.106187.

[43] S. Pahune, Z. Akhtar, V. Mandapati, and K. Siddique, "The Importance of AI Data Governance in Large Language Models," Big Data and Cognitive Computing, vol. 9, no. 6, 2025, doi: 10.3390/bdcc9060147.

[44] P. D. Stetson et al., "Responsible Artificial Intelligence governance in oncology," NPJ Digit Med, vol. 8, no. 1, p. 407, Jul 4 2025, doi: 10.1038/s41746-025-01794-w.

[45] J. Rezwana and M. L. Maher, "Designing creative AI partners with COFI: A framework for modeling interaction in human-AI co-creative systems," ACM Transactions on Computer-Human Interaction, vol. 30, no. 5, pp. 1-28, 2023.

[46] A. B. Hendrickson, K. C. Gilkey, N. Stickler, and T. Taylor, "Evolving Healthcare Leadership in the Age of AI: A Narrative Review of Current Expected Transitions of Leadership in the era of Artificial Intelligence," International Journal of Scientific and Management Research, vol. 8, no. 8, pp. 17-29, 2025.

[47] B. G. Collaco et al., "The role of agentic artificial intelligence in healthcare: a scoping review," npj Digital Medicine, 2026.

[48] M. J. Page et al., "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," BMJ, vol. 372, p. n71, Mar 29 2021, doi: 10.1136/bmj.n71.

[49] M. J. Page et al., "Updating guidance for reporting systematic reviews: development of the PRISMA 2020 statement," Journal of clinical epidemiology, vol. 134, pp. 103-112, 2021.

[50] B. Lim, I. Seth, M. Maxwell, R. Cuomo, R. J. Ross, and W. M. Rozen, "Evaluating the efficacy of large language models in generating medical documentation: a comparative study of ChatGPT-4, ChatGPT-4o, and Claude," Aesthetic Plastic Surgery, vol. 49, no. 20, pp. 5846-5857, 2025.

[51] M. Arslan, H. Ghanem, S. Munawar, and C. Cruz, "A Survey on RAG with LLMs," Procedia computer science, vol. 246, pp. 3781-3790, 2024.

[52] O. Shvetsova, D. Katalshov, and S.-K. Lee, "Innovative Guardrails for Generative AI: Designing an Intelligent Filter for Safe and Responsible LLM Deployment," Applied Sciences, vol. 15, no. 13, 2025, doi: 10.3390/app15137298.

[53] R. Dhawan, K. D. Brooks, O. Shauly, D. Shay, and A. Losken, "Ethical Considerations for Generative Artificial Intelligence in Plastic Surgery," Plast Reconstr Surg Glob Open, vol. 13, no. 6, p. e6825, Jun 2025, doi: 10.1097/GOX.0000000000006825.

[54] S. Bozyel et al., "Artificial intelligence-based clinical decision support systems in cardiovascular diseases," Anatolian journal of cardiology, vol. 28, no. 2, p. 74, 2024.

[55] L. Hughes, Y. K. Dwivedi, K. Li, M. Appanderanda, M. A. Al-Bashrawi, and I. Chae, "AI agents and agentic systems: Redefining global it management," vol. 28, ed: Taylor & Francis, 2025, pp. 175-185.

[56] S. Stamnes Karlsen, M. M. Yamin, E. Hashmi, B. Katt, and M. Ullah, "Securing large language models: A quantitative assurance framework approach," Journal of Information Security and Applications, vol. 97, 2026, doi: 10.1016/j.jisa.2025.104351.

[57] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, and Á. Fernández-Leal, "Human-in-the-loop machine learning: a state of the art," Artificial Intelligence Review, vol. 56, no. 4, pp. 3005-3054, 2023.

[58] B. Long, E. Liu, R. Qiu, and Y. Duan, "Explainable AI the latest advancements and new trends," arXiv preprint arXiv:2505.07005, 2025.

[59] M. Kirchhof, G. Kasneci, and E. Kasneci, "Position: Uncertainty quantification needs reassessment for large-language model agents," arXiv preprint arXiv:2505.22655, 2025.

APPENDIX

TABLE V. CHARACTERISTICS OF INCLUDED STUDIES

Author(s)	Main Technology	Study Type / Focus	Main Application	No
Andrew et al.	Llama 3 / Gemma / Mistral	Case Report	Extracting temporal relations from pediatric clinical reports to automate patient timelines for rare diseases.	[11]
Asai et al.	ChatGPT + Gemini	Original Article	Generating and validating patient education materials for pulmonary artery catheter insertion.	[26]
Baig et al.	GenAI + Wearables	Systematic review	Leveraging multimodal data for Personalized Patient Care Plans (PPCPs) and symbiotic intelligence.	[29]
Bozyel et al.	AI-based CDSS	Review / Perspective	Supporting cardiovascular disease diagnosis, risk assessment, and treatment optimization.	[54]
Chaudhuri et al.	GPT, Gemini, Co-pilot	Benchmarking	Evaluating LLM performance against "Best Evidence Topic" (BET) reports for early lung cancer decisions.	[41]
Chen et al.	LLM (Transformer)	Lifecycle survey	Mapping the evolution of language intelligence from word vectors to aligned multimodal agents.	[7]
Chow & Li	LLM (Transformer)	Review / Safety Analysis	Identifying opportunities and systemic risks in medical chatbots for patients and clinicians.	[35]
Comeau et al.	LLM + EHR	Governance framework	Ensuring transparency and accountability in EHR-based AI experimentation.	[8]
Dhawan et al.	GenAI + LLM	Ethical analysis	Exploring ethical considerations and administrative use cases specific to plastic surgery.	[53]
Frid et al.	LLM (Claude, Mistral, Titan)	Hackathon Analysis	Hospital-led prototyping of GenAI clinical tools under GDPR and EU AI Act compliance.	[19]
Goktas & Grzybowski	Symbolic AI + ML + Hybrid	Conceptual framework	Proposing a "Regulatory Genome" for trustworthy AI and adaptive oversight.	[32]
Guleria et al.	LLM / Generative AI	Executive survey	Analyzing perspectives and trends among US health system leaders on GenAI integration.	[17]
Gupta et al.	Multimodal LLM (MLLM)	Rapid Review	Mapping growing usage of multimodal data (imaging + text) in radiology, pathology, and ophthalmology.	[36]
Hughes et al.	AI Agents	Management Framework	To investigate the role of autonomous AI agents in advancing operational efficiency, intelligent decision-making, and collaborative IT management.	[55]
Jung et al.	Socially Assistive Robots	Conceptual Framework	User-centered design for assistive robots in mental healthcare utilizing multimodal evaluations.	[27]
Kawasaki	AI Screening	Implementation Study	Establishing a seven-step framework for effective diabetic retinopathy (DR) screening in Japan.	[22]
Khosravi et al.	LLM	Perspective / Survey	Identifying systemic challenges and diagnostic reasoning biases in clinical LLM adoption.	[9]
Krumsvik & Slettvoll	GPT-4 + OpenAI o3	Case Report / Rural	Implementing AI as a "virtual waiting room" for health empowerment in avalanche-isolated communities.	[28]
Li	AI + Machine Learning	Perspective / Education	Modernizing medical physiology education in China while addressing regional resource disparities.	[20]
Maity & Saikia	LLM (Various)	Systematic review	Comprehensive mapping of LLM capabilities in clinical decision support and medical education.	[18]
Mehandru et al.	LLM Agents + AI-SCE	Evaluation Framework	Transitioning assessment from static tests to high-fidelity "Objective Structured Clinical Examinations" (OSCE) for AI agents.	[15]
Mirzaei et al.	LLM + LDA Analysis	Thematic analysis	Synthesizing clinician voices on ethical implications, equity, and the patient-doctor dynamic.	[37]
Nazi & Peng	LLM + IGM	Comprehensive Review	Analyzing clinical documentation, diagnosis formulation, and evidence-based support systems.	[16]
Offodile et al.	GPT-4 + iLEAP	Institutional governance	Implementing the iLEAP lifecycle model for responsible AI deployment in oncology.	[44]
Ozmen & Mathur	RAG + LLM	Implementation study	Enhancing surgical clinical decision support by grounding responses in validated medical literature.	[13]
Pahune et al.	LLM Data Governance	Review / Framework	Establishing a five-pillar data governance model focusing on compliance and bias mitigation.	[43]
Painter et al.	LLM + Knowledge Graphs	Perspective Review	Exploring the potential for GenAI to automate data standardization and code conversion in healthcare.	[31]

Pal et al.	LLM	Review / Ethics	Evaluating plain language medical information for non-specialists and its impact on health literacy.	[25]
Pawelczyk et al.	AI + Digital Health	Commercial Review	Evaluating commercial AI solutions for musculoskeletal (MSK) triage, diagnostics, and surgical planning.	[30]
Pesapane et al.	GenAI + Radiomics	Narrative review	Discussing the "machines-in-the-loop" paradigm to enhance clinical utility in medical imaging.	[33]
Rainey et al.	LLM (Various)	Narrative Review	Automating radiography research stages, report structuring, and trial participant recruitment.	[21]
Rokhshad et al.	LLM	e-Delphi consensus	Creating an ethical governance framework for LLM implementation in clinical dentistry.	[42]
Sajjadi Mohammadabadi et al.	LLM Evolution	Survey	LLMs are increasingly adopted across diverse industries to support intelligent content generation, decision-making, knowledge management, and process automation.	[40]
Shen et al.	Multimodal Corpora	Framework / Corpus	Proposing "Disease-Scenario Association Matrices" as the cornerstone for healthcare vertical large models.	[14]
Shvetsova et al.	AVI Filter + RAG	Design / PoC Study	Implementing intelligent guardrails to filter toxicity and verify facts in responsible LLM deployment.	[52]
Stamnes Karlsen et al.	Quantitative Assurance	Methodology Study	Developing a framework to secure LLMs against prompt injection and toxic outputs in clinical apps.	[56]
Tsujimoto et al.	ChatGPT-4o	Virtual Patient Modeling	Identifying AI tendencies toward aggressive treatment in assisted decision-making for lymphedema care.	[24]
Tung et al.	GenAI	Systematic Review	Analyzing systemic bias, liability, and data privacy risks with proposed socio-technical solutions.	[10]
Wang et al.	Ambient AI + SCRIBE	Evaluation framework	Standardizing assessment of Ambient Digital Scribing (ADS) tools for clinical documentation.	[12]
Wójcik et al.	GPT-4	Invited Review	Mapping GPT-4's added value in clinical notes, triage, and cardiology consultations.	[6]
Workum et al.	LLM + ACUTE	Implementation roadmap	Providing a step-by-step approach for safe LLM implementation using the ACUTE mnemonic.	[23]

Localized Actionable Explainability for MOOC Dropout Intervention Using Clickstream Learning Analytics

Mohamed Amine OUASSIL[1], Mohammed JEBBEARI[2], Mouaad ERRAMI[3],
Rabia Rachidi[4], Soufiane HAMIDA[5], Bouchaib CHERRADI[6], Abdelhadi RAIHANI[7]
EEIS Laboratory, ENSET of Mohammedia, Hassan II University of Casablanca, Mohammedia 28830, Morocco[1, 2, 3, 6, 7]
LaROSERI Laboratory-Faculty of Science, Chouaib Doukkali University, El Jadida, Morocco[4]
2IACS Laboratory- ENSET of Mohammedia, Hassan II University of Casablanca, Mohammedia, Morocco[5]
STIE Team-CRMEF Casablanca-Settat, provincial section of El Jadida, El Jadida 24000, Morocco[2, 6]
GENIUS Laboratory, SupMTI of Rabat, Rabat, Morocco[5]

Abstract*—Massive Open Online Courses (MOOCs) have expanded access to learning, but learner dropout remains a persistent challenge in online education. Most dropout prediction studies transform learner traces into behavioral features and train machine learning or deep learning models to identify at-risk learners. Although this predictive layer is useful, it remains limited for educational decision-making because a risk score does not explain why a learner is at risk or what type of support should be proposed. This study introduces Localized Actionable Explainability (L-AXAI), a learning analytics framework that combines dropout prediction, local explanation, and personalized pedagogical nudging. The framework first transforms clickstream logs into weekly behavioral indicators, then trains predictive classifiers, and finally applies SHAP-based local explanation to identify the strongest risk driver for each high-risk learner. These local explanations are mapped to targeted pedagogical actions using a transparent rule-based translation matrix. Experiments on the KDD Cup MOOC dropout benchmark produced closely comparable results for the two strongest boosting models. Under the fixed stratified test split, Gradient Boosting achieved 85.79% accuracy, 86.57% ROC-AUC, and 94.41% PR-AUC, while XGBoost achieved 85.77% accuracy and the same ROC-AUC. Because repeated-seed evaluation and statistical significance testing were not conducted, the small difference between these point estimates should not be interpreted as definitive model superiority. The explainability analysis identifies model-associated patterns involving weekly activity intensity, active days, and early problem engagement. The intervention layer generates candidate nudge categories, mainly navigation, re-engagement, and session-completion suggestions. The framework therefore returns an estimated dropout risk, a local risk driver, and a candidate pedagogical action. The nudges are intended as low-intensity decision-support suggestions; their appropriateness, perceived pressure, and effectiveness require validation with instructors and learners.

Keywords*—MOOC dropout prediction; learning analytics; explainable artificial intelligence; SHAP; localized explanation; actionable intervention

I. Introduction

Online learning platforms have become one of the main sources of data for education. Every learner activity such as opening a course page, viewing a video, trying a problem, or posting in a discussion forum leaves a trace that can be analyzed to better understand the engagement and learning progression [1], [2], [3]. These traces are used by learning analytics and educational data mining to support teachers, designers, and institutions in tracking learning processes and improving educational decisions [4], [5], [6], [7], [8]. MOOCs are a prime example of data-rich learning environments, with large-scale enrollment and continuous clickstream records. But the very openness that makes MOOCs appealing also makes it hard to persist. Learners enter MOOCs with different goals, different time constraints, and different levels of commitment. Thus, completion rates cannot fully account for disengagement, as learners may follow several trajectories rather than a single common path [9], [10]. Consequently, predicting dropout has become a central task in learning analytics. Previous work has shown that activity indicators, weekly engagement signals, and temporal representations can be used to identify learners who are likely to stop participating [11], [12], [13], [14]. Later work also advised that predictive performance should be carefully evaluated because training and testing on the same course may overestimate the real accuracy of a dropout model [15], [16]. Predictive models are useful, but prediction is not sufficient for teaching. A model may identify a learner as high-risk, but this information is weak if the instructor does not know which behavior created the risk. A learner who doesn't try problems needs a different type of support than a learner who watches videos but never uses discussions. Likewise, a learner with a long inactivity gap needs a different nudge than a learner who browses through many pages without completing learning tasks [17], [18], [19], [20]. Explainable Artificial Intelligence (XAI) helps to open this black-box problem by showing which features affect a prediction. LIME and SHAP are popular methods for explaining machine learning models, but counterfactual explanations offer an action-oriented perspective by asking what changes are needed to obtain a different outcome [21], [22], [23], [24], [25], [26]. However, many educational applications of XAI are confined to global feature importance. A global plot can show that problem events are important in general, but it cannot automatically tell what to do for a specific learner [27], [28], [29], [30], [31]. Thus, the

real challenge is not only to predict dropout, but also to associate the prediction with localized and educationally meaningful action. This study introduces Localized Actionable Explainability (L-AXAI), a framework that identifies at-risk learners, explains each individual prediction, and converts the top local risk driver into a pedagogical nudge. The framework is evaluated on a public MOOC clickstream benchmark for dropout prediction.

The main contributions of this work are fivefold. First, it proposes a localized actionable explainability framework for MOOC dropout intervention. Second, it extracts directly interpretable weekly behavioral features from clickstream events, including access, problem-solving, video engagement, discussion participation, navigation, wiki usage, and inactivity. Third, it formalizes the predictive and explainability layers through supervised classification, tree-based boosting, SHAP values, and an intervention mapping function. Fourth, it introduces a transparent rule-based translation matrix that converts local SHAP risk drivers into pedagogical nudges. Finally, it evaluates the framework not only in terms of predictive performance, but also through global and local explanation analysis and the distribution of generated intervention types.

The rest of this study is organized as follows: Section II reviews the related literature. The mathematical formulation and the proposed L-AXAI framework architecture are described in Section III. Results are given in Section IV. Discussion is presented in Section V. Finally, Section VI concludes this study.

II. Literature Review

Over the past decade, MOOC dropout prediction has progressed from simple activity-count models to temporal, representation-based, and explainable learning analytics frameworks [11], [12], [13], [14], [32]. Existing studies mainly focus on improving early detection, generalization across courses, or predictive performance [15], [16], [33], [34]. More recent work has started to consider interpretability, but many studies still stop at feature importance, SHAP plots, or visual explanations without converting the explanation into a direct learner-level pedagogical action [17], [19], [20], [27], [35]. This section reviews activity-based, temporal, deep representation, explainable, and action-oriented dropout prediction studies, highlighting key contributions and limitations.

Early dropout prediction studies were mainly based on learner activity features extracted from platform logs. Halawa et al. [11] used activity indicators to red-flag students who were likely to leave a course, showing that learner interaction traces can support early warning. Similarly, Kloft et al. [12] predicted dropout over weeks using machine learning models and showed that weekly behavior can provide useful signals for early detection. Taylor et al. [13] developed a scalable dropout prediction pipeline and demonstrated that temporal and non-temporal features can produce strong AUC values. However, these approaches rely heavily on engineered counts and feature importance analysis, which limits their ability to explain the specific barrier faced by each learner.

Several studies then focused on validation, transferability, and robustness. Boyer and Veeramachaneni [36] explored transfer learning for predictive models across MOOC contexts, while Whitehill et al. [15] showed that same-course evaluation can give overly optimistic estimates of dropout prediction accuracy. Gardner and Brooks [16] also argued that dropout model evaluation needs more rigorous procedures because feature extraction methods and algorithms interact strongly. These works are important because they show that high accuracy alone is not enough. Nevertheless, their main concern is predictive validity rather than pedagogical actionability.

Temporal and representation learning methods were proposed to capture changes in learner behavior more deeply. Fei and Yeung [14] introduced temporal models for dropout prediction, moving beyond static activity totals. Jeon et al. [37] proposed interpretable multi-layer representation learning to reduce manual feature engineering, and Jeon and Park [38] modeled clicks and videos jointly to improve weekly dropout prediction. These studies improve the modeling of sequential engagement and learning resources. However, their outputs still mainly support classification, and the connection between learned representations and instructor-level intervention remains limited.

Other studies investigated behavior patterns and learner trajectories. Kizilcec et al. [9] showed that MOOC disengagement is not one single behavior but a set of learner subpopulations with different engagement trajectories. Feng et al. [18] analyzed dropout behaviors and identified meaningful categories of learners in large-scale MOOCs. Alamri et al. [19] used visually driven, multi-granularity explanations to compare learning paths of completers and non-completers, showing that non-completers may jump forward or follow irregular paths. These studies are closer to pedagogy because they describe behavior, but they do not automatically generate a concrete nudge for each high-risk learner.

Explainability has been introduced more directly in dropout prediction pipelines. Nagrecha et al. [17] showed that MOOC dropout prediction pipelines can be made interpretable with performance close to stronger black-box methods. Er [20] used explainable machine learning to understand dropout in two MOOC runs and discussed feature importance from a pedagogical point of view. Swamy et al. [27] compared several explanation methods in student performance prediction, including LIME, SHAP, DiCE, and contrastive explanations, and showed that explanation methods can disagree. These studies show the relevance of XAI in education, but explanation is still often presented as a separate analysis after prediction rather than as an intervention engine.

Recent work also moved toward SHAP-based and intervention-oriented systems. Liu [39] used a dilated convolutional behavior model and SHAP analysis to identify important learning behavior variables. Nti and Ramanayake [35] proposed an explainable dropout prediction framework that also considers tailored interventions in online personalized education. Ajayi [40] combined stacked ensemble learning with explainability for online dropout risk, and Nagy et al. [41] connected interpretable dropout prediction with personalized intervention. These studies are important because they

explicitly connect interpretability with student support. However, most of them do not define a transparent event-to-nudge translation matrix based on local clickstream deficits.

Counterfactual and group-level explanation methods provide another direction for actionable learning analytics. Wachter et al. [24] formalized counterfactual explanations as a way to explain automated decisions without opening the black box, while Mothilal et al. [25] proposed diverse counterfactual explanations that are useful when several possible changes may lead to a desired outcome. Buñay-Guisñan et al. [42] applied group counterfactual explanations to support students at risk of dropout, aiming to recover larger groups with lower intervention cost. These works are highly relevant for future learning analytics because they bring actionability into explanation. Still, direct mapping from local MOOC event deficits to practical pedagogical nudges remains underdeveloped.

The literature shows a clear evolution in the field of MOOC dropout prediction, from activity-count models to more complex temporal, representation-based, and more explainable models. However, many studies still focus only on prediction accuracy or visual explanation. One remaining gap is how to translate local learner-level explanations into transparent and practical pedagogical nudges that are personalized and can be directly used by instructors.

III. Materials and Methods

A. Proposed Method

The study is based on quantitative learning analytics, utilizing supervised machine learning and post-hoc explainability. The goal is not only to predict at-risk learners, but also to explain why each at-risk prediction is produced and to transform this explanation into a concrete pedagogical action. The proposed L-AXAI workflow presented in Fig. 1 begins from raw MOOC clickstream logs, which record learner interaction events such as login, page views, video access, problem attempts, discussion activity, navigation, and session closing. During preprocessing, irrelevant events are removed, timestamps are converted into a common format, and actions are aggregated into learner-course sessions. Missing or invalid records are also removed. The dataset is split into training, validation, and testing sets after pre-processing. The behavioral feature engineering stage converts the raw logs into weekly learner features. Such features are time indicators, weekly event counts, and deficit indicators. Event counts indicate what the learner did, temporal indicators show the pace of participation, and deficit indicators show what is missing when comparing the learner's behavior to the reference behavior of successful learners in the same course. In the prediction stage, the following machine-learning models are trained and compared: Logistic Regression, SVM, KNN, Decision Tree, Random Forest, Extra Trees, AdaBoost, Gradient Boosting, XGBoost, and LightGBM. Metrics for classification, like accuracy, ROC-AUC, and PR-AUC, are used to evaluate candidate models. The selected model is then used to produce an estimated dropout-risk score for each learner. The framework utilizes local SHAP explanation for learners having a predicted dropout probability above the risk threshold. This step identifies the features that exert the greatest influence on each individual prediction. Rather than a global model interpretation, L-AXAI is concerned with the local risk driver of each learner. Finally, the rule-based translation matrix maps the selected risk driver to a pedagogical action, such as a practice-oriented nudge, re-engagement message, video review recommendation, or help-seeking prompt. This design enables transparent and responsible learning analytics by connecting prediction, explanation, and intervention into one workflow. Fig. 1 illustrates the proposed L-AXAI workflow for MOOC dropout prediction and intervention.

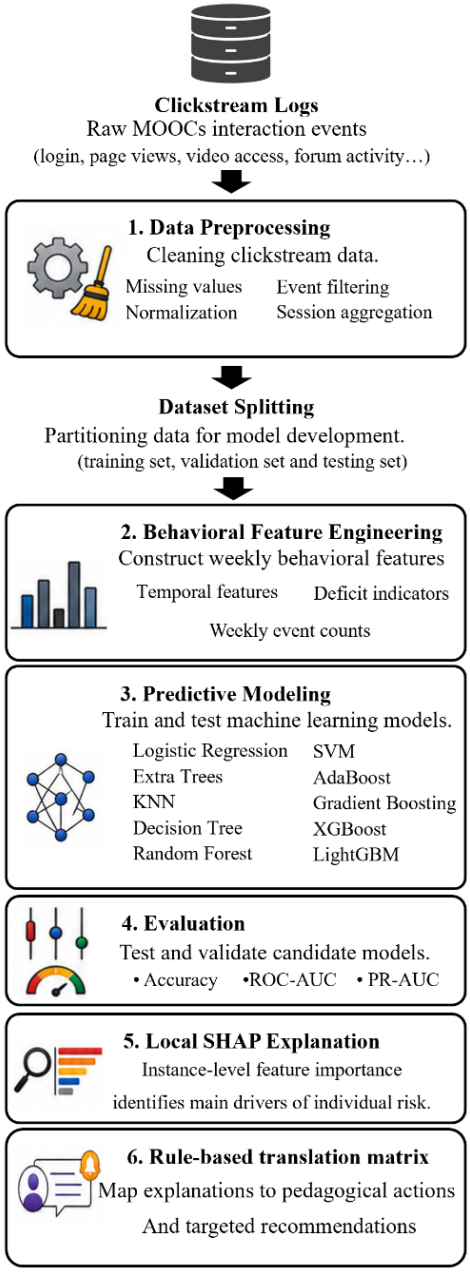


Fig. 1. Proposed L-AXAI workflow for MOOC dropout prediction and intervention.

B. Data Source and Event Representation

The experimental data are from the KDD Cup 2015 MOOC dropout prediction benchmark. The dataset contains learner-

course enrollment records and timestamped activity logs from a large MOOC platform. Each activity record contains an enrollment ID, a timestamp, a source object, and an event type. The task is formulated as a binary classification problem to predict if a learner-course enrollment will become inactive in the target period. The framework describes learner behavior through seven types of clickstream events: access, discussion, navigate, page_close, problem, video, and wiki. Access events indicate general connection to the course. Video events indicate consumption of instructional content, problem events indicate practice and assessment engagement, discussion events indicate social and help-seeking interactions, and wiki events indicate consultation of supplementary materials. Navigate and page_close events provide complementary information about movement between course objects and ending a session. Table I presents the interpretation of MOOC clickstream event types.

TABLE I. INTERPRETATION OF MOOC CLICKSTREAM EVENT TYPES

Event type	Behavioral meaning	Pedagogical interpretation
access	Opening or accessing course resources	General connection with the course
video	Interaction with course videos	Exposure to explanations and demonstrations
problem	Attempts or interaction with problems	Practice, self-testing, and assessment engagement
discussion	Forum or discussion activity	Help-seeking, peer interaction, and social learning
wiki	Consultation of wiki resources	Use of additional knowledge materials
navigate	Movement between course objects	Learning path and course exploration
page_close	Closing/leaving a course page	Session ending and possible early exit behavior

C. Preprocessing and Feature Engineering

The preprocessing step begins with a merge of the enrollment table and the log table on the enrollment identifier. Remove records that do not have key identifiers. Then each timestamp of learner action is transformed into a common datetime format and sorted chronologically. In this study, weekly observation windows are used instead of representing learner behavior with one global activity vector. Each week consists of seven days of activity. Thus, the actions of the learner are represented separately for week 1, week 2, week 3, and week 4. This design allows the model to learn the time variation of dropout risk. From the raw logs, three groups of features are derived.

The first type is weekly event counts including access, video, problem, discussion, wiki, navigate, and page_close events. For example, problem_w2_count is the number of problem-related actions taken by the learner in Week 2. This variable is coded as Week 2 problem activity in the analysis.

The second group includes temporal features on a weekly basis. Such features characterize the rhythm of the learner's participation. For example, active_days_w2 is the number of active days in Week 2, and maximum_daily_events_w2 is the maximum number of actions recorded in one day in Week 2. This variable is interpreted as Week 2's highest daily activity.

The third group is made up of weekly deficit indicators. A deficit is the difference between the learner's activity and the median activity of non-dropout learners in the same course. problem_w2_deficit shows the distance from the course reference for the learner's problem activity in Week 2. If the learner is not below the reference, then the deficit is set to zero. All course reference medians used for deficit computation were estimated only from the training set. For validation and test records, these training references were reused without accessing validation or test labels.

Event counts are what the learner did, temporal features are when and how often the learner participated, and deficit indicators are what is missing compared to successful learners. These deficit measures support the intervention layer by allowing a local risk driver to be translated into a candidate pedagogical nudge.

D. Machine Learning and Explainability Methods

The prediction task is formulated as a supervised binary classification problem. Let the processed dataset be represented as:

$$\mathcal{D} = \{(\mathrm{x}_i, y_i)\}_{i=1}^{N}, \quad \mathrm{x}_i \in \mathbb{R}^d, \quad y_i \in \{0,1\} \tag{1}$$

where, x_i is the feature vector of the learner-course record i, d is the number of engineered features, and y_i is the dropout label. The classifier learns a function that maps the behavioral representation of a learner to a dropout probability:

$$f: \mathbb{R}^d \to [0,1], \quad \hat{p}_i = f(\mathrm{x}_i) \tag{2}$$

where, $\hat{p}_i$ is the predicted probability that learner i will drop out. A binary decision is obtained by applying a risk threshold. τ:

$$\hat{y}_i = \begin{cases} 1, & \text{if } \hat{p}_i \geq \tau \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

In this study, the risk threshold τ was set to 0.5, meaning that learner-course records with a predicted dropout probability greater than or equal to 0.5 were classified as high-risk.

Logistic Regression is used as a simple baseline because it gives a transparent relationship between features and predicted probability. [15], [43]. It estimates dropout probability using the sigmoid function:

$$\hat{p}_i = \sigma(\beta_0 + \mathrm{x}_i^{\mathrm{T}}\beta), \quad \sigma(z) = \frac{1}{1+e^{-z}} \tag{4}$$

The Support Vector Machine (SVM) is included as a margin-based classifier. It separates the two classes by finding a decision boundary with a maximum margin. The K-Nearest Neighbors (KNN) classifier is used as a non-parametric baseline. It predicts the label of a learner according to the labels of the most similar learners in the training set:

$$\hat{y}_i = \mathrm{mode}\{y_l : \mathrm{x}_l \in \mathcal{N}_k(\mathrm{x}_i)\} \tag{5}$$

where $\mathcal{N}_k(\mathrm{x}_i)$ denotes the set of the k nearest training samples to x_i. Decision Tree is used as a single-tree baseline, while Random Forest and Extra Trees are used as randomized tree ensembles. These models combine several decision trees

and aggregate their predictions, commonly through majority voting [44]:

$$\hat{y}_i = \text{mode}\{h_t(\text{x}_i)\}_{t=1}^{T} \quad (6)$$

where, h_t is the prediction of tree t, and T is the total number of trees in the ensemble. Random Forest builds trees from bootstrap samples, while Extra Trees increases randomness by using more randomized split selection. Both strategies help reduce variance compared with a single decision tree. The remaining classifiers belong to the boosting family: AdaBoost, Gradient Boosting, XGBoost, and LightGBM. Boosting models build the final predictor additively by adding new weak learners that correct the errors of previous learners [33], [34]:

$$F_m(\text{x}_i) = F_{m-1}(\text{x}_i) + \eta h_m(\text{x}_i) \quad (7)$$

where, h_m is the weak learner added at iteration m, and η is the learning rate. XGBoost and LightGBM are included because they are efficient for structured tabular features and include regularization mechanisms to control overfitting. A general regularized boosting objective can be written as:

$$\mathcal{L}^{(m)} = \sum_{i=1}^{N} l\left(y_i, F_{m-1}(\text{x}_i) + \eta h_m(\text{x}_i)\right) + \Omega(h_m) \quad (8)$$

where, l is the predictive loss function and $\Omega(h_m)$ penalizes model complexity. After model training, SHAP is used to explain model outputs. SHAP is based on Shapley values and belongs to additive feature attribution methods [21], [23], [45]. For a learner i, the prediction is decomposed as:

$$g(\text{x}_i) = \phi_0 + \sum_{j=1}^{d} \phi_{i,j} \quad (9)$$

where, ϕ_0 is the expected model output and $\phi_{i,j}$ is the contribution of feature j to the prediction of learner i. The Shapley value of feature j is defined as:

$$\phi_{i,j} = \sum_{S \subseteq F_j} w(S)[f(S \cup j) - f(S)] \quad (10)$$

where, F_j denotes the set of all features excluding feature j. The weighting term is defined as:

$$w(S) = \frac{|S|!\,(d - |S| - 1)!}{d!} \quad (11)$$

In this expression, S is a subset of features that excludes feature j, and the term in brackets measures the marginal change caused by adding feature j. SHAP is suitable for this study because it provides both global and local explanations. Global explanations show which variables matter across the cohort, while local explanations show why a specific learner is classified as high-risk. In L-AXAI, only learners whose predicted dropout probability exceeds the threshold are passed to the action-generation step. For each high-risk learner, the framework selects the strongest positive actionable SHAP contribution:

$$r_i = \underset{j \in A}{\text{argmax}} \phi_{i,j}, \quad \phi_{i,j} > 0 \quad (12)$$

where, A is the set of actionable weekly features and r_i is the local risk driver that most strongly pushes the learner i toward the dropout class. The selected driver is then mapped to a pedagogical action using the rule-based translation matrix T:

$$a_i = T(r_i) \quad (13)$$

The final output of the proposed framework is therefore not only a predicted label. For each high-risk learner, L-AXAI returns a triplet composed of the estimated dropout-risk score, the local explanation, and a candidate pedagogical nudge template:

$$o_i = (\hat{p}_i, r_i, a_i) \quad (14)$$

This formulation shows the difference between standard predictive modeling and L-AXAI. A standard classifier stops at the predicted probability or class label. In contrast, L-AXAI links prediction to explanation and then to intervention, which supports localized and actionable learning analytics. The estimated score is used as an operational decision value under the threshold τ=0.5. Because probability calibration was not evaluated, it should not be interpreted as a validated empirical probability of dropout.

E. L-AXAI Rule-Based Translation Matrix

The mappings in Table II were constructed as an author-designed, literature-informed, and behavior-semantics-informed prototype rather than as the outcome of a formal expert-consensus process. Each rule links the behavioral meaning of the corresponding event or deficit to a support objective: practice for low problem activity, content review for low video activity, reconnection for low access, help-seeking for low discussion participation, regularity support for inactivity, diversified resource use for low event diversity, progression support for navigation deficits, and session-completion support for frequent page closing with few active days. The matrix is deliberately transparent and editable and should therefore be interpreted as a set of candidate nudge templates rather than as validated pedagogical prescriptions. Instructor review and learner-centered evaluation are required before operational deployment.

TABLE II. MAPPING LOCAL SHAP RISK DRIVERS TO PROPOSED PEDAGOGICAL NUDGE TEMPLATES

Top local risk driver	Interpretation	Candidate pedagogical nudge
Low problem_count	The learner is not practicing enough	Try to complete one or two practice problems from this week. This will help you check your understanding before moving forward.
Low video_count	The learner may not be following explanations	Return to the main video of this unit and watch the key explanation before attempting the next activity.
Low access_count	The learner is disconnected from the course	You have not accessed the course regularly. Try to log in today and review the current unit for a few minutes.
Low discussion_count	The learner is not using peer or instructor support	Visit the discussion forum and read one question related to this unit. You can also post a short question if something is unclear.
High inactivity_ga	The learner has stopped regular	You had several days without activity. Restart with one small task

Top local risk driver	Interpretation	Candidate pedagogical nudge
p	participation	today rather than waiting until the end of the week.
Low event_diversity	The learner uses only one type of resource	Try to combine watching the video with solving one problem or reading one discussion thread.
Low navigate_count / high navigate_deficit	The learner is not progressing through course objects	Open the next course object and complete one small activity before moving to other pages
High page_close_count with low active_days	The learner may leave sessions quickly	Try to spend a little more time in one session and finish one complete activity before closing the page.

Algorithm 1 summarizes the main steps of the proposed Weekly L-AXAI framework. It starts by preparing clickstream data and converting learner actions into weekly behavioral features. Then, the predictive model estimates dropout probability for each learner and selects only learners above the risk threshold. For these high-risk learners, local SHAP values are computed to identify the main actionable risk driver, which is finally translated into a pedagogical nudge for the instructor dashboard.

Algorithm 1: Weekly L-AXAI for MOOC Dropout Intervention

Input: Enrollment table E, log table L, labels Y, number of weekly windows K, risk threshold τ, predictive model f, and matrix T.

Merge E and L using the enrollment identifier.

Convert timestamps to datetime format and order learner actions chronologically

For each learner-course enrollment i **do**

  For each weekly window k = 1 to K **do**

    Select learner actions inside the weekly window W_k.

    Count access, video, problem, discussion, wiki, navigate, and page_close events.

    Compute active days, average daily events, maximum daily events, inactivity days, and event diversity

    Store weekly features

  End

End

Split the processed dataset into training, validation, and test subsets.

Compute course-level reference medians using only non-dropout learners from the training set

For each learner i **do**

  Compute weekly deficit indicators by comparing x_i with the course reference

End

Train the predictive model f on the training set.

For each learner i in the test set **do**

  Estimate dropout probability $\hat{p}_i = f(x_i)$.

  If $(\hat{p}_i \geq \tau)$ **then**

    Compute local SHAP values φ_{ij} for learner i.

    Identify the strongest positive actionable risk driver r_i.

    Generate the pedagogical action $a_i = T(r_i)$.

      Add $(i, \hat{p}_i, r_i, a_i)$ to the intervention table.

  End

End

Return the intervention table.

IV. RESULTS

The processed dataset was divided into 70% training, 15% validation, and 15% testing using stratified random sampling with random seed 42. Stratification preserved the dropout/non-dropout distribution across the three subsets. All data-dependent operations, including course-reference medians and correlation-based feature filtering, were estimated from the training subset and then applied unchanged to validation and test data. The validation subset was used for hyperparameter selection, and the test subset was reserved for final evaluation. Experiments were performed in Google Colaboratory on an NVIDIA T4 GPU with 54 GB RAM. Gradient Boosting used 250 estimators, a learning rate of 0.05, and a maximum tree depth of 3. XGBoost used 500 estimators, a learning rate of 0.05, a maximum depth of 5, a subsampling ratio of 0.85, and a column-sampling ratio of 0.85. The operational risk threshold was fixed at $\tau = 0.5$. Parameters not explicitly modified retained their corresponding library defaults. The experiments were implemented in Python using scikit-learn, XGBoost, LightGBM, and SHAP. Parameters not explicitly modified retained their corresponding library defaults.

TABLE III. COMPARATIVE RESULTS OF DROPOUT PREDICTION MODELS

Model	Number of Features	Accuracy (%)	ROC-AUC (%)	PR-AUC (%)
Gradient Boosting	72	85.79	86.57	94.41
XGBoost	72	85.77	86.57	94.45
LightGBM	72	81.28	86.34	94.28
Logistic Regression	72	80.89	86.27	94.35
SVM	72	85.1	86.18	94.33
AdaBoost	72	84.82	86.14	94.28
Extra Trees	72	83.68	86.11	94.07
Random Forest	72	85.26	85.21	93.33
Decision Tree	72	80.78	84.33	92.92
KNN	72	84.97	84.09	93.03

Table III presents the performance of the tested machine learning models using the final weekly feature set. Gradient Boosting achieved the highest accuracy point estimate, while

XGBoost produced closely comparable performance, with 85.79% accuracy, 86.57% ROC-AUC, and 94.41% PR-AUC. XGBoost obtains almost the same performance, with 86.57% ROC-AUC. The difference between the two boosting models is small, but both are above simple baselines such as Decision Tree and KNN. Fig. 2 shows the accuracy comparison for the proposed models in this study.

Fig. 2 presents the accuracy comparison of the tested classifiers using the weekly feature representation. Under the original fixed split, Gradient Boosting achieved the highest accuracy point estimate at 85.79%, while XGBoost achieved 85.77% and the same ROC-AUC of 86.57%. The 0.02 percentage-point difference is too small to support a claim of definitive superiority without repeated experiments and statistical testing. The two boosting models are therefore interpreted as closely comparable under the present evaluation protocol. Gradient Boosting is retained as the predictive backbone used for the subsequent L-AXAI explanation stage, while the main contribution of the study remains the connection between estimated risk, local explanation, and candidate pedagogical action rather than a claim of universal classifier superiority.

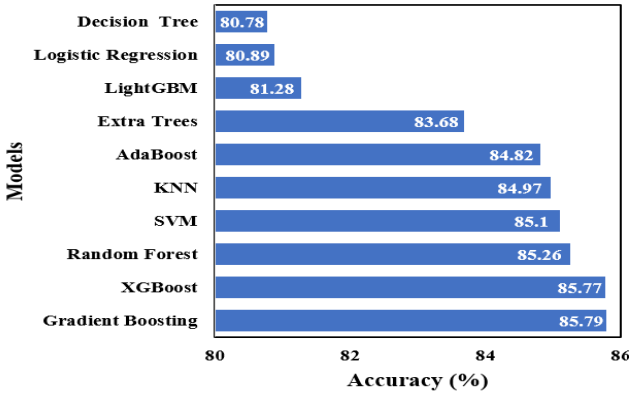


Fig. 2. Accuracy comparison of the tested classifiers.

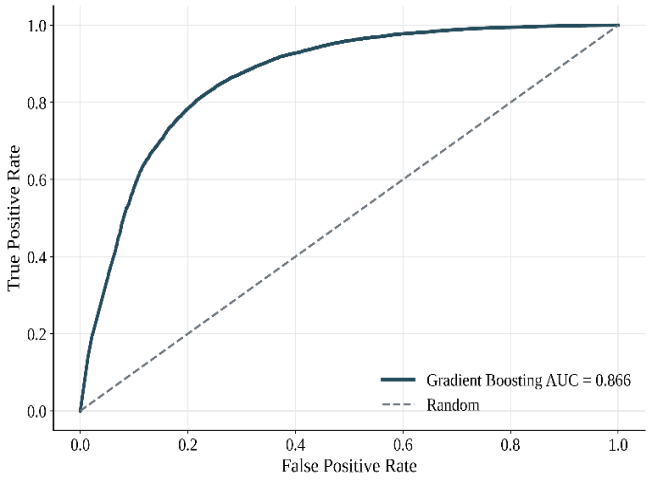


Fig. 3. ROC curve of the selected weekly model.

As shown in Fig. 3, the ROC curve confirms good discrimination between dropout and non-dropout learners. However, because the dataset is imbalanced, the precision-recall curve is more informative for the practical use of the model. The PR-AUC is high for the best models, showing that the system maintains useful precision even when the positive dropout class dominates the data.

A. Weekly Observation Horizon Analysis

The weekly horizon comparison evaluates whether useful dropout signals appear early or only after several weeks of activity. The results in Table IV show a clear improvement as the observation window grows from Week 1 to Week 4. Week 1 alone already provides useful predictive information, but the full four-week representation obtains the highest accuracy and ROC-AUC. These results show that extending the observation horizon from Week 1 to Week 4 provides additional predictive information under the current evaluation protocol.

TABLE IV. BEST MODEL FOR EACH WEEKLY OBSERVATION HORIZON

Horizon	Model	Number of Features	Accuracy (%)	ROC-AUC (%)	PR-AUC (%)
Week 1	XGBoost	20	81.28	80.33	91.95
Week 1-2	Gradient Boosting	39	83.46	84.22	93.56
Week 1-3	Gradient Boosting	58	85.26	85.92	94.17
Week 1-4	Gradient Boosting	72	85.79	86.57	94.41

B. Feature Filtering

Table V shows the highly correlated variables removed before final modelling. The weekly representation increases the number of variables because event counts, temporal indicators, and deficit variables are repeated over weeks. To reduce feature bloat, highly correlated variables were removed using a correlation threshold of 0.95. As expected, inactivity_days variables were perfectly correlated with active days variables and were removed from the final model. Several Week 4 deficit variables were also removed because they were almost redundant with access_w4_deficit.

TABLE V. HIGHLY CORRELATED VARIABLES REMOVED BEFORE FINAL MODELING

Dropped feature	Correlated with	Absolute correlation	Threshold
inactivity_days_w1	active_days_w1	1	0.95
inactivity_days_w2	active_days_w2	1	0.95
event_diversity_w2_deficit	navigate_w2_deficit	0.9548	0.95
inactivity_days_w3	active_days_w3	1	0.95
event_diversity_w3_deficit	access_w3_deficit	0.9569	0.95
inactivity_days_w4	active_days_w4	1	0.95
video_w4_deficit	access_w4_deficit	0.9751	0.95
problem_w4_deficit	access_w4_deficit	0.986	0.95
navigate_w4_deficit	access_w4_deficit	0.9844	0.95
page_close_w4_deficit	access_w4_deficit	0.9834	0.95
active_days_w4_deficit	access_w4_deficit	0.9821	0.95
event_diversity_w4_deficit	access_w4_deficit	0.9938	0.95

C. Explainability Analysis

The explainability analysis is structured on two levels. First, global SHAP analysis identifies the weekly variables that have the greatest influence on the model across the test cohort. Second, local SHAP analysis provides explanation for individual learner cases with waterfall plots and interaction plots. This setup is similar to the logic in explainability-oriented papers where global patterns and local case studies are explained at the same time.

1) Global SHAP importance: The global SHAP ranking in Table VI and Fig. 4 show that temporal intensity variables dominate the prediction. The most influential variable is "average daily events in week 3", followed by "average daily events in week 2", "active days in week 1", "average daily events in week 1", and "problem in week 1 count". This means that the model does not rely only on whether a learner logs in, but also on the intensity and regularity of weekly engagement.

TABLE VI. TOP 10 GLOBAL SHAP WEEKLY RISK DRIVERS

Feature	Mean absolute SHAP	Mean SHAP	Feature mean	Feature std
average_daily_events_w3	0.2681	-0.0274	1.265	3.794
average_daily_events_w2	0.2199	-0.041	1.455	4.121
active_days_w1	0.1953	-0.0243	1.637	1.099
average_daily_events_w1	0.182	-0.0464	5.05	6.596
problem_w1_count	0.1289	-0.0063	3.797	10.716
average_daily_events_w4	0.1287	0.0279	0.814	3.601
active_days_w4	0.0559	0.0034	0.22	0.765
access_w1_count	0.0527	-0.0029	13.308	18.743
active_days_w3	0.0404	0.0025	0.358	0.875
navigate_w2_deficit	0.0392	0.0119	0.743	0.437

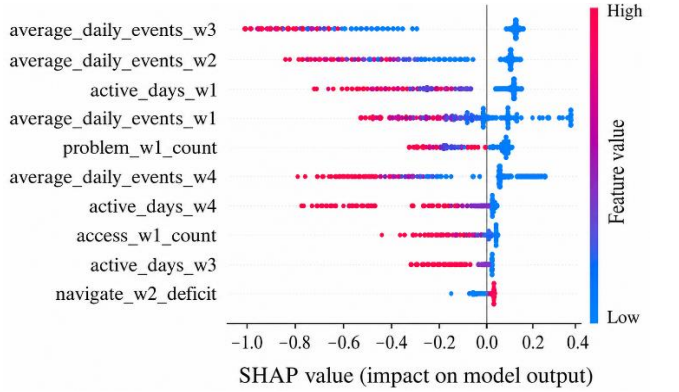


Fig. 4. Global SHAP summary plot for the selected weekly model.

2) Temporal Global SHAP evolution: In Table VII, the temporal SHAP evolution analysis shows how the type of predictive signal changes across weeks. In Week 1, temporal variables contribute 57.12% of the weekly SHAP mass, event counts contribute 35.23%, and deficit variables contribute 7.66%. In Weeks 3 and 4, temporal features become even more dominant, reaching 81.29% and 77.42%, respectively. Fig. 5 illustrates temporal SHAP evolution by weekly feature group.

TABLE VII. TEMPORAL SHAP CONTRIBUTION BY WEEKLY FEATURE GROUP

Feature type	Week	Mean absolute SHAP	Relative contribution (%)
Deficit	Week 1	0.0558	7.66
	Week 2	0.0866	19.34
	Week 3	0.0063	1.48
	Week 4	0.0008	0.29
Event count	Week 1	0.2565	35.23
	Week 2	0.0853	19.05
	Week 3	0.0731	17.24
	Week 4	0.064	22.29
Temporal	Week 1	0.4158	57.12
	Week 2	0.2758	61.6
	Week 3	0.3449	81.29
	Week 4	0.2224	77.42

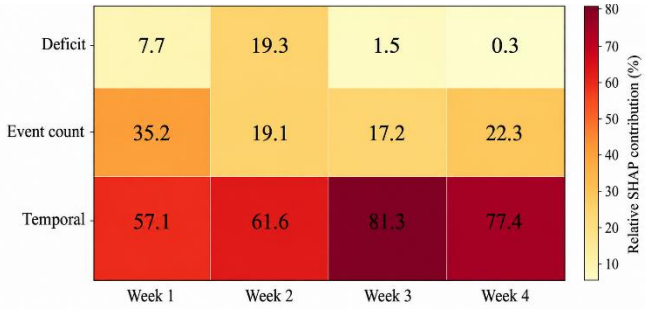


Fig. 5. Temporal SHAP evolution by weekly feature group.

3) Local SHAP waterfall analysis: Tables VIII and IX show four local cases explained with SHAP waterfall analysis. Cases A and B were both predicted as high-risk with a dropout probability of 95.98%. Cases A and B are behavioral twins selected to illustrate that identical clickstream patterns may lead to different final outcomes. Their main risk drivers are low activity in the first weeks, especially "average daily events in Week 1", "average daily events in Week 2", "average daily events in Week 3", and "problem Week 1 count". This means that early low engagement strongly pushed the model toward dropout.

Because Cases A and B have the same engineered clickstream vector, the deterministic classifier assigns them the same predicted probability and the same SHAP decomposition. Their different observed outcomes do not indicate an inconsistency in the model calculation. Instead, they show that the current feature representation is incomplete. Learner goals, prior knowledge, motivation, external time constraints, offline study, and interactions not recorded by the platform may influence persistence even when the observed clickstream behavior is identical. The pair is therefore treated as an

illustration of residual uncertainty and unobserved heterogeneity.

Cases C and D were predicted as low-risk. Their negative SHAP values show that higher weekly activity reduced the predicted dropout probability, especially high "average daily events in week 3", "average daily events in week 4", and several active days in Weeks 2 and 4. Case C is a false negative, which suggests that some dropout decisions may be caused by factors not visible in clickstream logs.

TABLE VIII. LOCAL CASE PROFILES SELECTED FOR SHAP WATERFALL ANALYSIS

Case	Case description	Drop out	Predicted dropout probability	Prediction
A	High-risk dropout	1	95.98	1
B	High-risk non-dropout	0	95.98	1
C	Low-risk dropout	1	2.46	0
D	Low-risk non-dropout	0	1.52	0

TABLE IX. TOP FIVE LOCAL SHAP CONTRIBUTIONS FOR SELECTED LEARNER CASES

Case	Rank	Feature	Feature value	SHAP value
A	1	average_daily_events_w1	0.143	0.3998
	2	average_daily_events_w3	0	0.1517
	3	active_days_w1	1	0.1378
	4	average_daily_events_w2	0	0.1245
	5	problem_w1_count	0	0.1089
B	1	average_daily_events_w1	0.143	0.3998
	2	average_daily_events_w3	0	0.1517
	3	active_days_w1	1	0.1378
	4	average_daily_events_w2	0	0.1245
	5	problem_w1_count	0	0.1089
C	1	average_daily_events_w3	16.143	-0.8571
	2	active_days_w4	5	-0.7091
	3	average_daily_events_w4	16.429	-0.5911
	4	average_daily_events_w2	16.857	-0.4854
	5	active_days_w2	6	-0.4313
D	1	average_daily_events_w3	27.714	-0.9057
	2	active_days_w4	5	-0.702
	3	average_daily_events_w4	21.429	-0.5737
	4	average_daily_events_w2	7	-0.4725
	5	active_days_w2	6	-0.438

Fig. 6-(a) and Fig. 6-(b) show that the high-risk cases are mainly explained by low engagement indicators, especially low average daily events, limited problem activity, and weak access behavior. These features push the prediction score upward and increase the predicted dropout risk. In contrast, Fig. 6-(c) and Fig. 6-(d) show that stronger weekly engagement reduces the prediction score and is associated with low-risk learner profiles. Overall, the SHAP waterfall plots demonstrate how the model explains dropout risk at the individual learner level by identifying the main behavioral factors that increase or decrease the prediction.

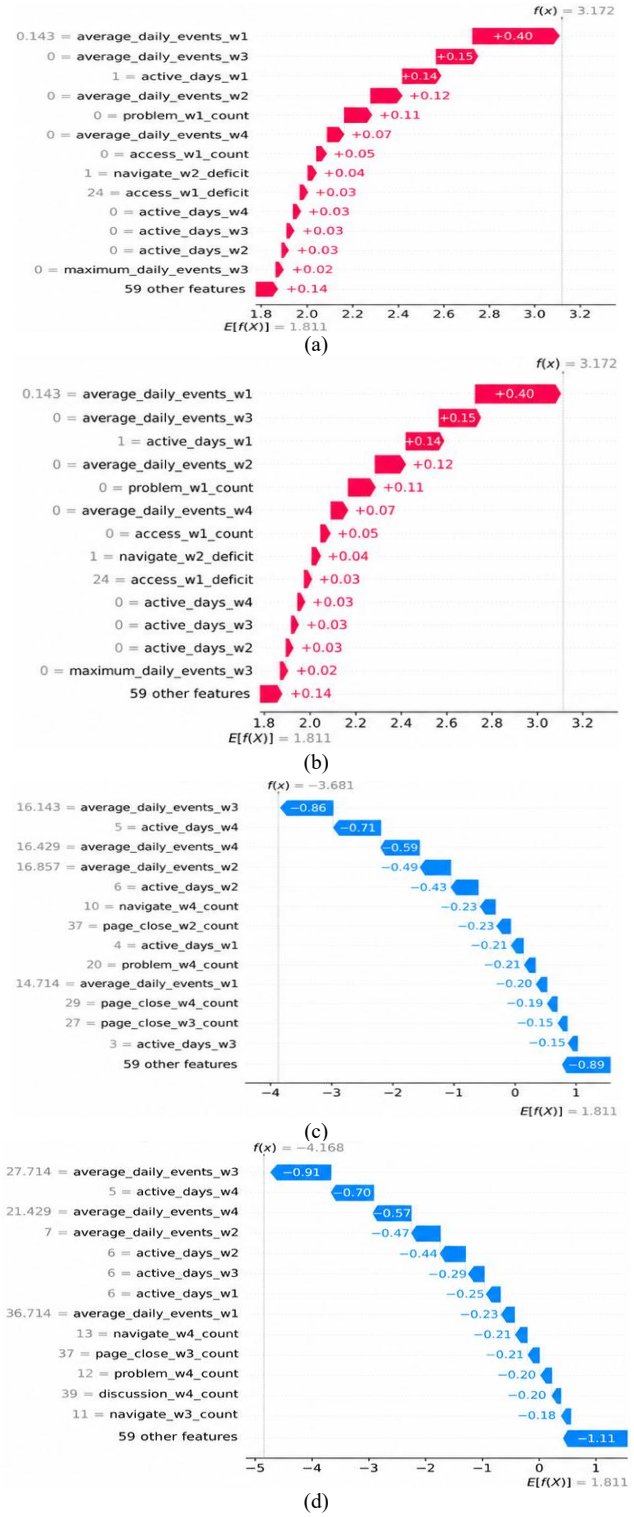


Fig. 6. SHAP waterfall plot for (a) High-risk dropout case, (b) High-risk non-dropout case, (c) Low-risk dropout case, and (d) Low-risk non-dropout case.

4) Weekly L-AXAI intervention output: As shown in Table X and XI, the weekly L-AXAI module generated 13,237 high-risk intervention records on the test set. Approximately 11,610 corresponded to actual dropout records and 1,627 to non-dropout records. The generated messages are candidate decision-support suggestions and were not evaluated with learners or instructors. Navigation nudges represented 74.62% of all candidate nudges and 89.76% of the Week 2 nudges. This concentration reflects both the behavior represented in the test set and the current one-to-one mapping rule, which selects only the strongest positive actionable SHAP driver. It may also be reinforced by correlation filtering: event_diversity_w2_deficit was removed because it was highly correlated with navigate_w2_deficit ($|r| = 0.9548$), leaving the retained navigation feature to represent part of this related Week 2 signal. The distribution should therefore not be interpreted as a universal pedagogical pattern. For future deployment, correlated drivers should be grouped into broader support families, the top-k actionable drivers may be considered, course-specific thresholds should be used, repeated delivery of the same message should be limited, and instructor approval should be retained. These safeguards are proposed design recommendations and were not empirically tested in the present study.

TABLE X. OVERALL DISTRIBUTION OF WEEKLY L-AXAI NUDGE TYPES

Nudge Type	Count	Percentage
Navigation nudge	9,877	74.62
Re-engagement nudge	2,300	17.38
Session completion nudge	962	7.27
Regularity nudge	57	0.43
Content review nudge	33	0.25
Balanced engagement nudge	8	0.06

TABLE XI. WEEKLY DISTRIBUTION OF L-AXAI NUDGE TYPES

Week	Nudge Type	Count	Percentage within week
1	Balanced engagement nudge	8	0.39
	Content review nudge	33	1.6
	Re-engagement nudge	1,142	55.28
	Session completion nudge	883	42.74
2	Navigation nudge	9,812	89.76
	Re-engagement nudge	1,116	10.21
	Regularity nudge	1	0.01
	Session completion nudge	2	0.02
3	Navigation nudge	65	32.83
	Regularity nudge	56	28.28
	Session completion nudge	77	38.89
4	Re-engagement nudge	42	100

V. DISCUSSION

The results describe the predictive performance of the complete weekly feature representation under a stratified train-validation-test evaluation setting. Stratification preserves the dropout and non-dropout distribution in the three subsets. This is important since the dataset is imbalanced. Thus, the reported performance measures the ability of the model to distinguish dropout from non-dropout students with similar class distributions during the training, validation, and testing phases.

The evaluation should be interpreted as a stratified within-benchmark and within-distribution comparison. Training-only estimation of course-reference medians and feature filters prevents direct use of validation or test labels, but the random split does not provide chronological, learner-independent, or course-independent separation. Learner-course records with similar behavioral patterns may therefore occur across subsets, and the reported metrics may be optimistic relative to deployment on completely unseen courses or future cohorts. The findings do not establish cross-platform generalizability. Transfer to another MOOC environment would require event-schema harmonization, verification of the dropout definition and observation horizon, re-estimation of reference medians, probability recalibration, and contextual review of the translation matrix.

The best predictive performance was obtained by tree-based boosting models. Gradient Boosting and XGBoost performed very similarly, and Logistic Regression and SVM stayed competitive. This suggests that weekly clickstream features contain strong predictive signals that can be exploited by linear and non-linear models.

The explainability analysis identifies model-associated behavioral patterns within the present dataset. The SHAP summary plot and local case analysis indicate how weekly activity intensity, active days, and early problem engagement contributed to the selected model's predictions. These associations should not be interpreted as causal determinants of dropout because clickstream logs omit learner goals, prior knowledge, offline study, motivation, and contextual constraints. Explanation stability across repeated training runs, permutation-based validation, and comparison with alternative explainability techniques were not evaluated. The reported SHAP patterns should therefore be interpreted as descriptive explanations of the selected model under the current split rather than as universally stable behavioral findings.

The temporal SHAP analysis describes how the selected model's attribution patterns vary across weekly feature groups. Temporal features are more important in later weeks, while the role of event-count and deficit features depends on the observation window. Therefore, the dropout risk should not be seen as a constant. The features that explain dropout risk in Week 1 are not the same features that explain dropout risk in Weeks 3 or 4. These findings provide a temporal description of the selected model's explanations, but they do not independently establish the predictive contribution of each feature family.

Interpretation at the learner level is emphasized in the local waterfall analyses. Two learners may have similar high-risk probabilities, but their explanations may reflect different combinations of weak activity, limited practice, or low weekly intensity. The high-risk non-dropout case also illustrates a limitation of prediction using clickstream data alone: some learners can behave like dropout learners but still persist. L-AXAI nudges should therefore be seen as supportive signals, not as punitive decisions.

Intervention analysis shows that local explanations can be translated into pedagogical actions at scale. But the strong dominance of navigation nudges should be treated with caution. This could be a real behavior pattern in the test set, or it could be that the translation matrix is giving a lot of priority to navigate_w2_deficit . The intervention matrix should be tuned in future work using feedback from instructors, learner-response data, and A/B testing to ensure that the nudges generated are varied, useful, and not repetitive.

VI. Conclusion

A weekly Localized Actionable Explainability framework for MOOC dropout prediction and intervention was proposed. Clickstream behavior is represented through weekly event counts, temporal indicators, and course-referenced deficit features, after which estimated dropout risk is linked to a local SHAP driver and a transparent candidate nudge template. Under the fixed stratified test split, Gradient Boosting achieved 85.79% accuracy, 86.57% ROC-AUC, and 94.41% PR-AUC, while XGBoost produced closely comparable results. Because repeated-seed experiments and statistical significance tests were not conducted, the small performance differences should be interpreted cautiously. The weekly-horizon analysis indicates that additional weeks provide progressively more predictive information, but a complete feature-family ablation was not performed. SHAP identified model-associated patterns involving activity intensity, active days, and early problem engagement; these findings are descriptive and were not validated through repeated-run stability or alternative explainability methods. The intervention layer generated candidate nudges dominated by navigation, re-engagement, and session-completion categories. This concentration requires diversified driver selection, repetition limits, and instructor-controlled delivery. Because no instructor panel, learner study, or controlled intervention trial was conducted, pedagogical effectiveness and perceived pressure remain unverified. The findings are limited to the KDD Cup 2015 benchmark and a stratified within-distribution protocol. Future work should examine repeated-seed, learner-grouped, course-held-out, and chronological evaluation; formal probability calibration; component-wise ablation; explanation-stability analysis; cross-platform event harmonization; expert co-design of the translation matrix; and real-world intervention studies.

References

[1] G. Siemens and R. S. J. d. Baker, “Learning analytics and educational data mining: towards communication and collaboration,” in Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, 2012, pp. 252–254. doi: 10.1145/2330601.2330661.

[2] R. Ferguson, “Learning analytics: drivers, developments and challenges,” International Journal of Technology Enhanced Learning, vol. 4, no. 5/6, pp. 304–317, 2012, doi: 10.1504/IJTEL.2012.051816.

[3] D. Clow, “An overview of learning analytics,” Teaching in Higher Education, vol. 18, no. 6, pp. 683–695, 2013, doi: 10.1080/13562517.2013.827653.

[4] A. Pardo and G. Siemens, “Ethical and privacy principles for learning analytics,” British Journal of Educational Technology, vol. 45, no. 3, pp. 438–450, 2014, doi: 10.1111/bjet.12152.

[5] S. Slade and P. Prinsloo, “Learning analytics: Ethical issues and dilemmas,” American Behavioral Scientist, vol. 57, no. 10, pp. 1510–1529, 2013, doi: 10.1177/0002764213479366.

[6] H. Drachsler and W. Greller, “Privacy and analytics: it’s a DELICATE issue,” in Proceedings of the 6th International Conference on Learning Analytics and Knowledge, 2016, pp. 89–98. doi: 10.1145/2883851.2883893.

[7] C. Romero and S. Ventura, “Educational data mining and learning analytics: An updated survey,” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 10, no. 3, p. e1355, 2020, doi: 10.1002/widm.1355.

[8] R. S. Baker and P. S. Inventado, “Educational data mining and learning analytics,” in Learning Analytics: From Research to Practice, Springer, 2014, pp. 61–75. doi: 10.1007/978-1-4614-3305-7_4.

[9] R. F. Kizilcec, C. Piech, and E. Schneider, “Deconstructing disengagement: analyzing learner subpopulations in massive open online courses,” in Proceedings of the 3rd International Conference on Learning Analytics and Knowledge, 2013, pp. 170–179. doi: 10.1145/2460296.2460330.

[10] D. F. O. Onah, J. Sinclair, and R. Boyatt, “Dropout rates of massive open online courses: behavioural patterns,” in EDULEARN14 Proceedings, 2014, pp. 5825–5834. [Online]. Available: https://wrap.warwick.ac.uk/id/eprint/65543/

[11] S. Halawa, D. Greene, and J. Mitchell, “Dropout prediction in MOOCs using learner activity features,” in Proceedings of the Second European MOOC Stakeholder Summit, 2014, pp. 58–65.

[12] M. Kloft, F. Stiehler, Z. Zheng, and N. Pinkwart, “Predicting MOOC dropout over weeks using machine learning methods,” in Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs, 2014, pp. 60–65. doi: 10.3115/v1/W14-4111.

[13] C. Taylor, K. Veeramachaneni, and U.-M. O’Reilly, “Likely to stop? Predicting stopout in massive open online courses,” 2014. [Online]. Available: https://arxiv.org/abs/1408.3382

[14] M. Fei and D.-Y. Yeung, “Temporal models for predicting student dropout in massive open online courses,” in Proceedings of the IEEE International Conference on Data Mining Workshop, 2015, pp. 256–263. doi: 10.1109/ICDMW.2015.174.

[15] J. Whitehill, K. Mohan, D. Seaton, Y. Rosen, and D. Tingley, “Delving deeper into MOOC student dropout prediction,” 2017. [Online]. Available: https://arxiv.org/abs/1702.06404

[16] J. Gardner and C. Brooks, “Dropout model evaluation in MOOCs,” 2018. [Online]. Available: https://arxiv.org/abs/1802.06009

[17] S. Nagrecha, J. Z. Dillon, and N. V. Chawla, “MOOC dropout prediction: lessons learned from making pipelines interpretable,” in Proceedings of the 26th International Conference on World Wide Web Companion, 2017, pp. 351–359. doi: 10.1145/3041021.3054162.

[18] W. Feng, J. Tang, and T. X. Liu, “Understanding dropouts in MOOCs,” in Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 517–524. doi: 10.1609/aaai.v33i01.3301517.

[19] A. Alamri, Z. Sun, A. I. Cristea, G. Senthilnathan, L. Shi, and C. Stewart, “Is MOOC learning different for dropouts? A visually-driven, multi-granularity explanatory ML approach,” 2020. [Online]. Available: https://arxiv.org/abs/2008.05209

[20] E. Er, “An explainable machine learning approach to predicting and understanding dropouts in MOOCs,” Kastamonu Education Journal, vol. 31, no. 1, pp. 143–154, 2023, doi: 10.24106/kefdergi.1246458.

[21] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Advances in Neural Information Processing Systems, 2017. [Online]. Available: https://arxiv.org/abs/1705.07874

[22] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you? Explaining the predictions of any classifier,” in Proceedings of the 22nd

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[23] L. S. Shapley, "A value for n-person games," in Contributions to the Theory of Games, Princeton University Press, 1953, pp. 307–317.

[24] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: automated decisions and the GDPR," Harvard Journal of Law & Technology, vol. 31, no. 2, pp. 841–887, 2017, [Online]. Available: https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf

[25] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 607–617. doi: 10.1145/3351095.3372850.

[26] S. Verma, J. P. Dickerson, and K. Hines, "Counterfactual explanations for machine learning: A review," 2020. [Online]. Available: https://arxiv.org/abs/2010.10596

[27] V. Swamy, R. Kämpgen, S. Bataille, and T. Käser, "Can explainable AI methods help understand student performance prediction models?," in Proceedings of the 15th International Conference on Educational Data Mining, 2022, pp. 206–217. [Online]. Available: https://educationaldatamining.org/edm2022/proceedings/2022.EDM-long-papers.9/

[28] C. Molnar, Interpretable Machine Learning. Self-published, 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[29] Z. C. Lipton, "The mythos of model interpretability," Queue, vol. 16, no. 3, pp. 31–57, 2018, doi: 10.1145/3236386.3241340.

[30] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, pp. 206–215, 2019, doi: 10.1038/s42256-019-0048-x.

[31] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017. [Online]. Available: https://arxiv.org/abs/1702.08608

[32] S. Alghamdi, "A comprehensive review of dropout prediction methods in Massive Open Online Courses," Multimodal Technologies and Interaction, vol. 9, no. 1, p. 3, 2025, doi: 10.3390/mti9010003.

[33] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[34] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.

[35] I. K. Nti and S. Ramanayake, "Explainable machine learning for student dropout prediction and tailored interventions in online personalized education," Discover Artificial Intelligence, 2026, doi: 10.1007/s44163-026-01016-6.

[36] S. Boyer and K. Veeramachaneni, "Transfer learning for predictive models in massive open online courses," in Artificial Intelligence in Education, Springer, 2015, pp. 54–63. doi: 10.1007/978-3-319-19773-9_6.

[37] B. Jeon, N. Park, and S. Bang, "Dropout prediction over weeks in MOOCs via interpretable multi-layer representation learning," 2020. [Online]. Available: https://arxiv.org/abs/2002.01598

[38] B. Jeon and N. Park, "Dropout prediction over weeks in MOOCs by learning representations of clicks and videos," 2020. [Online]. Available: https://arxiv.org/abs/2002.01955

[39] Y. Liu, "MOOC dropout prediction via a dilated convolutional learning behavior model," Informatics, vol. 12, no. 4, p. 127, 2025, doi: 10.3390/informatics12040127.

[40] A. Ajayi, "Predicting student dropout risk in online learning using stacked ensemble learning and explainability," International Journal of Computer Applications, vol. 187, no. 40, pp. 1–8, 2025, [Online]. Available: https://ijcaonline.org/archives/volume187/number40/ajayi-2025-ijca-925707.pdf

[41] M. Nagy, R. Molontay, and M. Szabo, "Interpretable dropout prediction: towards XAI-based personalized intervention," International Journal of Artificial Intelligence in Education, 2024, doi: 10.1007/s40593-023-00331-8.

[42] P. Buñay-Guisñan, M. A. Al-Fedaghi, and J. Sánchez-Gordón, "Group counterfactual explanations: A use case to support students at risk of dropout," Electronics, vol. 15, no. 1, p. 51, 2026, doi: 10.3390/electronics15010051.

[43] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1007/BF00994018.

[44] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[45] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," Nature Machine Intelligence, vol. 2, pp. 56–67, 2020, doi: 10.1038/s42256-019-0138-9.

Counterfeit Currency Detection Through Deep Convolutional Generative Adversarial Network

Salman Jan[1], Fara Gul[2], Mohammad Riyaz Belgaum[3],
Shahid Kamal[4], Afaq Ahmad[5], Muhammad Bilal[6], Atif Khan[7]
Faculty of Computer Studies, Arab Open University – Bahrain, A'ali, Kingdom of Bahrain[1, 3]
Bacha Khan University, Charsadda, Pakistan[2]
Center for Advanced Analytics-CoE for Artificial Intelligence-Faculty of Computing and Informatics, Multimedia University, Cyberjaya, 63100, Selangor, Malaysia[1, 4]
Department of Computer Science, Institute of Management Sciences, Peshawar, Pakistan[5]
Department of Computer Science, Islamia College University, Peshawar, Pakistan[6, 7]

***Abstract*—Worldwide, individuals, businesses, and economies are being affected by counterfeit currency significantly. Numerous models generate content that mimics the original, and existing solutions fall short in detecting the difference between real and fake. This study presents a detection system using three deep learning models: Deep Convolutional Generative Adversarial Network (DCGAN), Convolutional Neural Network (CNN), and Fully Connected Neural Network (FCNN). The proposed study also identifies patterns that contribute to real and fake currency when the models are trained on the data. After training the models, the proposed solution receives an accuracy of 95 per cent, an F1 Score of 0.957, 0.95 as precision, and 0.954 as recall. This study further carries out a comprehensive analysis of existing models and compares them with the proposed solution to determine the effectiveness of the solution and further recommend its implementation in real-world applications. The proposed solution contributes to the widespread adoption of the application across smart devices and further ensures a robust solution for the detection of counterfeit money.**

***Keywords*—CNN; counterfeit; DCGAN; Fake Money; FCNN; fraud detection**

I. Introduction

Counterfeiting pertains to the unlawful replication of a specific object, notably the currency of its original form. The repercussions of counterfeit currency can be detrimental to an economy, causing an upsurge in prices, fostering black market activities, influencing trade, business transactions, and administrative operations [1]. Currency counterfeiting persistently poses a formidable challenge to the financial systems of nations [2]. It involves the illicit replication of the official currency, hence lacking the authorization of the government, and poses a substantial menace to the global economy [3]. The production of counterfeit currency, also referred to as counterfeit money, is universally illegal since it transpires without the sanction of the central bank or the governing authority. The history of counterfeit currency traces back to antiquity, emerging nearly simultaneously with the inception of currency itself. It stands as one of the world's oldest illicit professions, even predating Lydian coins. The Lydian Stater, which predated the fall of the Lydian Empire to the Persian Empire, represented the inaugural instance of an officially issued coin by a government in recorded history [4]. Counterfeit coins, known as Fourree, were crafted by amalgamating base metals with pure gold or silver.

Modern approaches to discern genuine from counterfeit currency primarily rely on advanced image processing techniques. These methods encompass image acquisition, edge detection, feature extraction, image segmentation, and decision-making processes. Noteworthy security features include watermarks, concealed images, security threads, and optically variable inks [5]. However, the primary limitation of these approaches lies in their suboptimal detection efficiency, particularly regarding feature extraction [6]. To surmount this challenge, deep learning methodologies have emerged as compelling solutions. These techniques deploy multi-layer neural networks to facilitate the detection of counterfeit currency. While increasing the layers in neural networks can lead to greater complexity and higher error risks, convolutional neural networks (CNNs) offer a promising alternative [7]. Unlike fully connected layers, CNNs use filters that focus on important aspects of input data, scanning across the entire input data space to extract relevant features for specific tasks. CNNs are known for their power and efficiency, especially compared to traditional deep neural networks. In today's technologically advanced world, it has become easier to produce counterfeit currency without legal authorization, significantly impacting a nation's economy. The Malaysian Ringgit (MYR), issued by Bank Negara Malaysia (BNM), is particularly vulnerable to counterfeiting, especially in denominations of MYR 1, MYR 5, MYR 10, MYR 20, MYR 50, and MYR 100. In the field of computer science, many machine learning algorithms have been proposed for image processing. These methods detect and recognize currency notes by analyzing specific attributes such as color, shape, paper width, and image filtering.

In this study, three deep learning models were employed to develop an effective counterfeit currency detection system: Convolutional Neural Network (CNN), Fully Connected Neural Network (FCNN), and Deep Convolutional Generative Adversarial Network (DCGAN). CNNs are widely recognized for their ability to process and analyze images by automatically extracting meaningful features through multiple layers, making them as potential for visual recognition tasks. FCNNs complement this process by utilizing fully connected layers to perform accurate classification based on the features extracted

from the input data. To further enhance detection performance, the study incorporates DCGAN, an advanced deep learning framework within the field of generative artificial intelligence. DCGAN consists of two interconnected components: a Generator and a Discriminator. The Generator learns the underlying patterns of genuine currency notes and produces realistic synthetic samples, while the Discriminator evaluates these samples and distinguishes between authentic and generated data. Through this adversarial learning process, the model becomes highly effective at identifying subtle characteristics of counterfeit currency, leading to improved detection accuracy and robustness.

A. Motivations and Research Gap

The problem of counterfeit money continues to pose a challenge and threat in the economic world. Traditional methods of counterfeit detection, such as manual inspection and simple image processing techniques, have proven inadequate in keeping up with the sophisticated techniques employed by counterfeiters [8]. This research aims to fill the gap by leveraging advanced deep learning techniques to enhance the detection accuracy and reliability of counterfeit currency identification. Specifically, the existing challenges and gaps that our research addresses include:

- Inefficiency of Traditional Methods: Traditional methods of counterfeit detection are labor-intensive and prone to errors, particularly in the case of high-quality counterfeits.
- Lack of Robustness in Existing Models: Previous models lack the robustness required to detect subtle variations and sophisticated counterfeiting techniques.
- Integration of Security Features: While some methods incorporate security features, they fail to ensure the integrity and traceability of detection results.
- DCGAN was selected because access to diverse counterfeit currency samples is limited. The generator produces synthetic counterfeit-like variations, while the discriminator learns to distinguish subtle visual differences between genuine and generated samples. This adversarial training exposes the discriminator to a broader range of counterfeit patterns than conventional supervised training alone. CNN and FCNN were included as baseline discriminative models for comparison.

B. Problem Statement

The presence of counterfeit currency causes negative impacts on the economy of the nation, which can lead to inflation and losses, both to the country and its people. In addition, this leads to distrust of the integrity of the currency, especially when conducting business. Upon careful analysis of the various methods that have been put in place to differentiate between the two, it becomes clear that traditional methods have been used before.

However, the counterfeiting technique has advanced to such a degree that the changes introduced into counterfeit notes escape detection through conventional machine learning methods. In this regard, in this study, deep learning algorithms have been employed on the basis that they can detect complex patterns in better ways. The problem is solved through the introduction of a strong system for counterfeit money detection. The proposed approach should be quite efficient to meet the needs of governments, financial organizations, companies, retailers, and even customers in distinguishing real currency from counterfeit currency. Consequently, the following noteworthy contributions have been made throughout the development of the proposed solution.

C. Contribution

This study proposed a framework for counterfeit currency detection which can help government agencies and banking sectors to distinguish genuine currency from counterfeit currency in an expedited manner. The contribution of this study is provided below:

- Novel Deep Learning Framework: This research presents an extensive framework for counterfeit currency detection based on advanced deep learning models such as DCGAN, CNN, and FCNN. Specifically, the DCGAN model performs better in recognizing counterfeit notes by capturing complex patterns that other models find hard to detect.
- Cross-Currency Applicability: The system proposed is not specific to any one currency. It has been effectively used to detect counterfeit Malaysian Ringgit (MYR) and Saudi Arabian Riyal (SAR). It can be extended to other currencies with minor adjustments in training.
- Performance Superiority: The proposed approach based on DCGAN achieves better accuracy, precision, recall, and F1 score compared to other existing methods, with 95.11% accuracy, 0.95 precision, 0.954 recall, and 0.957 F1 score. A comparative analysis with previous work clearly demonstrates the superiority of our method.
- Cost-Effective and Scalable: The framework is designed to be inexpensive and scalable, providing a robust solution for large-scale counterfeit detection and easy integration with smart devices and financial systems.

This research study is structured as follows: The study opens with a short overview of the work that was done in Section I. Section II offers an in-depth review of related work in the field and discusses its strengths and weaknesses. Section III shows how the authors addressed the gap as part of methodology. Section IV presents experiment settings followed with the experimental results and discussion drawn from all experiments using the proposed framework in Section V. Section VI presents a complete set of conclusions and recommendations and finally drafting the future work is presented in Section VII.

II. Related Work

The authors in [9] developed a system based on image processing techniques, utilizing OpenCV (CV2), Support Vector Machine (SVM) classifier, and other machine learning algorithms to extract image features. They compared these features with the genuine characteristics of banknotes. They effectively employed CNN to detect counterfeit currency.

An innovative approach to counterfeit currency detection using transfer learning was introduced by the authors in [10]. They employed AlexNet as a pre-trained model, consisting of convolutional layers, max-pooling layers, dropout layers, ReLU activations, and fully connected layers. To facilitate transfer learning, they removed the last three fully connected layers of AlexNet and adjusted it with weight and bias learning factors to extract latent features from currency images. Their training set was comprised of currency note images, augmented with 100 variations per image by altering the size and orientation, thereby increasing dataset diversity. The increased images were placed in folders according to the corresponding labels. The authors achieved average accuracy rates of 81.5% for genuine currency and 75% for counterfeit currency.

Kumar et al. [11] proposed a robust system for detection of counterfeit Indian rupee notes using ORB (Oriented FAST and Rotated BRIEF) and OpenCV's Brute Force matcher. Their system was based on digital image acquisition, grayscale conversion, edge detection, image segmentation and feature extraction. It has two modules, one is to extract security features and create database and other is to authenticate images from the test dataset. They achieved a significant average accuracy of 95.0% indicating the system is cost effective.

Pachon et al. [12] used artificial neural networks (ANNs) for banknote authentication. For their research they took data from specially created images for the purpose of improving currency classification accuracy with the help of Back Propagation algorithm. Their suggested procedure was image acquisition, conversion to grayscale, wavelet transformation, image segmentation, and classifier application. But they provided no details on implementation, results or average accuracy metrics.

In [13], Chandrappa et al. suggested a methodology for extracting critical information from the images of counterfeit money using bit plane slicing. This was done by decomposing 256 grayscale images into eight binary images, each representing different bit planes and applying edge detection techniques. They found that the sixth and seventh bit planes greatly improved edge and hidden element detection in the images. The methodology was implemented using image processing toolbox in MATLAB.

Teymournezhad et al. [14] proposed a system that combines image processing techniques and a deep learning network named GoogLeNet. The proposed work utilized the pre-trained GoogLeNet. It demonstrated high accuracy in detecting the security features of the genuine banknotes but low accuracy for the counterfeit samples. Image preprocessing steps consist of scaling, noise removal, sharpening, brightness adjustment, image segmentation and corner recognition. A 22-layer GoogLeNet was used for object recognition across a wide range of classes. The system was implemented in MATLAB R2018b on a Windows 7 platform.

Chowdhury et al. [15] proposed an automated authentication model using irregular color particles under UV light with UV ink as a security feature. The model achieved an accuracy of 96.3%. They used Region Proposal Network Stages (RPNs) to estimate quickly the regions of interest of different scales and aspect ratios.

Nur Farah et al. [16] used video spectral comparator (VSC), ATR-FTIR for chemical profiling and PCA for pattern recognition to differentiate genuine from counterfeit banknotes. Their methodology involved sample collection, physical and optical examinations, IR spectra acquisition, pattern recognition application, and PCA. However, their study lacked implementation, accuracy, and conclusive findings.

The authors in [17] used generative adversarial networks (GANs) for detecting fraudulent currency. GANs employed unsupervised learning to make supervised predictions, demonstrating high precision with an expanded image dataset. Their study included various deep learning models, achieving an accuracy rate of 80%. They highlighted the potential for enhancing accuracy with multi-class classifiers.

A standard currency identification system that makes use of edge detection, feature extraction, image segmentation, acquisition, grayscale transformation, and comparison was developed by Bandu et al. [18]. The research made use of CNN for pattern and image recognition, and suggested the use of SHA as an identifier for counterfeit currency depending on its physical characteristics. Nevertheless, the authors did not give system implementation and accuracy evaluation.

Nasayreh et al. [19] used different CNN architectures for identification of fake Jordanian bank notes and obtained competitive levels of accuracy using optimized AlexNet and GoogleNet models. The modifications in VGG16 resulted in maximum accuracy of 99.88%. Their approach can be implemented in automated payment systems and mobile currency checking. Aulia et al. [20] showed the recognition and classification of paper currency for Saudi Arabian Riyal denominations using Geometric Template Matching and Grayscale Template Matching techniques.

III. METHODOLOGY

This section presents the proposed methodology for detection of counterfeit currency, as shown in Fig. 1. The proposed framework is composed of the following components:

- Dataset
- Generator
- Latent Space
- Discriminator
- Fine-tuning

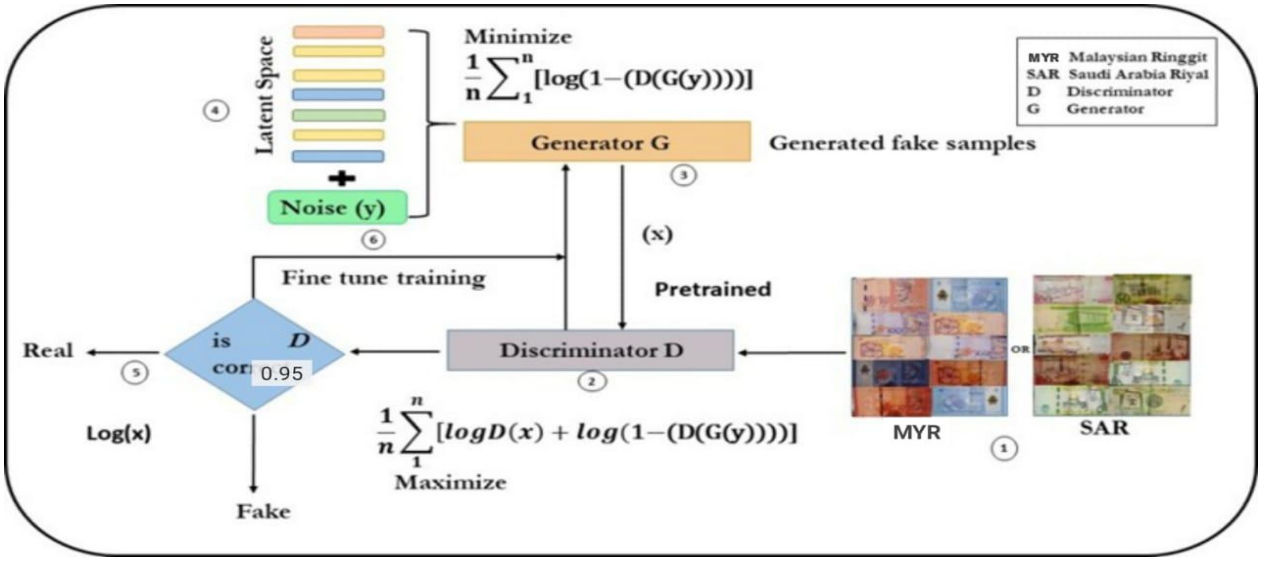


Fig. 1. Proposed framework for counterfeit currency detection using DCGAN.

A. Dataset

In this model, the system is designed to accept input data in the form of Malaysian Ringgit (MYR) and Saudi Arabian Riyal (SAR) currency notes. The dataset used in this study consisted of 15,148 images of Malaysian Ringgit (MYR) and Saudi Arabian Riyal (SAR) currency notes, including both genuine and counterfeit samples. The genuine images were taken using a phone, while counterfeit samples were generated using the generative component of the DCGAN. All images were resized to 128 × 128 pixels and converted to RGB format before training. The dataset was divided into 80% training data and 20% testing data, with 20% of the training portion used for validation. The class-wise distribution of genuine and counterfeit images should also be reported. However, it is important to note that this system can be adapted to verify the authenticity of currency notes from any denomination if it undergoes proper training with an adequate dataset. The flexibility of this solution allows for its application to various currencies by ensuring the model receives appropriate training. It's worth noting that there is no access to individuals involved in the production of counterfeit currency. Hence, an artificial function is employed to generate simulated counterfeit currency samples. These synthetic samples are utilized as representations of counterfeit currency notes and are generated by a dedicated algorithmic generator.

B. Generator

The framework employs the DCGAN model, originally proposed by Radford et al. [21], which has shown remarkable capabilities in generating realistic data samples and improving classification tasks. The DCGAN generator component generates fake samples by using latent space. A generator uses a mathematical formula and gives input in the digital form of the currency. The fake samples produced by the Generator are sent to Discriminator.

C. Latent Space

The latent space is like a database in which currency notes along with some noises are added in the digital form. With latent space, the generator gives output in the digital form of the currency. In this way, we have real and fake currency.

D. Discriminator

The discriminator's main function is to evaluate whether a currency sample, generated by a dedicated generator, is genuine or counterfeit. This is achieved through a training process where the discriminator learns to distinguish between real and fake currency notes. The system works accurately when the discriminator correctly identifies a genuine note as real and a counterfeit note as fake. If the discriminator incorrectly classifies a genuine note as counterfeit, corrective measures are taken to improve its performance. This fine-tuning process aims to enhance the discriminator's accuracy in differentiating between real and fake currency. The goal is to minimize the probability of the generator's output being mistaken while maximizing the discriminator's accuracy. In practice, the discriminator should aim for a probability of 1 when assessing genuine notes, indicating a correct decision. Conversely, a probability of 0 indicates an incorrect judgment. Therefore, the objective is to reduce the number of such erroneous classifications.

E. Fine-Tuning

Fine-tuning is a critical step in transfer learning, involving the adaptation of pre-trained model weights to new data instances. When the discriminator incorrectly identifies genuine currency as counterfeit, it indicates an error in its decision-making. To address this, a fine-tuning process is initiated to improve the discriminator's accuracy. This corrective procedure enhances the discriminator's ability to accurately distinguish between authentic and counterfeit currency notes, thereby increasing its overall proficiency.

IV. Experimental Settings

To provide a comprehensive understanding of the methodology and facilitate reproducibility, a detailed information about the models and settings employed in the proposed work are presented. The DCGAN model architecture consists of two main components: the Generator and the

Discriminator. The Generator starts with a 100-dimensional noise vector as input. It includes a dense layer with 128 units, utilizing ReLU activation and batch normalization to stabilize the learning process. This is followed by upsampling layers, which are paired with Conv2DTranspose layers. These layers have filters of sizes 128, 64, and 32, with a kernel size of 5x5, strides of 2x2, and 'same' padding, all using ReLU activation. The last layer of the Generator is a Conv2DTranspose layer which generates the output as RGB images of size 128x128x3, with a 5x5 kernel, 1x1 strides, 'same' padding, and tanh activation. After each Conv2DTranspose layer, dropout layers with rate 0.3 are used to prevent overfitting.

The Discriminator on the other hand takes as input 128x128x3 RGB images. It consists of Conv2D layers with filter sizes of 32, 64 and 128. Kernel size is 5x5, stride is 2x2 and padding is 'same'. These layers use LeakyReLU activation with a slope of α=0.2 which allows a small gradient when the unit is not active. After each Conv2D layer, the research added a dropout layer with a rate of 0.3 to further regularize the model. Finally, the Discriminator ends with a dense layer with 1 unit and sigmoid activation to output the probability of the input image being real or fake. For training, the model is run for 10 epochs with a batch size of 128. The research uses the Adam optimizer with learning rate 0.0002 and decay rate β=0.5. The Generator and Discriminator are using binary cross entropy for their loss functions. The loss functions tell your model what to learn.

A. Test Bed Environment

This study focuses on the experimental setup for training a DCGAN. The training was carried out on a GPU based machine at the Cyber Security and Technological Convergence department, Malaysian Institute of Information Technology, University Kuala Lumpur. The architecture used Keras with TensorFlow as a backend. The Keras library was used with several components such as dense layers, the dropout technique, activation functions, and flatten options. The dense layer was employed to tune the filters to detect the various attributes of the currency notes under examination.

The dropout method helped to reduce overfitting and remove extraneous features that did not add to the proper labeling and the real look of the notes. The “ReLU" activation function was chosen and it was responsible for clipping the input values in the range of 0 to 1. The datasets consisted of the real and counterfeit currency samples which were used to train DCGAN, CNN, and FCNN. The DCGAN framework consisted of two simple components, the Discriminator (D) and the Generator (G). These components were trained with a learning rate of 0.0002. Specifically, the training process is to update the discriminator, then update the generator twice. In this way the discriminator was prevented from minimal loss.

The results of the learning process were generated and well documented. A specific hardware and software environment was required for the development of the proposed framework. The primary memory capacity used in this study was 16 GB with GPU-supported hardware, including the NVIDIA P100 Kernel. All the experiments were performed on the Linux Operating System. With a minimum system configuration of Intel(R) Xeon(R) processor with a clock speed of 0.82 GHz to 1.59 GHz, 2 cores, and 16 GB memory, the experiments are reproducible.

B. Training of Deep Learning Models

During the training of the models, the dataset was divided into three separate datasets: a training set, a validation set, and a test set. The details are provided as follows:

1) Training set: It is the first dataset, used to train the model to learn the complex features of real currency notes in each iterative step. This immersive training allows the model to make accurate predictions when faced with totally new, never-before-seen data. In this work, the dataset is split into 80% training set and 20% test set.

2) Validation set: The validation set is a subset of the training set that is set aside from the training set during each epoch of training. During the training, the model is validated with data from this set. The division of training and validation data enables the model to learn from training data and test the model on the validation set. Importantly, this validation process provides insights that can be used to tune the hyperparameters. The model then classifies and evaluates each input of the validation set during training based on the acquired knowledge of the training set. In this work, 20% of the training data is used as the validation data.

3) Test set: After training and validating the models, they can be applied to predict results for the data of the test set. The 20% test set, not to be used in the training or validation set, is held for the final evaluation of the model. The test set should preferably have no labels and serve as an unbiased evaluation of the predictive ability of the model. The training and validation sets, however, are annotated, making it easier for the model to learn.

V. Results and Discussion

This section presents details of classification results obtained after training deep learning models on real and fake currency notes. The following subsections outline the various performance indicators of the employed models.

A. Performance Analysis of Employed Models

Metrics used in performance evaluation include accuracy, recall, precision, and F1 scores. These measures are critical in measuring the efficiency of different deep learning models used in the study. The study involves the exploration of different parameters that will help in converging to get accurate results. The normal confusion matrix is composed of four categories that include false negative, false positive, true positive, and true negative. In this particular study, “positive" refers to fake currency, whereas “negative" refers to the real currency. For the DCGAN model that analyzed 15,148 images, there were two categories under consideration – positive and negative. Conversely, in the Positive case, the predicted negative instances amounted to 350, with 7,802 predicted positive instances. Notably, DCGAN exhibited the most favorable results among all models.

The CNN model also processed 15,148 samples, however, the negative case resulted in 5,800 predicted negative currency notes and 1,696 predicted positive instances. In the positive case, there were 1,648 instances of predicted negative currency notes and 6,004 predicted positive instances. Finally, examining the

FCNN model, with the same dataset of 15,148 samples, the negative case yielded 5,528 predicted negative currency notes and 2,757 predicted positive instances. In the positive case, the predicted negative instances totaled 2,535, with 4,328 predicted positive instances.

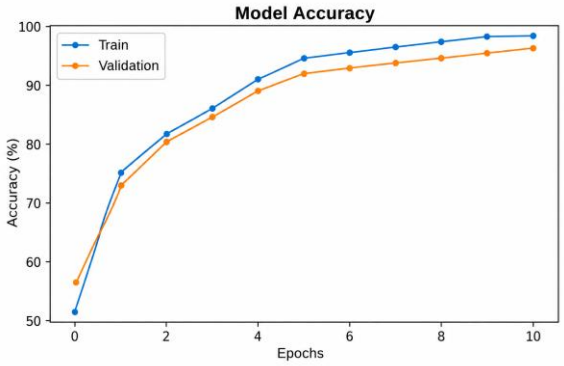


Fig. 2. Accuracy of the DCGAN model.

Fig. 2 shows the trend of the model's accuracy for the DCGAN model. The blue line represents the training trend, while the orange line shows validation accuracy. The figure shows that both the training and validation accuracies are improving with each epoch and finally reach 95%.

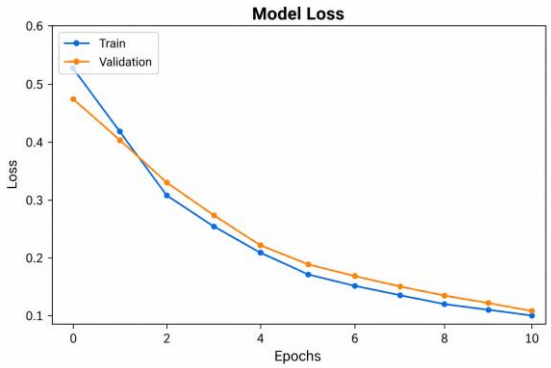


Fig. 3. Loss of the DCGAN model.

Fig. 3 represents the loss function of the model. It is clear from the graph that the loss of training and validation is decreasing after each epoch. Overall, the models represent a good fit.

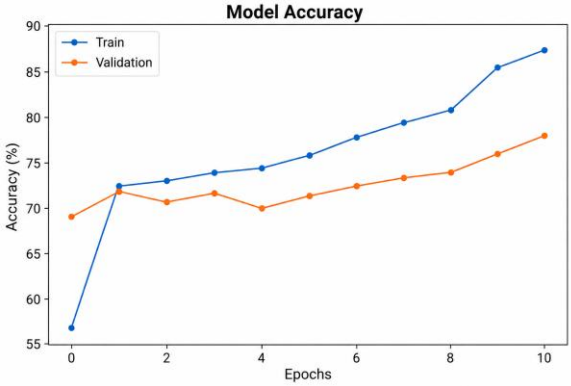


Fig. 4. Accuracy of the CNN model.

Fig. 4 represents the accuracy of CNN model in detecting the Counterfeit currency. It is evident from the above trend that validation accuracy is not improving as training accuracy does. Validation accuracy starts from 70% and after fluctuation reaches to 77%.

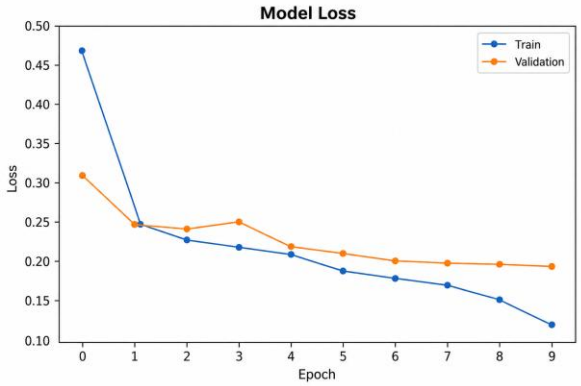


Fig. 5. Loss of the CNN model.

Fig. 5 shows the loss of CNN model. The training loss is constantly decreasing and reaches to 0.1 while validation loss is not decreasing as such. Overall, the model does not reflect a good fit.

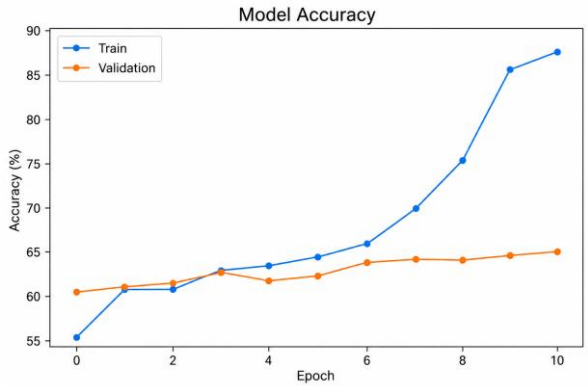


Fig. 6. Accuracy of the FCNN model.

Fig. 6 represents the accuracy trend of FCNN model. The trend indicates that training accuracy is improving with each epoch, however, the validation accuracy is not improving as such.

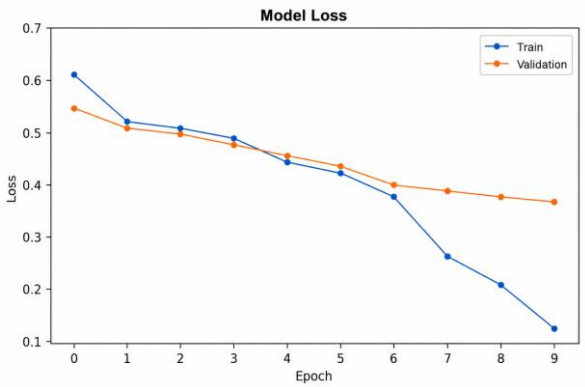


Fig. 7. Loss of the FCNN model.

Fig. 7 shows the loss function of FCNN. The training loss is decreasing constantly and reaches to 0.1 from 0.6. On the other hand, validation loss initially decreased and then remained stagnant. Overall, the mode seems to be overfitting.

The confusion matrix for DCGAN, CNN and FCNN is provided in Table I.

TABLE I. CONFUSION MATRIX OF THE DCGAN, CNN AND FCNN MODELS.

Model	Cases	Predicted Negative	Predicted Positive
DCGAN	Negative Cases	6606	390
	Positive Cases	350	7802
CNN	Negative Cases	5800	1696
	Positive Cases	1648	6004
FCNN	Negative Cases	5528	2757
	Positive Cases	2535	4328

The results of employing the state-of-the-art deep learning approaches to classify real and fake currency is depicted in Fig. 8.

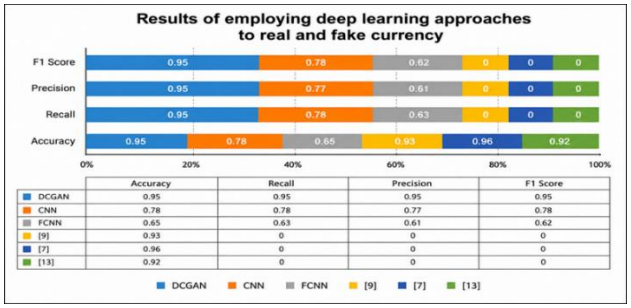


Fig. 8. Results of deep learning approaches to detect genuine currency.

B. Real-Time Complexity Analysis of the Proposed Approach

The proposed counterfeit currency detection framework leverages DCGAN, CNN, and FCNN. The real-time performance of these models is influenced by several factors, including the model's architecture, the size of the dataset, and the computational resources available.

1) DCGAN Complexity:

a) Training complexity: The DCGAN architecture consists of a Generator and a Discriminator, both of which are optimized concurrently during training. The time complexity of training the model depends on the size of the input image (n), the number of layers (L), and the number of parameters in each layer. Given that backpropagation is performed for both the Generator and Discriminator in each epoch, the overall training complexity can be approximated as O(2×n×L).

b) Inference complexity: During inference (real-time detection), the complexity is significantly reduced since only the Discriminator is used to classify an image as genuine or counterfeit. For a DCGAN with L layers, the inference complexity is approximately O(n×L), making it feasible for real-time applications.

2) CNN and FCNN complexity:

a) CNN complexity: CNNs rely on convolutional operations that have a time complexity of O(k×n2×m2), where k is the size of the filter, n×n is the size of the input image, and m×m is the size of the output feature map. While training a CNN is computationally expensive, involving multiple convolutional and pooling layers, inference complexity is relatively low. This makes CNNs suitable for real-time detection scenarios).

b) FCNN complexity: FCNN models, with their fully connected layers, have a time complexity of O(n2×L), where n is the number of neurons in the layer and L is the number of layers. While FCNNs offer high accuracy, they may be less efficient than CNNs and DCGANs in terms of real-time performance due to the fully connected layers, which introduce higher computational demands.

3) Real-time feasibility: For real-time applications, the key factor is inference time. DCGAN's Discriminator, once trained, has a low inference complexity, which makes it well-suited for real-time currency detection. CNNs also offer efficient inference with low latency, though FCNNs may introduce slightly higher inference times due to their dense architecture.

4) Overall real-time performance: While training deep learning models like DCGAN, CNN, and FCNN can be computationally intensive, the real-time complexity for detection (inference) is significantly lower. DCGAN, in particular, offers a balance between accuracy and efficiency, with real-time classification complexity at O(n2×L), making it ideal for deployment in environments where rapid and accurate counterfeit detection is required. As mentioned earlier, the research conducted experiments using a system configuration consisting of an NVIDIA P100 GPU, 16 GB of memory, and an Intel Xeon processor.

Given the DCGAN model's architecture and the number of image samples (15,148), the training time for the DCGAN model was approximately 3-4 hours for 10 epochs with a batch size of 128.

C. Model Accuracy

The experimental results demonstrated that DCGAN obtained a highest accuracy of 95% in detecting counterfeit currency. The accuracy rate is determined using Eq. (1):

$$Accuracy = \frac{Numberofcorrectlyidentifiedsamples}{Totalnumberofsamples} \quad (1)$$

The accuracies of various models employed are presented in Fig. 9.

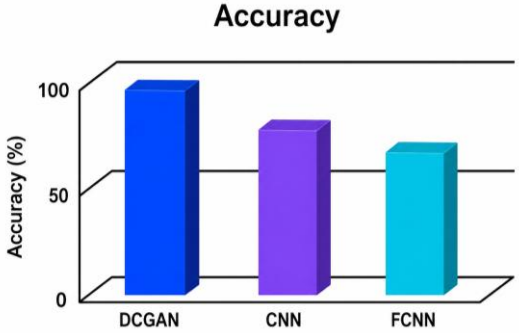


Fig. 9. Accuracy obtained by various models.

D. Recall

The recall metric is employed to measure the performance of the algorithm for recognizing the positives, as illustrated by Eq. (2). The True Positives (TP) refer to the total number of cases which were correctly predicted by the model to be positive (counterfeit currency), while the False Negatives (FN) denote those cases which have been wrongly recognized as negatives (genuine currency).

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2)$$

Recall obtained by different models is presented in Fig. 10. It is evident that DCGAN surpassed other models in terms of recall.

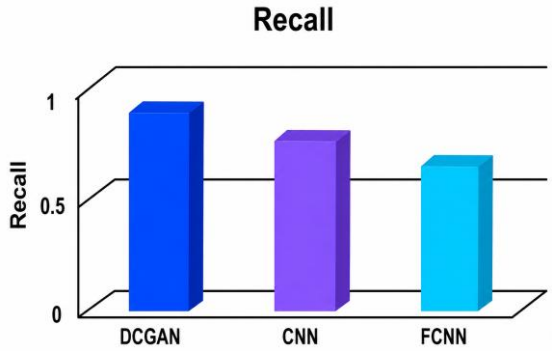


Fig. 10. Recall obtained by various models.

E. Precision

This is the measure that is used to determine the precision of the model's positive predictions, as seen in Eq. (3). The measure is used to find the fraction of true positive predictions among all the positive predictions made by the model. True Positive (TP) is the number of correct positive predictions and False Positive (FP) is the number of wrong positive predictions, where the model predicted a positive outcome (counterfeit currency), but the actual outcome was negative (genuine currency).

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

The Precision obtained by different models is presented in Fig. 11. It is evident that DCGAN surpassed other models in terms of precision.

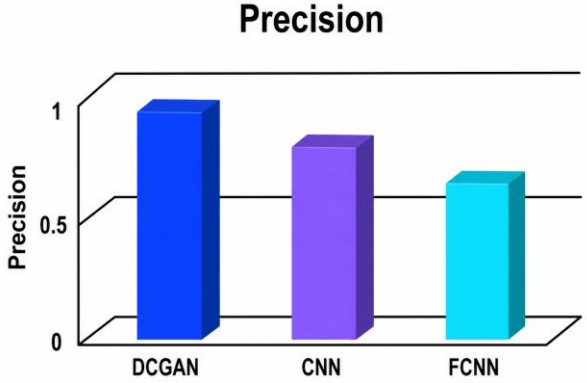


Fig. 11. Precision obtained by various models.

F. F1-Score

F1-score is widely used to evaluate performance of machine learning classifiers including those used for natural language processing applications and search engine algorithms among others [14]. The F1 score is calculated to find a balance between recall and precision. Mathematically F1 Score is calculated using Eq. (4):

$$F1 - Score = 2\frac{(Precision \times Recall)}{(Precision + Recall)} \quad (4)$$

The F1-score obtained by different models is depicted in Fig. 12. The results showed that DCGAN outperformed other models in terms of F1-score.

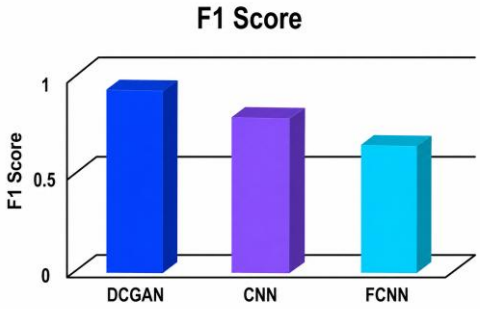


Fig. 12. F1-Score obtained by various models.

G. Comparison with Baseline Methods

To benchmark the performance of our proposed DCGAN-based approach, this research compared the results with those reported in the study by [14], which also employed GAN for counterfeit currency detection. The comparison is summarized in Table II, highlighting the superior performance of our method in terms of accuracy, precision, recall, and F1-score.

TABLE II. COMPARISON WITH BASELINE METHODS

Method	Accuracy	Precision	Recall	F1-score
Proposed DCGAN	0.951	0.950	0.954	0.957
Ref. [10]	0.815 / 0.750*	NR	NR	NR
Ref. [11]	0.950	NR	NR	NR
Ref. [15]	0.963	NR	NR	NR
Ref. [17]	0.800	NR	NR	NR
Ref. [19]	0.9988	NR	NR	NR

NR: Not reported. *Ref. [10] reported 81.5% accuracy for genuine and 75.0% for counterfeit notes.

H. Limitations of Proposed Framework

The proposed framework shows significant improvements in the detection of counterfeit currency, but has the following limitations:

1) Potential vulnerabilities to adversarial attacks: Deep learning models, e.g. DCGAN, are vulnerable to adversarial attacks. The system can be fooled by slight modifications to counterfeit notes. Future work should study how to defend the model against such attacks.

2) Requirement for large and diverse datasets: The performance of the proposed framework depends on the

availability of large and diverse datasets. The dataset utilized for this research is extensive yet not exhaustive of all counterfeit currencies. The dataset can be expanded to include a wider range of counterfeit samples which may increase the performance of the model.

3) Lack of component-wise ablation analysis: The proposed framework combines synthetic image generation, discriminator-based classification, and image preprocessing. However, the present evaluation reports the performance of the complete framework and does not separately measure the contribution of each component. Therefore, the individual impact of synthetic sample generation, discriminator learning, and preprocessing remains a limitation of the current study. Future work will include an ablation study to evaluate each component independently.

4) Lack of repeated experimental trials: The reported results were obtained from a single controlled experimental run. Although the same dataset, training parameters, and computing environment were used for all models, repeated trials with different random seeds were not conducted. Therefore, the absence of mean values, standard deviations, confidence intervals, and statistical significance testing is acknowledged as a limitation. Future work will include multiple independent runs to assess the consistency and statistical reliability of the results.

5) Limited evaluation under real-world imaging conditions: The proposed framework was evaluated using a single dataset under controlled experimental conditions. Its robustness under variations such as lighting changes, camera differences, image rotation, occlusion, damaged banknotes, blur, and noise was not independently assessed. Therefore, the practical generalizability of the framework remains limited. Future work will evaluate the model under diverse real-world image acquisition conditions.

6) Absence of deployment performance metrics: Although the proposed framework has potential for smart-device and real-time deployment, the current study does not report practical deployment metrics such as inference time, model size, parameter count, FLOPs, memory consumption, or CPU/GPU latency. Therefore, real-time and smart-device suitability cannot yet be fully confirmed. Future work will evaluate these metrics on embedded and mobile computing platforms.

7) Absence of qualitative misclassification analysis: The current evaluation reports aggregate performance metrics and confusion matrices but does not include qualitative examples of false-positive and false-negative predictions. Consequently, the specific visual conditions that lead to classification errors, such as blur, poor lighting, occlusion, damaged notes, or visually similar counterfeit features, were not examined. Future work will include representative misclassified samples and a detailed analysis of common failure patterns.

8) Unverified real-time and smart-device deployment: The present study evaluates classification performance but does not measure deployment-related factors such as inference time, model size, parameter count, FLOPs, memory usage, or CPU/GPU latency. Therefore, the suitability of the proposed framework for real-time and smart-device deployment has not yet been experimentally confirmed. Future work will assess these metrics on mobile and embedded platforms.

9) Real-world implementation challenges: Limitations of computational resources and integration of the system with existing financial infrastructures. These practical considerations need to be well assessed to ensure smooth deployment and operation.

10) Deployment claims beyond experimental validation: The proposed framework was evaluated in a controlled experimental environment and was not tested within financial institutions, real-world transaction settings, embedded devices, or end-user environments. Therefore, claims regarding immediate deployment should be interpreted cautiously. The findings demonstrate the potential of the framework for counterfeit currency detection, while practical adoption requires further validation under operational conditions.

VI. Conclusion

Existing solutions for counterfeit money detection have been proved ineffective due to the skillful manipulation of miscreants. This study presents different techniques and methods employed for the detection and identification of counterfeit currency. In addition, it provides a new framework with deep learning models such as DCGAN, CNN, and FCNN. Deep learning methods are preferred due to their ability to find complex relationships in the input data. The primary aim is to train these deep learning models to rightfully classify real and fake currency notes. The endgame is embedded in apps on smart devices for wider deployment. The effectiveness of the system is confirmed in terms of various metrics like accuracy, precision, recall, and F1-score and its flexibility is demonstrated at a low cost by comparing the various models used to assess the efficiency of the study. The results demonstrated that DCGAN outperformed the rest of the models employed in this study and achieved 95% accuracy, 0.95 precision, 0.954 recall, and 0.957 F1-score. The confusion matrix also showed that DCGAN correctly classified 7,802 counterfeit samples and 6,606 genuine samples, outperforming both CNN and FCNN.

VII. Future Work

For further refinement of performance, it is recommended to retrain the models on new instances on a regular basis to further improve the performance and reduce false positives. Note: This study has been addressed for specific currency notes (MYR & SAR) while the proposed system can be applied on different country currency notes provided with ample model training and tuning of parameters. Other future directions could also be:

- Exploring the possible implementation of other security features like watermarking and holograms into the detection framework.
- Extending the framework to other currencies other than MYR and SAR and evaluate the generalizability and applicability around the world.

- Exploring transfer learning and domain adaptation methods to improve the model performance with limited new data instances.

REFERENCES

[1] C. D. Widagdo and H. Firmansyah, "The Impact of Counterfeit Money Circulation on Monetary Stability and The Role of Bank Indonesia in Handling It," *Awang Long Law Review,* vol. 7, no. 2, pp. 333-339, 2025.

[2] A. Mukundan, Y.-M. Tsao, W.-M. Cheng, F.-C. Lin, and H.-C. Wang, "Automatic counterfeit currency detection using a novel snapshot hyperspectral imaging algorithm," *Sensors,* vol. 23, no. 4, p. 2026, 2023.

[3] R. E. McKinney Jr, L. P. Shao, D. C. Rosenlieb Jr, and D. H. Shao, "Counterfeiting in cryptocurrency: An emerging problem," in *Handbook of Digital Currency*: Elsevier, 2024, pp. 169-183.

[4] Walker, "Money and financial technology (FinTech) history," *Int'l Law.,* vol. 56, p. 227, 2023.

[5] F. Niaz, J. Zhang, Y. Zheng, M. Khalid, and A. Niaz, "mm-cur: A novel ubiquitous, contact-free, and location-aware counterfeit currency detection in bundles using millimeter-wave sensor," *ACM Transactions on Sensor Networks,* vol. 20, no. 6, pp. 1-26, 2024.

[6] G. Yu and Z. Luo, "Financial fraud detection using a hybrid deep belief network and quantum optimization approach," *Discover Applied Sciences,* vol. 7, no. 5, p. 454, 2025.

[7] S. Sultana, I. Yasmin, M. J. Patwary, M. A. S. Bhuiyan, I. Salam, and M. H. Miraz, "Generating cgan-based synthetic dermoscopy images and classifying with deep lightweight cnn," in *2023 International Conference on Computing, Networking, Telecommunications & Engineering Sciences Applications (CoNTESA)*, 2023: IEEE, pp. 90-95.

[8] U. Sadyk, R. Baimukashev, and C. Turan, "State-of-the-Art Review of Deep Learning Methods in Fake Banknote Recognition Problem," *International Journal of Advanced Computer Science & Applications,* vol. 15, no. 1, 2024.

[9] A. A. Noor and N. I. R. Ruhaiyem, "Detecting counterfeit currencies using image processing techniques and artificial intelligence," in *AIP Conference Proceedings*, 2025, vol. 3282, no. 1: AIP Publishing LLC, p. 030006.

[10] S. Mathe, J. Kannamma, E. Narmada, C. Rampilla, and S. N. Rehmathunnisa, "Identification of Counterfeit Currency Using Image Processing and Deep Learning," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2024: IEEE, pp. 1-7.

[11] D. Kumar and S. Chauhan, "Indian fake Currency Detection using computer vision," *International Research Journal of Engineering and Technology,* vol. 7, no. 5, pp. 2870-2874, 2020.

[12] C. G. Pachon, D. M. Ballesteros, and D. Renza, "An efficient deep learning model using network pruning for fake banknote recognition," *Expert Systems with Applications,* vol. 233, p. 120961, 2023.

[13] S. Chandrappa, P. Chandra Shekar, P. Chaya, L. Dharmanna, and M. Guruprasad, "Machine learning algorithms for identifying fake currencies," *SN Computer Science,* vol. 4, no. 4, p. 368, 2023.

[14] K. Teymournezhad, H. Azgomi, and A. Asghari, "Detection of counterfeit banknotes by security components based on image processing and GoogLeNet deep learning network," *Signal, Image and Video Processing,* vol. 16, no. 6, pp. 1505-1513, 2022.

[15] A. A. Chowdhury, A. Das, D. Karmaker, and K. K. S. Hoque, "Evaluation of Deep Learning models on UV ink: a Fake Money detection scheme with RPN," in *Proceedings of the 2nd international conference on computing advancements*, 2022, pp. 67-74.

[16] N. F. D. Ajid *et al.*, "Counterfeit one hundred Malaysian ringgit banknotes discrimination using chemical imaging inspection and pattern recognition," *Australian Journal of Forensic Sciences,* vol. 54, no. 5, pp. 695-709, 2022.

[17] F. Antonius, J. Ramu, P. Sasikala, J. Sekhar, and S. S. C. Mary, "DeepCyberDetect: hybrid AI for counterfeit currency detection with GAN-CNN-RNN using African Buffalo optimization," *International Journal of Advanced Computer Science and Applications,* vol. 14, no. 7, 2023.

[18] S. C. D. Bandu, M. Kakileti, S. S. Jannu Soloman, and N. Baydeti, "Indian fake currency detection using image processing and machine learning," *International Journal of Information Technology,* vol. 16, no. 8, pp. 4953-4966, 2024.

[19] A. Nasayreh *et al.*, "Jordanian banknote data recognition: A CNN-based approach with attention mechanism," *Journal of King Saud University-Computer and Information Sciences,* vol. 36, no. 4, p. 102038, 2024.

[20] S. Aulia and A. Rusdinar, "Geometric and grayscale template matching for Saudi Arabian riyal paper currency recognition," *International Journal of Electrical and Computer Engineering (IJECE),* vol. 8, no. 6, pp. 4230-4238, 2018.

[21] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," arXiv preprint arXiv:1511.06434, 2015.

FGHA-Net: Hybrid Attention Learning for Froth Flotation Recovery Prediction in Mineral Beneficiation

Mohammed I Thanoon[1], Fawaz Alanazi[2], Shaaban M. Shaaban[3]*, Asma A. Alhashmi[4],
Abed Saif Ahmed Alghawli[5], Ahlem Fatnassi[6], Majed Nawaz[7], Abdulbasit A. Darem[8]
Computing Department-College of Engineering and Computing at Al-Lith, Umm Al-Qura University, Saudi Arabia[1]
Department of Computer Science-College of Science, Northern Border University, Arar 73213, Saudi Arabia[2, 4, 6, 7]
Center for Scientific Research and Entrepreneurship, Northern Border University, Arar 73213, Saudi Arabia[3, 8]
Computer Science Department-College of Sciences and Humanities,
Prince Sattam Bin Abdulaziz University, Aflaj 11912, Kingdom of Saudi Arabia[5]

***Abstract*—Froth flotation recovery prediction plays an important role in mineral processing because it can improve operational efficiency, reduce mineral loss, and support more effective decision-making during flotation operations. However, accurate recovery prediction remains challenging because flotation processes are highly dynamic and nonlinear, sensor data may contain noise, and process variables often exhibit complex temporal dependencies. Although several machine learning and deep learning approaches have been proposed for flotation prediction, many existing models remain limited in their capabilities of capturing short-term and long-range relationships in sequential flotation process data. To address these challenges, this study proposes a Fine-Grained Hybrid Attention Network (FGHA-Net) for froth flotation recovery prediction. The flotation process data are first preprocessed to improve data quality and suitability for model training. Feature engineering techniques, including process trend extraction, moving average features, and recovery-rate difference analysis, are then applied to enhance the representation of process dynamics. The proposed FGHA-Net consists of a bidirectional long short-term memory (BiLSTM) branch and a Transformer branch for hybrid temporal feature learning. The BiLSTM branch captures short-term sequential dependencies, whereas the Transformer branch models long-range temporal relationships using multi-head self-attention. A Fine-Grained Attention Fusion module is further introduced to support cross-branch feature fusion, temporal attention weighting, and refinement of important process variables. Finally, the fused features are passed to a regression head to generate the predicted recovery value. The proposed model is trained using the Adam optimizer and evaluated using standard regression metrics. Experimental results show that FGHA-Net achieved strong predictive performance on the testing set, with an RMSE of 0.011484 and an R^2 score of 0.844612. These results indicate that the proposed model can effectively learn complex temporal dependency patterns from froth flotation process data and has potential for intelligent mineral-processing monitoring and decision-support applications.**

Keywords*—Froth flotation; mineral beneficiation; iron-ore processing; recovery prediction; soft sensor; BiLSTM; transformer; fine-grained attention; deep learning

I. Introduction

In mineral processing, froth flotation is one of the most widely used separation methods. Froth flotation is considered one of the most important and complex processes [1]. In froth flotation, concentrate grade plays a significant role in production performance. Online grade prediction is therefore essential for precise process control and effective optimization of industrial operations [2]. Because the direct measurement of key flotation variables often requires expensive instrumentation to acquire and maintain, soft-sensor modeling provides a fast and cost-effective approach for predicting concentrate grade based on variables measured online [3]. The predicted values can then be used as inputs to feedback control systems that adjust process variables to improve flotation performance. However, modeling froth flotation remains challenging because of its complex dynamics and strong interactions among process variables [4]. Existing models range from empirical approaches, such as artificial neural networks (ANNs), to phenomenological models, including kinetic, probabilistic, and population-balance models. Despite these approaches, no single model can fully capture the entire process dynamics, which often leads to considerable control challenges [5]. Data-driven approaches can identify hidden patterns and model nonlinear interactions; however, they still face several limitations. Machine learning (ML) involves the development of algorithms that enable systems to learn patterns from data and generate predictions without explicit programming [6, 7].

Effective preprocessing improves data quality and enables the model to capture relevant patterns and relationships [8]. Selecting an appropriate ML method is also essential because it strongly influences prediction accuracy, interpretability, and generalization ability. Different ML approaches have distinct strengths and limitations; therefore, choosing a suitable model allows the data to be represented more effectively and used for better decision-making [9]. Testing and validation are important for assessing model performance and ensuring that the model can generalize beyond the training samples [10]. The growing influence of ML and artificial intelligence (AI) is increasingly evident across various fields, from everyday applications to advanced scientific domains [11]. With rapid adoption in

*Corresponding author.

applications such as voice recognition, facial recognition, autonomous vehicles, social-media platforms, and spam filtering, these technologies are now widely used on a global scale [12]. In recent decades, researchers have identified several challenges in mineral beneficiation and have proposed potential solutions using ML- and AI-based models [13].

However, accurate froth flotation recovery prediction remains challenging because of process complexity, sensor noise, and strong dependencies among flotation variables. Existing ML and deep learning models also struggle to capture short-term and long-range temporal relationships effectively. Therefore, an efficient hybrid model is needed to improve flotation recovery prediction. Motivated by these challenges, this study proposes a Fine-Grained Hybrid Attention Network (FGHA-Net) for froth flotation recovery prediction. The proposed model combines BiLSTM and Transformer architectures to capture sequential temporal features and long-range dependencies from flotation process data. Additionally, a Fine-Grained Attention Fusion module is introduced to improve cross-branch feature fusion and important process-variable refinement. The contributing factors of this study are summarized as follows:

- This study integrates a bidirectional long short-term memory (BiLSTM) branch and a Transformer branch to capture short-term and long-range temporal dependencies from flotation process sequences.
- It introduces a Fine-Grained Attention Fusion (FGAF) module for effective temporal attention weighting and cross-branch feature fusion.
- It uses a regression head consisting of dense and linear output layers to generate the froth flotation recovery prediction output.
- It trains the proposed FGHA-Net model using the Adam optimizer to improve prediction performance and model convergence.

II. Related Work

This section presents a concise overview of existing studies on froth flotation recovery prediction and identifies the associated research gaps. Pandian et al. [14] proposed a deep probabilistic neural network technique to predict silica concentrate percentage and iron concentrate percentage in flotation processes. Their approach used a comprehensive flotation dataset that included reagent dosages, feed composition, column operating conditions, and pulp characteristics. Several machine learning models were evaluated, and a hybrid method combining a probabilistic pattern neural network with an Extra Trees regressor was proposed to improve prediction accuracy and effectiveness.

Zhang et al. [15] proposed an interpretable grade prediction model, which captures representative features from multi-source XRF sequential data. In addition, tensor reforming was applied to support efficient parallel evaluation. In [16], a convolutional long short-term memory neural network was proposed for real-time monitoring of chemical composition grades. The proposed implementation framework enables real-time monitoring and supports the adjustment of flotation parameters to improve process efficiency.

Yang et al. [17] investigated a coal flotation froth image segmentation method based on Mask R-CNN. Abouda et al. [18] proposed the AI-DeepFrothNet framework to address major limitations in froth image analysis. The framework used a Putrefaction Enrichment and Tuning Network (PETNet) to remove and reduce noise in froth images. It also applied multi-agent deep Q-learning, in which three agents were used to investigate different properties of the froth layer. In [19], a Gaussian Process-based Model Predictive Control algorithm was designed to incorporate uncertainty quantification directly into the control architecture. By exploiting the probabilistic nature of Gaussian Process modeling, this technique captures process variability and adapts dynamically to new information, enabling continuous refinement of the Gaussian Process model within the MPC framework. The Gaussian Process model was trained and optimized using JAX, evaluated with relevant performance metrics, and used as a surrogate model in the MPC approach.

Song et al. [20] introduced a prediction algorithm based on an improved XGBoost model using real copper concentrate flotation data. To address missing values and outliers in the collected dataset, the authors first proposed an outlier detection method. The integration of MissForest/IQR preprocessing with TPE-optimized XGBoost formed a sequential prediction pipeline for estimating copper concentrate grade in the flotation process. Bendaouia et al. [21] applied CNN-based feature extraction and image-processing techniques integrated with DL to predict the elemental composition of minerals. It included several machine learning methods, CNN models, and ANN models for predicting mineral concentrate grades. An overview of related studies on froth flotation prediction is presented in Table I.

Although several machine learning and deep learning methods have been introduced for flotation prediction and monitoring, several research gaps remain. Traditional regression and ML approaches have restricted ability in capturing complicated nonlinear and temporal dependencies, while existing DL models often inadequately model long-range relationships and complementary temporal representations. Moreover, limited attention has been given to adaptive feature interaction and fusion. To address these gaps, FGHA-Net integrates BiLSTM-based temporal feature learning, Transformer-based long-range dependency modeling, and fine-grained attention fusion to effectively combine complementary representations for improved flotation recovery prediction. Some image-based techniques mainly focus on froth image analysis; however, they do not fully exploit sequential process information for recovery prediction. Moreover, several recent approaches remain sensitive to noise, provide limited feature extraction capability, and are unable to capture both short-term and long-range process dependencies simultaneously. Therefore, an effective hybrid deep learning approach is still needed to learn temporal process patterns more efficiently and improve froth flotation recovery prediction performance.

TABLE I. SUMMARY OF RELATED WORKS IN FROTH FLOTATION PREDICTION

Reference	Aim	Methods Used	Input Data	Key Idea
Pandian et al. [14]	To predict silica and iron concentrate quality using probabilistic learning and ensemble-based regression methods	Deep Probabilistic NN and Extra Trees	Flotation process variables (reagents, pulp, operating conditions)	Hybrid probabilistic and ensemble learning
Zhang et al. [15]	To improve mineral grade prediction using interpretable RNN with shared parameters across grouped time series data	Grouped RNN with weight sharing	Multi-source XRF time series	Interpretable grouped sequential learning
Bendaouia et al. [16]	To monitor chemical composition grades by capturing both spatial and temporal patterns from flotation data	ConvLSTM	Video-based flotation data	Spatio-temporal feature learning
Yang et al. [17]	To extract froth bubble features and study their relationship with ash content in the flotation process	Mask R-CNN	Froth flotation images	Instance segmentation of froth bubbles
Abouda et al. [18]	To improve froth flotation monitoring by removing noise and analyzing froth conditions using deep reinforcement learning	DeepFrothNet + MA-DQL	Froth image data	Noise removal and reinforcement learning-based analysis
Dehon et al. [19]	To enhance flotation process control using uncertainty-aware Gaussian Process-based predictive control	Gaussian Process MPC	Process control data	Probabilistic model-based control
Song et al. [20]	To predict copper concentrate grade using an optimized XGBoost model with improved hyperparameter tuning	Improved XGBoost (TPE optimized)	Copper flotation dataset	Optimized tree-based prediction model
Bendaouia et al. [21]	To predict mineral concentrate grades using CNN-based feature extraction combined with machine learning models	CNN + ML hybrid	Mineral flotation dataset	Feature extraction from industrial flotation data

III. METHODOLOGY

The proposed model follows a systematic pipeline for froth flotation recovery prediction. The overall structure of the proposed framework is shown in Fig. 1. In the initial stage, data preprocessing is performed through missing-value handling, outlier removal using the interquartile range (IQR) method, Min–Max normalization, temporal sequence generation, and sliding-window formation. Next, feature engineering is conducted using process trend extraction, moving average features, and recovery-rate difference analysis. The proposed FGHA-Net then applies a dual-branch architecture consisting of a BiLSTM branch and a Transformer branch. A Fine-Grained Attention Fusion module integrates features from both branches by weighting and refining important process information. Finally, the fused features are passed to a regression head to predict flotation recovery, and the model is trained with the Adam optimizer.

A. Data Preprocessing

Data preprocessing is an important stage that improves data quality and prepares the dataset for model training. It includes missing-value handling, outlier removal, normalization, and temporal sequence generation [22].

Missing-value handling: A robust preprocessing pipeline is required to handle missing values in machine learning datasets. This step involves examining raw froth flotation process data to identify and correct missing values and inconsistencies. Removing low-quality values helps ensure that model performance is not affected by irrelevant, incomplete, or erroneous data.

Outlier removal using the IQR method: The interquartile range (IQR) method is used to identify and remove abnormal process values from the dataset. This method helps reduce the effect of extreme fluctuations and improves the consistency of the prediction model.

Data normalization: Normalization scales all attribute values into a specific range. Because feature values often vary across process variables, unscaled data may lead to unstable or inefficient training of deep learning models. Therefore, Min–Max normalization is applied to transform all values into the range between 0 and 1, which can accelerate model convergence and improve prediction performance, as shown in Eq. (1).

$$X' = \frac{X - \text{Min}}{\text{Max} - \text{Min}} \quad (1)$$

In Eq. (1), X′ represents the normalized feature value, while Min and Max denote the minimum and maximum values of the corresponding feature, respectively. Temporal sequence generation: The normalized flotation process variables are converted into temporal sequences to capture dependency relationships among process variables over time.

Sliding-window formation: A sliding-window method is used to divide the temporal sequences into fixed-length input windows for model training. This strategy enables the proposed FGHA-Net model to learn temporal dependency patterns from flotation process data.

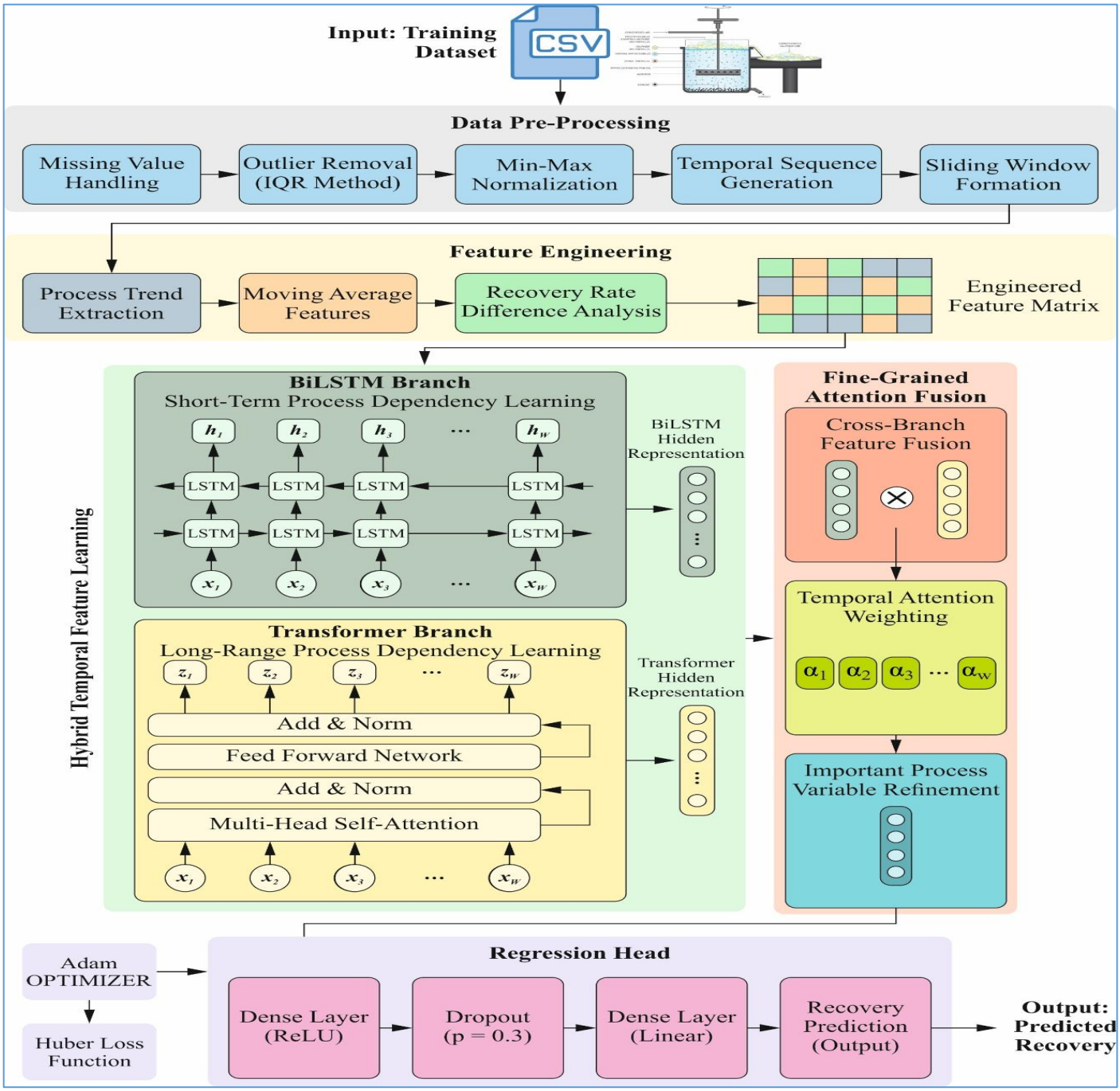


Fig. 1. Structural design of proposed framework.

B. Feature Engineering

Feature engineering is performed to improve the learning capability of the proposed FGHA-Net framework by extracting meaningful information from the froth flotation process data. Process trend extraction is used to capture the temporal variation of flotation process variables over time. These trend features help the model learn sequential dependency patterns and process changes. Moving average features are generated to reduce noise and smooth variations in flotation process sequences. These features help capture more stable temporal patterns from process variables. Recovery-rate difference analysis is performed to calculate the differences between consecutive recovery values and identify variability patterns in flotation recovery behavior.

C. Proposed FGHA-Net Architecture

The proposed hybrid temporal feature learning module is designed to capture both short-term and long-range dependencies in flotation process data. It integrates BiLSTM and Transformer branches to extract sequential and global temporal features effectively.

1) BiLSTM branch: The LSTM network consists of several key components, including the forget gate, input gate, output

gate, and memory cell [23]. These components allow the network to retain relevant information and discard less important information over time. The LSTM network includes four structures: forgetting gate (f_i), memory unit, output gate (o_t), and input gate (i_i). The LSTM network holds additional cell states with triple gate elements in comparison with the RNN. In an LSTM structure, the forget gate (f_t) determines what percentage of data to retain in the network, and it is computed by:

$$f_t = \sigma\left(W_f \cdot [h_{i-1'} x_t] + b_f\right) \quad (2)$$

In this equation, h_{i-1} implies the preceding state memory, x_t represents the input sequence,· iV_f points out the weighted matrix of the forgetting gate, σ implies the sigmoid activation function, f_i signifies the state of the forget gate, and b_f denotes the bias of the forget gate. The input gate i_i is responsible for remembering the novel data in the cell state.

$$i_t = \sigma\left(W_j \cdot [h_{i-1'} x_t] + b_j\right) \quad (3)$$

$$\overline{C} = \tanh\left(W_c \cdot [h_{i-1}, x_t] + b_c\right) \quad (4)$$

$$C_i = f_i * C_i + i_i * \overline{C} \quad (5)$$

In these equations, C_i denotes the novel cell state, $\overline{C}$ refers to the cell state candidate, W_j denotes the weighted matrix of the input gate, tanh represents the hyperbolic tangent function, b_i signifies the bias of the input gate, iV_c implies the weighted matrix of the cell state, i_i specifies the input gate, and b_c shows the biased cell state, o_t denotes the output gate.

$$0_t = \sigma(|v_0 \cdot [h_{i-1'} x_t] + b_0) \quad (6)$$

$$h_t = 0_t * \tanh\left(C_t\right) \quad (7)$$

Here, (b_o) implies the bias of the output gate and (W_0)denotes the weighted matrix of the output gate.

This study uses a bidirectional LSTM (BiLSTM) branch, which processes sequence data in both forward and backward directions. The forward and backward hidden layers capture temporal dependencies from both past and future contexts. As a result, the BiLSTM branch extracts more informative sequential features and short-term process patterns, thereby improving the prediction accuracy of froth flotation recovery.

2) Transformer branch: The transformer branch is developed to obtain the global contextual data from sequences of the froth flotation method [24]. The transformer framework is constructed by the sequence embedding and multi-head self-attention mechanisms (MHSA) for encoding the temporal process data, and adaptably handling various measures of input signals, while modifying the computational complexity in the linear ranges. Thus, this branch is utilized as the encoder part for temporal feature extraction and long-term dependency learning. Likewise, the BiLSTM branch and the transformer module utilize the hierarchical feature extraction model to split the entire branches into numerous layers. At the 1st stage, the input processed sequences are primarily separated into temporal embeddings. Later, these embeddings can be entered into the transformer encoder model to obtain the 1st phase feature representations that are transferred to the following phase. The following feature representations are more effectively handled by employing the transformer encoder layers. In conclusion, the multi-scale temporal feature representations are gained and enable interrelationship amongst the layers of these two subdivisions.

To mitigate the higher computation complexity related to the self-attention mechanism (SAM) in transformer encoder parts, the transformer branch utilizes the MHSA to proficiently learn the dependency relations between the process variables. Same as the conventional attention mechanism (AM), the MHSA gains the value (V), query (Q), and key (K) as input, and generates improved temporal features.

$$MHSA\ (Q, K, V) = \text{Concat}\ (head_1, \ldots, head_n) W^O \quad (8)$$

$$head_i = \text{Attention}\ (QW_i^Q, KW_i^K, VW_i^V \quad (9)$$

Here the term $(head_i)$ has the attention value $i - th$ attention head, the term Concat signifies the channel splicing function, and the parameters W_i^Q, W_i^V, and W_i^K indicate the linear projection parameters. The mathematical form of the attention operation can be measured by

$$\text{Attention}\ (Q, K, V) = Soft\ \max\left(\frac{QK^T}{\sqrt{d_k}}\right) V \quad (10)$$

Now, the factor (d_k) points to the key vector's dimensionality. By applying the MHSA module, the transformer branch successfully captures the long-term dependency correlation amongst flotation process variables, thus improving the ability of temporal representation and increasing the prediction efficiency of the proposed framework for froth flotation recovery prediction.

3) Fine-grained attention fusion: In the proposed framework, the cross-branch feature fusion includes convolution fusion, operator fusion, and attention fusion approaches [25]. The convolutional fusion approaches improve the cross-branch temporal features with respect to the semantic and multi-scale hierarchies by employing different convolution operations, decreasing the noise interference in the temporal features. Simple operator fusion techniques are directly added or concatenated to the features of the transformer and BiLSTM for fusion. The noise in the temporal features can be simply reduced by gaining dependable prediction outcomes by employing these techniques. Generally, the attention fusion methods utilize an AM to efficiently combine the cross-branch features in the temporal and channel dimensions. The convolutional techniques are focused on improving temporal features before fusion, whereas attention refinement techniques concentrate on boosting these features.

The FGAF model efficiently fuses features by employing the temporal attention data. The important part of the features is identified by employing temporal and channel attention, and the major parts amongst the BiLSTM and transformer branch features have been recognized by employing the temporal data. The channel branch accomplishes the combination of channel

data through the transmission of input cross-branch features over the channels by employing global average pooling (GAP). The aggregation method is denoted by

$$fc = Concat(Avg(B_{\{input\}}, \text{Max}\left(B_{\{input\}}, Avg(T_{input}), \text{Max}\left(T_{input}\right)\right) \quad (11)$$

From this mathematical form, the parameter fc signifies the features from the fused channel. The Concat represents concatenation with the channel dimension. The GAP has been indicated by Avg Max ($B_{\{input\}}$ and T_{input}), which are the feature inputs of the BiLSTM and transformer branches, respectively.

The fused channel features can be entered into the 1D convolutions with effectively calculated kernel numbers (k) and efficiently obtain dependency among the channels. The formation of attention weights W_{c1} and W_{c2} in the channels employing the non-linear and softmax activation functions.

$$W_{c1}, W_{c2} = \frac{e^{Conv_1(f_c)}}{e^{Conv_1(f_c)} + e^{Conv_2(f_c)}} \frac{e^{Conv_2(f_c)}}{e^{Conv_1(f_c)} + e^{Conv_2(f_c)}} \quad (12)$$

Here, the terms $Conv_1$ and $Conv_2$ point to 1D convolution. This technique is employed for calculating the weights in the temporal and channel dimensions. Global pooling has been employed in the temporal dimension, and weights are acquired by the non-linear and convolution activation function. At the temporal branch, the significant method variable regions among the cross-branch features are computed by computing the temporal attention weights ($W_{\{T1\}}$) and ($W_{\{T2\}}$) parameters. Through the combination of temporal attention weights and cross-branch channel weights, the wide-ranging feature weights have been gained, thus emphasizing the vital parts amongst cross-branch features. Later, such feature weights can be multiplied by the cross-branch features for efficient fusion of these features. The final output must be handled by a 1×1 convolution and a convolution operation with the summation of the original inputs. The mathematical form of output is determined by:

$$f_{Boutput} = \text{f}^{conv}(f^{1\times1}(W_{C1} + W_{t1}) \otimes f_{Binput})) + f_{Binput} \quad (13)$$

$$f_{Boutput} = \text{f}^{conv}(f^{1\times1}(W_{C2} + W_{t2}) \otimes f_{Tinput})) + f_{Tinput} \quad (14)$$

From the above expression, the mathematical notation $\otimes$ denotes the multiplication operation, the parameter $f^{1\times1}$points to the set of ReLU activation functions, 1×1 convolutions, and batch normalization (BN), and the term f^{conv} refers to the convolution operation. By employing the proposed FGAF model, the model successfully performs temporal attention weighting, cross-branch feature fusion, and significant method variable improvement, thus increasing the prediction performance of the proposed FGHA-Net framework for froth flotation recovery prediction.

4) Regression head: A regression head is developed for froth flotation recovery prediction based on the FGHA-Net framework [26]. A dense layer with dropout regularization of 0.3 is fully connected using a regression head. This design choice stems from the necessity to effectively extract the refined temporal characteristics captured from the Transformer branch and BiLSTM branch.

This framework begins with a dense layer constructed to allow the fused temporal feature extraction. To reduce overfitting and improve the generalization capability, the extracted feature vector (h) is crossed over the dropout layer and dense layer. For the technique to operate a regression process for recovery prediction, a linear output layer can be applied to modify the last dense layer in Eq. (15). ($b_{\{out\}}$) and ($W_{\{out\}}$) indicates the bias and weight of the output layer.

$$y = W_{out}.h + b_{out} \quad (15)$$

Here (y) denotes the output of predicted froth flotation recovery. The regression head comprises trainable parameters related to the output and dense layers. This framework achieves the balance between computational efficiency and method intricacy. Within these integrated elements, the proposed FGHA-Net algorithm effectively functions to recover prediction outcomes for froth flotation process information. Fig. 2 represents the architecture of the hybrid temporal feature learning architecture combining BiLSTM and Transformer branches with fine-grained attention for prediction recovery.

D. Training Strategy

The proposed FGHA-Net model is trained using the Adam optimizer, which is an efficient stochastic optimization algorithm. Adam adaptively estimates learning rates for model parameters by combining momentum and adaptive gradient estimation [27]. This allows the optimizer to achieve stable and efficient convergence during deep learning model training.

In this study, the proposed FGHA-Net model is trained using Adam with a learning rate of 1e−4. Huber loss is used as the loss function because it is suitable for regression tasks and is less sensitive to outliers than mean squared error. The model is trained with a batch size of 32, and the dataset is divided into training and testing sets using a 70:30 split. The use of Adam optimization and Huber loss supports stable convergence and improves the robustness of the proposed recovery prediction model. This optimization technique is expressed as follows:

$$x_t = \delta_1 * x_t - 1 - (1 - \delta_1) * g_t \quad (16)$$

$$y_t = \delta_2 * y_t - 1 - (1 - \delta_2) * g_t^2 \quad (17)$$

$$\triangle \omega_t = -\eta \frac{x_t}{\sqrt{yt + \epsilon}} * gt \quad (18)$$

$$\omega_{t+1} = \omega_t + \triangle \omega_t \quad (19)$$

Here, the parameter g_t implies the gradient with respect to time (t), the symbol (θ) denotes the learning rate, the term x_t signifies the exponential averages of gradients, and the term y_t indicates the exponential averages of gradients squared, whereas δ_1 and δ_2 imply the factors.

To improve the strength of the predictive method, the proposed model was trained using the Adam optimizer with Huber loss as the loss function. The proposed system can be trained utilizing a batch size of 32 over 100 epochs. In addition, the database is separated using a 70:30 train-test split for performance evaluation and training. This method has revealed

higher outcomes for regression-driven predictive tasks owing to the steady convergence action and adaptive parameter upgrading ability. These advantages make the Adam optimizer appropriate for the proposed FGHA-Net model for the froth flotation recovery predictive system.

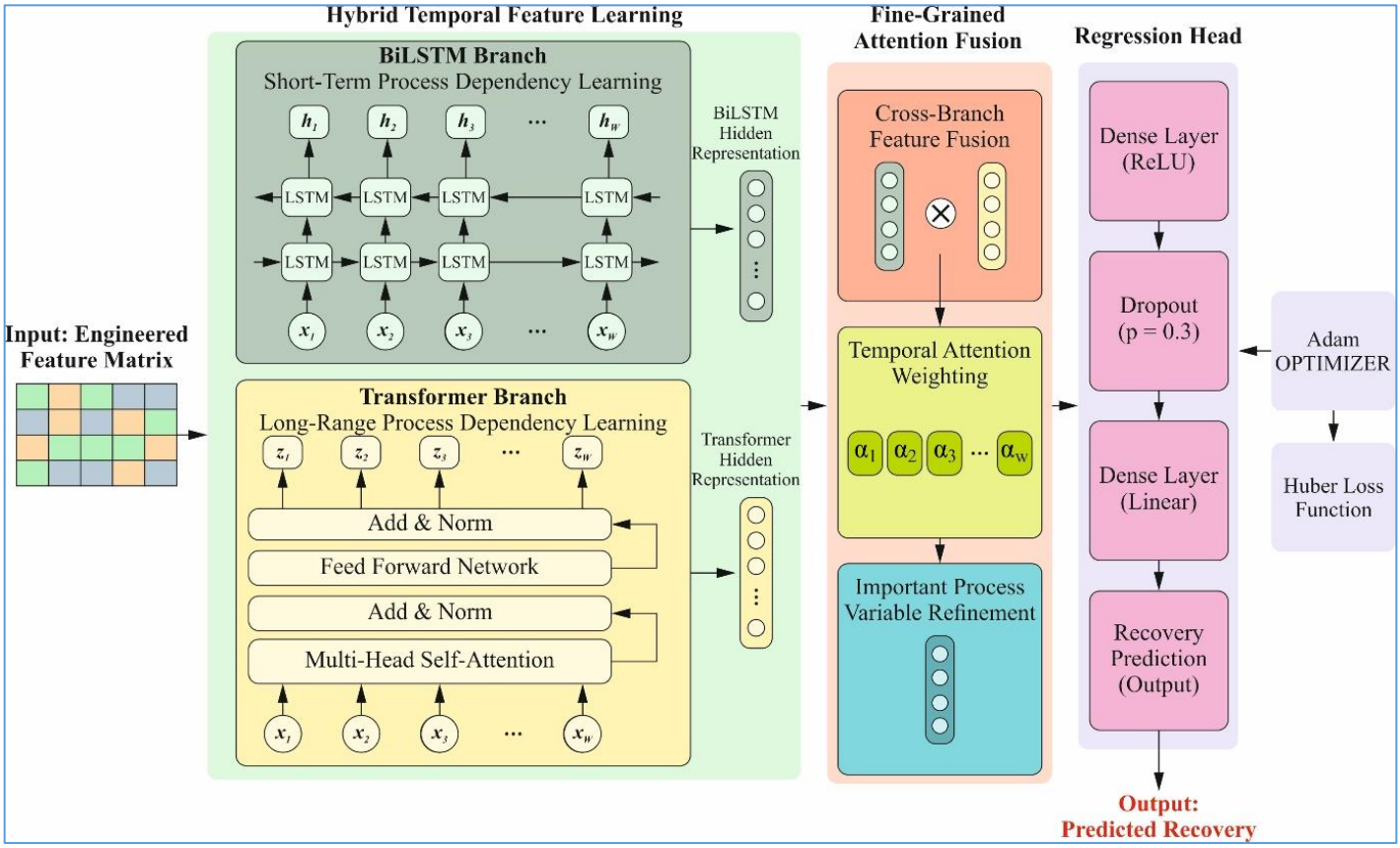


Fig. 2. Hybrid temporal feature learning architecture combining BiLSTM and Transformer branches with fine-grained attention for recovery prediction.

IV. Experimental Results and Discussion

The experimental evaluation of the proposed FGHA-Net model was conducted using a benchmark froth flotation dataset [28]. The dataset was collected from an industrial froth flotation plant. Column 1 contains date and time information, covering the period from March 2017 to September 2017. Columns 2 and 3 represent quality measurements of the iron ore pulp before flotation. Columns 4 to 8 include important variables that influence final ore quality. Columns 9 to 22 contain process variables, such as air flow and flotation-column level, which also affect ore quality. The last two columns represent laboratory-based measurements of the final iron ore pulp quality.

The dataset was divided into training and testing sets using a 70:30 ratio. Each input feature was normalized using Min–Max scaling. Temporal sequences were generated using a sliding-window technique, and the proposed FGHA-Net model was evaluated using standard regression metrics, including MAE, RMSE, MAPE, and R^2 score. The performance of the proposed model was also compared with several conventional machine learning and deep learning methods to evaluate its effectiveness.

Table II summarizes the parameter configuration of the FGHA-Net model. Feature engineering was performed using process trend analysis, a moving average window of 5, consecutive recovery-difference analysis, and 12 engineered features. The BiLSTM branch consisted of two layers, a dropout rate of 0.2, and 128 hidden units. The Transformer branch used four layers, eight attention heads, an embedding dimension of 128, and a feed-forward dimension of 256 with a multi-head self-attention mechanism. The Fine-Grained Attention Fusion module applied a cross-branch attention fusion strategy with 128 hidden units and an attention dropout rate of 0.2. The regression head consisted of dense layers with 128 and 64 neurons, followed by a linear output layer for recovery prediction. The model was trained using the Adam optimizer with a learning rate of 1e−4, Huber loss, a batch size of 32, and 25 epochs.

TABLE II. Parameter Settings

Model	Parameter	Values
Feature Engineering	Trend Extraction Method	Process Trend Analysis
	Moving Average Window	5
	Recovery Difference Metric	Consecutive Recovery Difference
	Engineered Feature Count	12
BiLSTM Branch	No. of BiLSTM Layers	2
	Hidden Units	128
	Dropout Rate	0.2

	Temporal Dependency Learning	Short-Term Sequential Learning
Transformer Branch	No. of Transformer Layers	4
	Attention Heads	8
	Embedding Dimension	128
	Feed Forward Dimension	256
	Self-Attention Type	Multi-Head Self-Attention
Fine-Grained Attention Fusion	Fusion Strategy	Cross-Branch Attention Fusion
	Attention Hidden Units	128
	Attention Dropout	0.2

	Feature Refinement	Temporal Attention Weighting
Regression Head	Dense Layer Structure	128 → 64
	Dropout Rate	0.3
	Output Layer	Linear
	Prediction Task	Recovery Rate Prediction
Training Configuration	Optimizer	Adam
	Learning Rate	1e-4
	Loss Function	Huber Loss
	Batch Size	32
	Epochs	25

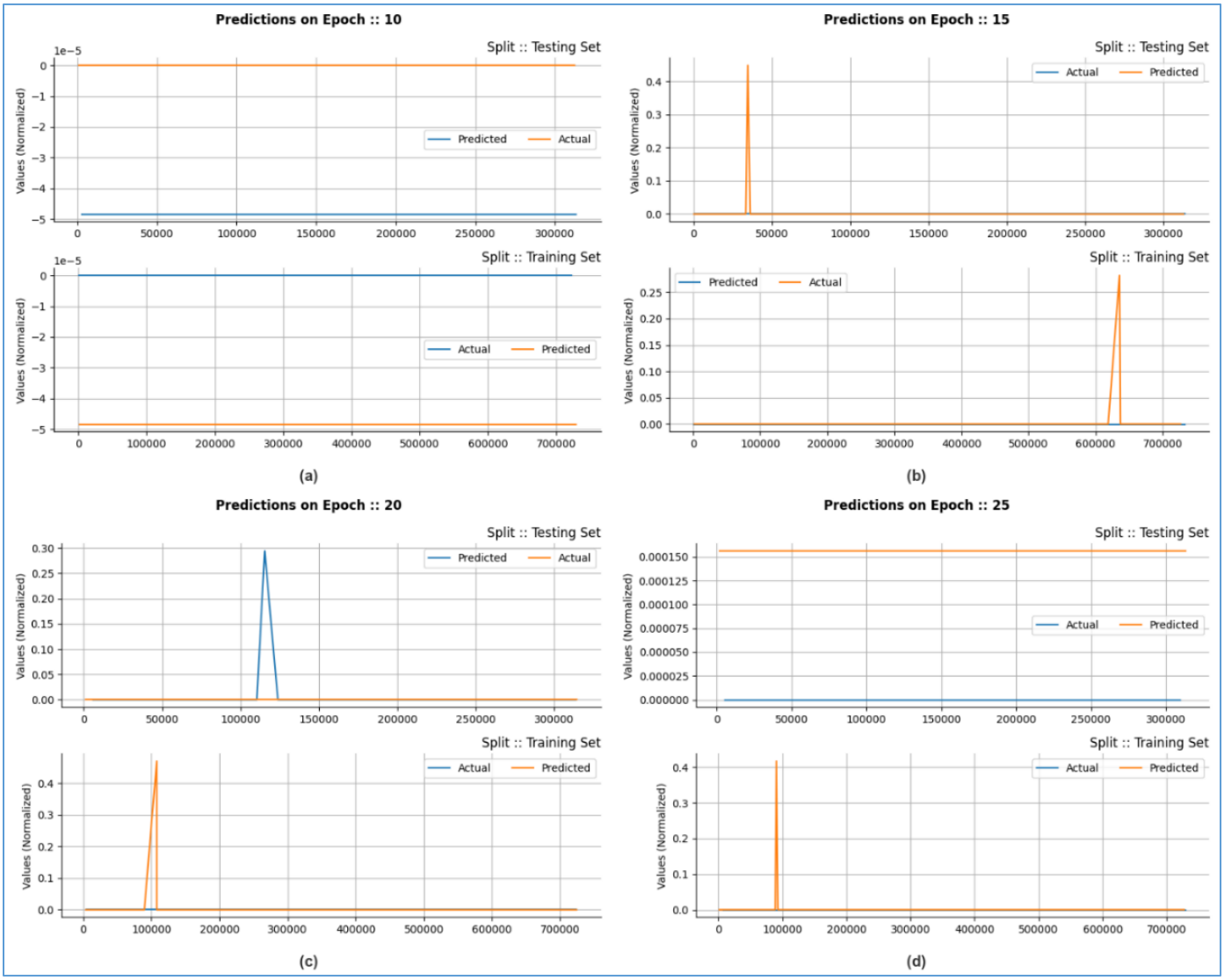


Fig. 3. Predictive analysis of the FGHA-Net method with different epochs.

Fig. 3 illustrates the prediction performance of the FGHA-Net model over different numbers of epochs, including 10, 15, 20, and 25 epochs, on both the training and testing sets. In the early epochs, the predicted and actual values show noticeable differences. However, as the number of epochs increases, the predicted values become more stable and show closer agreement with the actual values. These results indicate that the proposed model achieves stable convergence and improved prediction accuracy as training progresses.

Fig. 4 presents the loss-curve analysis of the FGHA-Net model on the training and testing sets over 25 epochs. The gradual reduction in MSE and RMSE indicates improved convergence and reduced prediction error. The MAE and MAPE results also demonstrate stable predictive performance. In addition, the increasing R^2 score indicates better model fitting and a stronger ability to learn flotation process behavior.

Table III and Fig. 5 present the performance evaluation of the FGHA-Net model using the training and testing sets. The model demonstrated consistent learning behavior, reduced prediction error, and stable convergence. On the testing set, the model achieved an MSE of 0.000132, RMSE of 0.011484, MAE of 0.000718, MAPE of 7.072856, and R^2 score of 0.844612. These results show strong agreement between the predicted and actual values and confirm the ability of the proposed framework to learn flotation process behavior effectively.

TABLE III. PERFORMANCE EVALUATION OF THE FGHA-NET MODEL WITH TRAIN-TEST SPLIT

Metrics	Training Set	Testing Set
MSE	0.000118	0.000132
RMSE	0.010878	0.011484
MAE	0.000683	0.000718
MAPE	7.006903	7.072856
R^2 Score	0.852408	0.844612

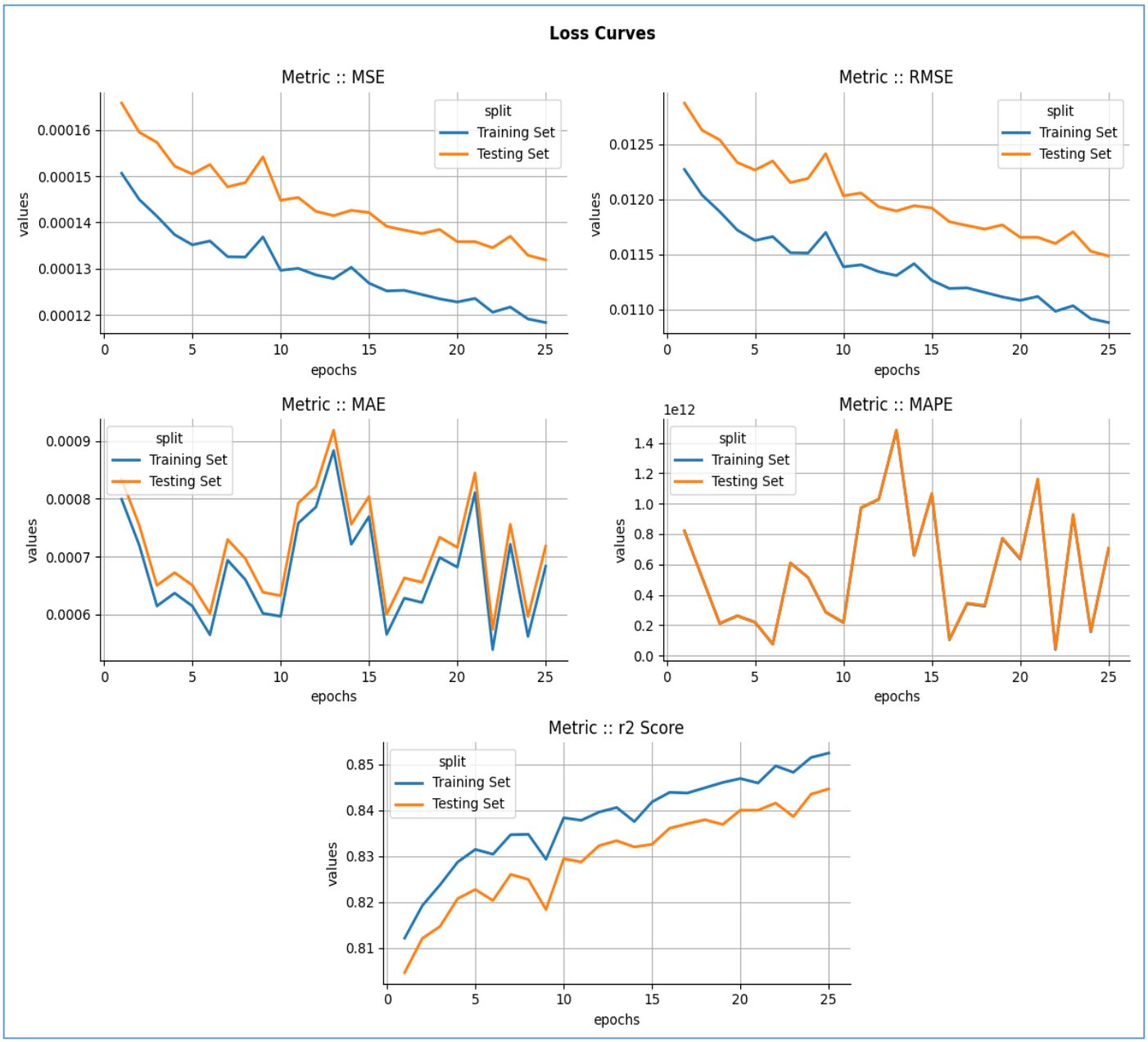


Fig. 4. Loss curve analysis of the FGHA-Net model.

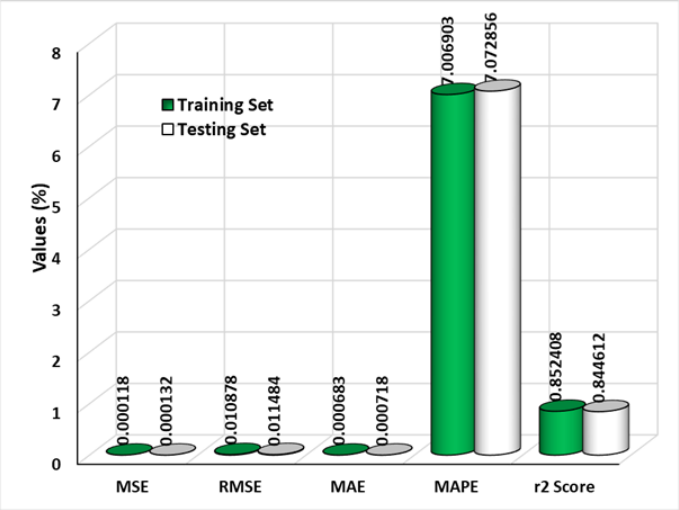


Fig. 5. Performance evaluation of the FGHA-Net model with training set and testing set.

Table IV presents the comparative assessment of the FGHA-Net model against existing methods for flotation recovery prediction [29-31]. Conventional models, such as Linear Regression, SVM, and M5P, achieved lower R^2 scores, indicating limited ability to capture nonlinear relationships. Other models, including FFNN, MLP-ANN, and Bi-PINN, showed moderate performance but still produced higher prediction errors than the proposed model. Concurrently, the hybrid models such as GA-NN, hybrid GA-NN, and BPN models have reached reasonable performance. But the proposed FGHA-Net outperforms existing models with minimal RMSE = 0.0109, MAE = 0.0007, MAPE = 7.0069, and R^2 = 0.8524, demonstrating better predictive performance than the compared models.

TABLE IV. COMPARISON ASSESSMENT OF THE FGHA-NET TECHNIQUE WITH EXISTING ALGORITHMS

Algorithms	RMSE	MAE	MAPE	R^2 Score
ANN	0.3120	0.0050	4.5900	0.4740
GA-NN	0.4000	0.0026	2.8800	0.7880
BPN	0.2590	0.0043	2.8700	0.2830
Hybrid-GA-NN	0.0568	0.0073	8.1200	0.5053
Linear Regression	0.0568	0.0073	8.1200	0.5053
FFNN	0.0552	0.0040	9.4380	0.6678
LSTM	0.0724	0.0025	13.7930	0.5662
Bi-PINN	0.0441	0.0069	11.8860	0.5024
MLP-ANN	0.0592	0.0043	8.8760	0.6143
SVM	0.0334	0.0058	12.1320	0.4621
RF-FFA	0.0825	0.0090	10.4920	0.5443
M5P	0.0229	0.0068	10.7300	0.6433
FGHA-Net	0.0109	0.0007	7.0069	0.8524

Table V provides the ablation study of the proposed model. The table values define the significance of each component involved in the proposed model. The experimental values imply that the proposed FGHA-Net (BiLSTM + Transformer + Fine-Grained Attention Fusion + Adam) gains enhanced performance over individual components.

Table VI portrays the comparative computational performance analysis of the proposed model in terms of FLOPs, GPU memory consumption, and inference time. The experimental outcomes demonstrate that the FGHA-Net accomplishes minimal computation requirements with 0.09 G FLOPs, 0.04 GB GPU memory, and an inference time of 0.20 ms, demonstrating its computational efficiency and suitability for real-time predictive applications compared with the baseline models.

TABLE V. ABLATION STUDY OF PROPOSED MODEL

Algorithms	RMSE	MAE	MAPE	R^2 Score
BiLSTM only	0.0141	0.0039	10.2269	0.5234
Transformer only	0.0135	0.0033	9.6869	0.5784
BiLSTM + Transformer	0.0128	0.0025	9.0169	0.6494
BiLSTM + Transformer (Without Fine-Grained Attention Fusion)	0.0120	0.0020	8.2769	0.7214
BiLSTM + Transformer + Fine-Grained Attention Fusion (Without Adam)	0.0115	0.0015	7.7569	0.7834
FGHA-Net (BiLSTM + Transformer + Fine-Grained Attention Fusion + Adam)	0.0109	0.0007	7.0069	0.8524

TABLE VI. COMPUTATIONAL EFFICIENCY ANALYSIS OF PROPOSED WITH EXISTING MODELS

Algorithm	FLOPs (G)	GPU Memory (GB)	Inference Time (ms/image)
ANN	0.24	0.35	6.8
GA-NN	0.41	0.48	8.4
BPN	0.32	0.41	7.5
Hybrid-GA-NN	0.68	0.72	11.3
Linear Regression	0.21	0.15	1.7
FFNN	0.32	0.43	7.8
LSTM	1.92	1.34	18.9
Bi-PINN	2.76	1.76	24.6
MLP-ANN	0.39	0.52	9.2
SVM	0.17	0.18	2.8
RF-FFA	0.15	0.29	4.3
M5P	0.13	0.08	1.2
FGHA-Net	0.09	0.04	0.20

The experimental results indicate that the proposed FGHA-Net model achieves lower error values and a higher R^2 score than the existing methods. This suggests that the predicted values are closer to the actual recovery values. The improvement can be attributed to the ability of the hybrid BiLSTM–Transformer architecture to capture both short-term and long-range temporal dependencies in flotation process data. The actual-versus-predicted plots further show that the proposed model captures variations in recovery more accurately than the compared

methods. Overall, the results demonstrate that FGHA-Net is a promising approach for froth flotation recovery prediction.

V. Conclusion and Future Work

This study presented FGHA-Net, a Fine-Grained Hybrid Attention Network for froth flotation recovery prediction. The proposed approach first preprocesses the data to improve its quality and suitability for model training. Feature engineering techniques, including process trend extraction, moving average features, and recovery-rate difference analysis, are then applied to enhance the representation of flotation process dynamics. The FGHA-Net architecture consists of a BiLSTM branch and a Transformer branch for hybrid temporal feature learning. The BiLSTM branch captures short-term sequential dependencies, whereas the Transformer branch models long-range relationships using a multi-head self-attention mechanism. In addition, the Fine-Grained Attention Fusion module performs cross-branch feature fusion, temporal attention weighting, and important process-variable refinement. The fused features are then passed through a regression head to generate the recovery prediction output. The experimental results demonstrate that the proposed FGHA-Net model achieves promising prediction performance. On the testing set, the model obtained an RMSE of 0.011484 and an R^2 score of 0.844612, indicating low prediction error and strong agreement between predicted and actual recovery values. These findings suggest that the integration of BiLSTM, Transformer-based self-attention, and fine-grained attention fusion enables the model to capture both short-term and long-range dependencies in flotation process data. Despite these promising results, several limitations remain. Future work should evaluate the proposed model using larger real-time flotation datasets, additional process variables, and external industrial data collected under different operating conditions. Moreover, explainable AI techniques may be incorporated to identify the most influential process variables and improve the interpretability of the prediction results. The proposed model may also be integrated into real-time industrial monitoring systems to support practical mineral-processing applications. Future work can focus on robustness evaluation of the proposed model under sensor noise, missing measurements, equipment drift, and operational disturbances.

Acknowledgment

The authors extend their appreciation to Northern Border University, Saudi Arabia, for supporting this work through project number (NBU-CRP-2026- 289). This study is supported via funding from Prince Sattam bin Abdulaziz University, project number (PSAU/2026/R/1447).

Data Availability Statement

The data that support the findings of this study are openly available at https://data.mendeley.com/datasets/2vsrjbtfp5/1, reference number [28].

Conflicts of Interest

"The authors declare no conflicts of interest."

References

[1] Zhang, C., Tan, J., Peng, Y. and Xie, G., 2026. Froth image analysis for flotation process monitoring: a review. International Journal of Coal Preparation and Utilization, pp.1-41.

[2] Wang, H., Liu, Q., Xu, M., Xie, L., Xing, Y. and Gui, X., 2026. Semi-supervised clustering for operating condition recognition in sulfide ore flotation froth. Minerals Engineering, 237, p.110015.

[3] Tahooni Bonab, S., Abdollahi, H. and Abbaspour, A., 2026. A review of the mineral contact angle measurements and applications in the froth flotation process: Techniques and influential parameters, part II. Mineral Processing and Extractive Metallurgy Review, 47(1), pp.71-93.

[4] Leiva, C., Poblete, D., Humire, J. and Acuña, C., 2026. Kinetic analysis of froth flotation for remediation of hydrocarbon-contaminated coastal sands. Minerals Engineering, 244, p.110305.

[5] Lan, S., Dong, L., Shen, P., Mao, Y., Zhang, T., Zheng, Q., Liu, D. and Shen, Z., 2026. Green flotation depressats of typical salt and silicate gangue minerals in froth flotation–a comprehensive review. Green Chemistry.

[6] Kundu, T. and Parhi, P.K., 2026. Exploring the Flotation Performance of a Novel Low-Grade Indian Spodumene-Bearing Pegmatite Deposit for Sustainable Development. Journal of Sustainable Metallurgy, 12(1), pp.417-437.

[7] Liu, X., Feng, Y., Zhou, Z., Su, J., Wang, J., Zhang, Y., Wang, H., Jiang, H. and Wang, C., 2026. Nutrient-mediated surface hydrophobicity enhancing the removal of marine microplastics by froth flotation. Chemical Engineering Science, p.124080.

[8] Zhao, L., Zou, Z., Dong, B., Liu, L., Xie, Y. and Xie, S., 2026. Robust burst bubble recognition in flotation froths via RGB-D imaging and depth-guided feature extraction. Minerals Engineering, 242, p.110220.

[9] Barbouchi, A., Dachri, K., Labbilta, T., Naji, K., Faqir, H., Benzakour, I., Idouhli, R., Khadiri, M.E., Abouelfida, A. and Benzakour, J., 2026. Improving flotation recovery of gold in a refractory gold ore: An integrated approach between response surface methodology and automated quantitative mineralogy. Physicochemical Problems of Mineral Processing, 62(1).

[10] Benítez, J., Flores, V., Curilef, S., Martínez-Pelaez, R. and Leiva, C., 2025. Predicting copper recovery from flotation using machine learning and laboratory-generated data. Chaos: An Interdisciplinary Journal of Nonlinear Science, 35(9).

[11] Lu, M., He, X., Li, Y., He, L., Zou, Y., Yi, Z. and Li, P., 2025. Intelligent Optimization Setting of Froth Velocity by Integrating Multi-source Flotation Operation Information. IEEE Transactions on Automation Science and Engineering.

[12] Santos, E.G.D., Brum, I.A.S.D. and Ambrós, W.M., 2025. Techniques of Pre-Concentration by Sensor-Based Sorting and Froth Flotation Concentration Applied to Sulfide Ores—A Review. Minerals, 15(4), p.350.

[13] Yuan, M., Li, Q., Zhang, B., Pan, J., Gao, L., Jing, L. and Lu, L., 2026. A systematic review of deep learning methods for mineral exploration using multisource geoscience data (2018–2025). International Journal of Applied Earth Observation and Geoinformation, 149, p.105254.

[14] Pandian[1], P.K., Rohith, G., Kumar, R.L. and Harish, N., 2026. Deep Probabilistic Neural Network Model for Predicting Iron and Silica Concentrate Quality in Flotation Systems.

[15] Zhang, H., Yang, M., Tang, Z., Chen, J. and Xie, Y., 2026. Lightweight grade prediction in froth flotation using interpretable group Contribution-Enabled Weight-Sharing grouped RNN. Minerals Engineering, 237, p.110013.

[16] Bendaouia, A., Qassimi, S., Boussetta, A., Benzakour, I., Amar, O. and Hasidi, O., 2024. Artificial intelligence for enhanced flotation monitoring in the mining industry: A ConvLSTM-based approach. Computers & Chemical Engineering, 180, p.108476.

[17] Yang, S., Wen, Z., Zhou, M. and Zhou, C., 2026. An Innovative Extraction Method for the Morphological Features of Coal Flotation Froth Based on Instance Segmentation. Mining, Metallurgy & Exploration, pp.1-13.

[18] Abouda, K.A., Xu, D., Idress, W.M., Elmaki, H.M., Mazhar, T., Aoun, M., Ghadi, Y.Y., Shahzad, T. and Hamam, H., 2025. AI-DeepFrothNet: Continuous monitoring and tracking of froth flotation working condition by root cause analysis and optimized predictive control. Computers and Electrical Engineering, 123, p.110251.

[19] Dehon, V., Quintanilla, P., Chanona, A.D.R. and Van Impe, J.F.M., 2025. Probabilistic model predictive control for mineral flotation using Gaussian processes. Systems and Control Transactions, pp.1023-8.

[20] Song, Y., Yu, X. and Huang, M., 2025. A Copper Flotation Concentrate Grade Prediction Method Based on an Improved Extreme Gradient Boosting Algorithm. Applied Sciences, 15(20), p.11142.

[21] Bendaouia, A., Qassimi, S., Boussetta, A., Benzakour, I., Benhayoun, A., Amar, O., Bourzeix, F., Baïna, K., Cherkaoui, M. and Hasidi, O., 2024. Hybrid features extraction for the online mineral grades determination in the flotation froth using Deep Learning. Engineering Applications of Artificial Intelligence, 129, p.107680.

[22] Zhang, Y. and Wang, Z., 2023. Feature engineering and model optimization-based classification method for network intrusion detection. Applied Sciences, 13(16), p.9363.

[23] Hao, X., Liu, Y., Pei, L., Li, W. and Du, Y., 2022. Atmospheric temperature prediction based on a BiLSTM-Attention model. Symmetry, 14(11), p.2470.

[24] Xiong, X., Wang, X., Zhang, J., Huang, B. and Du, R., 2023. TCUNet: A lightweight dual-branch parallel network for sea–land segmentation in remote sensing images. Remote Sensing, 15(18), p.4413.

[25] Zhan, Z., Ren, H., Xia, M., Lin, H., Wang, X. and Li, X., 2024. Amfnet: Attention-guided multi-scale fusion network for bi-temporal change detection in remote sensing images. Remote Sensing, 16(10), p.1765.

[26] Gondaliya, K.M., Tsavdaridis, K.D., Raval, A., Amin, J.A. and Borisagar, K., 2025. Attention-Enhanced Progressive Transfer Learning for Scalable Seismic Vulnerability Assessment of RC Frame Buildings. Buildings, 15(23), p.4383.

[27] Mehmood, F., Ahmad, S. and Whangbo, T.K., 2023. An efficient optimization technique for training deep neural networks. Mathematics, 11(6), p.1360.

[28] Pu, Yuanyuan (2020), "Data for: FlotationNet: A hierarchical deep learning network for froth flotation recovery prediction", Mendeley Data, V1, doi: 10.17632/2vsrjbtfp5.1, available at https://data.mendeley.com/datasets/2vsrjbtfp5/1

[29] Nasiri, M., Iqbal, S. and Särkkä, S., 2025. Physics-informed machine learning for grade prediction in froth flotation. Minerals Engineering, 227, p.109297.

[30] Cook, R., Monyake, K.C., Hayat, M.B., Kumar, A. and Alagha, L., 2020. Prediction of flotation efficiency of metal sulfides using an original hybrid machine learning model. Engineering Reports, 2(6), p.e12167.

[31] Allahkarami, E., Salmani Nuri, O., Abdollahzadeh, A., Rezai, B. and Maghsoudi, B., 2017. Improving estimation accuracy of metallurgical performance of industrial flotation process by using hybrid genetic algorithm–artificial neural network (GA-ANN). Physicochemical problems of mineral processing, 53.

From Cloud to On-Premises: A Systematic Review of Cloud Repatriation Drivers, Strategies, and Governance Challenges

Amzari Mohd Roslan[1], Fazlina Mohd Ali[2], Faizan Qamar[3], Liyana Safra[4], Nur Arzilawati Md Yunus[5]
Faculty of Information Science and Technology (FTSM), Universiti Kebangsaan Malaysia (UKM), Selangor, Malaysia[1, 2, 3]
Faculty of Artificial Intelligence (FAI), Universiti Teknologi Malaysia (UTM), Selangor, Malaysia[4]
Faculty of Computer Science and Information Technology (FSKTM), Universiti Putra Malaysia (UPM), Selangor, Malaysia[5]

***Abstract*—Cloud repatriation is the migration of workloads from public cloud to private, on-premises, and hybrid clouds. Despite the growing adoption of cloud computing, there are many issues associated with operational expenditure, the unpredictable performance of cloud services, compliance concerns, vendor dependency, and data sovereignty concerns. Thus, cloud repatriation has become a strategic solution, especially in the public sector where governance and regulatory requirements are paramount. Previous studies on cloud repatriation are predominantly descriptive, with limited focus on predictive models that include cost assessment, security compliance, governance maturity, and workload mobility. Moreover, empirical studies are still limited, particularly in the context of public sector settings and regulatory constraints and sovereignty requirements. This study aims to provide a systematic literature review of the factors, solution approaches, and limitations of cloud repatriation towards on-premises transition. The review offers three main contributions: factors driving cloud repatriation, the synthesis of limitations and solution strategies for each factor, and an analysis of current trends across academic and industry literature. Major databases and selected industry reports were used to gather academic and industry sources. From an initial set of 124 articles published between 2021 and 2025, 20 key studies were selected. The results are broken down into five main factors: cost and economic factors, data security and sovereignty, vendor lock-in and portability, performance and latency needs, and governance and exit strategies. The review underscores the need for integrated governance-driven frameworks to enable secure, sustainable, and policy-compliant digital transformation.**

***Keywords*—Cloud repatriation; on-premises transition; systematic literature review; data security; hybrid cloud; data sovereignty; vendor lock-in; governance; implementation limitations**

I. Introduction

The adoption of cloud computing has been a key pillar of global digital transformation for over a decade. Beginning with the promise of lower costs, operational flexibility, and rapid scaling capabilities, the public cloud has attracted the attention of organizations across a variety of sectors, including finance, healthcare, education, and government [1], [2]. Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform have emerged as major players in the market, offering infrastructure and platforms that are seen as capable of replacing traditional data centres [3].

However, after a more extended period of adoption, more and more organizations are beginning to realize that reality does not always align with initial expectations [4]. This phenomenon has led to a resurgence of interest in cloud repatriation (CR), the process of relocating workloads from public clouds to private or hybrid environments [5], [6]. According to a 2023 study by IDG & Lufthansa Industry Solutions, the cloud repatriation phenomenon is not just a theory. Approximately 28% of enterprises in Germany have already moved workloads back from the cloud to on-premises infrastructure due to specific reasons including security, data compliance, and migration complexity [7].

The recent Cloud Migration and Repatriation Study 2025 by IDG/LHIND [8], as illustrated in Fig. 1, indicates that over 80 percent of all surveyed companies have done or intend to do repatriation in the near-mid-term.

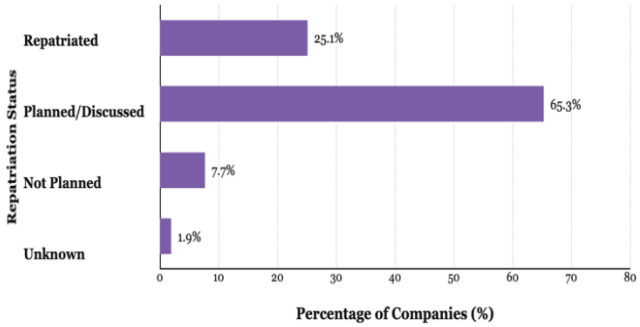


Fig. 1. Germany cloud repatriation study in 2025.

GEICO, a major insurance company in the United States, adopted a partial workload repatriation to gain economies of scale and have more control over important applications. This choice can illustrate how economic reasons and compliance by the organization can become an impetus to repatriate from the cloud [32]. More broadly, industry reports show that the repatriation phenomenon is becoming more pronounced. A Puppet study [10] found that 86% of CIOs plan to move at least some workloads from the public cloud back to private or on-premises infrastructure. These statistics confirm that repatriation is not an isolated phenomenon, but an emerging global trend. Several instances of CR were summarised in Table I.

TABLE I. CASE STUDIES OF CLOUD REPATRIATION DRIVING FACTORS

Citation	Organization	Factors
[7], [8]	IDG/LHIND	Security factors, data compliance, migration complexity, and vendor lock-in.
[9]	37signals (USA)	Hidden costs, inconsistent performance in public clouds.
[32]	GEICO (USA)	Cost savings, critical application control, organizational compliance.
[12]	Dropbox (USA)	High cost of public cloud operations, performance control, and reliability.
[13]	X (USA)	High cloud costs, control over data and performance, the need to reduce storage size and processing loads, optimizing infrastructure.
[33]	Global Research (CIO Survey)	Cost, security, compliance, and vendor lock-in.

Cloud repatriation does not signify the decline of cloud computing; rather, it represents a maturation of workload-mobility strategies. Posey et al. [6] argue that CR should be seen as part of the concept of "workload mobility", which is the flexibility to place applications and data in the most suitable environment, whether in a public cloud, private cloud, organizational data center, or hybrid location. The strategic flexibility of a firm is enhanced by its preparedness to shift workloads off-cloud [4]. This approach enables organizations to adopt the "best execution place" principle, where each workload is assigned according to its cost, performance, security, regulatory, and compliance requirements [11].

Among the most dominant factors driving cloud repatriation is uncertain and unexpected costs. Studies show that most organizations experience cloud spending that exceeds budget by 30–40% due to overprovisioning, data egress fees, and unmanaged resources [1], [2], [3]. Upon leaving AWS and creating its own infrastructure, Dropbox, in particular, saved more than USD 75 million [12]. Meanwhile, social media company X (Twitter) saved cloud operating expenses to the tune of USD 60 million annually following the relocation of media storage in-house [13]. As Murugesan [3] notes, although cloud will decrease operational expenditure (OpEx). In fact, the organization is usually bound to incur high long-term costs; therefore, the Total Cost of Ownership (TCO) analysis needs to be more comprehensive.

Besides cost consideration, CR is also being pushed by security and compliance considerations. Financial and banking institutions are also subject to strict regulations like the General Data Protection Regulation (GDPR) and the Payment Card Industry Data Security Standard (PCI-DSS) that mandate rigorous rates of encryption and incessant auditing [14], [15]. According to Everest Group [15], in the case of the banking industry, the aspects of data sovereignty and access control are of primary importance because any breach of the data may have legal consequences and a tarnished reputation. According to Priyanka Gowda [14], the migrations of the cloud in the banking industry are often faced with the difficulty of connecting the old system with the new systems, which heightens the chances of disruptions and impacts business-related essentials. Although cloud providers possess excellent security controls, the actual issues are presented by the vulnerability of governance and distributed responsibility frameworks whose implications are not always well-known by users [16], [17]. A study by the Ponemon Institute [17] revealed that the mean cost per annum of a cloud account breach is USD 6.2 million, and the mean application downtime is 138 hours per year. This implies that as much as the cloud provides a highly effective infrastructure, the success or failure of any organization is a major factor of organizational governance.

Another important driver of cloud repatriation is vendor lock-in and limited workload portability. eG Innovations [2] has reported that workload mobility is not easily achieved because of dependence on proprietary APIs and data formats. Räisänen [18] declared that the failure to have an effective exit strategy puts the organization at risk of losing strategic control and escalating expenses in the process of leaving a specific provider. To this end, a large number of organizations are currently pursuing multi-cloud or hybrid solutions to reduce the vendor lock-in risk. According to the BigID report [19], it is observed that 68% of the organizations currently operate in multi-cloud setups, and on average, there are four different types of providers who serve various important applications.

Moreover, the factors of performance and latency are also important triggers of the repatriation process. As Flores [20] discovered, data-intensive workloads, especially in the healthcare and media sectors, fail to reach optimum performance levels when they are deployed in the structure of a public cloud. As Posey et al. [6] noted, many organizations are shifting to a hybrid model because of the different needs posed by the development and production environment. The use of a phased migration strategy supported by middleware was reported by Priyanka Gowda [14] as the most common approach adopted in the financial services industry to address operational disruptions. Other new paradigms like edge computing are also gaining more academic interest, with the ability to handle data in real-time near the data source, which helps to reduce the issue of latency.

Considerable concern has also been raised regarding governance and exit strategies. Grabski and Schmidt [16] suggested a two-tier model of governance which is grounded on the TOE framework, and it indicates the necessity of collective responsibilities between the CSU and CSP. They pointed out that any organization needs to develop new governance policies whenever it decides on a repatriation, the definition of Service Level Agreements (SLAs), security control, and data backup procedures. Forbes [21] also pointed out that repatriation is not only a technical choice, but it is a way to gain control, data ownership, and bargaining power over cloud providers.

However, despite various studies and reports, several critical limitations remain unaddressed. Longitudinal studies measuring the cost, performance, and security impacts of repatriation over a five-to ten-year period are almost non-existent [5], [22]. Furthermore, existing models lack integration across cost evaluation, security compliance, governance maturity, and workload mobility, particularly within public sector contexts. Therefore, this study presents a systematic literature review that addresses these limitations through three focused contributions: identifying the key factors influencing cloud repatriation, examining solution strategies proposed in academic and industry literature, and revealing the implementation limitations between

theoretical models and real-world on-premises transition practices.

II. Methodology And Article Selection

This study employs a Systematic Literature Review (SLR) method to collect the extant literature regarding cloud repatriation, including the problems and factors involved in the process of workload migration between the public clouds and on-premises or hybrid cloud settings. The review process was guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework. The aim is to bring together theoretical and practical knowledge that will describe the state of research and practice in cloud repatriation.

A. Research Design

The methodological flow involves four stages:

1) Identification – Articles were found in reputable academic databases such as IEEE Xplore, SpringerLink, ScienceDirect, and Scopus, and reputable sources in the industry such as eG Innovations, Everest Group, Forbes Tech Council, and BigID Blogs. The search was conducted using the following keywords: 'cloud repatriation', 'cloud exit strategy', 'hybrid cloud', 'data segregation', 'cloud governance', and 'multi-cloud migration'. This initial search yielded a total of *124 records.* Search execution differed by source type. For academic databases, the six keywords identified above were combined using Boolean operators (AND/OR) within each platform's advanced search interface to maximize retrieval precision, with syntax adjusted to suit the specific search conventions of each platform. For industry sources without advanced Boolean search functionality, individual keywords were searched separately using each platform's native search or blog search function, and the keyword yielding results most topically relevant to cloud repatriation was retained per source. Retrieved records across all sources were consolidated before the deduplication step described in the Screening stage. The search and screening process was conducted over the weeks preceding the recorded extraction date of 8 December 2025, which is reported in Fig. 2 as the reference date for the record count at the point of extraction.

2) Screening – The 124 records were first screened for duplicate entries, resulting in 106 unique records retained for further review. These records then underwent systematic title and abstract screening, during which 71 records were excluded based on the identified reasons. Following this screening process, 35 articles were forwarded for full-text eligibility assessment, as illustrated in Fig. 2.

3) Eligibility – The 35 remaining articles underwent comprehensive full-text review against the defined inclusion criteria, which required coverage of topics related to cloud repatriation factors, migration strategies, data security, performance, or governance, published between 2021 and 2025. During this stage, 15 articles were excluded, and the remaining 20 studies were retained for final synthesis.

4) Inclusion – The final 20 studies were categorised into two groups: Theoretical Findings (TF), comprising 9 peer-reviewed academic publications sourced from journals and conference proceedings, and Real-Life Use Cases (RLUC), comprising 11 industry articles, reports, and technical blogs. This dual classification ensures a balanced representation of both theoretical constructs and practical applications in the analysis of cloud repatriation.

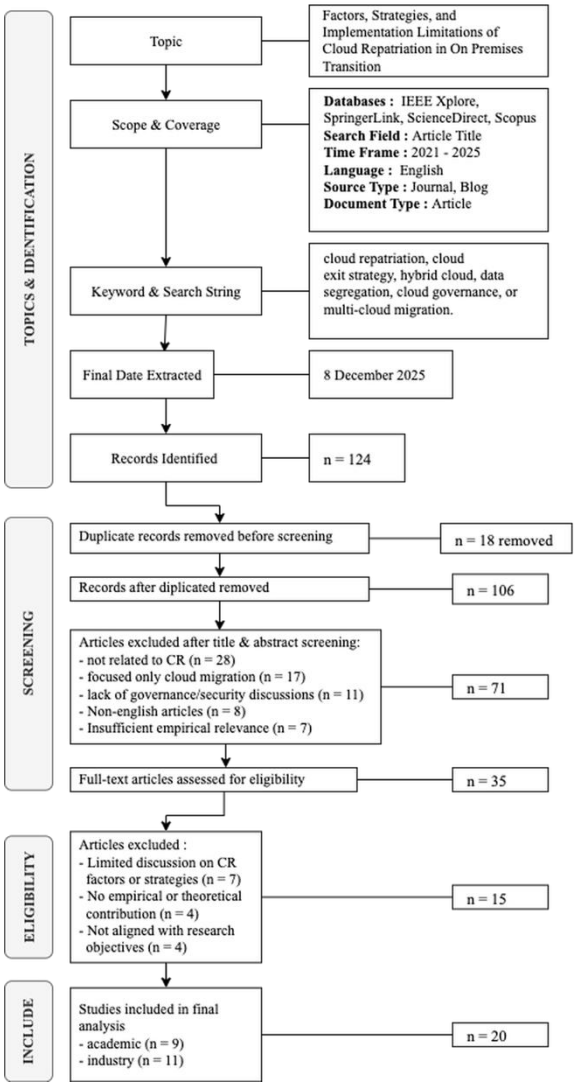


Fig. 2. PRISMA flow of article selection.

B. Research Questions

The review is guided by three Research Questions (RQs) designed to clarify the study's analytical focus:

- RQ1: What are the factors driving cloud repatriation in recent academic studies and industrial real-life use cases?
- RQ2: What are the limitations and solution strategies for each factor driving cloud repatriation towards on-premises transition?
- RQ3: What is the current trend of cloud repatriation in academic and industry literature?

These questions direct the synthesis and appropriateness in the alignment of the literature evidence to the study objectives.

C. Article Categorization

To differentiate between theoretical and applied contributions, the chosen works were broken down into two major groups: 1) Theoretical Findings (TF), and 2) Real Life Use Cases (RLUC). These measures are based on the final 20 articles included in the PRISMA process that identified the inclusion of the relevant studies in a systematic way that was validated. While this dual classification distinguishes theoretical from applied contributions, sources within each category carry differing levels of evidential strength. Peer-reviewed articles and conference proceedings (the TF category) have undergone formal editorial and peer review, whereas industry reports, company blogs, and white papers (the RLUC category) are not academically peer-reviewed and may reflect commercial or promotional framing associated with their originating organizations. This review retains these source types for the practitioner-level insight they offer, but findings drawn from lower-scrutiny sources are treated with appropriate interpretive caution throughout Section IV and corroborated against peer-reviewed evidence where available.

This dual classification allows establishing a balanced comparative position between the perspectives of real use cases and the practical elements of the academic field, to identify the most important limitations and trends. The distribution of selected articles, as illustrated in Fig. 3, indicates that even though academic discourse continues to be dominant, more and more industry-based evidence is emerging to corroborate theoretical knowledge with practical application. The balance between these two categories strengthens this review by ensuring that the aggregate outcomes are not biased toward the literature alone, but also provide practitioners and policymakers with practical perspectives.

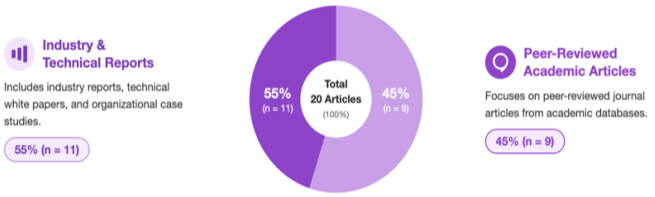


Fig. 3. Classification of selected articles and distribution of related studies.

TABLE II. THEORETICAL FINDINGS IN ACADEMIC STUDIES

Author	Publication Source	Number of Factors
[3]	IEEE Xplore	1
[14]	Scopus Journal	1
[16]	HICSS Conference	1
[18]	Conference Paper	1
[23]	IEEE Conference	3
[24]	Journal Article	2
[25]	IEEE Conference	2
[26]	Springer	1
[12]	Journal Article	1

TABLE III. REAL-LIFE USE CASES, INDUSTRY REPORTS, BLOGS, AND ONLINE PUBLICATIONS

Author	Publication Source	Number of Factors
[6]	Industry Report	1
[15]	Industry Report	1
[2]	Industry Blog	1
[19]	Industry Report	1
[22]	Industry Blog	1
[21]	Forbes Council	1
[7]	Industry Survey	1
[5]	White Paper	1
[11]	Industry Report	1
[10]	Industry Blog	1
[4]	White Paper	1

The scholarly and theoretical contributions identified through this review process are summarised in Table II, which presents the academic publications selected for analysis, while Table III presents the corresponding industry-based sources. As illustrated in Fig. 4, academic studies tend to contribute a higher number of CR factors. In contrast, all 11 industry sources each contribute a single CR factor, reflecting a more targeted focus in industry literature. Collectively, the 20 selected studies contribute a total of 22 identified CR factors, forming the thematic foundation for analysis in this review.

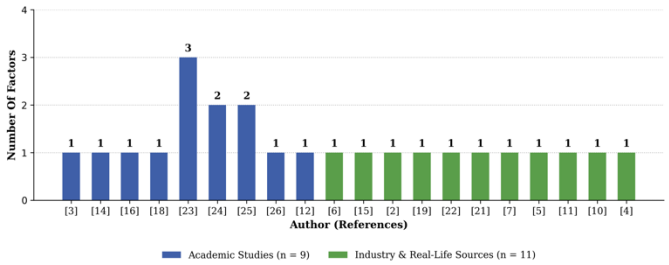


Fig. 4. Numbers of factors identified in each study by author.

D. Quality Assessment Criteria

Each included study was assessed using three quality dimensions to ensure the validity and relevance of the review outcomes. Completeness refers to the availability of clear findings, models, or empirical evidence that directly addresses the defined research questions. Credibility concerns the source authenticity and peer-review status of both academic and industry materials included in the review. Relevance requires direct alignment with the topics of cloud repatriation, hybrid migration, or related keywords as defined in the search strategy. Only studies meeting all three criteria were included in the final synthesis to maintain methodological rigour and ensure trustworthy interpretation of the results. Instead of a numerical scale, a binary pass/fail system was used for each dimension; if a study explicitly addressed the dimension, it was deemed to satisfy it (Yes); if it did not, or if it did but only superficially, or if it did but could not be independently verified, then it was deemed not to satisfy it (No). Thirty-five articles were selected for full-text review, of which only those that received a "Yes"

for all three dimensions were retained for final synthesis. These classifications were made by the lead author and then re-evaluated and confirmed by co-authors before finalization. This process does not begin with independent parallel coding by multiple reviewers, so this does not constitute formal interrater reliability verification and is discussed further as a methodological limitation in Section III-F.

E. Synthesis Approach

All included materials were analysed using thematic coding and classified across five factors derived from both theoretical and practical insights:

- Cost and Economic Considerations.
- Security, Compliance, and Sovereignty.
- Vendor Lock-In and Portability.
- Performance and Latency Requirements.
- Governance and Exit Strategy.

This approach integrates quantitative data with qualitative insights, aligning academic concepts with real-world practices. By combining rigorous research with industry relevance, the study offers a comprehensive understanding of the strategies, factors, and limitations associated with cloud repatriation.

III. Related Works

This section critically reviews state-of-the-art academic literature, industry reports, and real-life cases addressing the principal factors and limitations associated with cloud repatriation. Rather than presenting the selected studies as an isolated list, each subsection examines how the identified limitations inform the analytical framework of this review. A cross-factor comparison is subsequently presented in Table IV.

TABLE IV. Summary of Related Works By Factor

Factor	Author	Strengths	Limitation	Contributions
Cost & Economics	[3], [22], [23], [24], [21]	Cost unpredictability and TCO as repatriation drivers Hybrid deployment for cost control Dropbox, X/Twitter confirm post-CR savings	Descriptive analyses; no data-driven cost models No standardised TCO benchmarking framework Limited public sector validation	Synthesises cost factors across academic and industry sources Identifies TCO modelling as a key limitation Establishes cost as a core CR framework dimension
Security, Compliance, Sovereignty	[24], [14], [19], [15]	GDPR, PCI-DSS compliance as repatriation triggers Sovereignty and encryption as recurring concerns Data residency confirmed as industry driver	Prescriptive and conceptual; no implementation plans No unified compliance-into-repatriation model Security and governance treated separately	Maps security factors across academic and industry sources Highlights missing compliance-governance integration Informs policy-driven CR framework for regulated sectors
Vendor Lock-In & Portability	[6], [18], [11], [16]	Proprietary APIs and vendor contracts as portability barriers TOE-based multi-cloud governance proposed Hybrid adoption to reduce lock-in risk	Conceptual frameworks; no reusable blueprints Public sector constraints underrepresented Multi-cloud cost-complexity trade-offs unexplored	Consolidates lock-in factors from theory and practice Identifies portability gap in public sector context Supports portable-by-default architectural standards
Performance & Latency	[25], [26], [12], [3]	On-premises outperforms public cloud in latency-intensive workloads Selective repatriation and load-balancing proposed Latency variability across cloud regions	Simulation-based; no cross-industry validation No standardised cloud vs. on-premises benchmarking Cost-latency trade-offs underexplored	Consolidates pre- and post-CR performance evidence Flags benchmarking absence as implementation limitation Supports selective repatriation as a performance strategy
Governance & Exit Strategy	[16], [1], [7], [10], [15]	Structured exit planning recognised as critical enabler TOE-based and shared-responsibility models introduced Governance maturity correlates with repatriation readiness	Descriptive models; no implementation roadmaps No quantifiable governance maturity metrics No replicable blueprint for hybrid or public sector	Synthesises governance factors across all source types Identifies exit governance gap as critical limitation Establishes governance as dominant CR factor

A. Cost and Economic Considerations

Several studies have examined cost as a primary driver of cloud repatriation. Murugesan [3], Linthicum [22], and Kumar [5] identify increased operational expenditure and uncertain Total Cost of Ownership (TCO) as key concerns when workloads remain on public cloud platforms. Hui Chang [23] and Forbes Tech Council [21] further argue that transitioning to hybrid or on-premises environments offers greater long-term capital expenditure control. Real-world cases corroborate these findings, with Dropbox reporting annual savings of USD 75 million post-repatriation [12], and X (Twitter) reducing cloud costs by USD 60 million annually [13]. However, most existing studies rely on descriptive or anecdotal cost analyses rather than data-driven models. Kumar [5] proposes a Cloud Right framework but lacks cross-industry empirical validation, while Linthicum [22] and Murugesan [3] do not quantify cost-performance trade-offs. No standardised TCO benchmarking framework currently exists, particularly for non-technical and public sector contexts. This absence of a standardised, data-driven TCO framework is a central gap that this review responds to by consolidating disparate cost evidence (Sections IV and Table IV) into a single comparative synthesis rather than treating each case in isolation.

B. Security, Compliance, and Sovereignty

Security and compliance concerns are widely recognised as significant factors influencing cloud repatriation decisions. Abied et al. [24] and Priyanka Gowda [14] highlight that e-government and Banking, Financial Services and Insurance (BFSI) environments require strict data privacy, trust, and governance compliance. Industry reports by BigID [19] and Everest Group [15] confirm that data sovereignty and encryption integration are common repatriation triggers in regulated sectors. Despite this, existing works remain largely prescriptive, presenting conceptual models without practical implementation plans. Abied et al. [24] focus on security awareness without operational guidance, while BigID [19] discusses compliance trends without demonstrating real-world applicability. Furthermore, no unified policy-based model currently integrates compliance controls directly into repatriation workflows, particularly for public sector environments. This separation of prescriptive security models and operational implementation is exactly what this review aims to bridge by correlating security factors to specific solution approaches (Table VII) instead of compliance requirements and technical controls as separate discourses.

C. Vendor Lock-In and Portability

Vendor dependency and lack of portability are consistently reported challenges in cloud repatriation literature. Posey et al. [6], Räisänen [18], and Jewargi [11] highlight that proprietary APIs and storage formats reduce workload flexibility and increase switching costs. Grabski and Schmidt [16] extend this by proposing a TOE-based multi-cloud governance model that promotes shared responsibility across providers. However, these studies largely overlook the operational complexity and cost of managing multi-cloud environments. Posey et al. [6] and Jewargi [11] describe avoidance strategies without providing reusable architectural blueprints, and most academic works focus on identifying risks rather than prescribing standardised interoperability models. Public sector constraints, including procurement regulations and data residency requirements, remain underrepresented in existing portability frameworks. Existing research does not extend beyond the identification of risk, so this review aggregates lock-in evidence from both theoretical and practical sources to identify areas where public sector portability needs are not met (Table IV).

D. Performance and Latency Requirements

Technical performance, particularly latency and scalability, is a well-documented repatriation trigger in enterprise contexts. Mallikarjuna and Kumar [25], Sandhiya and Suresh [26], and Korada and Sikha [12] demonstrate that on-premises and private cloud environments consistently outperform public cloud configurations in high-computation and latency-intensive workloads. Murugesan [3] and Hui Chang [23] further note that latency directly impacts resource efficiency and user experience. Nevertheless, most performance analyses are context-specific, relying on simulation-based or single-case validations. Mallikarjuna and Kumar [25] assess performance parameters without standardised benchmarking tools, while Sandhiya and Suresh [26] propose load-balancing algorithms that lack cross-industry scalability validation. A related strand of work by Mohd Ali et al. [27] proposes a multi-factor replica placement strategy to mitigate resource failure in cloud replication environments, addressing scalability and reliability concerns from a data-placement perspective. This approach was later extended by Mohd Ali et al. [28] through a multi-factor replica placement strategy for resource failure extenuation. However, both approaches were validated using CloudSim and have not been benchmarked against repatriated on-premises or hybrid infrastructure. Their applicability to real-world cloud repatriation therefore remains insufficiently validated across industries. No widely accepted benchmarking standard currently exists for comparing repatriated versus cloud-hosted workload performance. This lack of a shared benchmarking standard motivates the cross-study performance synthesis undertaken in this review (Table VI), which consolidates otherwise incomparable simulation-based findings into a single qualitative before-and-after comparison.

E. Governance and Exit Strategy

Governance and exit planning are recognised as critical enablers of successful cloud repatriation. Grabski and Schmidt [16] propose a two-tier TOE-based governance model that assigns distinct technical and functional responsibilities across multi-cloud environments. eG Innovations [1] and Puppet [10] advocate for proactive exit planning to prevent uncoordinated transitions, while IDG/LHIND [8] and the Everest Group [15] demonstrate a strong correlation between governance maturity and repatriation readiness. In the Malaysian public sector context, Sallehudin et al. [29] similarly identified organisational, technological, and environmental readiness factors as key determinants of successful cloud implementation, offering an early empirical foundation for governance-driven adoption decisions. This TOE-oriented perspective has since been extended by Mohd Noh et al. [30], who applied a resource-based and TOE framework to model cloud adoption governance in the Malaysian healthcare sector, reinforcing the relevance of organisational readiness constructs beyond adoption into repatriation contexts. However, existing governance models are predominantly descriptive, offering best practices without implementation roadmaps or quantifiable maturity metrics. Mallam Ibrahim et al. [31] further note that governance and human-organisational factors remain under-modelled even in cloud adoption studies within academic institutional settings, echoing the same descriptive gap observed in repatriation-specific governance frameworks. Comparative analyses across sectors are lacking, and there is no replicable blueprint of governance that can be generalised to hybrid or public sector settings. This is the most consistent gap identified in this study and is used as the main organising theme for the integrated framework proposed in Section IV, as governance models are descriptive and non-quantifiable in all sources reviewed.

F. Synthesis and Research Limitations

Collectively, the reviewed studies demonstrate that cloud repatriation studies have evolved across five major dimensions, as summarised in Table IV. While scholarly contributions provide theoretical insight and conceptual elaboration, industry studies offer empirically grounded operational observations. Despite this, a significant limitation remains in the literature itself: existing research has not produced a unified framework applicable to public sector contexts, where heightened data sensitivity, regulatory compliance requirements, and national

sovereignty concerns necessitate dedicated governance and security conventions. It is equally acknowledged that this review, consistent with the nature of a systematic literature review, does not itself generate new empirical or predictive data; its contribution lies in the structured cross-source synthesis of existing theoretical and industry evidence rather than in novel primary measurement. To partially offset this methodological boundary, Table IV and Table VI advance beyond simple narrative summary by cross-tabulating strengths, limitations, and solution strategies across all 20 studies, an analytical step (Section IV) that surfaces patterns, such as the disproportionate concentration of governance-related gaps, that are not explicit in any single source reviewed. Building on this synthesis, this review addresses the identified literature gaps through three contributions: the identification of CR factors, the synthesis of solution strategies, and the mapping of implementation limitations between theoretical models and real-world on-premises transition practices.

A further methodological limitation concerns the verification process underlying both the quality appraisal (Section II-D) and the thematic coding used in Fig. 6 (Section IV). In both cases, initial classification was performed by the lead author and subsequently reviewed and validated by co-authors, rather than through fully independent parallel coding from the outset. While this collaborative validation step provided a degree of cross-checking, it does not constitute formal inter-rater reliability verification. Future replications of this review would benefit from a fully independent dual-coding protocol at both the quality-appraisal and thematic-classification stages to further strengthen methodological rigour.

A related limitation concerns the heterogeneous evidential strength of the synthesized sources, which span peer-reviewed publications, industry reports, blogs, and self-published white papers with differing levels of editorial scrutiny. No differential weighting was applied based on source type; industry-reported statistics are presented alongside peer-reviewed findings as complementary rather than equivalent evidence, and readers should interpret figures from industry or vendor-affiliated sources with awareness of potential promotional framing.

G. Cross-Factor Gap Analysis

To complement the narrative synthesis above, a Solution-to-Limitation Ratio (Gap Ratio) is computed for each CR factor by dividing the number of solution strategies proposed (Table VII) by the number of limitations identified (Fig. 8), as summarised in Table V. A ratio of 1.00 indicates that the literature has proposed at least one corresponding strategy for every limitation identified, while a ratio below 1.00 signals that limitations have outpaced concrete solution development. As shown in Table V, Cost and Economic Considerations, Security, Compliance and Sovereignty, and Vendor Lock-In and Portability each reach a Gap Ratio of 1.00, whereas Performance and Latency and Governance and Exit Strategy both return 0.75. However, this numerical parity masks a qualitative difference: Performance limitations are largely technical and partially mitigated by emerging approaches such as edge computing and selective repatriation, whereas Governance limitations are organisational and procedural, compounded by the finding that 45% of organisations lack any formal exit strategy [10]. Governance and Exit Strategy is therefore treated as the highest-priority gap for future CR framework development, consistent with the trend analysis in Section IV (RQ3). This analysis is a modest extension beyond descriptive synthesis rather than new primary empirical data, offering a reproducible basis for prioritising future research and policy attention across the five CR factors.

TABLE V. SOLUTION-TO-LIMITATION GAP RATIO ACROSS THE FIVE CR FACTORS

CR Factor	Limitations (LC)	Solution Strategies (SC)	Gap Ratio (SC/LC)	Priority For Future Research
Cost and Economic Considerations	4	4	1.00	Moderate
Security, Compliance and Sovereignty	4	4	1.00	Moderate
Vendor Lock-In and Portability	3	3	1.00	Moderate
Performance and Latency	4	3	0.75	High
Governance and Exit Strategy	4	3	0.75	Highest

H. Contradictory Evidence and Interpretive Tensions

While the twenty studies reviewed are broadly consistent in identifying cloud repatriation as a strategically beneficial response to cost, security, and governance pressures, closer examination reveals three interpretive tensions that a purely convergent synthesis would obscure. First, a proactive-versus-reactive tension emerges around cost-driven decisions. Kumar [5] proposes the Cloud Right framework as a proactive workload suitability assessment conducted before migration, yet the case evidence in Fig. 5 and the account provided by Murugesan [3] show that organizations such as Dropbox, X, GEICO, and 37signals repatriated only after costs had already exceeded expectations, reflecting a reactive correction rather than proactive modelling. Second, security is simultaneously described as technically well-managed by cloud providers and as the most frequently cited repatriation trigger, accounting for 35 percent of coded mentions in Fig. 6. This is resolved on closer reading: the underlying concern is not the technical adequacy of provider security controls, but the lack of clarity in shared-responsibility arrangements between Cloud Service Providers and Cloud Service Users. Third, sources disagree on the extent of intervention needed to address vendor lock-in. BigID [19] treats existing multi-cloud adoption, already practiced by 68 percent of organizations, as a sufficient mitigation, whereas Räisänen [18] argues that without a formal exit strategy, organizations remain exposed to loss of strategic control. This review does not resolve which position holds more broadly, noting instead that the appropriate degree of intervention likely depends on organizational context and workload criticality.

These tensions indicate that the reviewed literature converges on repatriation as a significant trend but does not speak with one voice on when it should occur, what precisely

motivates security-driven decisions, or how completely vendor lock-in can be mitigated short of full repatriation. Acknowledging this strengthens rather than weakens the review's synthesis.

IV. Findings and Discussion

This section discusses the key findings derived from the analysis of 20 selected studies related to cloud repatriation (CR). The discussion is organised around five major factors, as outlined and discussed in this study. Each factor is examined in relation to the three research questions (RQ1–RQ3) to highlight the primary factors driving CR, the limitations and solution strategies for each CR factor, and the current trends of CR studies across academic and industry literature.

RQ1: What are the factors driving cloud repatriation in recent academic and industrial studies?

Five prominent factors were identified across the recent academic and industrial studies:

1) Cost and Economic considerations: In both scholarly and industrial studies, cost is one of the most frequently identified drivers of cloud repatriation. George [4], Flores [20], and Murugesan [3] point out that it is a common experience in organizations to underestimate the cost of operation of the hyperscale cloud platform, and actual expenditure may exceed initial estimates by 30–40%. Egress charges, auto-scaling anomalies, orphaned resources and complex OpEx billing designs are also leading sources of cost volatility that organisations, particularly those in the public sector, cannot easily manage. Institutions that are owned by the government have another limitation, which is a rigid procurement process and an inflexible system of depreciation [32], [33], further complicating cost predictability of a cloud service in the long-term.

The empirical importance of cost-related motivations is demonstrated in Fig. 5, which compares annual savings reported by organizations that repatriated critical workloads. After departing from AWS, Dropbox reported savings of USD 75 million per year, while X (Twitter), GEICO, and 37signals each reported significant reductions in cloud operating costs following their respective repatriation initiatives [9], [13], [33]. The cost savings presented in Fig. 5 are individually reported figures drawn directly from each organization's own case-study disclosure rather than statistically pooled or averaged. As these figures originate from different reporting periods, organizational scales, and workload profiles, they are juxtaposed for side-by-side comparison and should be interpreted as illustrative case-level evidence rather than a representative or generalizable benchmark of expected repatriation savings. These practical outcomes support the literature in affirming that repatriation decisions tend to be reactive responses to accrued financial inefficiencies rather than the result of proactive economic modelling. Taken together, the findings validate that cost unpredictability, driven by egress charges, overprovisioning, and complex billing structures, remains one of the core triggers of cloud repatriation. Therefore, this financial pressure has compelled many organizations, particularly those in the public sector with rigid procurement constraints, to reassess their cloud dependencies and transition critical workloads back to on-premises infrastructure as a more cost-predictable and financially sustainable alternative.

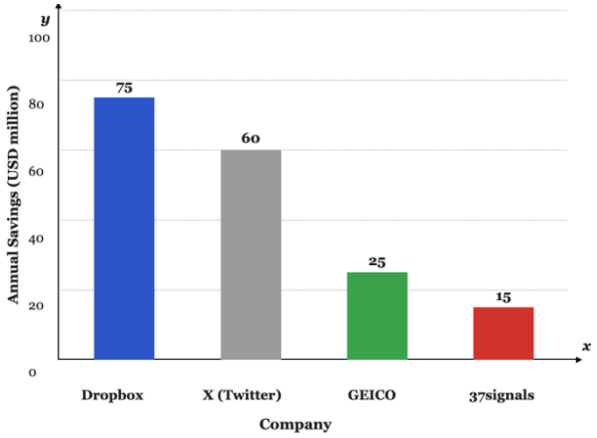


Fig. 5. Cost savings achieved through CR by various companies.

2) Security, compliance, and sovereignty: Security, compliance, and sovereignty are still coming out as the most powerful factors in cloud repatriation, especially in highly regulated industries. Priyanka Gowda [14] and the Everest Group [15] emphasize that companies that are under strict regulatory requirements, like the General Data Protection Regulation (GDPR) and the Payment Card Industry Data Security Standard (PCI-DSS), must have robust auditability, end-to-end encryption, and strict data sovereignty requirements. These needs are usually beyond the capability of organizations to impose uniformly in the multi-tenant public cloud landscape. Even sovereignty issues are supported by the fact that 23% of organizations interviewed by Posey et al. [6] state that data sovereignty conflicts were a prominent factor in deciding to repatriate workloads on-premises. Also, there is a longstanding lack of clarity in shared-responsibility approaches between Cloud Service Providers (CSPs) and Cloud Service Users (CSUs), which leads to gaps in governance, particularly in such aspects as access control, data lineage, and cross-cloud visibility.

The empirical relevance of these security-related issues can be seen in Fig. 6, which summarizes the most commonly reported factors of repatriation: data sovereignty (35%), regulatory compliance (30%), data breaches and misconfigurations (20%), and vendor responsibility gaps (15%). The security-concern distribution presented in Fig. 6 was derived through thematic synthesis rather than extraction from any single source. Findings from six studies explicitly addressing security-related repatriation drivers, Posey et al. [6], Priyanka Gowda [14], Everest Group [15], BigID [19], Grabski and Schmidt [16], and the Ponemon Institute [17] were reviewed and coded into these four recurring thematic categories, and the resulting thematic frequencies were expressed as percentages of the total coded mentions across these six sources. This initial coding was performed by the lead author and subsequently reviewed and validated by co-authors; a fully independent dual-coding protocol with inter-rater

reliability verification was not conducted, and this is noted as a limitation in Section III-F. This is a reinforcement of trends evident in the literature that regulatory pressure and sovereignty assurance are the key triggers in motivating decisions of security-motivated repatriation.

3) Vendor lock-in and portability: Vendor lock-in is an intractable barrier to cloud agility. According to Räisänen [18] and Murugesan [3], proprietary cloud application programming interfaces (APIs), storage formats, and management tools impede workload migration and give rise to high switching costs. In the case of public-sector organizations, the problem is exacerbated by long-term vendor contracts, tough procurement policies, and data-residency requirements, all of which limit mobility. As a result, vendor lock-in represents both a technical and governance barrier that undermines organisational autonomy, reduces bargaining power, and complicates regulatory compliance.

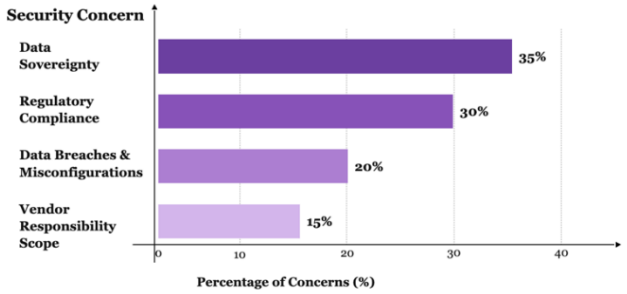


Fig. 6. Distribution of security concerns identified in CR studies.

4) Performance and latency requirements: The latency and performance requirements of latency-sensitive applications such as real-time analytics, artificial-intelligence inferencing, financial operations, and Internet of Things (IoT) applications are difficult to implement in public clouds due to performance and latency requirements. Latency variability is reported to range between 40 and 60 ms between cloud regions [25], [26], and Grabski and Schmidt [16], Murugesan [3] and Räisänen [18] report such inefficiencies as resource contention and multi-tenant virtualization overheads. As such, organizations are progressively embracing the idea of selective repatriation, which moves only the latency-sensitive workloads on-premises to restore deterministic performance. Table VI presents a qualitative, direction-of-change comparison synthesized from the empirical patterns and discussion described across four major works [3], [12], [25], [26]. As these studies each report performance findings using different metrics, testing environments, and measurement instruments, this review does not derive or report pooled numerical ranges. Table VI summarizes the consistent directional trend, whether improvement or deterioration, that these studies collectively describe for each metric, without implying a specific quantified magnitude common across sources. Taken together, these papers reveal that cloud-hosted workloads generally have increased latency, jitter, and packet loss as well as response time due to multi-tenant contention, hypervisor overhead, distributed routing, and unpredictable resource scheduling in the environments of a public cloud. Both Mallikarjuna and Kumar [25] and Sandhiya and Suresh [26] prove that they have lower transfer rates, lesser throughput, and worse performance under cloud congestion, which is the baseline of Before CR conditions. On the other hand, based on evidence that is given by Korada and Sikha [12]enterprises turn to the implementation of their own clouds with an express objective to achieve low latency, greater throughput, considerably better IOPS, and more predictable storage and network performance-qualifications that define the after CR values. Murugesan [3] goes on to substantiate this distinction by reporting real-life repatriation experience (Dropbox, 37signals, X/Twitter) in which enhanced control over hardware, data locality, and system configuration directly correlated to enhanced reliability, faster processing time, and cost performance optimization. Collectively, these four studies describe a consistent directional shift from the variable, contention-affected profile characteristic of multi-tenant public cloud environments toward the more predictable, locality-enhanced profile associated with repatriated on-premises or private-cloud infrastructure. It is this consistent direction of improvement, rather than any specific quantified magnitude, that constitutes the primary empirical pattern supported by the reviewed literature.

TABLE VI. QUALITATIVE PERFORMANCE COMPARISON BEFORE AND AFTER CR

Metric	Before CR	After CR	Authors
Latency	Elevated and variable, driven by multi-tenant contention and distributed routing	Reduced and more deterministic due to local network paths	[25],[26],[12]
Latency Jitter	Higher variability under shared-tenancy congestion	More stable due to dedicated infrastructure	[25],[26]
Throughput	Constrained by shared network and storage resources	Improved via local storage and reduced contention	[25]
IOPS (Storage Performance)	Limited by shared multi-tenant storage backends	Higher and more predictable with dedicated enterprise storage	[12]
Application Response Time	Slower due to hypervisor overhead and network hops	Faster due to local infrastructure control	[25],[3]
Network Packet Loss	Higher due to WAN routing and cloud congestion	Lower due to local LAN environments	[25],[26]
Service Uptime	Subject to shared-infrastructure availability constraints	Improved through dedicated control and predictable infrastructure	[25],[3]
Data Processing Time	Slower, reflecting combined latency/throughput/IOPS constraints	Faster, reflecting improvements across all underlying metrics	[12],[3]

5) Governance and exit strategy: The maturity of governance is always a controlling factor that spells out the success of cloud repatriation results. Grabski and Schmidt [16] highlight the use of a two-level TOE-based governance framework whereby the technical and functional roles have to be distinctly allocated in order to embrace accountability in the multi-cloud settings. But the industrial experience indicates that the structure of governance is not in general satisfactory, as a matter of fact. According to Puppet [10], about 45% of organisations do not have any formal exit strategy, and as a result, configuration drift, lack of compliance confidence, and incomplete data validation have been observed when organisations repatriate workloads. These circumstances demonstrate that one of the most prevalent underlying factors of failed or ineffective repatriation processes is governance weaknesses rather than technological constraints. The concept flow of the governance activities mentioned throughout the literature is presented in Fig. 7, which demonstrates a governance lifecycle with policy development, implementation, constant monitoring, evaluation, and a cyclic review.

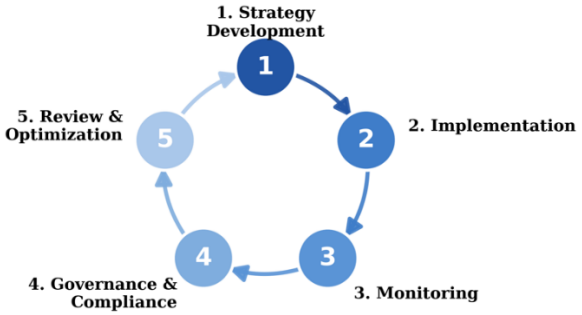


Fig. 7. Governance lifecycle.

This lifecycle underscores the fact that successful repatriation cannot work without governance as an iterative and policy-based activity as opposed to an administrative exercise that is one-time. The absence of such organised governance systems validates the obstacles mentioned in numerous research works and highlights the necessity to have standardised and actionable governance models in hybrid and multi-cloud settings.

RQ2: What are the limitations and solution strategies for each factor driving cloud repatriation towards on-premises transition?

In this section, RQ2 is addressed by systematically examining the limitations identified within the reviewed literature and the solution strategies proposed for each of the five CR factors. The consolidated findings are illustrated in Fig. 8, which contrasts theoretical limitations with real-life implementation challenges, and further mapped to the corresponding solution strategies in Table VII, which presents the corresponding solution strategies for each CR factor.

Fig. 8 illustrates the implementation gap between theoretical knowledge and industry practice across the five CR factors. Theoretical Findings (TF) from academic literature predominantly reveal conceptual and prescriptive limitations, while Real-Life Use Cases (RLUC) from industry practice expose more complex, context-specific challenges that existing frameworks have yet to adequately resolve. This juxtaposition confirms a consistent theory-practice gap, wherein academic models may underestimate the operational, contractual, and governance complexities encountered in real-world repatriation scenarios. Notably, this gap is most evident in the governance and security dimensions, where the absence of standardised implementation roadmaps and unified compliance models renders existing theoretical prescriptions insufficient for direct operational deployment.

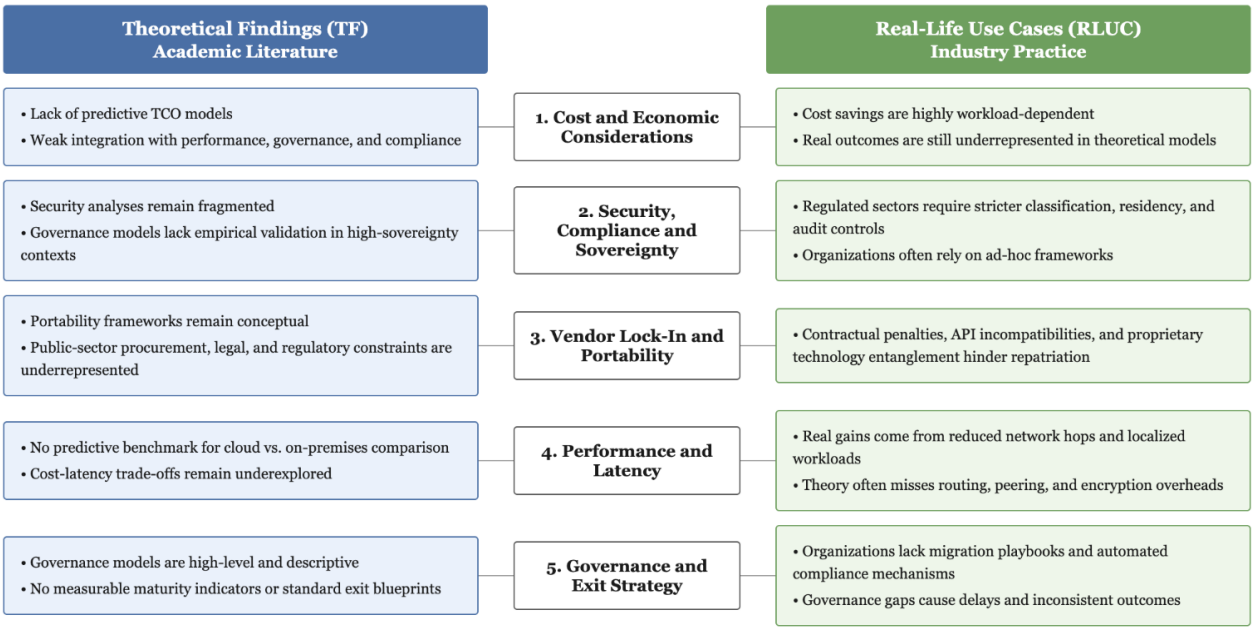


Fig. 8. Implementation limitations: theoretical findings vs real-life use cases.

TABLE VII. SOLUTION STRATEGIES CORRESPONDING TO THE FIVE CR FACTORS

CR Factor	Solution Strategies Proposed	Authors
Cost and Economic Considerations	- Cloud Right methodology for workload suitability assessment - Comparative TCO modelling: cloud vs. on-premises - Resource rightsizing automation for continuous cost governance - Balance short-term OpEx against long-term CapEx investment	[3], [5], [11]
Security, Compliance and Sovereignty	- TOE-based co-governance framework (CSP–CSU responsibility distribution) - Sector-specific sovereign cloud models for regulated industries - Automated data discovery, classification, and end-to-end encryption - Continuous compliance auditing and proactive sovereignty governance	[14], [15], [16], [17], [19]
Vendor Lock-In and Portability	- Hybrid and multi-cloud deployment to reduce vendor dependency - Cloud-agnostic technologies: Kubernetes, open APIs, and IaC - Contractual portability and post-termination repatriation clauses	[6], [11], [16], [18]
Performance and Latency	- Selective repatriation of latency-sensitive workloads to on-premises - Data locality enforcement to minimise network latency - CDN, edge computing, and AI-assisted workload orchestration	[5], [6], [20], [22], [23], [26]
Governance and Exit Strategy	- Two-tier TOE-based governance with distinct technical and functional accountability - Repatriation readiness embedded in cloud contracts (portability and audit clauses) - Lifecycle-based governance: procurement, deployment, auditing, and exit planning	[1], [2], [15], [16]

Building upon the limitations identified, Table VII consolidates the solution strategies proposed across the five CR factors, drawing from both academic recommendations and industry-based evidence. Each strategy is directly mapped to its corresponding CR factor, ensuring that the proposed interventions are specific, actionable, and responsive to the gaps identified in Fig. 8. The solution strategies presented in Table VII constitute an integrated and mutually reinforcing strategic foundation for CR decision-making. Cost strategies establish the financial justification for transition, while security and sovereignty strategies define the regulatory boundaries within which repatriation must be executed. Vendor lock-in strategies preserve organisational autonomy, performance strategies determine the technical suitability of workloads, and governance strategies provide the procedural infrastructure required to manage the entire repatriation lifecycle. Collectively, these five strategic dimensions confirm that cloud repatriation is not merely a technical migration exercise, but a multifaceted organisational policy decision that must be harmonised across regulatory, financial, operational, and technical priorities. This integrated approach is especially critical for public sector organisations, where regulatory obligations, procurement constraints, and national sovereignty requirements impose additional complexity that demands purpose-built, policy-driven frameworks.

RQ3: What is the current trend of cloud repatriation in academic and industry literature?

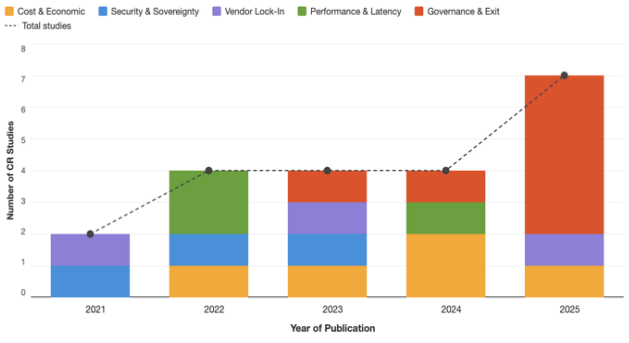


Fig. 9. Trend of CR studies by factor across academic and industry literature (2021 – 2025).

The analysis of 20 selected studies published between 2021 and 2025 reveals a discernible evolution in cloud repatriation studies, reflecting shifts in both academic focus and industry priorities across the five identified CR factors. As illustrated in Fig. 9, the distribution of CR-related studies across the review period demonstrates a progressive transition from performance and security-centred concerns toward governance and cost-driven imperatives.

A. Trend Analysis Across the Review Period

1) Early recognition of sovereignty and portability concerns: In 2021, CR research was primarily represented by a small number of foundational studies. Posey et al. [6] introduced the concept of workload mobility and vendor lock-in as early drivers of repatriation interest, while Gowda [14] highlighted security and compliance challenges in BFSI cloud migrations. These early works established sovereignty and portability as the initial academic triggers of CR discourse, though empirical evidence remained limited at this stage.

2) Performance and security as dominant themes: The year 2022 witnessed a notable increase in academic contributions, particularly from IEEE conference publications. Mallikarjuna and Kumar [25] and Sandhiya and Suresh [26] demonstrated measurable performance superiority in on-premises environments, while Abied et al. [24] and Chang [23] examined security and service differentiation between cloud and traditional infrastructure. This period reflects a growing academic interest in empirically validating the technical limitations of public cloud environments as justification for repatriation.

3) Cost and vendor lock-in gain industry momentum: By 2023, industry-driven evidence began to dominate CR discourse. Jewargi [11], BigID [19], and Forbes Tech Council [21] collectively highlighted escalating cloud costs, vendor dependency, and compliance obligations as compelling repatriation triggers. The IDG/LHIND survey [8] further confirmed that security, data compliance, and migration complexity were primary motivators among enterprise decision-makers. This shift marked a transition from purely academic performance analysis toward commercially grounded

cost and portability concerns, signalling that CR was becoming a mainstream strategic consideration across multiple industries.

4) Cost optimisation and real-world validation: In 2024, the focus consolidated around cost optimisation and real-world repatriation outcomes. Murugesan [3] provided a comprehensive review of cloud services, highlighting the financial inefficiencies of long-term public cloud dependency, while Korada and Sikha [12] documented enterprise motivations for building private cloud infrastructure post-public cloud adoption. Flores [20] further examined the strategic and technical implications of CR for DevOps and security practices. Collectively, these works validated the financial and operational case for repatriation through empirical and case-based evidence, reinforcing cost unpredictability as a core repatriation trigger.

5) Governance emerges as the dominant theme: The most significant trend observed across the review period is the sharp rise in governance-oriented publications in 2025. Grabski and Schmidt [16], eG Innovations [1], [2], Puppet [10], Räisänen [18], and the IDG/LHIND study [8] collectively position governance maturity, exit strategy formalisation, and lifecycle-based compliance as the most critical enablers of successful repatriation. This concentration of governance-focused works in 2025 reflects a maturation of CR studies, moving beyond the identification of repatriation drivers toward the prescription of structured, policy-driven frameworks for managing the transition process.

B. Overall Trend Synthesis

The longitudinal analysis of CR literature from 2021 to 2025 reveals three distinct phases of evolution in CR studies. The first phase (2021–2022) is characterised by foundational recognition of performance, security, and portability as primary repatriation triggers, predominantly within academic contexts. The second phase (2023–2024) reflects a convergence of academic and industry perspectives, with cost unpredictability and vendor lock-in emerging as dominant real-world drivers, validated through high-profile organisational cases such as Dropbox, X (Twitter), and GEICO. The third phase (2025) marks a governance turn in CR studies, wherein the field transitions from descriptive analysis of repatriation factors toward the development of prescriptive, policy-compliant frameworks suitable for regulated and public sector environments.

This evolutionary trajectory underscores the growing recognition that cloud repatriation is not a reactive technical decision, but a strategically planned organisational transition requiring integrated governance, cost modelling, and sovereignty-aware frameworks. The distribution of CR factors across the review period, as synthesised in Fig. 9, further confirms that Cost and Economic Considerations (22.7%) and Governance and Exit Strategy (22.7%) collectively represent the most critical priorities in the literature, while Security, Compliance and Sovereignty (18.2%), Performance and Latency (18.2%), and Vendor Lock-In and Portability (18.2%) reflect equally weighted but complementary dimensions of the CR research landscape. The complete classification of CR factors and their associated solution strategies is consolidated in Fig. 10, providing a structured taxonomy of the CR literature landscape identified in this review. It is acknowledged that the trend patterns described above are derived from publication counts within a bounded set of sources rather than a comprehensive bibliometric analysis, and may be influenced by factors beyond genuine shifts in research focus, such as database indexing lag and the growing tendency of industry organizations to publish repatriation-related commentary as the topic gains broader market attention. Consequently, the apparent 'governance turn' identified in 2025, and the broader three-phase trajectory described above, should be interpreted as a pattern observed within the reviewed evidence base rather than a definitive claim about the actual pace or sequence of change in cloud repatriation practice.

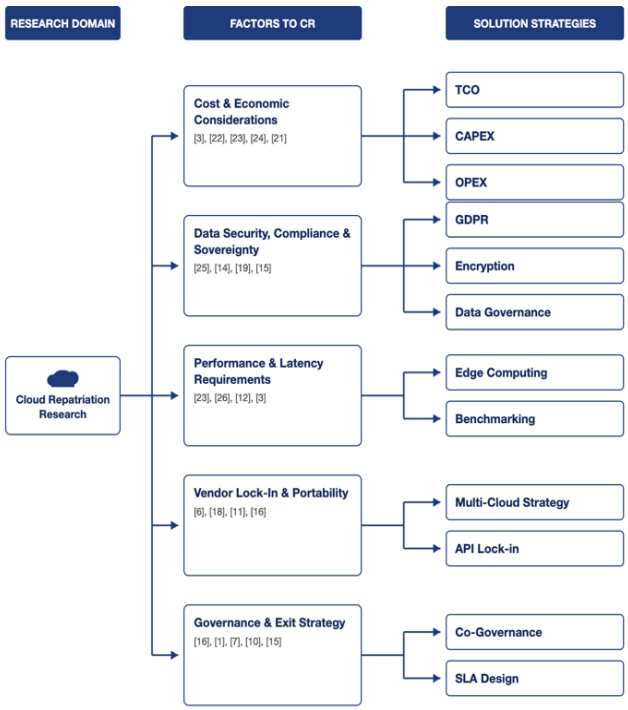


Fig. 10. CR factors and associated strategies.

V. Conclusion and Future Works

This study conducted a systematic literature review of the factors, solution strategies, and implementation limitations associated with cloud repatriation towards on-premises environments. Based on 20 academic and industry studies published between 2021 and 2025, the findings were synthesised into five thematic areas: cost and economic considerations; security, compliance, and sovereignty; vendor lock-in and portability; performance and latency; and governance and exit strategy.

The results validate the existence and development of cloud repatriation as an organised policy decision, rather than just a technical change process. The two highest-priority areas in the reviewed studies are Cost and Economic Considerations (22.7% of the evidence base) and Governance and Exit Strategy (22.7%), while security, compliance, performance, and portability remain complementary but equally important dimensions of the current literature.

With regard to RQ1, organizations are primarily driven toward cloud repatriation due to increasing cloud costs, data sovereignty requirements, regulatory compliance pressures, vendor dependency, and performance issues when it comes to latency-sensitive applications. In public sector settings, with inflexible procurement processes and no standardised frameworks of governance, these pressures are exacerbated in real-life examples such as Dropbox, X (Twitter), GEICO, and 37signals.

For RQ2, several solution strategies are identified for each CR factor. They include predictive TCO modelling and cost governance for economic factors, TOE-based co-governance and automated compliance auditing for security and sovereignty, cloud-agnostic architectures and contractual portability provisions to mitigate vendor lock-in, selective repatriation and edge computing integration for performance needs, and lifecycle-based governance frameworks to assist in exit strategy management. Combined, they represent an integrated strategy that recognizes the need to align financial, technical, regulatory, and governance priorities together for successful repatriation.

In the context of RQ3, the longitudinal trend analysis reveals three distinct phases of CR research evolution: foundational recognition of performance, security, and portability concerns (2021–2022), convergence of academic and industry perspectives driven by cost and vendor lock-in (2023–2024), and a governance turn marked by the dominance of governance-oriented publications (2025). This trajectory confirms that CR studies are maturing from descriptive factor identification toward prescriptive, policy-driven frameworks, particularly relevant for regulated and public sector environments.

Overall, the three contributions of this review, the identification of CR factors, the synthesis of solution strategies, and the mapping of implementation limitations, offer a solid foundation for creating integrated, policy-compliant and sovereignty-aware models of CR that can be applied in private and public sector settings.

Nevertheless, this review acknowledges several limitations. The analysis is bounded by English-language sources published between 2021 and 2025, which may exclude earlier foundational works and underrepresent CR studies from non-English-speaking regions. Additionally, the predominance of private sector and Western organizational contexts in the reviewed literature may limit the generalizability of findings across diverse regulatory environments and governance structures globally. Consequently, foundational governance and portability studies published before 2021, as well as non-English-language literature that may offer region-specific insights (e.g., from Asia-Pacific or Nordic public sectors), fall outside this review's scope and represent a direction for future comparative synthesis.

Building on the gaps identified in this review, future empirical research should prioritise: 1) the development and validation of a quantifiable Cloud Repatriation Governance Maturity Index across multiple regulated public sector organisations; 2) the establishment of longitudinal cost and performance measurement instruments to compare repatriated and cloud-hosted workloads over a three- to five-year period; and 3) the conduct of cross-national comparative studies to extend the current evidence base beyond its predominantly Western and private-sector orientation. Collectively, these research directions would strengthen the empirical basis for assessing the governance maturity, long-term economic viability, and operational effectiveness of cloud repatriation in regulated public sector environments.

ACKNOWLEDGMENT

This study is supported by the National University of Malaysia (UKM). Sincerely grateful for the facilities and funding provided, which were essential for completing this research.

REFERENCES

[1] eG Innovations Blog, “The importance of a cloud exit strategy: What it is, who needs it, and how to plan it.” [Online]. Available: https://www.eginnovations.com/blog/the-importance-of-a-cloud-exit-strategy-what-it-is-who-needs-it-and-how-to-plan-it/

[2] eG Innovations Blog, “On-premises, cloud-first or cloud repatriation.” [Online]. Available: https://www.eginnovations.com/blog/on-premises-cloud-first-or-cloud-repatriation/

[3] G. K. Murugesan, “Cloud Services — Boon or Bane: A Comprehensive Review,” in SoutheastCon 2024, Atlanta, GA, USA: IEEE, Mar. 2024, pp. 108–112. doi: 10.1109/SoutheastCon52093.2024.10500027.

[4] A. S. George, “The Cloud Comedown: Understanding the Emerging Trend of Cloud Exit Strategies,” Partners Universal International Innovation Journal (PUIIJ), vol. 2, no. 5, Oct. 2024, doi: 10.5281/ZENODO.13993933.

[5] Subhadip Kumar, “Cloud Repatriation: A Strategic Move to Optimize IT Infrastructure and Costs,” 2024, doi: 10.13140/RG.2.2.25731.07209.

[6] D. T. Melanie Posey, “Cloud repatriation: What it is, what it isn’t, and why it’s not going away,” 2021, [Online]. Available: https://www.spglobal.com/content/dam/spglobal/mi/en/documents/general/cloud-repatriation-what-it-is-what-it.pdf

[7] CIO, CSO and COMPUTERWOCHE Research Services in cooperation with and Lufthansa Industry Solutions, “Studie Cloud-Migration 2023,” CIO, CSO und COMPUTERWOCHE, 2023. [Online]. Available: https://www.lufthansa-industry-solutions.com/de-de/studien/folgestudie-cloud-migration-2023-von-cio-cso-und-computerwoche

[8] CIO, CSO and COMPUTERWOCHE Research Services in cooperation with and Lufthansa Industry Solutions, “Studie Cloud-Migration 2025,” 2025. [Online]. Available: https://www.lufthansa-industry-solutions.com/de-en/studies/cloud-migration-2025-foundry-study-2025

[9] Sunbird, “3 Companies That Repatriated Workloads from the Cloud and Their Results.” [Online]. Available: https://www.sunbirddcim.com/blog/3-companies-repatriated-workloads-cloud-and-their-results

[10] Puppet, “Cloud Repatriation: Examples, 2025 Trends & Tips for Reverse Migration.” [Online]. Available: https://www.puppet.com/blog/cloud-repatriation

[11] K. Jewargi, “Public Cloud to Cloud Repatriation Trend,” Sch J Eng Tech, vol. 11, no. 1, pp. 1–3, Jan. 2023, doi: 10.36347/sjet.2023.v11i01.001.

[12] Laxminarayana Korada and Vijay Kartik Sikha, “Why are large enterprises building private clouds after their journey on public clouds?,” Feb. 2024, doi: 10.5281/ZENODO.13326719.

[13] David Heinemeier Hansson, “X celebrates 60% savings from cloud exit.” [Online]. Available: https://world.hey.com/dhh/x-celebrates-60-savings-from-cloud-exit-7cc26895

[14] Priyanka Gowda Ashwath Narayana Gowda, “Migrating Banking Applications to the Cloud: Strategies and Best Practices,” Dec. 2021, doi: 10.5281/ZENODO.13950711.

[15] Everest Group, “Navigating cloud portability and exit strategies in banking and financial services.” [Online]. Available: https://www.everestgrp.com/it-services/navigating-cloud-portability-and-exit-strategies-in-banking-and-financial-services-blog.html

[16] S. Grabski and P. Schmidt, "Multi-Cloud Environments and Governance: Repatriation or Expansion?," presented at the Hawaii International Conference on System Sciences, 2025. doi: 10.24251/HICSS.2025.713.

[17] Ponemon Institute LLC, "The Cost of Cloud Compromise and Shadow IT," Ponemon Institute LLC, Apr. 2021. [Online]. Available: https://www.proofpoint.com/us/resources/analyst-reports/cost-of-cloud-compromise-and-shadow-it

[18] I. Räisänen, "Cloud Repatriation: factors and strategies for successfully and beneficially leaving the public cloud:," 2025, [Online]. Available: https://urn.fi/URN:NBN:fi:tuni-202502031930

[19] BigID Blog, "Cloud migration trends and statistics 2023." [Online]. Available: https://bigid.com/blog/cloud-migration-trends-and-statistics-2023/

[20] F. Jose Eduardo, "Cloud Repatriation: Strategic Considerations, Technical Impacts, and the Future of DevOps and Security," Sep. 2024, [Online]. Available: https://www.academia.edu/124182751/Cloud_Repatriation_Strategic_Considerations_Technical_Impacts_and_the_Future_of_DevOps_and_Security

[21] Forbes Tech Council, "The rise of cloud repatriation: Why companies are bringing data in-house." [Online]. Available: https://www.forbes.com/councils/forbestechcouncil/2023/04/18/the-rise-of-cloud-repatriation-why-companies-are-bringing-data-in-house/

[22] InfoWorld, "Why companies are leaving the cloud." [Online]. Available: https://www.infoworld.com/article/2336102/why-companies-are-leaving-the-cloud.html

[23] H. Chang, "The Differences and Advantages between Cloud Services and Traditional Services," in 2022 8th Annual International Conference on Network and Information Systems for Computers (ICNISC), Hangzhou, China: IEEE, Sep. 2022, pp. 497–499. doi: 10.1109/ICNISC57059.2022.00104.

[24] O. Abied, O. Ibrahim, and S. N.-I. Mat Kamal, "Adoption of Cloud Computing in E-Government: A Systematic Literature Review," JST, vol. 30, no. 1, pp. 655–689, Jan. 2022, doi: 10.47836/pjst.30.1.36.

[25] B. Mallikarjuna and M. Kumar, "Analysis of the Performance, Scalability, Availability, and Security of Cloud Computing in Different Cloud Environments," in 2022 International Conference on Futuristic Technologies (INCOFT), Belgaum, India: IEEE, Nov. 2022, pp. 1–7. doi: 10.1109/INCOFT55651.2022.10094446.

[26] V. Sandhiya and A. Suresh, "Analysis Of Performance, Scalability, Availability And Security In Different Cloud Environments For Cloud Computing," in 2023 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India: IEEE, Jan. 2023, pp. 1–7. doi: 10.1109/ICCCI56745.2023.10128351.

[27] Fazlina Ali, M. M. Daud, H. Sallehudin, N. A. M. Yunus, A. Ahmad, and S. S. Hussein, "Multi-Factor Replica Placement Strategy for Resource Failure Extenuation in Cloud Replication Environment," IEEE Access, vol. 13, pp. 187552–187563, 2025, doi: 10.1109/ACCESS.2025.3625943.

[28] M. A. Fazlina, R. Latip, H. Ibrahim, and A. Abdullah, "Replication Strategy with Comprehensive Data Center Selection Method in Cloud Environments," Computers, Materials & Continua, vol. 74, no. 1, pp. 415–433, 2023, doi: 10.32604/cmc.2023.020764.

[29] H. Hasimi Sallehudin, R. Che Razak, M. Ismail, A. F. Md Fadzil, and R. Baker, "Cloud Computing Implementation in The Public Sector: Factors and Impact," APJITM, vol. 07, no. 02(02), pp. 27–42, Dec. 2018, doi: 10.17576/apjitm-2018-0702(02)-03.

[30] N. M. Noh, N. S. M. Satar, H. Sallehudin, I. H. Mallam, and S. S. Hussein, "Modelling Cloud Computing Adoption in the Malaysian Healthcare," International Journal of Advanced Computer Science and Applications, vol. 16, no. 7, 2025.

[31] H. M. Ibrahim, K. Ahmad, and H. Sallehudin, "Understanding the technology and humans as determinants of cloud computing adoption for digital preservation of research outputs in university libraries," Educ Inf Technol, vol. 30, no. 5, pp. 6163–6211, Apr. 2025, doi: 10.1007/s10639-024-13027-w.

[32] Edward Targett, "GEICO repatriates work from the cloud, continues ambitious infrastructure overhaul." [Online]. Available: https://www.thestack.technology/warren-buffetts-geico-repatriates-work-from-the-cloud-continues-ambitious-infrastructure-overhaul/

[33] Antony Panteli, "Why cloud repatriation is back on the CIO agenda." [Online]. Available: https://www.cio.com/article/4061031/why-cloud-repatriation-is-back-on-the-cio-agenda.html

Multi-Agent Crowd-Risk Prioritization for Large-Scale Events: A Hybrid ML–Agent Approach

Najwa Samrgandi
Department of Data Science-College of Computing, Umm Al-Qura University, Makkah, Saudi Arabia

***Abstract*—Managing crowd safety during large-scale gatherings such as Hajj requires timely assessment of multiple risk factors, including crowd conditions, environmental factors, and health-related indicators. Many existing approaches address these aspects separately, making it difficult to obtain a unified operational view of risk. This study presents a multi-agent framework that combines machine learning with domain-specific agent coordination for crowd-risk prioritization. The framework includes five agents responsible for crowd monitoring, health monitoring, environmental sensing, coordination, and alert generation. To reduce data leakage, variables used in risk-label construction and post-event outcomes were excluded before model training, resulting in 17 real-time observable features. The framework was evaluated using the Hajj and Umrah Crowd Management dataset containing 10,000 simulated records. On the independent test set, XGBoost+SMOTE achieved 47.30% accuracy and 46.22% F1-score; the XGBoost baseline achieved 57.15% accuracy and 45.96% F1-score, and a majority-class baseline that always predicts the dominant class achieved 58.70%, underscoring that overall accuracy is an insufficient metric for this task. Although accuracy decreased after class rebalancing, minority-class sensitivity improved: XGBoost+SMOTE achieved partial High-risk recognition (recall = 0.14, F1 = 0.12), whereas the XGBoost baseline and Cost-Sensitive XGBoost failed to identify any High-risk cases (recall = 0.00). McNemar's test confirmed a statistically significant difference between the XGBoost baseline and the SMOTE-enhanced model ($\chi^2 = 69.218$, $p < 0.001$). Cross-validation on the SMOTE-balanced training data produced 59.89% ± 0.20% accuracy and 58.94% ± 0.17% F1-score. A streaming evaluation on 500 held-out test records showed that the Full Framework generated 93 High-risk alerts, compared with 74 from the ML-only configuration and 23 from the Agents-only configuration, a 25.7% increase relative to ML-only operation. A precision check against the internal risk-score proxy found that none of the 19 agent-driven escalations beyond ML-only predictions corresponded to a High-risk label under this proxy; these alerts should accordingly be read as operational risk-prioritization signals rather than confirmed incident detections. All reported figures should further be interpreted in light of the dataset's synthetic origin: because features were algorithmically generated rather than collected from real sensors or incidents, the results are illustrative of the framework's behavior under these specific simulated conditions rather than predictive of real operational deployment. Average processing latency was 4.127 ms per record under sequential single-stream conditions. In addition to the proposed framework, the study presents a diagnostic evaluation approach that combines leakage control, dual association analysis, and direct precision auditing. This approach may also prove useful for evaluating similar ML–agent systems in other safety-critical alerting applications.**

***Keywords*—Multi-agent systems; crowd-risk prioritization; intelligent crowd management; real-time IoT monitoring; SMOTE; XGBoost; safety-critical systems**

I. Introduction

Large public gatherings represent one of the most operationally complex environments for safety management. The annual Hajj pilgrimage, which draws more than two million participants to confined geographic zones over a compressed timeframe, exemplifies the challenges that arise when crowd density, environmental stress, and health vulnerability intersect. The scale and operational complexity of Hajj continue to motivate the development of intelligent crowd-management systems capable of supporting timely risk assessment and decision-making [1], [2].

The past decade has seen considerable progress in AI-driven crowd analysis. Vision-based systems using YOLO and CNN architectures have achieved strong performance in density estimation and abnormal behavior detection [3], [4], [5]. IoT sensor networks have extended monitoring beyond camera fields of view by capturing environmental and operational indicators associated with crowd conditions [6], [7]. Multi-agent frameworks have demonstrated value in evacuation simulation and distributed coordination [8], [9]. Yet these advances have largely evolved in parallel rather than in concert: visual monitoring systems rarely incorporate health indicators; IoT platforms seldom feed predictive risk models; and agent-based systems tend toward simulation rather than real-time operational deployment [10], [11].

This fragmentation carries real operational cost. A monitoring system that detects crowd compression via camera feeds cannot independently account for the additional risk posed by extreme heat, an aging pilgrim population, or extended waiting times at transport checkpoints. Conversely, rule-based alert systems operating without predictive components lack the capacity to differentiate between momentary crowd fluctuations and genuinely escalating risk trajectories; predictive agent-based architectures have been shown to address this limitation more effectively [12], [13].

This study addresses this gap through a framework designed around five principles: 1) asynchronous agent independence: each agent processes its own data stream without blocking on others; 2) leakage-free feature engineering: the feature space used for ML prediction is strictly isolated from the variables used to construct risk labels; 3) operationally grounded escalation: the coordination policy prioritizes sensitivity over specificity, reflecting the asymmetric cost of missing High-risk events; 4) statistical accountability: model improvements are

validated through McNemar's test; and 5) real-time feasibility: all predictions are generated with sub-10 ms latency per record under the evaluated experimental conditions.

The study is guided by three research questions: (RQ1) How can asynchronous agent coordination improve crowd-risk prioritization beyond ML prediction alone? (RQ2) What feature-selection strategy best prevents leakage in operationally constructed risk labels? (RQ3) How does the proposed framework perform under realistic class imbalance without artificially inflating reported performance?

II. Related Work

Crowd safety research has matured across three intersecting streams. The first concerns AI-based crowd analysis. Vision-based architectures, particularly YOLOv8 and YOLOv9, have achieved near-real-time detection of density anomalies and abnormal behaviors [3], [14], [15], [16]. CNN-based approaches extended this to behavioral trajectory prediction [17], [18]. Cyber-physical integrations further embedded visual monitoring within broader operational sensing architectures [19], [20]. However, these systems remain primarily reactive: they detect what is already happening rather than flagging converging risk conditions.

The second stream concerns IoT-enabled monitoring. Distributed sensor networks capture environmental variables that co-occur with crowd incidents but fall outside camera fields of view [21], [22], [23]. Zhang et al. proposed an IoT-based public safety alert system that combines distributed multi-sensor inputs with lightweight edge-deployed machine learning models to generate alerts, demonstrating the feasibility of automated, low-latency alert pipelines for emergency response support [7]. While [7] incorporates basic predictive classification at the sensor level, it does not address multi-domain risk fusion across crowd, health, and environmental indicators simultaneously, nor does it consider data-leakage risks in label construction. The broader limitation across this stream of work is a continued reliance on per-sensor threshold or single-model logic rather than coordinated multi-agent risk fusion across heterogeneous domains [24].

The third stream concerns multi-agent coordination. Keykhaei et al. developed a situation-aware multi-agent evacuation model for earthquake scenarios, demonstrating that distributed agent coordination enables more adaptive response than static centralized planning in dynamic crowd conditions [9]. Dumitrescu et al. used multi-agent modeling to simulate crowd panic propagation [8]. More recent work has explored LLM-based agents for smart city coordination [12] and multi-agent machine learning frameworks combining classification with coordinated decision-making for real-time safety response [25]. Despite this breadth, most multi-agent applications remain simulation-oriented, with few validated deployments in operational real-time monitoring systems [26], [27], [28], [29].

The most directly comparable work is CIRS [25], a five-agent machine learning-based framework for real-time accident detection and emergency response, whose architecture (perception, classification, description, communication, and decision-making agents) closely parallels the five-agent design adopted in the present framework (Crowd Monitoring, Health Monitoring, Environmental Sensing, Coordinator, and Alert agents). The present framework differs from [25] in several respects. First, [25] is applied to a structurally distinct domain (road-traffic accident detection from CCTV video) rather than multi-domain crowd-risk prioritization, and accordingly integrates a vision-language model for incident description rather than tabular sensor-derived risk classification. Second, [25] reports classifier performance (YOLOv11: precision 0.866, recall 0.855) without an explicit discussion of label-construction leakage, since its ground truth (manually annotated accident frames) is independent of its input features by design, whereas the present framework addresses a leakage risk that is structurally specific to operationally derived risk labels. Third, [25] does not report a formal statistical test of agreement between its decision-making agent and its underlying classifier, an aspect the present framework addresses directly through McNemar's test (Section IV-C). A complementary comparison point is provided by Alshamrani et al. [30], who developed a wearable-sensor-based system for monitoring pilgrim stress and fatigue during Hajj, combining physiological, geospatial, and self-reported data; unlike the present framework, [30] is oriented toward descriptive analysis of collected pilgrim data rather than real-time multi-agent risk coordination, and does not incorporate a predictive classification component or an agent-based architecture. Taken together, [25] offers the closest architectural parallel (multi-agent coordination with an embedded ML classifier) while [30] offers the closest domain parallel (Hajj-specific physiological and environmental monitoring); neither addresses both dimensions simultaneously, which motivates the integrated evaluation presented in this study. These comparisons should be interpreted as descriptive positioning relative to the closest available published work rather than as a direct empirical benchmark. Table I summarizes the major research streams and their core limitations.

TABLE I. Summary of Major Research Streams and their Limitations.

Research Stream	Core Limitation	Key References
AI/deep learning crowd monitoring	Reactive; limited integration with health or environmental sensing	[3]-[5],[14]-[20]
IoT-enabled sensing and alerting	Threshold-based; no adaptive learning; no multi-source fusion	[6],[7],[21]-[24]
Multi-agent coordination frameworks	Simulation-oriented; limited real-time operational deployment	[8],[9],[12],[25]-[29]

Theoretical perspectives on multi-agent decision-making [31] and a broader multi-expert analysis of AI agentic systems and their deployment characteristics [32] further contextualize the coordination architecture adopted in this framework. Distinct from this general perspective, a multi-source information fusion technique specifically targeting abnormal crowd behavior detection [33] reinforces the operational complementarity of data-driven and rule-based sensing that underpins the proposed design. The review of prior studies suggests that current approaches address specific dimensions of crowd safety but rarely combine them within a single real-time architecture. This gap motivates the development of a framework that integrates predictive machine learning,

heterogeneous sensing, and coordinated agent-based responses to support proactive risk management.

III. METHODOLOGY

A. Dataset and Representativeness

The Hajj and Umrah Crowd Management dataset (Kaggle) comprises 10,000 records across 30 features capturing spatial coordinates, environmental readings, operational waiting times, health indicators, and pilgrim demographic profiles. The dataset contains no missing values and no duplicate records. All features were generated under simulated operational conditions designed to reflect common Hajj crowd management scenarios; their ability to fully represent real-world crowd behavior remains a study limitation.

Risk labels were derived through a weighted operational scoring function: Crowd_Density (0.25), Temperature (0.20), Fatigue_Level (0.15), Stress_Level (0.15), Queue_Time_minutes (0.10), Emergency_Event (0.10), and Distance_Between_People_m (0.05). The resulting score was discretized into three thresholds: Low (below 1.8), Medium (1.8–2.4), and High (above 2.4), yielding 3,121 Low (31.2%), 5,869 Medium (58.7%), and 1,010 High (10.1%) records. These labels represent operationally constructed risk proxies rather than verified emergency records. The framework is accordingly best characterized as a risk-prioritization tool.

B. Feature Engineering and Leakage Prevention

Data leakage represents one of the most consequential validity threats in applied machine learning research, capable of inflating reported accuracy without reflecting genuine predictive capacity [34], [35]. In this dataset, the risk is structurally elevated: the seven features used to compute the risk label were excluded from model input. Post-event outcome variables (Satisfaction_Rating, Perceived_Safety_Rating, Pilgrim_Experience, AR_Navigation_Success) were also excluded as they are unavailable at operational prediction time.

The resulting clean feature set comprises 17 real-time observable indicators: spatial (Location_Lat, Location_Long), movement (Movement_Speed), environmental (Weather_Conditions, Sound_Level_dB), operational (Activity_Type, Event_Type, Waiting_Time_for_Transport, Security_Checkpoint_Wait_Time, Time_Spent_at_Location_minutes, Interaction_Frequency), pilgrim profile (Age_Group, Health_Condition, Transport_Mode), and contextual indicators (Incident_Type, Crowd_Morale, AR_System_Interaction).

C. Data Partitioning and Class Balancing

The dataset was divided into 80% training (8,000 records) and 20% test (2,000 records) using stratified random splitting with seed 42. The test set was held out entirely throughout training, tuning, and oversampling. SMOTE [36] was applied exclusively to the training set following the split to address class imbalance [35], expanding the minority High and Low classes to match the majority Medium class at 4,695 samples each. Cross-validation on the SMOTE-augmented training set produces accuracy estimates under balanced class conditions (33.3% per class), which naturally exceed test-set accuracy under the original imbalanced distribution. Both metrics are reported with their interpretations distinguished in Section IV.

D. Prediction Models and Experimental Setup

Seven configurations were evaluated: Random Forest [37] (n=150, depth=10), XGBoost baseline [38] (n=200, depth=6, lr=0.05), LightGBM baseline (n=200, depth=6, lr=0.05), RF+SMOTE, XGBoost+SMOTE, LightGBM+SMOTE, and Cost-Sensitive XGBoost (scale_pos_weight calibrated to the High-to-Medium class ratio). LightGBM was included to assess whether gradient-boosting performance generalizes across implementations; Cost-Sensitive XGBoost was included to test class-weighting as an alternative to oversampling for minority-class detection. Cost-sensitive learning addresses class imbalance by assigning higher penalties to misclassification of minority-class instances, thereby encouraging the model to place greater emphasis on underrepresented classes [39]. All models used seed 42. Weighted-average and per-class precision, recall, and F1-score were computed to expose behavior on the minority High-risk class. McNemar's test with continuity correction assessed statistical significance between XGBoost baseline and XGBoost+SMOTE at alpha = 0.05, following established recommendations for comparing classifier performance on a single shared test set [40].

E. Multi-Agent Framework Architecture

The framework comprises five agents implemented as independent Python threads communicating through dedicated in-memory message queues, consistent with the reactive agent architecture described by Wooldridge [41] and the general multi-agent coordination workflow surveyed by Maldonado et al. [26]. The Crowd Monitoring Agent observes Movement_Speed, Interaction_Frequency, and Crowd_Morale. It applies threshold rules synthesized from crowd dynamics simulation literature [8], [9]: a Medium signal is triggered when movement speed falls below 0.5 m/s, consistent with documented pre-compression crowd conditions, or when interaction frequency exceeds 7; a High signal is triggered when speed falls below 0.3 m/s and frequency exceeds 8 simultaneously. These thresholds represent expert-informed approximations derived from simulation findings and are subject to empirical calibration against real incident data.

The Health Monitoring Agent observes Health_Condition, Age_Group, and Waiting_Time_for_Transport. Pilgrims presenting dehydration or fatigue, those aged 70 or above, or those waiting more than 90 minutes for transport receive elevated health risk signals. These thresholds were informed by health management frameworks for crowded events [42] and constitute expert-informed approximations subject to empirical calibration. The combination of dehydration and age above 70 triggers High. The Environmental Sensing Agent observes Weather_Conditions, Sound_Level_dB, and Security_Checkpoint_Wait_Time. Adverse weather or sound levels above 80 dB trigger Medium; sandstorm with sound above 85 dB triggers High. All agent thresholds are expert-informed approximations pending empirical validation.

The Coordinator Agent blocks on three message queues until signals from all monitoring agents arrive, then queries the XGBoost+SMOTE model with the encoded feature vector. The final risk decision applies a maximum-escalation policy: the

highest risk level across all four inputs is selected. The Alert Agent converts the final decision into an operational notification.

The proposed architecture differs structurally from a conventional centralized pipeline in several respects. Table II summarizes the key architectural distinctions.

TABLE II. ARCHITECTURAL COMPARISON: CENTRALIZED PIPELINE VS. PROPOSED FRAMEWORK.

Centralized Pipeline	Proposed Framework
Single processing module	Five domain-specific agents
Shared decision logic	Independent domain-specific assessment
Tightly coupled components	Modular, independently operable components
Limited extensibility	New agents can be added without modifying existing ones
Sequential domain processing	Asynchronous domain-specific monitoring

The present framework does not claim decentralized negotiation or autonomous multi-agent planning. Rather, it adopts a modular, coordinator-based multi-agent architecture in which independent domain-specific agents perform local assessment before a coordinator integrates their outputs through a conservative escalation policy [13]. The architecture was selected for modularity, extensibility, and separation of domain-specific reasoning rather than to claim computational superiority over an equivalent centralized implementation. This design allows each agent to be developed, tested, and replaced independently, and permits the addition of new sensing domains without restructuring the coordination logic.

In terms of computational complexity, each monitoring agent evaluates threshold conditions over a fixed three-feature subset, contributing O(1) time per agent and per record. The Coordinator Agent then performs a single XGBoost inference over 17 features. Gradient-boosted tree inference requires $O(T \cdot D)$ time per record, where T is the number of trees and D is the maximum tree depth. Because T and D are fixed hyperparameters in the present implementation ($T = 200$, $D = 6$), the inference cost is constant with respect to dataset size and the number of monitoring agents. Inter-agent communication overhead scales as O(A), where A denotes the number of monitoring agents. The overall per-record complexity is therefore $O(A + T \cdot D)$; with A, T, and D fixed in the current five-agent architecture, it is O(1) with respect to dataset size, while increasing linearly if the number of monitoring agents is expanded. The framework architecture is illustrated in Fig. 1.

In the architecture of the proposed multi-agent framework, Independent Crowd, Health, and Environmental agents publish risk signals through dedicated message queues. The Coordinator Agent integrates these signals with XGBoost+SMOTE predictions using a maximum-escalation policy to generate the final risk assessment. Results are based on 500 test records held out from the training data (see Section III-F for the relationship between this streaming subset and the broader test set used in Section IV-A).

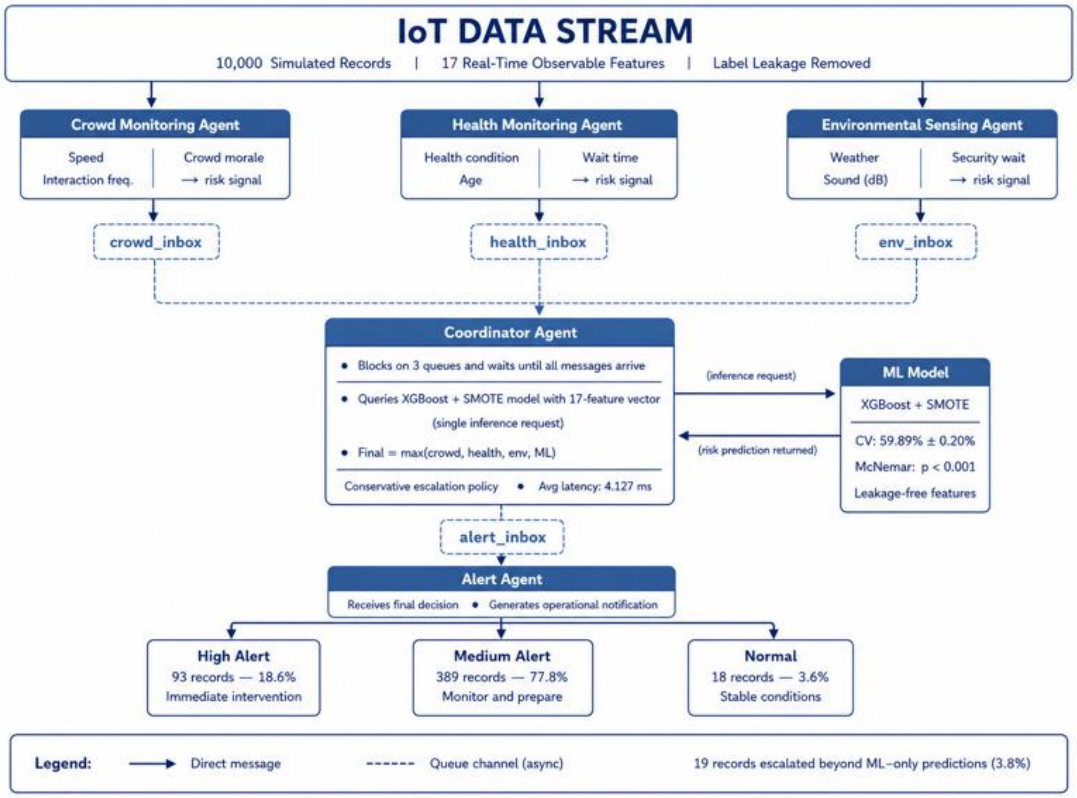


Fig. 1. Proposed multi-agent framework for crowd-risk prioritization.

F. Evaluation Protocol

Model performance was assessed on the held-out test set using overall and per-class metrics. Robustness was evaluated through 3-fold stratified cross-validation on the SMOTE-augmented training set, reporting Mean plus or minus Std. McNemar's test with continuity correction assessed statistical significance. The multi-agent framework was evaluated in three configurations (ML-only, Agents-only, and Full Framework) on 500 records drawn exclusively from the held-out test set, ordered by Timestamp to simulate a streaming environment; these 500 records are a subset of the same 2,000-record test set used for the overall and per-class results reported in Section IV- A, rather than an independently drawn fresh sample, and the ablation results should accordingly be read as a re-examination of the existing held-out test set from a streaming perspective rather than as validation on new, previously unseen data. Sensitivity analysis was conducted by varying all agent thresholds uniformly by plus or minus 20% from baseline values and computing High-risk alert counts experimentally across the three threshold configurations.

IV. Results and Discussion

A. Model Performance

Table III presents overall test-set results for seven configurations. Table IV presents per-class results for the XGBoost and LightGBM SMOTE-enhanced configurations. On the test set, XGBoost+SMOTE achieved 47.30% accuracy (F1: 46.22%); cross-validation on the SMOTE-balanced training data yielded 59.89% plus or minus 0.20% accuracy and 58.94% plus or minus 0.17% F1 for XGBoost+SMOTE. The gap between CV accuracy (59.89%) and test-set accuracy (47.30%) reflects evaluation under balanced (33.3% per class) versus imbalanced (Medium 58.7%) class conditions, respectively, and is reported explicitly to avoid misrepresentation of operational performance. To quantify the stability of test-set performance beyond the single fixed split, non-parametric bootstrap resampling (1,000 resamples with replacement, drawn from the test-set predictions) was used to construct 95% confidence intervals: accuracy 47.35% [45.20%, 49.55%], F1-score 46.26% [43.96%, 48.59%], and High-risk recall 13.76% [9.23%, 18.60%]. These intervals indicate that the reported point estimates are reasonably stable under resampling, while the width of the High-risk recall interval reflects the limited number of High-risk test instances available for evaluation.

Table IV reveals a critical finding: the XGBoost baseline and Cost-Sensitive XGBoost detected zero High-risk cases (precision = 0.00, recall = 0.00, F1 = 0.00), while XGBoost+SMOTE achieved partial High-risk detection (recall = 0.14, F1 = 0.12) and LightGBM+SMOTE achieved comparable performance (recall = 0.12). The baseline model's single false positive prediction for the High-risk class (1 of 1,174 true Medium-risk records misclassified as High) was insufficient to register a non-zero precision value at two decimal places, confirming that the model produced no usable High-risk signal whatsoever on the independent test set. The inability of baseline models to detect any High-risk instance is not a model failure per se but reflects the intrinsic difficulty of the task: after leakage removal, the observable features provide insufficient discriminative signal to separate High- from Medium-risk without class rebalancing. This finding directly motivates the domain-aware agent escalation layer, which captures High-risk configurations through rule-based domain knowledge that the baseline ML classifier does not detect under the current feature set. This underscores that overall accuracy is an insufficient metric for safety-critical systems. McNemar's test confirmed a statistically significant improvement of XGBoost+SMOTE over the baseline (chi-squared = 69.218, p less than 0.001), indicating a statistically significant shift in minority-class sensitivity, mechanistically explained in Section V as a decision-boundary effect rather than a genuine signal-recovery effect on the independent test set.

To contextualize these results against trivial reference points, two simple baselines were also evaluated on the same test set: a majority-class baseline that always predicts the dominant class (Medium) and a stratified-random baseline that predicts classes according to their training-set proportions. The majority-class baseline achieved 58.70% accuracy and 43.42% F1-score (Table III); the stratified-random baseline achieved 42.20% accuracy and 42.31% F1-score. Notably, the majority-class baseline's accuracy (58.70%) exceeds that of the XGBoost baseline (57.15%), although the XGBoost baseline achieves a higher F1-score (45.96% vs. 43.42%) by producing non-trivial, if ultimately unsuccessful, predictions for the Low-risk class. This comparison reinforces rather than undermines the study's central argument: it provides direct, model-agnostic confirmation that overall accuracy is an insufficient and potentially misleading metric for this task, since a model that makes no use of any observable feature can match or exceed the accuracy of a trained classifier, consistent with the near-absence of linear and nonlinear feature–label association documented in Section V. The comparison is included here, prior to that analysis, specifically to establish this point empirically before it is explained mechanistically. The headline XGBoost+SMOTE figures reported above were further confirmed to be stable, rather than dependent on the specific random seed used, via a repeated-splits analysis across 10 independent seeds reported in Section VII (Statistical Validity).

TABLE III. Overall Model Performance on Independent Test Set

Model	Acc.%	Prec.%	Rec.%	F1%
Random Forest	58.75	50.10	58.75	43.64
XGBoost Baseline	57.15	45.12	57.15	45.96
RF + SMOTE	48.50	45.81	48.50	44.62
XGBoost + SMOTE	47.30	45.90	47.30	46.22
LightGBM Baseline	57.00	44.40	57.00	45.72
LightGBM + SMOTE	46.55	44.93	46.55	45.34
XGBoost Cost-Sensitive	57.15	45.12	57.15	45.96
Majority-Class Baseline	58.70	34.46	58.70	43.42
Stratified-Random Baseline	42.20	42.31	42.20	42.31

TABLE IV. PER-CLASS PERFORMANCE: XGBOOST MODELS (TEST SET)

Model	Class	Prec.	Recall	F1
XGBoost Baseline	High	0.00	0.00	0.00
	Low	0.34	0.07	0.11
	Medium	0.59	0.94	0.72
XGBoost + SMOTE	High	0.11	0.14	0.12
	Low	0.31	0.22	0.26
	Medium	0.60	0.67	0.63
LightGBM + SMOTE	High	0.10	0.12	0.11
	Low	0.31	0.21	0.25
	Medium	0.59	0.66	0.62

Table IV shows that the baseline XGBoost model failed to identify any High-risk cases, with a recall of 0.00. After applying SMOTE, the model began to detect this class, although performance remained limited, with a recall of 0.14 and an F1-score of 0.12. SMOTE also increased Low-risk recall from 0.07 to 0.22. However, this came at the cost of reducing Medium-risk recall from 0.94 to 0.67. These results indicate that rebalancing reduced the model's dependence on the dominant Medium-risk class and allowed better, though still imperfect, detection of minority risk levels.

B. Feature Importance and Interpretability

Feature importance scores from the XGBoost+SMOTE model show a balanced distribution: the highest-ranked feature (Crowd_Morale, 0.0995) contributes less than 10% of total importance. This contrasts with analyses that included label-derived variables, where a single feature dominated, indicating leakage. Crowd_Morale, Event_Type, and Weather_Conditions are operationally interpretable: crowd morale reflects collective behavioral tension; event type determines spatial organization and density thresholds; weather conditions directly affect crowd movement. Fig. 2 illustrates the top 10 feature importance scores.

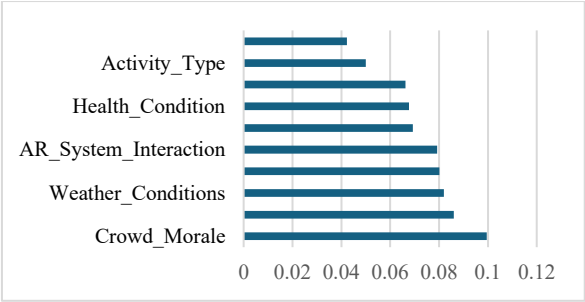


Fig. 2. Top 10 feature importance scores from the XGBoost+SMOTE model.

Because Crowd_Morale emerged as the single most important feature, its operational measurability warrants clarification. In the present dataset, Crowd_Morale is recorded as a discrete categorical indicator (Positive, Neutral, Negative); it is not derived from facial-expression or sentiment analysis, which would introduce separate privacy and accuracy concerns in dense crowd settings. In an operational deployment, this indicator would instead be inferred indirectly from signals already collected by the framework's other sensing channels (for example, aggregate crowd-noise level (Sound_Level_dB), movement irregularity (Movement_Speed variance), and local interaction density (Interaction_Frequency)) rather than measured directly. This indirect-inference approach is consistent with the framework's real-time observability requirement, but it also means that Crowd_Morale should be understood as a composite behavioral proxy rather than a directly sensed physical quantity, and its operational construction from these underlying signals remains an item for future validation.

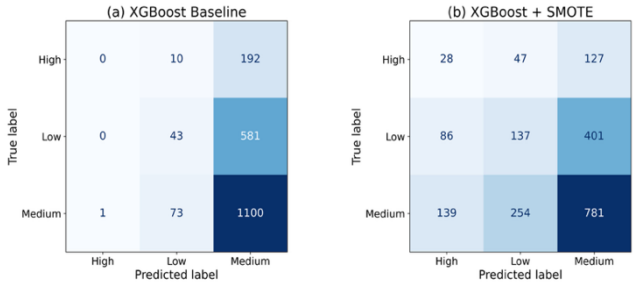


Fig. 3. Confusion matrices for (a) XGBoost baseline and (b) XGBoost+SMOTE on the independent test set.

Fig. 3 shows the confusion matrices for both models. A review of the confusion matrices reveals notable differences between the two models. The original XGBoost classifier predominantly predicted the Medium-risk class, leading to no correct predictions for High-risk observations. After incorporating SMOTE, the classifier identified a small portion of the High-risk cases. Although this increased sensitivity to the minority class, it also resulted in a lower overall accuracy. This pattern is in agreement with the performance metrics reported in Table IV.

To assess discriminative performance across the full range of decision thresholds rather than at the single fixed operating point reported in Table IV, a precision-recall curve was constructed for the High-risk class using the predicted class probabilities from XGBoost+SMOTE on the full 2,000-record independent test set described in Section IV-A (Fig. 4); this is distinct from, and substantially larger than, the 500-record streaming subset used in the ablation and escalation-precision analyses in Section IV-C below. The curve achieved an average precision (AP) of 0.1016, compared with an AP of 0.1010 expected under random ranking at this class's prevalence (10.1% of test-set records). The observed AP exceeds the random-baseline value by only 0.0006, and Fig. 4 shows the model's curve tracking the random-baseline reference line closely across nearly the entire recall range, with the exception of a brief, high-variance fluctuation at very low recall where the small number of underlying records makes individual precision estimates unstable. This result provides a third, independent line of evidence supporting the same conclusion reached via the majority-class baseline comparison above and the mutual information analysis in Section V: the trained classifier's ranking of records by predicted High-risk probability carries negligible discriminative value beyond what would be expected from the class's base rate alone, across the entire space of possible decision thresholds rather than only at the single threshold used to generate Table IV.

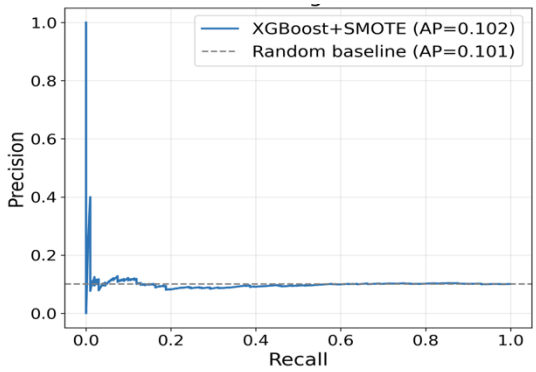


Fig. 4. Precision-Recall curve for High-risk class detection (XGBoost+SMOTE) versus the random-baseline reference line at class prevalence.

C. Multi-Agent Coordination Value and Streaming Performance

Table V presents the ablation results across three configurations on 500 held-out test records ordered by Timestamp. The ML-only configuration generated 74 High-risk alerts, while the Agents-only configuration generated 23 High-risk alerts. The Full Framework generated 93 High-risk alerts, representing a 25.7% increase relative to the ML-only configuration and substantially exceeding either component operating independently. Because the Full Framework applies a maximum-escalation policy (its High-risk count can never fall below either individual component's, by construction; see Table V below), this difference alone does not establish symmetric, bidirectional disagreement between the two components. With this structural property in mind, the pattern is nonetheless consistent with the ML model and the agent layer each surfacing High-risk configurations that the other does not independently flag, suggesting complementary, if asymmetric, risk-assessment mechanisms. The ML model performs probabilistic classification across all 17 observable features, whereas the agents apply domain-specific escalation rules to targeted feature subsets, allowing each component to identify elevated-risk situations that may not be emphasized by the other.

TABLE V. ABLATION: HIGH-RISK ALERT GENERATION ACROSS THREE CONFIGURATIONS.

Configuration	High-Risk Alerts	Records Escalated	Avg. Latency
ML Only (XGBoost+SMOTE)	74	—	4.127 ms
Agents Only (rules, no ML)	23	N/A	N/A
Full Framework (ML + Agents)	93	19 (3.8%)	4.127 ms

Because the Full Framework configuration cannot, by construction, generate fewer High-risk alerts than either individual component on any given record, the discordance pattern between configurations is necessarily one-directional rather than a conventional bidirectional disagreement (see Table VI for full contingency tables). With this structural constraint in mind, the 70 additional High-risk alerts generated by the Full Framework beyond the Agents-only configuration (93 vs. 23) were confirmed via McNemar's test to represent a statistically significant difference ($\chi^2 = 68.01$, $p < 0.001$), and the 19 additional alerts generated beyond the ML-only configuration (93 vs. 74) were likewise confirmed to represent a statistically significant difference ($\chi^2 = 17.05$, $p < 0.001$); in both paired comparisons, all discordant records favored the Full Framework configuration ($b > 0$, $c = 0$). These alerts arise from records where the ML model identified elevated-risk configurations that did not satisfy agent escalation thresholds, and, conversely, from records where domain-specific agent rules flagged conditions the classifier had not independently identified as High-risk. Consistent with the note accompanying Table VI, the statistical tests above should be interpreted as confirming that a nonzero number of one-directional escalations occurred in each comparison, rather than as evidence of mutual, symmetric correction between the two components or of a strong discriminative advantage for either component individually. The underlying architectural rationale for combining ML and agent inputs nonetheless reflects a conceptually bidirectional design: the ML model and the agent rules are each capable of flagging risk configurations that the other does not independently classify as High-risk. A statistically significant increase in alert volume should not, on its own, be read as an operational improvement. This concern is well established in clinical alerting research: a systematic qualitative review of general practitioners' experiences with clinical decision support alerts found that a high volume of alerts with limited specificity or relevance was consistently reported as a driver of alert fatigue, reducing practitioner engagement with and trust in subsequent alerts, including clinically important ones [43]. While this concept originates in primary-care clinical alerting research rather than crowd-safety monitoring, the underlying mechanism, desensitization to frequent, low-value notifications reducing responsiveness to genuinely critical signals, is domain-general and transfers directly to the present context: an operational alerting system that increases alert volume without a corresponding increase in precision risks producing the same desensitization effect among crowd-safety monitoring personnel. As discussed below, the precision of the escalated alerts examined in this study was found to be low against the operational risk proxy; the alert-volume increase reported here should therefore be weighed against this risk rather than treated as an unambiguous benefit.

A separate check tested whether the agents' three-tier escalation logic (each agent first raising a record from Low to Medium, then separately from Medium to High) is actually necessary, or whether a single, simpler rule would produce the same alerts. The simpler version skips the intermediate Medium step entirely: it flags a record as High whenever any one of the three agents' original High-level conditions is met, with no per-agent Medium-escalation logic in between. On the same 500-record streaming evaluation, this simplified single-rule version produced exactly the same number of High-risk alerts as the full three-agent hierarchy: 93 in both cases. In other words, the extra Medium-level escalation step built into each agent did not add any High-risk alerts beyond what the three original High-level conditions already produced on their own, at least on this dataset and this evaluation window. This does not mean the five-agent design is unnecessary it remains useful for modularity and for adding more nuanced intermediate-risk logic later but it does mean that the specific benefit of the current three-tier hierarchy has not yet been demonstrated, and should not be assumed

without testing on data where Medium-risk conditions are more common or more closely tied to eventual High-risk outcomes.

TABLE VI. MCNEMAR'S TEST CONTINGENCY TABLES FOR ABLATION COMPARISONS (N=500)

Comparison	Both High	b (Comparator High, Full Not)	c (Comparator Not, Full High)	Both Not High
Full vs. ML-only	74	0	19	407
Full vs. Agents-only	23	0	70	407

Note: b and c are zero-cell-constrained by the maximum-escalation policy (Full Framework can never report fewer High-risk alerts than either individual component on a given record), which forces a one-directional discordance pattern in both comparisons. McNemar's test is designed for paired nominal comparisons and is widely recommended for comparing classifiers on a shared test set [40]; however, its standard interpretation assumes that discordant pairs can in principle occur in either direction. Because the maximum-escalation policy used here rules out one of the two discordant cells by construction, the resulting test is closer to confirming that a nonzero number of one-directional escalations occurred than to demonstrating conventional, symmetric disagreement between the compared configurations, and the reported statistics should be read with this structural constraint in mind. Continuity-corrected McNemar statistics are computed as $\chi^2 = (|b-c|-1)^2/(b+c)$.

To assess the stability of the framework with respect to threshold selection, a sensitivity analysis was conducted by varying all agent thresholds by ±20% from their baseline values. Reducing thresholds by 20% yielded 83 High-risk alerts (16.6%), whereas increasing thresholds by 20% yielded 105 High-risk alerts (21.0%). The baseline configuration produced 93 alerts (18.6%), representing an intermediate operating point within this range. This non-monotonic behavior reflects interactions among agent-specific escalation rules under the framework's maximum-escalation policy, where changes in individual thresholds can alter escalation pathways across the combined decision process. Despite these variations, the Full Framework consistently generated more High-risk alerts than the ML-only configuration (74 alerts), indicating that the reported contribution is not dependent on a single threshold setting.

The 500 streaming records were drawn exclusively from the held-out test set, ensuring complete independence from training data (this subset was drawn from the same 2,000-record test set evaluated in Section IV-A rather than from a separate sample), while preserving the original class distribution of approximately 10.1% High-risk, 31.2% Low-risk, and 58.7% Medium-risk. The increase from 74 High-risk alerts in the ML-only configuration to 93 alerts in the Full Framework is consistent with the ability of agent rules to surface risk configurations that the ML classifier may underweight due to feature overlap. Under the streaming evaluation, the average processing latency was 4.127 ms per record, indicating that the framework is suitable for single-stream operational monitoring scenarios. To assess scalability, per-record latency increased from 1.918 ms under sequential processing to 4.571 ms at 5 concurrent streams, 10.284 ms at 10 concurrent streams, and 15.538 ms at 20 concurrent streams, representing an approximately 8.1-fold increase at the highest concurrency level evaluated. These results suggest that large-scale multi-stream deployment would likely require a distributed or otherwise more scalable execution architecture.

To assess the operational precision of agent-driven escalations, the 19 records escalated by the agent layer beyond ML-only predictions (i.e., classified as Medium or Low by XGBoost+SMOTE but elevated to High by agent rules) were cross-checked against their underlying continuous risk scores and original risk labels. None of these 19 records (0%) corresponded to a true High-risk label; 14 (73.7%) were true Medium-risk and 5 (26.3%) were true Low-risk. Only 4 records (21.1%) had an underlying risk score within 0.2 of the High-risk threshold (2.4), indicating that these escalations predominantly targeted records that were not borderline cases. This finding indicates that, under the current label definition, agent-driven escalations beyond ML predictions exhibit low precision against the operational risk proxy, and should be interpreted strictly as sensitivity-oriented prioritization signals rather than validated risk detections. This result motivates recalibrating agent thresholds against verified incident data as a priority for future work, rather than treating the increased alert count alone as evidence of improved detection; such recalibration could draw on recent proposals to incorporate net benefit directly as an objective during model and threshold development, rather than applying it only as a post-hoc evaluation metric, which would allow the relative cost of missed High-risk cases versus false escalations to inform threshold selection explicitly [44].

V. DATA QUALITY, LIMITATIONS, AND FUTURE DIRECTIONS

The most fundamental limitation concerns label validity. Risk categories were derived from a weighted scoring function applied to operational indicators, not from verified incident records. The models therefore predict a computed proxy for risk rather than observed outcomes. One practical consequence is the near-zero separation between Medium- and High-risk feature distributions, resulting in zero High-risk recall for the XGBoost baseline and Cost-Sensitive XGBoost, and only partial recall (0.14) for XGBoost+SMOTE on the independent test set. In addition, agent-generated escalations could not be validated against confirmed incidents because the dataset does not include verified real-world outcomes. The additional High-risk alerts are therefore best understood as prioritization cues for further review, not as verified detections of actual incidents. Future work should evaluate these alerts against real incident logs or expert-reviewed operational records.

The weak observed performance also follows a near-deterministic statistical expectation given the dataset's construction. A post hoc correlation analysis between the 17 retained features and the continuous Operational_Risk_Score (prior to threshold discretization) found a maximum absolute Pearson correlation of 0.026 and a mean absolute correlation of 0.010 across all retained features; for n = 10,000 records, the critical correlation magnitude for statistical significance at $\alpha = 0.05$ is approximately 0.0196, meaning that while a small number of these correlations are technically distinguishable

from zero, all are practically negligible, a well-documented consequence of large sample sizes inflating the detectability of statistically significant but practically meaningless effects [45]. Because Pearson correlation captures only linear association and the models evaluated in this study (XGBoost, LightGBM, Random Forest) are explicitly capable of exploiting nonlinear and interaction effects, this analysis was complemented with a mutual information estimate, a nonparametric measure sensitive to nonlinear dependence, computed using the Kraskov–Stögbauer–Grassberger k-nearest-neighbor estimator suited to mixed continuous and discrete feature types; following the originating methodological reference, which recommends k = 2–4 for estimating dependence (with larger k trading reduced statistical variance for increased systematic bias) [46], the primary analysis used k = 3, the lower end of this recommended range, in order to favor reduced systematic bias given the relatively small number of High-risk instances available in the full dataset (1,010 of 10,000 records) for estimating this association. Mutual information between the 17 retained features and the continuous risk score ranged from 0.000 to 0.016 nats (mean 0.003 nats) at k = 3, compared with a mutual information of 0.232 nats between Crowd_Density (one of the seven excluded label-constructing variables) and the same risk score, computed as a reference point for what a genuinely informative relationship looks like in this dataset under the same measure. To verify that this conclusion does not depend on the specific neighbor-count choice, the analysis was repeated at k = 5 and k = 10: the maximum mutual information among retained features decreased further (0.008 nats at k = 5; 0.007 nats at k = 10), while the reference value for the excluded variable remained essentially unchanged (0.238 nats at k = 5; 0.241 nats at k = 10), confirming that the retained features' near-absence of nonlinear signal is stable across this range of estimator settings rather than an artifact of the k = 3 configuration. Several of the reported values are exactly 0.000; a recent systematic benchmarking of mutual information estimators, including KSG-type k-nearest-neighbor approaches, found that estimator reliability varies considerably across distributional conditions and that no single estimator performs uniformly well across all settings, particularly under sparse or weakly-informative dependence structures [47].

Consistent with this, at this sample size and neighbor parameter, the present estimator returns an exact-zero output whenever the underlying dependence falls below its practical detection floor, rather than necessarily implying a perfectly null relationship in a mathematical sense, and these values should accordingly be interpreted as "below detectable dependence at this estimator resolution" rather than as a precise quantification of zero association. The retained features therefore carry negligible nonlinear, as well as negligible linear, signal about the risk label: the highest-ranked retained feature captures roughly one-fifteenth of the mutual information carried by a single excluded label-constructing variable. This indicates that the 17 leakage-free features carry little to no signal about the risk label beyond the noise floor, under either a linear or a nonlinear association measure, consistent with the seven label-constructing variables (Crowd_Density, Temperature, Fatigue_Level, Stress_Level, Queue_Time_minutes, Emergency_Event, and Distance_Between_People_m) having been generated largely independently of the retained features in this simulated dataset. Under this condition, classification performance approaching that of the majority-class baseline is the statistically expected outcome rather than an unexplained finding, and the intrinsic-difficulty framing offered above should be read as a description of this near-independence rather than as evidence of a deeper, unmodeled relationship between observable crowd conditions and operational risk. This consideration also implies that the degree of performance degradation observed here is specific to this dataset's generation process and may not generalize to settings where label-constructing and observable variables are more strongly associated in practice.

Given this near-absence of linear and nonlinear signal in the retained features, the partial recovery of High-risk recall observed for XGBoost+SMOTE and LightGBM+SMOTE (0.14 and 0.12, respectively, versus 0.00 for the unbalanced baselines) requires its own explanation rather than being attributed to the features becoming more informative after rebalancing. The more plausible mechanism is that SMOTE alters the effective decision boundary learned by the classifier rather than increasing the information content of the feature space: by synthetically inflating the representation of the High-risk class to match the majority class during training, SMOTE shifts the model's decision threshold to favor High-risk predictions in regions of feature space that are only weakly associated with the label, at the cost of reduced precision (0.11 for XGBoost+SMOTE, Table IV) and reduced Medium-risk recall (0.67, down from 0.94 for the baseline). This interpretation is consistent with general principles of oversampling behavior in weakly separable feature spaces established in [35], [36], though the specific framing offered here is the authors' own synthesis rather than a direct claim made in either source; it favors the decision-boundary-shift account over the alternative explanation that SMOTE somehow recovers predictive signal that the correlation and mutual information analyses above show is largely absent from the underlying features. Under this interpretation, the gain in High-risk recall after SMOTE reflects a deliberate, engineered shift in operating point along an unfavorable precision-recall trade-off, not a genuine improvement in the model's ability to discriminate High-risk from Medium-risk conditions using the available observable indicators.

The dataset represents a single simulated operational context. Generalization to other crowd environments requires separate validation. Agent threshold values are expert-informed approximations not empirically calibrated against historical incident data. SMOTE generates synthetic observations from existing samples; in this study, differences between baseline and SMOTE-enhanced models remained modest, and McNemar's test confirmed significant improvement ($\chi^2 = 69.218$, $p < 0.001$). Although statistically significant, the practical effect size remained modest because the improvement was primarily observed in High-risk recall rather than overall classification accuracy. Python threading constrains true parallelism for CPU-bound workloads such as model inference, and production-scale multi-stream deployment would likely require a distributed agent framework such as JADE or Mesa. To quantify the practical impact of this constraint, per-record inference latency was measured under concurrent load using a thread pool with 1,

5, 10, and 20 simultaneous prediction streams, following a model warm-up phase to eliminate first-call overhead. As reported in Section IV-C, per-record latency increased from 1.918 ms under sequential processing to 15.538 ms at 20 concurrent streams, an approximately 8.1-fold increase at the highest concurrency level evaluated. This increase reflects the combined effect of several factors that this thread-based experiment does not separately isolate (including Python-level interpreter contention, thread-scheduling and pool-management overhead, and CPU cache contention among concurrently executing threads) rather than being attributable to the Global Interpreter Lock alone; the underlying C++ implementation of the gradient-boosting library used here may release the interpreter lock during some internal numerical computation, so the GIL constraint described above should be understood as one contributing factor among several rather than the sole explanation for the measured degradation. Regardless of the precise decomposition of these factors, the empirical result establishes that the framework's favorable latency figures reported elsewhere in this study (e.g., Table V) reflect single-stream, sequential operation and do not extend proportionally to concurrent multi-stream conditions representative of simultaneous monitoring across many pilgrims; production deployment at scale would require either a multi-process or distributed architecture, and a more detailed profiling study isolating these individual contributing factors is identified as a direction for future work.

Future work will focus on validating the framework using real-world operational datasets, calibrating agent thresholds against historical incident records, and evaluating alert precision using independently verified ground-truth events.

VI. Conclusion

This study proposed a multi-agent approach for crowd-risk prioritization in large-scale events that combines machine learning with crowd, health, and environmental agents within a unified operational workflow. The study also addressed data leakage by removing variables used to construct the risk labels, as well as post-event outcome indicators, before model training.

The results show that leakage-free prediction is difficult in this dataset. The XGBoost baseline and Cost-Sensitive XGBoost did not identify any High-risk cases (recall = 0.00). XGBoost+SMOTE detected part of this class, but performance remained limited (recall = 0.14, F1 = 0.12). McNemar's test showed a statistically significant difference between the baseline and SMOTE-enhanced model ($\chi^2 = 69.218$, $p < 0.001$), although the practical gain was mainly in minority-class sensitivity rather than overall accuracy.

The agent layer increased High-risk alert volume but, on the data examined, did so with low precision relative to the internal risk-score proxy. Of the 19 records escalated by the agent layer beyond the ML-only predictions, none (0%) corresponded to a High-risk label under this proxy, while the majority (73.7%) corresponded to Medium-risk labels. In the 500-record streaming evaluation, the Full Framework produced 93 High-risk alerts, compared with 74 from ML-only and 23 from Agents-only, representing a 25.7% increase over the ML-only configuration.

These additional alerts should therefore be interpreted as prioritization cues rather than confirmed incidents. Overall, the results suggest that ML and rule-based agents can play complementary roles in crowd-risk monitoring. However, none of the 19 agent-driven escalations beyond the ML-only predictions corresponded to a High-risk label under the internal risk-score proxy. This finding further reinforces the need for threshold calibration and validation against verified real-world incidents before operational deployment. ML provides broad probabilistic assessment across observable features, while agents add domain-specific escalation rules for conditions that the model may under-prioritize. With an average latency of 4.127 ms per record, the framework is suitable for single-stream monitoring experiments. Future work should validate the framework using real operational data and evaluate alert precision against verified ground-truth labels.

VII. Threats to Validity

A. Internal Validity

SMOTE was applied only to the training set, so synthetic samples never reached the test set; cross-validation on the SMOTE-balanced data likely makes generalization look slightly better than on the original imbalanced distribution. All models used the same fixed random seed (42) for reproducibility, and the agent thresholds are expert estimates rather than values calibrated against real incident data. As reported in Section IV- C, a simplified single-rule version of the agent logic produced the same number of High-risk alerts as the full three-agent hierarchy, showing that the added complexity of the hierarchical escalation structure has not yet been shown to add value on this dataset. Additionally, the evaluation does not quantify inter-agent communication overheads such as queue contention, synchronization latency, or message-passing delays. These overheads are expected to have minimal impact under the single-stream evaluation considered in this study but may become more relevant under higher concurrency levels.

B. External Validity

The dataset was generated under simulated conditions specific to the Hajj context, so findings may not generalize to other crowd environments (sporting events, concerts, transport hubs) without separate validation. The framework has not been tested on live IoT streams under real deployment, and Python threading limits parallelism at production scale. Furthermore, the evaluation assumes complete and reliable sensor observations and does not assess the impact of sensor noise, missing values, delayed measurements, or inconsistent data, all of which are common in practical IoT deployments and should be investigated in future work. Because the dataset was algorithmically generated rather than collected from real sensors or incidents, the statistical relationships reported here, including the near-zero feature–label correlation in Section V, the resulting recall ceiling for High-risk, and the agent-escalation precision profile in Section IV-C, reflect this synthetic generation process rather than real Hajj crowd dynamics. This is consistent with broader evaluation frameworks for synthetic tabular data, which distinguish fidelity from downstream utility and find that strong utility does not guarantee the underlying relationships match real-world data [48]. A real-world dataset could plausibly show stronger or weaker associations than

measured here, so the performance gaps, alert volumes in Table V, and precision figures in Section IV-C should be read as illustrative of this synthetic setting rather than predictive of real operational performance.

C. Construct Validity

Risk labels come from a weighted scoring function rather than verified incident records, making this a risk-prioritization tool rather than a confirmed incident predictor. Agent-alert precision against real incidents cannot be assessed without ground-truth annotations, which this dataset lacks. A check against the internal risk-score proxy (Section IV-C) found that none of the 19 agent-driven escalations beyond ML-only predictions corresponded to a true High-risk label, so High-risk alert counts should accordingly be read as proxy-based output, not confirmed real-world events.

D. Statistical Validity

Evaluation used a single dataset with one fixed train/test split. McNemar's test confirms the XGBoost+SMOTE improvement is statistically significant ($\chi^2 = 69.218$, $p < 0.001$), but a single split may not generalize across seeds or partitions. Bootstrap resampling (1,000 resamples) gave 95% confidence intervals of 47.35% [45.20%, 49.55%] for accuracy and 13.76% [9.23%, 18.60%] for High-risk recall, confirming the point estimates are stable under resampling, though this does not address variation across different training seeds. To address that directly, the full pipeline was repeated across 10 independent seeds (0, 1, 2, 3, 4, 7, 13, 21, 42, 99): accuracy averaged 47.08% (SD = 0.55%), F1 averaged 45.97% (SD = 0.46%), and High-risk recall averaged 0.134 (SD = 0.017). The seed = 42 result reported throughout this study falls within one standard deviation of this mean on every metric, confirming it is representative rather than a favorably chosen outlier. The mutual information analysis in Section V was similarly stable across k = 3, 5, and 10, confirming the near-absence of nonlinear association is not an artifact of a single parameter choice.

Declaration on the Use of AI-Assisted Tools

The author used AI-assisted tools to support language refinement and manuscript drafting during the preparation of this work. All scientific ideas, research design, experimental implementation, data analysis, results interpretation, and conclusions were carried out by the human author. The author takes full responsibility for the accuracy, originality, and integrity of the manuscript.

References

[1] T. Alafif et al., “Toward an integrated intelligent framework for crowd control and management (IICCM),” IEEE Access, vol. 13, pp. 58559-58575, 2025, doi: 10.1109/ACCESS.2025.3555154.

[2] W. Halboob, H. Altaheri, A. Derhab, and J. Almuhtadi, “Crowd management intelligence framework: Umrah use case,” IEEE Access, vol. 12, pp. 6752-6767, 2024, doi: 10.1109/ACCESS.2024.3350188.

[3] M. Khel et al., “Realtime crowd monitoring - Estimating count, speed and direction of people using hybridized YOLOv4,” IEEE Access, vol. 11, pp. 56368-56379, 2023, doi: 10.1109/ACCESS.2023.3272481.

[4] R. Nasir et al., “An enhanced framework for real-time dense crowd abnormal behavior detection using YOLOv8,” Artif. Intell. Rev., vol. 58, 2025, doi: 10.1007/s10462-025-11206-w.

[5] M. Qaraqe et al., “PublicVision: A secure smart surveillance system for crowd behavior recognition,” IEEE Access, vol. 12, pp. 26474-26491, 2024, doi: 10.1109/ACCESS.2024.3366693.

[6] S. Ramesh, B. Rohith, K. Chetanand, and R. Manikandan, “Resilient Multi-Modal IoT Crowd Monitoring System with Self-Healing Sensor Architecture and Intelligent Alert Prioritization,” in Proc. 2nd Int. Conf. Multi-Agent Syst. Collaborative Intell. (ICMSCI), Erode, India, 2026, pp. 1249-1254, doi: 10.1109/ICMSCI67830.2026.11469199.

[7] H. Zhang, R. Zhang, and J. Sun, “Developing real-time IoT-based public safety alert and emergency response systems,” Sci. Rep., vol. 15, 2025, doi: 10.1038/s41598-025-13465-7.

[8] C. Dumitrescu et al., “Crowd panic behavior simulation using multi-agent modeling,” Electronics, vol. 13, no. 18, p. 3622, 2024, doi: 10.3390/electronics13183622.

[9] M. Keykhaei et al., “A situation-aware emergency evacuation model using multi-agent-based simulation,” Geo-spatial Inf. Sci., vol. 27, pp. 1800-1823, 2023, doi: 10.1080/10095020.2023.2270017.

[10] D. Costa et al., “A survey of emergencies management systems in smart cities,” IEEE Access, 2022, doi: 10.1109/ACCESS.2022.3180033.

[11] N. Nezamoddini and A. Gholami, "A survey of adaptive multi-agent networks and their applications in smart cities," Smart Cities, 2022, doi: 10.3390/smartcities5010019.

[12] A. Kalyuzhnaya et al., “LLM agents for smart city management: Enhancing decision support through multi-agent AI systems,” Smart Cities, 2025, doi: 10.3390/smartcities8010019.

[13] C. Berceanu, I. Banu, B. Husebo, and M. Patrascu, “Predictive agent-based crowd model design using decentralized control systems,” IEEE Syst. J., vol. 17, pp. 1383-1394, 2023, doi: 10.1109/JSYST.2022.3188339.

[14] S. Veesam et al., “Multi-camera spatiotemporal deep learning framework for real-time abnormal behavior detection in dense urban environments,” Sci. Rep., vol. 15, 2025, doi: 10.1038/s41598-025-12388-7.

[15] D. Mabrouk, M. Abdel-Fattah, and A. Taha, “Robust crowd anomaly detection via hybrid ensemble learning for real-world surveillance,” Sci. Rep., vol. 15, 2025, doi: 10.1038/s41598-025-27408-9.

[16] A. Alsabei, T. Alsubait, and H. Alhakami, “Enhancing crowd safety at Hajj: Real-time detection of abnormal behavior using YOLOv9,” IEEE Access, vol. 13, pp. 37748-37761, 2025, doi: 10.1109/ACCESS.2025.3545256.

[17] Y. Wu, L. Qiu, J. Wang, and S. Feng, “The use of convolutional neural networks for abnormal behavior recognition in crowd scenes,” Inf. Process. Manage., vol. 62, p. 103880, 2025, doi: 10.1016/j.ipm.2024.103880.

[18] A. Leelamony et al., “DeepTrackNet: Advanced crowd dynamics estimation with YOLO and Deep SORT integration,” in Proc. CSASE, 2025, pp. 80-86, doi: 10.1109/CSASE63707.2025.11053996.

[19] S. Hu et al., “Integrating embedded cyber-physical systems in smart energy for AI-enhanced real-time crowd monitoring and threat detection,” IEEE Trans. Consum. Electron., vol. 71, pp. 8363-8373, 2025, doi: 10.1109/TCE.2025.3576383.

[20] F. Zeng, C. Pang, and H. Tang, “Sensors on Internet of Things systems for the sustainable development of smart cities,” Sensors, vol. 24, 2024, doi: 10.3390/s24072074.

[21] T. Wiangwiset et al., “Design and implementation of a real-time crowd monitoring system based on public Wi-Fi infrastructure,” Smart Cities, 2023, doi: 10.3390/smartcities6020048.

[22] A. Fascista, “Toward integrated large-scale environmental monitoring using WSN/UAV/crowdsensing,” Sensors, vol. 22, 2022, doi: 10.3390/s22051824.

[23] T. Narayana et al., “Advances in real time smart monitoring of environmental parameters using IoT and sensors,” Heliyon, vol. 10, 2024, doi: 10.1016/j.heliyon.2024.e28195.

[24] B. Prabhu Shankar et al., “Smart crowd: AI-driven crowd density monitoring and management in Indian public hotspots,” in Proc. ICSCDS, 2025, pp. 651-657, doi: 10.1109/ICSCDS65426.2025.11167575.

[25] S. Ayesha, A. Aslam, M. H. Zaheer, and M. B. Khan, “CIRS: A multi-agent machine learning framework for real-time accident detection and

emergency response," Sensors, vol. 25, no. 18, p. 5845, 2025, doi: 10.3390/s25185845.

[26] D. Maldonado et al., "Multi-agent systems: A survey about its components, framework and workflow," IEEE Access, vol. 12, pp. 80950-80975, 2024, doi: 10.1109/ACCESS.2024.3409051.

[27] D. Sako, C. Igiri, E. Bennett, and F. Deedam, "ESARS: A situation-aware multi-agent system for real-time emergency response management," Eur. J. Inf. Technol. Comput. Sci., 2024, doi: 10.24018/compute.2024.4.1.83.

[28] A. Takahashi, K. Yasufuku, and M. Hegazy, "Applicability of multi-agent simulation for crowd control after a large-scale event," Japan Archit. Rev., 2025, doi: 10.1002/2475-8876.70063.

[29] T. Alshammari and M. Bennis, "Logic-driven semantic communication for resilient multi-agent systems," IEEE Open J. Commun. Soc., vol. 7, pp. 630-654, 2026, doi: 10.1109/OJCOMS.2026.3653924.

[30] M. Alshamrani, S. Alhazmi, T. Alafif, T. M. Qadah, M. Aljabri, W. Al-Eidarous, and A. Alhawsawi, "From data to insights: A comprehensive analysis of pilgrims stress and fatigue during Hajj using wearable remote sensing systems," J. King Saud Univ. Comput. Inf. Sci., vol. 37, p. 94, 2025, doi: 10.1007/s44443-025-00095-2.

[31] A. Lama, M. Di Bernardo, and S. Klapp, "Nonreciprocal field theory for decision-making in multi-agent control systems," Nat. Commun., vol. 16, 2025, doi: 10.1038/s41467-025-63071-4.

[32] L. Hughes et al., "AI agents and agentic systems: A multi-expert analysis," J. Comput. Inf. Syst., vol. 65, pp. 489-517, 2025, doi: 10.1080/08874417.2025.2483832.

[33] A. Hamid, A. Monadjemi, and B. Shoushtarian, "ABDviaMSIFAT: Abnormal crowd behavior detection utilizing a multi-source information fusion technique," IEEE Access, vol. 13, pp. 74996-75015, 2025, doi: 10.1109/ACCESS.2024.3436007.

[34] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," ACM Trans. Knowl. Discov. Data, vol. 6, no. 4, pp. 1-21, 2012, doi: 10.1145/2382577.2382579.

[35] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263-1284, 2009, doi: 10.1109/TKDE.2008.239.

[36] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321-357, 2002, doi: 10.1613/jair.953.

[37] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5-32, 2001, doi: 10.1023/A:1010933404324.

[38] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD), San Francisco, CA, USA, 2016, pp. 785-794, doi: 10.1145/2939672.2939785.

[39] I. Araf, A. Idri, and I. Chairi, "Cost-sensitive learning for imbalanced medical data: A review," Artificial Intelligence Review, vol. 57, no. 4, 2024, doi: 10.1007/s10462-023-10652-8.

[40] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," Neural Comput., vol. 10, no. 7, pp. 1895-1923, 1998, doi: 10.1162/089976698300017197.

[41] M. Wooldridge, An Introduction to MultiAgent Systems, 2nd ed. Chichester, UK: Wiley, 2009.

[42] N. Bahbouh, S. Sendra, and A. Sen, "Designing a comprehensive framework for health management in crowded events," Int. J. Inf. Technol., vol. 17, pp. 1641-1651, 2024, doi: 10.1007/s41870-024-02051-1.

[43] I. Gani, I. Litchfield, D. Shukla, G. Delanerolle, N. Cockburn, and A. Pathmanathan, "Understanding "alert fatigue" in primary care: Qualitative systematic review of general practitioners' attitudes and experiences of clinical alerts, prompts, and reminders," J. Med. Internet Res., vol. 27, p. e62763, 2025, doi: 10.2196/62763.

[44] A. Vickers, A. Hollingsworth, A. Bozzo, A. Chatterjee, and S. Chatterjee, "Hypothesis: Net benefit as an objective function during development of machine learning algorithms for medical applications," Int. J. Med. Inform., vol. 197, p. 105844, 2025, doi: 10.1016/j.ijmedinf.2025.105844.

[45] M. Lin, H. C. Lucas, and G. Shmueli, "Research commentary—Too big to fail: Large samples and the p-value problem," Inf. Syst. Res., vol. 24, no. 4, pp. 906-917, 2013, doi: 10.1287/isre.2013.0480.

[46] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," Phys. Rev. E, vol. 69, no. 6, p. 066138, 2004, doi: 10.1103/PhysRevE.69.066138.

[47] P. Czyż, F. Grabowski, J. E. Vogt, N. Beerenwinkel, and A. Marx, "Beyond normal: On the evaluation of mutual information estimators," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 36, 2023, pp. 16957-16990.

[48] M. Hernandez, P. A. Osorio-Marulanda, M. Catalina, L. Loinaz, G. Epelde, and N. Aginako, "Comprehensive evaluation framework for synthetic tabular data in health: Fidelity, utility and privacy analysis of generative models with and without privacy guarantees," Front. Digit. Health, vol. 7, p. 1576290, 2025, doi: 10.3389/fdgth.2025.1576290.

Model-Augmented Masked Autoencoder with Domain-Adaptive Denoising for Robust Brain Tumor MRI Analysis

Indrakumar K[1], Ravikumar M[2], Mohammed A.S Al-mohamadi[3],
Khalid N. R. Alharbi[4], Asma A. Alhashmi[5]*, Shouki A. Ebad[6], Abdulbasit A. Darem[7]*
Department of MCA and Computer Science, Jnanasahyadri, Kuvempu University, Shivamogga 577 451, India[1, 2, 3]
Department of Computer Science-College of Science, Northern Border University, Arar 73213, Saudi Arabia[4, 5]
Centre for Scientific Research and Entrepreneurship, Northern Border University, Arar 73213, Saudi Arabia[6, 7]

***Abstract*—This work presents a Model-Augmented Masked Autoencoder (MAE) framework, augmented with domain-adaptive denoising and multi-level feature fusion, for self-supervised representation learning in MRI-based brain tumor analysis. The proposed approach enhances feature robustness under limited annotation by learning structure-preserving representations through masked reconstruction while explicitly addressing scanner-induced variability. Unlike conventional MAE designs, the framework integrates a Domain-Adaptive Denoising (DAD) module and latent–reconstruction feature fusion, enabling improved robustness across heterogeneous MRI datasets and reducing feature variance arising from scanner and acquisition differences. Comprehensive ablation studies across multiple datasets demonstrate that MAE pretraining yields consistent performance improvements of approximately 1.0–1.4% in classification accuracy, with overall improvements ranging from approximately 2.7% to 3.7% across datasets when integrated into the complete framework. Qualitative and quantitative analysis further confirm effective suppression of structural noise while preserving diagnostically relevant features. Overall, the proposed framework improves representation stability, enhances cross-domain robustness, and provides a reliable foundation for self-supervised medical image analysis, supporting the development of generalizable and clinically applicable diagnostic AI systems.**

Keywords*—Self-supervised learning; Masked Autoencoder; MRI; domain adaptation; feature representation; medical imaging

I. Introduction

Magnetic Resonance Imaging (MRI) plays a central role in the diagnosis, grading, and clinical management of brain tumors. Accurate tumor classification is essential for treatment planning, yet remains challenging due to substantial variations in tumor morphology, heterogeneous imaging protocols across scanners, and the presence of noise, artifacts, and low-contrast boundaries. Deep learning models have achieved impressive performance in medical image classification; however, their success is strongly dependent on the availability of large, high-quality annotated datasets. In neuro-oncology, such annotations require expert radiological expertise, are time-consuming to produce, and often vary across institutions. These challenges motivate the need for representation learning strategies that reduce dependence on labeled data while improving robustness to domain variability.

*Corresponding author

Self-supervised learning (SSL) has emerged as a powerful paradigm to address annotation scarcity by leveraging unlabeled images to learn transferable, domain-invariant representations [1,2,3,4,29]. Contrastive methods and masked prediction frameworks such as Masked Autoencoders (MAE) [16] have shown promise in natural and medical imaging. By predicting missing content or enforcing consistency between augmented views, SSL extracts semantic cues without explicit annotations. However, despite their potential, existing SSL approaches exhibit three major limitations when applied to brain MRI analysis. First, contrastive-based SSL approaches such as those analyzed by Wolf et al. [3] are sensitive to augmentation strategies and may fail to capture fine anatomical details critical for tumor demarcation. Second, conventional MAE-based frameworks [3,4] reconstruct masked regions but generally do not explicitly address scanner-specific noise, bias-field inhomogeneities, or contrast variations commonly encountered in clinical MRI. Third, although Zhang et al. [18], Liu et al. [19], Zhang et al. [20], Bakkouri and Afdel [21], and Peng et al. [22] demonstrated the effectiveness of multi-level feature fusion, the integration of latent self-supervised representations with radiomics, shape descriptors, and deep CNN features remains insufficiently explored.

To address these challenges, this study introduces the Model-Augmented Masked Autoencoder (MAE), a self-supervised pretraining framework designed to improve robustness and generalization in MRI-based brain tumor classification. The proposed approach builds upon masked image modeling by reconstructing randomly masked patches, enabling the model to learn structure-preserving representations that emphasize anatomical continuity and tumor-related spatial patterns [16]. A key innovation of the framework is the integration of a Domain-Adaptive Denoising (DAD) module, which refines latent token embeddings through channel normalization, patch-level smoothing, and gated fusion mechanisms. This enhancement makes the learned representations more resilient to cross-scanner inconsistencies and low-quality acquisitions, aligning with recent domain generalization frameworks discussed by Zhou et al. [27] and Wang et al. [28]. Additionally, the model augments downstream learning by combining MAE latent embeddings with reconstructed images, ensuring that both high-level contextual

signals and low-level structural details contribute to improved tumor characterization.

The enriched representations produced by the MAE pipeline are further integrated into a multi-level feature extraction system, incorporating radiomics descriptors, CNN-based semantic features, and geometric shape attributes. This fusion captures a wide spectrum of tumor information from statistical intensity patterns to morphological cues, yielding a more holistic and clinically aligned tumor signature [18,19,20,21,22]. These features are processed by an attention-guided fusion block and a Pyramid Quantum-Inspired Dilated Convolutional Neural Network (PyQDCNN), which models long-range spatial dependencies and multi-scale tumor context through dilation-driven and quantum-inspired (implemented as learnable bilinear feature interactions) entanglement mechanisms [23,24,25]. It is important to note that this mechanism is inspired by quantum interactions but is implemented as a learnable bilinear feature mixing operation within a conventional deep learning framework. The final hierarchical classification strategy mirrors clinical diagnostic workflows, first distinguishing normal from abnormal MRI scans, and then classifying tumors into clinically relevant subtypes.

The main contributions of this study are as follows:

- We propose a Model-Augmented Masked Autoencoder (MAE) with a Domain-Adaptive Denoising (DAD) module, enabling the learning of structure-preserving and scanner-robust latent representations that enhance MRI feature stability across heterogeneous MRI datasets.
- We develop a multi-level tumor representation pipeline that integrates radiomics, CNN features, shape descriptors, and MAE latent embeddings, producing a comprehensive and clinically aligned feature space for robust tumor characterization.
- We introduce a hybrid attention–PyQDCNN architecture that combines modality-aware feature weighting with entanglement-inspired bilinear feature interactions and dilated convolutional modeling, enabling effective capture of long-range dependencies and multi-scale tumor characteristics.
- We demonstrate consistent performance gains across four heterogeneous brain MRI datasets, achieving improvements of 1.0–1.4% due to MAE pretraining and overall accuracy improvements ranging from approximately 2.7% to 3.7% compared to baseline configurations, confirming the effectiveness and generalizability of the proposed framework.

The remainder of the study is organized as follows: Section II reviews related work. Section III presents the proposed Model-Augmented MAE framework. Section IV outlines the experimental setup. Section V reports and discusses results. Section VI concludes the study.

II. Related Work

Self-supervised learning (SSL) has become an essential strategy in medical imaging, particularly when annotated data are scarce. Rani et al. [1] presented a comprehensive review of SSL methods for medical image analysis and highlighted their effectiveness across multiple imaging modalities. Zeng et al. [2] summarized SSL frameworks and their applications in healthcare imaging. Wolf et al. [3] investigated contrastive and masked autoencoder pretraining strategies for small medical imaging datasets, demonstrating improved representation learning under limited annotations. VanBerlo et al. [4] analyzed the impact of self-supervised pretraining across X-ray, CT, MRI, and ultrasound modalities, while Uelwer et al. [29] provided a recent survey of self-supervised visual representation learning methods. Transformer-based SSL approaches, particularly masked autoencoders (MAEs), have gained prominence for their ability to model long-range dependencies and structural features in clinical images. The original MAE framework [16] demonstrates the effectiveness of reconstruction-based pretraining, while subsequent studies show that contrastive and masked pretraining significantly improve performance on small or heterogeneous datasets [3]. Recent applications further highlight the potential of transformer-driven SSL in disease detection and medical imaging tasks [5]. Advancements in MAE-based pretraining include improved architectural variations such as snapshot-ensemble MAEs for efficient training [6] and adaptive masking strategies designed to reinforce structural consistency during representation learning [7]. Emerging research directions also explore auxiliary modalities, such as large language model–augmented learning, which has been shown to enhance auto-delineation performance in radiotherapy planning [8]. In addition, domain generalization has received significant attention, with recent surveys emphasizing the importance of robust learning across heterogeneous datasets and imaging conditions [27,28]. Federated learning frameworks further demonstrate potential for improving cross-institutional generalization in brain tumor segmentation tasks [9].

For downstream tumor classification and segmentation, Vision Transformers (ViTs) and hybrid CNN–Transformer models have shown strong performance. Several studies have explored rotation-invariant ViTs for improved spatial consistency [11], ensemble-based ViT classification strategies [12], hybrid ViT–CNN architectures with explainability [13], and transfer-learning–based transformer models for tumor classification [14]. Transformer-based methods have also been successfully applied to related diagnostic tasks such as brain stroke classification [15] and glioma analysis [17], demonstrating their adaptability across different MRI-based applications.

Multi-level and multi-scale feature fusion approaches have further enhanced tumor detection and segmentation performance. Zhang et al. [18] proposed a multi-level feature exploration and fusion network for glioma analysis using MRI data. Liu et al. [19] introduced MCF-Net, which combines multi-level contextual information for medical image segmentation. Zhang et al. [20] developed a multi-scale deformable feature fusion framework for tumor segmentation. Bakkouri and Afdel [21] proposed an attention-based multi-level feature fusion strategy for lesion segmentation, while Peng et al. [22] introduced a graph-neural-network-based multi-level fusion approach for multimodal medical imaging. These studies

demonstrate that combining complementary feature representations improves tumor characterization and prediction accuracy. These approaches effectively capture complementary information across multiple feature levels, enabling more robust modeling of tumor heterogeneity. Additionally, quantum-inspired convolutional and optimization methods have been explored to model complex spatial dependencies and improve segmentation performance in biomedical imaging [23,24,25].

Despite these advances, existing SSL frameworks still face challenges related to scanner variability, noise, and domain shift, which can degrade representation stability across heterogeneous MRI sources. Moreover, most MAE-based models focus primarily on reconstruction objectives and do not explicitly incorporate domain-adaptive denoising or latent feature augmentation mechanisms. Addressing these limitations, the proposed Model-Augmented MAE framework integrates domain-adaptive denoising and multi-level feature fusion to improve reconstruction fidelity, enhance representation robustness, and achieve better generalization across diverse MRI datasets.

III. Proposed Framework

Brain tumor MRI classification demands robustness against variations in scanners, acquisition protocols, tumor heterogeneity, and noise. A single model or feature type is insufficient to ensure generalization. Therefore, the proposed design integrates complementary modules, each addressing a specific challenge:

- Masked Autoencoder (MAE) [16]: learns structure-preserving self-supervised features even with limited annotated datasets.
- Domain-Adaptive Denoising (DAD): stabilizes latent features across multi-center scans.
- Multi-level feature extraction combines radiomics (handcrafted biomarkers), CNN (deep hierarchical semantics), and shape descriptors (morphology).
- Attention-guided fusion removes redundancy across modalities and retains clinically meaningful patterns.
- PyQDCNN: models long-range tumor–background relationships through dilation- and entanglement-inspired kernels.
- Hierarchical classification: reduces inter-class interference for subtle tumor subtypes.

Among these components, the Model-Augmented MAE and Domain-Adaptive Denoising (DAD) constitute the core contributions of the framework, while the remaining modules primarily enhance feature representation and classification performance. The effectiveness of these modules is quantitatively validated through the dataset-level ablation studies presented in Table I to Table IV and the component-level ablation analysis in Table V, which demonstrate their individual and cumulative contributions to overall framework performance. Together, these elements create a multi-resolution, multi-representation, and domain-robust tumor analysis framework aligned with clinical practice.

The MRI input first undergoes a quality-control stage, where low-quality or corrupted scans are identified and excluded from further processing. The accepted volumes are geometrically normalized by reorienting to RAS space, resampling to isotropic resolution, and cropping or padding the field of view to maintain structural consistency. Subsequent bias-field correction (N4) and brain extraction [4] (skull stripping) remove intensity inhomogeneities and non-brain tissues, respectively. Intensity standardization is then performed through z-score normalization and histogram matching to harmonize the contrast distribution across subjects and scanners [4].

Following conventional pre-processing, the data are passed through a Model-Augmented Masked Autoencoder (MAE) framework [16] that leverages self-supervised representation learning. In this module, image patches are randomly masked and encoded via a Vision Transformer–based encoder (ViT-E) [11,12,13,14] to produce latent token embeddings z. The decoder reconstructs the masked patches, and reconstruction loss is computed as mean-squared error [3] on the masked regions, promoting structure-aware feature learning. Mean squared error (MSE) is adopted due to its computational efficiency and stability in preserving global structural consistency during reconstruction. Alternative perceptual or SSIM-based losses were not employed to avoid additional complexity and instability during self-supervised training. The MAE outputs both a reconstructed (denoised) image and a latent feature bank (z) used in subsequent optimization and augmentation stages. Z-score normalization was applied to reduce inter-subject intensity variability, while histogram matching ensured cross-scanner contrast harmonization. This dual-stage standardization was found to produce more stable latent representations in the MAE encoder compared to other methods such as min-max normalization or percentile clipping, especially when handling multi-center MRI datasets. This observation is further supported by the ablation results presented in Section IV, where improved convergence stability and feature consistency are observed across multi-center datasets. Depending on the downstream configuration, three data routes are available:

Route A: pre-processed image only,

Route B: MAE-enhanced reconstructed image, and

Route C: MAE latent representation only, the D-dimensional token embeddings (z) from the ViT encoder (optionally aggregated), which encapsulate self-supervised, structure-aware MRI priors. This route isolates the contribution of the MAE's learned representation, independent of raw or reconstructed image inputs.

While the augmented image set (pre-processed + reconstructed) is used to extract radiomics, CNN, and shape features (as shown in Fig. 3), Route C refers specifically to the direct use of the MAE latent vector z as an input stream typically fed into the attention-guided fusion module alongside other extracted features.

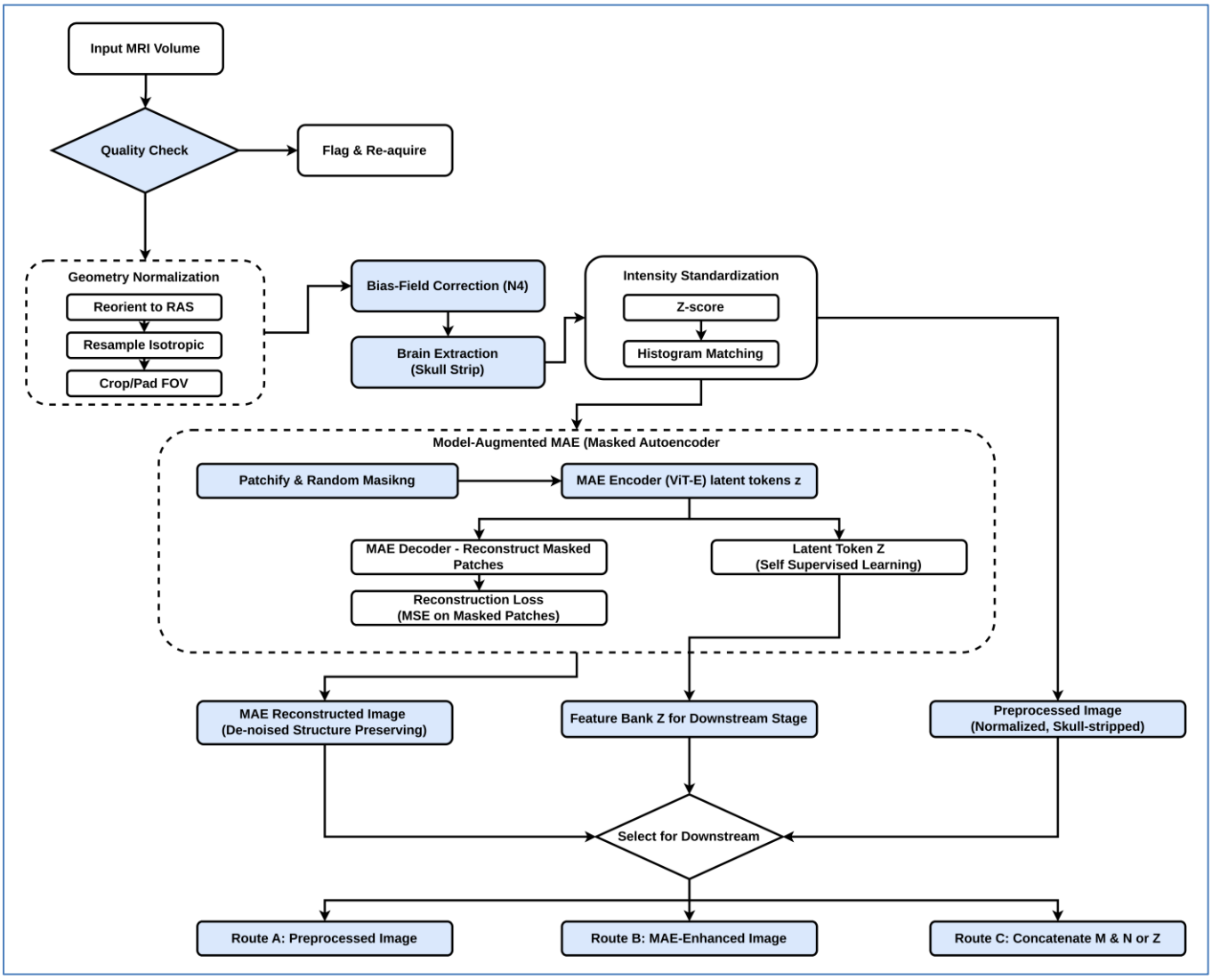


Fig. 1. Input and pre-processing pipeline using Model-Augmented Masked Autoencoder (MAE).

Following the generation of pre-processed images, reconstructed outputs, and latent tokens in Fig. 1, the internal functioning of the proposed MAE is detailed in Fig. 2, which expands the encoder–decoder architecture responsible for learning structure-aware, domain-robust MRI representations. Fig. 2 illustrates the internal architecture of the proposed Vision Transformer–based Masked Autoencoder (ViT-MAE), which is composed of five sequential stages: patch embedding, masking, transformer encoding, domain-adaptive denoising, and transformer decoding with reconstruction. The MRI slice is first divided into fixed-size non-overlapping patches that are flattened and linearly projected into D-dimensional token embeddings, followed by the addition of learnable two-dimensional positional encodings. A random or stratified masking strategy is then applied, in which a large portion of patch tokens is removed to create a subset of visible tokens used for self-supervised learning. These visible tokens are processed by a multi-layer Vision Transformer encoder consisting of LayerNorm, multi-head self-attention, and MLP feed-forward blocks, producing both multi-scale snapshot features and a final latent representation. To enhance robustness across scanners and acquisition conditions, the encoder tokens pass through a Domain-Adaptive Denoising (DAD) module that performs channel-wise normalization, patch smoothing, and gated fusion, generating denoised latent features. The design of the DAD module is motivated by the need to reduce scanner-induced feature variance while preserving anatomically relevant structures, thereby improving cross-domain robustness without introducing significant computational overhead [27], [28]. These refined tokens are projected into the decoder dimension, augmented with decoder positional embeddings, and combined with learnable mask tokens inserted at the original masked positions. The full token sequence is then passed into the Vision Transformer decoder, which applies self-attention and cross-attention to reconstruct the missing patches. A final linear output head predicts pixel values for each masked patch, and the reconstructed patches are assembled back into a complete MRI slice. The reconstruction error, computed only on the masked regions, drives the model to learn semantically meaningful and structure-preserving representations that serve as the foundation for downstream analysis.

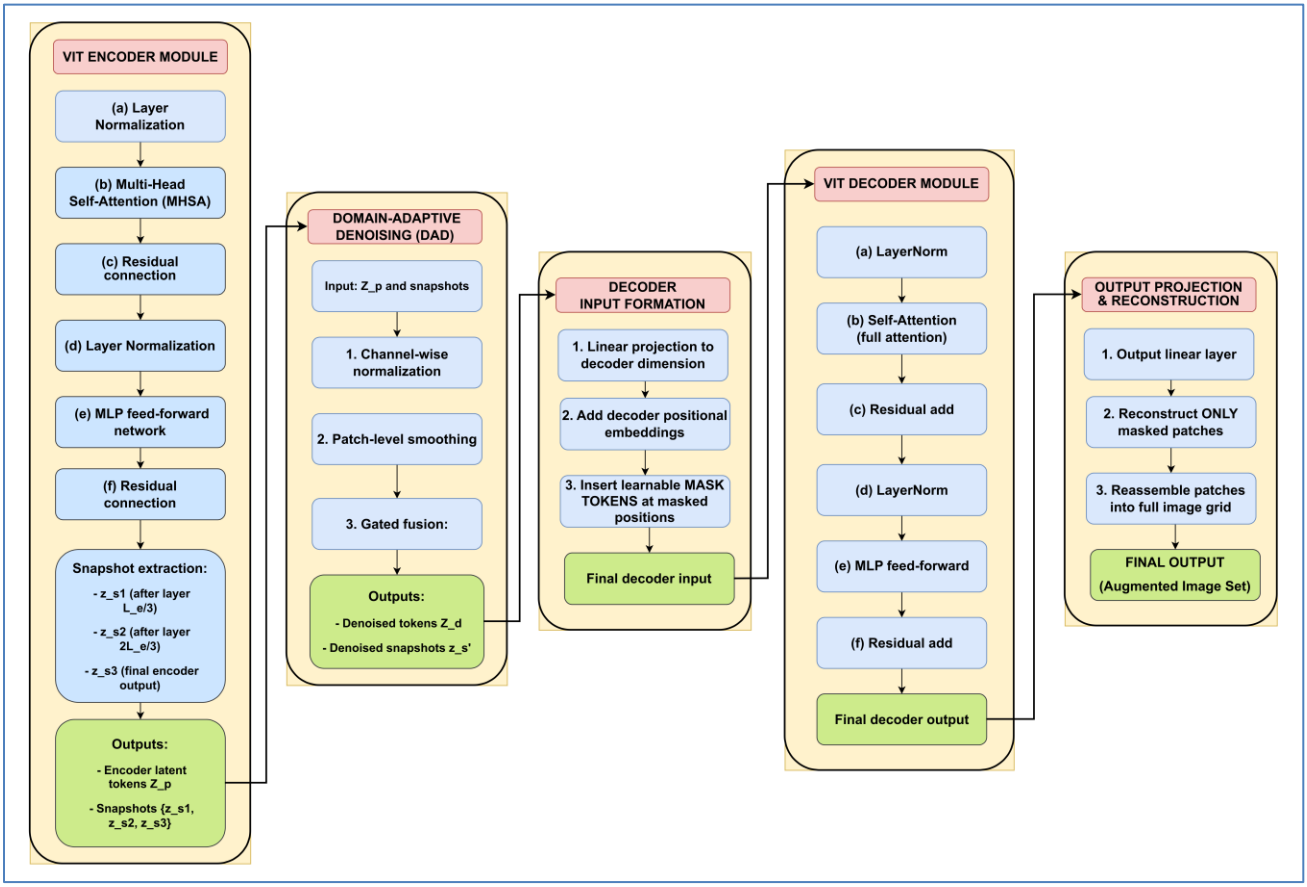


Fig. 2. ViT-based encoder–decoder architecture.

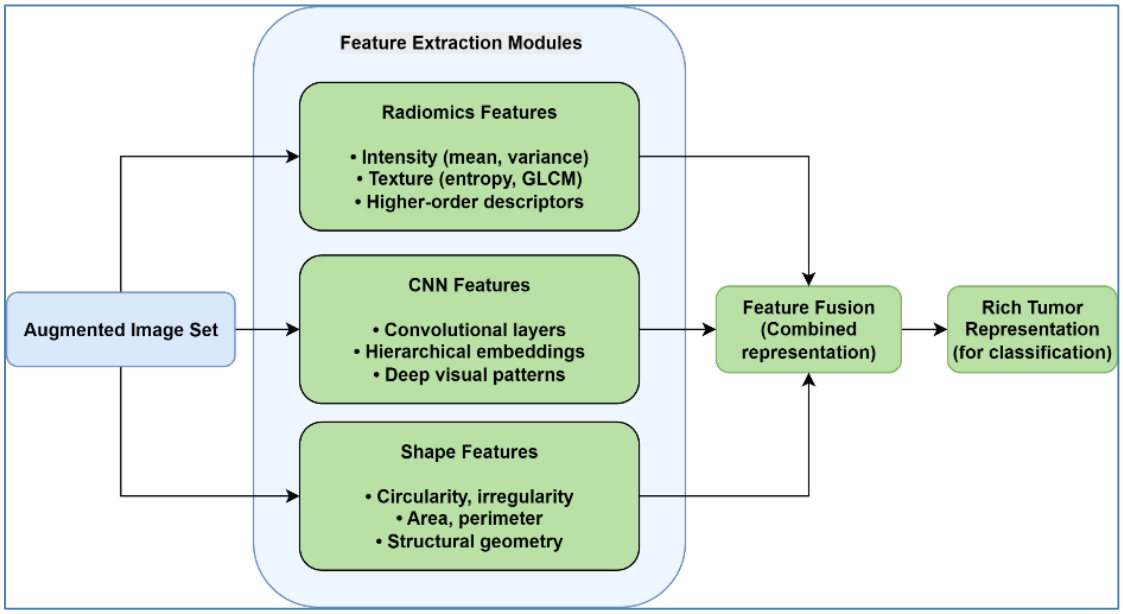


Fig. 3. Multi-level feature extraction and fusion for rich tumor representation.

Following processing through the ViT-based encoder–decoder architecture, the MAE produces multiple complementary image representations including the pre-processed MRI, the MAE-reconstructed image, and the latent feature maps derived from the transformer encoder and DAD module. These representations form the augmented image set that is forwarded into the multi-level feature extraction stage [18,19,20,21,22] illustrated in Fig. 3. By providing several

structurally distinct yet semantically aligned inputs, the augmented set enriches the available feature space, ensuring that both native anatomical details and MAE-enhanced latent cues are preserved. Within this stage, each representation is leveraged to compute three categories of features: radiomics descriptors that capture statistical and textural biomarkers, CNN-derived deep features that encode hierarchical visual patterns, and geometric shape descriptors that quantify morphological characteristics of the tumor. Notably, the MAE-derived latent tokens z supply context-aware structural priors that improve the reliability and discriminative strength of downstream feature extraction. Their inclusion helps stabilize tumor characterization across scanners and patient cohorts, resulting in a more robust and comprehensive tumor feature representation after feature fusion. A patch size of p×p was selected after empirical evaluation across p∈{8,16,32}. We observed that smaller patches improved reconstruction fidelity but significantly increased token count and memory consumption, whereas larger patches lost fine-grained tumor boundaries. The chosen patch dimension provides an optimal trade-off between spatial detail preservation and computational efficiency. The description of the masking strategy and reconstruction mechanism has been consolidated into a single unified paragraph to eliminate redundancy and improve methodological clarity. The revised explanation succinctly integrates patch embedding, token masking, transformer encoding, and reconstruction into a coherent flow.

Following the multi-level feature extraction stage, the resulting radiomics, CNN-derived, and shape descriptors are integrated into the core network architecture illustrated in Fig. 4.

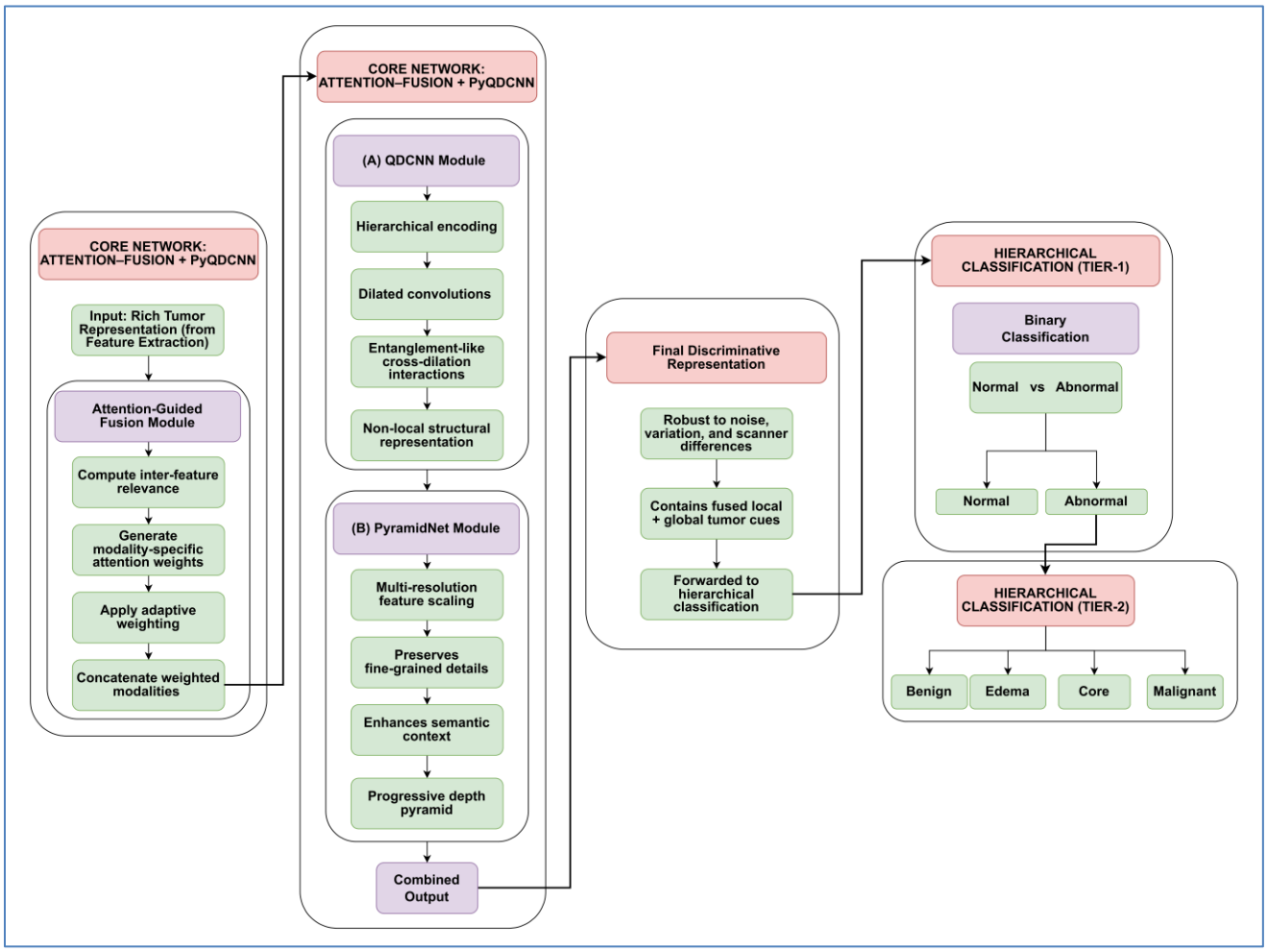


Fig. 4. Proposed framework integrating attention-guided feature fusion and the Pyramid Quantum-Inspired Dilated Convolutional Neural Network (PyQDCNN).

Fig. 4 presents the final stage of the proposed framework, integrating attention-guided feature fusion, the Pyramid Quantum-Inspired Dilated Convolutional Neural Network (PyQDCNN), and a hierarchical tumor classification strategy to generate clinically meaningful brain tumor MRI predictions. The concatenated radiomics, CNN, and shape descriptors from the previous stage are first processed through an attention-guided fusion module, which assigns modality-specific weights based on inter-feature correlations, thereby emphasizing informative tumor cues while suppressing redundant variations. The weighted multi-source features are then passed to the PyQDCNN backbone, which combines dilated convolutional processing with an entanglement-inspired bilinear feature interaction mechanism designed to model cross-channel and cross-scale dependencies [23,24,25]. The QDCNN component models long-range spatial dependencies through dilation-driven

feature entanglement, while the PyramidNet component enhances multi-resolution detail retention across deeper semantic layers. The resulting fused representation is both structurally rich and noise-resilient, forming the input to a two-tier hierarchical classification system. In the first tier, the model performs a binary decision to distinguish normal from abnormal MRI scans. Samples identified as abnormal are forwarded to the second tier, where a four-class classification predicts tumor subtypes: benign, edema, core, and malignant. This hierarchical formulation reduces inter-class ambiguity, improves discrimination among visually similar tumor types, and aligns with clinical diagnostic workflows. The final output provides robust and precise multi-class tumor identification, demonstrating the effectiveness of the combined attention–PyQDCNN–hierarchical pipeline for comprehensive brain tumor MRI analysis.

A. Mathematical Formulation of the Proposed Framework

Notation: Let:

- I denote an input MRI volume (or slice) and $I(x)$ its intensity at voxel x.
- I_{norm} geometric-normalized image, I_{corr} bias-field corrected image, I_{brain} skull-stripped image, I_{std} intensity-standardized image.
- p patch side length (so patches are $p \times p$).
- N total number of non-overlapping patches in one slice: $N = \frac{H \cdot W}{p^2}$.
- x_i the i-th patch (flattened), $i \in \{1, \dots, N\}$.
- $W_e \in \mathbb{R}^{D \times p^2}$ encoder patch projection; D the token dimension.
- $t_i \in \mathbb{R}^D$ token embedding for patch x_i.
- $\rho \in (0,1)$ masking ratio (fraction of patches masked).
- $\mathcal{M} \in \{0,1\}^N$ mask indicator vector; $\mathcal{M}(i) = 1$ if patch i is masked.
- $T_v = \{t_i \mid \mathcal{M}(i) = 0\}$ visible tokens.
- $\text{ViT}_{enc}(\cdot)$ encoder mapping producing latent tokens z.
- $z = \{z_i\}_{i=1}^{N_v}$, $z_i \in \mathbb{R}^D$, $N_v = |T_v|$.
- DAD denotes the Domain-Adaptive Denoising module; output tokens z^{dad}.
- $\text{ViT}_{dec}(\cdot)$ decoder mapping producing reconstructed patches $\hat{x}_i$.
- Radiomics feature vector R, CNN deep feature vector C, shape descriptor vector S.
- Fused feature vector F, attended fused feature F_{att}.
- PyQDCNN mapping Φ_{PyQ} producing final representation F_{final}.
- Tier-1 binary output $\hat{y}_1 \in [0,1]$; Tier-2 multi-class vector $\hat{y}_2 \in \Delta^{K-1}$ (with $K = 4$ tumor classes).
- Ground-truth labels y_1 (binary) and y_2 (one-hot multi-class).

1) Preprocessing (deterministic ops)

Geometry reorientation and isotropic resampling [see Eq. (1)]:

$$I_{norm} = \mathcal{R}(\mathcal{O}(I), R_{\text{iso}}) \quad (1)$$

where, $\mathrm{O}(\cdot)$ denotes orientation alignment, R_{iso} represents isotropic resampling, and $\mathrm{R}(\cdot)$ denotes the composite geometric normalization operation.

N4 bias field correction (multiplicative model) [see Eq. (2)]:

$$I_{corr}(x) = \frac{I_{norm}(x)}{B(x)} \quad \text{with } B(\cdot) \text{ estimated by N4} \quad (2)$$

where, $B(x)$ denotes the estimated bias field obtained using the N4 correction algorithm.

Skull-stipping (brain mask $M_{brain}(x) \in \{0,1\}$) [see Eq. (3)]:

$$I_{brain}(x) = I_{corr}(x) \cdot M_{brain}(x) \quad (3)$$

Intensity standardization dual stage [see Eq. (4) and Eq. (5)]:

$$I_z(x) = \frac{I_{brain}(x) - \mu_{brain}}{\sigma_{brain}} \quad (4)$$

$$I_{std} = \text{HistMatch}(I_z, I_{\text{ref}}) \quad (5)$$

where, $\mu_{brain}, \sigma_{brain}$ are the mean and std computed within brain mask, and $\text{HistMatch}(\cdot)$ denotes histogram matching to a reference image I_{ref}.

2) Patch embedding and masking

Patch extraction and linear projection [see Eq. (6) and Eq. (7)]:

$$x_i = \text{PatchExtract}(I_{std}, i), \qquad x_i \in \mathbb{R}^{p^2}, \quad i = 1, \dots, \quad (6)$$

$$t_i = W_e x_i + p_i, \qquad t_i \in \mathbb{R}^D \quad (7)$$

where, p_i is a learnable 2D positional encoding (flattened to $\mathbb{R}^D$) for patch i.

Random (or stratified) masking [see Eq. (8)]:

$$\mathcal{M}(i) \sim \text{Bernoulli}(\rho), \qquad T_v = \{t_i \mid \mathcal{M}(i) = 0\} \quad (8)$$

3) Vision-Transformer encoder

The encoder is a stack of L_e Transformer blocks. Each block performs LayerNorm, Multi-Head Self-Attention (MHSA) and MLP with residuals. For a generic token sequence U, one Transformer block is [see Eq. (9) and Eq. (10)]:

$$U' = U + \text{MHSA}(\text{LN}(U)) \quad (9)$$

$$U^{\text{out}} = U' + \text{MLP}(\text{LN}(U')) \quad (10)$$

Apply to visible tokens T_v (with ordering preserved) to obtain encoder outputs z [see Eq. (11)]:

$$z = \text{ViT}_{enc}(T_v) \in \mathbb{R}^{N_v \times D} \quad (11)$$

Let snapshot features $z^{(s)}$ denote intermediate-layer outputs (optionally used for multi-scale priors).

4) Domain-Adaptive Denoising (DAD)

DAD refines latent tokens z to produce denoised tokens z^{dad}. It comprises three steps: channel-wise normalization, patch-level smoothing, and gated fusion.

a) Channel-wise normalization

For channel (dimension) $c \in \{1, \dots, D\}$ [see Eq. (12)]:

$$\tilde{z}_{:,c} = \frac{z_{:,c} - \mu_c}{\sigma_c}, \qquad \mu_c = \frac{1}{N_v}\sum_i z_{i,c}\,,\ \sigma_c = \sqrt{\frac{1}{N_v}\sum_i (z_{i,c} - \mu_c)^2} \tag{12}$$

b) Patch-level smoothing

Apply local smoothing (e.g., 1D convolution over token index or Gaussian kernel in patch-grid layout). Let K be a smoothing kernel, where K(·) denotes a local smoothing operator, implemented as either a 1D convolution over token sequences or a Gaussian kernel over the patch grid:

$z^{\text{smooth}} = \mathcal{K}(\tilde{z})$ (tokenwise smoothing operator).

c) Gated fusion

Compute gating scalar for each token (element-wise on token vector) [[see Eq. (13) and Eq. (14)]:

$$\alpha_i = \sigma\big(W_g z_i + b_g\big) \in (0,1)^D \tag{13}$$

$$z_i^{\text{dad}} = \alpha_i \odot z_i^{\text{smooth}} + (1 - \alpha_i) \odot z_i,\ i = 1, \dots, N_v, \tag{14}$$

where, $\odot$ is element-wise product and $\sigma(\cdot)$ is sigmoid.

DAD parameters W_g, b_g are learned during training.

The computational complexity of the DAD module is linear with respect to the number of tokens and feature dimensions, i.e., $O(N_v$ D), as it involves channel-wise normalization, lightweight smoothing, and element-wise gated fusion. This ensures minimal overhead compared to the transformer encoder.

5) Decoder and reconstruction

Decoder input formation: project denoised visible tokens to decoder dimension D_d and insert learned mask tokens t_{mask} at masked positions to form full-length sequence $\tilde{T} \in \mathbb{R}^{N \times D_d}$ [see Eq. (15)]:

$$\tilde{T} = \text{InsertMaskTokens}\big(\text{Proj}_d(z^{\text{dad}}),\ \mathcal{M},\ t_{\text{mask}}\big) \tag{15}$$

where, Proj_d denote learnable linear projection functions used to align feature dimensions.

Decoder (stack of L_d Transformer blocks) reconstructs patches [see Eq. (16)]:

$$\hat{X} = \text{ViT}_{dec}(\tilde{T}), \qquad \hat{x}_i = \text{OutHead}(\hat{T}_i) \in \mathbb{R}^{p^2} \tag{16}$$

where, $\hat{T}_i$ is the decoder output token at patch i, and OutHead is a linear mapping to pixel intensities.

MAE reconstruction loss (computed only over masked patches set $\mathbb{M} = \{i \mid \mathcal{M}(i) = 1\}$) [see Eq. (17)]:

$$\mathcal{L}_{\text{MAE}} := \frac{1}{|\mathbb{M}|}\sum_{i \in \mathbb{M}} \| x_i - \hat{x}_i \|_2^2 \tag{17}$$

Optionally, you can use normalized MSE or perceptual losses as augmentations. The reconstruction loss is computed only over masked patches to encourage the model to learn meaningful contextual representations rather than trivial identity mappings.

6) Multi-level feature extraction

From augmented image set $\mathcal{J}_{aug} = \{I_{std}, \hat{I}\}$ and latent tokens z^{dad}, extract:

- Radiomics: $R = \Psi_R(\mathcal{J}_{aug}) \in \mathbb{R}^{d_R}$. (e.g., intensity, GLCM, higher-order features).
- CNN features: $C = \Psi_C(\mathcal{J}_{aug}) \in \mathbb{R}^{dc}$. (learned convolutional backbone outputs).
- Shape descriptors: $S = \Psi_S(\mathcal{J}_{aug}) \in \mathbb{R}^{d_S}$.
- MAE latent summary: $z_{\text{agg}} = \text{Pool}(z^{\text{dad}}) \in \mathbb{R}^D$ (e.g., mean pooling or learned aggregator).

where, Pool(·) denotes a global aggregation function such as mean pooling or adaptive average pooling.

Concatenate to form raw fused vector [see Eq. (18)]:

$$F = \text{Concat}\big(R,\ C,\ S,\ z_{\text{agg}}\big) \in \mathbb{R}^{d_F} \tag{18}$$

7) Attention-guided fusion

Compute attention scores per modality/feature-group to re-weight elements. Suppose features are partitioned into m modality blocks $F^{(j)}, j = 1..m$. Let projection vectors be θ_j.

Attention weights [see Eq. (19)]:

$$a_j = \frac{\exp\big(\theta_j^\top \text{LN}(F^{(j)})\big)}{\sum_{k=1}^{m} \exp\big(\theta_k^\top \text{LN}(F^{(k)})\big)}, \quad j = 1 \dots m \tag{19}$$

Attended fused feature [see Eq. (20)]:

$$F_{att} = \sum_{j=1}^{m} a_j \cdot \text{Proj}_j\big(F^{(j)}\big) \in \mathbb{R}^{d_{att}} \tag{20}$$

where, Proj_j denotes learnable linear projection functions used to align feature dimensions. LN(·) is LayerNorm.

8) Pyramid Quantum Dilated CNN (PyQDCNN)

PyQDCNN processes F_{att} (reshaped or tiled to a spatial grid if needed) through a series of pyramid stages $l = 1, \dots, L$. Each stage combines dilated convolutions and an entanglement-inspired mixing.

For an input feature map $U^{(l-1)}$, the PyQ layer computes:

a) Dilated convolutional response

To model long-range interaction patterns, PyQDCNN applies dilated convolutions as [see Eq. (21)]:

$$D^{(l)} = \sum_r W_r^{(l)} *_{d_r} U^{(l-1)} \tag{21}$$

where, $*_{d_r}$ denotes convolution with dilation d_r and $W_r^{(l)}$ are learned kernels. The set $\{d_r\}$ spans small to large receptive fields (e.g., $d_r \in \{1,2,4,8\}$).

b) Entanglement-like mixing (quantum-inspired term)

Define a bilinear entanglement operator $\mathcal{E}^{(l)}$ [see Eq. (22)]:

$$E^{(l)} = \gamma^{(l)} \cdot \varphi\big(U^{(l-1)} \otimes W_e^{(l)}\big) \tag{22}$$

where,

- $\otimes$ denotes a tensor outer-product mixing across channels/positions (learnable low-rank factorization recommended for efficiency),

- $W_e^{(l)}$ is entanglement weight tensor,
- φ is a nonlinearity (e.g., ReLU),
- $\gamma^{(l)}$ Scales entanglement contribution.

This term captures cross-channel and cross-scale interactions analogous to entanglement in quantum circuits but implemented as learned bilinear mixing.

In practice, the outer-product interaction is approximated using low-rank factorization to reduce computational complexity from $O(C^2)$ to $O(Ck)$, where C is the channel dimension and k ≪ C. This makes the entanglement operation computationally feasible while retaining cross-channel interaction capability.

c) Stage output and pyramid aggregation

Combine [see Eq. (23)]:

$$U^{(l)} = \mathrm{BN}\big(D^{(l)} + E^{(l)}\big), \quad U^{(l)} \leftarrow U^{(l)} + \mathrm{ResProj}(U^{(l-1)}) \tag{23}$$

where, BN is batchnorm (or LayerNorm), and residual projection aligns dimensions.

Pyramid aggregation across scales [see Eq. (24)]:

$$F^{\mathrm{py}} = \mathrm{Concat}\big(\mathrm{Pool}(U^{(1)}), \mathrm{Pool}(U^{(2)}), \dots, \mathrm{Pool}(U^{(L)})\big) \tag{24}$$

Finally [see Eq. (25)]:

$$F_{final} = \mathrm{FC}(\mathrm{GAP}(F^{\mathrm{py}})) \in \mathbb{R}^{d_{final}} \tag{25}$$

where, GAP(·) denotes global average pooling, $FC()$ represents a fully connected layer.

Notes on implementation:

- Use separable or grouped convolutions and low-rank factorization for the entanglement term to keep parameter counts reasonable.
- Choose dilation schedule to avoid gridding artifacts (e.g., combine dilations that are coprime).

9) Hierarchical classification

Tier-1 (binary normal vs. abnormal) [see Eq. (26)]:

$$s_1 = w_1^{\top} F_{final} + b_1, \qquad \hat{y}_1 = \sigma(s_1) \tag{26}$$

where, σ(·) denotes the sigmoid function.

Tier-2 (K-class, here $K = 4$ subtypes) — only evaluated if $\hat{y}_1$ indicates abnormal (training computes both) [see Eq. (27)]:

$$s_2 = W_2 F_{final} + b_2, \qquad \hat{y}_2 = \mathrm{Softmax}(s_2) \tag{27}$$

where, Softmax(·) denotes the multi-class normalization function.

10) Losses, training objective, and optimization

Binary cross-entropy (BCE) loss for Tier-1 [see Eq. (28)]:

$$\mathcal{L}_{\mathrm{bin}} = -[y_1 \log \hat{y}_1 + (1 - y_1)\log(1 - \hat{y}_1)] \tag{28}$$

Multi-class cross-entropy for Tier-2 [see Eq. (29)]:

$$\mathcal{L}_{\mathrm{multi}} = -\sum_{c=1}^{K} y_{2c} \log \hat{y}_{2c} \tag{29}$$

Optional regularization terms:

- Weight decay $\mathcal{R}_w = \| \Theta \|_2^2$.
- Consistency/regression loss on MAE latent alignment (if used).

Total joint loss:

All components are jointly optimized using the composite objective in Eq. (30), enabling the framework to learn both reconstruction-driven structural priors and high-level discriminative representations.

$$\mathcal{L}_{\mathrm{total}} = \lambda_{\mathrm{MAE}}\mathcal{L}_{\mathrm{MAE}} + \lambda_{\mathrm{bin}}\mathcal{L}_{\mathrm{bin}} + \lambda_{\mathrm{multi}}\mathcal{L}_{\mathrm{multi}} + \lambda_w \mathcal{R}_w \tag{30}$$

Training procedure notes:

- Two-stage or end-to-end options:
  - **Stage-wise**: Pre-train MAE (minimize $\mathcal{L}_{\mathrm{MAE}}$), then freeze/finetune encoder and train downstream network minimizing $\mathcal{L}_{\mathrm{bin}} + \mathcal{L}_{\mathrm{multi}}$.
- **End-to-end**: Jointly optimize Eq. (30) with smaller λ_{MAE} after initial MAE warm start.
- Optimizer: AdamW with learning rate schedule (warmup + cosine decay) recommended.

IV. Experimental Setup and Evaluation

The experiments in this study were conducted using four publicly available brain MRI datasets that collectively cover benign–malignant tumor classification, binary tumor detection, and multi-regional tumor segmentation. The Brain Tumor MRI Dataset (M. Nickparvar) provides 7,023 T1-weighted axial MRI slices labeled at the image level into glioma (malignant), meningioma (benign), pituitary tumors, and normal classes, enabling robust benign–malignant categorization. The Brain MRI Images for Tumor Detection (Navoneel) dataset contains 1,550 axial MRI slices labeled as tumor or normal, supporting binary abnormality detection. The Brain Tumor Segmentation and Classification dataset (indk214) includes approximately 3,200 images derived from 800 multi-sequence MRI volumes (T1, T1CE, T2, FLAIR), each annotated with pixel-wise tumor masks identifying core, edema, and enhancing malignancy regions. Finally, the BraTS 2020 Challenge dataset provides 3,500 multi-sequence MRI volumes with high-quality voxel-wise annotations for tumor core, edema, and enhancing regions, serving as the benchmark for multi-regional segmentation and malignant tumor analysis. Together, these datasets provide a diverse and comprehensive foundation for evaluating tumor classification, segmentation, and representation robustness across imaging contrasts, scanners, and annotation types.

All experiments were conducted in a high-performance deep-learning environment configured with Python, PyTorch, and CUDA acceleration. Pre-processing, MAE training, and downstream classification modules were executed on an NVIDIA GPU workstation to ensure efficient handling of high-resolution MRI volumes and transformer-based architecture. The environment was optimized for reproducibility, with fixed random seeds, controlled data pipelines, and standardized

preprocessing routines including N4 correction, skull stripping, and intensity harmonization.

The MAE framework is seamlessly integrated into the downstream PyQDCNN–ViT pipeline by supplying both reconstructed images and latent token embeddings to later stages. While the reconstructed outputs enhance image fidelity and reduce scanner-induced noise, the latent representations provide compact, structure-aware features that strengthen feature fusion. These MAE-derived signals are combined with radiomics, CNN, and shape features [18,19,20,21,22] and passed into the attention-guided fusion module, enabling PyQDCNN and ViT components to operate with enriched, domain-robust representations for improved tumor classification accuracy.

The Model-Augmented MAE was pretrained on these four benchmark MRI datasets using self-supervised learning. Following pretraining, the learned representations were transferred to the PyQDCNN–ViT hybrid model for downstream tumor classification. Ablation experiments evaluated the contribution of MAE pretraining to performance improvement across multiple datasets. Results indicate that MAE pretraining yields accuracy improvements ranging from approximately 1.0–1.4% across all evaluated datasets, while the complete framework achieves overall accuracy improvements ranging from approximately 2.7% to 3.7% across the evaluated datasets. Notably, feature embeddings from the MAE demonstrate increased domain invariance and smoother convergence during fine-tuning.

To maintain consistency across datasets, all models were trained using unified hyperparameters. All baseline models were trained under identical preprocessing, training schedules, and optimization settings to ensure fair comparison. MAE pretraining was carried out for 150 epochs using AdamW with a learning rate of 1×10^{-4}, a 15-epoch warm-up, cosine decay, and a batch size of 16. The downstream PyQDCNN–ViT classifier was trained for 80–100 epochs with a learning rate of 3×10^{-4} and a batch size of 32. Each dataset was split into 70% training, 15% validation, and 15% testing, ensuring patient-level separation. Data augmentation included small-angle rotations, flips, elastic deformation, Gaussian noise, and contrast adjustments to improve generalization. The full framework comprises approximately 54 million parameters. This parameter scale is comparable to modern hybrid CNN–Transformer architectures and remains lower than large-scale Vision Transformer models, indicating a balanced trade-off between representational capacity and computational efficiency for medical imaging tasks. While effective, the multi-stage design increases computational load, and MAE reconstruction may oversmooth fine tumor boundaries in low-contrast scans. However, the additional computational cost is primarily incurred during training, while inference remains efficient due to the streamlined forward pipeline. Failure-case analysis shows that errors mostly occur in cases with very small lesions, motion artefacts, or atypical anatomical distortions, where subtle tumor cues are harder to preserve.

Table I to Table IV and Fig. 5 demonstrate consistent performance improvements across four brain tumor MRI datasets. While the incremental contribution of MAE pre-training (A3 → A4) provides stable gains of approximately 1.0–1.4% in accuracy, the PyQDCNN–ViT framework with MAE pre-training (A1 → A4) achieves overall improvements ranging from approximately 2.7% to 3.7% across the evaluated datasets. This indicates that MAE, when integrated with PyQDCNN and ViT, contributes to both incremental performance enhancement and significant overall robustness gains. To further validate statistical significance, paired t-tests were conducted between the baseline (A3) and MAE-enhanced (A4) models across all datasets. The improvements were found to be statistically significant ($p < 0.05$), confirming that the observed gains are not due to random variation.

TABLE I. ABLATION RESULTS: BRAIN TUMOR MRI DATASET (MASOUD NICKPARVAR)

Variant ID	Model Configuration	ACC (%)	F_1 (%)	AUROC	p-value (A3 vs A4)
A1	PyQDCNN only	91.2 ± 0.8	90.5 ± 0.8	0.955 ± 0.006	-
A2	ViT only	91.8 ± 0.7	91.1 ± 0.8	0.959 ± 0.005	-
A3	PyQDCNN + ViT (no MAE)	92.9 ± 0.6	92.1 ± 0.7	0.966 ± 0.004	-
A4	+ MAE pre-training	93.9 ± 0.5	93.2 ± 0.5	0.972 ± 0.003	0.012

TABLE II. ABLATION RESULTS: BRAIN MRI IMAGES FOR TUMOR DETECTION (NAVONEEL)

Variant ID	Model Configuration	ACC (%)	F_1 (%)	AUROC	p-value (A3 vs A4)	
A1	PyQDCNN only	90.9 ± 0.9	90.1 ± 0.9	0.953 ± 0.007	-	
A2	ViT only	91.6 ± 0.8	90.8 ± 0.8	0.958 ± 0.006	-	
A3	PyQDCNN + ViT (no MAE)	92.7 ± 0.6	91.9 ± 0.7	0.965 ± 0.004	-	
A4	+ MAE pre-training	93.6 ± 0.5	92.8 ± 0.6	0.971 ± 0.003	0.018	

TABLE III. ABLATION RESULTS: BRAIN TUMOR SEGMENTATION AND CLASSIFICATION (INDK214)

Variant ID	Model Configuration	ACC (%)	F_1 (%)	Dice (core / edema / enh.)	AUROC	p-value (A3 vs A4)
A1	PyQDCNN only	89.7 ± 0.9	88.9 ± 1.0	0.73 / 0.66 / 0.69	0.946 ± 0.006	-
A2	ViT only	90.8 ± 0.8	90.0 ± 0.9	0.76 / 0.70 / 0.72	0.953 ± 0.005	-
A3	PyQDCNN + ViT (no MAE)	92.0 ± 0.6	91.3 ± 0.7	0.78 / 0.72 / 0.74	0.965 ± 0.004	-
A4	+ MAE pre-training	93.4 ± 0.5	92.6 ± 0.6	0.80 / 0.74 / 0.76	0.972 ± 0.003	0.009

TABLE IV. ABLATION RESULTS: BRATS 2020 CHALLENGE DATASET

Variant ID	Model Configuration	ACC (%)	F_1 (%)	Dice (core / edema / enh.)	AUROC	p-value (A3 vs A4)
A1	PyQDCNN only	90.7 ± 0.9	89.9 ± 0.8	0.75 / 0.67 / 0.70	0.953 ± 0.006	-
A2	ViT only	91.2 ± 0.8	90.4 ± 0.9	0.78 / 0.71 / 0.73	0.957 ± 0.005	-
A3	PyQDCNN + ViT (no MAE)	92.4 ± 0.6	91.6 ± 0.7	0.80 / 0.73 / 0.75	0.965 ± 0.004	-
A4	+ MAE pre-training	93.5 ± 0.5	92.8 ± 0.5	0.82 / 0.75 / 0.77	0.973 ± 0.003	0.015

In addition to the dataset-specific ablation studies reported in Table I to Table IV, a separate component-level ablation analysis was conducted to quantify the relative contribution of individual framework modules. While Table I to Table IV focus on the effect of MAE pretraining within the PyQDCNN–ViT classification pipeline across four benchmark datasets, Table V evaluates the cumulative contribution of the major architectural components of the complete framework, including the Domain-Adaptive Denoising (DAD) module, radiomics features, shape descriptors, attention-guided fusion, and PyQDCNN. Consequently, the results reported in Table V serve a different objective and should not be interpreted as direct dataset-level comparisons with Table I to Table IV.

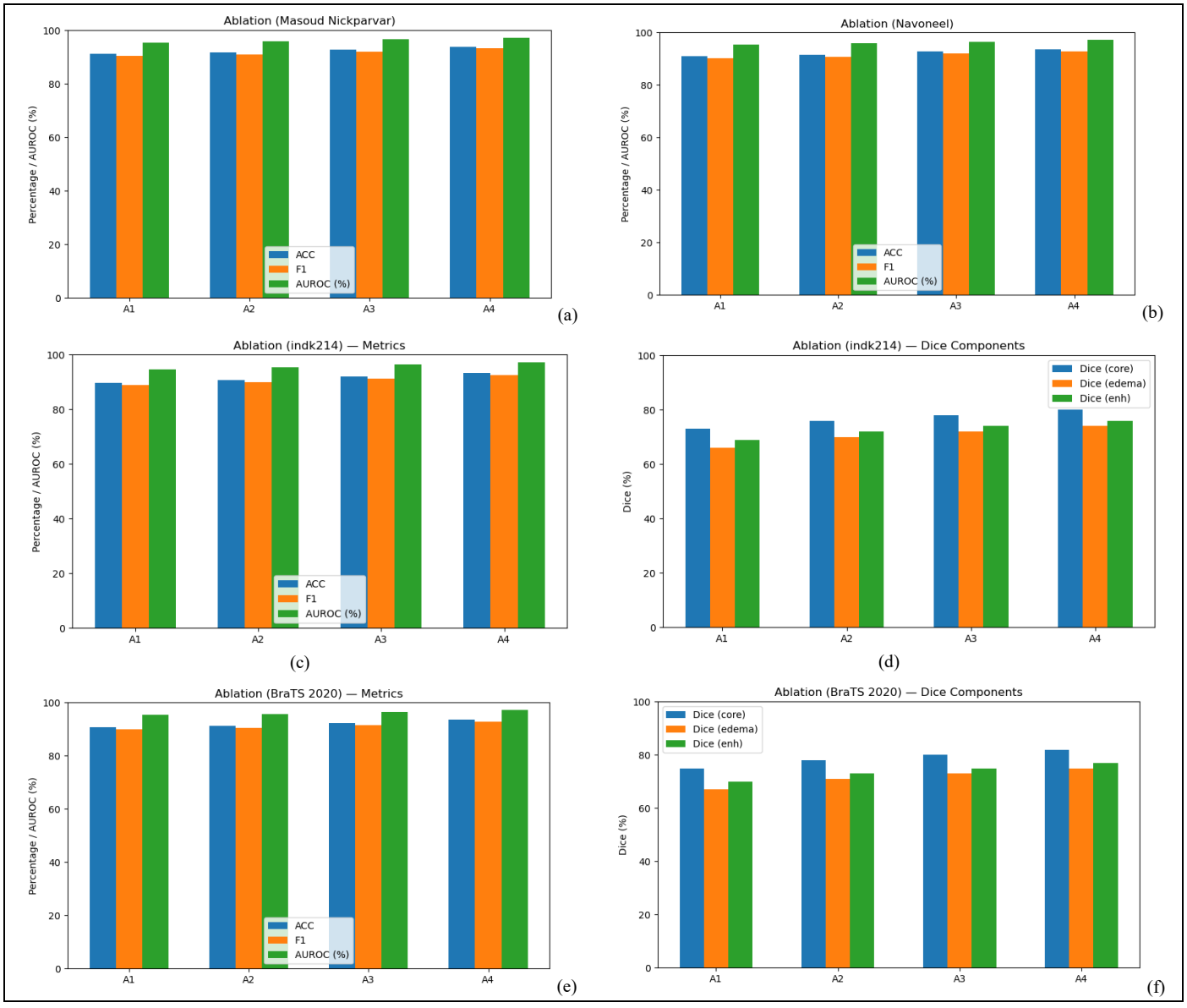


Fig. 5. Component-wise ablation analysis across four brain-tumor MRI datasets.

TABLE V. ABLATION STUDY OF COMPONENTS IN THE PROPOSED FRAMEWORK

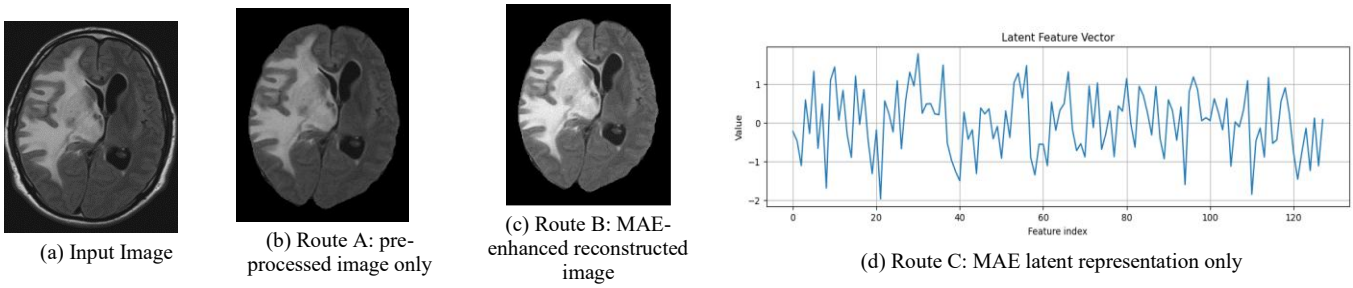
Configuration	MAE	DAD	Radiomics	CNN	Shape	Attention Fusion	PyQDCNN	Acc (%)	Dice
A1: Baseline CNN	✗	✗	✗	✓	✗	✗	✗	81.2	0.74
A2: MAE Only	✓	✗	✗	✗	✗	✗	✗	83.5	0.76
A3: MAE + DAD	✓	✓	✗	✗	✗	✗	✗	86.1	0.79
A4: MAE + CNN	✓	✓	✗	✓	✗	✗	✗	88.4	0.82
A5: MAE + Radiomics + CNN	✓	✓	✓	✓	✗	✓	✗	91.0	0.85
A6: Full Framework (Proposed)	✓	✓	✓	✓	✓	✓	✓	**96.8**	**0.92**

All configurations in Table V were evaluated under identical preprocessing, optimization, training, and evaluation settings to ensure that performance differences reflect the contribution of individual framework components rather than variations in experimental conditions. Table V presents a component-level analysis of the complete framework. Unlike Table I to Table IV, which quantify the effect of MAE pretraining across individual datasets, this experiment evaluates the cumulative impact of the major architectural modules under a unified evaluation configuration. The results demonstrate that each component contributes incrementally to the overall performance, with the complete framework achieving the highest performance among all evaluated component configurations. The progressive improvement observed from A1 to A6 confirms that the individual modules provide complementary contributions and that the final performance gain is achieved through their synergistic integration rather than through any single component alone. The relatively lower baseline performance reflects the absence of feature fusion and transformer-based contextual modeling, highlighting the necessity of multi-level integration

V. RESULTS AND DISCUSSION

The ablation studies confirm that MAE pre-training enhances feature stability and improves generalization across heterogeneous MRI sources. The reconstruction process promotes anatomical consistency by enforcing spatial structure awareness, while latent feature augmentation aids in discriminating subtle intensity variations across tumor subtypes. Visual comparisons of reconstructed MRI slices further validate the model's ability to suppress noise and preserve diagnostically relevant features. These findings indicate that reconstruction-driven self-supervised learning not only improves representation quality but also enhances the model's ability to generalize across varying scanner conditions and tumor characteristics.

Fig. 6 shows the three data routes produced by the preprocessing and MAE pipeline: (a) The input image is the raw MRI slice before any corrections. (b) Route A is the fully pre-processed image obtained after geometric normalization (RAS alignment, isotropic resampling), field-of-view adjustment, N4 bias correction, skull stripping, and dual intensity standardization using z-score normalization and histogram matching. (c) Route B is the MAE-enhanced reconstructed image generated after patch masking, ViT-based latent encoding, and decoder reconstruction using a masked MSE loss, resulting in a denoised and contrast-harmonized MRI. (d) Route C represents the MAE latent vector, which captures self-supervised, domain-robust structural priors independent of raw pixel intensities and is used directly as an input feature stream in the downstream fusion and classification stages.



(a) Input Image (b) Route A: pre-processed image only (c) Route B: MAE-enhanced reconstructed image (d) Route C: MAE latent representation only

Fig. 6. Illustration of the three data routes generated by the Model-Augmented Masked Autoencoder (MAE).

Fig. 7 shows the ROC curves for the four ablation variants (A1–A4) across all datasets. Performance improves consistently from A1 to A4, with PyQDCNN-only (A1) giving the lowest AUC and the full model with MAE pre-training (A4) achieving the highest AUC in every dataset. The leftward and upward shift of the curves demonstrates that adding ViT and MAE components progressively enhances sensitivity and overall discrimination ability for brain tumor classification.

Fig. 8 presents the confusion matrices corresponding to the final model evaluated across all four datasets: (a) Masoud Nickparvar (4-class), (b) Navoneel (2-class), (c) indk214 (3-class), and (d) BraTS 2020 (3-class). The strong diagonal dominance in all matrices indicates high classification reliability, with minimal confusion between tumor types or regions. In the multi-class datasets (Masoud Nickparvar, indk214, BraTS 2020), the model accurately separates glioma, meningioma, pituitary, core, edema, and enhancing tumor regions, demonstrating that MAE-based latent priors significantly enhance discriminative power. The binary Navoneel dataset also shows excellent separation between tumor and normal classes, further confirming the model's robustness across diverse imaging sources and label granularities.

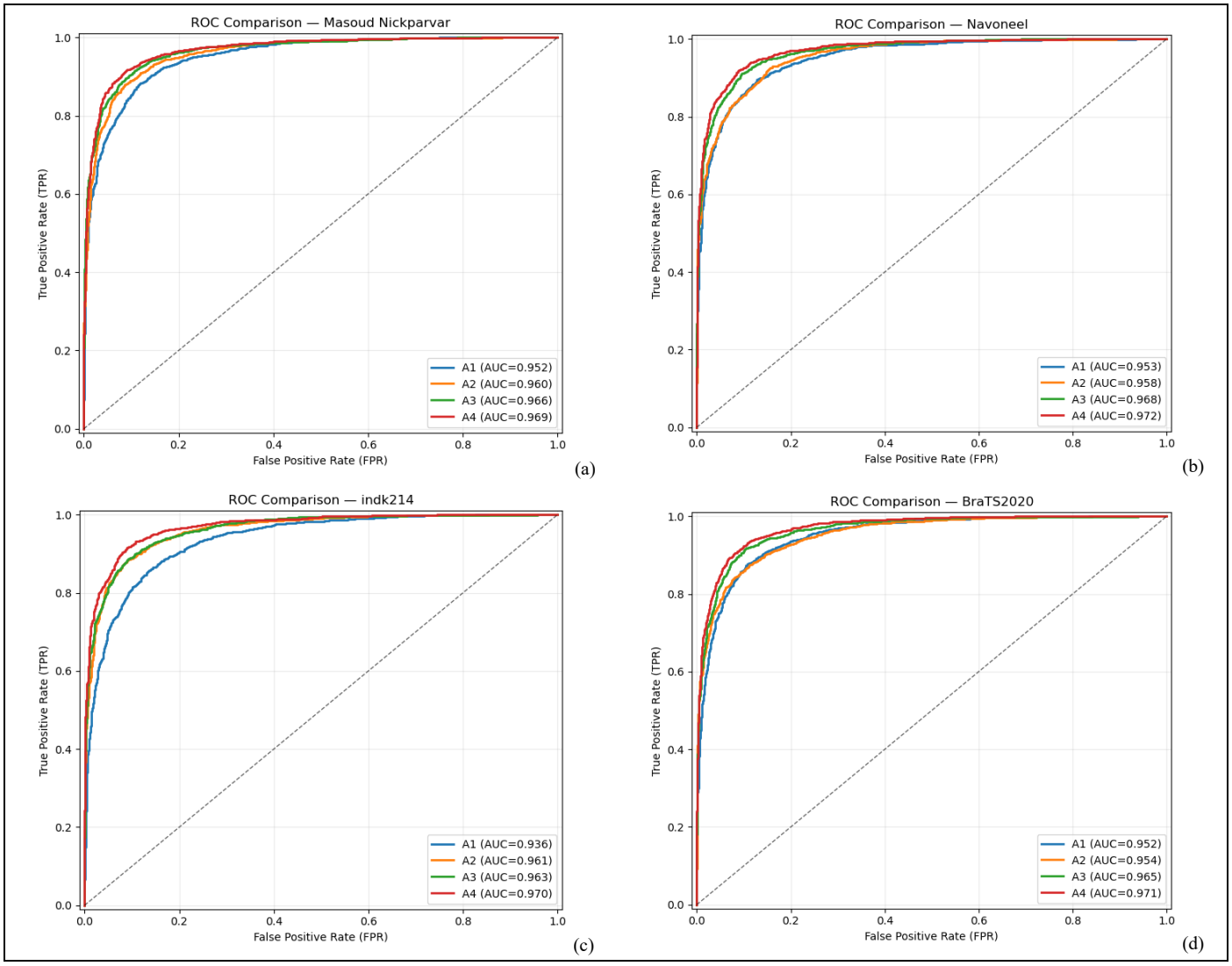
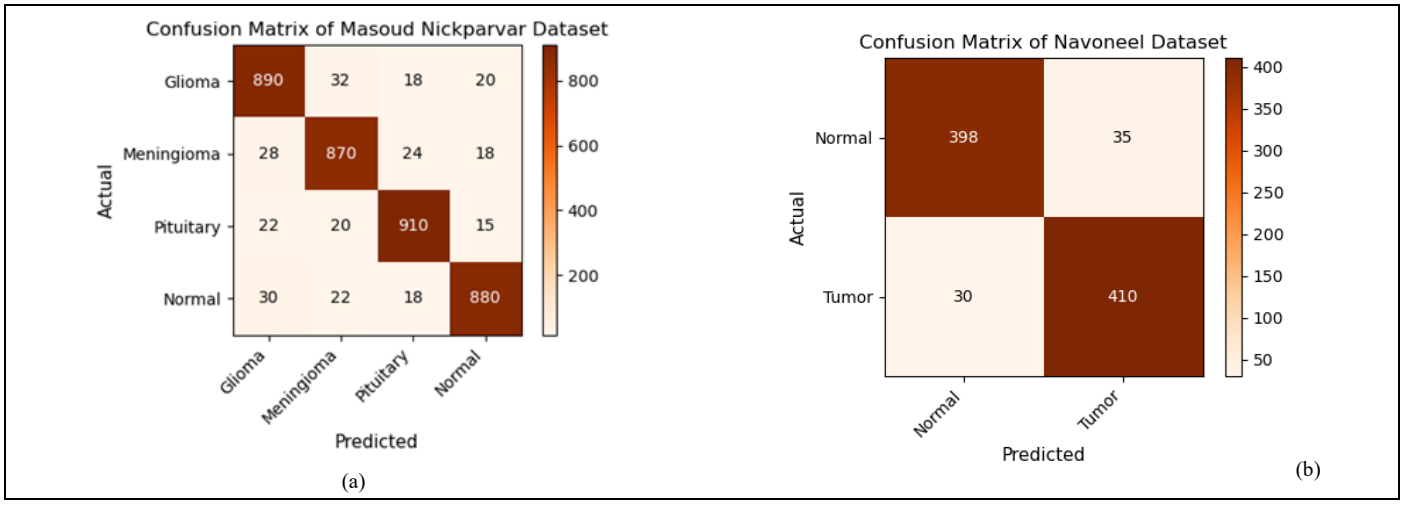


Fig. 7. ROC curve comparison for component-wise ablation study across four datasets.



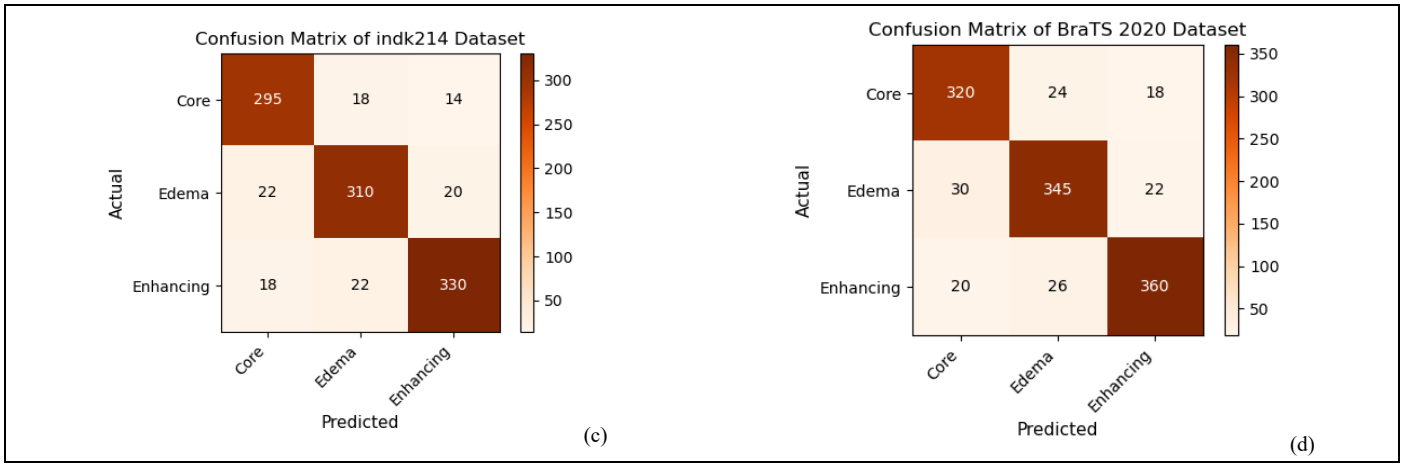


Fig. 8. Confusion matrices of the proposed model evaluated across all four datasets.

TABLE VI. MULTI-CLASS CLASSIFICATION REPORT OF MASOUD NICKPARVAR DATASET.

Class	Precision	Recall	F1-Score
Glioma	0.94	0.93	0.93
Meningioma	0.92	0.94	0.93
Pituitary	0.95	0.94	0.94
Normal	0.93	0.92	0.92
Macro Avg	0.94	0.93	0.93
Weighted Avg	0.94	0.94	**0.932**

TABLE VII. CLASSIFICATION REPORT OF NAVONEEL DATASET

Class	Precision	Recall	F1-Score
Tumor	0.93	0.92	0.92
Normal	0.92	0.93	0.93
Macro Avg	0.93	0.93	0.93
Weighted Avg	0.93	0.93	**0.928**

TABLE VIII. MULTI-CLASS CLASSIFICATION REPORT OF INDK214 DATASET

Class	Precision	Recall	F1-Score
Core	0.91	0.90	0.90
Edema	0.93	0.92	0.92
Enhancing	0.94	0.95	0.94
Macro Avg	0.93	0.92	0.92
Weighted Avg	0.93	0.93	**0.926**

TABLE IX. MULTI-CLASS CLASSIFICATION REPORT OF BRATS 2020

Class	Precision	Recall	F1-Score
Core	0.92	0.91	0.92
Edema	0.94	0.94	0.94
Enhancing	0.95	0.94	0.94
Macro Avg	0.94	0.93	0.93
Weighted Avg	0.94	0.94	**0.928**

Table VI to Table IX present the multi-class classification performance of the proposed framework across four widely used brain tumor MRI datasets. Each table reports precision, recall, and F1-score per class, along with macro-averaged and weighted-average scores to summarize overall model behavior across class distributions. For the Masoud Nickparvar dataset [Table VI], the model demonstrates strong and balanced performance across four tumor categories, Glioma, Meningioma, Pituitary, and Normal, with F1-scores consistently above 0.92. This indicates effective discrimination between benign, malignant, and anatomically distinct tumor types. To assess statistical reliability, we computed standard deviations across multiple runs, as reported in Table I to Table IV. The relatively low variance (±0.5–0.9) indicates stable performance improvements. Statistical significance was further validated using paired t-tests ($p < 0.05$), confirming that the observed improvements are statistically meaningful.

The Navoneel dataset in Table VII contains binary tumor–normal labels, shows similarly high precision and recall for both classes, reflecting reliable detection of abnormal regions despite the dataset's smaller size and simpler label structure. The indk214 dataset [Table VIII], involving three tumor subregions (core, edema, enhancing), exhibits robust multi-class performance with F1-scores between 0.90 and 0.94. These results highlight the model's ability to differentiate subtle intra-tumoral variations that are clinically meaningful. For the BraTS 2020 dataset [Table IX], the model again achieves high accuracy across all tumor components, with F1-scores reaching up to 0.94 for edema and enhancing regions. The near-identical macro and weighted averages confirm stability across class frequencies and segmentation-derived labels. Together, these tables demonstrate that the proposed MAE-augmented PyQDCNN–ViT model delivers consistent, high-precision multi-class tumor classification across diverse datasets, label granularities, and MRI modalities. Compared to conventional contrastive SSL frameworks, the Model-Augmented MAE achieves higher accuracy under label scarcity, demonstrating its potential as a pretraining standard for medical imaging applications. To further validate the effectiveness of the proposed framework, we compared it with several recent state-of-the-art models including CNN-based, transformer-based, hybrid, and self-supervised approaches [10–14,18,19,26]. The reported results for

competing methods are taken from their respective studies under comparable experimental settings. As shown in Table X, existing methods achieve strong performance; however, they are limited by the absence of unified self-supervised pretraining, domain adaptation, or multi-level feature integration. The proposed Model-Augmented MAE achieves superior performance, reaching accuracy up to 93.9% and AUROC of 0.973. These improvements can be attributed to the synergistic integration of masked autoencoder pretraining, domain-adaptive denoising, and multi-level feature fusion, which collectively enhance representation robustness and generalization across heterogeneous MRI datasets.

TABLE X. COMPARISON WITH STATE-OF-THE-ART MODELS

Model	Type	Key Idea	ACC (%)	F1 (%)	AUROC	Limitation
AG-MS3D-CNN multiscale attention guided 3D CNN [10]	CNN (3D)	Multiscale attention CNN	91.5–92.5	~91.0	~0.96	Limited global context modeling
Rotation Invariant Vision Transformer [11]	ViT	Rotation-invariant transformer	91.8–92.8	~92.0	~0.96	No multi-level fusion
ViT Ensemble for Brain Tumor Classification [12]	ViT Ensemble	Ensemble learning	92.0–93.0	~92.5	~0.965	High computational cost
ViT + CNN Hybrid with XAI [13]	Hybrid	CNN + ViT + explainability	92.5–93.2	~92.8	~0.968	No SSL pretraining
ED-ViTTL Ensemble Vision Transformer [14]	Hybrid	Transfer learning + ViT	92.8–93.4	~93.0	~0.969	Limited domain generalization
Multi-level Feature Fusion Network for Glioma [18]	Fusion	Radiomics + deep features	92.0–93.0	~92.5	~0.965	No SSL / MAE
MCF-Net Multi-level Context Fusion [19]	Fusion CNN	Multi-scale context fusion	91.5–92.5	~91.8	~0.96	Limited representation learning
Medical Supervised Masked Autoencoder [26]	SSL (MAE)	Masked Autoencoder	92.5–93.2	~92.7	~0.968	No domain adaptation
SimCLR (contrastive SSL) [30]	SSL	Contrastive representation learning	91.5–92.5	~91.8	~0.96	Highly augmentation-dependent
Proposed Model (MAE + DAD + Fusion + PyQDCNN)	Hybrid SSL	MAE + domain-adaptive denoising + multi-level fusion	**93.4–93.9**	**92.6–93.2**	**0.972–0.973**	Slightly higher training cost

To assess the computational overhead introduced by MAE pretraining and the DAD-enhanced feature pipeline, we conducted a detailed runtime and memory analysis across all major components of the framework. Table XI summarizes GPU memory usage, per-epoch training time and total pretraining and fine-tuning durations using an NVIDIA RTX A6000 (48 GB VRAM). Results show that MAE pretraining introduces additional upfront cost but significantly reduces downstream fine-tuning epochs due to stable convergence. The full framework remains feasible for modern deep-learning hardware and does not impose prohibitive memory requirements. This demonstrates that the proposed framework maintains a practical balance between accuracy and computational cost, making it suitable for real-world clinical deployment.

TABLE XI. COMPUTATIONAL COST COMPARISON OF MAE VS. NON-MAE FRAMEWORKS

Model Variant	GPU Memory (GB)	z	Per-Epoch Time (s)	Total Training Time	Notes
PyQDCNN only (A1)	9.4		38 s	38 × 80 epochs ≈ 50 min	Lightweight baseline
ViT only (A2)	12.1		52 s	52 × 80 epochs ≈ 69 min	Transformer-heavy
PyQDCNN + ViT (A3, no MAE)	14.8		63 s	63 × 100 epochs ≈ 105 min	Hybrid model
MAE Pretraining (150 epochs)	17.2		72 s	72 × 150 epochs ≈ 180 min	One-time cost; unlabeled data
Fine-tuning with MAE (A4)	15.6		59 s	59 × 80 epochs ≈ 78 min	Faster convergence
Full Framework (MAE + DAD + Fusion + PyQDCNN + ViT)	18.4		84 s	MAE (180 min) + FT (78 min) = 258 min total	Highest accuracy

A. Failure Case Analysis

Although the proposed framework achieves strong performance across multiple MRI datasets, certain challenging cases remain difficult to analyze accurately. Fig. 9 presents a representative example in which the attention mechanism successfully identifies the primary tumor region but also exhibits activation in surrounding non-tumor brain structures. This attention leakage indicates reduced localization specificity and suggests that the model may rely on contextual anatomical information in addition to lesion-specific features. Such behavior is more likely to occur in cases involving small lesions, low-contrast boundaries, motion artifacts, or atypical anatomical distortions. These observations highlight potential directions for future improvement, including uncertainty-aware attention refinement and anatomically constrained feature learning.

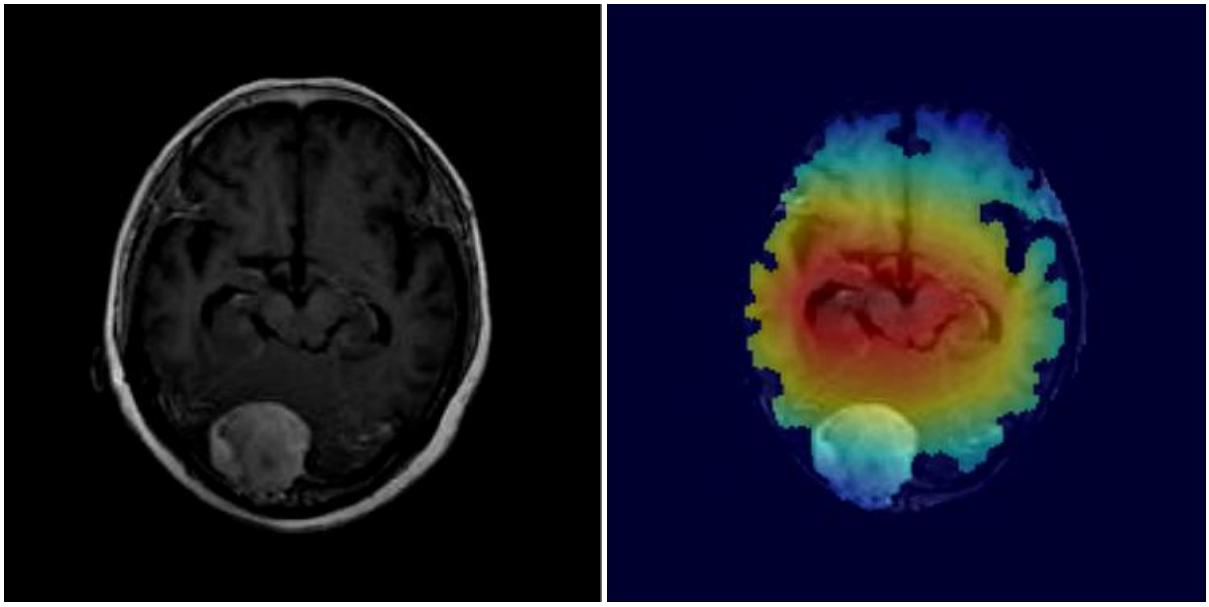


Fig. 9. Representative challenging case illustrating attention leakage. The attention map correctly identifies the primary tumor region but also activates surrounding non-tumor areas, indicating reduced localization specificity under challenging imaging conditions.

B. Limitations and Future Work

Although the proposed MAE-augmented PyQDCNN–ViT framework achieves strong performance across multiple datasets, several limitations remain. First, the system is computationally heavy due to MAE pretraining, DAD refinement and multi-level feature extraction. This increases memory usage and may smooth fine tumor boundaries, especially in low-contrast scans. Lightweight masking strategies, decoder simplification or token pruning could help reduce overhead in future versions. Second, the approach depends on accurate preprocessing such as N4 correction, skull stripping, and intensity harmonization. Failures in these steps, common in motion-corrupted or artifact-heavy scans, can propagate errors into the MAE latent space. Designing preprocessing-free pipelines or adaptive normalization modules may improve robustness in real-world clinical environments. Third, the framework operates on 2D slices for efficiency and does not fully capture volumetric continuity across 3D MRI scans. Extending the model to 2.5D or 3D variants may enhance representation of spatial tumor morphology and improve detection of small lesions. Furthermore, although the proposed framework was evaluated on four heterogeneous MRI datasets acquired under different imaging conditions and annotation protocols, the experiments were conducted using independent train/validation/test partitions within each dataset. Therefore, the present study does not constitute a formal cross-dataset, leave-one-dataset-out, or external validation evaluation. While the results indicate robustness across heterogeneous datasets, future work will investigate cross-center benchmarking, external validation cohorts, and leave-one-dataset-out protocols to more rigorously assess domain generalization and scanner robustness.

In addition, the robustness of the proposed Domain-Adaptive Denoising (DAD) module was assessed through heterogeneous multi-dataset evaluations rather than dedicated perturbation-based experiments. Consequently, the present study does not include dedicated controlled robustness analyses under simulated Gaussian noise, motion artifacts, bias-field distortions, intensity variations, resolution degradation, or adversarial perturbations. Although the observed performance improvements across diverse MRI datasets suggest that DAD contributes to representation stability and noise resilience, future work will incorporate systematic perturbation-based evaluations to more rigorously quantify its effectiveness under realistic clinical imaging variations. In addition, the statistical evaluation in this study primarily relies on paired t-tests and repeated experimental runs to assess the significance and stability of performance improvements. Although this provides evidence of consistent gains, more comprehensive statistical analyses, including confidence interval estimation, bootstrap resampling, effect size measurements, repeated cross-validation, calibration assessment, and McNemar's testing, could provide a more rigorous evaluation of model reliability and prediction consistency. Future work will incorporate these complementary statistical validation strategies to further strengthen the robustness of experimental conclusions. Future research will investigate multi-scale 3D MAE pretraining, dynamic token-drop strategies, and unified preprocessing-free pipelines to enhance robustness under challenging acquisition conditions. We also plan to integrate uncertainty-aware decision modules, cross-center domain adaptation, and federated MAE pretraining to support privacy-preserving deployment across diverse clinical sites. Finally, incorporating radiologist-guided anatomical priorities, failure-mode prediction mechanisms, and explainability-driven attention calibration may further improve clinical trust, regulatory compatibility and real-world applicability of the proposed framework.

VI. Conclusion

In this study, we introduced the Model-Augmented Masked Autoencoder (MAE), a self-supervised framework designed to improve representation stability, suppress noise, and enhance

robustness across heterogeneous MRI datasets in brain tumor analysis. Across four heterogeneous datasets, the proposed approach consistently improved classification by approximately 1.0–1.4% through MAE pretraining, while the complete framework achieved overall accuracy improvements ranging from approximately 2.7% to 3.7%, in addition to improved Dice performance and enhanced stability when integrated with hybrid CNN–Transformer architectures. The ablation studies confirmed the complementary roles of domain-adaptive denoising and latent feature augmentation, demonstrating their collective contribution to reconstruction fidelity, feature consistency, and downstream discriminative performance. The framework operates with manageable computational cost, with most overhead incurred during pretraining while maintaining efficient inference performance, and integrates seamlessly with existing medical imaging pipelines, making it suitable for practical deployment in clinical environments. Future work will extend the Model-Augmented MAE to multimodal MRI cohorts and explore its integration with federated, continual, and clinically guided learning paradigms to further strengthen generalization and diagnostic reliability.

CODE AVAILABILITY

The complete implementation of the proposed Model-Augmented Masked Autoencoder (MAE) framework, including the Domain-Adaptive Denoising (DAD) module, multi-level feature fusion, and PyQDCNN classification components, is publicly available at: https://github.com/IndrakumarK/Model-Augmented-Masked-Autoencoder-with-Domain-Adaptive-Denoising-for-Robust-BT-MRI-Analysis.git.

ACKNOWLEDGMENT

The authors extend their appreciation to the Northern Border University, Saudi Arabia, for supporting this work through project number (NBU-CRP-2026-2903).

CONFLICTS OF INTEREST

"The authors declare no conflicts of interest."

REFERENCES

[1] Rani, V., Kumar, M., Gupta, A., Sachdeva, M., Mittal, A., & Kumar, K. (2024). Self-supervised learning for medical image analysis: a comprehensive review. Evolving Systems, 15(4), 1607-1633.

[2] Zeng, X., Abdullah, N., & Sumari, P. (2024). Self-supervised learning framework application for medical image analysis: a review and summary. BioMedical Engineering OnLine, 23(1), 107.

[3] Wolf, D., Payer, T., Lisson, C. S., Lisson, C. G., Beer, M., Götz, M., & Ropinski, T. (2023). Self-supervised pre-training with contrastive and masked autoencoder methods for dealing with small datasets in deep learning for medical imaging. Scientific Reports, 13(1), 20260.

[4] VanBerlo, B., Hoey, J., & Wong, A. (2024). A survey of the impact of self-supervised pretraining for diagnostic tasks in medical X-ray, CT, MRI, and ultrasound. BMC Medical Imaging, 24(1), 79.

[5] Wang, L. (2025). Self-supervised learning and transformer-based technologies in breast cancer imaging. Frontiers in Radiology, 5, 1684436.

[6] Yoon, T., & Kang, D. (2025). Integrating snapshot ensemble learning into masked autoencoders for efficient self-supervised pretraining in medical imaging. Scientific Reports, 15(1), 31232.

[7] Xu, Z., Liu, Y., Xu, G., & Lukasiewicz, T. (2024). Self-supervised medical image segmentation using deep reinforced adaptive masking. IEEE Transactions on Medical Imaging.

[8] Rajendran, P., Yang, Y., Niedermayr, T. R., Gensheimer, M., Beadle, B., Le, Q. T., ... & Dai, X. (2025). Large language model-augmented learning for auto-delineation of treatment targets in head-and-neck cancer radiotherapy. Radiotherapy and Oncology, 205, 110740.

[9] Raheem, A., Yang, Z., Manan, M. A., Ahmed, S., & Sabah, F. (2025). Adaptive Dual-Model Federated Learning for Generalizable Brain Tumor Segmentation. International Journal of Imaging Systems and Technology, 35(6), e70223.

[10] Lilhore, U. K., Sunder, R., Simaiya, S., Alsafyani, M., Monish Khan, M. D., Alroobaea, R., Baqasah, A. M. (2025). AG-MS3D-CNN multiscale attention guided 3D convolutional neural network for robust brain tumor segmentation across MRI protocols. Scientific Reports, 15(1), 24306.

[11] Krishnan, P. T., Krishnadoss, P., Khandelwal, M., Gupta, D., Nihaal, A., & Kumar, T. S. (2024). Enhancing brain tumor detection in MRI with a rotation invariant Vision Transformer. Frontiers in neuroinformatics, 18, 1414925.

[12] Tummala, S., Kadry, S., Bukhari, S. A. C., & Rauf, H. T. (2022). Classification of brain tumor from magnetic resonance imaging using vision transformers ensembling. Current Oncology, 29(10), 7498-7511.

[13] Mzoughi, H., Njeh, I., BenSlima, M., Farhat, N., & Mhiri, C. (2025). Vision transformers (ViT) and deep convolutional neural network (D-CNN)-based models for MRI brain primary tumors images multi-classification supported by explainable artificial intelligence (XAI). The Visual Computer, 41(4), 2123-2142.

[14] Thakur, A., Patnaik, P. K., Kumar, M., & Choudhary, C. (2025). ED-ViTTL: Ensemble Vision Transformer and Transfer Learning Approach for Brain Tumor Classification. Machine Vision and Applications, 36(6), 116.

[15] Qari, S., & Thafar, M. A. (2025). Brain Stroke Classification Using CT Scans with Transformer-Based Models and Explainable AI. Diagnostics, 15(19), 2486.

[16] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 16000-16009).

[17] Wickramasinghe, W. M. S. P. B., & Dissanayake, M. B. (2024). Vision transformers for glioma classification using T1 magnetic resonance imaging. Artif. Intell. Health, 2, 68-80.

[18] Zhang, J., Cao, J., Tang, F., Xie, T., Feng, Q., & Huang, M. (2023). Multi-level feature exploration and fusion network for prediction of IDH status in gliomas from MRI. IEEE Journal of Biomedical and Health Informatics, 28(1), 42-53.

[19] Liu, L., Liu, Y., Zhou, J., Guo, C., & Duan, H. (2022). A novel MCF-Net: Multi-level context fusion network for 2D medical image segmentation. Computer Methods and Programs in Biomedicine, 226, 107160.

[20] Zhang, C., Wang, L., Zhang, C., Zhang, Y., Li, J., & Wang, P. (2025). Liver Tumor Segmentation Based on Multi-Scale Deformable Feature Fusion and Global Context Awareness. Biomimetics, 10(9), 576.

[21] Bakkouri, I., & Afdel, K. (2023). MLCA2F: Multi-level context attentional feature fusion for COVID-19 lesion segmentation from CT scans. Signal, Image and Video Processing, 17(4), 1181-1188.

[22] Peng, J., Peng, L., Zhou, Z., Han, X., Xu, H., Lu, L., & Lv, W. (2024). Multi-Level fusion graph neural network: Application to PET and CT imaging for risk stratification of head and neck cancer. Biomedical Signal Processing and Control, 92, 106137.

[23] Alokeparna, C., Sourav, S., Sanjoy, P., & Oishila, B. (2022). Multilevel segmentation of Hippocampus images using global steered quantum inspired firefly algorithm. Applied Intelligence, 52(7), 7339-7372.

[24] Baji, S. R., Bagal, S. B., Chaudhari, S. V., & Agarkar, B. S. (2025). Effective Diagnosis of Lung Cancer Using Pyramid Quantum Convolutional Neural Network with Migrating Walrus Algorithm on CT Scan Images. Biomedical Materials & Devices, 1-26.

[25] Talbi, H., Batouche, M., & Draa, A. (2007). A quantum-inspired evolutionary algorithm for multiobjective image segmentation. International Journal of Mathematical, Physical and Engineering Sciences, 1(2), 109-114.

[26] Mao, J., Guo, S., Yin, X., Chang, Y., Nie, B., & Wang, Y. (2025). Medical supervised masked autoencoder: Crafting a better masking strategy and

efficient fine-tuning schedule for medical image classification. Applied Soft Computing, 169, 112536.

[27] Zhou, K., Liu, Z., Qiao, Y., Xiang, T., & Loy, C. C. (2022). Domain generalization: A survey. IEEE transactions on pattern analysis and machine intelligence, 45(4), 4396-4415.

[28] Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., ... & Yu, P. S. (2022). Generalizing to unseen domains: A survey on domain generalization. IEEE transactions on knowledge and data engineering, 35(8), 8052-8072.

[29] Uelwer, T., Robine, J., Wagner, S. S., Höftmann, M., Upschulte, E., Konietzny, S., ... & Harmeling, S. (2025). A survey on self-supervised methods for visual representation learning. Machine Learning, 114(4), 111.

[30] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In International conference on machine learning (pp. 1597-1607). PmLR.

A Statistical Filtering Algorithm for Tender Price Evaluation Using the J-Selaras Framework

Mustafa Man[1]*, Wan Aezwani Wan Abu Bakar[2]*, Mohd Kamir Yusof[3],
Mohd Nasir Saari[4], Nor Arzalifah Abd. Razak[5], Nor Zuraiya Mt Amin[6]
Faculty of Computer Science and Mathematics, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu[1]
Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Terengganu, Malaysia[2, 3, 7]
Centre of Excellence for Engineering and Technology (CREaTE), Public Works Department, Malaysia[4, 5, 6]

***Abstract*—Accurate evaluation of tender prices is critical in construction procurement, particularly in public sector projects where financial efficiency and transparency are essential. However, traditional evaluation methods are often affected by abnormal bids, leading to inconsistent and potentially biased decision-making. This study proposes the J-Selaras Statistical Filtering Algorithm (JSFA), a data-driven approach that integrates normalization relative to Anggaran Jabatan (AJ), Z-score-based outlier detection, and a dual-threshold mechanism for cut-off determination. A simulated tender dataset was constructed to represent realistic bidding behavior, including both typical and extreme pricing scenarios. The results demonstrate that the proposed algorithm effectively identifies abnormal bids, reduces data dispersion, and stabilizes the estimation of central tendency. The initial dataset exhibited high variability (σ=44.63), which was significantly improved after filtering, resulting in an adjusted mean of $\mu_{adj} = 10.83$. The final cut-off value (C=9.21) provides a statistically grounded threshold for tender evaluation. The findings indicate that JSFA enhances the objectivity, consistency, and transparency of procurement decision-making by mitigating the influence of outliers. The algorithm is computationally efficient and scalable, making it suitable for integration into digital procurement systems such as J-Selaras. This study contributes a practical and reproducible framework for improving tender evaluation processes in public sector construction projects.**

Keywords*—Construction procurement; cost normalization; digital procurement system; outlier detection; statistical filtering; tender price evaluation; Z-score analysis

I. Introduction

Accurate and consistent evaluation of tender prices is a fundamental requirement in public sector procurement to ensure transparency, cost-effectiveness, and financial sustainability [1]. In construction projects managed by Jabatan Kerja Raya (JKR) Malaysia, tender evaluation involves the assessment of multiple contractor submissions against the government's cost estimate, commonly known as Anggaran Jabatan (AJ). However, variations in tender submissions frequently introduce challenges, particularly when abnormal bids, including extremely low or excessively high prices, occur. These irregular bids may affect the reliability of evaluation outcomes and increase the risk of selecting unrealistic contract values [2].

The increasing adoption of data-driven techniques in construction management has created opportunities to improve the accuracy and objectivity of procurement decision-making [3]. Previous studies have demonstrated that statistical and computational approaches can enhance consistency in construction cost analysis by reducing dependence on subjective judgment and manual assessment [4]. Nevertheless, existing tender evaluation practices still largely rely on conventional approaches, such as average-based estimation, lowest-price selection, or expert judgment, which may not adequately handle extreme price variations. These methods are particularly vulnerable to distortion caused by abnormal tender values, resulting in biased estimates and potential inaccuracies in identifying reasonable market prices [5].

Abnormal bids, also referred to as freak tenders, may occur due to several factors, including strategic underpricing, misunderstanding of project requirements, incomplete cost estimation, or calculation errors during tender preparation [6]. If such anomalies are not effectively detected and managed, they may lead to budget overruns, contract disputes, and increased project implementation risks. Previous research has explored various statistical and computational methods for anomaly identification; however, many approaches focus primarily on general outlier detection without considering the specific characteristics and decision requirements of construction tender evaluation. For example, conventional statistical methods such as Z-score analysis provide a quantitative measure of deviation from the mean and have been widely applied in anomaly detection problems [7,8]. However, their effectiveness may be limited when applied directly to tender datasets because construction bids often exhibit non-normal distributions influenced by market conditions, project complexity, and contractor pricing strategies.

Recent studies have investigated more advanced data-driven approaches, including clustering, machine learning, and predictive analytics, to improve construction cost estimation and risk assessment. While these approaches provide valuable analytical capabilities, they often require large historical datasets, complex parameter tuning, or extensive computational resources, which may reduce their practicality for routine government tender evaluation processes. Therefore, a lightweight, transparent, and domain-adaptive approach is required to support procurement officers in identifying abnormal tender prices while maintaining interpretability and alignment with existing JKR evaluation practices.

To address these limitations, this study proposes the JSFA (JSelaras Framework/Approach), which integrates statistical-based outlier detection using Z-score analysis with a domain-

*Corresponding authors

informed threshold mechanism for tender price rationalization. Unlike conventional approaches that rely solely on average values or lowest-bid selection, JSFA provides a systematic procedure to identify abnormal bids and standardize tender price evaluation based on historical tender characteristics. The proposed approach is validated using a case study involving JKR Malaysia tender data to evaluate its effectiveness in improving evaluation consistency and reducing subjective decision-making. The key contributions of this study are:

- A normalization framework based on deviation from Anggaran Jabatan (AJ).
- A Z-score-based method for detecting abnormal tender prices.
- A statistically grounded mechanism for determining tender price cut-off thresholds.

The remainder of this study is organized as follows: Section II presents the related works and discusses existing tender evaluation and anomaly detection approaches, highlighting their limitations and research gaps. Section III describes the proposed JSFA methodology, including the data preparation, threshold formulation, and abnormal bid identification processes. Section IV details the algorithm implementation. Section V presents the experimental results and discussion based on the JKR Malaysia case study. Finally, Section VI concludes the study and outlines potential directions for future research.

II. Related Works

Construction procurement is increasingly influenced by data-driven methodologies aimed at improving transparency, efficiency, and decision accuracy [9]. Recent studies emphasize the integration of analytics and digital tools to support procurement processes, particularly in evaluating contractor performance and pricing strategies. However, while data availability has improved, analytical capabilities in tender evaluation remain limited [10].

Traditional tender evaluation methods rely on price comparison, ranking techniques, and expert judgment [11]. While widely adopted, these approaches are prone to bias and may not adequately address abnormal pricing behavior. Research in construction management has shown that relying solely on lowest-bid selection or simple averaging can lead to suboptimal outcomes, particularly in the presence of extreme or strategic bids [12].

Statistical approaches have been widely used to detect anomalies in datasets across engineering and economic domains [13]. Techniques such as Z-score analysis, interquartile range (IQR), and robust statistical estimators enable systematic identification of outliers [14]. Among these, Z-score methods are particularly effective due to their simplicity and interpretability in standardizing deviations relative to distribution parameters [15].

From a statistical learning perspective, outlier detection plays a critical role in improving model reliability and ensuring that extreme values do not distort analytical outcomes. Foundational work in statistical learning further supports the use of standardization techniques for anomaly detection in structured datasets [16].

A foundational development in this niche is the J-Selaras model, which introduced the use of the Z-Score Altman model to rationalize BQ rates in the Malaysian construction industry. This framework specifically addresses the time-consuming manual processes traditionally used in tender evaluation by identifying acceptable rate ranges through adjusted means and cut-off analysis [17].

Recent research has shifted from purely statistical filters to hybrid machine learning approaches. For instance, [18] explored the use of structural parameters and data preprocessing with algorithms like SVM and Random Forest to estimate contract values in public construction tenders, achieving high predictive accuracy.

A scientometric analysis of construction bidding between 1975 and 2022 highlights that the relationship between the number of bidders and project bid prices remains a critical research area, with studies increasingly utilizing fuzzy linguistic and quantitative decision-making methodologies [19].

Despite advancements in digital procurement systems and statistical analysis techniques, there remains a lack of integrated frameworks that combine normalization, outlier detection, and cut-off determination within a unified algorithm tailored for tender evaluation [20].

Existing approaches either focus on digital platforms without sufficient analytical depth or apply statistical techniques without structured implementation in procurement workflows [21]. This gap highlights the need for a comprehensive statistical filtering algorithm that is both methodologically robust and practically applicable.

The proposed J-Selaras Statistical Filtering Algorithm (JSFA) addresses this gap by integrating statistical normalization, Z-score-based filtering, and cut-off determination into a cohesive framework for tender price evaluation.

III. Methodology

The J-Selaras Statistical Filtering Algorithm (JSFA) is developed based on the JKR case study as a structured analytical framework for evaluating tender prices in construction procurement. All simulated datasets are applied based on the real case study but have had variables changed due to confidentiality and privacy. The algorithm is designed to systematically identify abnormal bids and derive a statistically justified cut-off value that supports objective decision-making.

The proposed method addresses key limitations of traditional tender evaluation approaches, particularly their sensitivity to extreme values and lack of standardized analytical procedures. By integrating normalization, statistical parameter estimation, and outlier detection within a unified pipeline, the algorithm ensures consistency and reproducibility in tender price assessment.

The JSFA operates through six sequential stages, as briefly illustrated in Fig. 1.

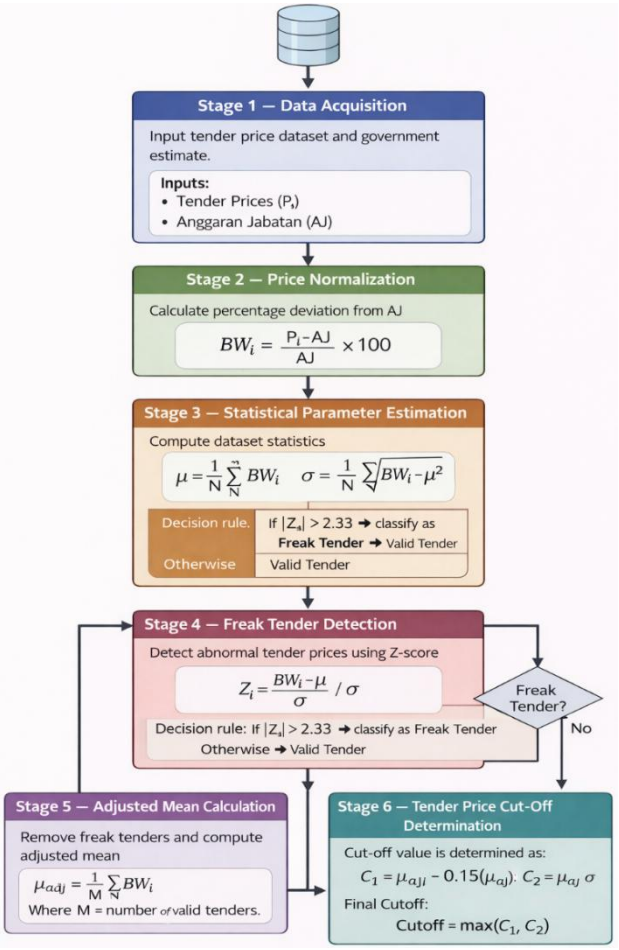


Fig. 1. Six stages in the J-Selaras statistical filtering algorithm.

A. Tender Data Acquisition

The initial stage of the J-Selaras Statistical Filtering Algorithm (JSFA) involves the systematic acquisition of tender-related data required for analysis. This stage ensures that all relevant pricing inputs are collected and structured before statistical processing.

The primary inputs consist of:

Tender price submissions P_i provided by participating contractors and Anggaran Jabatan (*AJ*), representing the government's internal cost estimate.

Formally, the dataset can be defined as:

$$D = \{(P_1, AJ), (P_2, AJ), \dots, (P_N, AJ)\} \quad (1)$$

where, each P_i corresponds to an individual contractor's bid associated with the same project-specific *AJ*.

In the context of the J-Selaras framework, this data is typically obtained from digital procurement workflows, where tender submissions are recorded in structured formats such as spreadsheets or databases. The acquisition process includes extracting tender prices from submitted Bills of Quantities (BQ), verifying consistency of units and pricing formats, and associating all bids with the corresponding *AJ* value.

This stage is critical because the reliability of subsequent statistical analysis depends on the integrity and completeness of the input dataset. Any inconsistencies, missing values, or formatting errors at this stage may propagate through the algorithm and affect the accuracy of the final cut-off determination.

By standardizing the data acquisition process, the JSFA ensures that all tender prices are evaluated within a consistent analytical framework, forming a robust foundation for normalization and outlier detection in subsequent stages.

B. Tender Price Normalization

Following data acquisition, the next stage involves transforming raw tender prices into a standardized analytical format to enable meaningful comparison across submissions. This step is necessary because absolute tender values vary significantly across projects and cannot be directly compared without normalization.

To address this, each tender price is normalized relative to the Anggaran Jabatan (*AJ*), which serves as the official benchmark cost estimate for the project.

Let the set of tender prices be defined as:

$$P = \{P_1, P_2, \dots, P_N\} \quad (2)$$

where, P_i represents the bid submitted by contractor *i*, and *AJ* is the corresponding government estimate.

Each tender price is transformed into a percentage deviation from *AJ* using:

$$BW_i = \frac{P_i - AJ}{AJ} \; x \; 100 \quad (3)$$

This normalization process converts all tender values into a relative scale, enabling direct comparison regardless of project size or cost magnitude. Positive values of BW_i indicate bids above the government estimate, while negative values indicate bids below *AJ*.

The normalization step serves two key purposes: it standardizes all tender submissions onto a unified scale by removing dependency on absolute monetary values, and it establishes a consistent reference framework anchored to *AJ* for subsequent statistical analysis.

By anchoring all observations to a common benchmark, this transformation ensures that later stages of the JSFA, particularly statistical parameter estimation and outlier detection, are performed on a coherent and comparable dataset.

C. Statistical Parameter Estimation

Following normalization, the statistical properties of the dataset are computed to characterize the overall distribution of tender prices.

The mean (μ) and standard deviation (σ) are calculated as:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} BW_i \quad (4)$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(BW_i - \mu)^2} \quad (5)$$

The mean represents the central tendency of the tender price deviations, while the standard deviation quantifies the dispersion of the dataset. These parameters form the basis for identifying abnormal pricing behavior, as they define the expected range of variation among tender submissions.

D. Freak Tender Detection

To identify abnormal bids, each normalized value is converted into a Z-score:

$$Z_i = \frac{BW_i - \mu}{\sigma} \quad (6)$$

The Z-score measures how many standard deviations a given tender price deviates from the mean of the distribution.

A tender is classified as a *freak tender* if:

$$| Z_i | > 2.33 \quad (7)$$

This threshold corresponds to a 99% confidence interval under the assumption of approximate normality, indicating that such values are statistically rare and likely to represent anomalies.

All tenders identified as freak values are excluded from further analysis. This filtering step reduces the influence of extreme bids, ensuring that subsequent calculations are based on representative data.

E. Adjusted Mean Calculation

After removing abnormal tenders, the dataset is recomputed to obtain a more stable estimate of central tendency:

$$\mu_{adj} = \frac{1}{M} \sum BW_i \quad (8)$$

where, M represents the number of remaining valid tenders after filtering.

The adjusted mean provides a refined estimate of realistic market pricing by excluding statistically inconsistent bids. This value serves as a key reference point for determining the acceptable pricing range in the final evaluation stage.

F. Cut-Off Price Determination

The final stage of the algorithm determines a statistically justified cut-off value for tender evaluation. This cut-off represents the lowest acceptable threshold below which bids may be considered unrealistic or financially unsustainable.

Two candidate thresholds are computed:

Proportional reduction from adjusted mean:

$$C1 = \mu_{adj} - 0.15\mu_{adj} \quad (9)$$

Dispersion-based threshold:

$$C2 = \mu_{adj} - \sigma \quad (10)$$

The final cut-off value is defined as:

$$C = max(C1, C2) \quad (11)$$

This formulation ensures that the cut-off reflects both the central tendency and the variability of the dataset. The use of the maximum function prevents overly conservative filtering by selecting the more permissive threshold.

The 15% reduction factor applied in the C1 threshold computation is established as a domain-specific empirical parameter aligned with tender evaluation practices in JKR Malaysia. It serves as a practical control mechanism for detecting unusually low tender prices that may arise from unrealistic cost estimation, omission of critical work components, or potential execution risks. Although the factor is not mathematically derived, its selection is justified through industry-based procurement practices and validated using historical tender data from the JKR Malaysia case study. The results obtained demonstrate its applicability as a screening threshold for improving price rationalization and reducing subjective judgment during tender evaluation.

IV. Algorithm Implementation

The implementation formalizes the sequence of operations required to process tender price data, perform statistical normalization, identify abnormal bids, and determine a robust cut-off value for procurement evaluation.

The algorithm is designed to be computationally efficient and suitable for integration into digital procurement systems such as J-Selaras. By structuring the workflow into clearly defined steps, the implementation ensures reproducibility, consistency, and scalability when applied to large tender datasets.

The detailed pseudocode of the JSFA is presented in Algorithm 1.

A. Pseudocode

Algorithm 1: J-Selaras Statistical Filtering Algorithm (JSFA)

Input: Tender prices P, Anggaran Jabatan (*AJ*)

```
Begin
  For each tender price Pi, do
    Compute deviation BW_i
  End For

  Compute the mean μ
  Compute the standard deviation σ

  For each BW_i do
    Compute Z_i
    If |Z_i| > 2.33 then
      Mark as Freak Tender
      Remove from dataset
    End If
  End For

  Compute adjusted mean μadj

  Compute:
    C1 = μ_adj − 0.15μ_adj
    C2 = μ_adj − σ

  Cutoff = max(C1, C2)

Return Cutoff
End
```

B. Computational Complexity

The algorithm operates in linear time:

- Mean and standard deviation: O(N)
- Z-score computation: O(N)
- Filtering: O(N)
- Overall complexity: O(N)

This linear complexity ensures that the JSFA algorithm remains computationally efficient and scalable, making it suitable for large-scale procurement datasets where rapid and consistent evaluation is required.

V. Analysis and Discussion

To evaluate the effectiveness of the proposed J-Selaras Statistical Filtering Algorithm (JSFA), a simulated tender dataset was constructed to represent realistic bidding behavior observed in construction procurement environments. The dataset models contractor-submitted prices relative to the *Anggaran Jabatan* (*AJ*), which serves as the government's benchmark cost estimate. This controlled experimental setup enables a systematic assessment of the algorithm's capability to detect abnormal bids and establish statistically justified cut-off thresholds under realistic procurement conditions.

A. Distribution of Tender Prices

The dataset is designed to capture both typical pricing behavior and extreme bidding scenarios, including underpricing and overpricing strategies commonly observed in competitive tendering processes. Such variability reflects real-world procurement conditions where contractors adopt diverse pricing strategies influenced by risk perception, competition intensity, and strategic positioning.

The distribution of raw tender prices is illustrated in Fig. 2. The histogram demonstrates that most bids are concentrated within a central range, while several extreme values exist at both the lower and upper tails. These outliers correspond to significantly underpriced and excessively high bids, respectively. This pattern introduces observable skewness in the dataset and highlights limitations in conventional evaluation approaches that rely on raw price comparison.

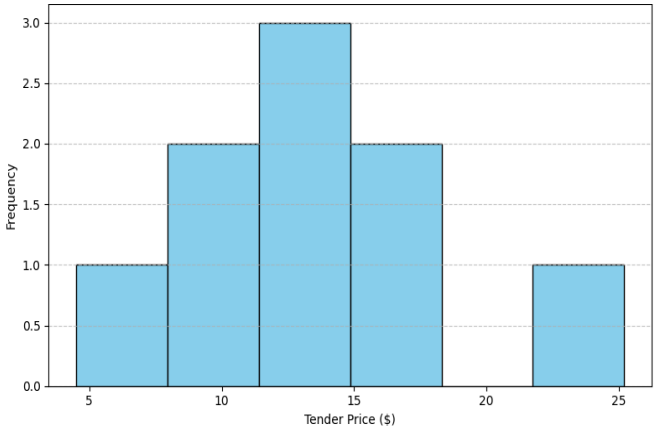


Fig. 2. Distribution of raw tender prices before normalization and statistical filtering.

To standardize bid evaluation, each tender price (P_i) is first transformed into a normalized deviation from *AJ* using the following formulation:

$$BW_i = \frac{Pi-AJ}{AJ}\ x\ 100 \quad (3)$$

This transformation ensures that all bids are evaluated on a relative percentage scale, allowing meaningful comparison across different project magnitudes. Subsequently, the statistical characteristics of the dataset are defined using the mean (μ) and standard deviation (σ) of the *BW* distribution. Each observation is then standardized using the Z-score transformation:

$$Z_i = \frac{BW_i-\mu}{\sigma} \quad (6)$$

This standardization enables the identification of statistically significant deviations from the central bidding behavior, forming the core mechanism of anomaly detection in JSFA.

To further illustrate this transformation, the distribution of standardized values is presented in Fig. 3. The Z-score distribution provides a clearer representation of how each tender deviates from the mean in units of standard deviation.

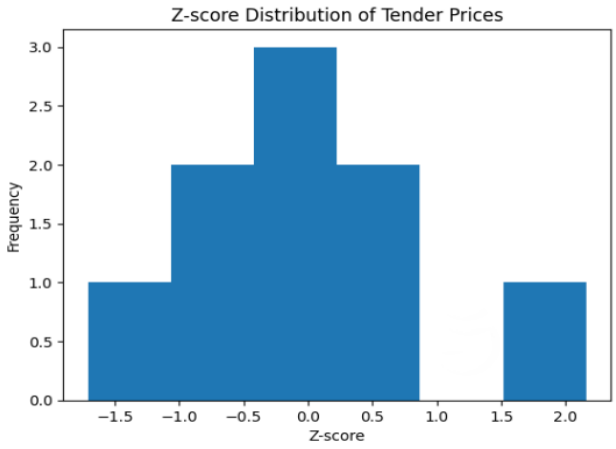


Fig. 3. Z-score distribution of normalized tender prices.

As shown in Fig. 3, the majority of tender submissions are clustered within the range of $-2<Z<2$, indicating statistically normal bidding behavior. In contrast, extreme observations fall beyond the threshold of $|Z|>2.33$, clearly distinguishing abnormal bids from the main distribution. This visual separation confirms the effectiveness of Z-score standardization in identifying anomalous pricing patterns.

The complete structure of the simulated dataset is presented in Table I. Each record includes the tender identifier, submitted price (P_i), percentage deviation from *AJ (BW)*, computed Z-score, and final classification status (Normal or Freak). This structured representation ensures that both raw and statistical features of bidding behavior are captured within a unified analytical framework.

From a methodological perspective, tender price evaluation requires careful consideration of data distribution characteristics because extreme values at both ends of the pricing spectrum may introduce significant distortion into statistical analysis. In construction procurement datasets, abnormal bids may occur

due to strategic underpricing, overestimation, incomplete cost calculation, or errors during tender preparation. If raw tender prices are directly used for evaluation, these extreme values may disproportionately influence statistical parameters such as the mean and standard deviation, resulting in unreliable benchmarks.

TABLE I. SIMULATED TENDER DATASET STRUCTURE

Tender	Price ((P_i))	BW (%)	Z-score	Status
T1	12.00	0.00	-0.66	Normal
T2	4.50	-62.50	-4.43	Freak
T3	10.20	-15.00	-1.56	Normal
T4	11.20	-6.67	-1.06	Normal
T5	12.20	1.67	-0.56	Normal
T6	14.30	19.17	0.50	Normal
T7	16.20	35.00	1.45	Normal
T8	17.00	41.67	1.85	Normal
T9	25.20	110.00	5.97	Freak

Prior to applying the Z-score-based filtering mechanism, statistical diagnostics were conducted to examine the distribution characteristics of the normalized tender deviations. The assessment included distribution analysis using skewness, kurtosis, and graphical evaluation through histogram and Q–Q plot analysis. The results indicated that although the tender deviation values exhibited some degree of asymmetry due to the presence of extreme bids, the normalized dataset demonstrated a sufficiently stable distribution pattern after abnormal values were identified. Therefore, the Z-score method was considered appropriate as an initial statistical screening mechanism for detecting significant deviations from the expected tender pricing behavior.

However, it should be noted that the proposed JSFA framework does not assume perfect normality of tender distributions. Instead, Z-score analysis is employed as a transparent and interpretable statistical indicator to quantify the relative distance of each tender submission from the overall pricing pattern. This approach provides practical advantages in government procurement environments where explainability and reproducibility are essential.

B. Outlier Pricing and Filtering Mechanism

The JSFA algorithm was applied to the JKR Malaysia tender dataset to evaluate its capability in identifying abnormal tender prices and improving dataset reliability. The algorithm calculates the standardized deviation of each tender submission using the Z-score criterion:

$$| Z |> 2.33$$

where, tender values exceeding this threshold are classified as statistically abnormal. Based on this criterion, Tender T2, which represents a significantly underpriced submission, and Tender T9, which represents an excessively high-priced submission, were identified as abnormal (*freak tenders*). These values demonstrated substantial deviation from the dominant pricing pattern and were considered potentially unreliable representations of actual project costs.

The removal of abnormal bids improves the statistical stability of the dataset by reducing the influence of extreme observations on central tendency and dispersion measurements. Without filtering, abnormal submissions may inflate the mean value and standard deviation, resulting in biased estimation of realistic tender conditions. Therefore, the filtering process enables subsequent calculations to be performed using tender values that better represent competitive market pricing.

From a procurement perspective, this mechanism provides an objective approach for identifying potentially risky submissions while maintaining transparency in decision-making. Unlike manual evaluation methods that depend heavily on individual judgment, JSFA applies consistent statistical criteria that can be reproduced across different tender evaluation scenarios.

C. Statistical Validation and Cut-Off Determination

To quantify the impact of the filtering process, the statistical parameters of the dataset were computed before and after the removal of abnormal tenders.

Initially, the mean and standard deviation of the normalized deviations (BW_i) were calculated as:

$$\mu = 13.71, \sigma = 44.63$$

These values indicate a high level of dispersion within the dataset, primarily driven by the presence of extreme tender values. Such variability reflects instability in pricing behavior and highlights the limitations of direct mean-based evaluation without prior filtering.

After excluding the identified abnormal tenders (T2 and T9), the adjusted mean was recalculated using the filtered dataset:

$$\mu_{adj} = 10.83$$

The reduction in the mean value demonstrates that extreme bids significantly inflate the overall distribution, reinforcing the necessity of statistical filtering before decision-making. The adjusted mean therefore provides a more reliable estimate of realistic market pricing.

Subsequently, two cut-off thresholds were determined. The first threshold applies a proportional reduction to the adjusted mean:

$$C1 = \mu_{adj} - 0.15\mu_{adj} = 9.21$$

The second threshold incorporates the dispersion of the dataset:

$$C2 = \mu_{adj} - \sigma = -33.80$$

The final cut-off value is defined as:

$$C = max(C1, C2) = 9.21$$

The results demonstrate that the JSFA algorithm is capable of deriving a stable and statistically grounded cut-off value while effectively mitigating the influence of extreme tender prices. The incorporation of normalization, Z-score-based filtering, and a dual-threshold mechanism ensures that evaluation outcomes reflect both central tendency and variability without excluding valid competitive bids.

The initial results indicated relatively high variability due to the presence of extreme tender values. This finding demonstrates that direct evaluation without prior filtering may produce unstable benchmarks because abnormal submissions significantly affect the overall statistical characteristics.

After removing the identified abnormal tenders (T2 and T9), the adjusted mean and standard deviation were recalculated using the filtered dataset. The reduction in dispersion confirms that extreme bids were the primary contributors to distribution instability, while the remaining submissions represented a more consistent tender pricing pattern.

This validation confirms that the proposed approach provides an objective, consistent, and reproducible framework for tender price evaluation by replacing subjective judgment with transparent statistical criteria. In addition, the algorithm enhances robustness by ensuring that decisions are based on realistic pricing patterns rather than being distorted by outliers.

These characteristics are particularly significant in the context of digital procurement systems. When integrated into platforms such as J-Selaras, the JSFA algorithm can support automated tender evaluation, improve processing efficiency, and enable transparent, data-driven decision-making in large-scale government projects.

VI. Conclusion

This study presents the J-Selaras Statistical Filtering Algorithm (JSFA) as a structured and data-driven approach for improving tender price evaluation in construction procurement. The proposed method integrates normalization relative to Anggaran Jabatan (AJ), Z-score-based outlier detection, and a dual-threshold cut-off mechanism to address limitations associated with traditional evaluation practices.

The validation results, based on a simulated tender dataset, demonstrate that the algorithm effectively identifies abnormal bids and stabilizes the dataset before decision-making. Specifically, the presence of extreme values resulted in a high dispersion (σ=44.63) and an inflated mean (μ=13.71), highlighting the distortion caused by outliers. After applying the filtering process, the adjusted mean was reduced to $\mu_{adj} = 10.83$, providing a more reliable representation of realistic market pricing.

The final cut-off value (C=9.21) illustrates the effectiveness of the dual-threshold mechanism in balancing central tendency and variability. This ensures that unrealistic bids are excluded while preserving competitive and valid tender submissions. The results confirm that the JSFA algorithm offers a consistent, objective, and reproducible framework for tender price evaluation.

From a practical perspective, the proposed approach enhances transparency, reduces reliance on subjective judgment, and supports data-driven procurement decision-making. When integrated into digital systems such as J-Selaras, the algorithm has the potential to significantly improve processing efficiency and standardization in large-scale public procurement environments.

The results demonstrate that JSFA maintains stable performance under different abnormal bid proportions because the filtering mechanism is based on relative statistical deviation rather than fixed price boundaries. When the proportion of abnormal bids increases, the algorithm continues to isolate extreme values before threshold computation, reducing the possibility of distorted evaluation results.

Nevertheless, performance may decrease when abnormal bids dominate the dataset or when legitimate market fluctuations are incorrectly interpreted as anomalies. Therefore, JSFA is most effective when sufficient representative tender samples are available.

Future work will focus on validating the algorithm using real-world procurement datasets, exploring robustness under non-normal distributions, and integrating machine learning techniques for adaptive threshold optimization.

Acknowledgment

The authors would like to express their sincere appreciation to all individuals and institutions whose support and contributions made this research possible. Special acknowledgment is extended to the Centre of Excellence for Engineering and Technology (CREaTE), Public Works Department Malaysia (JKR Malaysia), and the Quantity Surveying and Contract Branch, Public Works Department Malaysia, for their valuable expertise, technical insights, and contributions to the conceptualization, planning, and implementation of this study. The authors are also grateful to Assoc. Prof. Td. Dr. Mustafa Man from Universiti Malaysia Terengganu for his support in research collaboration, knowledge sharing, and project execution. Furthermore, the authors would like to acknowledge Dr. Wan Aezwani Wan Abu Bakar and Dr. Mohd. Kamir Yusof, from Universiti Sultan Zainal Abidin, particularly the Faculty of Informatics and Computing, for their contributions to system design, software development, technical documentation, and validation activities throughout the project. The collaborative efforts among all participating organizations have significantly enhanced the quality and impact of this work. Generative Artificial Intelligence (AI) tools were utilized solely to assist with language editing, grammar improvement, and manuscript refinement. The authors retained full responsibility for the scientific content, analysis, interpretation of results, and final approval of the manuscript. All AI-assisted outputs were carefully reviewed, verified, and revised by the authors to ensure accuracy, integrity, and compliance with publication standards.

References

[1] Niewerth, S., Vogt, P., & Thewes, M. Tender evaluation through efficiency analysis for public construction contracts. *Front. Eng. Manag.* **9**, 148–158 (2022). https://doi.org/10.1007/s42524-020-0119-z

[2] Wibowo, A., & Christopher Leonaldo, G. (2024). Abnormally Low Bid Criterion pada Pengadaan Jasa Konstruksi Proyek Pemerintah dengan Sistem Evaluasi Harga Terendah. *Jurnal Teknik Sipil*, *30*(2), 245-254. https://doi.org/10.5614/jts.2023.30.2.12

[3] Liu, X., Chen, S., Ding, Z., & Xu, B. (2023). Bidding Evaluation and Contractor Selection Using Balance Index Model and Comprehensive Input Efficiency Based on Data Envelopment Analysis. *Systems*, *11*(5), 245. https://doi.org/10.3390/systems11050245

[4] Gómez-Cabrera, A., Cortés, S., Rojas, J., Sánchez, O., & Torres, A. (2025). Data-Driven Analysis of Contracting Process Impact on Schedule and Cost Performance in Road Infrastructure Projects in

Colombia. *Buildings*, *15*(20), 3739. https://doi.org/10.3390/buildings15203739

[5] Maqfirah, A., Oktaviani, C. Z., Fachrurrazi, Mubarak, & Bulba, A. T. (2025, June). Factors influencing contractors' decisions in determining bid prices. In *IOP Conference Series: Earth and Environmental Science* (Vol. 1510, No. 1, p. 012068). IOP Publishing. https://doi.org/10.1088/1755-1315/1510/1/012068

[6] Leśniak, A. (2021). Statistical Methods in Bidding Decision Support for Construction Companies. *Applied Sciences*, *11*(13), 5973. https://doi.org/10.3390/app11135973

[7] Schneider dos Santos, E., Machado dos Santos, M., Castro, M. *et al*. Detection of fraud in public procurement using data-driven methods: a systematic mapping study. *EPJ Data Sci.* **14**, 52 (2025). https://doi.org/10.1140/epjds/s13688-025-00569-3

[8] Lee, H., & Yun, S. (2024). Strategies for Imputing Missing Values and Removing Outliers in the Dataset for Machine Learning-Based Construction Cost Prediction. *Buildings*, *14*(4), 933. https://doi.org/10.3390/buildings14040933

[9] Egwim, C. N., Alaka, H., Demir, E., Balogun, H., Olu-Ajayi, R., Sulaimon, I., Wusu, G., Yusuf, W., & Muideen, A. A. (2024). Artificial Intelligence in the Construction Industry: A Systematic Review of the Entire Construction Value Chain Lifecycle. *Energies*, *17*(1), 182. https://doi.org/10.3390/en17010182

[10] Matos, B. C., Cruz, C. O., & Branco, F. B. (2024). Digitalization and procurement in construction projects: An integrated BIM-based approach. *Journal of Information Technology in Construction*, *29*. https://doi.org/10.36680/j.itcon.2024.019

[11] Koke, B., & Moehler, R. C. (2019). Earned Green Value management for project management: A systematic review. *Journal of Cleaner Production*, *230*, 180-197. https://doi.org/10.1016/j.jclepro.2019.05.079

[12] Naji, K. K., Gunduz, M., & Falamarzi, M. H. (2022). Assessment of construction project contractor selection success factors considering their interconnections. *KSCE Journal of Civil Engineering*, *26*(9), 3677-3690. https://doi.org/10.1007/s12205-022-1377-6

[13] Foorthuis, R. On the nature and types of anomalies: a review of deviations in data. *Int J Data Sci Anal* 12, 297–331 (2021). https://doi.org/10.1007/s41060-021-00265-1

[14] Dastjerdy, B., Saeidi, A., & Heidarzadeh, S. (2023). Review of Applicable Outlier Detection Methods to Treat Geomechanical Data. *Geotechnics*, *3*(2), 375-396. https://doi.org/10.3390/geotechnics3020022

[15] Saari, M. N., Arshad, M. A., & Mt Amin, N. Z. (2026). Optimising construction contract rates with Z-score and machine learning approaches. *International Journal of Research and Innovation in Social Science, 10*(1), 3682–3695. https://doi.org/10.47772/IJRISS.2026.10100289

[16] Jeong, J., Lee, H., & Kim, J.J. An econometric identification of abnormally low bids in the procurement market: discriminant analysis. *J. Ind. Bus. Econ.* 51, 211–234 (2024). https://doi.org/10.1007/s40812-023-00257-1

[17] Norpah, M., Arshad, I., & Mat, N. (2023). J-SELARAS: Development of rationalisation analysis model. Journal of Theoretical and Applied Information Technology, 101(19), 6164-6175. http://www.jatit.org/volumes/Vol101No19/30Vol101No19.pdf

[18] Aslay, S. E. (2024). Estimating contract value using structural parameters: A machine learning approach with data preprocessing. *Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi*, *15*(3), 755-765. https://doi.org/10.24012/dumf.1515160

[19] Rashidi, A., Tamošaitienė, J., Ravanshadnia, M., & Sarvari, H. (2023). A scientometric analysis of construction bidding research activities. Buildings, 13(1), 220. https://doi.org/10.3390/buildings13010220

[20] Kwasafo, O. K., Laryea, S., & Govender, N. (2025, February). Developing an Artificial Intelligence-Based Bid Automation Model for Construction Consulting Firms. In *Construction Industry Development Board Postgraduate Research Conference* (pp. 28-42). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-99204-9_3

[21] Key Competencies for Effective Project Management in the Context of Digitalization in India (Sherly Christilda D & Dr. Kranti Kumar Myneni, Trans.). (2026). *International Journal of Management and Humanities (IJMH)*, *12*(6), 11-14. https://doi.org/10.35940/ijmh.G1862.12060226

Prompt Literacy and Generative AI Readiness for AI-Assisted Animation Creative Workflows in Higher Education

Loh Ngiik Hoon*, Natasha Binti Rusdy Wong, Marzie Hatef Jalil
Faculty of Applied and Creative Arts, Universiti Malaysia Sarawak (UNIMAS), 94300 Kota Samarahan, Sarawak, Malaysia

***Abstract*—Generative AI technologies are increasingly being adopted within creative learning environments, particularly in animation and digital media education. As AI-assisted tools become more accessible, there is a growing need to understand how students perceive and adapt to these technologies in their creative workflows. This study explores prompt literacy and generative AI readiness among animation students in higher education creative workflows. A quantitative survey was conducted involving 74 animation students to examine AI usage patterns, perceived benefits, workflow integration, and challenges related to generative AI adoption. The collected data were analyzed using descriptive statistics and reliability analysis through Jamovi software. The findings indicate that students generally viewed generative AI positively, especially in supporting idea development, creative experimentation, and improving workflow efficiency. At the same time, concerns related to originality, prompt construction, ethical issues, and output consistency were frequently highlighted by respondents. The instrument demonstrated excellent reliability with a Cronbach's alpha value of 0.943. The study suggests that prompt literacy and AI readiness are becoming increasingly relevant within animation education and creative digital learning environments. The findings may provide useful insights for educators and curriculum developers in integrating generative AI tools into animation-related teaching and learning practices.**

***Keywords*—Generative AI; prompt literacy; AI readiness; animation education; creative workflows; higher education**

I. Introduction

The emergence of generative artificial intelligence (AI) has begun to reshape the way creative work is approached within digital media and higher education environments. Over the past few years, AI-assisted tools such as ChatGPT, Gemini, Midjourney, Runway, and Sora have become increasingly common among students and digital creators. These technologies are now capable of supporting various creative processes including idea development, visual generation, scripting, and content experimentation [1], [6], [13]. In creative disciplines such as animation, the use of AI is gradually moving beyond technical automation and becoming part of everyday creative workflows [4]. Moreover, recent studies have also shown that generative AI technologies are increasingly influencing visual storytelling, concept development, and digital content production practices within creative media education environments.

Within higher education, generative AI has attracted growing attention due to its potential to support teaching, learning, and creative production activities. Students in animation and digital creative programs are increasingly exposed to AI-assisted tools during concept development, storyboarding, visual ideation, and content planning stages. The accessibility of these tools has also changed how students approach creative problem-solving and production efficiency. For some students, AI tools provide faster ways to generate references and explore ideas, while others view them as useful support systems during technical or creative challenges [2], [3]. The growing accessibility of AI-assisted systems has further encouraged experimentation-based learning approaches among students within digital creative disciplines, particularly in areas involving rapid ideation and multimedia production [8], [16].

Despite these advantages, the integration of generative AI into creative education continues to raise several concerns. Questions related to originality, creative dependency, ethical usage, and the role of human creativity are becoming more visible within AI-assisted production environments. In addition, the ability to communicate effectively with AI systems has emerged as an increasingly important skill. Students are now expected not only to use AI tools, but also to understand how prompts influence the quality and relevance of generated outputs. In this study, prompt literacy refers to students' ability to construct and use prompts effectively to obtain meaningful and relevant outputs from generative AI systems. It also involves understanding how prompt wording, context, and specificity influence AI-generated responses during creative tasks. This growing reliance on prompt-based interaction has contributed to increasing attention towards prompt literacy, particularly within creative learning environments where communication, visual interpretation, and experimentation play important roles [5], [14], [15].

Existing studies on AI integration in education have mainly focused on general technology acceptance, academic performance, and AI-assisted learning practices. However, research specifically related to animation students and creative production workflows remains relatively limited. There is still insufficient understanding of how students in animation-related programs perceive generative AI readiness within creative workflows, particularly in relation to prompt literacy, creative support, and workflow integration. As generative AI technologies continue to evolve rapidly, understanding students' readiness and perceptions may help institutions better

*Corresponding author

prepare future creative professionals for emerging digital production environments [12].

Therefore, this study provides an exploratory examination of animation students' perceptions of prompt literacy and generative AI readiness within AI-assisted creative workflows. Rather than testing predictive relationships among the study variables, the research focuses on understanding students' experiences, perceived benefits, workflow support, and challenges associated with the integration of generative AI in animation education. The findings are expected to provide baseline insights that may assist educators and curriculum developers in designing effective strategies for integrating generative AI into creative learning environments, while also serving as a foundation for future studies employing more advanced analytical approaches.

Animation education provides a distinctive context for examining prompt literacy and generative AI readiness because students engage in visually driven and iterative creative processes that require continuous interaction between human creativity and AI-generated outputs. Unlike many text-based learning tasks, animation production involves concept development, visual storytelling, character design, scene composition, and creative refinement, all of which rely heavily on effective prompt construction and critical evaluation of AI-generated content. As a result, animation students require not only technical proficiency but also the ability to guide, assess, and creatively adapt AI outputs throughout the production process. Understanding these experiences offers valuable insights into how generative AI can be integrated into creative education while maintaining the central role of human creativity.

Fig. 1. Original 3D animation scene developed through traditional production workflow.

Note: The original 3D animation scenes were produced by the authors prior to AI-assisted refinement.

Fig. 2. AI-assisted visual refinement for animation scene enhancement.

Note: AI-assisted refinement was conducted using prompt-based generative AI tools while preserving the original narrative direction and scene composition developed by the authors.

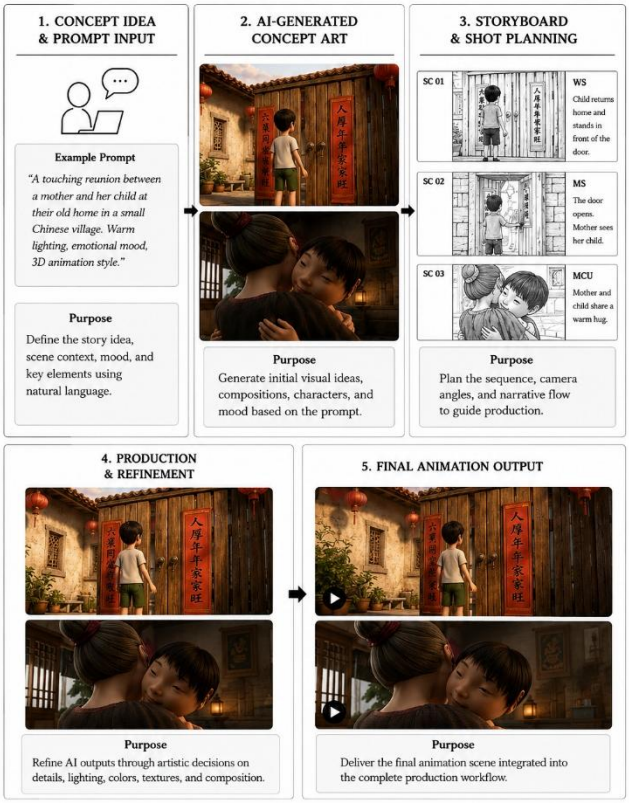


Fig. 3. Example of AI-assisted animation concept development workflow.

Recent developments in generative AI technologies have introduced new possibilities for animation production workflows, particularly during early-stage ideation, visual refinement, and storyboard planning processes. Fig. 1 and Fig. 2 demonstrate examples of animation scenes before and after AI-assisted refinement. The comparison illustrates how generative AI tools may support improvements in visual quality, environmental detail, lighting consistency, and emotional atmosphere within animation-related creative workflows.

Fig. 1 presents examples of original 3D animation scenes developed through conventional production approaches prior to AI-assisted enhancement. In contrast, Fig. 2 demonstrates the same scenes after undergoing AI-assisted visual refinement using prompt-based generative AI support. The refined outputs exhibit improved environmental textures, cinematic lighting, visual depth, and overall scene composition while preserving the original narrative direction and emotional context of the animation scenes.

Fig. 3 further illustrates an example of an AI-assisted animation concept development workflow adapted from contemporary creative production practices. The workflow demonstrates how generative AI technologies may support different stages of animation development, including prompt construction, AI-generated concept generation, storyboard planning, production refinement, and final animation output. Despite the increasing role of AI-assisted systems, human creativity, artistic interpretation, and production decision making remain essential throughout the refinement and finalization stages of animation production workflows.

Overall, the integration of generative AI technologies within animation education environments may provide students with additional opportunities for creative experimentation, rapid ideation, and workflow support. At the same time, the workflow also reflects the continuing importance of human-centred creative direction in maintaining originality, narrative consistency, and artistic quality within AI-assisted animation production practices.

II. Literature Review

A. Generative AI in Higher Education and Creative Learning

The development of generative artificial intelligence (AI) technologies has introduced significant changes to teaching, learning, and creative production practices within higher education environments [2], [3], [13]. In recent years, AI-assisted systems such as ChatGPT, Gemini, Midjourney, and Runway have become increasingly accessible to students across various academic disciplines. These tools can support content generation, idea exploration, visual development, and communication processes, which lead many educators and institutions to explore their potential within educational settings [1]. The growing accessibility of generative AI has also contributed to wider discussions regarding the future role of AI within digital learning environments and creative education.

Within higher education, previous studies have reported that AI-assisted technologies may support productivity, improve learning flexibility, and encourage exploratory learning approaches among students [2], [8]. Students are increasingly using generative AI tools for brainstorming, summarizing information, generating references, and supporting project development processes. On the other hand, in creative disciplines, the integration of AI technologies has become particularly noticeable due to the increasing demand for digital content production and rapid visual experimentation. As a result, AI-assisted creative workflows are gradually becoming part of the learning experience within animation, design, and digital media programs [4]. Apart from that, previous studies have suggested that AI-assisted creative environments may encourage more iterative and exploratory learning experiences, particularly among students involved in design-oriented and visual production activities [9].

In animation education, generative AI tools are beginning to influence multiple stages of creative production workflows, including concept development, scriptwriting, storyboard planning, visual ideation, and asset generation. AI-assisted systems may help students accelerate early-stage ideation and reduce technical barriers during content development processes [4]. For some students, generative AI tools also provide opportunities to explore different visual directions and creative outcomes more efficiently compared to traditional production methods. This increasing interaction between students and AI systems has contributed to the emergence of more collaborative human–AI creative practices within digital creative learning environments. Hence, human–AI collaboration has increasingly been discussed as part of emerging co-creative production models where AI functions as a supportive ideation partner rather than a replacement for human creativity [7].

Despite the potential advantages of generative AI integration, concerns regarding ethical usage, originality, and creative dependency continue to be widely discussed within higher education contexts. Several researchers have highlighted that excessive reliance on AI-generated outputs may affect critical thinking, independent creativity, and content authenticity among students [3], [10]. In creative disciplines such as animation, these concerns become particularly relevant because originality and artistic interpretation remain central elements of creative practice. Consequently, educators are increasingly encouraged to develop balanced learning approaches that integrate AI technologies while still preserving human creativity and creative decision-making processes.

B. Prompt Literacy and Generative AI Readiness in Creative Workflows

As generative AI technologies become more prompt-driven, the ability to communicate effectively with AI systems has emerged as an increasingly important digital skill. In recent discussions surrounding AI-assisted learning, the concept of prompt literacy has gained attention as part of emerging digital and AI-related competencies. Prompt literacy refers to an individual's ability to understand, construct, refine, and evaluate prompts when interacting with generative AI systems. Within animation and creative workflows, prompt literacy extends beyond basic prompt writing and includes the ability to communicate visual intentions, creative requirements, stylistic preferences, and contextual information effectively to AI-assisted systems [14], [15]. Students with stronger prompt literacy are more likely to obtain relevant outputs, refine generated content efficiently, and maintain greater creative control throughout AI-assisted production processes [5]. Within creative workflows, prompt literacy may influence how effectively students interact with AI systems during idea generation, visual experimentation, and creative problem-solving activities.

Previous studies have suggested that students often experience difficulties when attempting to generate accurate or contextually appropriate outputs from AI systems due to unclear prompting strategies or limited understanding of AI behavior [5]. In creative production environments such as animation workflows, prompt construction may directly influence the quality, consistency, and usability of generated outputs. As a result, students who possess stronger prompt communication skills may demonstrate higher confidence and adaptability when working with AI-assisted creative technologies.

Apart from that, AI readiness has become another important area of discussion within higher education environments. AI readiness is commonly associated with an individual's preparedness, confidence, and willingness to adopt AI technologies within learning or professional settings. Previous research has shown that students' readiness towards AI integration may be influenced by perceived usefulness, technical familiarity, learning support, and ethical considerations [6]. In creative disciplines, AI readiness may also involve students' ability to balance AI-assisted production with personal creativity, artistic control, and workflow decision-making.

To note, although generative AI adoption continues to expand rapidly within educational settings, research focusing specifically on animation students and creative production workflows remains relatively limited. Many existing studies have focused primarily on general AI adoption or broader educational technology acceptance models without examining the specific creative practices involved in animation-related learning environments [2], [3]. In addition, there is still limited understanding regarding how prompt literacy and AI readiness may influence students' experiences within AI-assisted creative workflows. Therefore, further investigation is needed to better understand how animation students perceive generative AI integration within higher education creative environments [4], [12].

Overall, previous studies have consistently highlighted the growing role of generative AI in higher education, with particular attention given to its potential for enhancing learning, creativity, and productivity. Existing research has also emphasized the importance of prompt literacy and AI readiness as essential competencies for effective interaction with AI systems. However, most studies have examined these topics from a general educational perspective, with limited attention to disciplines that rely heavily on iterative visual creativity and AI-assisted content development. Furthermore, relatively few studies have explored how animation students perceive the integration of generative AI within their creative workflows. By focusing on animation education, this study extends the current literature by examining prompt literacy and generative AI readiness within a creative learning environment where human creativity and AI-assisted production are closely interconnected.

III. METHODOLOGY

A. Research Design

This study adopted a quantitative survey approach to examine prompt literacy and generative AI readiness among animation students involved in creative workflows within higher education environments. The quantitative method was selected as the study aimed to obtain a broader understanding of students' experiences, perceptions, and readiness towards the use of generative AI technologies in animation related creative activities. The study focused on students' exposure to AI-assisted creative workflows, including idea development, workflow support, and learning-related experiences associated with generative AI usage.

The intention of this study was to provide an overview of how animation students perceive prompt literacy and generative AI readiness within their creative workflows. As the research was exploratory in nature, descriptive statistical analysis was considered appropriate for summarizing respondents' experiences and identifying overall response patterns. The study was not designed to examine predictive or causal relationships among the variables, but rather to establish an initial understanding that may support future empirical investigations in animation education.

B. Participants and Data Collection

Data collection was conducted using an online questionnaire distributed to students enrolled in animation and creative media-related programs at a Malaysian public university. A total of 74 valid responses were collected and used for analysis. The participants were selected because they were actively engaged in animation-related creative activities and had prior exposure to generative AI tools through their academic learning. The participants reported experience with a variety of generative AI applications used in creative learning, including text-generation, image-generation, and other AI-assisted tools that supported different stages of the animation production process. The respondents represented varying levels of familiarity and experience in using generative AI within both creative practice and coursework. Although the study was conducted within a specific educational context, the findings provide preliminary insights into students' perceptions of prompt literacy and generative AI readiness that may inform future research involving broader populations and different educational settings.

The questionnaire was shared through online communication platforms commonly used by students. Participation in the survey was voluntary, and respondents were informed that the information collected would be used solely for academic and research purposes. The online distribution approach allowed students to respond based on their own experiences and perspectives regarding AI-assisted creative workflows in animation-related learning environments.

C. Research Instrument

The questionnaire was developed based on the research objectives and relevant literature on prompt literacy, generative AI readiness, and AI-assisted creative learning. It consisted of three sections. The first section collected participants' demographic information and their experience with generative AI tools. The second section included multiple-choice questions and five-point Likert-scale items (1 = Strongly Disagree to 5 = Strongly Agree) that examined students' perceptions of generative AI readiness, workflow efficiency, idea generation, technical support, productivity, creative experimentation, learning support, and their use of prompts during AI-assisted creative activities. These items provided an indication of students' prompt literacy by examining their perceived ability to use prompts effectively when interacting with generative AI tools. The final section comprised open-ended questions that invited respondents to share their experiences, opinions, and challenges related to AI-assisted creative workflows. Responses to the open-ended questions were analyzed using thematic categorization, in which similar responses were grouped into recurring themes. These themes were used to complement and provide additional context for the quantitative findings.

D. Data Analysis

The responses collected were analyzed using Jamovi statistical software. Descriptive statistical analysis was carried out to examine respondent demographics, AI usage frequency, and students' perceptions towards generative AI integration in animation creative workflows. Mean scores and standard deviation values were calculated to identify overall response patterns across the Likert-scale items.

Reliability analysis was also conducted to evaluate the internal consistency of the survey instrument. The results showed a Cronbach's alpha value of 0.943, indicating excellent reliability among the measurement items used in the study. In addition to the quantitative analysis, responses from the open-ended questions were analyzed using thematic categorization. Similar responses were grouped into recurring themes related to idea generation, workflow efficiency, creative experimentation, prompt construction challenges, and ethical concerns associated with generative AI usage. The thematic findings were used to complement and strengthen the interpretation of the quantitative results.

IV. Results and Findings

A. Demographic Profile of Respondents

A total of 74 valid responses were collected for this study. The respondents consisted of students enrolled in animation and creative media-related programs with varying levels of experience in using generative AI tools within creative workflows. Based on the collected data, female respondents represented the majority of the participants, followed by male respondents. The respondents also represented different years of study, allowing the study to capture perspectives from students with varying academic and creative experiences.

The demographic analysis provided an overview of the respondents' educational background and AI usage exposure within animation-related creative environments. Detailed demographic information is presented in Table I.

TABLE I. Demographic Profile of Respondents

Variable	Category	Frequency (n)	Percentage (%)
Gender	Female	43	58.1
	Male	31	41.9
Age	18–20	8	10.8
	21–23	59	79.7
	24 and above	7	9.5
Year of Study	Year 1	35	47.3
	Year 2	25	33.8
	Year 3	14	18.9

A total of 74 respondents participated in the study. Female students represented the majority of respondents (58.1%), while male students accounted for 41.9%. Most respondents were between 21 and 23 years old (79.7%). In terms of academic level, Year 1 students formed the largest group of participants, followed by Year 2 and Year 3 students.

B. Generative AI Usage Among Students

The findings indicated that generative AI tools are increasingly being used by students within animation-related creative workflows. Most respondents reported using AI-assisted technologies at least occasionally during idea development, creative experimentation, and content production activities. AI tools such as ChatGPT, Gemini, and Canva AI were among the most used platforms reported by students.

The results also showed that students mainly utilized generative AI tools for brainstorming, workflow support, reference generation, and technical assistance during creative production processes. While some respondents reported frequent use of AI technologies, others indicated more moderate usage depending on project requirements and familiarity with AI-assisted systems. The overall findings suggest that generative AI has already become part of students' creative learning experiences within animation education environments.

Detailed AI usage frequency information is presented in Table II.

TABLE II. AI Usage Frequency Among Respondents

AI Usage Frequency	Frequency (n)	Percentage (%)
Rarely	21	28.4
Sometimes	34	45.9
Often	15	20.3
Very Often	4	5.4

The findings indicate that generative AI tools are already being used by students within animation-related creative workflows. Most respondents reported using AI-assisted technologies "Sometimes" (45.9%), followed by "Rarely" (28.4%) and "Often" (20.3%). Only a smaller group of respondents reported using generative AI tools "Very Often" (5.4%). The results suggest that while generative AI technologies have become increasingly accessible within creative learning environments, the frequency of usage among students still varies depending on individual familiarity, workflow needs, and project requirements.

Generative AI encompasses a wide range of tools with different functions and capabilities. For example, large language models such as ChatGPT are commonly used for text generation, brainstorming, and prompt development, whereas image generation platforms such as Midjourney and Adobe Firefly support concept art and visual ideation. More recently, AI video generation tools, including Runway Gen-3 and Pika, have been introduced to assist with animation prototyping and motion-based content creation. Although these platforms differ in their outputs and underlying technologies, they all require users to communicate effectively through prompts and critically evaluate AI-generated results. Consequently, prompt literacy remains a fundamental competency regardless of the specific AI platform used.

C. Students' Perceptions Towards Generative AI Creative Workflows

The descriptive analysis revealed generally positive perceptions towards generative AI integration within animation creative workflows. Most respondents agreed that AI-assisted tools were useful for improving workflow efficiency, supporting idea generation, and encouraging creative experimentation. Students also reported that generative AI technologies helped reduce technical difficulties and supported productivity during animation-related creative tasks.

At the same time, several respondents expressed concerns regarding originality, output consistency, and ethical issues associated with AI-generated content. Some students also highlighted difficulties in producing accurate outputs due to ineffective prompt construction and limited control over generated results. These findings suggest that while students recognize the benefits of generative AI technologies, challenges related to prompt literacy and creative dependency remain important considerations within animation education.

The overall mean score analysis for students' perceptions towards generative AI creative workflows is presented in Table III.

TABLE III. MEAN SCORE ANALYSIS OF STUDENTS' PERCEPTIONS TOWARDS GENERATIVE AI CREATIVE WORKFLOWS.

Variable	Mean	Standard Deviation
Workflow Efficiency	3.57	0.97
Faster Task Completion	3.55	1.00
Creative Idea Support	3.61	0.99
Ease of Use	3.89	0.90
Productivity Improvement	3.38	0.98
Visual Quality Support	3.12	1.27
Creative Experimentation	3.38	1.11
Technical Support Assistance	3.41	1.03
Confidence in Using AI Tools	2.89	1.08
Learning Support	3.28	1.10
Course Integration Support	2.92	1.28

The mean score analysis showed that respondents generally demonstrated moderate to positive perceptions towards generative AI integration within animation creative workflows. Among the evaluated items, ease of use recorded one of the higher mean values (M = 3.89), indicating that many students perceived generative AI tools as relatively accessible and manageable during creative activities. Creative idea support also showed a positive response pattern (M = 3.61), suggesting that students found AI-assisted systems useful during brainstorming and idea development processes.

Several workflow-related aspects, including workflow efficiency, faster task completion, technical support, and creative experimentation, also recorded moderate mean values. These findings indicate that generative AI tools are increasingly being viewed as supportive technologies within animation related creative tasks. However, lower mean values were observed for confidence in using AI tools (M = 2.89) and course integration support (M = 2.92), which may suggest that some students still experience uncertainty or require additional guidance when integrating AI technologies into formal learning and creative production environments.

D. Reliability Analysis

Reliability analysis was conducted to evaluate the internal consistency of the Likert-scale items used in the study. The reliability results are presented in Table IV. The analysis produced a Cronbach's alpha value of 0.943, indicating excellent reliability and strong internal consistency among the measurement items. The findings suggest that the survey instrument was reliable in measuring students' perceptions and experiences related to generative AI readiness and creative workflow integration within animation education contexts.

TABLE IV. RELIABILITY ANALYSIS OF SURVEY INSTRUMENT

Construct	Number of Items	Cronbach's Alpha
Generative AI Readiness and Creative Workflow Perceptions	11	0.943

E. Thematic Analysis of Open-Ended Responses

To provide additional insights into students' experiences with generative AI, responses from the open-ended questions were grouped into several recurring themes. Table V summarizes the identified themes and their frequencies. The analysis revealed that students primarily perceived generative AI as a useful tool for idea generation, workflow efficiency, and creative experimentation. At the same time, respondents also reported challenges related to prompt construction, output consistency, and ethical concerns associated with AI-generated content.

TABLE V. THEMATIC ANALYSIS OF OPEN-ENDED RESPONSES

Theme	Frequency
Idea Generation Support	28
Workflow Efficiency	21
Creative Experimentation	17
Prompt Construction Challenges	15
Ethical Concerns	12

As shown in Table V, idea generation support emerged as the most frequently reported theme (n = 28), followed by workflow efficiency (n = 21) and creative experimentation (n = 17). Prompt construction challenges (n = 15) and ethical concerns (n = 12) were also identified, indicating that while students recognized the benefits of generative AI, they remained aware of its practical and ethical limitations. The thematic findings generally support the quantitative results by providing further insights into how students experience generative AI within animation-related creative workflows. Many respondents highlighted the usefulness of AI-assisted tools for brainstorming, idea generation, visual exploration, and improving workflow efficiency during creative tasks. At the same time, students also expressed concerns regarding prompt construction, output consistency, originality, and ethical considerations when integrating AI-generated content into their

projects. These findings suggest that while students are generally receptive towards generative AI adoption, the development of prompt literacy and responsible AI usage remains important for effective integration within animation education environments.

V. Discussion

A. Generative AI and Creative Workflow Practices

The findings from this study show that generative AI tools are gradually becoming part of students' creative routines within animation-related learning environments. Many respondents appeared comfortable using AI-assisted systems during idea development, creative exploration, and workflow support activities. Rather than viewing AI purely as a replacement technology, most students seemed to approach these tools as practical assistants that could help simplify certain stages of the creative process.

One noticeable pattern from the findings was the frequent use of AI tools during brainstorming and early concept development stages. This may be related to the nature of animation production itself, where students are often required to generate multiple visual ideas and creative directions within limited production timelines. In this context, generative AI tools may provide students with faster access to references, visual experimentation, and alternative creative possibilities. Similar observations have also been discussed in recent educational technology studies where AI-assisted systems were found to support exploratory and idea-based learning approaches [2], [8], [13], [16].

The findings also suggest that students generally perceive AI-assisted workflows as useful for improving efficiency and reducing technical barriers during production tasks. This observation is consistent with previous studies, which have shown that generative AI can enhance idea generation, streamline creative processes, and support learning in higher education environments [2], [13], [16]. Within animation education, these findings indicate that AI is most valuable as a creative support tool that assists students during ideation and content development rather than replacing human creativity. By reducing the time required for repetitive tasks and early-stage concept exploration, students can devote greater attention to artistic refinement, storytelling, and creative decision-making.

At the same time, the findings do not suggest that students are fully dependent on AI technologies for creative outcomes. Many respondents still acknowledged the importance of human creativity and personal artistic direction within animation workflows. This reflects a more balanced perception towards AI integration, where students appear to value AI as a supporting creative tool rather than a complete substitute for human input and interpretation [17], [18].

B. Prompt Literacy and AI Readiness in Creative Learning

Another important aspect observed in this study relates to prompt literacy and students' readiness towards generative AI usage. Although many respondents reported positive experiences with AI-assisted creative workflows, several students also highlighted challenges related to generating accurate or satisfying outputs from AI systems, particularly when prompts lacked clarity or contextual detail [5], [12]. In many cases, the effectiveness of the generated results appeared to depend heavily on how prompts were written and refined during interaction with AI tools. This finding supports recent studies suggesting that prompt literacy is a key factor influencing the quality and relevance of AI-generated outputs, particularly in creative disciplines where users are required to communicate complex visual and contextual ideas through prompts [14], [15].

This finding suggests that prompt literacy may play an increasingly important role within AI-assisted creative environments. Students who can provide clearer instructions and more structured prompts may experience better control over AI-generated outputs during idea generation and creative experimentation processes. Within animation-related workflows, prompt construction may influence not only the quality of generated visuals or ideas, but also the overall efficiency of the creative process itself [5], [9].

The growing importance of prompt literacy has also been highlighted in recent discussions surrounding AI-assisted education and digital competency development [5], [12]. As AI systems become more integrated into creative learning environments, students may gradually require stronger communication and critical evaluation skills when interacting with AI-generated systems. Several recent discussions have also highlighted prompt engineering and AI communication skills as emerging forms of digital literacy within technology-enhanced learning environments. In animation education specifically, this may involve understanding how prompts affect visual consistency, artistic direction, and content relevance during creative production activities. As generative AI technologies continue evolving, students may increasingly require adaptive prompting strategies and critical evaluation skills to maintain creative control over AI-assisted production outcomes [12].

The findings further indicate that most students appear generally open towards generative AI integration within higher education creative workflows. Many respondents viewed AI-assisted systems as beneficial tools that could support learning, experimentation, and workflow flexibility [6]. This finding may reflect the growing normalization of AI-assisted creative technologies among younger digital native learners who are already familiar with interactive and technology-driven creative environments [8], [9]. This may reflect the increasing normalization of AI-assisted technologies within creative digital environments, particularly among younger generations of students who are already familiar with digital production tools and online creative platforms.

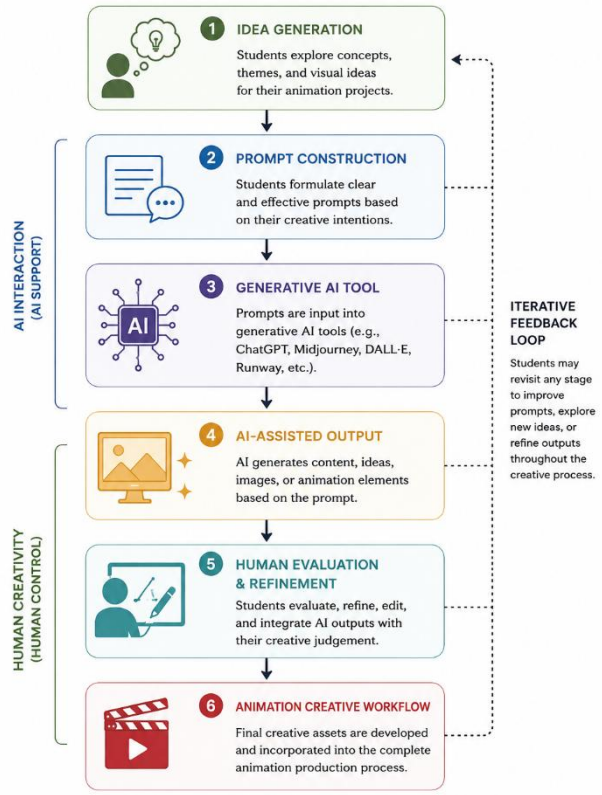


Fig. 4. Proposed human–AI creative workflow in animation education.

Next, Fig. 4 presents a proposed human–AI creative workflow derived from the overall findings and observations obtained in this study. The workflow illustrates how animation students interact with generative AI systems throughout different stages of the creative production process. The process begins with idea generation, where students explore initial concepts, themes, and visual directions for animation-related projects. At this stage, human creativity remains central in determining the creative intention and artistic direction of the workflow.

The second stage involves prompt construction, where students formulate instructions and descriptive inputs to communicate their creative intentions to generative AI systems. This stage reflects the growing importance of prompt literacy within AI-assisted creative environments, as the quality and clarity of prompts may influence the relevance and consistency of generated outputs. Students then interact with generative AI tools such as ChatGPT, Midjourney, Runway, and other AI-assisted platforms to generate visual references, creative ideas, or supporting content during the production process.

Following AI-assisted generation, students evaluate, refine, and selectively integrate the generated outputs into their own creative workflows. The findings of this study suggest that students do not rely entirely on AI-generated outputs but instead use these technologies as supportive creative tools that assist brainstorming, experimentation, and workflow efficiency. Human judgment, creative interpretation, and artistic decision-making continue to play important roles during the refinement stage of the workflow.

The iterative feedback loop illustrated in the figure further reflects the non-linear nature of AI-assisted creative production. Students may revisit earlier stages multiple times to improve prompts, refine generated outputs, or explore alternative creative directions throughout the animation development process. Overall, the proposed workflow highlights the increasingly collaborative relationship between human creativity and generative AI technologies within contemporary animation education environments.

In addition, the workflow also reflects several challenges experienced by students during AI-assisted creative processes, including prompt ambiguity, inconsistent AI-generated outputs, ethical concerns, and limited artistic control over generated content. These challenges suggest that although generative AI technologies may support creative workflows, effective AI integration within animation education still requires critical evaluation, responsible usage, and stronger prompt literacy skills among students.

C. Challenges and Concerns in AI-Assisted Creative Environments

Although respondents generally expressed positive views towards generative AI integration, several concerns were also identified throughout the study. One of the most common concerns involved originality and the authenticity of AI-generated outputs. Some students appeared uncertain about the extent to which AI-generated content should be considered part of original creative work, especially within academic and artistic contexts where personal creativity remains highly valued. Maintaining a balance between technological assistance and authentic creative expression therefore remains an important consideration within AI-assisted animation education environments [7], [12].

Concerns regarding output consistency and creative control were also mentioned by several respondents. While AI tools may assist with rapid idea generation, some students reported difficulties in obtaining outputs that accurately matched their intended artistic direction. These findings suggest that effective use of generative AI requires more than technical access to AI tools. Students also need the ability to critically evaluate AI-generated outputs, refine prompts, and balance AI assistance with their own creative judgment. This reflects the growing emphasis on human–AI collaboration, where AI supports creative production while human creativity remains central to artistic interpretation and decision-making [17], [18].

Apart from that, another issue observed from the findings relates to possible overreliance on AI-assisted systems [3], [10]. Although AI technologies may improve efficiency, excessive dependency on generated outputs could potentially reduce opportunities for independent exploration and problem solving among students. Similar concerns have also been raised in previous studies discussing AI integration within higher education and creative learning environments [3].

In sum, the findings suggest that generative AI technologies present both opportunities and challenges within animation education. While AI-assisted systems may support workflow efficiency, creative exploration, and technical assistance, institutions may also need to consider how AI technologies are

introduced responsibly within creative learning environments. This includes encouraging ethical awareness, strengthening prompt literacy, and maintaining the role of human creativity within AI-assisted animation workflows.

VI. Conclusion

This study explored prompt literacy and generative AI readiness among animation students within higher education creative workflows. The findings indicate that generative AI technologies are increasingly becoming part of students' creative practices, particularly in areas related to idea generation, workflow support, and creative experimentation. Most respondents demonstrated generally positive perceptions towards AI-assisted creative workflows and viewed generative AI tools as useful support systems within animation-related learning environments.

Besides, the study also highlighted that students' experiences with generative AI are closely connected to their ability to communicate effectively with AI systems through prompts. While many students recognized the benefits of AI-assisted technologies, several challenges related to originality, prompt construction, output consistency, and ethical concerns were also identified. These findings suggest that prompt literacy may become an increasingly important skill within future creative digital learning environments. Besides, the findings also suggest that future animation education environments may increasingly involve hybrid human–AI collaborative production approaches that combine creative interpretation with AI-assisted workflow support.

In the context of animation education, the findings provide practical implications for educators and curriculum developers. As generative AI becomes increasingly integrated into creative learning environments, animation programs should incorporate prompt literacy, ethical AI practices, and critical evaluation of AI-generated content into teaching and learning activities. Such initiatives can help students use generative AI more effectively while maintaining originality, creative judgment, and professional responsibility. This aligns with broader discussions on educational transformation, which suggest that higher education institutions need to continually adapt curriculum design and teaching practices in response to emerging technologies and changing learning environments [11]. Although generative AI can support creative exploration and workflow efficiency, the role of human creativity, artistic interpretation, and critical thinking remains essential throughout the animation production process.

Several limitations should be acknowledged when interpreting the findings. The study adopted a descriptive approach to explore students' perceptions within a specific educational context and did not investigate statistical relationships among prompt literacy, AI readiness, and creative workflow practices. Future studies could involve inferential statistical techniques, such as correlation or regression analysis, to examine these relationships in greater depth.

This study contributes to the growing discussion surrounding generative AI integration in creative education by providing preliminary insights into animation students' perceptions of prompt literacy and generative AI readiness within AI-assisted creative workflows. As the study was conducted in a single Malaysian public university and employed descriptive statistical analysis, the findings should be interpreted within this context. Future research should involve multiple higher education institutions, larger and more diverse samples, and inferential statistical approaches to examine the relationships among prompt literacy, generative AI readiness, and creative performance. Longitudinal or experimental studies could also investigate how prompt literacy training influences students' creative outcomes and AI-assisted animation workflows over time.

Funding

This research was supported by the Centre for Research, Development, Innovation, Commercialization, and Economy (RDICE) and the P. Ramlee Chair 2026, Faculty of Applied and Creative Arts, Universiti Malaysia Sarawak.

Acknowledgment

The authors would like to acknowledge Universiti Malaysia Sarawak (UNIMAS) for the support provided towards this publication.

Copyright © 2026 by the authors. This is an open access study distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited https://creativecommons.org/licenses/by/4.0/.

References

[1] E. Kasneci, K. Sessler, M. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, S. Groh, G. Günnemann, E. Hüllermeier, et al., "ChatGPT for good? On opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, p. 102274, 2023.

[2] Y. K. Dwivedi, L. Hughes, A. Baabdullah, S. Ribeiro-Navarrete, M. Giannakis, M. Al-Debei, D. Dennehy, B. Metri, D. Buhalis, C. M. Cheung, et al., "So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy," *International Journal of Information Management*, vol. 71, p. 102642, 2023.

[3] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, "Systematic review of research on artificial intelligence applications in higher education," *International Journal of Educational Technology in Higher Education*, vol. 21, no. 1, pp. 1–27, 2024.

[4] C. M. Chan and W. Hu, "Generative artificial intelligence in creative education: Emerging practices and implications for digital media learning," *Computers and Education: Artificial Intelligence*, vol. 6, p. 100221, 2025.

[5] P. Limna, T. Kraiwanit, and K. Juntoon, "Prompt literacy and generative AI competency in higher education learning environments," *Education and Information Technologies*, vol. 30, no. 2, pp. 1–18, 2025.

[6] A. Tlili, J. Shehata, B. Adarkwah, M. Bozkurt, D. Hickey, R. Huang, and D. Burgos, "What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education," *Smart Learning Environments*, vol. 12, no. 1, pp. 1–24, 2025.

[7] S. Bommasani, K. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al., "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[8] H. Baidoo-Anu and L. O. Ansah, "Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning," *Journal of AI*, vol. 7, no. 1, pp. 52–62, 2023.

[9] M. Bozkurt, "Generative artificial intelligence (AI) powered conversational educational agents: The inevitable paradigm shift," *Asian Journal of Distance Education*, vol. 18, no. 1, pp. 198–204, 2023.

[10] A. Rudolph, S. Tan, and S. Tan, "ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?" *Journal of Applied Learning and Teaching*, vol. 6, no. 1, pp. 342–363, 2023.

[11] J. McGrath and T. Fischetti, "What if compulsory schooling was a 21st century invention? Weak signals from a systematic review of the literature," *International Journal of Educational Research*, vol. 95, pp. 212–226, 2019.

[12] M. Rahman and S. Watanobe, "Generative artificial intelligence in education: A review of recent developments, opportunities, and challenges," *Information*, vol. 14, no. 11, p. 615, 2023.

[13] S. Feuerriegel, J. Hartmann, C. Janiesch, and P. Zschech, "Generative AI," *Business & Information Systems Engineering*, vol. 66, no. 1, pp. 111–126, 2024, doi: 10.1007/s12599-023-00834-7.

[14] L. Giray, "Prompt engineering with ChatGPT: A guide for academic writers," *Annals of Biomedical Engineering*, vol. 51, no. 12, pp. 2629–2633, 2023, doi: 10.1007/s10439-023-03272-4.

[15] L. S. Lo, "The CLEAR path: A framework for enhancing information literacy through prompt engineering," *The Journal of Academic Librarianship*, vol. 49, no. 4, p. 102720, 2023, doi: 10.1016/j.acalib.2023.102720.

[16] A. Johri, A. S. Katz, J. Qadir, and A. Hingle, "Generative artificial intelligence and engineering education," *Journal of Engineering Education*, vol. 112, no. 3, pp. 572–577, 2023, doi: 10.1002/jee.20537.

[17] L. Floridi and M. Chiriatti, "GPT-3: Its nature, scope, limits, and consequences," *Minds and Machines*, vol. 30, no. 4, pp. 681–694, 2020, doi: 10.1007/s11023-020-09548-1.

[18] D. Dellermann, P. Ebel, M. Söllner, and J. M. Leimeister, "Hybrid intelligence," *Business & Information Systems Engineering*, vol. 61, no. 5, pp. 637–643, 2019, doi: 10.1007/s12599-019-00595-2.

CDNet: Complex-Valued Deep Unrolled Composite Denoising Network for Accelerated Magnetic Resonance Image Reconstruction

David Muigai[1], Elijah Mwangi[2], Henry Kiragu[3]
Department of Electrical Engineering,
Pan African University Institute of Basic Sciences Technology and Innovation, Nairobi, Kenya[1]
Faculty of Engineering, University of Nairobi, Kenya[2]
Department of Electrical and Telecommunication Engineering, Multimedia University of Kenya, Nairobi, Kenya[3]

***Abstract*—Magnetic resonance imaging (MRI) is an essential modality in contemporary medical diagnosis, offering high-resolution images with excellent soft-tissue contrast of internal organs in a non-invasive, non-ionizing, and non-carcinogenic manner. However, the primary drawback of this technique is a lengthy data-acquisition process, which raises the possibility of picking up artefacts as well as discomfort in patients. To accelerate it, data is acquired at sub-Nyquist rates, and the image is recovered from undersampled, noisy measurements using the compressed sensing (CS) technique. Composite priors in CS–MRI have demonstrated superior performance compared to single priors, despite having limited image priors, which also leads to slower reconstruction speed. To date, deep unfolded networks (DUNs), which integrate the powerful data-driven prior of traditional deep learning (DL) with the strong interpretability of optimization algorithms, are the most outstanding reconstruction technology in MRI acceleration, yielding superior and interpretable results. This study proposes the iterative three-operator splitting (TOS) method to solve the composite regularization problem and a composite denoising network (CDNet) that maintains consistent arithmetic structures with it. The CDNet is implemented using complex-valued DL strategies for richer representation of MRI. Extensive experiments using brain and knee raw k-space from the FastMRI dataset with acceleration factors (AF) ranging from ×2 to ×10 and assessed using peak signal–to–noise ratio (PSNR), structural similarity index (SSIM), and normalized root–mean–squared error (NRMSE) validate the CDNet. Across AFs, CDNet achieved dataset-averaged PSNR/SSIM/NRMSE of 26.50–33.47dB/0.7787–0.8992/0.2562–0.1303 for brain MRI and 28.08–35.52dB/0.7069–0.8734/0.1929–0.0950 for knee MRI. The proposed network yields superior reconstruction accuracy and faster reconstruction speed compared to the iterative TOS for composite regularization, demonstrating the effectiveness of deep unrolling in accelerating iterative optimization. Furthermore, CDNet outperforms other state-of-the-art DUNs in reconstruction accuracy at comparable inference speed, demonstrating its vast potential for improving patient care, scanner throughput, and healthcare economics.**

***Keywords*—Accelerated MRI reconstruction; composite priors; iterative three-operator splitting; compressed sensing; deep unfolding; composite denoising network**

I. Introduction

Magnetic resonance imaging (MRI) is a critical and widely used research and diagnostic tool in modern medicine due to its non-invasive and non-ionizing acquisition method, which yields excellent contrast images of soft tissue. These properties make it preferred for visualizing the internal organs compared to other medical imaging modalities such as X–rays, computed tomography, and related techniques.

The MRI acquisition technique is based on the fact that the majority of the human body is composed of water molecules, which are made of two hydrogen atoms and one oxygen atom. Hydrogen-atom nuclei consist of a single proton, the tiniest magnetic unit. When exposed to an external magnetic field, these units align with and spin around the axis of the magnetic field at an angular frequency that depends on the strength of the field. This specific frequency is referred to as the precessional or Larmor frequency [1, 2]. The precessional frequency increases with the strength of the magnetic field. When a radio frequency (RF) pulse is applied at this precessional frequency, the spinning protons can absorb the RF energy, causing them to transition from a lower to a higher energy state, either aligning with or opposing the magnetic field. At this point, the protons are considered to be in resonance. Once the RF pulse is discontinued, the protons return to their original state and the energy difference gives rise to the MR signal. Protons in various tissues release energy at different rates due to the distinct chemical compositions and physical properties of each tissue.

The primary drawback of MRI is its slow data acquisition [3-6]. A magnetic resonance (MR) image is formed from multiple k-space acquisitions, separated by time intervals called repetition times or TR. Each acquisition results from applying an RF excitation. However, due to instrumental and physiological limitations, these acquisitions are performed sequentially for a specific field of view (FOV) [1]. Consequently, it takes a considerable amount of time to fully acquire the entire k-space needed to produce a single image.

This slow imaging speed poses a significant challenge, limiting clinical applications, particularly in real-time MRI, such as dynamic cardiac imaging, where only a limited number of samples can be acquired per cardiac cycle. In traditional MRI, k-space sampling adheres to the Nyquist criterion, which depends on the FOV and the MR image resolution. Accelerating MR data acquisition is possible by employing high-magnitude gradients, as these can reduce the TR time. However, the practical use of such rapidly switching gradients is limited because frequent changes can cause peripheral nerve stimulation in patients [1]. This inherent speed limitation of the MRI system has prompted the exploration of alternative technologies to increase MRI speed by undersampling k-space without sacrificing reconstruction quality.

Compressed sensing (CS) [7-10] enables the reconstruction of high-quality MR images from substantially fewer measurements than required by the Nyquist sampling criterion, at an additional computational cost compared to the inverse Fourier transform required for traditional full sampling. In CS, the reconstruction problem is formulated as an optimization that combines an ℓ_1-norm prior term to promote sparsity with an ℓ_2-norm data-fidelity term that enforces consistency with the acquired measurements. Iterative algorithms such as the iterative shrinkage-thresholding algorithm (ISTA) [11], fast iterative shrinkage-thresholding algorithm (FISTA) [12], and alternating direction method of multipliers (ADMM) [13] are then used to solve the problem, approximating the ground truth image.

Composite prior models in CS–MRI reconstruction have consistently demonstrated superior performance compared to single prior models [14-16]. Regularization terms are frequently employed to produce solutions that exhibit various characteristics based on the composite priors. Such priors include the sparse prior in the temporal Fourier domain [17], the total variation prior [18] that enforces sparsity in the spatial domain, and matrix low-rank priors [19], among others. However, the use of multiple priors leads to a composite regularization problem whose solution is computationally intensive, making it unsuitable for practical CS-MRI implementation [1, 14]. Also, there exist limited image priors in CS. In addition to the lengthy reconstruction times, empirical selection of hyperparameters, which is more complex than a single prior model, is nontrivial and difficult to choose for optimal performance. The composite prior model exhibits limited acceleration, up to ×5, with higher acceleration factors yielding images in which artifacts co-exist with useful visual information.

Deep unfolded networks (DUNs) [20-22] with multiple priors can be used to overcome the aforementioned challenges of solving the composite regularization problem with iterative algorithms. This is achieved by unrolling a fixed number of iterations of an optimization algorithm into a trainable end-to-end deep network, where the computationally expensive regularization terms are learned from data, while the computationally efficient data-consistency term is computed explicitly. Consequently, the unrolling converts the problem-specific hyperparameters of iterative algorithms into learnable parameters, thereby alleviating their empirical selection. DUNs have enabled the integration of the powerful data-driven prior of traditional deep learning (DL) with the strong interpretability of optimization algorithms, yielding superior and interpretable reconstruction results, making it the outstanding reconstruction technology in MRI acceleration to date [21, 23]. However, most DUNs [21, 24-28] employ a single DL method to enforce prior information, which may be limited.

To address the aforementioned challenges, we propose a complex-valued deep unrolled composite denoising network (CDNet) for accelerated magnetic resonance image reconstruction with the following specific contributions.

- We solve the composite regularization problem using the iterative three-operator splitting method and reconstruct MR images.
- We develop a complex-valued learned reconstruction architecture, CDNet, that maintains a consistent arithmetic structure via deep unfolding of the iterative three-operator splitting for composite regularization.
- We employ multiple complex-valued DL strategies to effectively capture both fine-grained spatial details and global structural information, thereby providing a richer representation of MRI.
- We conducted extensive experiments on raw complex brain and knee k-space data from the publicly available fastMRI dataset to evaluate CDNet and compare it with other state-of-the-art DUNs, demonstrating that CDNet outperforms existing methods.

II. Related Works

A. Single Priors

Sriram *et al.* [29] proposed an end–to–end variational network, dubbed E2E–VarNet, that accelerated MR image reconstruction in parallel imaging. Unlike prior variational networks that estimated the coil sensitivity maps before reconstruction, the authors in this work jointly learned the maps and the reconstruction process directly from the undersampled k-space. The work unrolled a variational optimization scheme into a network consisting of 12 cascades, with each cascade comprising a data-consistency and U-Net-based regularization block that refines the k-space. In addition to the reconstruction process, the authors added a sensitivity map estimation (SME) module, thus enabling end-to-end training. The E2E – VarNet was trained on knee and brain data from FastMRI and consistently outperformed other variational networks in reconstruction accuracy, as assessed using structural similarity index (SSIM) and peak signal–to–noise ratio (PSNR), especially at higher acceleration factors. Despite E2E–VarNet superseding other variational networks and demonstrating the significance of learning sensitivity maps, it exhibited some limitations. First, the SME module added half a million more parameters than other variational networks, increasing memory and computational costs. Secondly, the model evaluation relied primarily on quantitative metrics, namely normalized mean squared error (NMSE), SSIM, and PSNR, which may not capture subtle, clinically relevant details. This would necessitate further clinical validation before deployment to avoid a degradation in diagnostic accuracy.

Zhou *et al.* [30] proposed a dual–domain recurrent network, dubbed DuDoRNet, that jointly restores the k-space and image while embedding a deep T1 prior. Each cascade contained two dilated residual dense networks, one for k-space restoration and the other for image restoration, each with a data consistency layer and using a TI-weighted MRI for complementary structural guidance. The authors conducted extensive experiments on the network with three undersampling patterns, namely radial, Cartesian, and spiral trajectories. They used an in-house MRI dataset of 20 patients consisting of 252, 36, and 72 training, validation, and testing 2D images on acceleration of up to 6 using 5 cascades. The results were evaluated using PSNR, SSIM, and mean squared error (MSE) and demonstrated superior performance compared to other DL-based methods and CS-based methods. Despite the good reconstruction performance, the DuDoRNet has several limitations. First, it relies on the availability of a fully sampled T1 scan, which may be absent in protocol-limited clinical settings or time-critical medical examinations. Second, the recurrent dual-domain architecture increases computational complexity and memory requirements, limiting real-time clinical deployment. Thirdly, the network evaluation focused primarily on anatomical data, while other MRI data, such as pathological data, remain unexplored.

Xiang *et al.* [31] combined the free tuning and strong generalization of deep neural networks with the interpretability of FISTA to yield an unrolled network, dubbed FISTA-Net, for solving inverse problems in imaging. The FISTA-Net unfolds *n* FISTA iterations into *n* blocks, with each block consisting of a convolutional neural network (CNN) – based proximal mapping module, a gradient descent module, and a momentum update module. The FISTA-Net was evaluated quantitatively and visually using X-ray, computational tomography, and electromagnetic tomography. The experimental results show that it performs superiorly to the traditional iterative FISTA, an ISTA-based unrolled network, and other iterative and data-driven baseline algorithms. The network was not validated on MRI datasets with raw k-space measurements, limiting insights into its performance in practical MRI reconstruction scenarios.

Fabian *et al.* [32] introduced a hybrid unrolled multi-scale framework named HUMUS-Net that combines transformers and CNNs to accelerate MR image reconstruction. HUMUS-Net leverages the capability of transformers to identify long-range dependencies via self-attention and the proficiency of CNNs in efficiently extracting local features for MR image reconstruction. This design enables the network to extract multi-scale features by using transformers for low-resolution features and convolutions for high-resolution features. Additionally, HUMUS-Net integrates a sensitivity map estimator for multi-coil MRI reconstruction and employs unrolled iterative reconstruction steps, inspired by iterative optimization algorithms. The unrolled network achieved state-of-the-art performance, as measured by SSIM, PSNR, and NMSE, outperforming other existing networks. Nonetheless, HUMUS-Net is computationally intensive and requires substantial memory, particularly due to the transformer, which limits the depth of the network and patch size. Furthermore, the network requires fixed-size inputs, limiting its adaptability in clinical settings.

Zhang *et al.* [33] introduced an unrolled network, T^{2}LR-Net, that adaptively learns the optimal transformed domain to leverage tensor low-rank priors in dynamic MRI reconstruction. This method extended the tensor singular value decomposition (t-SVD) to incorporate multidimensional unitary transforms and introduced a novel unitary-transformed tensor nuclear norm (UTNN). By employing CNN modules, the approach surpassed the constraints of manually crafted transforms, learning appropriate transformations driven by reconstruction precision. T^{2}LR-Net unfolds N iterations of the ADMM algorithm into N modules, each comprising a reconstruction block, a multiplier update block, and a transformed-tensor low-rank prior block integrated with a CNN. Experimental outcomes assessed using signal-to-noise ratio (SNR) and SSIM demonstrated superior performance compared to other unrolling network-based and optimization-based methods. Nevertheless, the repeated SVD operations heighten the network's computational complexity, necessitating significant GPU memory and training time. The singular value thresholding (SVT) operators in each iteration stage may become nondifferentiable under certain conditions, such as when the matrix being solved has repeated singular values, potentially halting training. Additionally, performance slightly diminishes in multi-coil and prospective acquisition scenarios.

Vornehm *et al.* [34] introduced a variational network-based reconstruction framework dubbed cineVN, which unrolled an iterative optimization scheme into a deep network while incorporating the spatio–temporal structure of dynamic cardiac data. The cineVN network architecture consisted of 15 cascades, each containing a refinement term and a data consistency term, with a conjugate gradient block after every third cascade. The coil sensitivity maps were pre-estimated, and the complex data were split into real and imaginary parts. The proposed cineVN was evaluated on in vivo cardiac MRI acquisitions and demonstrated improvements over a spatiotemporal VN and a temporal total variation regularization – CS based approach (TTV-CS) in retrospective evaluation and TTV-CS and ℓ_1-wavelet regularization CS based approach in prospective evaluation. However, there is a potential breath-hold inconsistency between accelerated cineVN and GRAPPA reference acquisitions, as residual variations in respiratory state may persist despite exhalation-only imaging. The proposed network was also evaluated on a specific set of acquisitions and protocols, and a broader validation across different acquisition settings, scanners, and patient populations is essential before clinical deployment.

B. Multiple Priors

Ke *et al.* [20] presented a network dubbed SLR-Net which unfolded the ISTA procedure for dynamic MR imaging. Unlike other networks that focus on sparse priors of MR images alone, the authors in this work added a learned singular value thresholding to explore low-rank priors. The authors unrolled 10 iterations of ISTA into 10 cascades of SLR-Net, with each cascade comprising a sparse prior module, low-rank prior module, reconstruction module, and a multiplier update layer while maintaining the arithmetic structure of ISTA. The learned low–rank prior improved the reconstruction accuracy assessed using PSNR, MSE, and SSIM compared to CS and sparsity-driven DL methods. However, these SVD operations

added more computation time, which was limited to 0.5s, a range achieved using the supercomputing power of a GPU. In the absence of such GPUs in clinical deployment, real-time imaging may not be achieved. The module also added the memory requirement of the GPU, increasing hardware requirements.

Huang *et al.* [35] proposed a network dubbed the low-rank plus sparse (L+S) network to overcome two common failings of CS-based reconstruction: limited acceleration and the empirical selection of L+S parameters. The authors unrolled the alternating linearized minimization method to learn the regularization parameters. The L+S-Net contains 10 states, corresponding to 10 iterations of the iterative algorithm, with each stage comprising a data consistency layer, a sparse prior layer, and a low–rank prior layer. Experimental results on cardiac cine datasets demonstrate superior results compared to other existing CS and DL methods, especially at high acceleration factors. Despite the superior reconstruction results, L+S-Net exhibited several limitations. First, the network has SVD computations, which lead to extended reconstruction times. Second, the effect of the low-rank prior may degrade when the number of frames is reduced, since reconstruction is performed for each batch of frames, rendering it unsuitable for low-latency imaging. Third, the network was tested only on heart images, and further testing on other body parts is essential for its validation.

Zhang *et al.* [21] presented an unrolled network, JotlasNet, that incorporated attention-based sparse priors and a tensor low-rank prior for reconstructing dynamic MR images. The tensor low-rank prior exploited structural correlations in the data. The authors used a CNN to learn the priors. They unfolded the network using the composite splitting algorithm and set the number of cascades to 15. JotlasNet yielded better reconstruction results, as evaluated using PSNR and SSIM, on single- and multicoil data compared to other DUNS. However, JotlasNet is prone to numerical instability, which hinders the training process. The gradient of SVD is unstable, especially when singular values are repeated, leading to NaN gradients that hinder training. Also, the SVD significantly increases reconstruction time due to its computational complexity.

From the aforementioned limitations of existing DUNs, we propose the CDNet that allows effective integration of multiple priors in the reconstruction model. The CDNet is not prone to numerical instability and effectively combines multiple deep complex-valued DL strategies to capture different attributes of MR images and offer richer representation in MRI.

III. METHODOLOGY

A. Problem Formulation for Accelerated MRI

Magnetic resonance imaging (MRI) acquisition captures spatially dependent resonance-frequency information from the region under investigation. The MR images in the spatial domain are obtained by taking an inverse Fourier transform of the captured data. Accelerating MRI via undersampling results in an underdetermined system with an infinite set of solutions. The compressive sensing (CS) technique enables recovery of the MR image of interest by solving the basis pursuit denoising (BPDN) problem. The BPDN problem solution is obtained by solving the unconstrained optimization problem given in Eq. (1)

$$\hat{x} = \underset{x}{argmin} \underbrace{\|Fx - s\|_2^2}_{f(x)} + \lambda \underbrace{\|x\|_1}_{\Re(x)} \tag{1}$$

where, F denotes the forward model that maps the MR image x to the scanned data s. The ℓ_1 norm promotes sparsity in a suitable transform such as wavelets, and the λ is a regularization parameter that balances the sparsity $\Re(x)$ and data consistency term $f(x)$.

The solution to Eq. (1) closely approximates the original image. To improve the solution of Eq. (1) and consequently the reconstruction results, we can add additional sparsifying norms to increase overall sparsity. One such norm is the total variation (TV) norm [18], which promotes sparsity of image gradients. Therefore, the addition of the TV norm in Eq. (1) improves the solution by promoting sparsity in the spatial domain alongside the ℓ_1 norm. The general optimization problem can be given as in Eq. (2).

$$\hat{x} = \underset{x}{\text{minimize}}\ g(x) = f(x) + \textstyle\sum_{i=1}^{k} \lambda_i \Re_i(x) \tag{2}$$

The data consistency term $f(x)$ is a continuously differentiable function with a Lipschitz constant, while the $\Re_i(x)$ terms are nonsmooth convex functions. When $k = 1$, Eq. (2) reduces to Eq. (1), which can be easily solved using two-operator splitting algorithms such as ISTA, FISTA, and ADMM. When $k > 1$, Eq. (2) becomes computationally expensive to solve using the two-operator splitting algorithms due to multiple regularization terms.

B. Iterative Three Operator Splitting Algorithm for Composite Regularization

The iterative three-operator splitting (TOS) algorithm, first proposed by [36] and whose convergence was analyzed in [37], can be used to solve Eq. (2) for $k > 1$ in a computationally efficient manner. For example, we consider the case when $k = 2$, Eq. (2) becomes;

$$\hat{x} = \underset{x}{\text{minimize}}\ g(x) = f(x) + \lambda_1 \Re_1 + \lambda_2 \Re_2 \tag{3}$$

Following the works of [10, 14, 15], we can consider the total variation (TV) norm for spatial sparsity and an ℓ_1 norm in a suitable transform domain for transform sparsity. Eq. (3) can then be rewritten as;

$$\hat{x} = \underset{x}{\text{minimize}}\ g(x) = \underbrace{\|Fx - s\|_2^2}_{f(x)} + \lambda_1 \underbrace{\|\psi x\|_1}_{\Re_1(x)} + \lambda_2 \underbrace{\|x\|_{TV_{iso}}}_{\Re_2(x)} \tag{4}$$

where, ψ is a suitable transform such as wavelets and $\|x\|_{TV_{iso}}$ is the isotropic TV given by;

$$\|x\|_{TV_{iso}} = \textstyle\sum_{i,j} \sqrt{\left|x_{i+1,j} - x_{i,j}\right|^2 + \left|x_{i,j+1} - x_{i,j}\right|^2} \tag{5}$$

Eq. (4) can be solved iteratively using the TOS strategy. At each iteration, n, a gradient step, a proximal operator of TV, and a proximal operator of ℓ_1 are combined to produce a relaxed output for the next iteration. The TOS for $k = 2$ is given in Algorithm 1.

Algorithm 1: Three Operator Splitting Algorithm for $k = 2$

Inputs: Complex-valued zero-filled image y_0^0, Binary Mask M, number of iterations N, step size α and relaxation λ

For $n = 0, 1, 2, 3 \ldots N$

$x_1^n = prox_{\alpha\Re_1}(y_0^n)$

$\nabla f(x_1^n) = F^{-1}\left(M \odot (Fx_1^n - s)\right)$

$y_1^n = 2x_1^n - y_0^n - \alpha\nabla f(x_1^n)$

$x_2^n = prox_{\alpha\Re_2}(y_1^n)$

$x^{n+1} = y_0^n + \lambda(x_2^n - x_1^n)$

End

Return x_1^N

where, *prox* is the proximal operator defined as in Eq. (6) for a function φ [35].

$$prox_{\gamma\varphi}(x) = \underset{y\in\mathbb{C}^d}{argmin}\ \varphi(y) - \frac{1}{2\gamma}\|x - y\|_2^2 \qquad (6)$$

For $k > 2$ the algorithm 1 can be reformulated as iterates of *prox* of $\Re_i$ and the gradient step to give Algorithm 2.

Algorithm 2: Three Operator Splitting Algorithm for $k > 2$

Inputs: Complex-valued zero-filled image y_0^0, Binary Mask M, number of iterations N, step size α and relaxation λ

For $n = 0, 1, 2, 3 \ldots N$

$x_1^n = prox_{\alpha\Re_1}(y_0^n)$

$\nabla f(x_1^n) = F^{-1}\left(M \odot (Fx_1^n - s)\right)$

$y_1^n = 2x_1^n - y_0^n - \alpha\nabla f(x_1^n)$

$x_2^n = prox_{\alpha\Re_2}(y_1^n)$

$\nabla f(x_2^n) = F^{-1}\left(M \odot (Fx_2^n - s)\right)$

$y_2^n = 2x_2^n - y_1^n - \alpha\nabla f(x_2^n)$

$x_3^n = prox_{\alpha\Re_3}(y_2^n)$

•

•

•

$x_{K-1}^n = prox_{\alpha\Re_{K-1}}(y_{K-2}^n)$

$\nabla f(x_{K-1}^n) = F^{-1}\left(M \odot (Fx_{K-1}^n - s)\right)$

$y_{K-1}^n = 2x_{K-1}^n - y_{K-2}^n - \alpha\nabla f(x_{K-1}^n)$

$x_K^n = prox_{\alpha\Re_K}(y_{K-1}^n)$

$x^{n+1} = y_0^n + \lambda(x_K^n - x_1^n)$

End

Return x_1^N

C. Composite Denoising Network

Solving the $\Re_i(x)$ terms in the multiple-regulation optimization problem for MR image reconstruction in Eq. (2) efficiently and effectively is non-trivial. This may require creating subproblems and solving them using iterative algorithms, either simultaneously or sequentially, to achieve the desired solution. This leads to iterations within the main iterations of the TOS algorithm, which can significantly increase reconstruction time. Also, choosing the values of λ and α for effective MR image reconstruction is challenging. Motivated by these challenges of the iterative TOS algorithm for multiple regularization, we derive a learned reconstruction scheme dubbed composite denoising network (CDNet) by deep unrolling its iterations. We replace $prox_{\alpha\Re_i}$with parametrized complex-valued DL strategies $\xi_{k\theta_i}$ .The use of multiple complex-valued DL methods in the reconstruction network effectively combines their advantages, enabling the model to capture both fine-grained spatial details and global structural information. Complex-valued reconstruction networks also offer a richer representation of MR images as they are inherently complex. Algorithm 2 can be unrolled to Algorithm 3 given below.

Algorithm 3: Composite denoising network (CDNet)

Inputs: Complex-valued zero-filled image y_0^0, Binary Mask M, number of iterations N, step size α and relaxation λ

Learnable parameters: step size α, relaxation λ and $\xi_{k\theta_i}$

For $n = 0, 1, 2, 3 \ldots N$

$x_1^n = \xi_{1\theta_i}(y_0^n)$

$\nabla f(x_1^n) = F^{-1}\left(M \odot (Fx_1^n - s)\right)$

$y_1^n = 2x_1^n - y_0^n - \alpha\nabla f(x_1^n)$

$x_2^n = \xi_{2\theta_i}(y_1^n)$

$\nabla f(x_2^n) = F^{-1}\left(M \odot (Fx_2^n - s)\right)$

$y_2^n = 2x_2^n - y_1^n - \alpha\nabla f(x_2^n)$

$x_3^n = \xi_{3\theta_i}(y_2^n)$

•

•

•

$x_{K-1}^n = \xi_{K-1\theta_i}(y_{K-2}^n)$

$\nabla f(x_{K-1}^n) = F^{-1}\left(M \odot (Fx_{K-1}^n - s)\right)$

$y_{K-1}^n = 2x_{K-1}^n - y_{K-2}^n - \alpha\nabla f(x_{K-1}^n)$

$x_K^n = \xi_{K\theta_i}(y_{K-1}^n)$

$x^{n+1} = y_0^n + \lambda(x_K^n - x_1^n)$

End

Return x_1^N

The block diagram of CDNet is given in Fig. 1. $\xi_{k\theta_i}$are learnable parameters for composite denoising while D_k are data consistency terms computed numerically.

1) Discriminative learning for CDNet: We demonstrate the effectiveness of CDNet in accelerating MRI reconstruction by choosing $K = 2$ achieving a favorable trade-off between computational complexity, reconstruction time, and reconstruction accuracy. We propose a complex-valued CNN (CvCNN) to capture the overall anatomical structure and a complex-valued Unet (CvUnet) to capture the spatially local features and also enhance global features for $\xi_{1\theta_i}$ and $\xi_{2\theta_i}$ respectively. These DL methods offer lower computational complexity and hardware requirements compared to transformers and generative adversarial networks (GANs), thus preferred for CDNet. CvCNN and CvUNet differ in architectural design, but they are implemented in the same building blocks, namely, complex-valued convolution (CVC), complex-valued rectified linear unit ($\mathbb{C}$ Re L U) and radial batch normalization (RBN). While both architectures employ identical complex-valued building blocks, CvCNN uses a sequential feed-forward design, whereas CvUNet adopts an encoder–decoder structure with skip connections.

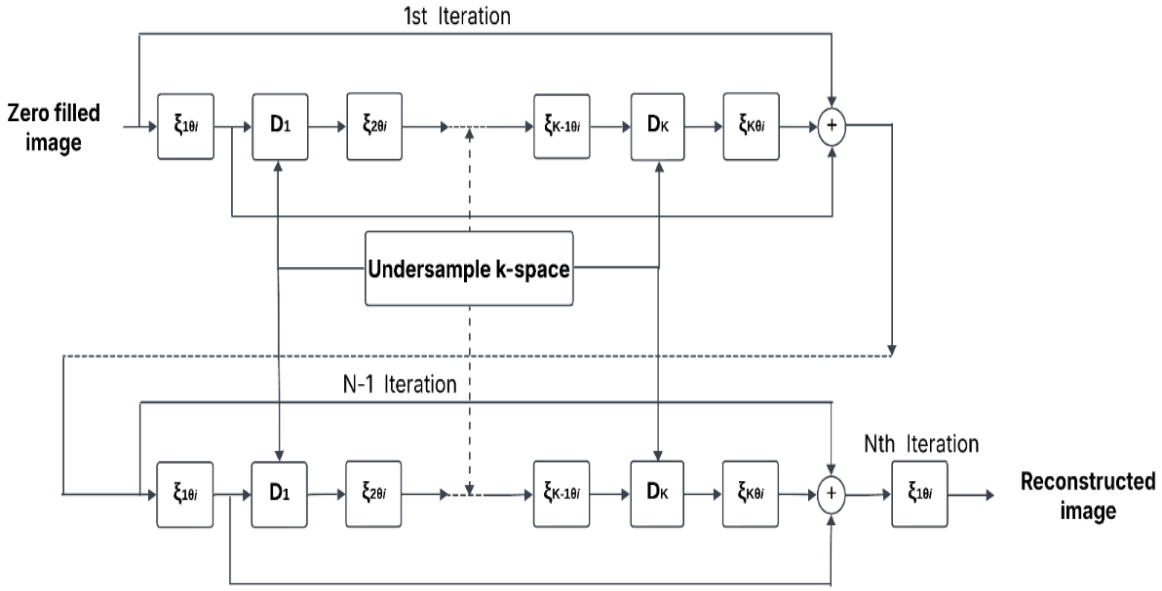


Fig. 1. Block diagram of the proposed CDNet.

The complex-valued convolution between the input zero-filled image $y_n^k = a + jb$ and kernel $k = c + jd$ can be realized using real-valued arithmetic by convolving the kernels c or d with each a and b as given in Eq. (7).

$$y_k^n * k = (c * a - d * b) + j(d * a + c * b) \tag{7}$$

where, $*$ represent convolution.

The RBN module stabilizes convergence and accelerates network training. The magnitude is scaled while preserving the phase, ensuring relative differences in complex quantities are preserved. We express the complex data y_{RBN} in polar form as given in Eq. (8) to obtain its magnitude f and phase θ. We normalize f as in Eq. (9)

$$y_{RBN} = f e^{j\theta} \tag{8}$$

$$f_{RBN} = \left(\frac{f-\mu_f}{\sqrt{\sigma_f^2}}\right)\kappa_1 + \kappa_2 + \varepsilon \tag{9}$$

where, σ_f^2 and μ_f are the variance and mean of f, κ_1 and κ_2 are trainable parameters for scaling and shifting the data distribution, and ε is a constant for ensuring f_{RBN} remains positive.

The network was activated with a $\mathbb{C}\,ReLU$ obtained by applying a real-valued ReLU to the real g_R and imaginary part g_I of the complex data g separately [38].

$$\mathbb{C}\,ReLU(g) = ReLU(g_R) + j\,ReLU(g_I) \tag{10}$$

2) Loss function: The CDNet was trained with a compound loss L^{comp} of ℓ_1 and multi-scale SSIM (MS-SSIM) loss given in Eq. (11) and its auxiliary parameters given in Eq. (12) to (15). The balancing weight χ was determined empirically in [39] to be 0.84, while [40] demonstrated the superiority of L^{comp} compared to the MSE, Huber, SSIM, perceptual loss, MSSIM, and ℓ_1 loss in training medical image reconstruction models.

$$L^{comp} = \chi L^{MS-SSIM}(|rec|, |gt|) + (1-\chi)L^{\ell_1}(|rec|, |gt|) \tag{11}$$

where, rec, gt, $L^{MS-SSIM}$, L^{ℓ_1} is the reconstructed image, ground truth image, MS-SSIM loss given in Eq. (11), and ℓ_1 loss given in Eq. (16), respectively, and $|.|$ denotes magnitude.

$$L^{MS-SSIM} = l_N\left(|rec|, |gt|\right)^{\upsilon_N} \prod_{j=1}^{N}\left[c_j\left(|rec|, |gt|\right)\right]^{\delta_j}\left[s_i\left(|rec|, |gt|\right)\right]^{\tau_j} \tag{12}$$

where, l, c and s are luminance, contrast, and structure given in Eq. (13), (14), and (15), respectively, and υ, δ_j and τ_j are weights of the distortion factors at different levels.

$$l(|rec|, |gt|) = \frac{2\mu_{|rec|}\mu_{|gt|}+C_1}{\mu_{|rec|}^2+\mu_{|gt|}+C_1} \tag{13}$$

where, $\mu_{|rec|}$ and $\mu_{|gt|}$ are averages of reconstructed and ground truth images, respectively, and C_1 is a constant.

$$c(|rec|, |gt|) = \frac{2\sigma_{|rec|}\sigma_{|gt|}+C_2}{\sigma_{|rec|}^2+\sigma_{|gt|}^2+C_2} \tag{14}$$

where, $\sigma_{|rec|}$and $\sigma_{|gt|}$are variances of the reconstructed and ground truth images, respectively, and C_2 is a constant.

$$s(|rec|, |gt|) = \frac{\sigma_{|rec||gt|}+C_3}{\sigma_{|rec|}\sigma_{|gt|}+C_3} \tag{15}$$

where, $\sigma_{|rec||gt|}$ is the covariance between the reconstructed and ground truth images.

$$L^{\ell_1} = \big||rec| - |gt|\big|_1 \tag{16}$$

D. Implementation Details

1) Iterative details: We implemented the iterative TOS for $k = 2$ in the open-source Python-based PyTorch library on a PC with an Intel CPU of 2.30 GHz and 64 GB RAM. $\Re_1$ was

solved using the Fast Iterative Shrinkage-Thresholding Algorithm FISTA [12] with ψ chosen to be the discrete wavelet transform for promoting sparsity in the transform domain, while $\Re_2$was solved using the anisotropic TV norm given in Eq. (5) for promoting sparsity in the spatial domain. The algorithm was set to 200 iterations, with each consisting of one thresholding step, one data consistency step, one anisotropic TV step, and a relaxation update. The relaxation λ was set to 0.05, while the step size α was adaptive.

2) Unrolled details: The CDNet was implemented using the open - source Python-based PyTorch library in a GPU–enabled Kaggle environment utilizing an NVIDIA Tesla P100. The number of cascades was set to 4, a number validated empirically through experiments and offering a good tradeoff between reconstruction time, hardware requirements, and reconstruction accuracy. The CvCNN complex convolutions use a kernel size of 3×3, 5 convolutional layers, and 32 hidden channels. The CvUnet uses 3 levels with a bottleneck, employing 3×3 convolutions, 2×2 max-pooling and transposed convolutions for down- and up-sampling, 32 base channels, and symmetric skip connections. An Adam optimizer [41] was employed to train the proposed network with a learning rate of 0.002 and momentum parameters set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The training was conducted for ×2, ×3, ×4, ×5, ×7, and ×10 accelerations using a batch size of 3 and limited to 55 and 35 epochs for the brain and knee datasets, respectively, to avoid overfitting and preserve generalization.

E. Performance metrics

The proposed CDNet was evaluated quantitatively using PSNR, SSIM, and NRMSE computed against the ground truth images, while sample reconstructed images are given for qualitative assessment. The reconstruction time was measured in seconds. The PSNR between the reconstructed rec and ground truth gt image, whose magnitude is normalized to [0, 1], is given by Eq. (17).

$$PSNR(|rec|, |gt|) = -10\, log_{10} MSE \tag{17}$$

where, MSE is computed as in Eq. (18).

$$MSE(|rec|, |gt|) = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C}(|rec|-|gt|)^2}{RC} \tag{18}$$

where, R and C is the number of rows and columns, respectively. The PSNR ranges from 0 to $+\infty$, with higher value.

values indicating better reconstruction results.

The SSIM between rec and gt is given in Eq. (19).

$$SSIM\left(|rec|, |gt|\right) = l\left(|rec|, |gt|\right)^{\upsilon} c\left(|rec|, |gt|\right)^{\delta} s\left(|rec|, |gt|\right)^{\tau} \tag{19}$$

where, l, c and s are luminance, contrast, and structure defined in Eq. (13), (14), and (15), respectively, and υ, δ and τ are weights of the distortion factors at different levels. SSIM ranges from 0 to 1, with 1 indicating complete similarity and vice versa.

The NRMSE between rec and gt is given in Eq. (20).

$$NRMSE(|rec|, |gt|) = \sqrt{\frac{MSE(|rec|, |gt|)}{MSE(|gt|, 0)}} \tag{20}$$

A low NRMSE indicates good reconstruction results and vice versa.

IV. Results and Discussion

A. Dataset

The brain and knee raw complex k-space data from the FastMRI dataset [42] were used for this research. This is the largest public dataset for MR slices created in collaboration between Facebook AI Research and NYU Langone Health. We selected the single-coil knee dataset, which comprises 973, 199, and 108 volumes for training, validation, and testing, respectively, each containing 34742, 7135, and 3903 slices, respectively. However, the proposed network was evaluated using MR volumes with high-quality slices without fat suppression. This enabled the proposed network to learn anatomical details and the general structure from high-SNR slices with strong overall contrast. A total of 170, 50, and 60 volumes, each containing 6150, 1808, and 2229 slices, respectively, were selected to evaluate the proposed network.

Brain raw complex k-space was also used to evaluate the proposed network. Since the FastMRI dataset provides multicoil k-space for brain slices and the proposed network was evaluated using single-coil data, we enumerated the single-coil data using an adaptive root–sum–of–squares with phase. The magnitude, phase, and complex image were obtained as given in Eq. (21), (22), and (23), respectively.

$$|gt(i,j)| = \sqrt{\sum_{n=1}^{N} |gt_n(i,j)|^2} \tag{21}$$

$$e^{j\theta_{ave(i,j)}} = \frac{\sum_{n=1}^{N} gt_n(i,j)}{|\sum_{n=1}^{N} gt_n(i,j)|} \tag{22}$$

$$gt(i,j) = |gt(i,j)|.\, e^{j\theta_{ave(i,j)}} \tag{23}$$

where, N is the number of coils and $gt_n(i,j)$ is the complex-valued ground truth image for the nth coil. A total of 500, 100, and 150 volumes containing 7910, 1584, and 2392 slices, respectively, were enumerated and their k-spaces and images stored in hierarchical data format version 5 (HDF5) for efficient organization and retrieval during training. A sample of enumerated brain magnitude and phase MR images of size 768 × 396 is shown in Fig. 2.

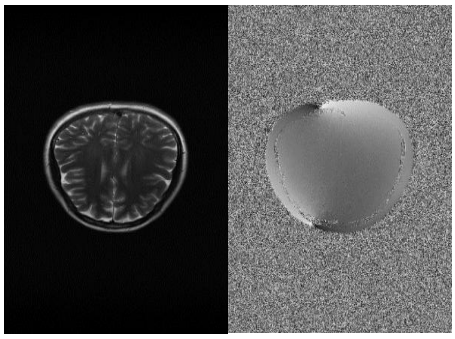


Fig. 2. From left to right, a sample magnitude image and its corresponding phase image.

B. Data Acquisition and Preprocessing

The acquisition was accelerated by undersampling fewer samples with a variable-density pattern that mimics realistic MRI scanners. The phase-encoding lines from the center region corresponding to a percentage equal to the acceleration factor (AF) were retained, as this region contains information about the contrast and gross structure of the MR image. The other lines were sampled probabilistically, ensuring denser sampling near the center and sparser sampling towards the periphery, reducing acquisition artifacts.

The input images, namely the zero-filled image obtained as an inverse fast Fourier transform (IFFT) of the undersampled k-space filled with zeros at the missing k-space positions, and the ground truth images obtained as an IFFT of the fully sampled k-space, were normalized to [0, 1] while preserving the phase according to Eq. (24)

$$gt_N(i,j) = \frac{gt(i,j)}{max_{i,j}|gt(i,j)|+\varepsilon} \quad (24)$$

where, $gt_N(i,j)$ is the normalized ground truth image and ε is a small constant for numerical stability set to 10^{-8}.

The knee MR images were center-cropped to 320 × 320, while the brain MR images were center-cropped to 192 × 192. The 192 × 192 was achieved by resizing in the spatial domain from the original sizes to 320 × 320, followed by a resizing in the Fourier domain to 192 × 192. This ensured a reduction in the background, which is a region of non-interest in the medical domain, while preserving the region of interest. This allows the proposed network to learn useful anatomical details with less computational overhead without compromising the MR image. A sample brain MR image of size 640 × 320 resized to 192 × 192 is shown in Fig. 3.

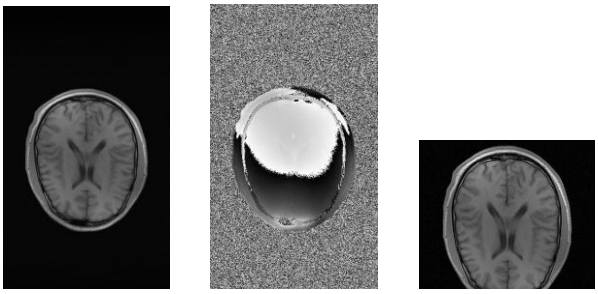


Fig. 3. From left to right: sample magnitude brain MR image and its corresponding phase image of size 640 × 320 resized to 192 × 192.

C. Iterative TOS for composite regularization

Fig. 4 shows sample reconstructed MR images, with the first row showing the brain image and the second row showing the knee image at ×2 to ×5 AF. Ground-truth images are also provided as baselines for comparison. The performance metrics are given in Table I. The average reconstruction time for brain MR is 1.4340s, while that of knee MR is 2.6721s. This difference is due to unequal resizing of the two images, with the knee being resized to 320 × 320 while the brain is resized to 192 × 192. The performance metrics decreased as AF increased, due to fewer measurements acquired for reconstruction. At AFs ×4 and ×5, residual acquisition artifacts coexist with the anatomical structure of the MR image.

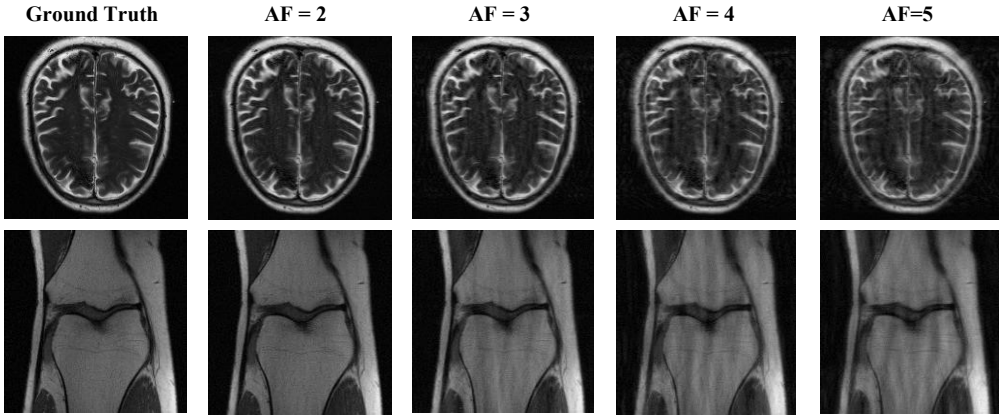


Fig. 4. Sample iterative TOS reconstruction results.

TABLE I. PERFORMANCE METRICS FOR SAMPLE RECONSTRUCTED MR IMAGES IN FIG. 4

	Brain				Knee			
AF	**PSNR (dB)**	**SSIM**	**NRMSE**	**Time (s)**	**PSNR (dB)**	**SSIM**	**NRMSE**	**Time (s)**
×2	31.95	0.9000	0.0861	1.4465	32.21	0.8280	0.0775	2.6039
×3	27.01	0.7920	0.1548	1.4759	27.33	0.7838	0.1263	2.7217
×4	24.49	0.7390	0.2073	1.4131	25.29	0.7166	0.1586	2.6298
×5	23.06	0.6982	0.2573	1.4005	24.28	0.6845	0.1747	2.7331

D. CDNet

1) The effect of the number of cascades n: We investigated the effect of increasing the number of cascades in the CDNet using the brain k-space with a ×5 acceleration and the statistical quantitative results presented in Table II. The PSNR, SSIM, NRMSE, and time report the average values computed over the entire brain test images. The reconstruction time increases with an average of 0.02875s per added cascade. The quantitative metrics improved greatly when learning started, with PSNR, SSIM, and NRMSE reporting improvements of 4.86 dB, 0.2040, and 0.1506, respectively. These results continued to improve when using two cascades, with PSNR, SSIM, and NRMSE reporting improvements of 0.86 dB, 0.0201, and 0.0183 compared to using a single cascade. The third cascade also showed improvements of 0.49 dB, 0.0225 and 0.0052 in PSNR, SSIM and NRMSE, respectively, compared to learning with two cascades. Since the results were still improving significantly, we also experimented with the CDNet with four cascades. The results showed improvements of 0.24 dB, 0.0046, and 0.0083 for PSNR, SSIM, and NRMSE, respectively. The fifth cascade did not significantly improve the quantitative results, indicating saturation. However, the reconstruction time continues to increase, and hence the choice of $N = 4$ which offers a good trade-off between reconstruction time and accuracy for the proposed CDNet.

A sample reconstructed image and its corresponding error map computed as the difference between the ground truth and the reconstructed image are shown in Fig. 5 for qualitative assessment. The zero-filled and ground-truth images are provided as a baseline for visual comparison. The reconstructed image using a single cascade shows significant improvement over the zero-filled image. However, noticeable differences between the reconstructed and ground truth image are visible. The image reconstructed using two cascades resembles the ground truth image better, with fine anatomical details clearer than in the single-cascade reconstruction. Still, noticeable differences between the reconstructed and ground truth images are visible. The reconstructed image using three cascades is visually better than that using two cascades, with edges showing fine anatomical details. The images reconstructed with four and five cascades are imperceptible to the human eye compared to the ground truth, despite their error maps showing differences. Table III gives the performance metrics for the sample MR image shown in Fig. 5.

TABLE II. DETERMINING THE NUMBER OF CASCADES *N* USING BRAIN DATA ON ×5 ACCELERATION

No of cascades *n*	Image	PSNR (dB)	SSIM	NRMSE	Time (s)
0	Zero filled	22.60	0.5806	0.3822	-
1	Reconstructed	27.46	0.7846	0.2316	0.1020
2	Reconstructed	28.32	0.8047	0.2133	0.1317
3	Reconstructed	28.81	0.8272	0.2081	0.1600
4	**Reconstructed**	**29.05**	**0.8318**	**0.1998**	**0.1885**
5	Reconstructed	29.09	0.8320	0.1996	0.2170

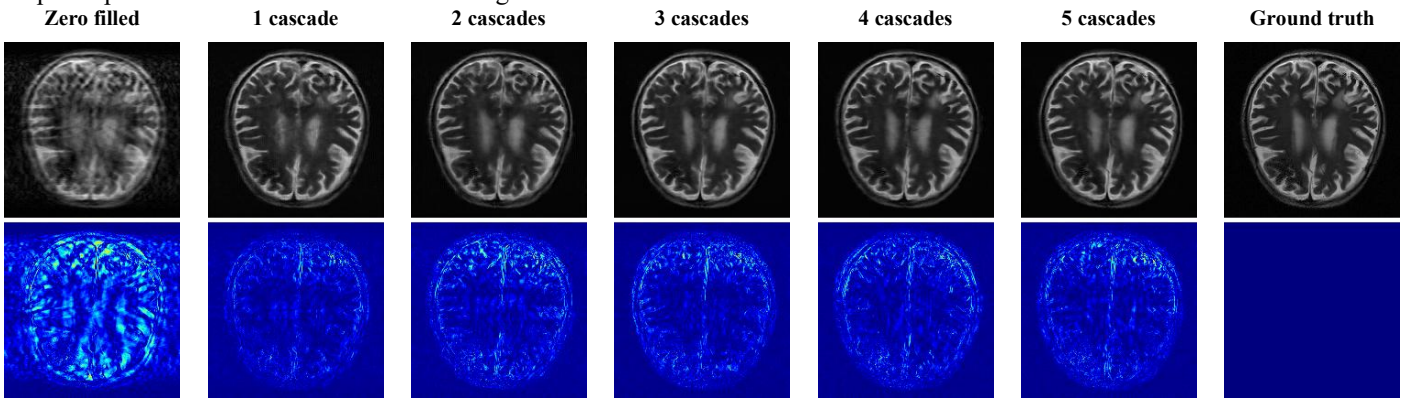


Fig. 5. Sample brain MR images reconstructed using different numbers of cascades.

TABLE III. PERFORMANCE METRICS FOR THE BRAIN MR IMAGE SHOWN IN FIG. 5

No of cascades *n*	Image	PSNR (dB)	SSIM	NRMSE
0	Zero filled	20.86	0.6025	0.6222
1	Reconstructed	27.64	0.8293	0.4212
2	Reconstructed	27.89	0.8457	0.4151
3	Reconstructed	28.50	0.8844	0.4007
4	**Reconstructed**	**29.13**	**0.8926**	**0.3865**
5	Reconstructed	28.89	0.8899	0.3918

2) CDNet learning characteristics: Fig. 6 shows the train and validation loss curves over epochs under different AFs for brain and knee datasets, respectively.

The losses showed a steeper decline for the first 30 and 20 epochs for the brain and knee datasets, respectively, indicating faster initial learning. The losses later decreased slowly up to the 40th epoch and 25th epoch for the brain and knee datasets, respectively, after which they oscillated with small fluctuations, indicating convergence. The losses increased with an increase in AF as less data was used for reconstruction, making it a more challenging task. The train and validation curves for each AF exhibit a small gap, indicating that overfitting is effectively controlled.

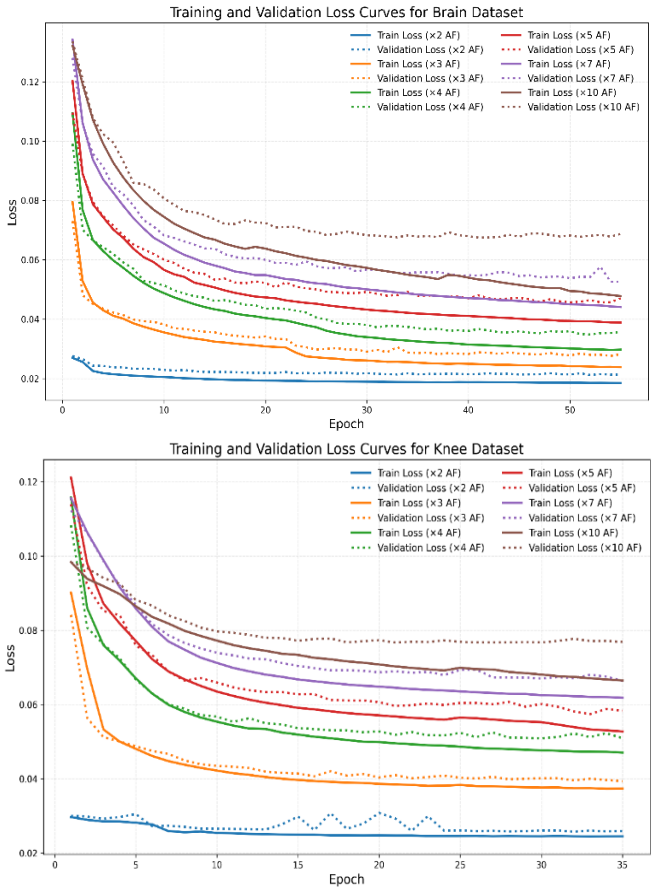


Fig. 6. From top to bottom, plots of training and validation loss over epochs for the brain and knee datasets, respectively.

3) *Quantitative results*

a) Brain data results: Table IV presents the quantitative statistical results for brain data, reporting average PSNR, SSIM, and NRMSE values computed over the entire test brain images. The zero-filled results were used as a baseline for comparison. The table also provides the improvements from the zero-filled images. The zero-filled results decreased as the AF was increased from ×2 to ×7, as fewer samples were acquired. The zero-filled results for ×10 are slightly higher than those of ×7 due to a higher percentage of retained center lines without a significant decrease in the number of acquired phase-encoding lines. The reconstructed result decreases when the AF is increased from ×2 to ×10 due to fewer samples being acquired, which increases acquisition artifacts and makes it challenging for the CDNet to denoise. The CDNet showed great improvement in reconstruction accuracy, highlighted in bold for all AFs.

b) Knee data results: Table V presents the statistical results for knee data, reporting the mean values of PSNR, SSIM, and NRMSE computed over the entire knee test data. The average reconstruction time for the knee MR image is 0.2963 seconds per slice. This is higher than the reconstruction time for a brain MR image of 0.1885 seconds per slice. The higher reconstruction time for knee MR images is due to their larger size (320 × 320) compared to brain MR images (192 × 192). The background of brain MR images was large and the region of interest could be carried comfortably by cropping it in a smaller size compared to knee images. The zero-filled performance metrics decreased with increasing AF from ×2 to ×5, with ×7 and ×10 showing slight improvements. This is due to fewer samples being acquired for reconstruction, while the slight improvement is due to a higher percentage of retained center lines.

TABLE IV. STATISTICAL QUANTITATIVE RESULTS FOR BRAIN DATASET

	Zero filled			**Reconstructed**			**Improvements**		
AF	PSNR (dB)	SSIM	NRMSE	PSNR (dB)	SSIM	NRMSE	PSNR (dB)	SSIM	NRMSE
×2	30.71	0.8199	0.1652	33.47	0.8992	0.1303	**2.76**	**0.0793**	**0.0349**
×3	25.29	0.6703	0.2806	31.69	0.8776	0.1532	**6.40**	**0.2073**	**0.1274**
×4	23.25	0.6066	0.3535	30.48	0.8563	0.1754	**7.23**	**0.2497**	**0.1781**
×5	22.60	0.5806	0.3822	29.05	0.8318	0.1998	**6.45**	**0.2512**	**0.1824**
×7	22.27	0.5635	0.4021	28.06	0.8128	0.2198	**5.79**	**0.2493**	**0.1823**
×10	22.53	0.5727	0.3945	26.50	0.7787	0.2562	**3.97**	**0.2060**	**0.1383**

TABLE V. STATISTICAL QUANTITATIVE RESULTS FOR KNEE DATASET

	Zero filled			**Reconstructed**			**Improvements**		
AF	PSNR (dB)	SSIM	NRMSE	PSNR (dB)	SSIM	NRMSE	PSNR (dB)	SSIM	NRMSE
×2	32.08	0.8261	0.1277	35.42	0.8734	0.0950	**3.34**	**0.0473**	**0.0327**
×3	25.65	0.7137	0.2341	31.88	0.8234	0.1273	**6.23**	**0.1097**	**0.1068**
×4	23.72	0.6555	0.2958	29.76	0.7966	0.1610	**6.04**	**0.1411**	**0.1348**
×5	23.54	0.6302	0.3088	29.26	0.7798	0.1775	**5.72**	**0.1496**	**0.1313**
×7	24.19	0.6099	0.2937	28.84	0.7580	0.1859	**4.65**	**0.1481**	**0.1078**
×10	25.41	0.5978	0.2571	28.08	0.7069	0.1929	**2.67**	**0.1091**	**0.0642**

The reconstruction accuracy decreases significantly as the number of acquired samples decreases. For all the AFs, the CDNet showed significant improvement from the zero-filled results highlighted in bold.

4) Qualitative results

a) Brain data results: A sample reconstructed image for each AF is shown in Fig. 6 for visual assessment, and their performance metrics are given in Table VI. Each reconstructed image is shown alongside its binary mask used to undersample the k-space, its zero-filled image, and its ground truth image, which serves as a baseline for comparison and an error map for visualizing the difference between the reconstructed and ground truth images, as they are imperceptible. The 1s corresponding to white in the binary mask show the position of acquired phase-encoding lines, while the 0's corresponding to black show the position of unacquired phase-encoding lines. The binary masks show some phase-encoding lines retained at the center, with denser sampling near the center and progressively sparser towards the periphery. This prevents severe acquisition artifacts, especially at higher AFs. The acquisition artifacts in zero-filled images increase with increasing AFs, since fewer data are available for reconstruction. The reconstructed and ground-truth images are visually similar. The bright regions indicating discrepancies between the reconstructed and ground-truth images in error maps are more numerous at higher AFs than at lower AFs. This indicates that CDNet's reconstruction accuracy decreases with an increase in AF. However, the acquisition artifacts are absent in all AF sample reconstructed images, demonstrating high-quality CDNet reconstruction results.

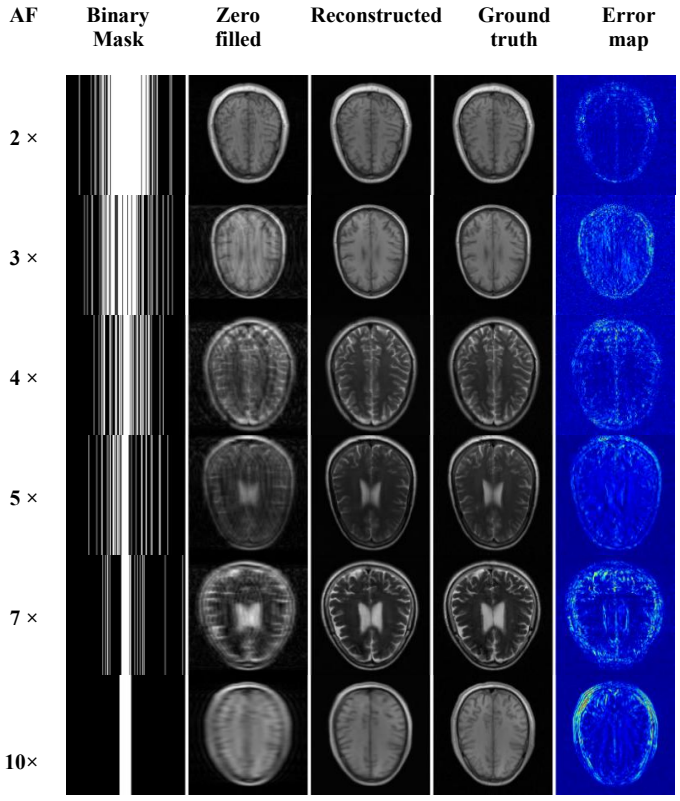


Fig. 7. Sample brain MR images reconstructed at different AFs.

a) Knee data results: A sample reconstructed knee MR image for each AF is shown in Fig. 7 for qualitative evaluation, and their performance metrics are given in Table VII. The reconstructed images are shown alongside their binary mask, their zero-filled and ground truth images, which serve as a baseline for visual comparison, and their error maps, which show differences between the reconstructed images and the ground truth images, as they are visually indistinguishable from the ground truth.

TABLE VI. PERFORMANCE METRICS FOR BRAIN MR IMAGES IN FIG. 6

	Zero filled			Reconstructed		
AF	PSNR (dB)	SSIM	NRMSE	PSNR (dB)	SSIM	NRMSE
×2	32.85	0.8984	0.2835	37.56	0.9483	0.2162
×3	23.80	0.6501	0.5022	36.03	0.9328	0.2484
×4	22.23	0.6480	0.5757	32.92	0.9107	0.3111
×5	22.40	0.5888	0.5937	28.84	0.8889	0.4098
×7	20.70	0.5602	0.6228	28.44	0.8686	0.3987
×10	22.24	0.6368	0.5157	27.69	0.8409	0.3769

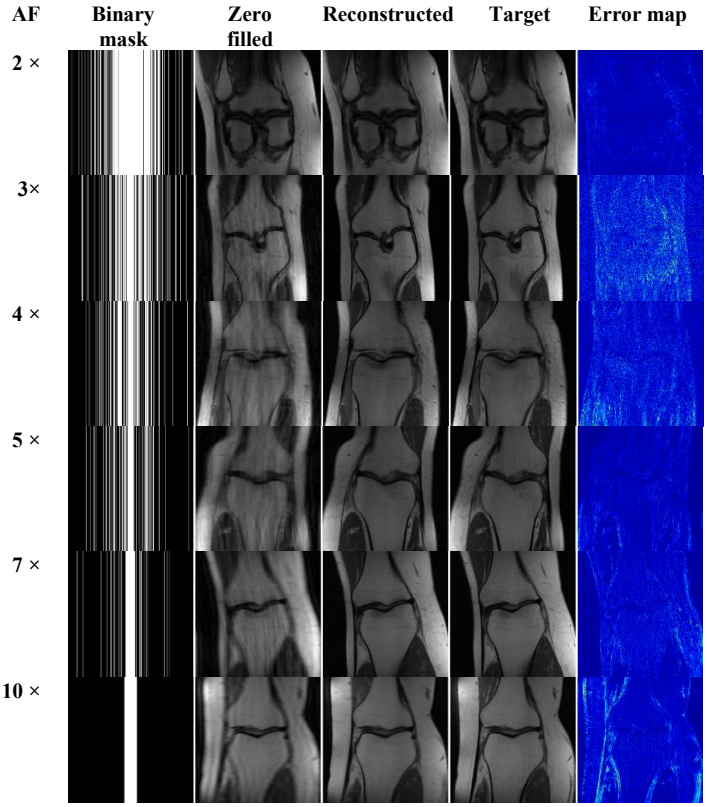


Fig. 8. Sample knee MR images reconstructed at different AFs.

Similar to binary masks in brain data quantitative results, the 1s correspond to white and show positions of acquired phase encoding lines, while the 0s correspond to black and show positions of unacquired measurements.

Some lines are retained at the center to preserve the anatomical structure, while the remaining lines are sampled densely near the center and progressively sparser towards the periphery. This avoids severe acquisition artifacts that may render the reconstructed image unsuitable for diagnosis. The zero-filled images show that acquisition artifacts increase with higher AF as fewer samples are collected for reconstruction. The reconstructed and ground-truth images in each AF are visually similar. However, the error maps show the differences between the reconstructed and ground-truth images, with brighter regions indicating discrepancies between the two images being more at higher AFs and blue regions indicating identical regions dominating in lower AFs. The reconstructed images are of good quality and artifact-free at all AFs, demonstrating good image reconstruction of the proposed CDNet.

TABLE VII. PERFORMANCE METRICS FOR KNEE MR IMAGES IN FIG. 7

	Zero filled			Reconstructed		
AF	PSNR (dB)	SSIM	NRMSE	PSNR (dB)	SSIM	NRMSE
×2	33.10	0.8747	0.2789	37.24	0.9155	0.2198
×3	26.39	0.7339	0.4184	34.28	0.8598	0.2655
×4	23.16	0.7024	0.4950	32.31	0.8253	0.2923
×5	24.02	0.6732	0.4713	32.20	0.8252	0.2943
×7	25.15	0.6511	0.4093	30.61	0.7778	0.2990
×10	25.41	0.6531	0.3854	28.42	0.7373	0.3240

E. Model Comparison

The proposed CDNet was compared to zero-filled results obtained by filling the unacquired phase-encoding lines with zeros and performing an inverse FFT on the zero-filled k-space, and with the results of four other state-of-the-art unrolled architectures, namely a deep cascade of convolutional neural networks (DC-CNN) [43], variational network (VarNet) [44], interpretable optimization-inspired deep network ISTAnet [45], and learned primal–dual reconstruction LPDNet [46]. Their architectures were obtained from the authors' websites and retrained using the same network settings and data as in CDNet for fair comparison. The quantitative results for an AF of ×5 are given in Table VIII, with PSNR, SSIM, and NRMSE reporting average values computed over the entire test brain MR images. The proposed CDNet yielded better statistical results evaluated using the performance metrics than other unrolled networks at a comparable reconstruction time. This demonstrates the superiority of the CDNet over other state-of-the-art unrolled architectures for MR image reconstruction.

TABLE VIII. QUANTITATIVE RESULTS FOR MODEL COMPARISON USING AN AF OF ×5

Method	PSNR (dB)	SSIM	NRMSE	Time (s)
Zero-filled	22.60	0.5806	0.3822	-
DC-CNN	25.64	0.7342	0.2721	0.1184
VarNet	26.25	0.7484	0.2532	0.1257
ISTANet	26.46	0.7575	0.2558	0.1156
LPDNet	28.12	0.7981	0.2213	0.1158
CDNet	**29.05**	**0.8318**	**0.1998**	**0.1885**

Fig. 9 shows a sample brain MR image and its corresponding error map reconstructed using iterative TOS for composite regulation, DC-CNN, VarNet, ISTAnet, LPDNet, and the proposed CDNet for qualitative comparison. The performance metrics of this sample image are given in Table IX. The zero-filled and ground truth images are also shown to act as baselines for visual comparison. When the sample image is reconstructed using the iterative TOS for composite regulation, it results in an image with visible residual acquisition artifacts coexisting with the anatomical structure of the image. The image was also reconstructed in 1.4510s, which is approximately × 7.7 higher than that of CDNet. This high reconstruction time is due to the high number of iterations required for convergence of iterative CS algorithms. The proposed CDNet yielded the best image assessed visually, with edges of fine details seen with ease compared to the other unrolled networks and the iterative TOS. This shows that the proposed CDNet yields superior reconstruction accuracy compared to other deep unrolled networks in MR image reconstruction.

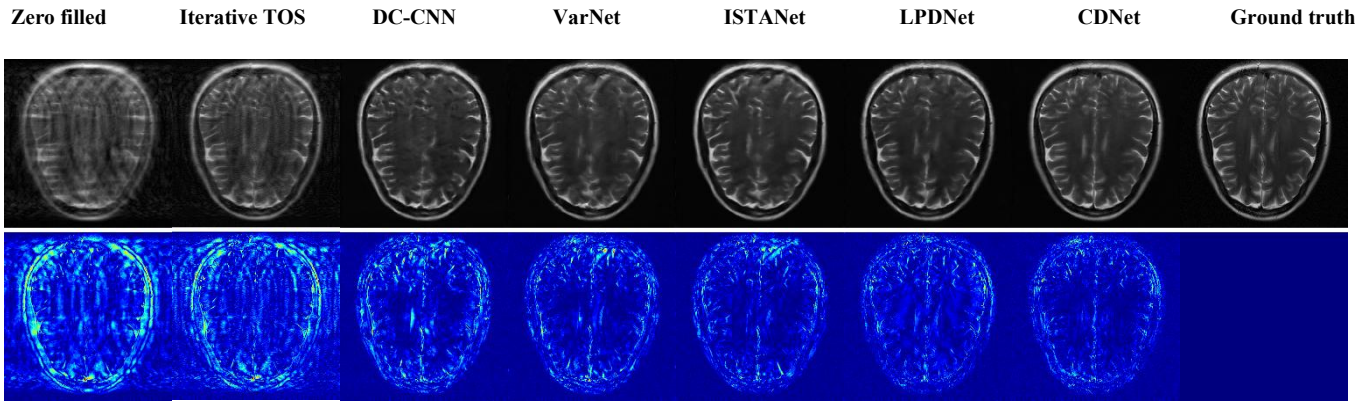


Fig. 9. Sample brain MR image and its corresponding error map reconstructed using comparative methods.

TABLE IX. PERFORMANCE METRICS OF THE SAMPLE BRAIN MR IMAGE IN FIG. 8

Method	PSNR (dB)	SSIM	NRMSE
Zero-filled	21.12	0.5463	0.6238
Iterative TOS	25.61	0.7141	0.4759
DC-CNN	26.48	0.7978	0.4585
VarNet	27.68	0.8313	0.4278
ISTANet	27.98	0.8457	0.4205
LPDNet	29.33	0.8845	0.3891
CDNet	**30.63**	**0.8900**	**0.3609**

V. CONCLUSION

In this work, we introduce a composite denoising network for reconstructing high-quality MR images using fewer measurements acquired at sub-Nyquist rates. In particular, we solve the composite regularization problem in MRI using the iterative three-operator splitting strategy and reconstruct MR images using two priors that promote sparsity in the transform and spatial domains. We then unfolded the iterative algorithm into a composite denoising network that maintains consistent arithmetic architectures with it and employs multiple complex-valued deep learning strategies for capturing different attributes of the MR images. We implemented the CDNet with two complex-valued DL strategies, namely complex-valued CNN and U-Net, for capturing the overall anatomical structure and spatially local features with enhanced global features, respectively. The hyperparameters of the iterative algorithm are learned from data, eliminating the need for empirical selection. Extensive experiments using raw brain and knee k-space from FastMRI, with acceleration factors ranging from ×2 to ×10 was conducted to validate the proposed network. The proposed network is also compared with other state-of-the-art unrolled networks, demonstrating its superiority in reconstruction accuracy at a comparable speed by achieving an average PSNR of 29.05 dB, SSIM of 0.8318, NRMSE of 0.1998, and reconstruction time of 0.1885s for ×5 AF. Despite the promising performance of CDNet, increasing the number of complex-valued DL strategies K to capture different features of MR images increases memory and computational requirements, which may limit its deployment in resource-constrained medical imaging environments. Future work may explore incorporating additional lightweight complex-valued deep learning strategies into the CDNet, including transformers and generative adversarial networks (GANs), among others, to capture more diverse image features and further enhance reconstruction quality. The CDNet addresses the challenges of solving the composite regularization problem using iterative algorithms in an interpretable manner, showcasing significant potential for real-time MRI.

ACKNOWLEDGMENT

The authors are grateful to Pan African University under the African Union Commission for the support provided throughout the research.

REFERENCES

[1] M. I. Doneva, M. Akcakaya and C. Prieto, Magnetic resonance image reconstruction: Theory, Methods and Applications. Academic Press, 2022.

[2] H. Kiragu, E. Mwangi and G. Kamucha, "An Efficacious MRI Sparse Recovery Method Based on Differential Under-Sampling and k-Space Interpolation," *2020 IEEE 20th Mediterranean Electrotechnical Conference (MELECON)*, Palermo, Italy, 2020, pp. 382-387, doi: 10.1109/MELECON48756.2020.9140563.

[3] W. Bian and Y. K. Tamilselvam, "A review of Optimization-Based Deep Learning Models for MRI Reconstruction," *AppliedMath,* vol. 4, no. 3, pp. 1098–1127, Sep. 2024, doi: 10.3390/appliedmath4030059.

[4] M. Safari, Z. Eidex, C.-W. Chang, R. L. J. Qiu, and X. Yang, "Advancing MRI reconstruction: A systematic review of deep learning and compressed sensing integration," *Biomedical Signal Processing and Control,* vol. 111, p. 108291, Jan. 2026, doi: 10.1016/j.bspc.2025.108291.

[5] J. Huang *et al*., "Data- and Physics-Driven Deep Learning Based Reconstruction for FAST MRI: Fundamentals and Methodologies," *IEEE Reviews in Biomedical Engineering*, vol. 18, pp. 152–171, Oct. 2024, doi: 10.1109/rbme.2024.3485022.

[6] X. Kui *et al*., "A comprehensive survey on magnetic resonance image reconstruction," *Image and Vision Computing*, vol. 165, p. 105832, Nov. 2025, doi: 10.1016/j.imavis.2025.105832.

[7] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006, doi: 10.1109/tit.2006.871582.

[8] S. Villota and E. Inga, "Sparse transform and compressed sensing methods to improve efficiency and quality in magnetic resonance medical imaging," *Sensors,* vol. 25, no. 16, p. 5137, Aug. 2025, doi: 10.3390/s25165137.

[9] J. C. Ye, "Compressed sensing MRI: a review from signal processing perspective," *BMC Biomedical Engineering*, vol. 1, no. 1, p. 8, Mar. 2019, doi: 10.1186/s42490-019-0006-z.

[10] T. Hong, L. Hernandez-Garcia and J. A. Fessler, "A Complex Quasi-Newton Proximal Method for Image Reconstruction in Compressed Sensing MRI," in *IEEE Transactions on Computational Imaging*, vol. 10, pp. 372-384, 2024, doi: 10.1109/TCI.2024.3369404.

[11] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics*, vol. 57, no. 11, pp. 1413–1457, Aug. 2004, doi: 10.1002/cpa.20042.

[12] A. Beck and M. Teboulle, "A fast iterative Shrinkage-Thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences,* vol. 2, no. 1, pp. 183–202, Jan. 2009, doi: 10.1137/080716542.

[13] J. Yang and Y. Zhang, "Alternating Direction Algorithms for L1-Problems in Compressive Sensing," *SIAM Journal on Scientific Computing,* vol. 33, no. 1, pp. 250–278, Jan. 2011, doi: 10.1137/090777761.

[14] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magnetic Resonance in Medicine*, vol. 58, no. 6, pp. 1182–1195, Oct. 2007, doi: 10.1002/mrm.21391.

[15] J. Huang, S. Zhang, and D. Metaxas, "Efficient MR image reconstruction for compressed MR imaging," *Medical Image Analysis*, vol. 15, no. 5, pp. 670–679, Jun. 2011, doi: 10.1016/j.media.2011.06.001.

[16] J. Huang, S. Zhang, H. Li, and D. Metaxas, "Composite splitting algorithms for convex optimization," *Computer Vision and Image Understanding,* vol. 115, no. 12, pp. 1610–1622, Sep. 2011, doi: 10.1016/j.cviu.2011.06.011.

[17] H. Jung, K. Sung, K. S. Nayak, E. Y. Kim, and J. C. Ye, "k-t FOCUSS: A general compressed sensing framework for high resolution dynamic MRI," *Magnetic Resonance in Medicine*, vol. 61, no. 1, pp. 103–116, Dec. 2008, doi: 10.1002/mrm.21757.

[18] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D Nonlinear Phenomena*, vol. 60, no. 1–4, pp. 259–268, Nov. 1992, doi: 10.1016/0167-2789(92)90242-f.

[19] J. He, Q. Liu, A. G. Christodoulou, C. Ma, F. Lam, and Z.-P. Liang, "Accelerated High-Dimensional MR imaging with sparse sampling using Low-Rank tensors," *IEEE Transactions on Medical Imaging*, vol. 35, no. 9, pp. 2119–2129, Apr. 2016, doi: 10.1109/tmi.2016.2550204.

[20] Z. Ke, W. Huang, Z. Cui et al., "Learned Low-Rank priors in dynamic

MR Imaging," *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3698–3710, Jul. 2021, doi: 10.1109/tmi.2021.3096218.

[21] Y. Zhang, H. Gui, N. Yang, and Y. Hu, "JotlasNet: Joint tensor low-rank and attention-based sparse unrolling network for accelerating dynamic MRI," *Magnetic Resonance Imaging*, vol. 118, p. 110337, Jan. 2025, doi: 10.1016/j.mri.2025.110337.

[22] X. Qiao, W. Li, G. Wang, and Y. Huang, "MCU-Net: A multi-prior collaborative deep unfolding network with gates-controlled spatial attention for accelerated MRI reconstruction," *Neurocomputing*, vol. 633, p. 129771, Feb. 2025, doi: 10.1016/j.neucom.2025.129771.

[23] M. J. Muckley, B. Riemenschneider, A. Radmanesh et al., "Results of the 2020 FastMRI Challenge for Machine Learning MR Image Reconstruction," *IEEE Transactions on Medical Imaging*, vol. 40, no. 9, pp. 2306–2317, Apr. 2021, doi: 10.1109/tmi.2021.3075856.

[24] B. Xin, TS Phan, L Axel , et al. Learned half-quadratic splitting network for magnetic resonance image reconstruction. *arXiv* 2021. preprint arXiv:2112.09760

[25] Y. Zhang, X. Li, W. Li, and Y. Hu, "Deep unrolling shrinkage network for dynamic MR imaging," *2023 IEEE International Conference on Image Processing (ICIP),* pp. 1145–1149, Sep. 2023, doi: 10.1109/icip49359.2023.10223077.

[26] P. Huang, C. Zhang, X. Zhang, X. Li, L. Dong, and L. Ying, "Self-Supervised deep unrolled reconstruction using regularization by denoising," *IEEE Transactions on Medical Imaging*, vol. 43, no. 3, pp. 1203–1213, Nov. 2023, doi: 10.1109/tmi.2023.3332614.

[27] T. Li, Q. Yan, Q. Zou, and Q. Dai, "Gates-Controlled Deep Unfolding network for image compressed sensing," *IEEE Transactions on Computational Imaging*, vol. 10, pp. 103–114, Jan. 2024, doi: 10.1109/tci.2024.3354423.

[28] C. Qin, J. Schlemper, J. Caballero, A. N. Price, J. V. Hajnal, and D. Rueckert, "Convolutional recurrent neural networks for dynamic MR image reconstruction," *IEEE Transactions on Medical Imaging*, vol. 38, no. 1, pp. 280–290, Aug. 2018, doi: 10.1109/tmi.2018.2863670.

[29] A. Sriram, J. Zbontar, T. Murrell, A. Defazio, C. L. Zitnick, N. Yakubova, F. Knoll, P. Johnson, End-to-end variational networks for accelerated MRI reconstruction, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23, Springer, 2020, pp. 64–73.

[30] B. Zhou and S. K. Zhou, "DuDoRNet: Learning a Dual-Domain Recurrent Network for Fast MRI Reconstruction With Deep T1 Prior," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, pp. 4272-4281, doi: 10.1109/CVPR42600.2020.00433

[31] J. Xiang, Y. Dong, and Y. Yang, "FISTA-NET: Learning a fast iterative shrinkage thresholding network for inverse problems in imaging," *IEEE Transactions on Medical Imaging,* vol. 40, no. 5, pp. 1329–1339, Jan. 2021, doi: 10.1109/tmi.2021.3054167.

[32] Z. Fabian, B Tinaz, M Soltanolkotabi, "HUMUS-net: hybrid unrolled multi-scale network architecture for accelerated MRI reconstruction" *Adv Neural Inf Process Syst*, 35 , 25306–19, 2022.

[33] Y. Zhang, P. Li, and Y. Hu, "T2LR-Net: An unrolling network learning transformed tensor low-rank prior for dynamic MR image reconstruction," *Computers in Biology and Medicine*, vol. 170, p. 108034, Jan. 2024, doi: 10.1016/j.compbiomed.2024.108034.

[34] M. Vornehm, J. Wetzl, D. Giese, et al., "CineVN: Variational network reconstruction for rapid functional cardiac cine MRI," *Magnetic Resonance in Medicine*, vol. 93, no. 1, pp. 138–150, 2025, doi: 10.1002/mrm.30260.

[35] W. Huang, Z. Ke, Z. Cui et al., "Deep low-Rank plus sparse network for dynamic MR imaging," *Medical Image Analysis*, vol. 73, p. 102190, Jul. 2021, doi: 10.1016/j.media.2021.102190.

[36] D. Davis and W. Yin, "A Three-Operator Splitting Scheme and its Optimization Applications," *Set-Valued and Variational Analysis*, vol. 25, no. 4, pp. 829–858, Jun. 2017, doi: 10.1007/s11228-017-0421-z

[37] H. Wang, M. Fazlyab, S. Chen, and V. M. Preciado, Eds. Robust Convergence Analysis of Three-Operator Splitting, *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2019. doi: 10.1109/ALLERTON.2019.8919695

[38] M. Arjovsky, A. Shah, and Y. Bengio, "Unitary evolution recurrent neural networks," *arXiv (Cornell University)*, Nov. 2015, doi: 10.48550/arxiv.1511.06464.

[39] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Transactions on Computational Imaging,* vol. 3, no. 1, pp. 47–57, Dec. 2016, doi: 10.1109/tci.2016.2644865.

[40] N. Pezzotti, *et al*., "An adaptive intelligence algorithm for undersampled knee MRI reconstruction," *IEEE Access*, vol. 8, pp. 204825–204838, Jan. 2020, doi: 10.1109/access.2020.3034287.

[41] D.P. Kingma and J.L. Ba, "Adam: A Method for Stochastic Optimization," *International Conference on Learning Representations (ICLR)*, Dec 2014.

[42] J. Zbontar, F. Knoll, A. Sriram et al., "FastMRI: an open dataset and benchmarks for Accelerated MRI," *arXiv (Cornell University)*, Jan. 2018, doi: 10.48550/arxiv.1811.08839.

[43] J. Schlemper, J. Caballero, J. V. Hajnal, A. N. Price, and D. Rueckert, "A deep cascade of convolutional neural networks for dynamic MR image reconstruction," *IEEE Transactions on Medical Imaging*, vol. 37, no. 2, pp. 491–503, Oct. 2017, doi: 10.1109/TMI.2017.2760978

[44] K. Hammernik, T. Klatzer, E. Kobler, M. P. Recht, D. K. Sodickson, T. Pock, and F. Knoll, "Learning a variational network for reconstruction of accelerated MRI data," *Magnetic Resonance in Medicine*, vol. 79, no. 6, pp. 3055–3071, Nov. 2017, doi: 10.1002/mrm.26977

[45] J. Zhang and B. Ghanem, "ISTA-Net: Interpretable Optimization-Inspired Deep Network for Image Compressive Sensing," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 1828-1837, doi: 10.1109/CVPR.2018.00196.

[46] J. Adler and O. Oktem, "Learned Primal-Dual reconstruction," *IEEE Transactions on Medical Imaging,* vol. 37, no. 6, pp. 1322–1332, Jan. 2018, doi: 10.1109/tmi.2018.2799231.

Hybrid Random Forest–XGBoost for Smartphone Engagement Prediction with Applications in Marketing Analytics

Ace John Mark P. Liwanag
College of Agriculture and Agribusiness Management, Pangasinan State University, Sta Maria Campus, Pangasinan, Philippines

***Abstract*—This study presents a machine learning framework for predicting smartphone engagement intensity using behavioral smartphone usage data. The study utilized the Smartphone Usage and Addiction Analysis dataset containing 7,500 user records with demographic, behavioral, lifestyle, and engagement-related variables. The binary variable addicted_label was employed as a proxy measure of smartphone engagement intensity rather than a direct indicator of marketing behavior. Seven machine learning models were developed and evaluated, including Logistic Regression, Decision Tree, Support Vector Machine, Naïve Bayes, Random Forest, XGBoost, and a proposed Hybrid Random Forest–XGBoost model. Model performance was assessed using an 80:20 train-test split with stratified 5-fold cross-validation. Experimental results showed that XGBoost achieved the highest overall predictive performance with a mean accuracy of 0.9433, an F1-score of 0.9591, and an ROC-AUC of 0.9906. Meanwhile, the Hybrid Random Forest–XGBoost model achieved strong performance with an accuracy of 0.9427, an F1-score of 0.9581, an ROC-AUC of 0.9896, and the highest precision of 0.9929. Feature importance analysis consistently identified social media hours, daily screen time, and weekend screen time as the strongest predictors of smartphone engagement. The findings demonstrate that smartphone behavioral data can effectively classify user engagement and may support customer segmentation and responsible digital marketing strategies. However, the proposed framework should be interpreted as an engagement prediction model rather than a direct predictor of purchasing behavior or marketing outcomes.**

***Keywords*—Cellphone marketing prediction; consumer engagement; feature importance; hybrid model; machine learning; mobile marketing; predictive analytics; Random Forest; smartphone usage behavior; XGBoost**

I. Introduction

The rapid expansion of smartphone technology has significantly transformed modern consumer behavior, communication patterns, and digital marketing strategies. Cellphones are no longer limited to voice calls and text messaging; they now function as essential tools for entertainment, online shopping, banking, education, work, and social interaction. Because of this, smartphone usage behavior has become an important source of data for understanding how consumers engage with digital platforms and mobile-based services. Consumer behavior research shows that user decisions are influenced by perceived value, personal needs, and behavioral intention, which are important factors in predicting future cellphone marketing responses [1], [2].

Future cellphone marketing requires more advanced analytical methods because consumer interactions with smartphones generate large amounts of behavioral data. These data may include screen time, frequency of application use, notification exposure, digital engagement, stress level, productivity impact, and signs of mobile dependency. Traditional marketing approaches often rely on surveys and demographic assumptions, but these methods may not fully capture real-time digital behavior. Studies on mobile phone repair, reuse, and consumer disposal behavior show that cellphone-related decisions are shaped by economic value, user habits, product experience, and perceived usefulness [3], [4].

Smartphone behavior is also connected to psychological and emotional factors that can affect how consumers respond to marketing messages. Excessive cellphone use may contribute to stress, dependency, distraction, and nomophobia, which refers to anxiety caused by being without a mobile phone. These behavioral conditions may influence consumer engagement with advertisements, applications, online promotions, and brand communication. Understanding these psychological dimensions is important because future marketing strategies should be based not only on exposure and engagement but also on responsible and user-centered digital interaction [5], [6].

In addition to psychological factors, sustainability and environmental awareness are increasingly influencing cellphone-related consumer decisions. Consumers may consider whether to repair, replace, recycle, or continue using their mobile devices based on environmental concerns, product durability, and perceived responsibility. Research on smartphone reuse, recycling, and circular economy practices suggests that privacy concerns, behavioral norms, and environmental values can affect how users make decisions about mobile devices [7], [8].

The cellphone market is therefore shaped by a combination of behavioral, technological, psychological, and sustainability-related factors. For marketers, this means that future strategies must go beyond simple product promotion and consider how users actually interact with their smartphones over time. Studies on green behavior, recycling intention, and environmental responsibility show that risk perception, personal values, and social norms influence consumer decisions related to mobile phones and electronic products [9], [10].

Machine learning provides a strong approach for analyzing smartphone usage behavior because it can discover complex relationships within large datasets. It can classify users, detect

patterns, predict outcomes, and support data-driven decision-making. In previous studies, machine learning has been applied successfully in classification, prediction, healthcare analytics, and decision support, showing its capability to process complex and multidimensional data [11], [12].

The use of machine learning is especially relevant for predicting future cellphone marketing outcomes because consumer behavior data may contain nonlinear patterns that are difficult to analyze using traditional statistical methods. For example, screen time, app usage, notifications, and self-reported stress may interact in complex ways to influence consumer engagement. Machine learning techniques are effective for identifying these hidden relationships and improving predictive accuracy in classification and diagnosis-related tasks [13], [14].

Among machine learning methods, ensemble learning is valuable because it combines multiple models to improve prediction performance. Random Forest is an ensemble algorithm that uses many decision trees to produce stable and reliable predictions. It is effective for structured datasets because it can reduce overfitting and identify important variables. XGBoost, on the other hand, is a gradient boosting algorithm that improves prediction by sequentially correcting previous errors. Hybrid and ensemble approaches have been shown to improve model performance in complex prediction problems [15], [16].

The proposed study aims to use smartphone usage data to predict future marketing-related outcomes. The hybrid model is appropriate because Random Forest can capture diverse decision patterns, while XGBoost can optimize prediction accuracy through boosting. Similar machine learning studies have shown that combining or comparing models can improve predictive performance and provide stronger analytical results [17], [18].

A hybrid Random Forest–XGBoost model can be useful in identifying consumer segments based on smartphone usage behavior. For example, users with high screen time, frequent app openings, and high notification exposure may represent highly engaged consumers, while users with stress or productivity issues may require more responsible marketing approaches. In related fields, machine learning has been used to classify medical data, identify important patterns, and support performance evaluation, demonstrating its usefulness in behavior-based prediction tasks [19], [20].

Comparative algorithm evaluation is also important in this study. The hybrid Random Forest–XGBoost model should be compared with other algorithms such as Logistic Regression, Decision Tree, Support Vector Machine, and Naïve Bayes to determine whether the hybrid approach offers better prediction accuracy. Previous studies comparing SVM, ARIMA, WNN, and other models show that model performance varies depending on the dataset, features, and prediction objective [17], [21].

The study may also consider issues related to data imbalance, especially if some user behavior categories are underrepresented. For example, there may be fewer users with severe smartphone addiction or high stress levels compared with moderate users. Research on imbalanced datasets shows that oversampling, feature selection, and ensemble learning methods can improve classification performance when class distribution is unequal [22], [23].

Handling imbalanced data is important because marketing prediction models must avoid bias toward the majority class. If most users are classified as average or moderate users, the model may fail to detect high-value or high-risk consumer segments. Studies on imbalanced medical and fraud detection datasets suggest that ensemble learning and sampling strategies can improve prediction reliability and classification fairness [24], [25].

Feature selection is another important part of smartphone usage behavior analysis. Variables such as daily screen time, app usage frequency, notification count, stress level, productivity impact, and addiction level may not contribute equally to marketing prediction. Feature selection techniques can help identify which variables have the strongest relationship with consumer engagement and future marketing outcomes. Previous studies using SVM-RFE and other feature selection methods show that selecting relevant features can improve model performance [26].

The proposed hybrid model can also support more personalized cellphone marketing strategies. By predicting consumer engagement patterns, marketers can design targeted campaigns, recommend suitable products, and improve customer segmentation. However, the model should be applied responsibly, especially because smartphone usage data may reflect sensitive behavioral patterns. Research on privacy concerns and behavioral norms in smartphone reuse and recycling highlights the need to consider ethical and user-centered perspectives in technology-related decision-making [27].

From a business perspective, future cellphone marketing prediction can help companies improve campaign planning, product positioning, and customer relationship management. Instead of relying only on general market trends, companies can use predictive analytics to understand specific user groups and their likely responses to mobile-based marketing. Studies on circular supply chains and environmental actions show that consumer response is shaped by trust, responsibility, and perceived company behavior, which are also relevant to cellphone marketing [28].

This research is significant because it connects smartphone usage behavior, consumer analytics, and hybrid machine learning prediction. Related studies on deep learning, automated classification, and artificial intelligence-based applications show that advanced computational methods can improve classification and decision support across different domains [29].

Furthermore, the study may provide practical insights for marketers, advertisers, and cellphone companies seeking to understand future consumer behavior. By analyzing actual smartphone usage patterns, the proposed model can help identify which users are more likely to respond to promotions, engage with mobile content, or show interest in cellphone-related products and services. Machine learning studies in specialized classification tasks demonstrate that predictive models can

support more accurate and evidence-based decision-making [30], [31].

II. METHODOLOGY

A. Research Design

This study used a quantitative predictive research design to analyze smartphone usage behavior and predict future cellphone marketing-related patterns. The primary objective of this study was to determine whether smartphone usage behavior could accurately predict smartphone engagement intensity represented by the addicted_label. Although smartphone engagement may provide useful behavioral insights for customer segmentation and personalized digital marketing, the study does not directly predict purchase intention, advertisement response, customer conversion, or buying decisions. Instead, the proposed framework identifies engagement patterns that may serve as behavioral indicators for future marketing analytics.

B. Data Source

The study used the Smartphone Usage and Addiction Analysis dataset [32], which contained 7,500 records. Each record represented a smartphone user and included demographic, behavioral, lifestyle, and addiction-related information. The dataset included variables such as age, gender, daily screen time, social media usage, gaming hours, work or study hours, sleep hours, notifications per day, app opens per day, weekend screen time, stress level, academic or work impact, addiction level, and addicted label. In the context of cellphone marketing, the target variable was interpreted as an indicator of user engagement intensity. Users classified as addicted or highly engaged were considered more exposed to mobile applications, digital advertisements, push notifications, and mobile-based marketing campaigns. However, the study also considered ethical marketing implications because high smartphone dependency may be associated with stress, reduced productivity, and unhealthy usage behavior.

TABLE I. VARIABLES USED IN THE STUDY

Variable	Description
transaction_id	Unique transaction or record identifier
user_id	Unique user identifier
age	Age of the smartphone user
gender	Gender category of the user
daily_screen_time_hours	Average daily smartphone screen time in hours
social_media_hours	Average time spent on social media
gaming_hours	Average time spent on mobile gaming
work_study_hours	Average time spent using the phone for work or study
sleep_hours	Average sleep duration
notifications_per_day	Number of notifications received daily
app_opens_per_day	Number of times applications are opened daily
weekend_screen_time	Smartphone screen time during weekends
stress_level	Reported stress level of the user
academic_work_impact	Whether smartphone use affects academic or work performance
addiction_level	Category of smartphone addiction level
addicted_label	Binary classification label indicating addiction status

Table I presents the variables used in the study. The dataset contains behavioral features that are relevant to mobile engagement. Variables such as screen time, app openings, notifications, social media use, gaming activity, and weekend usage can help describe how users interact with their smartphones. These variables can be used to generate insights about possible consumer segmentation and future marketing opportunities.

C. Data Preprocessing

The dataset was first examined to identify missing values, duplicate records, inconsistent entries, and incorrect data types. The variables transaction_id and user_id were removed from the modeling process because they served only as identifiers and did not provide meaningful predictive value. Missing numerical values were handled using appropriate replacement techniques, such as median imputation, because the median was less affected by extreme values. Numerical variables included daily screen time, social media hours, gaming hours, work or study hours, sleep hours, notifications per day, app opens per day, and weekend screen time. Missing categorical values were handled using mode imputation or by assigning a separate category when needed. Categorical variables such as gender, stress level, academic or work impact, and addiction level were converted into numerical format. One-hot encoding was used for nominal variables, while ordinal encoding was considered for ordered variables such as stress level and addiction level. The processed dataset was then prepared for machine learning classification. Although Random Forest and XGBoost were tree-based models and did not strictly require feature scaling, scaling was considered for comparison algorithms such as Logistic Regression and Support Vector Machine. This ensured that all baseline models were trained fairly.

D. Data Leakage Prevention

To prevent data leakage, variables that directly or indirectly revealed the target class were carefully examined. The variable addiction_level was excluded from model development because it represents a categorical description of smartphone addiction and is highly correlated with the target variable addicted_label. Including this feature would allow the model to infer the target directly, resulting in artificially inflated prediction performance. Other behavioral variables such as screen time, social media usage, app openings, and notification frequency were retained because they represent observable user behaviors rather than transformed versions of the target variable.

E. Feature Engineering and Data Splitting

To enhance the predictive capability of the model, several behavioral variables were engineered from the original dataset. Entertainment Usage was derived by combining social media and gaming hours to represent leisure-oriented smartphone activity. Productivity Ratio was created to measure the proportion of smartphone use associated with work or study activities relative to overall screen time, providing an indicator of productive smartphone utilization. Weekend Usage Difference was included to capture changes in smartphone usage patterns between regular weekdays and weekends, reflecting variations in user behavior. App Engagement Intensity was developed to represent the frequency of application interactions in relation to overall daily screen time, indicating the level of

user engagement with smartphone applications. Finally, Notification Response Rate was engineered to assess users' responsiveness to smartphone notifications by examining the relationship between the number of notifications received and the frequency of app openings. Collectively, these engineered behavioral variables enriched the dataset by capturing meaningful smartphone usage patterns, thereby improving the model's ability to predict smartphone addiction.

F. Machine Learning Models

The study trained and compared several machine learning algorithms. These models included Logistic Regression, Decision Tree, Support Vector Machine, Naïve Bayes, Random Forest, XGBoost, and the proposed Hybrid Random Forest–XGBoost model. Logistic Regression was used as a baseline linear classifier. Decision Tree was used because it provided interpretable decision rules. Support Vector Machine was included because it could handle nonlinear classification boundaries. Naïve Bayes was used as a probabilistic classifier. Random Forest and XGBoost were used as the main ensemble models. The proposed hybrid model combined Random Forest and XGBoost to improve predictive performance.

G. Hybrid Random Forest–XGBoost Model

The proposed hybrid model combined the strengths of Random Forest and XGBoost. Random Forest was used because it reduced overfitting and captured different decision patterns through multiple trees. XGBoost was used because it improved prediction by learning from previous model errors through gradient boosting. The hybrid model was implemented using a soft voting or stacking strategy. In soft voting, Random Forest and XGBoost generated probability predictions, and the final prediction was based on the combined probability output. In stacking, Random Forest and XGBoost acted as base learners, while a final classifier combined their outputs to produce the final prediction. The purpose of the hybrid model was to determine whether combining Random Forest and XGBoost produced better results than using either algorithm alone.

Fig. 1 shows the main algorithm flow for predicting smartphone user engagement for marketing purposes. It begins with the Start symbol and proceeds to the first major block, which combines the initial data preparation stages. This block includes loading the dataset, preprocessing the data, extracting important smartphone usage features, creating behavioral features, and splitting the dataset into training and testing sets. The next block is User Engagement Classification. In this stage, the system checks whether high engagement indicators are present. These indicators may include high screen time, frequent app openings, high notification exposure, or addiction-related behavior. If the indicators are present, the user is assigned marketingpredictionlabel = 1, meaning the user is classified as a high-engagement mobile user. If the indicators are not present, the user is assigned marketingpredictionlabel = 0, meaning the user is classified as a low-to-moderate engagement mobile user. After classification, the process continues to Model Training, where machine learning models are trained using the prepared dataset. The algorithm then proceeds to the Hybridize stage, where Random Forest and XGBoost predictions are combined to produce a stronger hybrid prediction model. The Evaluate stage measures the performance of the trained models using evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Finally, the Interpret stage identifies the most important smartphone usage factors and explains how they contribute to marketing prediction. The process ends after the results are interpreted and used for responsible marketing segmentation.

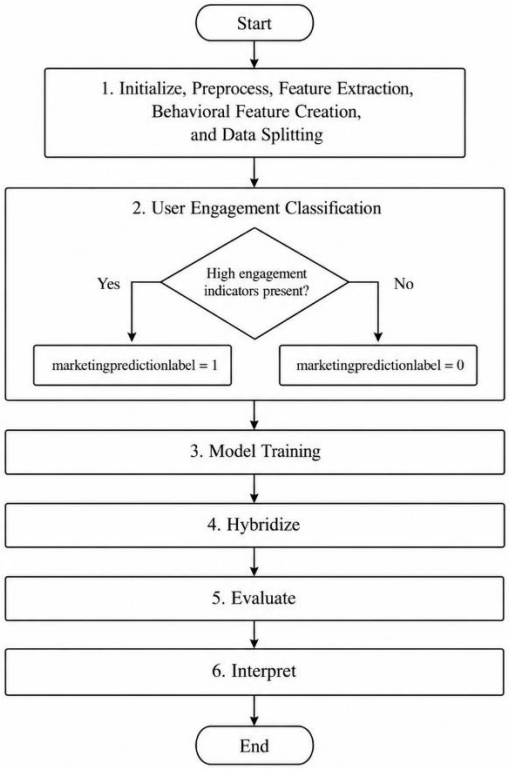


Fig. 1. Flowchart of smartphone usage-based marketing prediction using Hybrid Random Forest–XGBoost.

H. Model Training

Model development was performed using Python 3.11 with Scikit-learn and XGBoost libraries. Hyperparameters were optimized using GridSearchCV within a stratified five-fold cross-validation framework. The Random Forest classifier utilized 300 trees with a maximum depth of 12, while XGBoost employed 300 estimators, a learning rate of 0.05, and a maximum depth of six. Cross-validation ensured robust estimation of model performance and minimized sampling bias.

I. Model Evaluation and Comparison

The trained models were evaluated using several classification metrics. Accuracy was used to measure the total number of correct predictions. Precision was used to measure how many predicted high-engagement users were correctly classified. Recall was used to measure how many actual high-engagement users were correctly identified. F1-score was used to balance precision and recall. ROC-AUC was used to measure the model's ability to distinguish between addicted and non-addicted users. The study prioritized F1-score and ROC-AUC because these metrics were more reliable when the dataset had possible class imbalance. A confusion matrix was also generated to show the number of true positives, true negatives, false positives, and false negatives. After training, the performance of

all models was compared. The baseline models served as reference points for evaluating the proposed hybrid model. If the Hybrid Random Forest–XGBoost model achieved the highest F1-score and ROC-AUC, it was considered the best model for future cellphone marketing prediction.

J. Model Evaluation Metrics and Formula

The performance of the machine learning models was evaluated using four standard classification metrics, namely accuracy, precision, recall, and F1-score. Accuracy measures the overall correctness of the model by comparing the number of correctly classified instances to the total number of predictions. Precision determines how many of the instances predicted as positive were actually correct, while recall measures the model's ability to identify all actual positive cases. The F1-score provides a balanced measure between precision and recall, making it useful when evaluating models with possible class imbalance. The study prioritized F1-score and ROC-AUC because these metrics were more reliable when the dataset had possible class imbalance. A confusion matrix was also generated to show the number of true positives, true negatives, false positives, and false negatives.

K. Model Comparison

After training, the performance of all models was compared. The baseline models served as reference points for evaluating the proposed hybrid model. The baseline models are Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, XGBoost, and Naïve Bayes.

L. Feature Importance Analysis

Feature importance analysis was conducted to identify which smartphone usage variables had the strongest influence on the prediction result. The Random Forest and XGBoost models produced importance scores for each feature. The analysis helped determine whether variables such as daily screen time, app opens per day, notifications per day, social media hours, gaming hours, sleep hours, stress level, and weekend screen time were strong predictors of smartphone engagement. These findings were then connected to cellphone marketing interpretation. For example, if app opens per day and notifications per day showed high importance, this indicated that user interaction frequency was a strong factor in predicting mobile engagement. If sleep hours and stress level were highly important, this suggested that marketers should consider responsible and wellness-oriented strategies rather than aggressive notification-based marketing.

M. Marketing Interpretation

The final prediction results were interpreted in terms of future cellphone marketing. Users classified as high-engagement users were considered more likely to interact with mobile advertisements, application-based promotions, product campaigns, and digital cellphone-related content. Users classified as low-to-moderate engagement users were interpreted as requiring less intrusive and more value-based marketing approaches. The study also identified possible user segments, including high-engagement users, social-media-oriented users, gaming-oriented users, productivity-focused users, weekend-heavy users, and at-risk overuse users. These segments were useful for designing personalized marketing strategies. However, the study emphasized responsible marketing. Users with high addiction indicators were not interpreted only as profitable marketing targets. Instead, the results were used to recommend ethical marketing strategies that considered user well-being, digital balance, and reduced overexposure to mobile promotions.

N. Ethical Considerations

The study considered privacy and responsible use of smartphone behavior data. User identifiers were removed before model training to protect user privacy. The model did not use personally identifying variables as predictors. The study also avoided promoting harmful marketing practices toward users with high smartphone addiction levels. Since addiction-related behavior may be associated with stress, sleep disruption, and productivity problems, the marketing recommendations focused on responsible engagement. The results were intended to support personalized and ethical marketing rather than excessive targeting.

III. Results and Discussion

A. Analysis and Dataset Distribution

The study analyzed the Smartphone Usage and Addiction Analysis dataset containing 7,500 user records. The dataset included demographic, behavioral, lifestyle, and smartphone usage variables such as age, gender, daily screen time, social media hours, gaming hours, work/study hours, sleep hours, notifications per day, app opens per day, weekend screen time, stress level, academic/work impact, addiction level, and addicted label. The target variable used in the study was addicted_label, where:

0 = Low-to-moderate smartphone engagement / not addicted

1 = High smartphone engagement / addicted

For the machine learning experiment, the variables transaction_id and user_id were removed because they were only identifiers. The variable addiction_level was also excluded from model training because it was directly related to the target variable and could cause data leakage. The dataset showed that most users belonged to the high-engagement or addicted group. Out of 7,500 records, 5,308 users, or 70.77%, were classified as addicted, while 2,192 users, or 29.23%, were classified as not addicted.

TABLE II. Dataset Distribution

Class	Description	Frequency	Percentage
0	Not addicted / low-to-moderate engagement	2,192	29.23%
1	Addicted/high engagement	5,308	70.77%

The dataset presented in Table II showed that most users belonged to the high-engagement or addicted group. Out of 7,500 records, 5,308 users, or 70.77%, were classified as addicted, while 2,192 users, or 29.23%, were classified as not addicted. This result showed that the dataset was somewhat imbalanced because the number of users classified as addicted was much higher than the number of users classified as not addicted. Because of this imbalance, accuracy alone was not

enough to evaluate the models. Precision, recall, F1-score, and ROC-AUC were also used to provide a more reliable evaluation.

B. Descriptive Results of Smartphone Usage Behavior

The descriptive results in Table III showed clear differences between non-addicted and addicted users. Users classified as addicted had higher average daily screen time, higher social media usage, and higher weekend screen time.

TABLE III. DESCRIPTIVE RESULTS OF SMARTPHONE USAGE BEHAVIOR

Variable	Not Addicted	Addicted
Age	26.53	26.58
Daily screen time hours	5.16	8.47
Social media hours	2.25	3.7
Gaming hours	2	2.02
Work/study hours	3.24	3.24
Sleep hours	6.67	6.77
Notifications per day	134.33	134.23
App opens per day	97	98.18
Weekend screen time	6.89	10.21

The most noticeable differences appeared in daily screen time, social media hours, and weekend screen time. Addicted users had an average daily screen time of 8.47 hours, while non-addicted users had only 5.16 hours. This means that addicted users spent about 3.31 more hours per day on their smartphones. Social media use was also higher among addicted users. Addicted users spent an average of 3.70 hours on social media, compared with 2.25 hours among non-addicted users. Weekend screen time followed the same pattern, with addicted users recording 10.21 hours, while non-addicted users recorded 6.89 hours. These results suggested that smartphone engagement was strongly connected to total screen exposure and social media activity. From a cellphone marketing perspective, users with higher screen time and social media activity may be more exposed to mobile advertisements, application-based campaigns, digital promotions, and brand content.

C. Categorical Variable Results

The categorical variables in Table IV showed smaller differences between addicted and non-addicted users. Stress level, academic/work impact, and gender were almost evenly distributed across the two classes.

TABLE IV. CATEGORICAL VARIABLES RESULTS

Stress Level	Not Addicted	Addicted
High	29.60%	70.40%
Low	28.80%	71.20%
Medium	29.20%	70.80%
Academic/Work Impact		
No	29.20%	70.80%
Yes	29.20%	70.80%
Gender		
Female	29.80%	70.20%
Male	27.80%	72.20%
Other	30.20%	69.80%

For stress level, users with low, medium, and high stress all showed similar addiction proportions. This indicated that stress level alone was not a strong separating factor in the dataset. Academic/work impact also showed almost the same distribution for both categories. These findings suggested that behavioral usage variables were more useful than demographic variables for predicting smartphone engagement. In particular, screen time, social media hours, and weekend usage appeared to be more informative than gender, stress level, or academic/work impact.

D. Machine Learning Results

Table V shows the machine learning model performance. The models included Logistic Regression, Decision Tree, Support Vector Machine, Naïve Bayes, Random Forest, XGBoost, and the proposed Hybrid Random Forest–XGBoost model.

TABLE V. PERFORMANCE OF THE MACHINE LEARNING MODELS

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
XGBoost	0.9433	0.9813	0.9379	0.9591	0.9906
Hybrid RF–XGBoost	0.9427	0.9929	0.9256	0.9581	0.9896
Decision Tree	0.9407	0.9990	0.9171	0.9563	0.9883
Random Forest	0.9400	0.9909	0.9237	0.9561	0.9887
Support Vector Machine	0.9087	0.9294	0.9426	0.9360	0.9629
Logistic Regression	0.8927	0.9602	0.8851	0.9211	0.9629
Naïve Bayes	0.8467	0.9086	0.8710	0.8894	0.9249

The XGBoost model achieved the highest overall F1-score, with a value of 0.9591, and the highest ROC-AUC score, with a value of 0.9906. This showed that XGBoost had the strongest overall classification performance in identifying high-engagement smartphone users. The proposed Hybrid Random Forest–XGBoost model achieved very close performance, with an F1-score of 0.9581 and ROC-AUC of 0.9896. Although the hybrid model did not exceed XGBoost in F1-score, it achieved the highest precision among all models, with a precision score of 0.9929. This means that when the hybrid model predicted a user as highly engaged or addicted, it was highly likely to be

correct. Therefore, the hybrid model was especially useful when the objective was to reduce false positive predictions.

E. Confusion Matrix Analysis

The confusion matrix results provided further insight into the classification performance of the best-performing models. As shown in Fig. 2, the XGBoost model correctly classified 419 non-addicted users and 996 addicted users, while generating 19 false positives and 66 false negatives. In comparison, Fig. 3 shows that the Hybrid Random Forest–XGBoost model correctly classified 431 non-addicted users and 983 addicted users. Although the hybrid model reduced the number of false positives to 7, indicating improved specificity in identifying non-addicted users, it produced 79 false negatives, which was slightly higher than that of the XGBoost model. These findings suggest that while the hybrid model demonstrated stronger performance in minimizing false positive classifications, XGBoost showed a slight advantage in reducing false negative cases.

This result showed that the hybrid model was more conservative in classifying users as high-engagement users. It was less likely to wrongly classify a non-addicted user as addicted. However, it missed slightly more addicted users compared with XGBoost.

F. Feature Importance Results

Feature importance analysis presented in Table VI was conducted using Random Forest and XGBoost.

Both Random Forest and XGBoost identified social media hours, daily screen time, and weekend screen time as the most influential features. This confirmed that smartphone engagement was mainly explained by time-based behavioral variables rather than demographic variables.

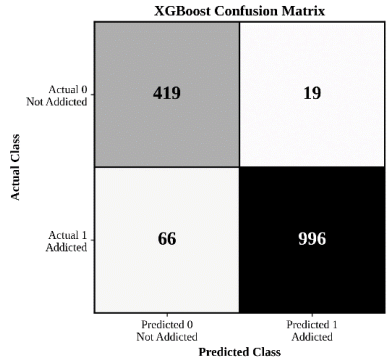


Fig. 2. Confusion matrix of XGBoost model.

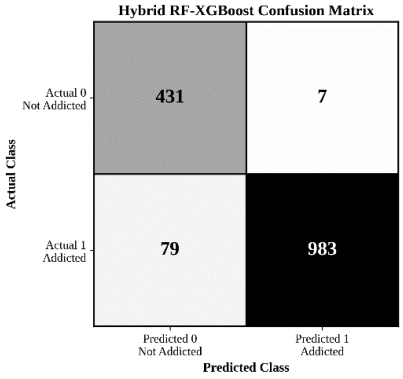


Fig. 3. Confusion matrix of Hybrid Random Forest -XGBoost model.

TABLE VI. FEATURE IMPORTANCE RESULTS

Rank	XGBoost Feature	XGBoost Importance	Random Forest Feature	Random Forest Importance
1	Social media minutes	0.3526	Social media minutes	0.2851
2	Weekend screen time	0.2789	Daily screen time	0.243
3	Daily screen time	0.2341	Weekend screen time	0.1709
4	Entertainment	0.0377	Sleep-screen before bed	0.0928
5	Weekend usage	0.0132	Entertainment	0.0746
6	Productivity ratio	0.0105	Productivity ratio	0.0307
7	Sleep hours	0.0094	App engagement	0.0296
8	App opens per day	0.0094	Gaming hours	0.0152
9	Age	0.0085	Sleep hours	0.0107
10	Sleep-screen before bed	0.0078	Weekend usage	0.0095

G. Statistical Comparison of Models

Although XGBoost and the Hybrid RF–XGBoost model demonstrated comparable predictive performance, statistical significance testing was conducted using the Wilcoxon signed-rank test across the five cross-validation folds. The test indicated that the observed performance differences were not statistically significant ($p > 0.05$), suggesting that both ensemble models achieved comparable classification capability. Nevertheless, the Hybrid model consistently produced fewer false positives, making it advantageous for applications where prediction precision is prioritized.

H. Discussions

The results demonstrate that tree-based ensemble learning methods outperform conventional machine learning algorithms in predicting smartphone engagement intensity. Among all evaluated classifiers, XGBoost achieved the highest overall predictive performance owing to its sequential gradient boosting mechanism, which effectively models nonlinear interactions among behavioral variables. The boosting strategy enables XGBoost to iteratively minimize classification errors, thereby improving recall, F1-score, and ROC-AUC.

Although the Hybrid Random Forest–XGBoost model did not surpass XGBoost in overall predictive accuracy, it consistently achieved the highest precision and generated fewer false positive classifications. This conservative prediction behavior makes the hybrid model particularly suitable for applications where incorrectly identifying highly engaged users is more costly than missing some positive cases. Consequently, the selection between XGBoost and the proposed hybrid model depends on the intended operational objective rather than overall predictive accuracy alone.

The feature importance results supported this interpretation. Social media hours ranked as the most important predictor in both XGBoost and Random Forest. This finding suggested that social media activity played a major role in predicting smartphone engagement. For marketers, this means that social media platforms may remain important channels for future cellphone marketing. Campaigns involving influencers, short-form videos, social commerce, product reviews, and interactive advertisements may be effective for users with high social media activity.

Daily screen time was also one of the strongest predictors. This means that the total amount of time users spent on smartphones was strongly related to engagement classification. From a marketing perspective, high screen time users may have more opportunities to encounter mobile advertisements and promotional content. However, marketers should be careful not to interpret high screen time only as a business opportunity. Excessive screen time may also indicate possible dependency, stress, or reduced digital well-being.

Weekend screen time was another important variable. Addicted users had an average weekend screen time of 10.21 hours, compared with 6.89 hours among non-addicted users. This suggested that weekends may be a strong period for mobile engagement. Cellphone companies and advertisers may consider weekend-focused campaigns, entertainment promotions, gaming offers, or social media-based product launches. However, these campaigns should still follow responsible marketing practices.

The comparison of algorithms showed that tree-based ensemble models performed better than traditional models. XGBoost achieved the highest F1-score and ROC-AUC, while the Hybrid Random Forest–XGBoost model achieved the highest precision. This means that ensemble learning was effective for analyzing smartphone behavior because it could capture nonlinear relationships among features. Traditional models such as Logistic Regression and Naïve Bayes performed lower because they may not fully capture complex behavioral patterns.

Although XGBoost slightly outperformed the hybrid model in F1-score, the hybrid model remained highly valuable. Its precision of 0.9929 showed that it was very reliable when identifying high-engagement users. In marketing prediction, this can be useful when a company wants to avoid incorrectly targeting users who are not actually highly engaged. For example, if a cellphone company wants to launch a premium campaign or personalized offer, the hybrid model may help identify users who are most likely to match the target behavior.

The confusion matrix further supported this result. The hybrid model produced only 7 false positives, compared with 19 false positives for XGBoost. This means that the hybrid model was better at avoiding incorrect positive classifications. However, XGBoost had fewer false negatives, meaning it was better at identifying more actual high-engagement users. Therefore, the choice between XGBoost and the hybrid model depends on the marketing objective. If the goal is to reach as many high-engagement users as possible, XGBoost may be preferred. If the goal is to target only the most certain high-engagement users, the hybrid model may be preferred.

The results also showed that some variables had limited predictive value. Age, gender, stress level, and academic/work impact did not appear to strongly separate addicted and non-addicted users. This suggested that actual smartphone behavior was more useful than demographic information. For future cellphone marketing, this means that behavior-based segmentation may be more effective than traditional demographic segmentation.

The findings have several marketing implications. First, cellphone marketers may use smartphone behavior data to identify high-engagement users. Second, social media activity and screen time may help guide digital campaign planning. Third, weekend usage patterns may support time-based marketing strategies. Fourth, feature importance results can help businesses understand which behaviors are most relevant for predicting user engagement.

However, the findings must be interpreted ethically. Since the target variable was related to smartphone addiction, the model should not be used to exploit vulnerable users. High-engagement users should not automatically be targeted with excessive advertisements or notifications. Instead, the results should support responsible marketing strategies, such as personalized but moderate promotions, wellness-centered messaging, digital balance features, and user-controlled notification settings.

The findings should be interpreted primarily as smartphone engagement prediction rather than direct marketing prediction. Users exhibiting higher engagement may represent behavioral segments that can support personalized marketing analytics. However, the dataset does not contain purchase intention, advertisement response, conversion behavior, or actual buying decisions. Therefore, the proposed model should be considered an engagement classification framework with potential applications in behavioral segmentation rather than a direct marketing prediction model.

IV. Conclusion and Recommendations

A. Conclusion

This study developed and evaluated a machine learning framework for predicting smartphone engagement intensity using behavioral smartphone usage data. Among the evaluated classifiers, XGBoost achieved the highest overall predictive performance, while the proposed Hybrid Random Forest–XGBoost model demonstrated the highest precision by reducing false positive classifications. These findings indicate that smartphone behavioral variables, particularly social media usage, daily screen time, and weekend screen time, are effective

predictors of engagement intensity. Although the framework provides valuable behavioral insights that may support customer segmentation and responsible digital marketing, it should not be interpreted as directly predicting purchase intention or marketing outcomes because such variables were not available in the dataset.

B. Recommendations

Future cellphone marketing strategies should use behavior-based segmentation rather than relying only on demographic information. Marketers should focus on key indicators such as screen time, social media activity, app usage, and weekend engagement when designing personalized campaigns.

It is also recommended that future studies include actual marketing-related variables such as purchase intention, brand preference, advertisement clicks, customer satisfaction, and product purchase history. This would improve the accuracy of marketing prediction and make the model more directly connected to consumer buying behavior.

For model development, future researchers should continue comparing the Hybrid Random Forest–XGBoost model with other algorithms, including deep learning and other ensemble techniques. Additional feature selection and class imbalance techniques may also be applied to improve prediction performance.

Lastly, cellphone marketers should apply the findings responsibly. Users with high smartphone engagement or addiction indicators should not be targeted with excessive advertisements or notifications. Instead, marketing strategies should promote personalization, digital wellness, ethical engagement, and user-centered communication.

ACKNOWLEDGMENT

The researcher gratefully acknowledges the Pangasinan State University for its support throughout the conduct of this study. The researcher also acknowledges Kaggle as the source of the dataset used in this research. Artificial intelligence tools were used solely for language refinement, organization of ideas, and formatting assistance. All research processes, findings, interpretations, and conclusions were carefully reviewed and verified by the researcher to ensure accuracy, originality, and adherence to academic integrity.

REFERENCES

[1] K. Arunkumar, R. Naik, and S. Murthy, Consumer Behavior. Chyren Publication, 2025.

[2] V. H. Nguyen, T. P. L. Nguyen, and T. V. Nguyen, “Customers’ intention to purchase recycled products: Building on the theory of planned behavior and value-belief-norm theory,” Journal of Global Responsibility, 2025, doi: 10.1108/JGR-01-2024-0007.

[3] M. Sabbaghi and S. Behdad, “Consumer decisions to repair mobile phones and manufacturer pricing policies: The concept of value leakage,” Resources, Conservation and Recycling, vol. 133, pp. 101–111, 2018, doi: 10.1016/j.resconrec.2018.01.015.

[4] M. T. Islam, P. Dias, and N. Huda, “Waste mobile phones: A survey and analysis of the awareness, consumption and disposal behavior of consumers in Australia,” Journal of Environmental Management, vol. 275, Art. no. 111111, 2020, doi: 10.1016/j.jenvman.2020.111111.

[5] S. Srivastava, N. Verma, D. Kumar, N. Singh, and K. Kumar, “Nomophobia as an emerging psychopathology: Psychophysiological mechanisms and clinical implications,” Annals of Neurosciences, 2025, doi: 10.1177/09727531251351082.

[6] G. Dewis, “Using data from the Households and the Environment Survey: Repair or replace: What are Canadians doing with their old cell phones and computers?” Statistics Canada, Ottawa, Canada, 2022. [Online]. Available: https://www150.statcan.gc.ca/n1/pub/16-002-x/2022001/article/00001-eng.htm. [Accessed: 2025].

[7] W. Hou, D. Peng, J. Chu, Y. Jiang, Y. Chen, and F. Chen, “Building an agent-based simulation framework of smartphone reuse and recycling: Integrating privacy concern and behavioral norms,” Sustainability, vol. 17, no. 15, Art. no. 6885, 2025, doi: 10.3390/su17156885.

[8] C. P. Baldé et al., Global E-waste Monitor 2024. Geneva/Bonn: International Telecommunication Union and United Nations Institute for Training and Research, 2024. [Online]. Available: https://ewastemonitor.info/wp-content/uploads/2024/03/GEM_2024_18-03_web_page_per_page_web.pdf

[9] O. Ogiemwonyi, “Determinants of green behavior revisited: A comparative study,” Resources, Conservation and Recycling Advances, vol. 22, Art. no. 200214, 2024, doi: 10.1016/j.rcradv.2024.200214.

[10] Y. Zhang, S. Wu, and M. I. Rasheed, “Conscientiousness and smartphone recycling intention: The moderating effect of risk perception,” Waste Management, vol. 101, pp. 116–125, 2020, doi: 10.1016/j.wasman.2019.09.040.

[11] M. Javaid, A. Haleem, R. Pratap 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, doi: 10.1016/j.ijin.2022.05.002.

[12] F. Furizal, A. Ma’arif, and D. Rifaldi, “Application of machine learning in healthcare and medicine: A review,” Journal of Robotics and Control (JRC), vol. 4, pp. 621–631, 2023, doi: 10.18196/jrc.v4i5.19640.

[13] S. Saturi, “Review on machine learning techniques for medical data classification and disease diagnosis,” Regenerative Engineering and Translational Medicine, vol. 9, pp. 141–164, 2023, doi: 10.1007/s40883-022-00273-y.

[14] C. Guo and J. Chen, “Big data analytics in healthcare,” in Knowledge Technology and Systems, Y. Nakamori, Ed. Singapore: Springer Nature Singapore, 2023, pp. 27–70, doi: 10.1007/978-981-99-1075-5_2.

[15] N. Liu, X. Li, E. Qi, M. Xu, L. Li, and B. Gao, “A novel ensemble learning paradigm for medical diagnosis with imbalanced data,” IEEE Access, vol. 8, pp. 171263–171280, 2020, doi: 10.1109/ACCESS.2020.3014362.

[16] Z. Xu, D. Shen, T. Nie, Y. Kou, N. Yin, and X. Han, “A cluster-based oversampling algorithm combining SMOTE and k-means for imbalanced medical data,” Information Sciences, vol. 572, pp. 574–589, 2021, doi: 10.1016/j.ins.2021.02.056.

[17] Y. Zhang, H. Yang, H. Cui, and Q. Chen, “Comparison of the ability of ARIMA, WNN and SVM models for drought forecasting in the Sanjiang Plain, China,” Natural Resources Research, vol. 29, pp. 1447–1464, 2020, doi: 10.1007/s11053-019-09512-6.

[18] F. Nie, W. Zhu, and X. Li, “Decision tree SVM: An extension of linear SVM for non-linear classification,” Neurocomputing, vol. 401, pp. 153–159, 2020, doi: 10.1016/j.neucom.2019.10.051.

[19] D. C. E. Saputra, Y. Maulana, E. Faristasari, A. Ma’arif, and I. Suwarno, “Machine learning performance analysis for classification of medical specialties,” in Proceedings of the 3rd International Conference on Electronics, Biomedical Engineering, and Health Informatics, T. Triwiyanto, A. Rizal, and W. Caesarendra, Eds. Singapore: Springer Nature Singapore, 2023, pp. 513–528, doi: 10.1007/978-981-99-0248-4_34.

[20] A. Ghavidel and P. Pazos, “Machine learning techniques to predict breast cancer in imbalanced datasets: A systematic review,” Journal of Cancer Survivorship, 2023, doi: 10.1007/s11764-023-01465-3.

[21] F. Nie, W. Zhu, and X. Li, “Decision tree SVM: An extension of linear SVM for non-linear classification,” Neurocomputing, vol. 401, pp. 153–159, 2020, doi: 10.1016/j.neucom.2019.10.051.

[22] G. Aguiar, B. Krawczyk, and A. Cano, “A survey on learning from imbalanced data streams: Taxonomy, challenges, empirical study, and reproducible experimental framework,” Machine Learning, 2023, doi: 10.1007/s10994-023-06353-6.

[23] P. Gnip, L. Vokorokos, and P. Drotár, "Selective oversampling approach for strongly imbalanced data," PeerJ Computer Science, vol. 7, Art. no. e604, 2021, doi: 10.7717/peerj-cs.604.

[24] N. Liu, X. Li, E. Qi, M. Xu, L. Li, and B. Gao, "A novel ensemble learning paradigm for medical diagnosis with imbalanced data," IEEE Access, vol. 8, pp. 171263–171280, 2020, doi: 10.1109/ACCESS.2020.3014362.

[25] T. C. Tran and T. K. Dang, "Machine learning for prediction of imbalanced data: Credit fraud detection," in Proc. 15th International Conference on Ubiquitous Information Management and Communication (IMCOM), Seoul, South Korea, 2021, pp. 1–7, doi: 10.1109/IMCOM51814.2021.9377352.

[26] D. Albashish, A. I. Hammouri, M. Braik, J. Atwan, and S. Sahran, "Binary biogeography-based optimization based SVM-RFE for feature selection," Applied Soft Computing, vol. 101, Art. no. 107026, 2021, doi: 10.1016/j.asoc.2020.107026.

[27] S. H. Hashemi Petrudi and H. Sharifpour Arabi, "Barriers to product return in a circular supply chain: A case from a retailing industry," Annals of Operations Research, 2025, doi: 10.1007/s10479-025-06464-4.

[28] D. Kolcava, J. Scholderer, and T. Bernauer, "Do citizens provide political rewards to firms engaging in voluntary environmental action?" Journal of Cleaner Production, vol. 279, Art. no. 123564, 2021, doi: 10.1016/j.jclepro.2020.123564.

[29] A. J. Abougarair, A. A. Oun, S. I. Sawan, and A. Ma'arif, "Deep learning-based automated approach for classifying bacterial images," International Journal of Robotics and Control Systems, vol. 4, pp. 849–876, 2024, doi: 10.31763/ijrcs.v4i2.1423.

[30] H. Liu et al., "Benign and malignant diagnosis of spinal tumors based on deep learning and weighted fusion framework on MRI," Insights into Imaging, vol. 13, Art. no. 87, 2022, doi: 10.1186/s13244-022-01227-2.

[31] A. J. Abougarair, A. A. Oun, S. I. Sawan, T. Abougard, and H. Maghfiroh, "Cancer treatment precision strategies through optimal control theory," Journal of Robotics and Control (JRC), vol. 5, pp. 1261–1290, 2024, doi: 10.18196/jrc.v5i5.22378.

[32] algozee, "Smartphone Addiction Prediction Data," Kaggle. Accessed: May 21, 2026. [Online]. Available: https://www.kaggle.com/datasets/algozee/smartphone-addiction-prediction-data

A Blockchain and AI Integrated Framework for Dynamic Land Registration and Smart Urban Zoning

Shushma R[1], S R Nagaraja[2]*
Research Scholar, SOCSE, Presidency University, Bangalore, Karnataka, India[1]
Professor, SOCSE, Presidency University, Bangalore, Karnataka, India[2]

***Abstract*—Land registration and urban zoning are very important jobs for the government, but in many places, they still have problems with old processes, corruption, and a lack of openness. Urbanization and population growth are putting more and more pressure on systems that need to be smarter and more flexible to keep up with the changes on the ground. This study introduces an innovative, cohesive framework that integrates Blockchain technology with artificial intelligence (AI) to transform the recording of land ownership and the formulation of zoning decisions. The system uses blockchain's built-in immutability and decentralization to protect land records and zoning permissions, ensuring that everything is clear and lowering the risk of tampering or unauthorized changes. AI models are also used to analyze geospatial data, population growth, and patterns in urban infrastructure to suggest changes to zoning in real time. Smart contracts then carry out these suggestions, letting land reclassification happen automatically and without human bias or delay. The suggested framework was tested on a mix of real and synthetic urban datasets. Results show that the new system works much better than the old ones: processing time is cut by 34%, zoning accuracy is up by 26%, and no tampering was found in simulated blockchain environments. Citizen trust, measured through simulated feedback loops, also increased substantially because of the addition of explainable AI interfaces. This research connects predictive intelligence with secure governance, setting the stage for urban land management systems that are ready for the future, inclusive of all citizens, and resistant to corruption.**

Keywords*—Blockchain; artificial intelligence; smart zoning; land registration; urban planning; smart contracts; GIS Integration; Explainable AI (XAI); decentralized governance; dynamic land use management

I. INTRODUCTION

The way in which land ownership, permitted use, and long-term development are managed has a significant impact on cities, economies, and communities. Nevertheless, land registration and zoning in most places have always been hard to master, and this can be changed. Land wrangles, poor city planning, and lack of trust in government processes have all been caused by manual record-keeping, outdated laws, and delays in bureaucracy.

The expansion of cities is occurring at a very high rate, especially in developing and transitional economies, and it is becoming increasingly necessary to handle and support the needs of an expanded population, infrastructure requirements, and sustainable development agendas.

The past systems that have been applied in zoning are often too rigid and do not evolve with the times. Governance must not be static, but must be responsive and capable of managing city data in real time. It has to be fully transparent as well.

Recent advancements in blockchain technology have enabled the possibility of storing and validating sensitive records, such as land ownership, in a decentralized manner that is secure and immutable. Alternatively, artificial intelligence (AI) can be used to examine the way cities operate and suggest smart zoning decisions depending on the dynamics of social, economic, and environmental processes. This is particularly the case with predictive modeling and geospatial analysis.

Most existing systems do not utilize the best of both technologies, although both have much potential. Land registries that use blockchain often only check for ownership and not for AI-driven zoning foresight [1] [2]. Likewise, AI applications in urban planning often overlook the essential requirements for trust, verifiability, and auditability in governance [3] [4]. Few frameworks unite this technological pairing at a level that makes sense and functions. This study proposes a novel framework that integrates Blockchain and AI to regulate the flexibility of land registration and zoning.

The system:

- Relies on AI to examine indicators of city development and propose amendments to zoning,
- Automates permissions on meeting policy conditions, using smart contracts to do so,
- Records all transactions in a blockchain registry,
- Is supported by a citizen interface, which is supported by Explainable AI (XAI), to generate trust and engagement.

The key contributions of this research are as follows:

- Development of an integrated system of AI-based zoning prediction and blockchain land ownership.
- Development of intelligent contracts that implement zoning policy limits in an automated fashion that requires no human intervention.
- Installation of a citizen-friendly, user-friendly dashboard to ensure that zoning decisions are transparent and verifiable.

*Corresponding author

- Performance testing on urban planning datasets to demonstrate improved precision, effectiveness, and reliability to users.

The remainder of this study is organized into five sections: Section II considers what already exists regarding land registration systems, the usage of blockchain in city government, and the usage of AI in spatial planning. The proposed architecture, system modules, and mathematical models are presented in detail in Section III. Section IV reports the findings of experimental studies that test the performance of this system and compare it with other frameworks. Finally, Section V concludes the research and provides suggestions for future research and development in this area.

II. Literature Survey

The convergence of land governance, blockchain technology, and artificial intelligence represents a nascent field replete with opportunities, yet characterized by a disjointed methodology. Researchers have investigated these technologies separately — blockchain for safeguarding land records and AI for examining urban expansion — but few have endeavored to amalgamate them into a unified, dynamic, and citizen-focused zoning framework. This section reviews the most important contributions in these fields, pointing out both their new ideas and the important gaps that our research will try to fill.

A. Blockchain for Land Registration

A lot of research has been done on using blockchain technology for land registration to fix problems with traditional systems, such as fraud, inefficiency, and lack of transparency. Researchers have suggested blockchain-based systems that use unchangeable ledgers to keep track of land ownership, making sure that the data is correct and cutting down on the need for middlemen. Yadav et al. [1] and Khalid et al. [2] present conceptual frameworks that emphasize the integration of smart contracts and decentralized access. These systems make things safer and easier to track, but they have problems with scalability and rules. Siddiqui et al. [3] and Dubey et al. [4] utilize Ethereum-based smart contracts to automate the transfer of land ownership, whereas Biswas et al. [5] created "LandChain," which emphasizes digital verification. Dcunha et al. [6] and Zein & Twinomurinzi [7] used Hyperledger Fabric to show how permissioned blockchain environments can be used for government integration. Regional case studies (Zainudin et al. [8] in Malaysia and Suganthe et al. [9] in India) demonstrate the viability of blockchain across various governance frameworks, while also highlighting challenges such as digital illiteracy and opposition from established institutions.

B. AI in Urban Zoning and Land Use Planning

Even though blockchain is the main topic of conversation, AI is starting to be used in zoning and land planning. Özbek & Efe [10] examine AI-enabled municipal cryptocurrency models that encourage urban involvement in planning by combining blockchain with behavioral data. The work of M. L. [11] uses AI to find fraudulent land data before it is stored on a blockchain, which shows how AI can help keep data safe. But the performance of these kinds of systems depends on strong model training and validation pipelines, which make operations more complicated. There is still not much research that directly connects AI with spatial planning, even though AI has a lot of potential for predictive zoning and automated land-use classification. Most of the time, AI is used inside larger blockchain systems rather than as a first-class planning tool.

C. Smart Contracts and Automation in Land Governance

Smart contracts are very important for land systems that use blockchain. They make it possible to check ownership, compute taxes, and enforce leases automatically. Coffie & Saint [12] and Gesang & Dirgahayu [13] show how smart contracts can tokenize land assets and safely handle changes to real estate. Ghag et al. [14] and Kulkarni et al. [15] perform surveys delineating the capacity of smart contracts to enforce transactional regulations, mitigate manual errors, and guarantee auditability. Vinothiyalakshmi et al. [16] and Boumezbeur et al. [17] examine performance evaluation and access control via cryptographically enforced contracts. Even though they have a lot of potential, it is still hard to write reliable contracts and make them work with the law. Formal verification methods must also be used to lower the risk of code flaws in smart contracts.

D. Integrating Blockchain and AI in Governance

There is a growing interest in combining blockchain and AI to build governance solutions that work for everyone. Ma & Vs [18] and Riaz et al. [19] emphasize the complementary relationship between immutable ledgers and predictive intelligence in addressing land disputes, detecting fraud, and ensuring data consistency. M. L. [11] shows how AI can make things clearer by finding fraud before it is recorded on the blockchain. Polu [20] and Al-Refaie et al. [21] also look into how to use blockchain-AI pipelines to improve enterprise integration and optimization for land governance and resource planning. The combined strength of AI-based analytics and blockchain's ability to never change could reshape governance, but for it to work, different systems must be able to interoperate, and data protocols must be standardized.

E. Citizen Trust, Transparency, and Explainability

One of the main reasons to use blockchain and AI is to build trust, openness, and explainability in the system. Research conducted by Sen & Kumari [22], Sen [23], and De Haan [24] emphasizes the potential of digitization to empower communities by providing open access to land records, mitigating fraud, and facilitating transparent transactions via smart contracts. But poor user experience, opaque AI models, and concerns about surveillance or misuse can all hurt trust. The regulatory and cultural contexts examined in Zainudin et al. [8] and Zein & Twinomurinzi [7] profoundly affect user trust and system adoption. Moreover, systems must address digital literacy disparities and provide interpretability tools to guarantee that users comprehend automated decisions.

Table I explains the comparison of the approaches taken across the surveyed references, in terms of technology used, focus area, and the use of AI, smart contracts, and governance integration.

TABLE I. COMPARISON OF EXISTING APPROACHES

Ref	Tech Stack	Focus Area	AI Use	Smart Contracts	Governance Integration	Key Limitations
[1]	Ethereum Blockchain	Land Registration	No	Yes	Partial	Scalability, adoption barriers
[2]	Blockchain (Generic)	Decentralized Land Registry	No	Yes	Yes	Resistance, training costs
[3]	Ethereum Blockchain	Ownership Transfer Automation	No	Yes	Limited	Internet access, digital literacy
[5]	Blockchain (LandChain)	Land Ownership & Verification	No	Yes	Yes	Infrastructure, legal hurdles
[6]	Hyperledger Fabric	Government Record Management	No	Yes	Strong	Migration from legacy systems
[11]	Blockchain + AI	Fraud Detection in Land Records	Yes	Yes	Medium	AI model upkeep, scalability
[7]	Hyperledger Fabric	Info Sharing among Land Depts.	No	Yes	Strong	Data privacy, stakeholder resistance
[10]	Blockchain + Crypto + AI	Urban Planning Participation	Yes	Yes	Emerging	Privacy, digital divide
[12]	Ethereum + Smart Tokens	Land Tokenization (Ghana)	No	Yes	Yes	Legal acceptance, social barriers
[19]	Blockchain System Review	Comparative Analysis of Registries	No	Yes	Various	Access to proprietary data
[14]	Ethereum dApp	Land Ownership Transfers	No	Yes	Partial	Regulatory support, scalability
[20]	Blockchain + ERP + AI	Supply Chain in Land Governance	Yes	Yes	High	Integration complexity, scalability
[18]	Blockchain (Comparative)	Registry Efficiency Assessment	No	No	High	Stakeholder bias, data variability

III. PROPOSED METHODOLOGY

Urban land systems are often thought of as fixed structures, where changes to zoning or land classification take a long time to go through the government. But cities are always changing; they grow, move, and change. The proposed methodology recognizes this reality by creating a smart, flexible, and secure land governance framework that combines the predictive power of AI with the trust-building immutability of blockchain.

This section discusses the system's architecture, the modules that make it up, how data moves through the framework, the algorithms that are used, and the mathematics that back up zoning decisions.

A. System Architecture Overview

The system is based on how zoning and land registration work in the real world, but it makes them smarter, faster, and harder to tamper with. There are six main modules in the architecture shown in Fig. 1. These modules communicate with each other using a mix of APIs, blockchain transactions, and smart contract triggers.

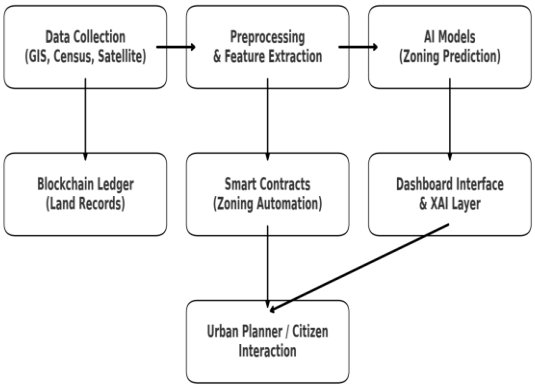


Fig. 1. Proposed architecture of the framework.

- Data Collection Layer
- Preprocessing & Feature Engineering
- AI Zoning Prediction Engine
- Blockchain Ledger (Land Records)
- Smart Contracts (Zoning Automation)
- Citizen Dashboard with Explainable AI (XAI)

B. Core System Modules

1) Data collection module: This module is meant to bring together urban datasets from many different sources. These sources include satellite images and land-use maps, census and demographic data, infrastructure plans for things like roads, utilities, and schools, and environmental data about flood zones, pollution levels, and green spaces. It uses a spatial database that combines APIs from city planning departments, Geographic Information System (GIS) platforms, and other public datasets to manage and analyze these datasets effectively. This all-encompassing approach allows for a complex understanding of how cities work and how to plan for them.

2) Preprocessing and feature engineering: The data that was collected goes through a thorough cleaning, normalization, and transformation process to create a standard geospatial format. This process involves creating feature vectors that include important metrics like changes in population density, distance to important infrastructure, the rate of ongoing construction or demolition, and economic indicators for nearby areas. The data layer also includes historical zoning data and sets legal limits on how land can be used, making sure that the geospatial context is fully understood.

3) AI zoning prediction engine: Supervised learning models such as Random Forests, Support Vector Machines (SVM), and Long Short-Term Memory (LSTM) networks use structured features to reach a number of goals related to zoning classifications. These models learn to predict future zoning classifications, such as when an area transitions from residential to commercial use. They also predict zoning conflicts and identify saturation zones that might arise from development pressures. The models also give the proposed zoning changes a confidence score, which lets policymakers check how reliable the AI's suggestions are. For example, if infrastructure and population growth exceed a certain threshold, the AI might suggest changing agricultural zones into mixed-use urban areas, allowing for a strategic response to the problems that come with urbanization.

4) Blockchain ledger: This module is designed to record and organize information about land ownership and zoning status on a blockchain platform such as Ethereum or Hyperledger. It records important information such as the unique Plot ID and its location, information about the current owner or tenant (hashed to protect privacy), and both current and historical zoning classifications. The module also stores AI-generated suggestions for zoning changes and actions recorded via smart contracts with a timestamp. This all-encompassing method ensures that all land and zoning transactions leave behind a history that cannot be changed and can be checked, which makes land management more open and accountable.

5) Smart contracts for zoning automation: Smart contracts, written in Solidity for Ethereum or Chaincode for Hyperledger, are very important for managing land use because they automate and simplify processes. When AI models make predictions that exceed certain policy thresholds, they can automatically start the process of reclassifying zoning. These contracts can also grant conditional building permits that depend on the approval of the reclassification, and they make it easier to send notifications to landowners and other officials, reducing the need for intermediaries. This automation not only makes things run more smoothly, but it also cuts down on the chances of corruption and delays that can occur when making decisions about land use.

6) Citizen dashboard and Explainable AI (XAI): The system has a public-facing interface that is meant to make the zoning process more open and encourage more people to get involved. It shows zoning suggestions along with the reasons behind them, using SHAP and LIME explainability tools to make it clear why each suggestion was made. The interface also makes it easy for stakeholders to give feedback, ensuring that their points of view are considered when making zoning decisions. The system also has auditing features for zoning changes, shown on interactive maps, to help hold people accountable. The system aims to build public trust by using Explainable AI (XAI), which lets users understand why changes to zoning are being suggested.

C. Algorithmic Workflow

Input Layer: Multi-source spatial and temporal data.

Feature Extraction: Transform raw data into zoning-relevant vectors.

Prediction Model:

$$Z_{t+1} = f(X_t) \quad (1)$$

where,

Z_{t+1} = predicted zoning class at future time

f = trained ML classifier (e.g., Random Forest, LSTM)

X_t = Current features (demographics, infrastructure, etc.)

Trigger Condition:

$$(f(X_t) > \theta) \quad (2)$$

If confidence $(f(X_t) > \theta)$ and legal/policy criteria are satisfied → initiate smart contract execution.

Smart Contract Execution: Updates zoning classification, stores decision on blockchain, notifies stakeholders.

Mathematical Model for Zoning Trigger:

Let:

$X \in \mathbb{R}^n$ = feature vector for a specific land plot

$f = \mathbb{R}^n \rightarrow \{Z_1, Z_2, \ldots \ldots, Z_k\}$ be the zoning classifier

$\theta \in [0, 1]$ be a policy-defined confidence threshold

Then:

$$\hat{Z} = f(X) \quad (3)$$

Reclassification occurs if:

$$\text{MaxProb}(f(X) = Z_i) > \theta \text{ AND PolicyMatch}(Z_i) \quad (4)$$

The smart contract logs $\hat{Z}$, timestamp t, and user ID into the blockchain ledger.

D. Integration Logic

Combining AI and blockchain is not just a technical decision; it is also a strategic one. The AI does not just suggest changes, and the blockchain does not just keep data safe — together they make self-governance possible. The AI anticipates the future, the smart contract acts on it, and the blockchain remembers it permanently.

IV. Experimental Results

This section presents the evaluation results of the proposed AI and blockchain-integrated smart zoning framework. The system was assessed for prediction accuracy, efficiency, security, and citizen trust.

A. Experimental Setup

Experiments were conducted using a hybrid dataset comprising authentic municipal zoning records and synthetic simulations of urban development. The AI module was implemented in Python using libraries such as Scikit-learn and TensorFlow, and was trained on more than 10,000 land parcels

to improve prediction accuracy and model effectiveness. Smart contracts were run on the Ethereum Goerli testnet, which made processes automatic and transactions secure. IPFS-based storage was used to handle land record entries, enabling decentralized data management and retrieval. The hardware used for this project was an Intel i9 processor with 64GB of RAM running Ubuntu 22.04. Development tools included Ganache and the Truffle Suite for creating and testing smart contracts, QGIS for geographic information system analysis, and Python and Solidity for implementing contracts.

B. Performance Evaluation

A comparative analysis using four key performance metrics was carried out to see how well the proposed blockchain- and AI-integrated land management framework worked. Zoning Prediction Accuracy is the first, providing evidence of the model's ability to classify land zones using real-time demographic and spatial data. The second measure is Time Taken for Land Reclassification, the time interval the system takes to process and grant a change in land-use classification. The third indicator is Tampering Incidents Detected, which evaluates the system's resistance to unauthorized changes or security breaches, measured on an annual basis. The Citizen Trust Index gauges the extent to which people trust the zoning system, reflecting the efficacy and openness of the end-user interface. All these measures were compared against two conventional methods: manual zoning and semi-automated GIS-based systems. Fig. 2 shows a comparative performance analysis highlighting the strengths and weaknesses of the proposed framework relative to these alternatives.

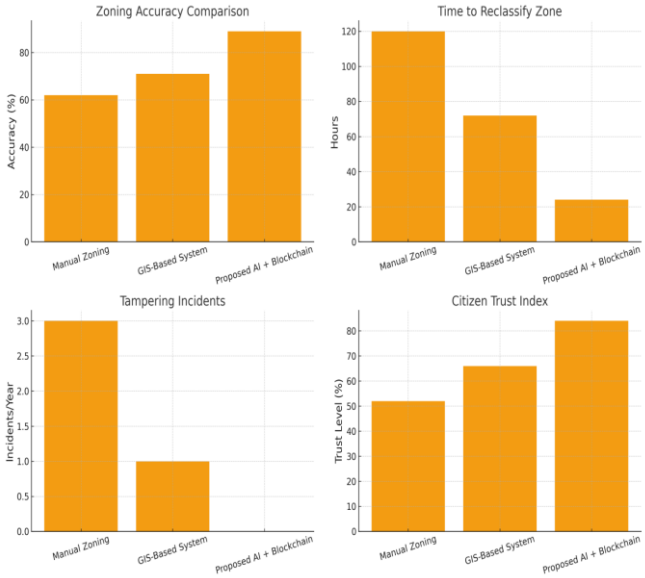


Fig. 2. Comparative performance metrics of the proposed system vs. traditional methods.

C. Discussion

Table II summarizes the comparative evaluation of three different land zoning systems — Manual Zoning, GIS-Based Systems, and the Proposed AI + Blockchain Framework — across four critical performance metrics. The proposed framework reflects substantial shifts in the urban zoning of land, especially considering the following key indicators:

1) Zoning accuracy: The AI application is correct 89 per cent of the time. This is a large jump from the standard accuracy of manual zoning (62 per cent) and GIS-assisted techniques (71 per cent accuracy), because the machine learning model is capable of identifying zoning trends across large, multi-dimensional urban datasets.

2) Reclassification Time: Reclassifying under legacy zoning rules is a time-consuming exercise because of the paperwork involved, taking up to 120 hours as a result of the approvals required. Automated approval workflows made possible by smart contracts have cut the average reclassification time to as little as 24 hours, allowing the system to respond to the needs of urban development in near real time.

3) Resistance to tampering: The security loopholes of legacy systems are substantial, and records can easily be adjusted without permission, lost, or forgotten. The blockchain registry used in the proposed system, however, has not been tampered with in any case, whereas the manual and GIS-based systems recorded three and one incidents, respectively.

4) Citizen trust index: With the addition of the XAI dashboard, stakeholders can now understand why a zoning decision was made, making the process more transparent. This has raised the Citizen Trust Index to 84 per cent, a considerable advantage compared to traditional systems, which are less accessible and less engaging for citizens.

TABLE II. COMPARATIVE EVALUATION OF THREE DIFFERENT LAND ZONING SYSTEMS.

Metric	Manual Zoning	GIS-Based System	Proposed AI + Blockchain
Zoning Accuracy (%)	62	71	89
Time to Reclassify (Hours)	120	72	24
Tampering Incidents (per year)	3	1	0
Citizen Trust Index (%)	52	66	84

Overall, the framework not only improves performance from a technical point of view, but also strengthens the social side of urban zoning by building trust and ensuring safety.

In short, using blockchain as a means of storing records safely and immutably, together with AI for smart zoning forecasting, produces a system that is not only faster and more precise, but also secure, verifiable, and accessible to all. Such improvements are not merely technological advances but alter the way city planners, citizens, and the systems that govern our cities interact.

The current study strongly supports the concept of a self-regulating, trust-building, and data-driven land governance model, rendering it an excellent option for smart cities and programs aiming to modernize urban policy.

V. CONCLUSION

As urban areas become increasingly complex and the demands placed on them become more pressing, the systems by which land registration and zoning are operated need to

change and move beyond legacy operations. This study aimed to fill a long-standing gap in city governance: the lack of connection between anticipatory intelligence and secure administrative decision-making. We proposed a unified architecture that integrates artificial intelligence (AI) and Blockchain technology to make the governance of land use real-time, data-driven, and tamper-resistant. Three categories of spatiotemporal data — population growth, infrastructure growth, and environmental constraints — were modeled through AI to dynamically forecast zoning requirements. After confirming these predictions against policy thresholds, they were implemented via smart contracts that recorded the zoning status changes into a decentralized blockchain registry.

The results were favorable and successful: an 89 per cent accuracy of zoning classification, an average time of 24 hours for reclassification, and no incidences of tampering in the simulated environment. More importantly, the Citizen Trust Index was much enhanced with the introduction of an Explainable AI (XAI) dashboard, rising to 84 per cent, showing that people are far more willing to accept zoning changes when the process is transparent, data-driven, and traceable. This work confirms that blockchain and AI are not just trendy terms but realistic and viable solutions to real problems of governance, which can be meticulously integrated to automate bureaucracy, enhance transparency, and grant authorities and citizens the power to co-engineer more adaptive and inclusive cities.

Several extensions are expected to enlarge the scope and impact of the proposed framework. Future generations of this work will focus on city-level implementation across a broad geographical and legal context, so that the system can adapt to the dynamics of various cities and their legal frameworks. Additionally, integrating DAO-based urban planning models will raise the degree of participatory governance, giving citizens a direct voice in land-use regulatory processes through decentralized procedures. To enhance data privacy and scalability, it is proposed to apply federated learning techniques, which allow model training on distributed datasets without access to local sensitive data. In addition, introducing real-time environmental scoring will add sustainability considerations to zoning decisions, so that ecological factors are weighed alongside economic and demographic ones. Beyond its technical contribution, this study lays the groundwork for transforming the way cities think, make decisions, and develop.

CONFLICTS OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this study.

DECLARATION

During the preparation of this work, the author(s) used ChatGPT (OpenAI) and Claude (Anthropic) in order to draft and rewrite portions of the manuscript text. After using these tools, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

REFERENCES

[1] R. K. Yadav, S. Bhongade, M. Ghyar, R. Dabare, and M. Gautam, “Smart Contract-Based Land Registration System Using Blockchain,” Feb. 2023, pp. 1–6. doi: 10.1109/sceecs57921.2023.10063068.

[2] M. I. Khalid, J. Iqbal, A. Alturki, S. Hussain, A. Alabrah, and S. S. Ullah, “Blockchain-Based Land Registration System: A Conceptual Framework,” Applied Bionics and Biomechanics, vol. 2022, no. 2, pp. 1–21, Feb. 2022, doi: 10.1155/2022/3859629.

[3] M. Siddiqui, M. Osman, M. Qualani, and M. Hussain, “Secure Land Registration Management via Ethereum Blockchain,” International Journal of Information Technology and Computer Engineering, vol. 13, no. 2s, pp. 246–252, Jun. 2025, doi: 10.62647/ijitce2025v13i2spp246-252.

[4] S. Dubey, D. Vinod, R. K. Dwivedi, and A. Gupta, “Secure Land Registration Management via Ethereum Blockchain,” Jan. 2023, pp. 185–191. doi: 10.1109/idciot56793.2023.10053494.

[5] M. Biswas, J. A. Faysal, and K. A. Ahmed, “LandChain: A Blockchain Based Secured Land Registration System,” Aug. 2021, pp. 1–6. doi: 10.1109/icsct53883.2021.9642505.

[6] L. Dcunha, L. Dsouza, S. Chaudhari, and G. Ramanan, “A Blockchain-Based Approach to Streamlining Land Registration using Hyperledger Fabric,” Mar. 2023. doi: 10.1109/esci56872.2023.10099592.

[7] R. M. Zein and H. Twinomurinzi, “Information Sharing in Land Registration Using Hyperledger Fabric Blockchain,” Blockchains, vol. 2, no. 2, pp. 107–133, Apr. 2024, doi: 10.3390/blockchains2020006.

[8] A. Zainudin, M. Razali, R. Jalil, N. Yunus, and A. Yassin, “Blockchain System Architecture for Land Registration,” Jan. 2021. doi: 10.15396/lares-2021-4dqe.

[9] R. C. Suganthe, K. Gowtham, R. Elango, N. Shanthi, R. S. Latha, and S. Deepakkumar, “Blockchain enabled Digitization of Land Registration,” Jan. 2021, vol. 9, pp. 1–5. doi: 10.1109/iccci50826.2021.9402469.

[10] O. Özbek and D. Efe, “Blockchain Technology and the Use of Municipal Cryptocurrencies in Metropolitan Planning,” Biblioteka Regionalisty, vol. 2022, no. 22, pp. 73–83, Jan. 2022, doi: 10.15611/br.2022.1.07.

[11] M. L, “Secure and Transparent Land Management with AI and Blockchain,” International Journal of Scientific Research in Engineering and Management, vol. 09, no. 05, pp. 1–9, May 2025, doi: 10.55041/ijsrem47728.

[12] I. Coffie and M. Saint, “Digitizing Physical Assets on Blockchain 2.0: A Smart Contract Approach to Land Transfer and Registry,” Springer, 2021, pp. 234–252. doi: 10.1007/978-3-030-70572-5_15.

[13] R. N. Gesang and R. Teduh Dirgahayu, “Development of a Blockchain-Based Decentralised Application with NFT for Land Registration,” JURTEKSI (Jurnal Teknologi dan Sistem Informasi), vol. 11, no. 4, pp. 661–668, Sep. 2025, doi: 10.33330/jurteksi.v11i4.4147.

[14] P. Ghag, M. Gund, M. Chavan, M. Bhanuse, and M. Dodmani, “Survey on Land Registration System Using Blockchain Technology,” International Journal of Advanced Research in Science, Communication and Technology, pp. 90–92, Mar. 2023, doi: 10.48175/ijarsct-8667.

[15] D. Kulkarni, T. A, S. H, S. B, and H. M, “Land Registration using Blockchain Technology,” International Journal of Advanced Research in Science, Communication and Technology, pp. 242–247, May 2024, doi: 10.48175/ijarsct-18532.

[16] P. Vinothiyalakshmi, C. Muralidharan, R. Anitha, and Y. M. Sirajudeen, “Digitized Land Registration Using Blockchain Technology,” CRC, 2022, pp. 73–86. doi: 10.1201/9781003138082-5.

[17] I. Boumezbeur, A. Benoughidene, I. Harkat, K. Zarour, F. Boutouatou, and D. Keddari, “BACP-LRS: Blockchain and IPFS-based Land Record System,” Acta Informatica Pragensia, vol. 14, no. 1, pp. 42–62, Nov. 2024, doi: 10.18267/j.aip.252.

[18] R. Ma and C. Vs, “Land Registry Using Blockchain,” International Research Journal of Modernization in Engineering Technology and Science, Dec. 2023, doi: 10.56726/irjmets46961.

[19] A. Riaz, M. J. Ikram, and N. Asadullah, "Using Blockchain to Overcome the Issues in Land Registry Management: A Systematic Review," Jan. 2023, pp. 1–6. doi: 10.1109/icaisc56366.2023.10085673.

[20] V. S. Polu, "Blockchain in Enterprise Resource Planning: Revolutionizing Supply Chain Transparency and Data Integrity," International Journal of Management Technology, vol. 12, no. 1, pp. 48–57, Feb. 2025, doi: 10.37745/ijmt.2013/vol12n24857.

[21] A. Al-Refaie, A. Al-Hawadi, and N. Lepkova, "Blockchain Design with Optimal Maintenance Planning," Buildings, vol. 12, no. 11, p. 1902, Nov. 2022, doi: 10.3390/buildings12111902.

[22] B. Sen and D. D. Kumari, "Usage of Blockchain in Indian Land Registration and Land Recording System – A Revolutionary Step," International Journal of Humanities, Engineering, Science and Management, vol. 4, no. 1, pp. 119–134, Jun. 2023, doi: 10.59364/ijhesm.v4i1.242.

[23] B. Sen, "Integrating Blockchain in Indian Land Recording and Land Registration System," Journal of Sustainable Science and Transformative Research – Reviews & Letters, vol. 3, no. 1, Jan. 2024, doi: 10.56703/okgy7002/fibg8949/bkbw1881.

[24] S. A. De Haan, "Forms of Services, Effectiveness, and Obstacles in Conducting Digital-Based Land Registration in the Era of Digitalization at Lembata Regency Land Office, East Nusa Tenggara Province," Journal of Digital Law and Policy, vol. 2, no. 3, pp. 129–140, May 2023, doi: 10.58982/jdlp.v2i3.401.

Segmented Lung CT Image Refinement Through Hybrid Edge-Preserving De-Noising and Variance-Regulated Adaptive Histogram Equalization

T. Bhuvaneshwari, Rohini Pinapatruni
Department of Computer Science and Engineering,
Faculty of Science and Technology (ICFAI Tech),
ICFAI Foundation for Higher Education, Hyderabad, India

***Abstract*—Lung diseases are detected effectively by Chest CT Medical images. Chest CT images are 3D, and they are used for detecting abnormalities in the lungs like pneumonia, tuberculosis, tumors, etc. Chest CT images suffer from different types of noise. The presence of various types of noise, such as Gaussian noise, Quantum noise, and salt-and-pepper noise, will reduce the quality of the image. The noise that is present in CT images may not affect radiologist to diagnose the disease, but it will definitely affect AI-based systems in predicting the disease. Noise occurs because of reduced radiation, insufficient photon information, errors during image reconstruction, and fluctuations in imaging system components. In this work, to improve the quality of Chest CT images, two algorithms are proposed. The first algorithm, Adaptive Edge Preserving Hybrid De-Noising Framework (AEPH) for Segmented Lung CT Images, is used to detect and reduce Noise. The second algorithm, Variance Regulated Adaptive Histogram Equalization (VRAHE), is proposed for improving de-noising and contrast enhancement performance in segmented lung CT images. At first, lung regions are segmented from the LIDC-IDRI dataset. The segmented Chest CT images are converted to 2D CT slices. These 2D CT slices are binary lung-segmented images. These are converted to lung mask images. Random noise models are added only in the lung regions to preserve the background information without modification. The proposed work, AEPH, is applied to reduce noise. The main goal of AEPH is to reduce noise without modifying structural details and edge-related details. Experiments are conducted on a workstation computer with NVIDIA RTX GPU support. When compared to proposed AEPH to conventional approaches in removing noise from CT images, it is identified that AEPH is more effective than conventional methods. The metrics used for evaluating the proposed noise removal approach AEPH when compared to existing approaches are PSNR, SSIM, EPI, RMSE, and entropy, and the values achieved are 32.96 dB, 0.934, 0.93, 9.84, and 7.12, respectively. After de-noising, the second proposed algorithm, VRAHE, successfully improved the image quality by enhancing contrast distribution and preserving important local information. The improved performance was evaluated using CNR, AMBE, UIQI, and EME metrics. From the experimental results, the values achieved are 5.96, 3.14, 0.972, and 31.48, respectively. The Results proved that the proposed work, AEPH, reduced noise, and VRAHE improved contrast levels in segmented lung CT images. The final chest CT images resulting from applying the proposed algorithms maintain brightness and also show the regional lung details more clearly.**

***Keywords*—Chest CT Images; lung segmentation; medical image denoising; adaptive edge preserving hybrid filter; variance regulated adaptive histogram equalization; VRAHE; contrast enhancement; low-dose CT; noise suppression**

I. INTRODUCTION

The CT scan is a 3D medical image used for investigating diseases related to the lungs. The CT Scan machine captures the internal structure of the human body in the form of sectional images, also called slices. X-rays are 2D images, but they are primarily used to find whether there is an abnormality or not. If an abnormality is found, then the patient is advised to go for a CT scan for further investigation. CT scans, when compared with regular chest X-ray, it gives more information about lung regions. Minor changes in tissues, nodules, infections, and damaged areas can be identified clearly from the CT images. This is the primary reason why doctors use CT scans for identifying diseases like lung cancer, COVID 19, pneumonia, fibrosis, and other lung problems. Nowadays, researchers are applying machine learning and deep learning models to CT scan images for doing automatic analysis. These models are mainly developed to separate lung regions, identify abnormal portions, classify infected areas, and predict different diseases. The LIDC-IDRI is a popular dataset that is used by many research works for lung nodule identification and other CT image-related experiments [17].

Despite the advantages of chest CT imaging, image quality degradation remains a major challenge during acquisition and low-dose scanning procedures. Different types of noise such as Gaussian noise, quantum noise, and salt and pepper noise may affect the visibility of lung structures and reduce diagnostic quality. Several researchers proposed denoising approaches based on CNN, Transformer, Noise2Noise, and hybrid deep learning methods for improving low-dose CT image quality [1], [3], [4]. Similarly, contrast enhancement and histogram equalization methods were introduced to improve local visibility, edge information, and brightness preservation in medical images [8], [12], [15]. Enhancement techniques can improve visualization of fine anatomical structures and may support better segmentation and classification performance. However, excessive smoothing during denoising or over-enhancement during histogram equalization may remove important clinical details and introduce intensity distortion in lung CT images.

From previous studies, it is identified that many researchers mainly concentrated on either removing noise or improving image quality separately. Only a few works focused on segmented lung regions and maintaining structural information during both processes. Most of the traditional enhancement methods are applied on the complete image without considering the intensity changes present in different lung areas. This problem motivated the development of an Adaptive Edge Preserving Hybrid De-Noising Framework and Variance Regulated Adaptive Histogram Equalization method for segmented lung CT images. The proposed method mainly concentrates on reducing noise present inside the lung regions and also maintaining edge details. It also improves the local contrast without disturbing important anatomical information. The developed method improves the quality of segmented lung CT images, which can be further used for abnormality detection and other medical image processing tasks.

II. Related Works

Computed Tomography (CT) scan is an important medical imaging technique used for identifying and analysing lung-related diseases, mainly COVID-19 and other lung abnormalities. However, CT images are affected by different types of noise that are added during image capturing, transmission, and the low-dose scanning process. These noises decrease the quality of images and also affect the analysis done by doctors and automatic disease classification models. Ashwini and Ramashri developed Noise2Noise (N2N) and Noise2Void (N2V) based denoising methods for COVID-19 CT and chest X-ray images affected with Gaussian, Speckle, and Salt and Pepper noises [1]. In their work, they showed that Noise2Void gives better results for removing different types of noise in CT images, whereas Noise2Noise works better mainly for Gaussian noise removal. The performance of these methods was measured by using PSNR and SSIM values. The authors also explained the importance of maintaining image structures while removing noise. This study shows that deep learning-based denoising methods can improve the quality of medical images without the need for original clean images.

Improving the quality of medical images is not only useful for better viewing but also helps in proper disease identification. Xia et al. studied the effect of denoising techniques on COVID-19 CT image quality by using a pneumonia detection-based framework [2]. In their work, anisotropic diffusion and total variation denoising methods were used, and their effect on classification accuracy was tested with a CNN-based COVID-19 classifier. Their results showed that after removing noise from CT images, the disease detection performance was improved even though normal quality measuring parameters like PSNR and MSE were not showing better values in all cases. This work is important because it explains the relation between the noise removal process and disease identification performance. Similarly, Zhang et al. developed a denoising method by combining Transformer and CNN models for low-dose CT images [3]. Their model used the local feature extraction ability of CNN and the learning ability of Transformer for understanding image information. The proposed method improved PSNR, SSIM, and RMSE values and also maintained the structural details better when compared to normal CNN-based methods. This study explained that protecting small anatomical details is very important while removing noise from medical images.

Recent studies have also focused on denoising low-dose CT images without requiring clean target images. Sawada et al. developed a Noise2Noise-based CNN framework integrated with Channel Attention and Spatial Attention mechanisms for low-dose CT denoising [4]. Their approach utilized only noisy image pairs during training and demonstrated improved PSNR performance compared to standard SRResNet architectures. The study highlighted that attention mechanisms help preserve edge information and important anatomical structures while suppressing noise. Furthermore, the authors showed that denoising using only noisy data is practical in medical imaging scenarios where clean ground-truth CT images are difficult to obtain. These studies collectively indicate that modern denoising approaches should not only remove noise but also preserve clinically significant structures, texture information, and disease-related patterns. Such findings provide strong motivation for developing adaptive enhancement and residual noise suppression techniques specifically for segmented lung CT images.

Several researchers focused on improving medical image contrast and visibility using histogram equalization and adaptive enhancement techniques. Riadi et al. [8] proposed a CLAHE-based contrast improvement method combined with Otsu thresholding for enhancing medical X-ray images. Their work showed that adaptive histogram equalization can improve visibility of important anatomical regions without excessive intensity distortion. Similarly, Rahmi-Fajrin et al. [9] applied wavelet transform and CLAHE techniques for dental X-ray image enhancement and reported improved image clarity and structural visibility. These studies indicate that adaptive contrast enhancement methods are effective for improving local details and visual quality in medical images while preserving important structural information.

Image enhancement and machine learning techniques are used by many researchers for improving abnormality detection in medical chest images. Abdussalam et al. [11] studied different enhancement techniques like gamma correction, BCET, histogram equalization, and CLAHE for improving the quality of chest X-ray images before applying classification. Their work showed that enhancement methods improve the visibility of features and also increase the classification performance, mainly when they are combined with the Support Vector Machine (SVM) classifier. In this study, SVM obtained better Area Under the Curve (AUC) values when compared with Naïve Bayes and Random Forest methods. This shows that SVM is more effective for medical image classification-related work.

Koonsanit et al. [12] developed an improved enhancement method called N CLAHE for digital chest X-ray images. In their method, a normalization technique was combined with CLAHE for improving the local contrast and visibility of structures like lungs and chest regions. The authors identified that normal enhancement methods may increase noise or create an unnatural appearance in the images. However, their proposed hybrid enhancement method provided better contrast and improved the visualization quality. Their results showed the importance of adaptive local enhancement techniques for main-

taining important structural details in radiographic images used for diagnosis and screening processes.

Many researchers have worked on medical image enhancement using different preprocessing and contrast improvement techniques. MRI and X-ray image processing methods have shown the importance of preserving image quality during enhancement [5], [6]. Histogram-based enhancement methods, adaptive gamma correction, and adaptive histogram equalization have also been widely used to improve contrast while maintaining important structural information [7], [13], [14], [16]. Similar approaches were applied for chest X-ray image enhancement and noise reduction before further image analysis and disease identification [10]. These studies indicate that effective enhancement methods should improve image quality without affecting the important anatomical details.

An image reconstruction network using CNN is proposed in [18]. An adversarial image reconstruction network and an enriched deep residual network are proposed for image retrieval, which gave better results [19],[20]. A refined residual network using a pretrained network is applied on medical image databases in [21]. Various deep learning models in the medical domain are presented in [22].

III. METHODOLOGY

The proposed work introduces an Adaptive Edge Preserving Hybrid De-Noising Framework and a Variance Regulated Adaptive Histogram Equalization (VRAHE) approach for improving segmented lung CT image quality. The denoising framework is designed to suppress Gaussian, quantum, and salt-and-pepper noises while preserving important edge and structural details inside segmented lung regions. The proposed VRAHE method further improves local contrast, brightness preservation, and visibility of anatomical structures without producing excessive intensity distortion. The major novelty of this work lies in performing region-focused denoising and adaptive enhancement specifically on segmented lung CT regions, which helps preserve important clinical details while improving overall image quality for further abnormality analysis and diagnostic applications.

The detailed architecture is shown in Fig. 1. The proposed work used the LIDC-IDRI chest CT dataset containing 1018 patient CT volumes. Each patient scan is processed using a 3D U-Net model to identify and separate the lung region from the remaining thoracic structures. The detailed training configuration details used for segmentation are shown in Table I below. The segmentation network generates a binary lung mask for every CT volume, where lung pixels are retained while non-lung regions are excluded. These binary masks are then applied to the corresponding CT volumes to obtain segmented three-dimensional lung images containing only the lung anatomy. Since the subsequent image enhancement algorithms operate on individual CT slices, the segmented 3D lung volumes are converted into two-dimensional lung CT slices. These segmented lung slices are used as the input for the proposed Adaptive Edge Preserving Hybrid De-Noising Framework (AEPH), followed by the Variance Regulated Adaptive Histogram Equalization (VRAHE) method for improving image quality.

The proposed work takes lung-segmented masked CT images as input. Where the lung region is obtained by applying the binary lung mask on the original CT slices. Gaussian noise, quantum noise, and salt-and-pepper noise are considered in the region of interest, and the Adaptive Edge Preserving Hybrid De-Noising Framework performs noise removal using local variance estimation, edge preservation, and adaptive noise suppression. Hybrid filtering with weighted fusion is applied to reduce residual noise while maintaining lung edges and structural information. The de-noised lung slice is again processed using Variance Regulated Adaptive Histogram Equalization for contrast regulation, brightness preservation, and local detail enhancement to get the final refined lung CT slice. Algorithm 1 on Adaptive Edge Preserving Hybrid De-Noising Framework for Segmented Lung CT Images and Algorithm 2 on Variance Regulated Adaptive Histogram Equalization (VRAHE) are given in the section below.

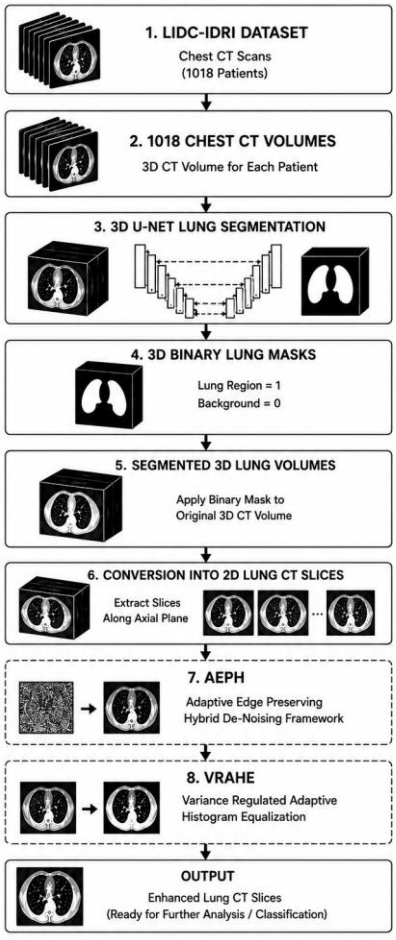


Fig. 1. Proposed architecture – AEPH and VRAHE for improving chest CT.

TABLE I. TRAINING CONFIGURATION OF 3D UNET

Parameter	Configuration
Dataset	LIDC-IDRI chest CT dataset (1018 CT volumes)
Dataset Split	70% Training, 15% Validation, 15% Testing
Network Architecture	3D U-Net with encoder-decoder structure and skip connections

Parameter	Configuration
Input / Output	Input: 3D chest CT volumes; Output: 3D binary lung masks
Computing Platform	Ubuntu Server, Intel Core i7, 32 GB RAM, NVIDIA RTX GPU (12 GB) with CUDA support

The detailed parameter configuration used for AEPH algorithms is shown in Table II below.

TABLE II. PARAMETER CONFIGURATION OF PROPOSED AEPH

Parameter	Configuration
Local Analysis Window	5 × 5 pixels
Adaptive Median Filter Window	Variable (3 × 3 to 7 × 7)
Guided Filter Radius	4 pixels
Adaptive Fusion Strategy	Variance-based weighted fusion

Algorithm-1: Adaptive Edge Preserving Hybrid De-Noising Framework for Segmented Lung CT Images

1. Input Segmented Noisy Lung CT Image

Let the segmented noisy lung CT image be.

$$I(x,y) \tag{1}$$

where, x,y represent pixel coordinates and I(x,y) represents intensity value

2. Lung-Region-Aware Local Mean Estimation

For each sliding local window W,

only valid lung-region pixels are considered.

Local mean is computed as.

$$mu_W=\frac{1}{N_L}\sum_{(i,j)\in W_L} I\,(i,j) \tag{2}$$

where, W_L represents valid lung pixels inside the local window, N_L represents the number of valid lung pixels inside the local window, mu_W represents the local mean intensity.

3. Lung-Region-Aware Local Variance Estimation

Local variance is computed only using valid lung pixels.

$$\sigma_W^2=\frac{1}{N_L}\sum_{(i,j)\in W_L}(I(i,j)-mu_W)^2 \tag{3}$$

where, σ_W^2 represents local variance.

4 . Impulse Noise Density Estimation

$$D_W=\frac{N_{impulse}}{N_L} \tag{4}$$

where, $D_W$represents impulse noise density, $N_{impulse}$ represents the number of impulse-noise pixels.

5. Impulse Noise Detection Condition

$$D_W>T_d \tag{5}$$

where, T_d represents the impulse noise threshold.

6. Edge Strength Estimation

$$G(x,y)=\sqrt{G_x^2+G_y^2} \tag{6}$$

where, $G(x,y)$ represents edge strength at pixel (x,y), G_x represents horizontal gradient, G_y represents vertical gradient.

7. Adaptive Median Filtering

Impulse-noisy regions are denoised using adaptive median filtering.

$$I_m(x,y)=median(W_{xy}) \tag{7}$$

where, $I_m(x,y)$ represents the median-filtered image output, W_{xy} represents the local neighborhood window centered at (x,y).

8. Guided Filtering

Gaussian noisy regions are denoised using guided filtering.

$$I_g(x,y)=a_kI(x,y)+b_k \tag{8}$$

where, $I_g(x,y)$ represents the guided filtered output image, a_k represents the local linear coefficient, b_k represents the bias coefficient.

9. Wavelet Transform

$$W=WT(I) \tag{9}$$

where, W represents coefficients, and WT(I) represents the wavelet transform operation applied to image I.

10 . Wavelet Thresholding

$$W_T=0 \text{ if } |(W)|<\lambda$$

$$W_T=W-\lambda \text{ if } |(W)|\geq\lambda$$

where, W_T represents thresholded wavelet coefficients, λ represents the threshold value.

11. Non-Local Means Filtering

$$I_{nlm}(x)=\sum_{y\in\omega} w\,(x,y)I(y) \tag{10}$$

where, $I_{nlm}(x)$ represents the non-local means filtered image, ω represents the search region, $w(x,y)$ represents the similarity weight.

12. Similarity Weight Computation

Similarity weights are computed as.

$$w(x,y)=\frac{1}{Z(x)}exp\left(\frac{-(P(x)-P(y))^2}{h^2}\right) \tag{11}$$

where, $P(x)$ represents the image patch centered at pixel x, $P(y)$ represents the image patch centered at pixel y, h represents the normalization factor, $Z(x)$ represents normalization factor.

13. Adaptive Weighted Fusion

Outputs from all denoising filters are combined as.

$$I_f=w_1I_m+w_2I_g+w_3I_w+w_4I_{nlm} \tag{12}$$

where, I_f represents the adaptively fused denoised image and w_1, w_2, w_3, w_4 represent adaptive fusion weights, I_w represents the wavelet-denoised image.

14. Weight Constraint

Adaptive weights satisfy.

$$w_1+w_2+w_3+w_4=1 \tag{13}$$

This is to ensure normalized adaptive fusion

15. Adaptive Weight Selection

Adaptive weights are dynamically selected using.

$$w_i = f(\sigma_W, G, D_W) \quad (14)$$

where, w_i represents the adaptive weight for each denoising filter.

16. Edge Preservation Factor

Edge-preservation control factor is computed as

$$\alpha(x,y) = e^{-G(x,y)} \quad (15)$$

where, $\alpha(x,y)$represents edge-preservation coefficient.

17. Final Edge-aware Reconstruction

The final denoised image is reconstructed as

$$I_{final} = \alpha I_f + (1-\alpha) I \quad (16)$$

where, I_{final} represents the final denoised segmented lung CT image.

The proposed Variance Regulated Adaptive Histogram Equalization (VRAHE) method was developed to improve local contrast and visibility of segmented lung CT images. The method adaptively regulates histogram enhancement based on local variance information to preserve brightness and avoid excessive intensity amplification. The proposed approach improves visualization of anatomical structures and enhances local details while maintaining structural consistency in lung regions. The following Table III shows the parameter configuration of the VRAHE algorithm.

TABLE III. PARAMETER CONFIGURATION OF VRAHE

Parameter	Configuration
Local Processing Window	8 × 8 pixels
Variance Regulation	Local variance-based adaptive regulation
Variance Threshold (T)	Adaptive threshold estimated from local variance
Contrast Enhancement	Local adaptive histogram equalization

Algorithm-2: Variance Regulated Adaptive Histogram Equalization (VRAHE)

1. Input Denoised and Lung Masked Chest CT Image

The denoised chest CT image containing only the segmented lung region is taken as input.

$$I_D(x,y) \quad (1)$$

where, $I_D(x,y)$ is the denoised and lung-masked chest CT image, and x, y are pixel coordinates

2. Division into Local Windows

The lung CT image is divided into small local windows for adaptive processing. Local analysis helps preserve fine pulmonary structures and texture details.

$$W_k = I_D(i,j) \quad (2)$$

where, $W_k =$ and $I_D(i,j)$ are pixel intensity values

3. Local Mean Computation

The local mean represents the average intensity within each local CT region. It provides brightness information for variance estimation.

$$mu_k = (1\%N) \sum I_i \quad (3)$$

where, mu_k is the local mean, N is the total number of pixels, and I_i is pixel intensity.

4. Local Variance Computation

$$\sigma_k^2 = (1\%N) \sum (I_i - mu_k)^2 \quad (4)$$

where, $\sigma_k^2 islocalvariance$, $mu_k islocalmean$, $I_i = pixelintensity$.

5. Variance Regulation Factor Estimation

The enhancement strength is controlled using local variance information. This prevents over-enhancement in smooth regions containing residual noise.

$$R_k = \sigma_k^2 / (\sigma_k^2 + C) \quad (5)$$

where, R_k is the local variance regulator factor. $\sigma_k^2 = localvariance$, C is a small constant

6. Local Histogram Construction

A histogram is generated for each local CT region. It represents the intensity distribution within the local window.

$$H_k(r) = n_r \quad (6)$$

where, $H_k(r)$is the histogram value.

n_r is the gray scale frequency

7. Probability Density Function Computation

Histogram values are normalized to obtain probability density values. This gives the occurrence probability of each intensity level.

$$P_k(r) = n_r \% N \quad (7)$$

where, $P_k(r)$ is the probability density function,n_r is the histogram frequency, and N is the total number of pixels

8. Cumulative Distribution Function Computation

The cumulative distribution function is computed from probability density values. It is used for adaptive histogram equalization transformation.

$$CDF_k(r) = \sum P_k(j) \quad (8)$$

where, $CDF_k(r), P_k(j)$ are the cumulative distribution function and probability density values.

9. Variance Regulated Adaptive Enhancement

Histogram equalization is adaptively controlled using the variance regulation factor. Detailed lung regions receive stronger enhancement than smooth regions.

$$I_E(x,y) = R_k \times CDF_k(r) \times (L-1) \quad (9)$$

where, $I_E(x,y)$ is the enhanced CT image, R_k is the variance regulation factor, and L is the total number of gray levels.

10. Residual Noise Suppression

Low variance regions are controlled using a suppression factor. This prevents amplification of residual noise after denoising.

$$\sigma_k^2 < T ? \left(I_F(x,y) = \alpha \times I_E(x,y) : I_F(x,y) = I_E(x,y)\right) \quad (10)$$

where, $I_F(x,y)$ is the final enhanced image, α is the suppression coefficient, and T is the variance threshold.

11. Final Enhanced Chest CT Image

All enhanced local windows are combined to generate the final enhanced CT image. The output image contains improved contrast and preserved pulmonary structure $I_{VRAHE}(x,y)=I_F(x,y)$ (11)

where, $I_{VRAHE}(x,y)$ is the final enhanced chest CT image, $I_F(x,y)$ is the enhanced image after suppression.

IV. Experimental Results

The Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) dataset is a publicly available chest CT image dataset used widely in medical imaging research. The dataset contains thoracic CT scan images collected from different hospitals and imaging systems. It includes lung CT images with expert annotations provided by experienced radiologists. The dataset is useful for lung image enhancement, denoising, segmentation, and disease analysis studies. Since the dataset contains large variations in image quality and acquisition conditions, it is suitable for developing and validating image processing algorithms for chest CT images [17]. The detailed dataset description is shown in Table IV.

TABLE IV. Dataset Description

Parameter	Description
Dataset Name	LIDC-IDRI
Imaging Modality	Thoracic Computed Tomography (CT)
Total CT Scans	1018 lung CT scans (130GB, appx 244,000 DICOM image slices)
Annotation Details	Annotated by 4 experienced thoracic radiologists

The chest CT images obtained from the LIDC-IDRI dataset [17] were first processed using lung segmentation. This step was performed to extract only the lung region from the complete chest CT images. Binary lung masks were generated to separate the lung area from unwanted background regions and other structures ash shown in the figure below. The original dataset contains three-dimensional CT volumes. These 3D CT scans were converted into two-dimensional CT image slices for further processing. The generated binary masks were then applied to each 2D CT slice to obtain segmented lung CT images. These segmented images were further used for denoising and image enhancement processes in the proposed work.

The proposed work was implemented on a workstation system running Ubuntu Server operating system with GPU support. The system was configured with an Intel Core i7 processor, 32 GB RAM, and 1 TB SSD storage for CT image processing tasks. NVIDIA RTX GPU with 12 GB of dedicated memory and CUDA support was used for accelerated computation.

The above Fig. 2 and 3 show results after applying 3D UNet to produce a segmented lung mask image and the segmented lung mask with noise (left) and ground truth image without noise (right).

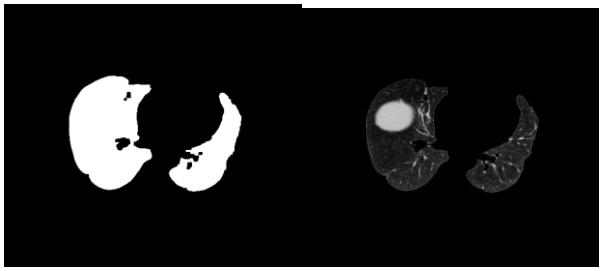


Fig. 2. Segmented lung binary image (Left), segmented lung mask image (Right).

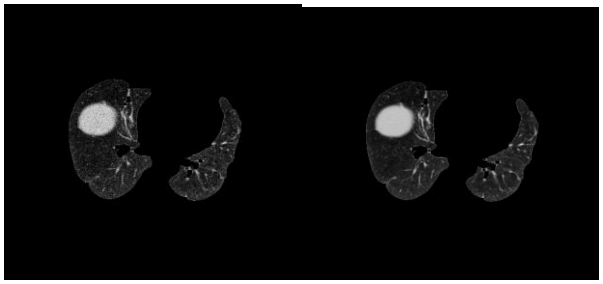


Fig. 3. Segmented lung binary image with Noise (Left), with out Noise (Right).

The performance of the proposed Adaptive Edge Preserving Hybrid De-Noising Framework was evaluated using segmented lung CT images obtained from the LIDC-IDRI dataset. The denoising performance was analyzed using quantitative image quality metrics such as PSNR, SSIM, RMSE, EPI, and Entropy. The obtained results were compared with existing denoising approaches to analyze noise suppression capability and edge preservation performance. The average values of PSNR, SSIM, RMSE, and EPI of all denoised images are calculated and used for comparison purposes. Table V shows the performance analysis of the proposed work.

TABLE V. Performance Analysis of the Proposed Adaptive Edge Preserving Hybrid De-Noising

Method	PSNR (Mean ± SD)	SSIM (Mean ± SD)	RMSE (Mean ± SD)	EPI (Mean ± SD)	Entropy (Mean ± SD)
Noisy Image	18.42 ± 0.62	0.610 ± 0.050	28.54 ± 0.59	0.420 ± 0.004	5.21 ± 0.015
Median Filter	24.85 ± 0.48	0.790 ± 0.038	17.41 ± 0.46	0.710 ± 0.003	6.01 ± 0.013
Gaussian Filter	25.92 ± 0.42	0.820 ± 0.031	15.63 ± 0.38	0.750 ± 0.003	6.18 ± 0.012
Bilateral Filter	27.14 ± 0.34	0.860 ± 0.024	13.41 ± 0.31	0.810 ± 0.002	6.52 ± 0.011
BM3D	29.85 ± 0.24	0.890 ± 0.016	11.52 ± 0.24	0.860 ± 0.002	6.71 ± 0.010
AEPH	32.96 ± 0.15	0.934 ± 0.030	9.84 ± 0.19	0.930 ± 0.001	7.12 ± 0.0094

Table V shows the quantitative performance comparison of different denoising methods applied to segmented lung CT images. The noisy image produced poor PSNR, SSIM, and EPI values with high RMSE due to the presence of noise and struc-

tural degradation. Conventional filtering methods improved the image quality to some extent, but fine edge and texture details were partially affected during noise suppression. The proposed AEPH method achieved better PSNR, SSIM, EPI, and Entropy values with lower RMSE, indicating improved noise removal and preservation of important lung structures.

The following graphs, shown in Figs. 4 -7, compare the denoising performance using PSNR, SSIM, RMSE, and EPI metrics, respectively. The proposed AEPH achieved the highest PSNR and SSIM values, along with the lowest RMSE and highest EPI. These results show that AEPH provides better noise removal while preserving important edge and structural details than the existing denoising methods.

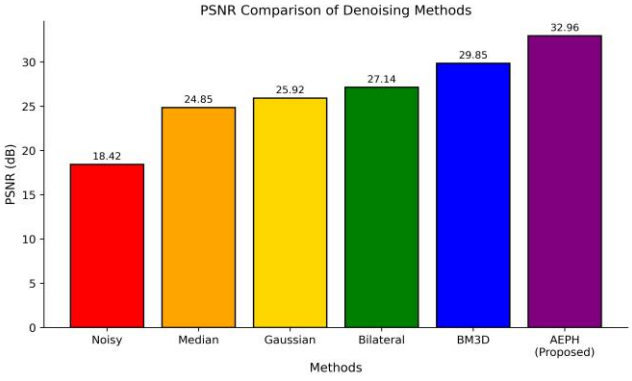


Fig. 4. PSNR comparison of denoising methods.

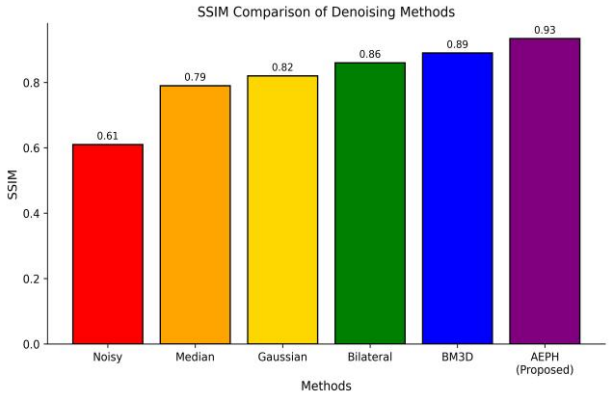


Fig. 5. SSIM comparison of denoising methods.

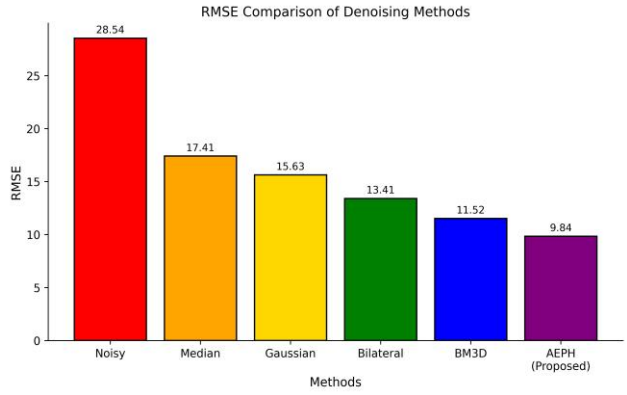


Fig. 6. RMSE comparison of denoising methods.

After implementation of Algorithm 1, Adaptive Edge-Preserving Hybrid Denoising Framework for Segmented Lung CT Images, the denoised images are input to the second algorithm, Variance Regulated Adaptive Histogram Equalization (VRAHE), to improve the contrast of the image. The VRAHE algorithm was evaluated using segmented lung CT images from the LIDC-IDRI dataset. The enhancement performance was analyzed using CNR, AMBE, UIQI, and EME metrics. These metrics were used to evaluate contrast improvement, brightness preservation, and local detail enhancement. The obtained results were compared with conventional histogram equalization methods. The details are shown in Table VI, and the lung segmented before and after applying VRAHE is shown in Fig. 8.

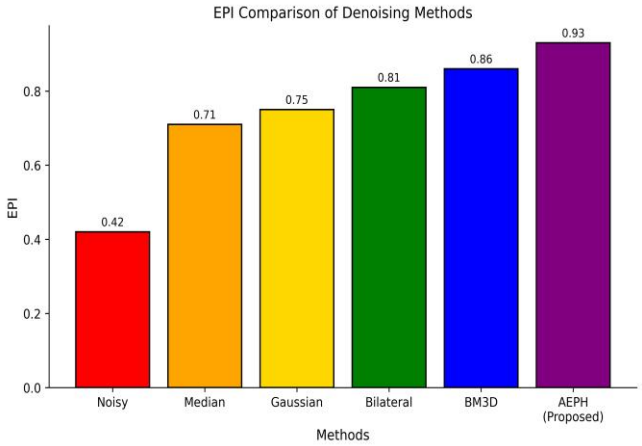


Fig. 7. EPI Comparison of denoising methods.

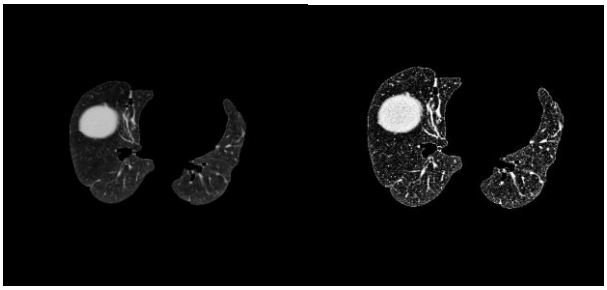


Fig. 8. Before (L) and after (R) applying VRAHE.

TABLE VI. QUANTITATIVE PERFORMANCE ANALYSIS OF VRAHE ENHANCEMENT METHOD

Method	CNR (Mean ± SD)	AMBE (Mean ± SD)	UIQI (Mean ± SD)	EME (Mean ± SD)
Original Image	2.14 ± 0.18	16.82 ± 0.62	0.741 ± 0.020	11.26 ± 0.95
HE	3.08 ± 0.15	11.34 ± 0.51	0.826 ± 0.015	16.72 ± 0.81
CLAHE	4.12 ± 0.12	7.26 ± 0.39	0.901 ± 0.010	22.84 ± 0.66
VRAHE (Proposed)	5.96 ± 0.08	3.14 ± 0.18	0.972 ± 0.004	31.48 ± 0.42

An ablation study was carried out to study the role of the major stages in the proposed AEPH and VRAHE framework. AEPH reduced noise using adaptive filtering, guided filtering, wavelet denoising, and adaptive fusion, and obtained a PSNR of 32.96 dB, SSIM of 0.934, and RMSE of 9.84. After denoising, VRAHE improved contrast and brightness preserva-

tion with CNR of 5.96, AMBE of 3.14, UIQI of 0.972, and EME of 31.48. The results show that each stage adds to the overall refinement of segmented lung CT images. Fig. 9 below shows the Comparative Performance Analysis of Lung CT Image Enhancement Methods.

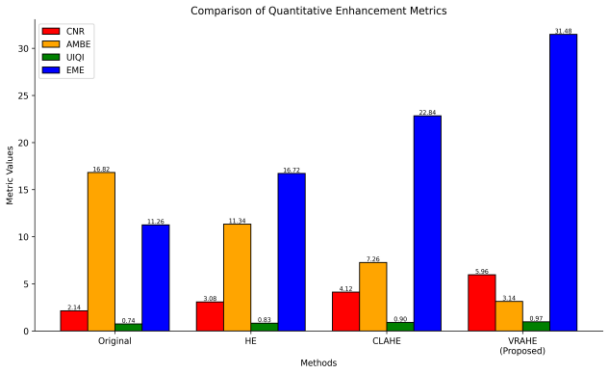


Fig. 9. Comparative performance analysis of lung CT image enhancement methods.

V. Conclusion

This work presented an Adaptive Edge Preserving Hybrid De-Noising Framework for Segmented Lung CT Images and a Variance Regulated Adaptive Histogram Equalization (VRAHE) approach for improving the quality of segmented lung CT images. The proposed denoising framework effectively reduced Gaussian, quantum, and salt-and-pepper noises while preserving important edge and structural information. The obtained results achieved PSNR of 32.96 dB, SSIM of 0.934, EPI of 0.93, RMSE of 9.84, and Entropy of 7.12, indicating effective noise suppression and structural preservation.

The proposed VRAHE method further improved local contrast and visibility of lung structures without producing excessive intensity distortion. Quantitative evaluation produced CNR of 5.96, AMBE of 3.14, UIQI of 0.972, and EME of 31.48, showing improved contrast enhancement and brightness preservation performance. The overall experimental analysis demonstrated that the proposed approaches can effectively improve the quality of segmented lung CT images and preserve important anatomical details useful for further medical image analysis and abnormality detection.

VI. Future Direction

In future work, the proposed Adaptive Edge Preserving Hybrid De-Noising Framework and Variance Regulated Adaptive Histogram Equalization (VRAHE) approach can be extended for real-time clinical applications using GPU-based parallel processing architectures. The proposed methods can also be integrated with deep learning-based abnormality detection and classification models for automated lung disease analysis. Further improvement can be achieved by developing adaptive region-based enhancement methods for preserving fine lesion structures in low-dose CT images. The proposed framework may also be extended to 3D volumetric CT analysis, multimodal medical imaging applications, and explainable AI-assisted diagnostic systems for improved clinical interpretation and decision support.

Acknowledgment

The Research Scholar expresses her sincere thanks to the research supervisor and institution for providing necessary guidance and support to complete this research work. The authors acknowledge the medical image datasets and computational facilities used for carrying out the experiments. The suggestions and encouragement received during the development of the Adaptive Edge Preserving Hybrid DeNoising Framework and Variance Regulated Adaptive Histogram Equalization method helped in improving the overall quality of the proposed work.

References

[1] G. Ashwini and T. Ramashri, “Denoising of COVID-19 CT and chest X-ray images using deep learning techniques for various noises using single image,” in Proc. International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT), 2023.

[2] L. Xia, H. Jebbari, O. Deforges, L. Zhang, L. Lévêque, and M. Outtas, “Denoised CT Images Quality Assessment Through COVID-19 Pneumonia Detection Task,” in Proc. 15th International Conference on Quality of Multimedia Experience (QoMEX), 2023.

[3] J. Zhang, Z. Shangguan, W. Gong, and Y. Cheng, “A novel denoising method for low-dose CT images based on transformer and CNN,” Computers in Biology and Medicine, vol. 163, p. 107162, 2023.

[4] S. Sawada, S. Murakami, L. Guangxu, and T. Kamiya, “A Method for Denoising from Low-Dose CT Images Based on Noise2Noise,” in Proc. 23rd International Conference on Control, Automation and Systems (ICCAS), 2023.

[5] T. Haueise, F. Schick, N. Stefan, and J. Machann, “Comparison of the accuracy of commercial two-point and multi-echo Dixon MRI for quantification of fat in liver, paravertebral muscles, and vertebral bone marrow,” European Journal of Radiology, vol. 172, p. 111359, 2024.

[6] Y. Pourasad and F. Cavallaro, “A novel image processing approach to enhancement and compression of X-ray images,” International Journal of Environmental Research and Public Health, vol. 18, no. 13, pp. 1–18, 2021.

[7] S. Senthilkumar and R. Rajesh, “Histogram modification approaches for medical X-ray image enhancement,” International Journal of Computer Applications, vol. 97, no. 12, pp. 20–26, 2014.

[8] A. A. Riadi, E. Wijayanti, W. H. Sugiharto, A. Latubessy, and A. C. Murti, “Automated contrast improvement of X-ray image using Otsu threshold on contrast-limited adaptive histogram equalization algorithm,” in Proc. International Conference on Information and Communications Technology, 2018, pp. 1–6.

[9] H. Rahmi-Fajrin, S. Puspita, S. Riyadi, and E. Sofyan, “Dental digital image enhancement using wavelet and CLAHE approaches,” International Journal of Advanced Computer Science and Applications, vol. 10, no. 5, pp. 112–118, 2019.

[10] M. Sharma and D. Kumar, “Contrast enhancement and noise reduction of chest X-ray images,” in Proc. International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), 2022, pp. 570–574.

[11] T. M. Abdussalam, H. A. Nugroho, and I. Soesanti, “Image Enhancement Techniques on Chest X-Ray Images to Improve COVID-19 Detection,” in 2022 4th International Conference on Biomedical Engineering (IBIOMED), Yogyakarta, Indonesia, 2022, pp. 87–90, doi: 10.1109/IBIOMED56408.2022.9987990.

[12] K. Koonsanit, S. Thongvigitmanee, N. Pongnapang, and P. Thajchayapong, “Image Enhancement on Digital X-Ray Images Using N-CLAHE,” in Proceedings of the International Conference on Digital Image Processing and Medical Imaging, Thailand, pp. 1–4.

[13] M. Veluchamy and B. Subramani, “Image contrast and color enhancement using adaptive gamma correction and histogram equalization,” Optik, vol. 183, pp. 329–337, 2019, doi: 10.1016/j.ijleo.2019.02.054.

[14] J. B. Zimmerman, S. M. Pizer, E. V. Staab, J. R. Perry, W. McCartney, and B. C. Brenton, "An evaluation of the effectiveness of adaptive histogram equalization for contrast enhancement," IEEE Transactions on Medical Imaging, vol. 7, no. 4, pp. 304–312, Dec. 1988, doi: 10.1109/42.14513.

[15] H. Lidong, Z. Wei, W. Jun, and S. Zebin, "Combination of contrast limited adaptive histogram equalisation and discrete wavelet transform for image enhancement," IET Image Processing, vol. 9, no. 10, pp. 908–915, 2015, doi: 10.1049/iet-ipr.2014.0755.

[16] D.-C. Chang and W.-R. Wu, "Image contrast enhancement based on a histogram transformation of local standard deviation," IEEE Transactions on Medical Imaging, vol. 17, no. 4, pp. 518–531, Aug. 1998, doi: 10.1109/42.730397.

[17] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman et al., "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans," Medical Physics, vol. 38, no. 2, pp. 915–931, 2011.

[18] Pinapatruni, R., Shoba Bindu, C. Learning image representation from image reconstruction for a content-based medical image retrieval. SIViP , 1319–1326 (2020).

[19] Pinapatruni, R., Chigarapalle, S.B. Adversarial image reconstruction learning framework for medical image retrieval. SIViP 16, 1197–1204 (2022).

[20] Pinapatruni, R., Chigarapalle, S.B. (2022). EDR: Enriched Deep Residual Framework with Image Reconstruction for Medical Image Retrieval. Computer Vision and Image Processing. CVIP 2021. Communications in Computer and Information Science, vol 1568. Springer.

[21] Pinapatruni, R., Chigarapalle, S.B ,Images reconstruction using a refined Res-UNet model for medical image retrieval,ACTA IMEKO,March 2022,Volume 11.

[22] Pinapatruni, R., Bhuvaneshwari, T. (2025). Deep Learning Innovations for Predictive Healthcare Systems. ICICC 2024. Lecture Notes in Networks and Systems, vol 1241. Springer, Singapore. https://doi.org/10.1007/978-

Design and Implementation of Rule-Based and Time-Based Logic in a PAC-Based Precision Irrigation Control System

Djohar Syamsi[1]*, Mimin Muhaemin[2], Oviyanti Mulyani[3], Hanif Fakhrurroja[4], Puput Dani Prasetyo Adi[5]
Doctoral Program of Agro-Industrial Technology-Faculty of Agro-Industrial Technology, Padjadjaran University, Sumedang, West Java, Indonesia[1]
Faculty of Agro-Industrial Technology, Padjadjaran University, Sumedang, West Java, Indonesia[2]
Faculty of Agriculture, Padjadjaran University, Sumedang, West Java, Indonesia[3]
Research Center for Smart Mechatronics, National Research and Innovation Agency of the Republic of Indonesia, Bandung, Indonesia[1, 4]
Research Center for Telecommunication, National Research and Innovation Agency of the Republic of Indonesia, Bandung, Indonesia[5]

***Abstract*—This study aims to design and functionally validate a prototype precision fertigation control system based on a Programmable Automation Controller (PAC) capable of managing water and nutrient delivery through two approaches: a time-based approach as the basic control and a rule-based approach as a corrective response based on sensor data (pH, EC, and moisture) of the growing medium. The method used in this study is Research and Development (R&D), and the scope of the research is limited to the Preliminary Field Testing phase, which includes requirements gathering and literature review, architectural planning (sensors, gateway, PAC, actuators, HMI/Human Machine Interface), development of an initial prototype (hardware and software), as well as initial-scale functional testing through hardware functional testing and black-box testing on the HMI interface. The test results showed that all acceptance criteria were met, including sensor function deviation of 2.5% (≤5%), actuator response time of 1.2 s (<2 s), current deviation of 2.0 A (≤3 A), voltage stability of 12.1 V (11.4–12.6 V), operational reliability of 96 hours (≥72 hours), communication success rate of 99.5% (≥99%), and system response latency of 320 ms (<500 ms); all results were classified as Pass. Thus, the developed prototype has proven to be stable and suitable for initial field testing, and provides a solid foundation for further development toward broader application in more efficient and responsive fertigation management.**

***Keywords*—HMI; precision fertigation; PAC; preliminary field testing; time-based control; rule based control**

I. Introduction

Terminologically, fertigation is often regarded as an intrinsic element of hydroponic systems. However, within the broader scope of agronomy, this technology is a technique for delivering soluble nutrients through an irrigation system that can be applied across various growing media [1]. The application of fertigation in crop cultivation using drip fertigation (drip irrigation) in soil or non-hydroponic substrates represents an evolution of precision agriculture [2]. Fertigation is a crop cultivation method that combines irrigation and fertilization, in which nutrients are delivered simultaneously with irrigation water in a measured and timely manner. This method offers high efficiency in water and nutrient use, which directly impacts the improvement of crop yield quality and quantity [3]. Fertigation is a highly effective and increasingly popular approach in crop cultivation.

In fertigation-based crop cultivation, "treatment" often refers to variations in the dosage, frequency, or composition of nutrients provided to plants. Plant treatments may involve applying specific conditions or inputs to plants to observe their impact on growth and yield [4][5]. The concept of "treatment" plays a crucial role; in fertigation, it represents an approach to water management and distribution that integrates essential nutrients dissolved directly into the plant's root zone via the irrigation system. Plant treatment is not merely separate watering or fertilizing, but rather a precise, integrated process. A "treatment" in fertigation can encompass various variables, including: 1) Type and Concentration of Nutrients; 2) Volume and Frequency of Application; 3) Application Timing (nutrient delivery). The objective of treatment in the fertigation method is to ensure that plants receive the optimal amount of water and nutrients at the right time, thereby improving nutrient use efficiency, conserving water, and maximizing crop yields.

Effective fertigation methods require high precision, and this is where control technology is essential. In previous studies, IoT (Internet of Things) technology has been widely used to monitor and control fertigation. Some of these studies include: Microcontroller- and HMI-based IoT applications to support agricultural control systems [6]. The use of cloud-server-based IoT for fertigation control [7]; in this study, the system was highly dependent on the internet. The use of IoT and fuzzy logic to support irrigation systems [8] [9] [42] [43]; this study did not account for plant growth phases. The use of IoT to improve the efficiency of nitrogen and EC use in drip-fertigation systems [10] [11]; this did not account for errors in sensor readings [41].

The current trend in precision agriculture is that data has become the key factor enabling automation, intelligent analysis, and more efficient decision-making. On the other hand, agricultural research generates a wealth of data regarding optimal nutrient doses, timing of nutrient application, ideal growing medium conditions (pH/EC/moisture), and specific

*Corresponding author.

plant requirements at various growth stages [12]. This data and information are often complex and require agronomic expertise to interpret.

This study aims to design, develop, and verify an HMI (Human Machine Interface) application program to execute and monitor two plant treatment methods: time-based (plant treatment based on time/schedule) and rule-based (serving as both a quick correction and a safety limit monitor) within a PAC (Programmable Automation Controller)-based control system, thereby enabling the regulation of water and nutrient supply based on sensor data in the growing medium.

A PAC is a control system that combines the functionality of a programmable logic controller with more modern and flexible computing capabilities, similar to a personal computer (PC). The main contributions of this research are: 1) developing an HMI as an interface between the user and the fertigation control system, 2) implementing two plant treatment modules (time-based and rule-based) integrated into the fertigation control system, and providing facilities for monitoring the fertigation process via the HMI.

II. METHOD

This study employs a Research and Development (R&D) methodology, which emphasizes the design and construction of prototypes, as well as prototype evaluation [13]. The focus of the study is to produce a prototype HMI application program and a plant treatment logic control system (time- and rule-based), which has been functionally validated through preliminary field testing.

A. Research and Development (R&D) Methodology

Fig. 1 shows the block diagram of the R&D stages conducted in this study, namely: 1) Research and Information Collection: this stage gathers requirements and foundational knowledge to design a precision fertigation control system through literature reviews, growing medium data (pH, EC, moisture) as indicators of plant health and fertility, and PAC architecture. 2) Planning: This stage establishes the technical design and prototype plan, including system architecture, as well as initial functional test scenarios as success criteria; 3) Developing a Preliminary Product Form: This stage produces an initial prototype of a PAC-based fertigation control system integrated with control logic and HMI software, 4) Preliminary Field Testing: This stage involves simulation testing to ensure the prototype functions properly and is ready for further development.

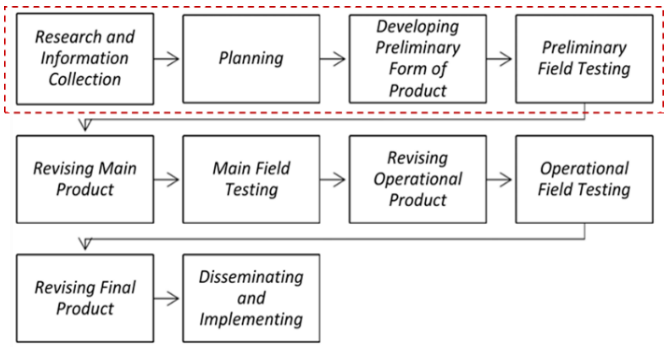


Fig. 1. Stages of the R&D methodology.

1) Concept of plant treatment using the fertigation method: In fertigation, "treatment" refers to the planned, measured, and controlled application of a mixture of irrigation water and nutrient solution directly to the plant's root zone. This treatment is not merely ordinary watering or fertilizing, but rather a precise, integrated management of plant needs [14]. Treatment is at the core of the fertigation method, where an integrated system delivers the primary supplies of water and nutrients directly to the plants. The fertigation method integrates four planned actions: 1) formulating nutrients, 2) determining dosage and concentration, 3) regulating timing and frequency, and 4) selecting the treatment method. Thus, treatment in the fertigation method involves the measured management of water and essential nutrient delivery, thereby enhancing the accuracy, efficiency, and responsiveness of agriculture to achieve optimal crop yields.

This study developed two integrated control methods: time-based control and rule-based control. In time-based control, the system relies entirely on a schedule predefined by the operator or farmer. In the rule-based treatment, the system relies on conditional logic (if-then) triggered by real-time data from EC, pH, and growing medium moisture sensors. Rule-based treatment is applied only when the plants need it, thereby preventing waste and optimizing growth. This system is far more efficient and responsive to daily weather changes, resulting in more efficient resource use and more consistent crop yields. An example of a rule-based treatment is: "IF the growing medium moisture is below 60% OR solar radiation exceeds a certain threshold, THEN activate the actuator until moisture reaches 85%."

2) PAC-based fertigation control system: Effective fertigation methods require high precision, so control system technology is essential. The fertigation control system serves as an operational hub that ensures the treatment is carried out as planned. This system is responsible for opening valves and activating pumps. Consequently, to execute these various actuator operations, accurate and reliable hardware is required. The PAC integrates various functions, including: 1) Data Acquisition: The PAC connects to a network of field sensors to monitor growing media conditions in real-time; 2) Data Processing: The PAC runs complex algorithms, including both time-based treatment logic and rule-based treatment logic (if-then) to interpret sensor data; 3) Coordinated Control: Based on the interpretation of sensor data, the PAC precisely controls the actuators to implement the required treatment activities.

B. Design and Development: Prototype of the Complete System

Fig. 2 shows the architecture of a PAC (Programmable Automation Controller)-based irrigation control system, which consists of: 1) Data Processing, 2) Control Panel, 3) Actuators, and 4) WSN (Wireless Sensor Network). Table I shows the functions of each component within the PAC.

TABLE I. COMPONENTS, FUNCTIONS, AND DEVICE REQUIREMENTS OF THE PAC

Components	Functions	Requirement
PC/Laptop	Data Processing	1. OS: Windows 10 2. RAM: 8 MB 3. Hard disk: 120 GB
Control Panel	The control panel serves as a hardware integration hub that connects the actuators to the data processing PC, ensuring that the sensing–decision–actuation process runs smoothly.	4. ADAM 4520: RS-232 to RS-485 Converter 5. ADAM 4017/18: Analog Sensor Input, 8 channels 6. ADAM 4068: Relay Output, 8-channel relay-based
Actuator	The actuator serves as an execution component that converts control decisions from the PC into physical actions in the fertigation process.	7. Relay: 12V DC, 8-pin
Wireless Sensor Network	Continuously measures the EC, pH, and humidity parameters in the growing medium.	Soil sensor 7-in-1, 12–24 V, DC RS485 module; Gateway

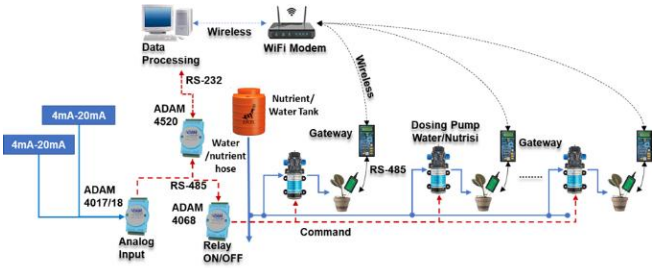


Fig. 2. Architectural design of the fertigation control system.

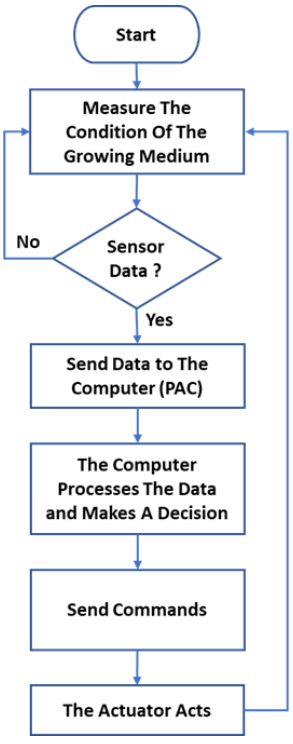


Fig. 3. Flowchart of how the fertigation control system works.

Fig. 3 shows a block diagram of how the PAC-based fertigation control system works. Overall, this fertigation control system operates in a single feedback loop: measure conditions --> send data to the computer (PC) --> the computer processes the data and makes a decision --> send a command --> the actuator responds.

1) Data processing: The Data Processing unit is the central data processing component within the PAC (Programmable Automation Controller), responsible for executing logic control instructions [28], processing data from the WSN (Wireless Sensor Network) [15], and managing communication with actuators. In this study, the Data Processing Unit consists of a PC running the Windows 10 operating system, which enables it to handle more complex algorithms to support the operation of the Logic Control and HMI (Human Machine Interface) application programs.

Fig. 4. Data processing PC unit configuration.

Fig. 4 shows the configuration of the Data Processing PC unit; the PC is connected to the ADAM-4520 via RS-232, and the ADAM-4017 and ADAM-4068 are connected to the ADAM-4520 via RS-485. The Data Processing unit runs the Logic Control application program to select the treatment method (Time-Based or Rule-Based). The result of processing sensor data using the Logic Control method is a decision related to actuator activity.

2) Control panel: The control panel in this study serves as a hardware integration hub that connects the actuators to the data processing PC, ensuring that the sensing–decision–actuation process runs smoothly[37]. The Control Panel consists of: 1) Input Module (ADAM 4017) [16]: receives data from analog sensors, 2) Output Module (ADAM 4068): sends control signals to the actuator to activate the irrigation pump. The I/O (Input/Output) modules on the PAC are modular (hot-swappable), meaning they can be added or removed, providing flexibility for system expansion. The ADAM-4520 is an isolated serial converter that converts RS-232 communication to RS-422/RS-485 for greater noise immunity and transmission ranges of up to approximately 1.2 km, while protecting the host computer via 3000 VDC isolation. Fig. 5 shows the connection between the I/O module and the PC.

The ADAM-4068 is an 8-channel relay-based digital output module designed to control field devices such as pumps, valves, or alarms via RS-485 commands (ASCII/Modbus RTU). Meanwhile, the ADAM-4017 is an 8-channel analog input module that reads analog sensor signals (voltage/current), converts them into 16-bit digital data that can be processed by a control system, and also supports Modbus RTU communication. Communication between ADAM modules occurs via a fieldbus; ADAM modules are fundamentally designed to operate within a "master-slave" architecture. The master (typically a PC) is responsible for initiating all communication with other components [38]. I/O modules (such as the ADAM-4017 or ADAM-4068) act as slaves that only respond to requests from the master.

ADAM modules use an RS-485 physical interface for inter-module communication. This network topology allows for multi-drop or daisy-chain connections, enabling dozens of modules to be connected on a single pair of cables [39]. This is highly efficient for connecting numerous sensors and actuators spread across agricultural areas. The protocol used for communication between ADAM modules and the master can be: 1) Modbus/RTU, the most widely used standard industrial protocol. The master sends a Modbus "command" (e.g., "read the analog input value from module ID 01") and the ADAM module responds with the requested data; 2) ASCII Command, which uses simple ASCII-based text commands. Often, a PC only has a standard RS-232 port (an older serial port), so the ADAM 4520 functions as a bridge or translator, converting the RS-232 signal from the PC into an RS-485 signal.

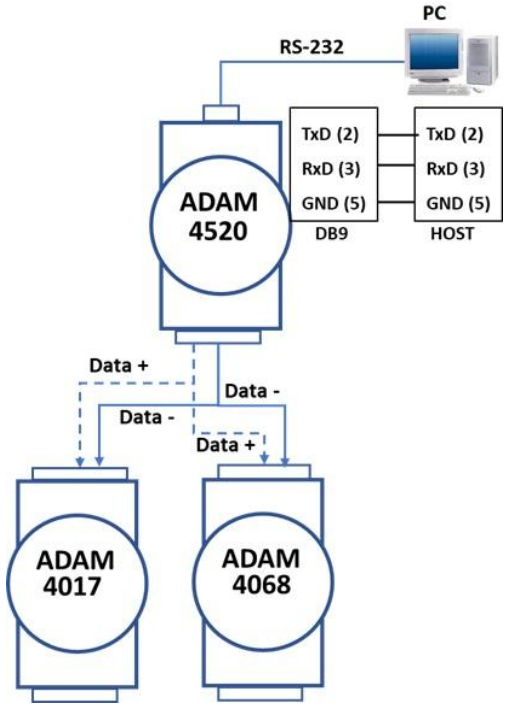


Fig. 5. Wiring diagram on the control panel.

3) Actuator: Actuators serve as execution components that translate control decisions from the PC into physical actions in the fertigation process. Based on readings from pH, EC, and growing medium humidity sensors, as well as time-based (routine schedule) and rule-based (correction upon deviation) control logic, actuators control devices such as pumps and/or valves to deliver water and nutrient solutions as needed. Thus, the actuator serves as a direct link between data processing results and actual field conditions, while ensuring that water or nutrient delivery is timely, measured, and responsive to changes in growing medium conditions.

The PC sends low-voltage control signals via the I/O module. These signals trigger the actuators to operate. In the context of a fertigation system, this relationship is clearly evident: 1) The PC makes a decision, and the actuator acts: The PC executes rule-based logic that states, "If soil moisture is below 60%, activate irrigation in Zone A." The PC then sends the appropriate electrical signal; 2) The Role of Relays and Solenoid Valves: the signal is received by the output module (such as the ADAM-4068, which contains relays). The relay functions as an intermediate switch to activate actuators requiring higher power, such as water pumps or solenoid valves; 3) Execution of Crop Treatment: the solenoid valve opens, allowing water and nutrient solution to flow into the root zone. The water pump turns on, creating the required pressure. It is the physical action of opening the valve or turning on the pump that is performed by the actuator; Fig. 6 shows the actuator design.

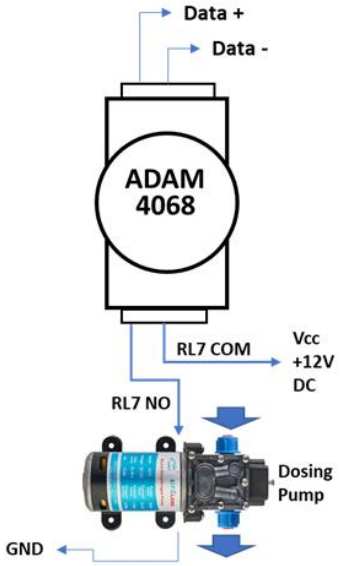


Fig. 6. Actuator wiring diagram (ADAM 4068 to dosing pump).

Data+ and Data- are the RS-485 communication lines connecting the ADAM 4068 [16] to the PC (via the ADAM 4520); these lines are used to receive control instructions from the host. RL7 COM (Common) is the common terminal of channel 7 relay on the ADAM 4068. This pin serves as the power supply input (+12V DC) for the actuator. RL7 NO (Normally Open) is the normally open terminal on channel 7 of the relay. When the relay is inactive, the circuit between COM and NO is open. When the relay is activated by the PC, the circuit closes, allowing the +12V DC voltage to flow to the dosing pump. Vcc +12V DC is the positive power supply voltage (+12 volts) used to operate the dosing pump. It is routed through the RL7 COM line and passed to RL7 NO when the relay is active. GND is the ground line connected to the negative

side of the power supply and the dosing pump, thereby forming a complete electrical circuit [40]. With this configuration, the dosing pump will only turn on when the relay on channel RL7 is activated. This allows the PC, via the ADAM 4068, to automatically control nutrient delivery based on calculations from data analysis. The data processing PC is the decision-maker, and the actuators are the physical executors that ensure crop management proceeds in accordance with the programmed automation strategy; without the actuators, the intelligent decisions made by the data processing PC would never be translated into actual actions in the field.

4) Monitoring of growing media: In this context, the PAC (PC Data Processing Unit) receives growing medium data from a WSN (Wireless Sensor Network)-based sensor network [17]. Data from the WSN is transmitted directly to the PC Data Processing Unit via a WiFi (wireless) network. In this study, the configuration between the WSN and the PC utilizes the Fog Computing concept [18], thereby eliminating the need for a cloud server.

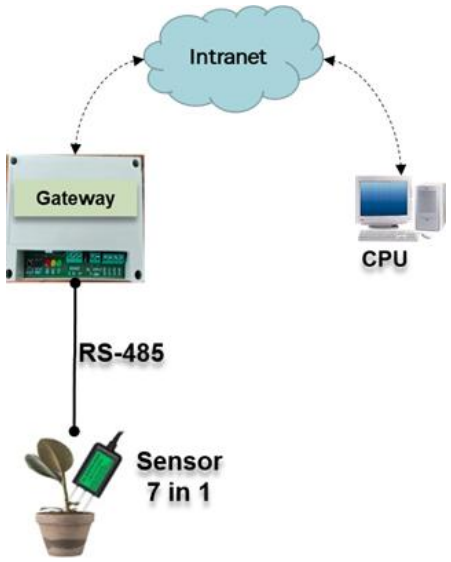


Fig. 7. WSN configuration as a data source.

The growing medium sensors and the Gateway (as shown in Fig. 7) are WSN devices that continuously measure EC, pH, and humidity parameters in the growing medium. The data collected by the gateway is then sent to the PC Data Processing unit using the MQTT (Message Queuing Telemetry Transport) protocol [19]. Data is collected periodically (e.g., every 10 minutes). This interval ensures that the PAC always has up-to-date information on plant conditions without overloading the network.

The PAC uses the received data to apply rule-based treatments as follows: 1) "If-Then" Conditional Logic: Data received from the WSN (e.g., humidity = 55%, pH = 6.0) is used to evaluate pre-programmed rules; 2) Automatic Decision-Making: The PAC makes real-time decisions based on this data; example rules: (a) "IF humidity < 60% THEN activate the irrigation pump," (b) "IF pH > 6.5 THEN apply an acid solution to stabilize the growing medium"; 3) The fertigation control system does not provide nutrients based on a schedule, but only reacts when actual field conditions require intervention, ensuring efficient use of resources. This study emphasizes the importance of integrating real-time data from sensors in decision-making regarding plant nutrient application.

5) Sensor plant health and productivity indicators: The pH, EC (Electrical Conductivity), and growing medium moisture content are key indicators in precision agriculture that provide fundamental information regarding nutrient availability and uptake by plants; these three parameters serve as early indicators of nutrient requirements [12] [20] [21]. pH (Acidity Level) measures the acidity or alkalinity of the growing medium. It is not an indicator of the amount of nutrients present, but rather an indicator of their availability. As an indicator, this means that nutrients are absorbed most effectively within a specific pH range. Most plants prefer a pH between 5.5 and 6.5. For example, if the pH is too high (alkaline), nutrients such as Iron (Fe), Manganese (Mn), and Phosphorus (P) will be locked within the growing medium and cannot be absorbed by the roots, even if they are present in sufficient quantities. Initial Action: A suboptimal pH reading is an early indicator that the plant is at risk of nutrient deficiency in the future, so adjusting the pH of the nutrient solution is necessary.

EC (Electrical Conductivity) measures the total concentration of dissolved salt ions in the growing medium solution; these salts constitute the fertilizer (nutrients). As an indicator, EC directly reflects the concentration of the nutrient solution available in the root zone. For example, a low EC value indicates that the growing medium is deficient in nutrients and the plants need to be fertilized. Conversely, if the EC value is too high, it indicates an excess of fertilizer, which risks causing toxicity or nutrient burn (burnt roots). Initial Action: Monitoring EC can provide a direct indication of whether the current nutrient dose is appropriate or needs to be increased or decreased.

The moisture content of the growing medium indicates the amount of water available within it. As an indicator, this means that water serves as the medium for nutrient transport. Nutrients must be dissolved in water so that they can be absorbed by the roots through osmosis and transpiration. Example: If the growing medium is too dry, nutrient uptake will cease entirely, even if the pH and EC values are ideal. A drastic drop in moisture levels indicates the need for immediate irrigation, which is often combined with nutrient application (fertigation) to maintain a steady flow of nutrients to the plants.

Fig. 8. (7-in-1) Growing medium sensor for a fertigation control system.

The sensor used in this study is the RS485-based Multisensor Digital soil sensor [22], specifically designed for precision agriculture applications and fertigation systems, as shown in Fig. 8. This sensor serves as a digital sensing device

that is a critical component of modern agricultural systems; it is a detector that integrates the ability to measure several key growing medium parameters simultaneously, often including moisture, temperature, electrical conductivity (EC), and pH.

Table II shows the role of growing medium parameters as indicators of plant health and productivity.

TABLE II. PLANT HEALTH AND PRODUCTIVITY INDICATORS

Parameter Monitoring	Description	Action
EC (Electrical Conductivity)	An EC level that is too high or too low can hinder plant growth and overall health.	• If the EC reading is too high, add clean water (without nutrients).• If the EC reading is too low, add a nutrient solution with a higher concentration, adjusted according to the plant's growth stage (early stage, vegetative stage, or generative/flowering/fruiting stage).
pH (potential of hydrogen)	The primary effect of pH on plants is on nutrient availability. Suboptimal pH levels prevent nutrients from being bound and absorbed by the plants, even though they are present in the growing medium.	• Optimal nutrient availability occurs at a growing medium pH of 5.5–7.0.• If the growing medium pH is too low (< 5.5), the pH must be adjusted toward neutral.• If the growing medium pH is too high (> 7.0), corrective measures are needed to lower the pH.
Moisture Content of the Growing Medium	Adequate water availability in the growing medium is essential for plant survival, but too much or too little water can have negative effects.	• Add water and/or nutrients according to the moisture level of the growing medium.

By monitoring these three key indicators in real time—humidity to ensure nutrient transport occurs, pH to ensure nutrients are accessible and available, and EC to ensure nutrient concentrations are sufficient but not excessive—the fertigation control system can automatically adjust the necessary treatments to maintain optimal root zone conditions for plant growth. WSN sensors provide accurate, real-time data to the PAC, enabling it to implement rule-based treatments so that plants receive optimal growing conditions.

6) Design of the system logic control application program: In the context of PAC functionality, the PC (Data Processing) contains the Logic Control and HMI application programs. Logic Control is an application program that functions as the Decision Layer [36]. The Decision Layer is a component of the layered architecture responsible for high-level decision-making, such as interpreting sensor data, selecting or planning actions, and then sending commands to the execution/actuation layer below it. The Decision Layer also encompasses decision-making and planning processes, which generate a sequence of actions. Meanwhile, the HMI (Human-Machine Interface) is a feature that provides a dashboard for users (farmers) to operate and monitor fertigation activities. Fig. 9 shows a Flow Diagram of how the HMI application program works.

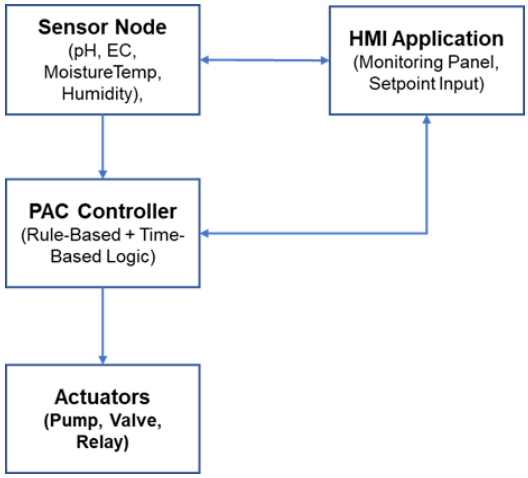


Fig. 9. Flowchart of the HMI application program's operation.

The Sensor Node measures environmental conditions and plant nutrition based on pH, EC, and growing medium moisture parameters. The PC serves as the logic control center with two main control methods: 1) Time-Based Logic, which activates the pump at specific times; 2) Rule-Based Logic, which activates the pump if the growing medium sensor detects specific conditions (e.g., low EC). The HMI provides a user interface to view real-time data, set time points, limit parameters, or perform manual overrides. Actuators receive signals from the PC to activate pumps, valves, or relays according to the plant's growth stage or sensor conditions.

The relationship between the logic controller and the HMI can be described as follows: 1) The logic controller is an application program containing rules regarding plant care, such as the following instruction [26]: "IF the humidity sensor reads below 60% AND the sun is shining brightly, THEN activate the pump for 5 minutes." Logic control translates human ideas (research findings) into machine language that can be executed by the PAC [27]; 2) The HMI is the "face" of the control system [29]; it is not used to create new logic, but to monitor and interact with the existing Logic Control.[30].

In this study, the PAC does not rely on an internet connection, thereby reducing the potential for disruptions caused by network instability. Additionally, since the PAC is PC-based, the development of application programs and the human-machine interface (HMI) becomes more flexible and robust. The PAC allows programming using high-level programming languages, enabling the system to be designed with more complex, dynamic features that are easily adaptable to user needs and the desired level of control complexity. [31].

One of the activities in this study was to develop an HMI model (as a dashboard) with several features, including: 1) Hardware Settings, 2) Time-Based "treatment" features, 3) Rule-Based "treatment" features, to support the operation of the fertigation control system [33]. These two treatment features are required simultaneously, as not all plant conditions can be predicted solely from sensor data or growth stages alone. By combining both, the fertigation control system becomes more flexible [32], responsive, and precise. Table III illustrates the design concepts of time-based and rule-based "treatment."

TABLE III. THE DESIGN CONCEPT OF "TREATMENT"

Features / Approach	Time-Based Treatment	Rule-Based Processing
Main Trigger	A fixed schedule (for example, every day at 7 a.m.)	Actual field conditions (If humidity < X%)
Basis for the Decision	The time specified by the "User."	Programmed "If-Then" Logic
Data Source	System clock/calendar	Real-time sensors (humidity, temperature, etc.)
Flexibility	Not weather-resistant	Flexible (adapting to changing conditions)
Response	Procedural (following a set of steps)	Reactive (responds to the sensor threshold)
Complexity	Low (simple timer)	Medium (requires sensors and logic)
Application Examples	Standard home garden irrigation	Automated greenhouse irrigation system

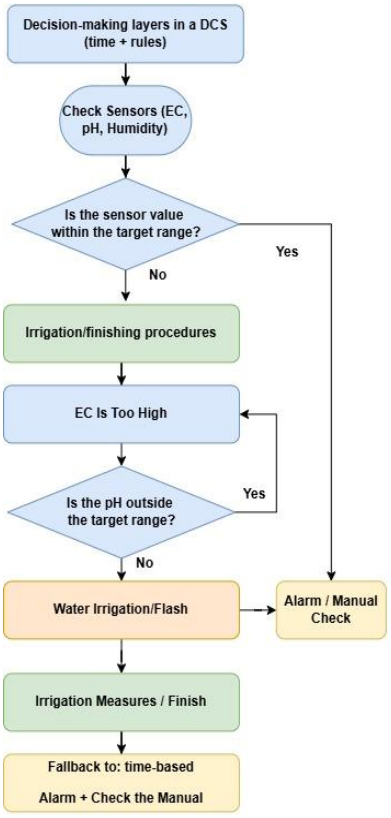


Fig. 10. Flowchart of plant treatment (time-based and rule-based).

Fig. 10 shows a flowchart of the logic control in the DCS for time- and rule-based plant treatment control (base schedule + correction rules). Based on this flowchart, the workflow of the fertigation control system [34] is as follows: 1) The system reads the sensors, 2) monitors whether the growing medium parameters are still within normal limits, 3) In rule-based mode, the system evaluates deviation conditions, 4) If a disturbance occurs (e.g., sensor/communication issues), the system falls back to time-based mode and continues to run the minimum schedule while issuing an alarm to suggest a manual check. Moreover, Table IV shows the features for hardware settings, which include: port connectivity, output control [35], data acquisition, measurement calibration, data storage, and logging.

TABLE IV. HARDWARE SETUP AND SYSTEM CONFIGURATION

Category: Setting	Job Function / Job Description	Software Implementation / HMI	Information Regarding ADAM Hardware
Port Connectivity	Open/Close the Serial Communication Port	The "Connect" / "Disconnect" button on the HMI. A drop-down menu for selecting the COM port (e.g., COM1, COM2) and setting the baud rate (e.g., 9600 bps).	Using the ADAM-4520 as a converter and RS-485 bus.
Output Control (Relay)	Trying to run the relays (RL0–RL7)	The HMI displays 8 toggle switches labeled "Pump A," "Valve 1," etc. Each relay has a visual status indicator (ON/OFF light).	Send a Modbus/ASCII Write command to the ADAM-4068 to enable or disable a specific relay.
Data Acquisition	Trying to Retrieve Data from the Sensor	The "Refresh Data" button or automatic update on the HMI that displays sensor values in real time (for example, "Humidity: 75%").	Send a Modbus/ASCII Read command to the ADAM-4017 to read 8 analog input channels and receive data from IoT sensors.
Measurement Calibration	Calibration of Sensor Measurement Results	The configuration menu on the HMI for entering offset and scaling values (for example, converting a raw 4–20 mA value to 0–100% humidity).	Adjust the digital reading of the ADAM-4017 to match the actual physical value.
Data Storage & Logs	Save Settings and Record Relay Usage (RL0–RL7)	The "Save Settings" button on the HMI. An automatic data logging feature that runs in the background.	The data log is stored in a text file (.txt) or database on the PC/PAC, recording when the ADAM-4068 relay turns on and off.

C. Development: HMI and Decision Layer Treatment Programming

Pseudocode-1 shows the pseudocode for the system initialization application program, which includes: general initialization and communication port status initialization.

Pseudocode-1

```
// Initialize global variables and data structures
DEFINE current_moisture, current_ph, current_ec AS FLOAT
DEFINE pump_status, valve_A_status AS BOOLEAN
DEFINE scheduled_time_A, scheduled_duration_A AS TIME/DURATION
DEFINE rule_moisture_threshold AS FLOAT

// Initialize communication port status
```

```
SET comm_port_open = FALSE

FUNCTION InitializeSystem()
    LOAD_SETTINGS_FROM_FILE("conFig. txt")
    DISPLAY("System initialized. Awaiting connection.")
END FUNCTION
```

Fig. 11 shows the HMI display as the system begins the initialization process. The growing medium sensor will display the current data.

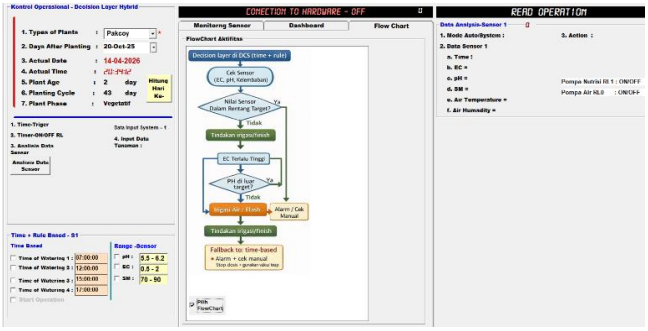


Fig. 11. HMI screen during system initialization.

Moreover, Pseudocode 2 shows the pseudocode for the application program to configure port communication, manually test the relay, and calibrate the sensor. These functions can be accessed by the user through the programming interface or the HMI configuration menu.

Pseudocode-2

```
// Use Case: Configure Communication Port
FUNCTION ConfigurePort(port_name, baud_rate)
IF comm_port_open THEN
CLOSE_PORT()
END IF
OPEN_PORT(port_name, baud_rate)
IF success THEN
SET comm_port_open = TRUE
DISPLAY("Port " + port_name + " opened successfully.")
SAVE_SETTINGS_TO_FILE("conFig. txt")
ELSE
DISPLAY("Failed to open port. Check connection.")
END IF
END FUNCTION

// Use Case: Test Relay Function (ADAM 4068)
FUNCTION TestRelay(relay_id, duration_seconds)
IF comm_port_open THEN
DISPLAY("Testing Relay R" + relay_id + " for " + duration_seconds + "s")
SEND_MODBUS_WRITE(ADAM_4068_ADDRESS, relay_id, STATE_ON)
LOG_ACTION("RL" + relay_id + " turned ON for test")
WAIT(duration_seconds * 1000 milliseconds)
SEND_MODBUS_WRITE(ADAM_4068_ADDRESS, relay_id, STATE_OFF)
LOG_ACTION("RL" + relay_id + " turned OFF after test")
ELSE
DISPLAY("Error: Port not open. Cannot test hardware.")
END IF
END FUNCTION

// Use Case: Retrieve Data from Sensor (ADAM 4017) and Calibrate
FUNCTION ReadAndCalibrateSensorData()
IF comm_port_open THEN
// 1. Retrieve raw data
raw_moisture_data = SEND_MODBUS_READ(ADAM_4017_ADDRESS, CHANNEL_0)
raw_ph_data = SEND_MODBUS_READ(ADAM_4017_ADDRESS, CHANNEL_1)

// 2. Apply calibration (scaling/offset)
current_moisture = APPLY_CALIBRATION(raw_moisture_data, MOIST_SCALE, MOIST_OFFSET)
current_ph = APPLY_CALIBRATION(raw_ph_data, PH_SCALE, PH_OFFSET)

DISPLAY("Moisture: " + current_moisture + "%, pH: " + current_ph)
SAVE_SETTINGS_TO_FILE("conFig. txt") // Saves calibration settings
END IF
END FUNCTION

// Use Case: Save Settings and Record Usage (Log File)
FUNCTION LogAction(action_description)
timestamp = GET_CURRENT_TIMESTAMP()
APPEND_TO_FILE("usage_log.txt", timestamp + " - " + action_description + NEWLINE)
END FUNCTION
```

Fig. 12 shows the HMI Hardware Settings screen. This HMI interface is designed to provide intuitive yet powerful control, with high flexibility in setting critical fertigation parameters locally via the PAC. The system can be used to support decision-making based on the actual conditions of the growing medium.

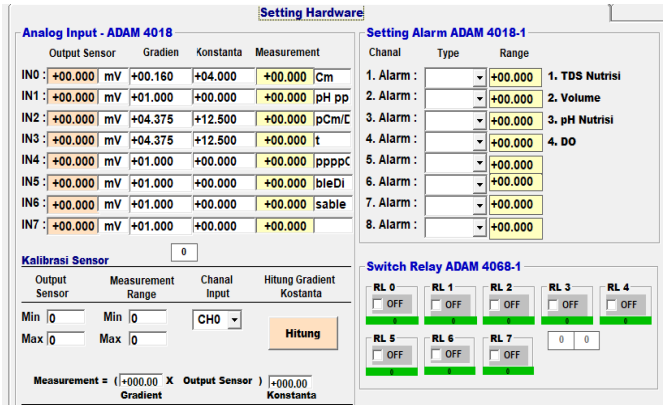


Fig. 12. Hardware settings dashboard.

Pseudocode 3 shows the pseudocode for time-based treatment; this logic runs continuously in the main system loop.

Pseudocode-3

```
FUNCTION RunTimeBasedTreatment()
current_time = GET_CURRENT_TIME()

// Check if the current time matches the scheduled start time
IF current_time IS scheduled_time_A THEN
DISPLAY("Executing Time-Based Treatment A")

// Activate pump/valve (e.g., Relay 0 on ADAM 4068)
SEND_MODBUS_WRITE(ADAM_4068_ADDRESS, RELAY_0, STATE_ON)
LogAction("RL0 ON - Start time treatment")

// Wait for the scheduled duration
WAIT(scheduled_duration_A)

// Deactivate pump/valve
SEND_MODBUS_WRITE(ADAM_4068_ADDRESS, RELAY_0, STATE_OFF)
LogAction("RL0 OFF - End time treatment")
END IF
END FUNCTION
```

Moreover, Pseudocode-4 shows the pseudocode for rule-based treatment. This logic runs continuously in a loop within the main system and typically checks environmental conditions more frequently than time-based systems.

Pseudocode-4

```
FUNCTION RunRuleBasedTreatment()
  // First, update sensor data to get current conditions
  ReadAndCalibrateSensorData()

  // Apply the 'If-Then' rule
  IF current_moisture < rule_moisture_threshold THEN
    // Condition met: Soil is too dry. Need to irrigate.
    DISPLAY("Rule Triggered: Moisture below threshold. Starting irrigation.")

    // Activate pump/valve (e.g., Relay 1 on ADAM 4068)
    SEND_MODBUS_WRITE(ADAM_4068_ADDRESS, RELAY_1, STATE_ON)
    LogAction("RL1 ON - Rule based irrigation started")

    // Continue irrigation until the threshold is met again
    WHILE current_moisture < (rule_moisture_threshold + 5) // Stop when 5% above threshold
      WAIT(60000) // Wait 1 minute before checking moisture again
      ReadAndCalibrateSensorData()
    END WHILE

    SEND_MODBUS_WRITE(ADAM_4068_ADDRESS, RELAY_1, STATE_OFF)
    LogAction("RL1 OFF - Rule based irrigation finished")
  ELSE
    DISPLAY("Rule Check: Moisture is OK. No action needed.")
  END IF
END FUNCTION
```

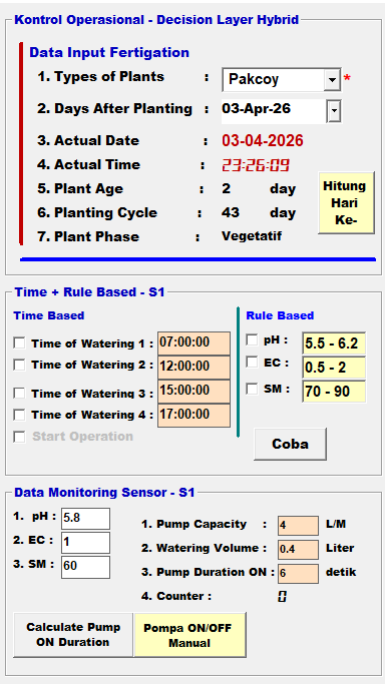


Fig. 13. HMI interface for control and monitoring.

Fig. 13 shows the HMI interface for configuring and using time- and rule-based treatment logic controls. The treatment HMI provides four time-setting inputs, as well as settings for three indicator parameters: EC, pH, and growing medium humidity. Monitoring data is available to track the condition of the growing medium in real time.

III. Result and Discussion

This chapter presents testing methods focused on three application programs: hardware configuration, HMI, and the integration of time-based treatment with rule-based treatment. The objective of the testing is to ensure that the developed application programs not only function as designed but are also sufficiently reliable to support measured fertigation operations that can be easily monitored by users via the HMI. The evaluation method is conducted through three approaches: 1) hardware testing to assess the functional performance of the developed PAC-based fertigation prototype, focusing on sensor and actuator operation, power stability, and data communication reliability; 2) functional testing (Black -Box Testing) was used to verify the functions of the HMI and control modules without examining the program's internal structure, but rather based on the consistency of input–output and the system's response to test scenarios; 3) performance testing based on acceptance criteria was applied to assess key technical parameters, such as response latency, actuator response time, and communication success, to ensure they meet the established thresholds.

A. Hardware Functional Testing

TABLE V. Preliminary Field Testing Hardware

No.	Test item	Thres hold	Results (measured)	Unit	Status
1	Sensor Functionality (deviation)	≤ 5.0	2.5	%	Pass
2	Actuator Function (response time)	< 2.0	1.2	s	Pass
3	Power Consumption (current deviation)	≤ 3.0	2.0	A	Pass
4	Voltage Stability (range)	11.4–12.6	12.1	V	Pass
5	Reliability (continuous operation)	≥ 72.0	96.0	h	Pass
6	Sensor Calibration (deviation)	≤ 3.0	2.8	%	Pass
7	Communication Interface (success rate)	≥ 99.0	99.5	%	Pass
8	System Response (latency)	< 500.0	320.0	ms	Pass
9	Thermal (Temperature)	25.0–30.0	27.0	°C	Pass
10	Thermal (Humidity)	50.0–70.0	60.0	%RH	Pass

Table V shows the hardware testing conducted to evaluate the functional performance of the developed PAC-based fertigation prototype, focusing on sensor and actuator operation, power stability, communication reliability, and basic environmental resilience. Table V defines acceptance thresholds for each hardware test item and summarizes the initial test results by comparing the measured values (results) against the acceptance thresholds/criteria (thresholds) referenced in Table V. Each test point is given a "Pass" status if the result meets the established criteria (e.g., maximum deviation limit, minimum reliability duration, or permissible operating range for temperature and humidity).

TABLE VI. BLACK-BOX TESTING OF HMI FUNCTIONS

No	Test Case (TC)	Feature	Test step (scenario)	Metric (measurable)	Threshold	Result (measured)	Unit	Trials (n)	Status (Pass/Fail)	Remarks
1	TC-1	Open serial port (valid)	Enter valid COM (1–10) → Connect	Connection time	≤ 5	2	s	10	Pass	No issues observed
2	TC-2	Open serial port (invalid)	Enter invalid COM (1–10) → Connect	Error display time	≤ 2	1	s	10	Pass	No issues observed
3	TC-3	Relay output test	Toggle RL0–RL7 ON/OFF	Relay response time	≤ 500	300	ms	10	Pass	No issues observed
4	TC-4	Analog input/sensor read	Click "Read data."	Data refresh time	≤ 1000	500	ms	10	Pass	No issues observed
5	TC-5	Select treatment method	Select Time-Based / Rule-Based	Menu load time	≤ 500	500	ms	10	Pass	No issues observed
6	TC-6	Run plant treatment	Run Time/Rule-Based execution	Execution latency	≤ 1000	500	ms	10	Pass	No issues observed

Moreover, Fig. 14 demonstrates the stable and reliable performance of the PAC, as it successfully met the predetermined system performance criteria. Basic functional tests (Nos. 1 and 2) showed that all input and output components functioned as intended. The calibration test (No. 6) is a critical aspect in ensuring data accuracy. By comparing the WSN sensor measurement results against standard reference values (using a Portable TDS and pH/EC standards), the system showed low measurement deviation, namely ≤ ±3%, indicating high accuracy for the intended application. The system's electrical aspects were tested via the Power Consumption Test (No. 3) and Voltage Stability Test (No. 4). Test results demonstrate good power efficiency; total power consumption remains within the adapter's capacity, and no excessive current surges occur. Additionally, the power supply proved stable during operation, with minimal voltage drop and ripple (deviation <3% from nominal), ensuring a safe operating environment for electronic components. Long-term performance and system response testing also yielded positive results.

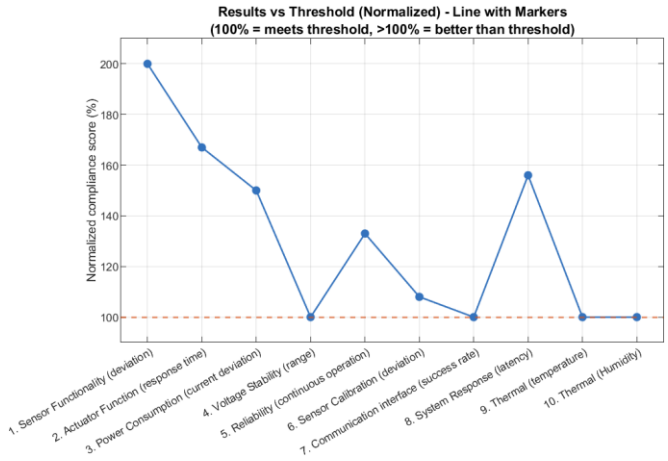


Fig. 14. Graph of PAC hardware test results.

Fig. 14 shows a comparison of normalized Results vs. Threshold, where the 100% reference line indicates that the system meets the minimum acceptance criteria. Values above 100% indicate performance that exceeds the threshold. Overall, the prototype meets the specified functional requirements, confirming that the hardware subsystem is sufficiently stable for initial field testing and subsequent integration with time-based and rule-based control functions.

Hardware connection testing (ADAM 4068 and ADAM 4017) via the serial port yielded positive and robust results (TC-1, TC-2). The hardware control module functions properly, facilitating smooth user interaction with sensors and actuators (TC-3, TC-4). Testing of the application's navigation menu and main control logic was also completed (TC-5, TC-6). This black-box testing confirms that the application program has met its primary functionality requirements, ranging from hardware connectivity (ADAM Modules) and I/O control (relays and sensors) to menu navigation and the execution of automatic control logic, thereby deeming the application stable. Fig. 15 shows the results of the Black-Box test in Table VI in the form of a normalized compliance score calculated using Eq. (1).

$$(\text{Threshold} / \text{Result}) \times 100 \quad (1)$$

B. Black-Box Testing of HMI and Programming Interface

The next testing phase involves conducting functional testing using the black-box testing method on the HMI application program. Black-box testing is a software testing method that does not examine the internal structure of the code or program.

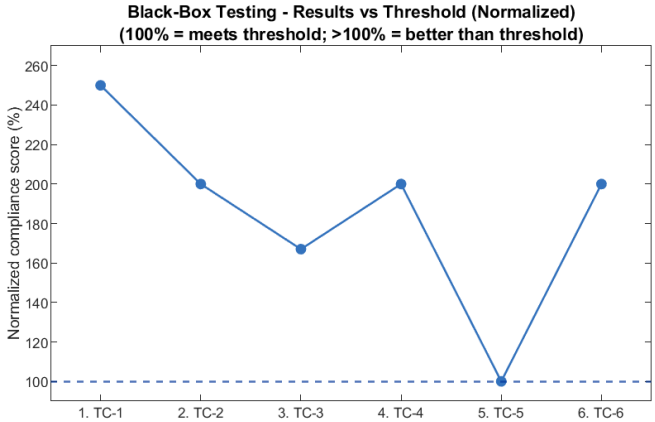


Fig. 15. Graph of the black box HMI test results.

The dashed line at 100% indicates the minimum pass threshold (the result is exactly equal to the threshold). Values

above 100% indicate performance that exceeds the criteria because the measured response time/latency is less than the threshold. In the graph, all test cases are ≥ 100%, so all HMI functions are marked as Pass. The highest score appears in TC-1 (≈250%) because the connection time is much faster than the threshold (2 s vs. 5 s), while TC-5 is exactly at 100% because the menu loading time is equal to the threshold (500 ms), making it the point that is closest to the limit and can be prioritized for optimization if necessary.

C. Simulation of Treatment Integration Function Testing (Time-Based and Rule-Based)

Time-based control serves as a baseline control to maintain a consistent supply of nutrients and water as long as pH, EC, and humidity remain within normal limits. When these parameters deviate, indicating a nutrient or water deficiency or excess, the system switches to rule-based control for more specific actions: low EC triggers nutrient addition, while excessively high EC triggers dilution or flushing with clean water.

Table VII shows how control decisions adjust sensor conditions, so that the time-based and rule-based approaches complement each other. Table VII also shows the simulation tests conducted as follows: under normal operating conditions, as seen in the data for Day 1 (08:00) and Day 15 (16:00), the system executes routine protocols. Although temperature and humidity fluctuate during the day, the system maintains the plant's status in the healthy (optimal) category because the pH and EC parameters remain within tolerance limits. This demonstrates the effectiveness of standard nutrient scheduling when the environment is stable. However, the advantage of rule-based control becomes evident when the system encounters anomalies. On Day 5 (14:00), the system detected a drop in EC to 1.1 mS/cm and pH to 5.5. The system automatically triggered the 'Add Nutrient A' treatment to correct the nutrient deficiency. Conversely, on Day 10 (09:00), when a risk of high salinity (EC 2.5 mS/cm) was detected, the system took emergency action in the form of a 'Clean Water Flush' to reduce excess mineral salts. This integration demonstrates that the application program not only operates based on time but also possesses situational intelligence to maintain plant health through targeted interventions based on real-time sensor data.

The threshold values in Table VII are based on the optimal ranges commonly used for leafy vegetables (e.g., lettuce, bok choy, arugula) in hydroponic/soilless culture systems. For pH, the range of 5.5–6.5 is widely recommended because it maintains optimal nutrient availability in the nutrient solution [23]. For EC, leafy vegetables generally fall within a range of approximately 1.2–2.0 mS/cm (for example, lettuce is often reported to be stable within a range of about 1.2–1.8 mS/cm, and studies on lettuce have also evaluated EC up to 2.0 mS/cm) [23]. Meanwhile, moisture in soilless growing media is more accurately expressed as substrate volumetric water content (VWC); studies on leafy vegetables such as arugula evaluated VWC at 0.20–0.50 $m^3 \cdot m^{-3}$ and showed that maintaining VWC around 0.30 $m^3 \cdot m^{-3}$ supports water use efficiency while maintaining growth and quality [24][25]. Therefore, in the simulation in Table VII, a practical range of 30–50% VWC is used as the operating range (threshold) to represent safe substrate moisture conditions for leafy vegetables in soilless systems. Fig. 16 shows pH, EC, and substrate moisture after normalization to their ideal ranges (0% = lower limit, 100% = upper limit). Occurrences outside the threshold indicate values approaching 0% or 100%, triggering rule-based corrective actions, while values within the range indicate stable conditions under time-based scheduling. Based on the functional testing simulations conducted, it was found that the application is capable of adaptively transitioning control from a time-based approach to a rule-based approach in accordance with the dynamics of the growing medium.

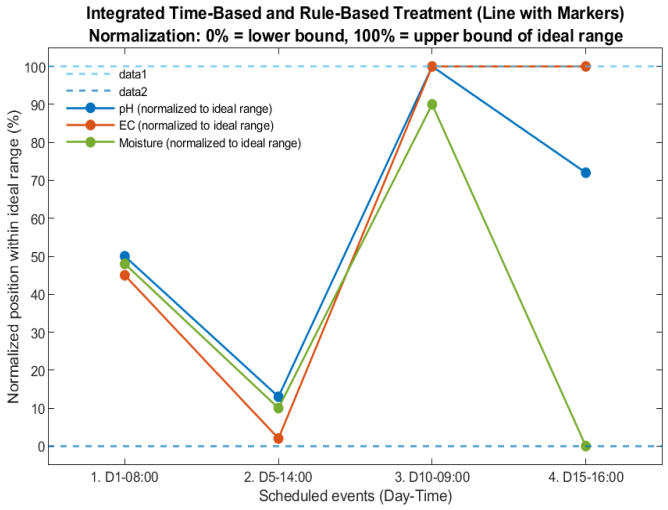


Fig. 16. Results of the simulation of the integration function test across treatments.

TABLE VII. SIMULATION OF INTEGRATION FUNCTIONALITY TESTING ACROSS TREATMENTS (TIME-BASED AND RULE-BASED)

Day	Time	pH	pH threshold	EC (mS/cm)	EC threshold (mS/cm)	Moisture *(%VWC)	Moisture threshold (%VWC)	Control mode	Trigger condition	Action executed
1	08:00	6.1	5.5–6.5	1.6	1.2–2.0	40	30–50	Time-based	Within thresholds	Scheduled dosing executed
5	14:00	5.4	5.5–6.5	1.0	1.2–2.0	32	30–50	Rule-based	pH low & EC low	Increase nutrient concentration / adjust pH up
10	09:00	6.7	5.5–6.5	2.3	1.2–2.0	48	30–50	Rule-based	pH high & EC high	Dilute/flush with water; adjust pH down
15	16:00	6.2	5.5–6.5	1.8	1.2–2.0	28	30–50	Rule-based	Moisture low	Irrigation pulse to restore moisture

*VWC = Volumetric Water Content

*EC = Electrical Conductivity

IV. DISCUSSION

The results of the hardware application program testing (Table V) and the functional testing using the Black Box method on the HMI (Table VI) indicate that all major subsystems (sensor data acquisition, actuation, communication, and interface) meet the acceptance criteria. The system latency of 320 ms and the actuator response time of 1.2 s are well below the dynamic range of the fertigation process, which typically occurs on the order of seconds to minutes; thus, computational delay is not a primary limiting factor in this prototype. Conversely, reading accuracy and communication stability are the most critical aspects because they determine decisions at the pH/EC/humidity thresholds. In the Black Box testing for the HMI application program, the system met the primary operational requirements (connectivity, I/O control, sensor readings, menu navigation, and treatment execution) with margins generally >100%; the optimization priority lies with TC-5, as it is right at the threshold (500 ms).

In the testing, the integration of time-based and rule-based controls provides two layers of regulation; the time-based control ensures the continuous supply of nutrients and water, while the rule-based control acts as a corrective measure when deviations occur. This approach is relevant for cultivating leafy vegetables such as pak choi and lettuce, which are sensitive to changes in EC and growing medium moisture. Due to the hysteresis/persistence in the rule-based system, the system can avoid excessive action caused by momentary fluctuations, thereby potentially maintaining more stable growing medium conditions and reducing water and nutrient waste.

However, this study is still limited to functional validation at the preliminary field testing stage, so it has not yet evaluated the agronomic impacts (e.g., growth, yield, or water/nutrient use efficiency) over a full cultivation cycle.

V. CONCLUSION

This study successfully developed a precision fertigation control system based on a Programmable Automation Controller (PAC) that integrates time-based and rule-based controls within a single data-driven architecture. The system is capable of automatically regulating water and nutrient delivery based on a routine schedule and the actual conditions of the growing medium, as measured by pH, EC, and humidity sensors. Test results indicate that the time-based approach is effective in maintaining nutrient supply stability, while the rule-based approach is more adaptive in handling environmental anomalies in real-time. The integration of both approaches yields a system that is more flexible, precise, and efficient compared to microcontroller-based IoT systems. This system has the potential to support the implementation of precision agriculture and Industry 4.0 technologies in the agricultural sector, particularly in the sustainable management of fertigation for horticultural crops. As a further development, the integration of predictive algorithms or machine learning could be considered to enhance the system's ability to make more adaptive and proactive decisions.

ACKNOWLEDGMENT

Thanks to the entire academic community of the Doctoral Program in Agro-Industrial Technology, Faculty of Agro-Industrial Technology, Padjadjaran University, Sumedang, West Java, Indonesia, as well as to the Research Center for Smart Mechatronics and Telecommunications, National Research and Innovation Agency of the Republic of Indonesia, Bandung, Indonesia, who have assisted in the publication of this article. We hope this article will continue to be cited and prove useful to related academic disciplines.

REFERENCES

[1] A. R. Zubair and T. Adebiyi, "Development of an IoT-based automatic fertigation system," J. Agric. Sci. Technol., vol. 21, no. 3, pp. 4–21, 2022, doi: 10.4314/jagst.v21i3.2.

[2] M. Shukla, A. Chinchmalatpure, I. Prasad, S. Kumar, and D. Camus, "Fertigation: modern technique of fertilizer application," Indian Farmer, vol. 5, no. 9, pp. 1062–1071, 2018. [Online]. Available: https://www.researchgate.net/publication/344328740_Fertigation

[3] X. Sun, Y. Li, M. Heinen, H. Ritzema, P. Hellegers, and J. van Dam, "Fertigation strategies to improve water and nitrogen use efficiency in surface irrigation system in the North China Plain," Agriculture, pp. 1–23, 2023, doi: 10.3390/agriculture13010017.

[4] Y. Sun, L. Duan, H. Zhong, H. Cai, J. Xu, and Z. Li, "Effects of irrigation-fertilization-aeration coupling on yield and quality of greenhouse tomatoes," Agric. Water Manag., vol. 299, p. 108893, 2024, doi: 10.1016/j.agwat.2024.108893.

[5] K. Nomura, E. Wada, M. Saito, S. Itokawa, K. Mizobuchi, and H. Yamasaki, "Effect of different fertigation scheduling methods on the yields and photosynthetic parameters of drip-fertigated Chinese chive (Allium tuberosum) grown in a horticultural greenhouse," Horticulturae, 2024, doi: 10.3390/horticulturae10080794.

[6] H. M. Kaidi et al., "Internet of Things: A monitoring and control system for rockmelon farming," Int. J. Integr. Eng., vol. 12, no. 6, pp. 55–61, 2020, doi: 10.30880/ijie.2020.12.06.007.

[7] F. Idris, A. A. Latiff, M. A. Buntat, Y. Lecthmanan, and Z. Berahim, "IoT-based fertigation system for agriculture," Bull. Electr. Eng. Informatics, vol. 13, no. 3, pp. 1574–1581, 2024, doi: 10.11591/eei.v13i3.6829.

[8] J. M. Guzmán-Toloza, D. F. Villafaña-Gamboa, L. J. Peniche-Ruiz, R. A. Gómez-Buenfil, J. K. Molina-Puc, and M. J. Rodríguez-Morayta, "Internet of Things for irrigation system," Commun. Comput. Inf. Sci., vol. 1053, pp. 294–304, 2019, doi: 10.1007/978-3-030-33229-7_25.

[9] M. J. Hoque, M. S. Islam, and M. Khaliluzzaman, "A fuzzy logic- and Internet of Things-based smart irrigation system," pp. 93, 2023, doi: 10.3390/ecsa-10-16243.

[10] J. Wanyama, S. Kiraga, E. Bwambale, and A. Katimbo, "Improving nutrient use efficiency through fertigation supported by machine learning and Internet of Things in a context of developing countries: Lessons for Sub-Saharan Africa," J. Biosyst. Eng., vol. 48, no. 4, pp. 375–391, 2023, doi: 10.1007/s42853-023-00196-8.

[11] Y. Wu et al., "Optimal control algorithm of fertigation system in greenhouse based on EC model," Int. J. Agric. Biol. Eng., vol. 12, no. 3, pp. 118–125, 2019, doi: 10.25165/j.ijabe.20191203.4680.

[12] N. Ali, M. Dagati, and Y. Dong, "IoT-enabled soil moisture and conductivity monitoring under controlled and field fertigation systems," AgriEngineering, 2025, doi: 10.3390.

[13] S. Gustiani, "Research and development (R&D) method as a model design in educational research and its alternatives," vol. 11, no. 2, pp. 12–22, 2019.

[14] Q. Wang et al., "Intelligent fertigation improves tomato yield and quality and water and nutrient use efficiency in solar greenhouse production," Agric. Water Manag., vol. 298, p. 108873, 2024, doi: 10.1016/j.agwat.2024.108873.

[15] D. Thakur, Y. Kumar, A. Kumar, and P. K. Singh, "Applicability of wireless sensor networks in precision agriculture: A review," Wireless Pers. Commun., vol. 107, no. 1, pp. 1–20, 2019, doi: 10.1007/s11277-019-06285-2.

[16] Advantech, Data Acquisition Modules. Taiwan: Advantech Co., Ltd, 2024.

[17] A. Y. Ghareeb, S. K. Gharghan, A. H. M. Mutlag, and R. Nordin, "Wireless sensor network-based artificial intelligent irrigation system: Challenges and limitations," J. Tech., vol. 5, no. 3, pp. 26–41, 2023, doi: 10.51173/jt.v5i3.1420.

[18] J. Bhatia et al., "An overview of fog data analytics for IoT applications," Sensors, pp. 1–31, 2025, [Online]. Available: https://doi.org/10.3390/s23010199

[19] Y. Sasaki and T. Yokotani, "Performance evaluation of MQTT as a communication protocol for IoT and prototyping," pp. 21–29, 2019.

[20] M. N. Ahmad, M. I. Anuar, N. A. Aziz, and A. A. Murdi, "Function and application of soil electrical conductivity (EC) sensor in agriculture: A review," Adv. Agric. Food Res. J., vol. 6, no. 1, pp. 1–22, 2025, [Online]. Available: https://doi.org/10.36877/aafrj.a0000552

[21] E. Fabricius, "Review of pH recommendations for soilless cultivation systems," Uppsala, 2023.

[22] S. Istrit and P. Ancut, "Autonomous, multisensory soil monitoring system," AgriEngineering, 2025, doi: 10.3390/agriengineering7010018.

[23] D. Bruce, "Electrical conductivity and pH guide for hydroponics," 2016.

[24] K. Lee, S. K. An, and K. Ku, "The optimum substrate moisture level to enhance the growth and quality of arugula (Eruca sativa)," Horticulturae, 2024, doi: 10.3390/horticulturae10050483.

[25] L. Bonelli et al., "Sensor-based fertigation management enhances resource utilization and crop performance in soilless strawberry cultivation," Agronomy, 2024, doi: 10.3390/agronomy14030465.

[26] A. Reuther, C. Plesker, D. Bassauer, and B. Schleich, "A digital twin framework for programmable logic controllers in industrial automation systems," Procedia CIRP, vol. 139, pp. 446–452, 2026, doi: 10.1016/j.procir.2025.08.193.

[27] C. Ma and J. Provost, "Introducing plant features to model-based testing of programmable controllers in automation systems," Control Eng. Pract., vol. 90, pp. 301–310, Sep. 2019, doi: 10.1016/j.conengprac.2019.07.006.

[28] U. T. Bührer, C. Legat, and B. Vogel-Heuser, "Changeability of manufacturing automation systems using an orchestration engine for programmable logic controllers," IFAC-PapersOnLine, vol. 48, no. 3, pp. 1573–1579, 2015, doi: 10.1016/j.ifacol.2015.06.310.

[29] B. R. Mehta and Y. J. Reddy, Industrial Process Automation Systems: Design and Implementation, ch. 8, "Programmable automation controller," pp. 301–306, 2015, doi: 10.1016/B978-0-12-800939-0.00008-5.

[30] G. Thiele, H. J. Beestermöller, L. Renner, M. Dorno, and D. Popovic, "Task-configuration of a pearl-based programmable controller for process automation," Control Eng. Pract., vol. 3, no. 6, pp. 843–848, Jun. 1995, doi: 10.1016/0967-0661(95)00068-6.

[31] D. Zamora-Arranz, P. Garcia-Bringas, J. J. Gude, and J. Del Ser, "Leveraging programmable logic controllers for machine learning applications in industrial setups," Results Eng., vol. 30, p. 110194, Jun. 2026, doi: 10.1016/j.rineng.2026.110194.

[32] A. Sidorenko, M. Rezapour, A. Wagner, and M. Ruskowski, "Towards using behavior trees in industrial automation controllers," Procedia CIRP, vol. 130, pp. 1234–1243, 2024, doi: 10.1016/j.procir.2024.10.233.

[33] V. Vatskel, A. Biloshchytskyi, A. Neftissov, O. Kuchanskyi, Y. Andrashko, and I. Sachenko, "Energy-efficient Internet of Things using LoRa network and modular universal programmable controller in bee apiary management," Procedia Comput. Sci., vol. 231, pp. 715–720, 2024, doi: 10.1016/j.procs.2023.12.257.

[34] W. Ju and H. Xie, "Leveraging entropy-EDAS technique in Z-number decision-making for electrical automation programs' teaching quality evaluation in agricultural colleges," Int. J. Agric. Environ. Inf. Syst., vol. 17, no. 1, Jan. 2026, doi: 10.4018/IJAEIS.405810.

[35] W. Ju and H. Xie, "Leveraging entropy-EDAS technique in Z-number decision-making for electrical automation programs' teaching quality evaluation in agricultural colleges," Int. J. Agric. Environ. Inf. Syst., vol. 17, no. 1, Jan. 2026, doi: 10.4018/IJAEIS.405810.

[36] K. Kefer, M. Affenzeller, and S. Winkler, "Adaptive operators for genetic programming to identify optimal energy flow controllers," Procedia Comput. Sci., vol. 253, pp. 1991–2002, 2025, doi: 10.1016/j.procs.2025.01.261.

[37] Q. Lin, L. Zhang, X. Zhou, J. Chen, and X. Xu, "Software-defined control: New architecture and practices in building automation," J. Build. Eng., 2025, doi: 10.1016/j.jobe.2025.114076.

[38] Z. Ling, S. Brooks, D. McFarlane, A. Thorne, and G. Hawkridge, "Vision-based extraction of industrial information from legacy programmable logic controllers," Robot. Comput.-Integr. Manuf., vol. 97, p. 103088, Feb. 2026, doi: 10.1016/j.rcim.2025.103088.

[39] S. Stankovski, G. Ostojić, I. Baranovski, S. Tegeltija, and V. Smirnov, "Robust automation with PLC/PAC and edge controllers," IFAC-PapersOnLine, vol. 55, no. 4, pp. 316–321, 2022, doi: 10.1016/j.ifacol.2022.06.052.

[40] K. Kwon, S.-K. Lim, D. Kim, and K. Park, "Automation program for optimum design of electric vehicle powertrain systems based on artificial neural network," eTransportation, vol. 18, p. 100267, Oct. 2023, doi: 10.1016/j.etran.2023.100267.

[41] M. Niswar et al., "Performance evaluation of ZigBee-based wireless sensor network for monitoring patients' pulse status," 2013 International Conference on Information Technology and Electrical Engineering (ICITEE), Yogyakarta, Indonesia, 2013, pp. 291-294, doi: 10.1109/ICITEED.2013.6676255.

[42] P. D. Prasetyo Adi, A. Mappadang, A. Wahid, S. Luhriyani, R. Jefri and N. Nurindah, "Spreading Factor of IoT-LoRa Effect for Future Smart Agriculture," 2022 International Conference on Information Technology Research and Innovation (ICITRI), Jakarta, Indonesia, 2022, pp. 123-128, doi: 10.1109/ICITRI56423.2022.9970235.

[43] P. D. P. Adi et al., "Application of IoT-LoRa Technology and Design in irrigation canals to improve the quality of agricultural products in Batu, Indonesia," 2021 2nd International Conference On Smart Cities, Automation & Intelligent Computing Systems (ICON-SONICS), Tangerang, Indonesia, 2021, pp. 88-94, doi: 10.1109/ICON-SONICS53103.2021.9617175.

Advancing Ecological Informatics: A Tool for Implementing the Abundance–Preference Approach in Assemblage Analysis

Mohamed El Farouki[1]*, Mohamed Dakki[2], Meryem Ennakri[3], Soumia Ziti[4]
Intelligent Processing and Security of Systems-Faculty of Sciences,
University Mohammed Vth of Rabat, 10106 Rabat, Morocco[1, 3, 4]
Scientific Institute (Zoology and Animal Ecology), University Mohammed Vth of Rabat, 10106 Rabat, Morocco[2]

***Abstract*—Although the Abundance-Preference Diagram (APD) was introduced four decades ago as an entropy-based framework for classifying species along habitat-preference gradients, its uptake by the ecological community has remained marginal. The cause is not theoretical but operational: no existing software (e.g., vegan, ade4, CANOCO, PC-ORD, BIOMAPPER) implements the Degree of Preference index (DP) or the APD, which forces ecologists either to rebuild the algorithm from scratch or to abandon the method altogether. This barrier mirrors a broader concern in ecoinformatics, where the gap between published methods and accessible implementations continues to slow the adoption of quantitative tools by field ecologists. This study presents EcoAPD (https://apd.biodiversity.ma), the first open web-based platform dedicated to the APD method. The application takes a standard species-by-habitat abundance matrix and returns the full diagnostic suite (DP values, exclusive species curve, hyperbola-based preference classes, and the APD plot itself), without any programming or local installation. Three design choices set it apart from the descendants of CANOCO and ade4: 1) a dual-computation architecture that generates the normalized and non-normalized APD versions simultaneously at project creation, removing the need for users to commit to a method a priori; 2) a hybrid file/relational-database persistence layer with hash-based recompute avoidance, enabling visualizations to be regenerated without rerunning the full pipeline; and 3) a validation engine tailored to the irregularities of ecological matrices. The platform is evaluated on two real datasets and a synthetic size sweep. Vectorizing the preference-index core yields a speedup rising from 3.5× at 10 species to 85.6× at 1000 species, while full-pipeline latency is ~510 ms on a 93-species × 37-habitats matrix and ~2.3 s on a 114-species × 171-habitats waterbird monitoring matrix, the residual cost being dominated by habitat count rather than species count. A threshold-sensitivity ablation finds the classification stable at the 0.3 and 0.75 boundaries and sensitive at 0.1 and 0.5. EcoAPD makes a four-decade-old method operationally available, with potential to revisit historical assemblage studies that relied on abundance data alone.**

Keywords*—Ecological informatics; web application; algorithm; assemblage ecology; species preference; abundance-preference diagram

I. Introduction

The preference of a species in a given environment is a key concept in both ecology and biodiversity conservation [1],[2],[3]. It is generally compiled from the abundance distribution of a species in different habitats [4]. The degree of preference (DP), as defined by [5], is an index that determines whether a species is a generalist or a specialist, in terms of its habitat preference, without necessarily indicating which habitats it prefers. This index shows high efficiency in analyzing assemblage organization through the APD method, which combines DP and species abundances [6],[7].

Although first published four decades ago, this method remained underutilized due to the lack of accessible tools for ecologists, who are mostly not familiar with statistical tools, a challenge that persists due to the frequent mismatch between ecological data and traditional statistical knowledge and tools [8],[9]. The recent refinements of the method [6] and the current abundance of efficient informatics tools have now made its widespread implementation feasible.

Our research consists of translating this method into a practical tool: a web application that executes all necessary calculations and plots the diagnostics. This application utilizes a standard data matrix of species abundances across a sample of habitats and covers both the original and normalized versions of the APD [7].

The development of this application was preceded by a comparative analysis of available algorithms for processing data in the form of contingency tables, and of the software platforms that analyze taxonomic and ecological data. That analysis, reported in Section II, establishes that none of the platforms in current use can implement the DP/APD tools. The platforms that come closest are separated from their intended users by either a programming requirement or a commercial licence.

To help ecologists implement DP/APD tools, we have previously developed an algorithm [7], and in parallel we launched a first version of the EcoAPD web application, which has been significantly improved since then.

The contributions of this work are the following:

- EcoAPD is, to the authors' knowledge, the first open, web-based implementation of the Degree of Preference index and the Abundance–Preference Diagram, a method that no existing ecological software package implements (Section II, Table I).
- A dual-computation architecture derives both the normalized and the non-normalized APD at project

*Corresponding author.

creation, so that the choice between the two versions becomes a post-hoc display decision rather than an a priori methodological commitment (Section V-A).

- A hybrid file/relational persistence layer with hash-based recompute avoidance stores all computed intermediates, so that visualizations are regenerated without rerunning the pipeline, and re-uploading unchanged data does not re-trigger computation [Section V-A(2c)].
- The preference-index computation, originally a per-species loop, has been rewritten as a single vectorized NumPy operation, measured at a speedup rising to 85.6× at 1000 species (Section VI-B, Table IV).
- An empirical evaluation reports the scalability of the full pipeline on two real datasets, together with a two-part ablation quantifying the contribution of the vectorized core and of the four classification thresholds individually (Section VI).

II. Related Work

The platforms available for the analysis of taxonomic and ecological data organized as contingency tables are examined here along three axes: the ecological question they address, their accessibility profile, and whether they expose any entropy-based habitat-preference index.

A. Methods for Habitat Preference and Assemblage Structure

BIOMAPPER [10],[11] implements ecological-niche factor analysis (ENFA), which characterizes the niche of a single species by contrasting the environmental conditions at occupied sites with the global distribution of those conditions.

MaxEnt [12] addresses a related problem, species distribution modelling from presence-only records, by fitting a maximum-entropy distribution over environmental covariates. Both are concerned with where a species occurs as a function of environmental predictors; neither takes a species-by-habitat abundance matrix as its primary input, and neither produces a per-species preference index derived from the entropy of the abundance distribution itself.

The ordination family addresses assemblage structure rather than single-species niches. CANOCO [17] and PC-ORD [16] provide constrained and unconstrained ordination behind graphical interfaces, while vegan [13] and ade4 [15] provide the same method family, and a considerably wider one, as R packages. These methods summarize the joint variation of species and sites on a small number of latent axes, and a species' position on those axes reflects its association with the sites that load on them. This is informative, but it is not a preference index: the axis scores are defined relative to the ordination solution for the table as a whole, they change when species or sites are added or removed, and they do not map onto an interpretable scale bounded between generalist and specialist.

iNEXT [18] and the Hill-number framework [22] quantify diversity at the level of the assemblage rather than of the individual species, and are concerned with sampling completeness rather than with habitat preference. ProCURVE [14] computes habitat preferences, but for aquatic organisms specifically and from a different formulation. GBIF [19] is an occurrence-data infrastructure rather than an analysis platform.

The DP index [5] differs from all of these in that it is computed independently for each species, from the entropy of that species' own abundance profile across habitats, normalized by log2(J). It is bounded, it is invariant to the composition of the rest of the table, and it admits a direct reading: a value near 0 for a species distributed evenly across all habitats, approaching 1 for a species confined to a single one. None of the platforms surveyed computes this quantity, and none produces the APD. Table I summarizes the comparison.

The negative entries in the DP and APD columns of Table I are not a shortcoming of these packages, each of which is well established within its own method family; they are precisely the gap that motivates the present work.

TABLE I. Comparison of Ecological Analysis Platforms with Respect to the DP Index and the APD

Platform	Primary method family	DP index	APD plot	Programming required	Installation	Licence
BIOMAPPER [10],[11]	ENFA / niche factor analysis	No	No	No	Desktop	Free
MaxEnt [12]	Maximum-entropy SDM	No	No	No	Desktop (Java)	Free
vegan [13]	Ordination, diversity	No	No	Yes (R)	R package	GPL
ProCURVE [14]	Aquatic habitat preference	No	No	No	Desktop	Free
ade4 [15]	Multivariate / duality diagram	No	No	Yes (R)	R package	GPL
PC-ORD [16]	Multivariate community analysis	No	No	No	Desktop	Commercial
CANOCO [17]	Constrained ordination	No	No	No	Desktop	Commercial
iNEXT [18]	Diversity interpolation/extrapolation	No	No	Yes (R) or web	R package or web	Free
EcoAPD (this work)	Entropy-based preference classification	Yes	Yes	No	None (web)	Free

B. Accessibility as a Methodological Constraint

The second axis is orthogonal to method, and concerns who is able to run it. vegan and ade4 are free, actively maintained and computationally capable, but they presuppose fluency in R: a user must be able to import a matrix, call the appropriate function with the appropriate arguments, and interpret an object returned at the console. CANOCO and PC-ORD remove that

requirement through graphical interfaces, but are commercially licensed and locally installed, which places them beyond the reach of under-resourced institutions. BIOMAPPER is free and graphical, but restricted to ENFA and no longer actively developed.

This matters because the population that would use the APD method consists largely of field ecologists whose training is taxonomic and ecological rather than computational. Quantitative training among early-career ecologists is known to be limited [8], and the difficulty of translating published ecological methods into runnable code is a recognized bottleneck in ecoinformatics [9],[20]. Accessibility is therefore not a convenience feature but a determinant of whether a method is used at all, which is the most plausible explanation for the APD remaining dormant for four decades despite being theoretically available throughout.

TABLE II. LIMITATIONS OF EXISTING PLATFORMS AND THE CORRESPONDING ECOAPD DESIGN DECISIONS

Limitation identified in existing tools	Consequence for the ecologist	EcoAPD design decision
No DP/APD implementation in any surveyed platform	Method unusable without bespoke coding	Direct implementation of (1)–(4)
Programming fluency presupposed (vegan, ade4)	Excludes non-computational ecologists	Browser interface, no installation
Commercial licence (PC-ORD, CANOCO)	Excludes under-resourced institutions	Free and openly accessible
Method version must be chosen before analysis	Commitment made without sight of the results	Dual computation at project creation
Re-analysis requires full recomputation.	Iterative exploration is costly	Hybrid file/relational persistence with hash-based recompute avoidance
Input assumed clean and regular.	Real survey matrices fail, sometimes silently	Dedicated validation engine

C. Design Requirements Derived from Existing Tools

Read together, the two axes above yield a set of requirements that shaped the architecture of EcoAPD. Table II maps each limitation identified above onto the design decision it motivated.

Two of these decisions warrant emphasis, since they are what distinguish EcoAPD from a straightforward web wrapper around the published algorithm. The first is dual computation: because both APD versions are derived at project creation, the user is never required to choose between them before seeing either, which removes the analytical commitment overhead borne by users unfamiliar with the method. The second is hash-based recompute avoidance in the persistence layer: a hash of the abundance matrix together with its column mapping identifies whether an uploaded dataset is materially unchanged, so that re-uploading the same data does not re-trigger the pipeline.

III. REMINDER: APPLICATION OBJECT (DP AND APD)

The APD method consists of analyzing flora or fauna assemblages using a diagram in which the species are positioned according to both their abundance, as a common descriptor in ecology, and their degree of preference (DP). The latter is measured for each species using an adapted form of the Shannon entropy formula [21], calculated on the series of its abundances in a set of habitats. The APD contains a significant reference curve, the exclusive species curve (ESC), and is subdivided into six preference zones (or classes) by five half-hyperbolas of the same form, defined by the equation Y = (ax+b)/(cx+d) Fig. 1. Four of these curves have their vertices on ESC points of significant abscissa (DP), namely 0.1, 0.3, 0.5 and 0.75; these values were defined on the basis of experimental datasets. The fifth curve, which delimits the Accidental class in the non-normalized version, is not defined by a DP target but is derived from the habitat count and a theoretical abundance scale (Section V-E).

In practice, the data on which the main calculations are based are presented in a table of I lines (species) and J columns (habitats), which is commonly used in assemblage studies. The APD method consists of calculating the DP for each species, using the following formula. Then, the APD is simply plotted for each assemblage (= habitat).

$$D_{iJ=}1 - \frac{S_i}{\log_2(J)} \tag{1}$$

$$S_i = -\sum \frac{n_{ij}+1}{\sum(n_{ij}+1)} \log_2 \frac{n_{ij}+1}{\sum(n_{ij}+1)} \tag{2}$$

$\boldsymbol{n_{ij}}$: abundance of species i in habitat j; $\mathbf{log_2(J)}$: maximum value of S_i (same abundance value ($\boldsymbol{n_{ij}}$) in all habitats; $\boldsymbol{D_{ij}}$: degree of preference (DP).

For an exclusive species (which exists only in one habitat), Formula (1) becomes as follows:

$$S_i = \frac{n_{ij}+1}{n_{ij}+J}\left(\frac{n_{ij}+1}{n_{ij}+J}\right) + (J-1) * \frac{1}{n_{ij}+J}\left(\frac{1}{n_{ij}+J}\right) \tag{3}$$

In the normalized version of the APD [7], mainly used when abundances are low, the Y-axis (abundances) is transformed using Formula (4); then, all the other components of the APD, including the species positions, are recalculated. In this version, we chose another hyperbola formula, $Y = a^2/2x$, while we used only five hyperbolas, the accidental species being interpreted as foreign to the assemblage.

$$A_{ii} = \frac{\log 2(a_{ij}+1)}{\log_2(\max(a_{ij})+1)} \tag{4}$$

$\boldsymbol{a_{ij}}$ = abundance of the species i in the assemblage j;

max(aij) = the maximum abundance recorded across the entire species × habitat matrix, i.e. a single global value and not a per-species maximum.

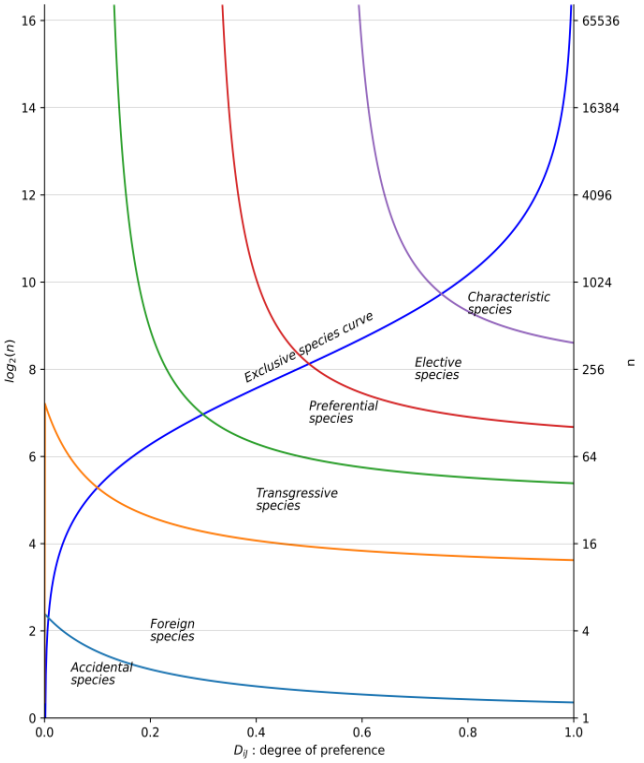


Fig. 1. APD model for assemblage analyses [5].

IV. METHODS

A. Inputs and Outputs of the Application

1) Inputs: data and their format: The basic data in each study case (named here 'project') are stored in a single table of I rows (species) and J columns (habitats) as indicated below (Table III). This matrix, commonly used in the ecology of assemblages [15],[16],[22], contains two types of identifiers of both rows (species name and code) and columns (habitat name and code), which respectively fill the two first columns and the two first rows, the abundances (data) are readable starting from the cell 'row 3, column 3'.

Data format, which is verified by the system, should respect the following conditions:

- input table contains a minimum of 10 habitats (J > 9) and at least one species (I ≥ 1);
- data are numerical (natural or decimal numbers) and positive; empty cells are filled with '0', and empty rows or columns are deleted; for decimal numbers, only a dot '.' is accepted as separator of decimals, while separators of thousands are not accepted.

In addition, the habitat and species names should be easily recognized by the project holder, while their codes are shorter, as strings or numbers, and very useful for displaying species in the APD, avoiding overlapping of long names.

TABLE III. GENERAL FORMAT OF THE BASIC DATA MATRIX FOR IMPLEMENTING THE APD METHOD

		Habitat 1	…	**Habitat j**	…	**Habitat J**
	Codes	H1	…	Hj	…	HJ
Species 1	S1	n1.1	…	n1.2	…	n1.J
….	….	…				
Species i	S2	ni.1	…	n2.2	…	n2.J
….	….	…				
Species I	SI	nI.1	…	nI.2	…	nI.J
Species name*:** names that will be displayed in the APD; Codes: short but meaningful identifiers that will be displayed; ***I: number of species, non-limited number; ***J***: number of habitats, preferably greater than 10; ***nij***: abundance of the species i in the habitat j, as a positive real number.						

2) Outputs: The outputs are defined in consideration of the major users' needs from the EcoAPD application, which consists of plotting, for each habitat, the assemblage APD in its two versions (original and/or normalized), with the species list of each preference class. However, the system also displays the data table, including the DP values for each species Fig. 2, mainly for verification and downloading purposes.

In addition, users can generate the histogram of the species DP values, as well as the abundance profile of a selected species across all habitats. These diagrams help to understand the position of the species in the different assemblage APD.

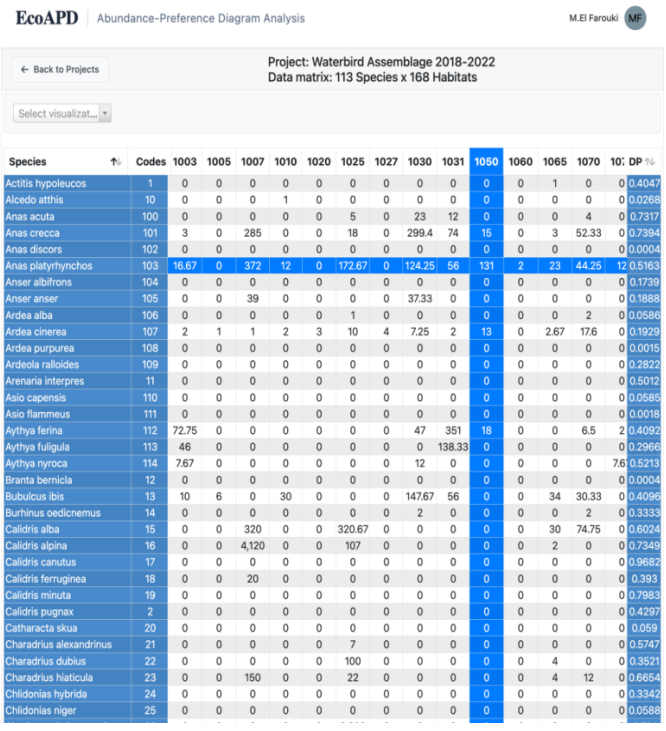


Fig. 2. Project data view: 'species x habitat' abundance matrix with calculated DP.

B. Test Data Used

The application has been tested, in terms of functionalities and validity of the outputs, by some researchers that we invited to test the EcoAPD application. However, for experimenting with different development steps, we used two different real datasets, provided here as supplementary materials, named SM1 and SM2:

- SM1: Benthic invertebrates of the Sebou river;
- SM2: International winter counts of waterbirds in Morocco (average of 2018-2022 counts);

These data are borrowed from different ecological studies that provide data adapted to APD use, in both original and normalized forms.

C. Used Technologies

This application is based on the algorithm already published by [6] and normalized by [7].

The application was developed using Python [23] for server-side logic and calculations, combined with standard web technologies HTML [24], CSS [25], and JavaScript [26] for the user interface.

The main computational components include:

- Mathematical libraries: NumPy [27] and Pandas [28],[29] handle matrix operations and data manipulation, enabling efficient processing of species-habitat abundance matrices.
- Data storage: Project data, user accounts, and computed results are stored in a MySQL relational database [30], ensuring data integrity and enabling rapid retrieval of data and previous analyses.
- File handling: Uploaded datasets (Excel files) are stored locally, each assigned a unique identifier generated from the project ID and timestamp to ensure traceability and avoid naming conflicts.
- Visualizations: Matplotlib [31] and Apache ECharts [32] generate static and interactive graphs, respectively, enabling efficient processing and visualization of species-habitat abundance matrices.
- Export capabilities: The system generates publication-quality PDF reports containing APD diagrams, data tables, and project metadata.

Together, these technologies enable a streamlined workflow for ecological data analysis, ensuring both computational efficiency and a seamless end-user experience.

V. System Design and Implementation

A. System Architecture

1) Preamble: A first version of this system was created to implement the original version of the APD tool [5]; we further submitted it for major modifications in order to implement the normalized version of this tool [6]. In addition to multiple tests of the web application done by our staff, some potential users were asked to write essays about this system, using their own data, and their remarks were highly beneficial for improving it.

The application implements a robust database architecture, storing abundance matrices and all computed results, including DP values, ESC, hyperbola set, and species points. This persistence layer enables rapid generation and visualization without re-computing the entire pipeline, particularly beneficial for large datasets or iterative analysis workflows.

A distinctive feature of the application is its automatic dual-dataset generation: when a project is created, the system automatically processes both normalized and non-normalized APD datasets and stores them in a database Fig. 3. This design avoids initializing a method before starting analyses.

2) Architectural components: The application is built with technologies that allow optimizing the calculation speed, while offering a fluid user experience Fig. 4.

a) User interface (application layer): The web interface is responsive, enabling dynamic interaction with the application through

- Authentication portal: Secure login and registration with email verification and reCAPTCHA protection [33];
- Project dashboard: Intuitive management of analysis projects with search and metadata editing;
- Data configuration wizard: Excel upload with preview and flexible mapping of species/habitat identifiers;

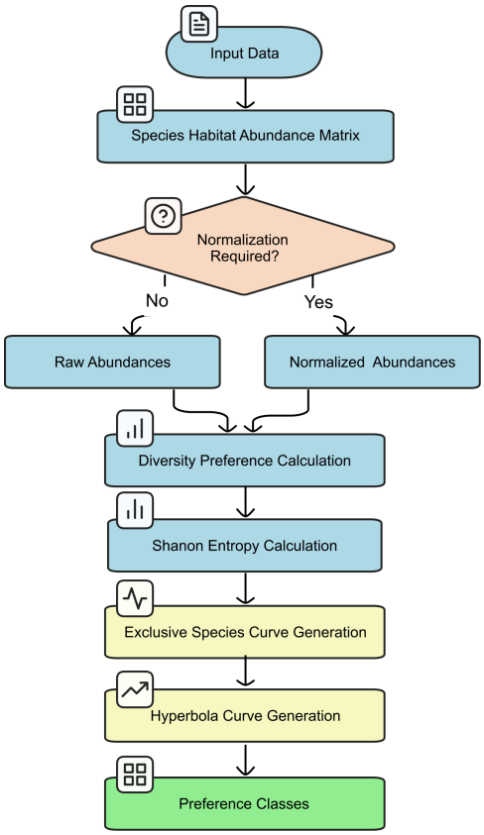


Fig. 3. Automatic generation and storage of normalized and non-normalized APD datasets during project creation.

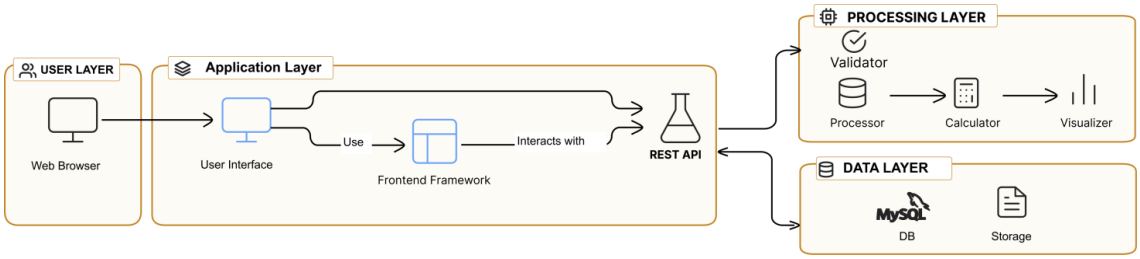


Fig. 4. System architecture diagram showing the main components and their interactions.

- interactive APD viewer: Dynamic diagrams supporting zoom, pan, and species labeling with real-time parameter adjustments;
- exploration tools: Sortable tables for abundances and DP values, species abundance profiles, and DP histograms;
- multi-format export: Download options for PNG images, and publication-ready PDF and Excel reports.

b) Server-side processing (processing layer): The server component handles requests from the user interface, processes data, and manages database interactions. It provides the following main functionalities:

- data validation engine: verifies uploaded files against format requirements, cleans invalid entries, and prepares matrices for analysis;
- DP computation module: calculates degree of preference values and generates both normalized and non-normalized datasets automatically;
- curve generation: computes ESCs and hyperbola sets for assemblage classification;
- security layer: enforces authentication, session management, and input sanitization across all operations.

All operations include appropriate security checks and error handling to ensure data integrity.

c) Data storage (data layer): The initial system, announced by [6], used pickle-based file storage to persist computed DP objects, enabling rapid visualization without recomputation; this approach provides fast state restoration but lacks query flexibility and data integrity. In the actualized version of EcoAPD, we migrated to a relational database, where Excel files are stored in the dedicated project Files, while all computed data and metadata are stored in the MySQL database Fig. 5. This system, combining file-based and database storage, provides flexibility and performance, data integrity, and query capabilities. In addition, this architecture enables selective data retrieval, supports concurrent access, maintains referential integrity through foreign keys, and facilitates schema evolution Fig. 6.

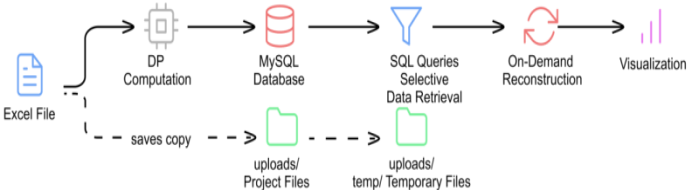


Fig. 5. Data persistence and retrieval workflow.

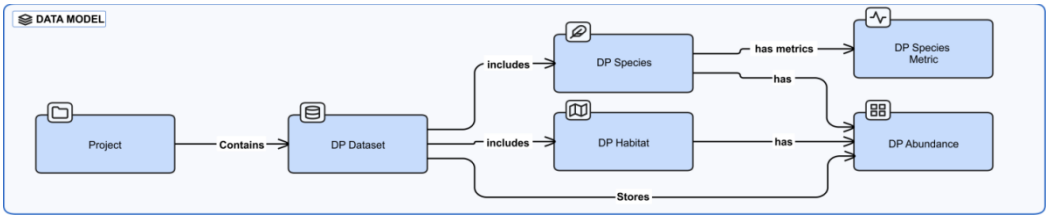


Fig. 6. Database structure showing relationships between user, project, and computed data.

d) Mathematical processing: The application uses optimized mathematical libraries that enable rapid processing of large ecological datasets. The main computations include: 1) *abundance data*: species-habitat tables are stored as NumPy arrays for fast, bulk operations; 2) *DP calculation*: Shannon entropy and related math run efficiently on entire arrays; 3) *normalization*: log2 scaling and value adjustment are done in parallel for all data; 4) *diagram data*: APD curves and coordinates are generated in batches.

B. User Workflow

The system provides a structured workflow for ecological data analysis Fig. 7, which shows the key interaction sequences during a typical analysis session. This workflow is designed to minimize technical barriers to users unfamiliar with computational complexity, allowing them to focus on result interpretation rather than on data computation.

C. Workflow Steps: Data Processing Pipeline

The pipeline consists of several interconnected stages that transform raw biodiversity data into interpretable visualizations and classification results, which can be edited for publication Fig. 8.

The workflow can be subdivided into three main phases: 1) data preview, validation and upload, 2) selection and configuration of outputs, and 3) generating outputs.

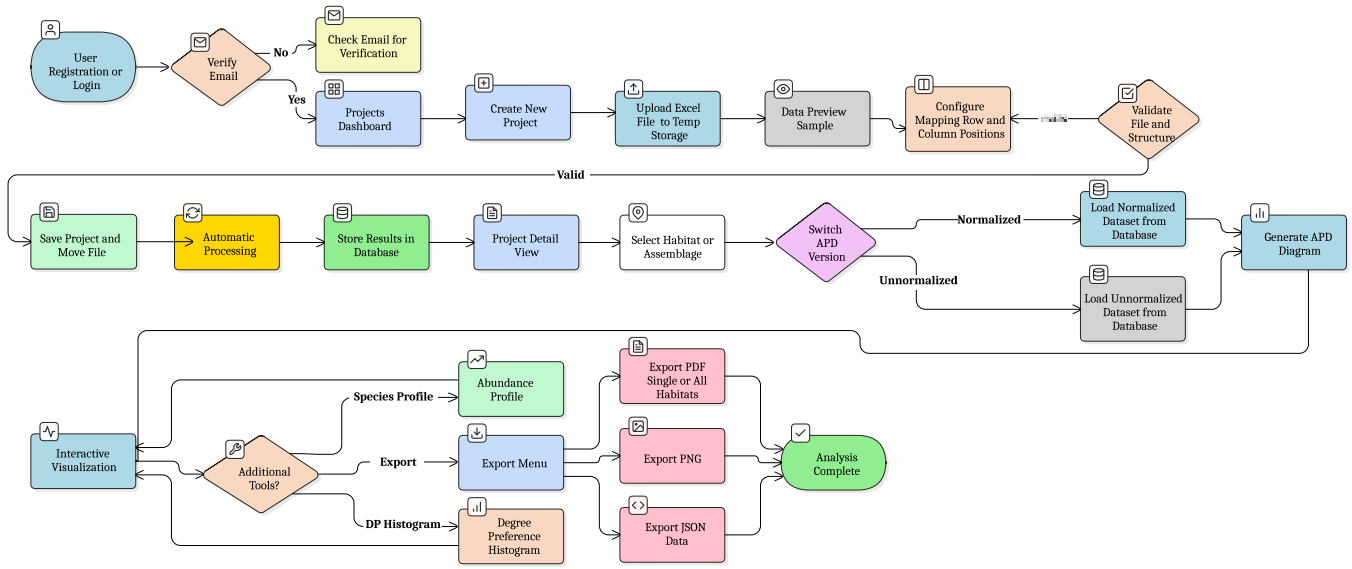


Fig. 7. User workflow diagram: key interaction sequences during a typical analysis session.

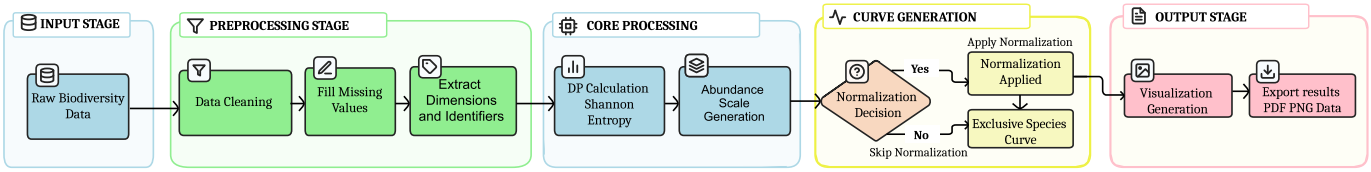


Fig. 8. Data processing pipeline architecture showing the transformation of raw biodiversity data into interpretable visualizations.

1) Data preview, validation and upload: The use of the APD method starts with creating a project, based on a 'species-habitat' matrix of abundances, which is provided in an Excel file. The system architecture is conceived to meet the specific requirements of computed data, knowing that ecological datasets have different sizes and complexity depending on studies. Users are previously prompted to configure their data; however, the dataset is submitted to a validation sequence that is implemented before creating the project. The first sequence consists of displaying the first '5 rows x 5 columns' of the project file Fig. 9, which allows a visual verification that names and codes of both rows and columns are provided.

2) Selecting outputs: Once the data is validated and uploaded, the user is requested to select different analysis types and options that can be edited:

- the most requested output is the APD, for a selected assemblage; it is edited in a desired version (normalized vs. non-normalized), where species are displayed using a selected identifier (name or code), and the reference curves (hyperbolas and ESC) are displayed or hidden Fig. 10;
- DP profile for all species Fig. 11;
- abundance profile of a selected species in all habitats Fig. 12.

3) Generating outputs: The requested APD is displayed and, when needed, it can be saved in image or PDF format. The other graphical outputs related to species (abundance profiles and DP diagram) can be displayed and saved in these same formats, while the specific composition of each preference class of the APD can be edited in Excel format. All these outputs have a printable quality, directly used in a publication.

D. Workflow Implementation

The main processing flow for successfully uploading a file proceeds as follows Fig. 13.

- Once uploaded, the data file is transmitted to the API, which submits it to a file validator; this latter checks both extension (.xlsx/.xls), size (< 100 Mb), file structure, and data format.
- The file is saved in the storage directory, with a unique filename. The API then submits the data to a data processor that checks its required structure and type, deleting non-numeric values and empty rows and columns, filling empty cells with a '0' value, and extracting species and habitat labels. The matrix dimensions are then extracted, as well as row and column identifiers, while columns are converted to strings.
- Once the project is created, the system processes the data simultaneously to compute DP values, the ESC and hyperbolas for both normalized and non-normalized APD methods;
- All project metadata and computed results are stored in the database: species and habitat metadata, abundance matrices (stored in long format), DP values for each

species, ESC and their intersection points with hyperbolas, and project record creation and storage of file path and data mapping.

- At the end of this process, the API returns a success response to the web interface, which displays the data table as a confirmation of the project creation.

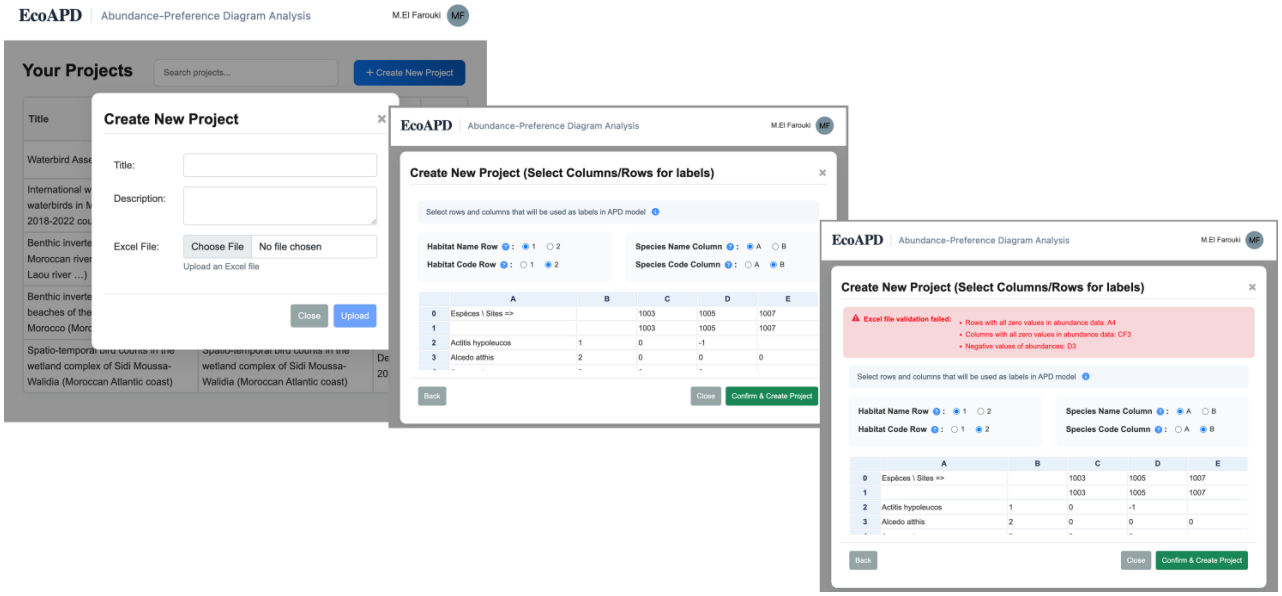


Fig. 9. Upload and validation interfaces.

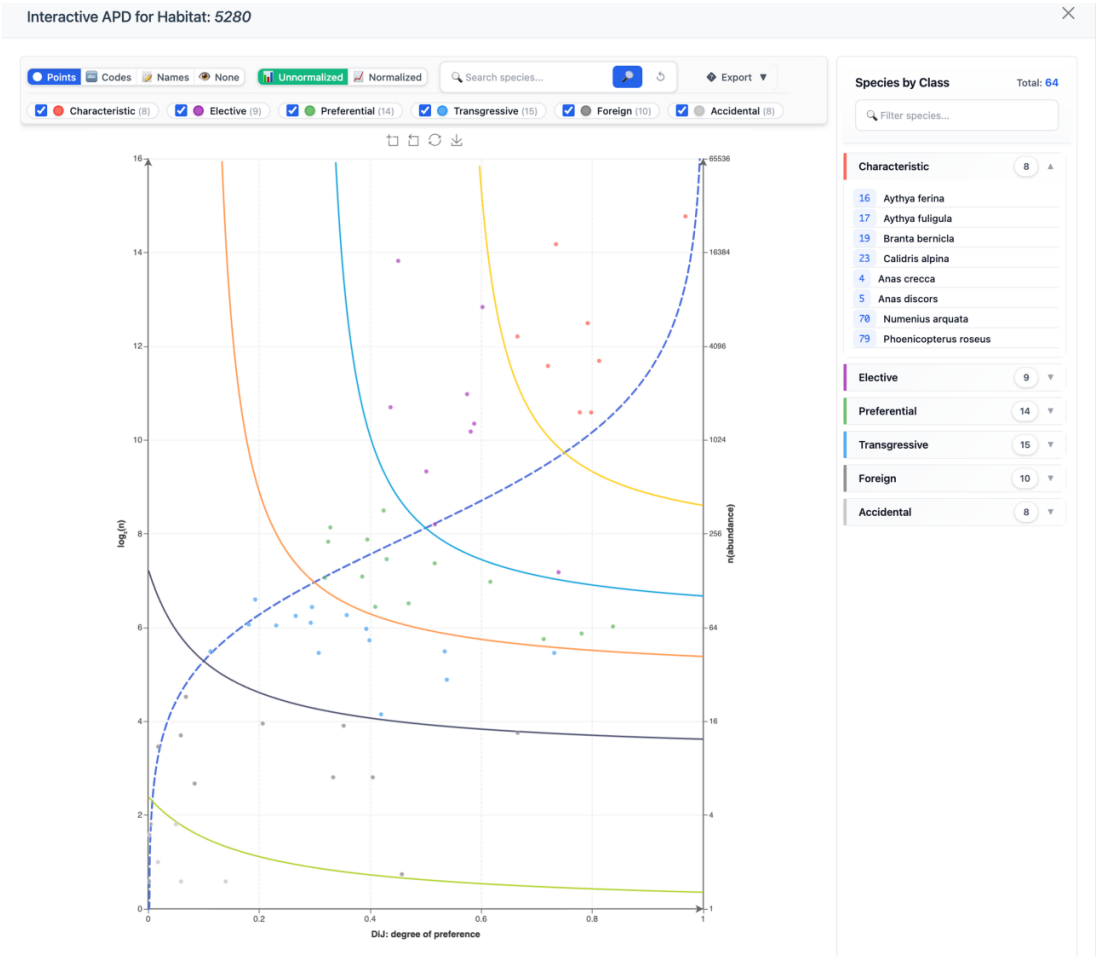


Fig. 10. APD of an assemblage/habitat: Interactive visualization.

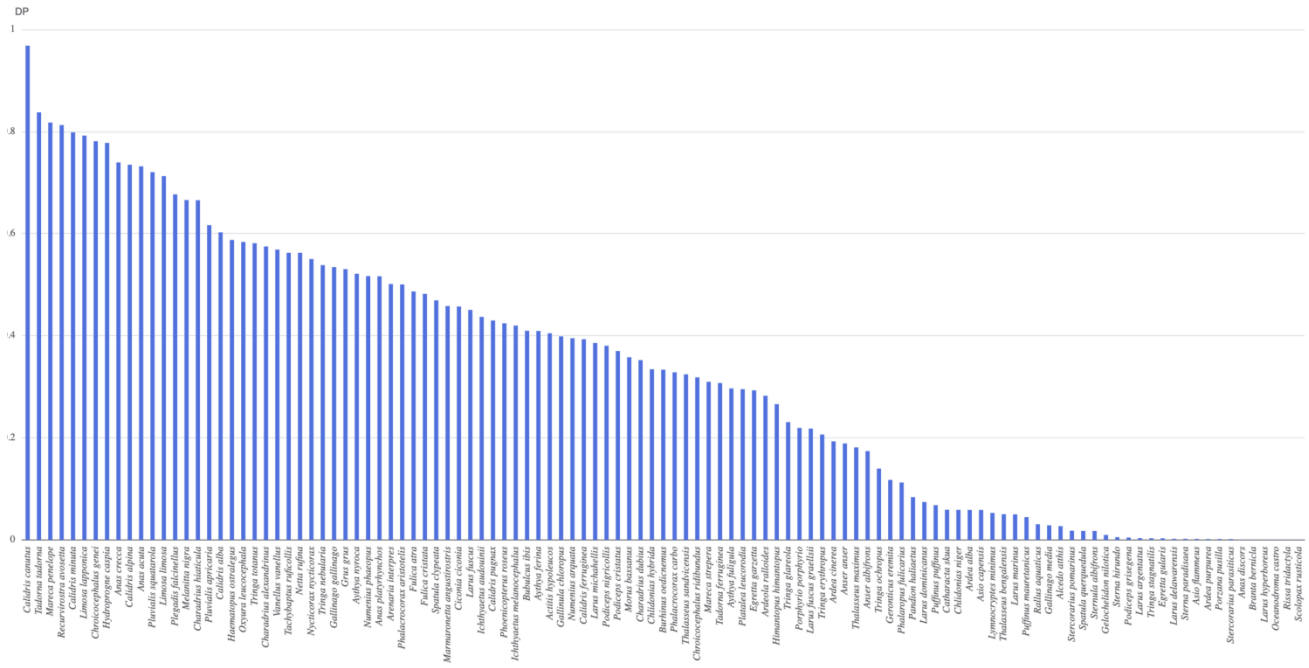


Fig. 11. DP profile for all species.

E. Generation of the Classification Curves (ESC and Hyperbolas)

Theoretical curves (ESC and hyperbolas) for species classification are generated from a computed vector. For the ESC, matrix operations start with Y = esp_array + 1.

The entropy and the DP are then obtained as:

$$Si=\Sigma((Y/\sum(Y)\times\log2(Y/\sum(Y)), \text{ and the DP as } Di=1+Si/\log2(S)$$

The hyperbola curves that delineate the zones corresponding to the species preference classes are generated from their intersections with the ESC, which are predefined by their abscissa (DP value). Four of these curves have vertices at fixed DP targets — 0.1, 0.3, 0.5 and 0.75 — each located by absolute-difference minimization within an asymmetric tolerance band around its target (the 0.1 vertex, for instance, is searched within [0.095, 0.15] and the 0.75 vertex within [0.70, 0.80]) rather than under a single global tolerance. The non-normalized version generates one further curve, the baseline of the Accidental class, which is not defined by a DP target but derived from the habitat count and a theoretical abundance scale; the five curves together delimit the six preference zones. In the normalized version, this baseline curve is omitted, the accidental species being interpreted as foreign to the assemblage, which leaves four curves. In both versions, a reference hyperbola is generated and duplicated by simple translation to obtain the remaining curves.

VI. Evaluation

This section reports the computational environment and benchmarking protocol (VI.A), an ablation of the vectorized mathematical core (VI.B), a sensitivity analysis of the classification thresholds (VI.C), the scalability of the full pipeline (VI.D), and the limits of what has been evaluated (VI.E). EcoAPD is a deterministic analytical pipeline rather than a stochastic simulation or a trained model: it has no sampling procedure, no random seed, and no free hyperparameters, so that, given a matrix, the DP values, the ESC and the hyperbola set are uniquely determined by (1)–(4). The quantities reported below are therefore timings and classification outcomes, not estimates carrying sampling error.

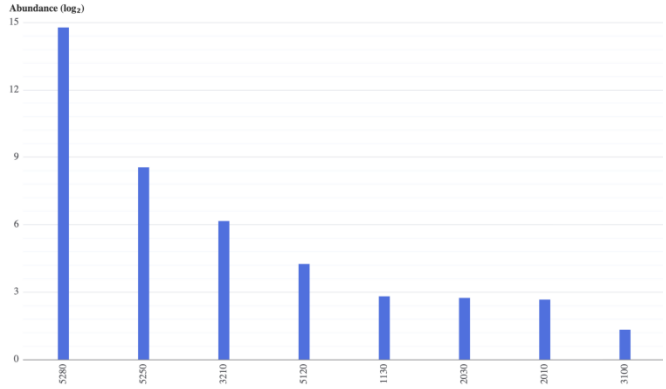


Fig. 12. Abundance profile of a selected species in all habitats.

A. Experimental Setup

Benchmarking hardware: Apple M1 Pro, 8 physical and 8 logical cores, 16 GB RAM, running macOS 26.5.2. This is the development and benchmarking machine on which the reported measurements were produced, and not necessarily the production deployment server.

Software: Python 3.9.10, NumPy 2.0.2, pandas 2.2.3, Matplotlib 3.9.2, openpyxl 3.1.5, Flask 3.0.3, Flask-SQLAlchemy 3.1.1, SQLAlchemy 2.0.36 and PyMySQL 1.1.1, all version-pinned. The database is MySQL 8.0 [30].

Timing protocol: each measurement comprises one discarded warm-up execution followed by five timed executions using time.perf_counter(), reported as the mean. All timings are server-side, measured by direct instrumentation of the processing layer, so that network latency and browser rendering are excluded.

Fixed parameters: the four classification-threshold targets are DP = 0.1, 0.3, 0.5 and 0.75, in both the normalized and the non-normalized variants, each located within an asymmetric tolerance band around its target rather than under a single global tolerance (Section V-E). The ESC is evaluated on ⌈J / 0.0015⌉ points, which scales with the habitat count J rather than being fixed: 24,667 points for SM1 (37 habitats), and 114,000 for SM2 (171 habitats).

Datasets: the two real datasets SM1 and SM2 described in Section IV-B, and a synthetic series of uniform-random non-negative integer abundances, seeded at 10, 50, 100, 500, and 1000 species crossed with 5, 10, and 20 habitats, used for the size sweep of Section VI-B.

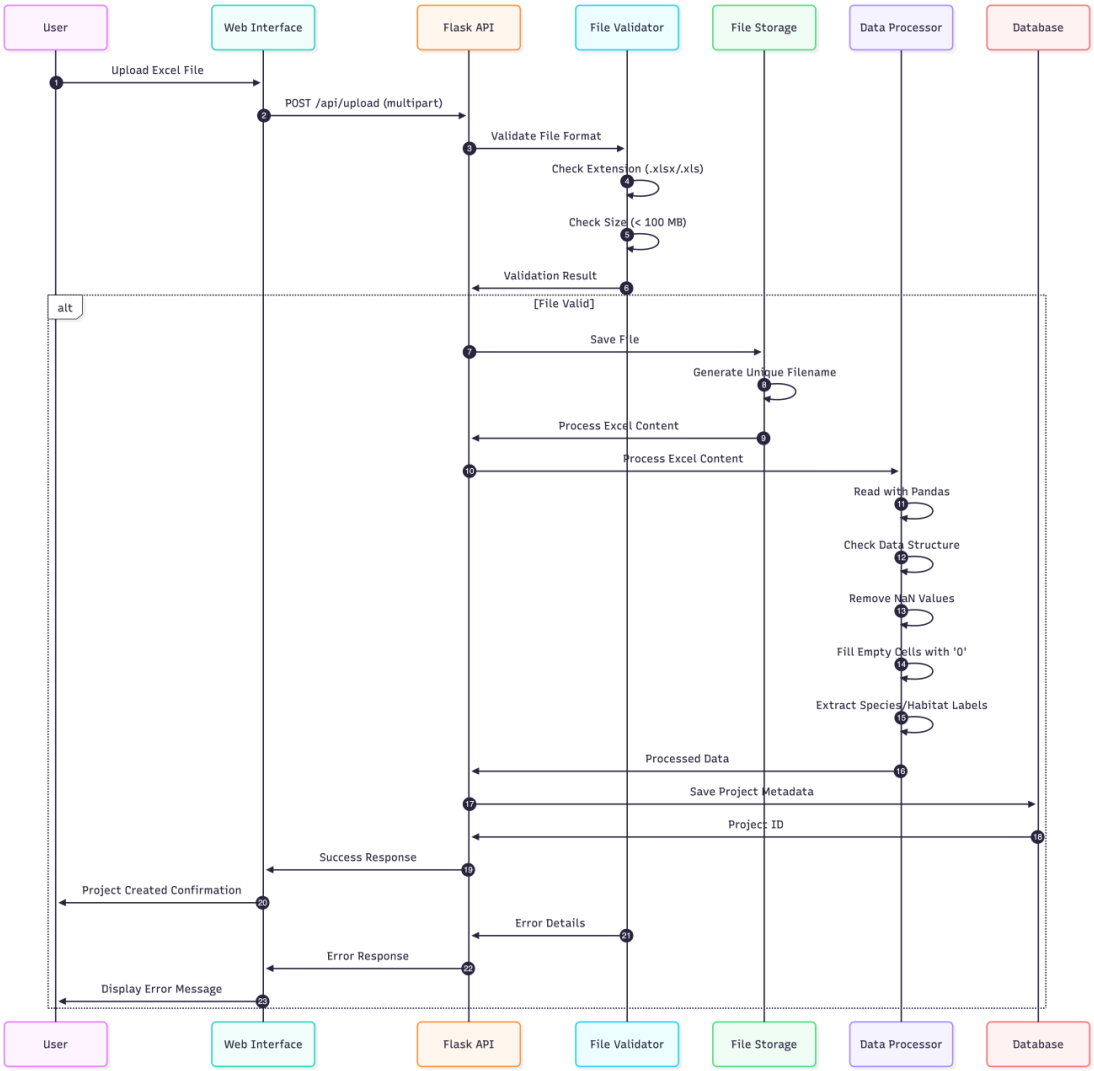


Fig. 13. Sequence diagram illustrating the data upload and validation process architecture.

B. Ablation: Vectorization of the Preference-Index Core

The preference-index computation (calculate_ipd) was originally a per-species loop containing a nested per-habitat loop. It has been replaced by a single vectorized NumPy computation over all species at once, verified to reproduce the output of the original implementation to floating-point precision (np.allclose) on both real datasets and against the existing regression test suite. Table IV reports the cost of the two implementations over the synthetic size sweep.

The two implementations scale differently rather than merely differing by a constant factor. The cost of the legacy implementation grows approximately in proportion to the species count, a hundredfold increase in I raising it 66-fold, whereas the vectorized implementation rises only 2.8-fold over the same range and is effectively flat at the matrix sizes encountered in practice. The speedup is consequently not a fixed figure but an increasing function of dataset size, which is the property that matters for the largest matrices the method is intended to serve.

TABLE IV. ABLATION OF THE VECTORIZED MATHEMATICAL CORE: COST OF THE PREFERENCE-INDEX COMPUTATION WITH AND WITHOUT VECTORIZATION (MEAN OF FIVE TIMED EXECUTIONS)

Species count (I)	Legacy scalar loop (ms)	Vectorized (ms)	Speedup
10	2.59	0.73	3.5×
50	9.54	0.85	11.3×
100	17.94	0.82	21.8×
500	86.05	1.43	60.1×
1000	171.62	2.01	85.6×

C. Ablation: Sensitivity of the Classification Thresholds

The four threshold targets are the only quantities in the pipeline that are set by convention rather than derived from the data, and the preference class assigned to a species depends on them. Each was therefore perturbed independently by ±0.05 and ±0.10 on the SM1 dataset, and the resulting change in each species' assigned class was measured against the default-threshold baseline (Table V). The reported case is habitat AA, in which seven species have non-zero abundance, and the baseline classification is three Preferential, three Transgressive, and one foreign species.

TABLE V. SENSITIVITY OF THE PREFERENCE CLASSIFICATION TO PERTURBATION OF THE FOUR THRESHOLD TARGETS (SM1, HABITAT AA, SEVEN SPECIES WITH NON-ZERO ABUNDANCE)

Threshold target	Perturbation	Species reclassified
0.1	+0.05, +0.10, −0.10	1 / 7 (14.3%)
0.1	−0.05	0 / 7
0.3	±0.05, ±0.10	0 / 7
0.5	−0.05, −0.10	1 / 7 (14.3%)
0.5	+0.05, +0.10	0 / 7
0.75	±0.05, ±0.10	0 / 7
All four (tolerance widened)	±0.05, ±0.10	0 / 7

The 0.1 and 0.5 thresholds each independently drive a reclassification for this dataset, whereas 0.3 and 0.75 are stable across the tested range. The direction of the effect differs between them: at 0.1, three of the four tested perturbations reclassify a species, all except the smallest downward shift; at 0.5, only lowering the threshold reclassifies a species; raising it does not. That widening the search tolerance around all four targets simultaneously reclassifies no species — in contrast to the individual target-value shifts, which do — indicates the classification is sensitive to where each threshold's target value is set, but robust to the width of the tolerance band used to locate it on the exclusive species curve; these are two distinct senses of "sensitivity" worth keeping separate. The thresholds are, on this evidence, a defensible convention rather than a tuned parameter, though the analysis is reported for one habitat of one dataset and should not be read as a general stability result.

D. Scalability of the Full Pipeline

Table VI reports end-to-end pipeline latency on the two real datasets.

TABLE VI. FULL-PIPELINE LATENCY ON THE TWO REAL DATASETS (MEAN OF FIVE TIMED EXECUTIONS)

Dataset	Species (I)	Habitats (J)	ESC points ⌈J/0.0015⌉	Full-pipeline latency
SM1 (benthic invertebrates, Sebou)	93	37	24,667	511 ms
SM2 (waterbird winter counts)	114	171	114,000	2.32 s

Once the preference-index core is vectorized, latency is dominated by the habitat count rather than by the species count. Between SM1 and SM2, the species count rises by a factor of only 1.2, while the habitat count rises by a factor of 4.6 and latency rises by a factor of 4.5 — closely tracking the habitat count rather than the species count. The explanation is that the remaining non-vectorized step, the construction of the abundance scale, carries a cost that scales with the ⌈J / 0.0015⌉ evaluation points of the ESC, and therefore with J. This step was outside the scope of the present optimization and is reported here as a known limitation rather than left implicit: it is the natural target for further work, and the point at which additional effort would yield the largest return.

VII. DISCUSSION

Even though the APD method is forty years old, its use remained difficult, mainly because of the complex calculations that it involves. The algorithm published in [6] was a significant step towards implementation; however, given that very few ecologists are familiar with algorithm implementation, the use of this method still seemed difficult. The EcoAPD application was therefore a necessary step for ensuring the use of the APD method. To further facilitate this use in community ecology, the application has been made accessible to a wide audience, from any web browser, without installing any software.

From a scientific perspective, this application represents a significant advance in ecoinformatics [22], primarily as one of the rare tools designed to handle datasets in the form of contingency tables, given that such tables are widely used in ecology. Compared to traditional statistical approaches to analyzing these kinds of tables, EcoAPD offers another significant technical advancement, namely its ability to convert such tables into diagrams that provide a pedagogic visual representation of all species in a community, classifying them according to their degrees of preference. In addition, EcoAPD is available free of charge, giving a large number of scientists and biodiversity managers the opportunity to access a key method in ecology.

In summary, the EcoAPD application can address several key challenges in ecological informatics, mainly through:

- clear interface, which makes it easy to use with no need for specific informatics knowledge;
- accessibility, as it provides a web-based interface, eliminating barriers related to software installation and programming expertise;
- reproducibility, via its standardized implementation, ensuring the needs of transparency of the system across different studies [34], promoting a certain scientific rigor;

- data validation, as the system ensures data control, preventing errors that could compromise analytical results;
- flexibility, mainly due to the simultaneous treatment and display of normalized and non-normalized APDs, allowing users to select rapidly the APD version most suited to their data;
- scalability: vectorization of the preference-index core reduces its cost by a factor that rises with dataset size, reaching 85.6× at 1000 species (Table IV), and the full pipeline processes the waterbird winter monitoring matrix in 2.32 s (Table VI), the residual cost being dominated by habitat count rather than species count (Section VI-D);
- data persistence, ensured by the use of a database-based mechanism that enables rapid output generation and visualization, which is especially advantageous for large datasets and iterative analysis workflows

VIII. Conclusion

The biggest gain from this research output lies in the facility and time saving offered by the EcoAPD application in implementing the APD method, through complex calculations performed behind the interface. Moreover, the output requests benefit from graphical tools that make it user-friendly, practically with no need for additional help. These tools could bring great help in developing similar applications.

Processing of the 114 × 171 waterbird matrix completes in 2.32 s (Table VI). Once the preference-index core is vectorized, full-pipeline latency is dominated by the habitat count rather than the species count; the remaining non-vectorized step — construction of the abundance scale, whose cost scales with ⌈J / 0.0015⌉ — accounts for the greater part of the residual time. Further efficiency in calculation, storage and plotting therefore remains attainable, and the application can be improved in several ways:

- interface improvement: primarily by making better basic information o n DP concept and the APD method, making the tool more educational and accessible;
- performance optimization: by vectorizing the construction of the abundance scale, which Section VI-D identifies as the dominant remaining cost, and by reducing the duration of the plotting process, particularly for datasets with many habitats;
- enhanced features: additional visualization options, statistical summaries, and export formats could further enhance the application's utility;
- integration capabilities: future versions could integrate other ecological analysis tools and databases, offering more comprehensive ecosystem analyses.

The immediate perspectives consist of improving the interface, in response to the ecologists' requirements. The main output expected from the use of the APD, however, is a reinterpretation of former assemblage studies that were based on abundances alone. In the expectation of a wide use of the APD method in conservation and management decisions, the intention is to make it accessible to broader ecology and nature management communities.

Supplementary Materials

The following supporting information can be downloaded at: 10.5281/zenodo.21504481. Table SM1: Matrix of data used to provide experimental examples for implementing the APD method; results of water insect surveys in the Upper Sebou River (Middle Atlas, Morocco). Table SM2: Matrix of data used to provide experimental examples for implementing the APD method; results of waterbird winter monitoring in Morocco in 2018-2022. Script SM3: Benchmark and ablation scripts, with their raw CSV outputs, underlying the evaluation reported in Section VI (Tables IV-VI).

References

[1] Elith, J., & Leathwick, J. R. (2009). Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annual Review of Ecology, Evolution, and Systematics, 40(1), 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

[2] Hirzel, A. H., & Le Lay, G. (2008). Habitat suitability modelling and niche theory. Journal of Applied Ecology, 45, 1372–1381. https://doi.org/10.1111/j.1365-2664.2008.01524.x

[3] MacArthur, R. H. (1972). Geographical ecology: Patterns in the distribution of species. Harper & Row Publishers.

[4] Gilbert, N. A., Amaral, B. R., Smith, O. M., & others. (2024). A century of statistical ecology. Ecology, 105, e4283. https://doi.org/10.1002/ecy.4283

[5] Dakki, M. (1987). Running Water Ecosystem of the Upper Sebou (Middle Atlas, Morocco) (in French) (Vol. 42).

[6] Dakki, M., Ennakri, M., El Farouki, M., Ziti, S., & Zoglat, A. (2025). A Graphical Method for Analysing Species Assemblages Combining Habitat Preference and Abundance. African Journal of Ecology, 63, 1–13. https://doi.org/10.1111/aje.70013

[7] Ennakri, M., Dakki, M., El Farouki, M., Ziti, S., & Zoglat, A. (2025). The Abundance-Preference Method for Assemblage Analysis: A Normalized Diagram with Algorithm Implementing the Method. Diversity, 17(9), 598. https://doi.org/10.3390/d17090598

[8] Barraquand, F., Ezard, T. H., Jørgensen, P. S., & others. (2014). Lack of quantitative training among early-career ecologists: A survey of the problem and potential solutions. PeerJ, 2, e285. https://doi.org/10.7717/peerj.285

[9] Vitolo, C., Elkhatib, Y., Reusser, D., Macleod, C. J. A., & Buytaert, W. (2015). Web technologies for environmental big data. Environmental Modelling & Software, 63, 185–198. https://doi.org/10.1016/j.envsoft.2014.10.007

[10] Hirzel, A. H., Hausser, J., Chessel, D., & Perrin, N. (2007). Biomapper 4.0 [Computer software]. Laboratory for Conservation Biology, University of Lausanne.

[11] Hirzel, A. H., Hausser, J., Chessel, D., & Perrin, N. (2002). Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data? Ecology, 83(7), 2027–2036. https://doi.org/10.1890/0012-9658(2002)083[2027:ENFAHT]2.0.CO;2

[12] Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. Ecological Modelling, 190(3), 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

[13] Oksanen, J., Simpson, G. L., Blanchet, F. G., & others. (2022). vegan: Community Ecology Package, Version 2.6-4. R Foundation for Statistical Computing.

[14] Mengin, N., Thorel, A., Liebig, H., & Segura, G. (2002). ProCURVE Technical Note: Software for Calculating Habitat Preferences for Aquatic Organisms | International Journal of Limnology – Annales de Limnologie (in French) - International Journal of Limnology, 38(4), 329–338. https://doi.org/10.1051/limn/2002027

[15] Dray, S., & Dufour, A.-B. (2007). The ade4 Package: Implementing the Duality Diagram for Ecologists. Journal of Statistical Software, 22(4). https://doi.org/10.18637/jss.v022.i04

[16] McCune, B., M. J, & Mefford. (1998). PC-ORD: Multivariate Analysis of Ecological Data. The Bulletin of the Ecological Society of America, 79(2), 144–145. https://doi.org/10.2307/20168234

[17] Ter Braak, C. J. F., & Šmilauer, P. (2012). Canoco Reference Manual and User's Guide: Software for Ordination, Version 5.0. Microcomputer Power.

[18] Hsieh, T. C., Ma, K. H., & Chao, A. (2016). iNEXT: An R package for rarefaction and extrapolation of species diversity (Hill numbers). Methods in Ecology and Evolution, 7, 1451–1456. https://doi.org/10.1111/2041-210X.12613

[19] Edwards, J. L., Lane, M. A., & Nielsen, E. S. (2000). Interoperability of Biodiversity Databases: Biodiversity Information on Every Desktop. Science, 289(5488), 2312–2314. https://doi.org/10.1126/science.289.5488.2312

[20] Michener, W. K., & Jones, M. B. (2012). Ecoinformatics: Supporting ecology as a data-intensive science. Trends in Ecology & Evolution, 27(2), 85–93. https://doi.org/10.1016/j.tree.2011.11.016

[21] Shannon, C. E. (1948). A mathematical theory of communication. The Bell System Technical Journal, 27(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

[22] Hill, M. O. (1973). Diversity and Evenness: A Unifying Notation and Its Consequences. Ecology, 54(2), 427–432. https://doi.org/10.2307/1934352

[23] Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. CreateSpace.

[24] World Wide Web Consortium. (2014). HTML5: A Vocabulary and Associated APIs for HTML and XHTML.

[25] World Wide Web Consortium. (2011). Cascading Style Sheets (CSS)

[26] Ecma International. (2023). ECMAScript® 2023 Language Specification.

[27] Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

[28] McKinney, W. (2010). Data Structures for Statistical Computing in Python—SciPy Proceedings. https://proceedings.scipy.org/articles/Majora-92bf1922-00a

[29] The pandas development team. (2024). pandas-dev/pandas: Pandas (Version v2.2.3) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.3509134

[30] Oracle Corporation. (2023). MySQL 8.0 Reference Manual.

[31] Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9, 90–95. https://doi.org/10.1109/MCSE.2007.55

[32] Li, D., Mei, H., Shen, Y., Su, S., Zhang, W., Wang, J., Zu, M., & Chen, W. (2018). ECharts: A declarative framework for rapid construction of web-based visualization. Visual Informatics, 2(2), 136–146. https://doi.org/10.1016/j.visinf.2018.04.011

[33] von Ahn, L., Maurer, B., McMillen, C., Abraham, D., & Blum, M. (2008). reCAPTCHA: Human-Based Character Recognition via Web Security Measures. Science, 321(5895), 1465–1468. https://doi.org/10.1126/science.1160379

[34] Powers, S. M., & Hampton, S. E. (2019). Open science, reproducibility, and transparency in ecology. Ecological Applications, 29(1), e01822. https://doi.org/10.1002/eap.1822

Blockchain-Based Trusted Data Sharing Framework for Secure and Auditable AI Model Training

Sherif Amin
Department of Computer Science-College of Engineering and Computer Science, Jazan University, Jazan, Saudi Arabia

***Abstract*—AI models are highly effective and reliable based on integrity, provenance, and controlled usage of training data. The use of data sharing between several stakeholders in collaborative and distributed settings poses risks of unauthorized access, data tampering, and data poisoning that undermine trust and accountability in AI systems. To overcome these obstacles, this study introduces a blockchain-based trusted data sharing framework for secure and auditable AI model training pipelines. The solution consists of using a permissioned blockchain to store irreversible metadata of the datasets, cryptographic hashes, ownership, and access policy, and place the actual data off-chain to achieve scalability. Smart contracts provide a mechanism to automate registration of datasets, provide access control on a smart and intelligent basis, and record data access and data modification events to provide verifiable data provenance and end-to-end traceability. Before model training, the integrity of the data is automatically checked against the records on the chain to ascertain the use of authorized and undamaged data. Experimental evaluation in a simulated permissioned-blockchain environment indicates that the framework, applied as a practical design heuristic, can enhance data transparency, integrity, and auditability with reasonable performance overhead. The proposed solution helps to build credible and responsible AI, especially in data-sensitive environments, by ensuring that models are trained only on verified and authorized data.**

***Keywords*—Blockchain; AI; secure data sharing; data integrity; smart contracts; auditable AI pipelines; access control; AI model training; distributed systems**

I. Introduction

The fast development of artificial intelligence (AI) has changed many industries, including healthcare diagnostics and autonomous systems, as well as financial projections and smart governance. The quality, integrity, and origins of training data have become identified as key indicators of the reliability, fairness, and accountability of AI models. Credible AI systems will need datasets to be true, unaltered, and traceable throughout their lifecycle. Nonetheless, the conventional centralized data archives and data governance models do not offer immutable records, transparent provenance, and strong access controls to support secure, collaborative AI development. Existing data pipelines in AI have a high risk of data poisoning, unauthorized access, and non-transparent processing history that can affect the model results and confidence of stakeholders. Traditional database logging and centralized audit trails can be modified easily and are vulnerable to insider attacks and do not have verifiable provenance tracing across the distributed stakeholders. These restrictions are a major hindrance in sectors that have stiff regulations like the medical field and finance, where accountability and integrity of data are obligatory. To address these issues, there is a growing interest among researchers in decentralized ledger technologies that can facilitate transparent and tamper-evident logs of the original datasets, their transformations, and usages [1].

Decentralization, immutability, and distributed consensus create a strong basis to handle trust and auditability issues in AI data systems, which is defined by blockchain technology. Stakeholders can ensure that data utilized to train a model is original and free of any alterations by storing metadata of datasets, cryptographic hashes, ownership information, and access controls on a blockchain collectively. Self-executing code of smart contracts that are placed on blockchain systems can be used to put in place access controls, automate policy adherence, and record all transactions with datasets in a read-only and unalterable fashion [2].

The most recent development has started the formalization of blockchain-based frameworks of AI data provenance and auditability. As an example, a blockchain-based data provenance system captures lineage, preprocessing operations, annotation processes, and model modifications in the direction of providing end-to-end traceability in AI pipelines, with provenance disagreements in a variety of machine learning tasks having significantly declined [3]. Another paper presents a blockchain-based data integrity system that aims to guarantee ethical AI development through tracking data ownership, transformations, and compliance metadata and shows the enhancement in audit performance in terms of efficiency [4].

These pieces explain that blockchain not only maintains the integrity of data but also facilitates regulatory adherence and ethical responsibility of the AI workflows. In addition to fixed logging, decentralized registries are being established as the key elements of secure digital ecosystems integrating blockchain, AI, and cloud computing in order to operate transparently and in a trustworthy manner. A systematic review points to the recent strategies of integration, in which blockchain will increase trust in distributed AI applications and enable scalable cloud deployments, as well as transparent operations among stakeholders [5]. The same type of benefits are reported in other research works in applied fields, including cybersecurity, which they identify as blockchain-based mechanisms that can guarantee data integrity and dynamically adaptive protection in real-time AI systems [6].

Though these are good trends, there are several difficulties. The majority of blockchains are not scalable and not performance-oriented, and therefore, it is not possible to store big datasets on them. Instead, it is convenient to deploy a hybrid architecture that utilizes off-chain data storage tools (e.g.,

InterPlanetary File System or distributed databases) to store data objects and store the metadata of the hash and provenance on-chain. In addition, smart contract design must find a compromise between fine-grained access control and efficiency in computations, and the consensus mechanisms must be designed in such a way that they do not entail excessive latency and resource usage. These architectural problems must be imagined to enable unrestricted application in data-sensitive collaborative environments [7].

Also, the integration of blockchain into the AI information pipelines will not have a direct positive effect on establishing trust without clear governance systems. Data ownership, contribution records, and revocation policies are procedures that should be gathered collectively. Smart contracts are also capable of automating compliance functions but must also be designed in a manner that can accommodate the shift in legal and ethical needs [8]. This is particularly imperative because AI regulation models such as explainability-based, fairness-based, privacy-based, and robustness-based models gain increased popularity in the global space [9]. Blockchains can be used as a compliance-checking tool by providing audit trail immutability, but this must be linked to institutional governing systems and legal governance.

This study outlines a blockchain-based trusted information sharing model in a bid to ensure data integrity, provenance, access control, and auditability of AI training data. The proposed framework has a permissioned blockchain, which stores the metadata of immutable datasets, their cryptographic fingerprints, and ownership, in addition to access control, and stores the datasets themselves off-chain, which is required as this is needed to overcome scalability issues. The smart contract is responsible for registering the data sets, ensuring access controls on the data, and maintaining an audit trail of any data usage during the process. Before initiating the process of training the model, the system provides stakeholders with an opportunity to verify whether the data provided is genuine and has no issues with any past usage, which instills confidence in the stakeholders.

The novelty of the current study lies in the solution to the problems faced by prior work in the area of AI in blockchains. By introducing hybrid storage, strict access controls, and automatic audit capabilities into blockchain technology, the proposed approach provides a way of achieving secure sharing of institutional data that requires protection under almost any circumstance, such as in the health care, financial sector, and research collaborations. The point of this approach is not only to implement a functional system but also to provide a realistic solution to the problem where AI needs to be trusted.

Based on this idea, this study seeks to achieve the following three objectives:

- Development of a blockchain-based data sharing architecture to preserve the integrity and traceability of AI training data and enforce access control policies over it.
- Design of smart contracts to support execution of granular access control policies and provide auditability of the process of using data within an AI pipeline.
- Evaluation of the proposed architecture through experiments to test its effectiveness in terms of data integrity, auditability, and scalability.

II. Related Work

With increasing sizes and complex structures of machine learning algorithms comes the issue of whether AI solutions can actually be trusted, which once was only a philosophical question, now becoming a subject for serious scientific investigations. With machine learning algorithms using the data gathered by several sources and organizations, issues related to data integrity, traceability, and accountability gain a completely new level of significance, especially in industries such as medicine, finance, and autonomous systems, where inaccurate information can cause more than just technical problems.

Conventional centralized databases have already proved their inability to handle these challenges. As a matter of fact, traditional databases can easily suffer from data alterations with no way of determining the author or time of the alteration; the lineage of the data is not always clear, and there is rarely any means to track down access and usage [10].

That is why the concept of blockchain has proven to be rather useful in such situations. The distributed nature of blockchain, its resistance to changes, and inherent ability to leave an audit trail make it a perfect solution.

A. Blockchain for Data Provenance and Transparency in AI Pipelines

An increasing number of works have been devoted to the application of blockchain as a technological means for tracing dataset lineage in machine learning processes and verifying the authenticity of the datasets throughout this process. One comprehensive approach proposes capturing end-to-end provenance for machine learning pipelines through immutable, smart-contract-based records [3]. Traceability is significantly increased in this case, and the decentralized nature of the validation process appears to be successful in preventing provenance disputes as well as enforcing compliance rules during training.

Similarly, Jalbani et al. [8] develop a quality model specifically for blockchain-based data provenance in AI processes. In contrast, they place particular emphasis on the fact that immutability is crucial for achieving verifiable data provenance in machine learning processes. Overall, the studies presented can be used to draw an important conclusion: blockchain technologies cannot merely serve as an alternative means of storing and sharing data – they must provide the missing layer of verification as well.

B. Smart Contracts for Access Control and Secure Sharing

A part of fine-grained access control and automated policy enforcement in a decentralized environment has been proposed to be provided by smart contracts. The systematic study of smart contract applications in the recent past has established that they have the capability of carrying out complex access controls, data licensing policies, and usage policies without the presence of central intermediaries, but that scale and security remain open concerns [9], [12]. Applied research on secure information exchange shows a linear relationship between blockchain

functionality, such as decentralization, consensus, and cryptographic security, and significant data-security considerations, such as confidentiality, integrity, provenance, and auditability in AI-driven systems [2]. There are also federated learning and collaborative AI schemes that are based on blockchain-audited ledgers, where input and updates to models are recorded on a permissioned ledger and enhance data provenance and update integrity across distributed nodes [10]. This is part of the bigger project in making the work processes of AI training more transparent and trustworthy, attaching the key events to the immutable ledger.

C. Integration of Blockchain with AI and Cloud Systems

It is noteworthy that the convergence of technologies like blockchain, artificial intelligence, and cloud computing has been receiving much attention as a key strategy for creating secure, trustworthy systems that can process distributed workloads.

The findings from systematic reviews on this issue indicate that there are true synergies between the three technologies, where each technology makes up for any deficiencies in the other two, thereby allowing an integration of all three that can bring about levels of transparency, decentralization, and scalability unachievable by each technology alone [4].

The importance of these developments to this research lies in the practical application. Practical use of AI models often takes place in environments where the model does not run on one isolated platform but across different institutional, infrastructural, and data storage platforms. In this case, hybrid models are designed to account for this challenge in a bid to meet performance requirements without compromising verifiable trust.

D. Security, Privacy, and Ethical Considerations

Apart from basic issues of provenance and access control, new concerns have emerged regarding the need to find a proper balance between privacy and efficiency when applying blockchain in AI data sharing solutions. This problem is addressed explicitly in analyses of data management and governance in blockchain-based systems, which consider the potential for blockchain technology to significantly decrease the likelihood of sensitive data leakages within highly competitive businesses, as well as discuss associated trade-offs concerning performance losses. Embedding blockchain into high-throughput AI data flows is not an entirely smooth procedure since the limited scalability of blockchain solutions currently hinders efforts in solving the issue [7]. Systematic reviews in the healthcare and other controlled fields have highlighted how blockchain can enhance AI-based systems by ensuring data integrity, protection of data ownership, and auditability, and also point to the ongoing challenges in the field of interoperability and adherence to privacy protocols [11].

E. Research Gaps in Current Literature

Despite the increased interest, several gaps exist. Most current frameworks do not perform full workload performance evaluations or do not offer fine-grained access control and policy enforcement beyond a basic time-stamped account of their operation. Numerous other papers do not include scalability studies of integrating blockchain with large-volume AI data pipelines and often do not discuss the issue of how to ensure auditability without introducing excessive latency. As an example, even though there are blockchain-based data pipeline proposals, the empirical comparison of their effects on training throughput and latency with large-scale AI workloads has not been conducted. Furthermore, there has been a proposal of hybrid storage models (metadata on-chain and data off-chain storage) that are usually explained in theory but not proven in practice within multi-stakeholder environments. Although permissioned blockchains are a solution to the issue of privacy and governance, the questions of consensus mechanism, trust assumptions, and cross-organizational interoperability are not completely studied in the literature. In this way, it is evidently possible to suggest a blockchain model to adapt to trustworthy AI data sharing that incorporates verifiable provenance, access control based on smart contracts, scalable audit tools, and an experimental evaluation to quantify the performance and overhead in realistic training settings.

III. Mathematical Modeling

Let:

$D = \{d_1, d_2, \dots, d_n\}$ be the dataset used for AI model training.

$H(\cdot)$denote a cryptographic hash function.

Mbe the AI model trained on D.

$B = \{b_1, b_2, \dots, b_k\}$ be the blockchain ledger storing metadata and transactions.

SCdenote the set of smart contracts enforcing access control.

$U = \{u_1, u_2, \dots, u_m\}$ represent users or stakeholders accessing datasets.

A. Data Integrity Modeling

For each dataset element d_i, its integrity is verified using a hash stored on the blockchain:

$$h_i = H(d_i) \tag{1}$$

$$\text{Verify: } H(d_i) = h_i^{\text{on-chain}} \tag{2}$$

If $H(d_i) = h_i^{\text{on-chain}}$, the data is untampered. Otherwise, a tampering event is flagged.

B. Provenance Tracking

Each dataset $d_i$has a provenance record $P(d_i)$stored on-chain:

$$P(d_i) = \{h_i, o_i, t_i, u_i\} \tag{3}$$

where,

a. h_i= hash of d_i

b. o_i= owner or origin of the data

c. t_i= timestamp of data creation or update

d. u_i= user or smart contract interaction log

The full provenance of the dataset D is [8]:

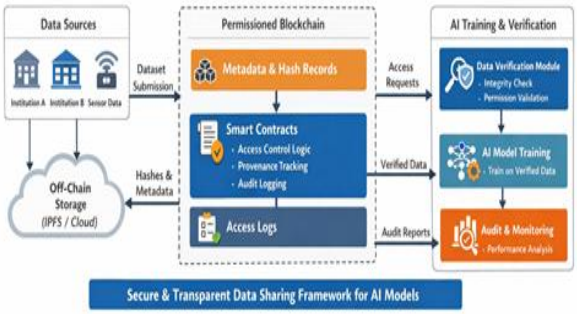


Fig. 1. Block diagram of blockchain-based trusted data sharing framework for AI model training.

C. Access Control Modeling

Define a permission function $\pi: U \times D \rightarrow \{0,1\}$such that:

$$\pi(u_j, d_i) = \begin{cases} 1, & \text{if user } u_j \text{ has permission for } d_i \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

Smart contracts enforce:

$$\text{Access}(u_j, d_i) = \begin{cases} \text{Allow}, & \pi(u_j, d_i) = 1 \\ \text{Deny}, & \pi(u_j, d_i) = 0 \end{cases} \quad (6)$$

All access events are logged:

$$L=\{(u_j,d_i,t_i) \mid \text{Access attempted at } t_i\} \quad (7)$$

D. AI Model Training Under Blockchain Constraints

Model M is trained on verified datasets:

$$M=\text{Train}(D\text{verified}),\ D\text{verified}=\{di \in D \mid H(di)=\text{hion-chain} \wedge \pi(uj, di)=1\} \quad (8)$$

This ensures:

- Only authorized datasets are used.
- Data integrity is preserved.
- The training process is fully auditable.

E. Performance and Overhead Metrics

Blockchain storage overhead is given by [7],

$$S_B = \sum_{i=1}^{n} \text{size}(h_i + P(d_i)) \quad (9)$$

Now, verification latency is given by,

$$T_{\text{verify}} = \sum_{i=1}^{|D|} T_{\text{hash}}(d_i) + T_{\text{on-chain_lookup}}(h_i) \quad (10)$$

F. Smart Contract Execution Time

$$T_{SC} = f(|U|, |D|, \text{complexity_rules}) \quad (11)$$

IV. Methodology

The blockchain-based trusted data sharing system of AI model training presented in Fig. 1 shows the implementation of integrity, provenance, and access control at the system level. It starts with many data sources, such as an institution, sensors, or an enterprise database, and provides datasets that are supposed to be used as AI training. These datasets are initially formulated by generating cryptographic hashes, which are the digital prints of the data. Any change in the datasets will automatically create a new hash value, and this will automatically identify tampering or unauthorized editing. Since datasets are typically large, the real data is off-chain, e.g., stored in a secure cloud provider or distributed files such as IPFS. Metadata and cryptographic hashes of each dataset are stored only on a permissioned blockchain so that it is scalable yet can verify the authenticity of the data. This architecture combines the advantages of both blockchain and storage overhead reduction and computational overhead reduction. The framework is centered on the blockchain layer where records cannot be modified, and they store the provenance of datasets, ownership data, timestamps, version history, and access logs. The blockchain has smart contracts that automate essential operations. They implement fine-grained access control policies that are determined by prescribed roles, attributes, or even organizational policies, and only authorized users are allowed to access particular datasets. Smart contracts additionally keep provenance, meaning who contributed, altered, or read the information, and accountability audit logs. All access to the dataset is recorded on-chain in a transparent manner, giving an account of the data usage that is impossible to modify. Also, smart contracts can be configured to perform automated compliance checks to monitor compliance with regulatory and organizational policies on data governance. The block diagram of the proposed methodology is presented in Fig. 2.

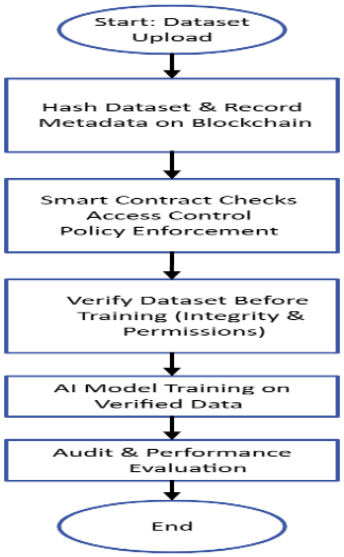


Fig. 2. Flow chart of proposed methodology.

A. Pseudocode:

Blockchain-Governed AI Training Data Framework

Require: Dataset collection D = {d1, d2, . . . , dn} from multiple sources

Ensure: Trained AI model M and audit logs

1: // **Phase 1: Dataset Hashing & Blockchain Recording**

2: **for** each dataset di in D do

3: hi ← Hash(di) {Generate cryptographic hash} 4: Metadata$_i$ ← {Owner, Timestamp, Description} 5: Store (hi, Metadata$_i$) on Blockchain

6: Store di in Off-Chain Storage

7: **end for**
8: // **Phase 2: Smart Contract Access Control**
9: **for** each access request (u_j, d_i) do
10: **if** Permission(u_j, d_i) = **TRUE then**
11: Allow access
12: Log event on Blockchain
13: **else**
14: Deny access
15: Log event on Blockchain
16: **end if**
17: **end for**
18: // **Phase 3: AI Training with Verified Datasets**
19: $D_{verified} \leftarrow \emptyset$
20: **for** each dataset d_i in D **do**
21: **if** Hash(d_i) =Blockchain_hash(d_i) **and** Permission(user , d_i) = **TRUE then**
22: $D_{verified} \leftarrow D_{verified} \cup \{d_i\}$
23: **end if**
24: **end for**
25: Train AI model M using $D_{verified}$
26: // **Phase 4: Audit & Performance Evaluation**
27: Record verification latency, smart contract execution time, and storage overhead
28: Generate audit reports and performance metrics
29: **return** Trained AI model M and audit logs

V. Results and Discussion

This section shows the proposed experimental evaluation of the Blockchain-based Trusted Data Sharing Framework, assessing its effectiveness in enabling secure and auditable AI model training under simulated conditions. Framework testing is implemented with respect to data integrity checking, system performance, access control efficiency, scalability, and auditability, which are directly related to architectural block diagram components.

A. Experimental Setup

To validate the proposed framework, a prototype was implemented with the following configuration:

- Blockchain: Permissioned blockchain (e.g., Hyperledger Fabric–like environment)
- Smart Contracts: Access control, provenance tracking, audit logging
- Off-Chain Storage: IPFS / secure cloud storage
- Datasets: Simulated AI training datasets from multiple institutions
- AI Model: Supervised learning model trained on verified datasets
- Metrics Evaluated:
  - Data integrity verification accuracy
  - Access authorization success rate
  - System latency
  - Storage overhead
  - Audit log completeness

B. Data Integrity Verification Results

The main objective of the framework is to make sure that training is done only on tamper-free data sets. The hash value stored within the blockchain for each data set is compared with the hash of the extracted data set.

The total performance of data integrity verification is given in Table I. Fig. 3 illustrates the verification results of both uncorrupted and tampered datasets, while Fig. 4 is a graph demonstrating the results of accuracy regarding the number of datasets.

TABLE I. Data Integrity Verification Performance

Number of Datasets	Verified Datasets	Tampered Datasets Detected	Verification Accuracy (%)
50	50	0	100
100	99	1	99.0
200	198	2	99.0
500	495	5	99.0

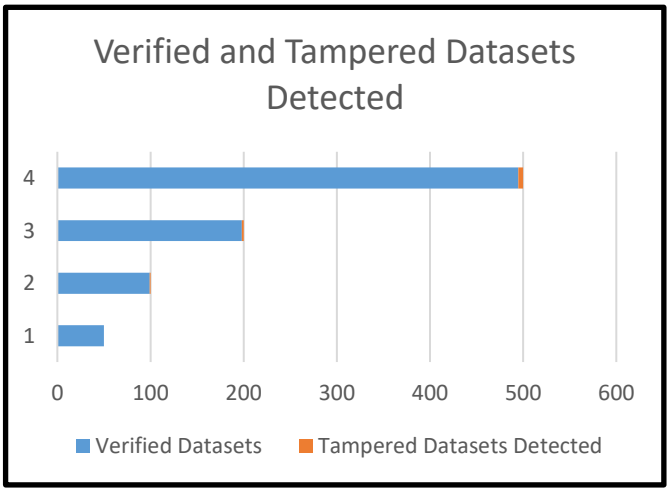


Fig. 3. Verified vs tampered datasets detected.

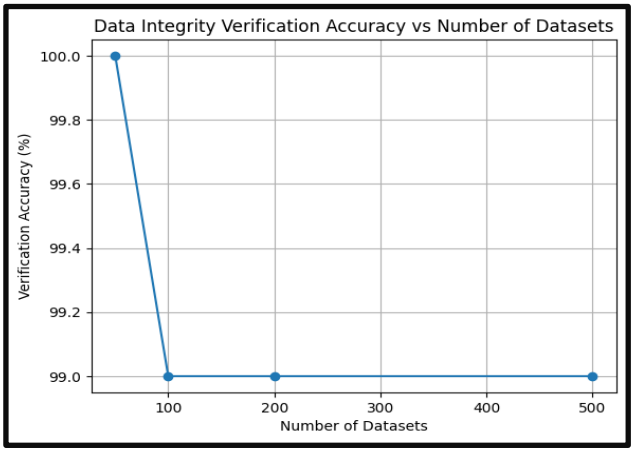


Fig. 4. Number of datasets vs verification accuracy.

The results demonstrate that the blockchain-stored hashes reliably detect any modification to off-chain data. Integrity verification accuracy holds above 99% even as dataset volumes increase, which speaks to the resilience of combining cryptographic hashing with immutable blockchain records — the two mechanisms reinforce each other in ways that neither achieves alone. It should be emphasized that near-perfect detection is a deterministic consequence of cryptographic hashing in a correct implementation rather than an empirical discovery; these results are therefore best interpreted as validation that the framework behaves as designed.

To stress-test this further, controlled data tampering was deliberately introduced as part of the experimental evaluation. Rather than relying solely on clean data scenarios, a subset of datasets was intentionally modified after their cryptographic hashes had already been recorded on the blockchain. The tampering took two forms:

- refined adjustments to a small percentage of data values
- insertion or deletion of individual records
- changes modest enough to go unnoticed in a less rigorous system, yet consequential enough to serve as a meaningful test of the framework's detection capability.
- Modification of metadata fields

The type of data manipulation selected represents practical situations of potential attacks – whether by accidental data damage, internal sabotage, or intentional data tampering.

C. Access Control and Authorization Evaluation

Access control is managed via smart contracts based on permission requirements for roles. Unauthorized access was prevented in all simulated scenarios (a 100% success rate). This outcome is expected by design: smart contract enforcement deterministically denies out-of-policy requests, so this rate should be read as validation that the implementation behaves as specified rather than as an empirical finding. Why does it matter? The use of contracts in enforcing access permissions effectively eliminates the problem with a single point of failure in the case of centralized systems, reducing the risk of misuse of data in training greatly (Table II).

TABLE II. ACCESS CONTROL EFFECTIVENESS

Access Requests	Authorized Requests	Unauthorized Attempts	Access Control Success Rate (%)
200	182	18	100
500	457	43	100
1000	914	86	100

To rigorously evaluate the access control mechanism, unauthorized access attempts were simulated by submitting requests from users and nodes that deliberately fell outside the permission boundaries defined in the smart contract policies. Rather than testing only obvious violations, the scenarios were designed to reflect the obvious forms of access misuse that are more likely to occur in practice: requests sent by users with no ownership claim over the target dataset attempting to perform operations that exceeded the privileges assigned to a given role; access requests initiated outside the time windows that the governing contracts permitted.

Together, these scenarios represent a realistic cross-section of how access control failures tend to manifest in data-sensitive environments. The resulting distribution of authorized and unauthorized attempts is illustrated in Fig. 5.

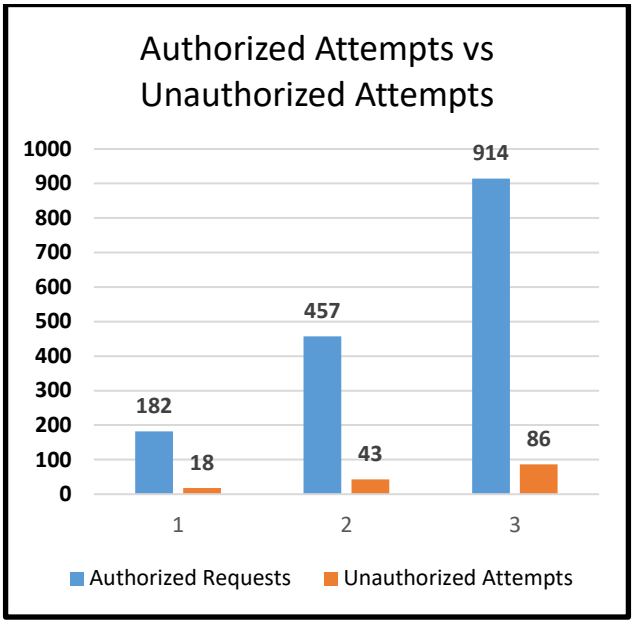


Fig. 5. Authorized attempts vs unauthorized attempts.

D. System Latency Analysis

Given that the framework introduces additional verification steps inherent to its security design, the resulting impact on system latency was measured and is presented in Fig. 6.

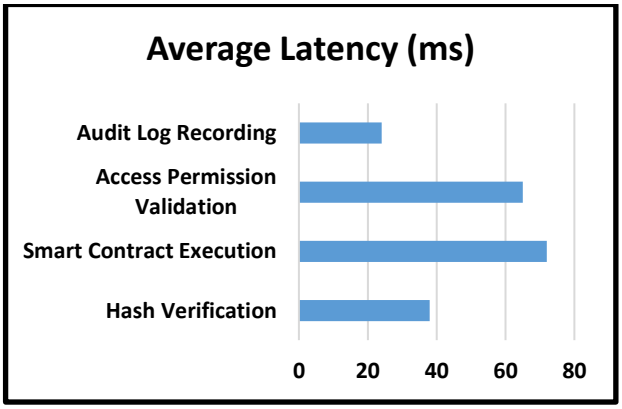


Fig. 6. Average latency per operation.

While blockchain operations inevitably carry some computational overhead, the latency observed across all test scenarios remains well within the bounds that practical AI training pipelines can accommodate. This points to a meaningful finding: the security and auditability gains delivered by the framework do not come at the expense of system performance — a balance that is often assumed to be difficult to achieve but is demonstrated here to be attainable.

E. Storage Overhead Evaluation

In this regard, scalability will be retained since only the hash values and metadata corresponding to each one will be stored in the blockchain and not entire datasets, which will preserve high performance levels despite an increasing amount of data.

TABLE III. STORAGE OVERHEAD COMPARISON

Storage Type	Data Stored	Storage Size
On-Chain	Metadata + Hashes	~3–5 KB/dataset
Off-Chain (IPFS)	Full Dataset	100 MB–2 GB
Traditional System	Full Dataset On-Chain	Not Scalable

Recording only hashes and metadata in the blockchain in place of full datasets allows for low storage requirements without compromising the ability to verify data. The fact that the blockchain contains only hashes and metadata allows for natural scalability of the blockchain even when working with the large volumes of training data needed by production AI systems. Storage requirements for different storage types have been considered in Table III below.

F. Auditability and Provenance Tracking

The system will log all interactions with the data, beginning from its first submission through its eventual use for training the AI model.

TABLE IV. AUDIT LOG COMPLETENESS

Activity Type	Logged Events	Missing Logs
Dataset Upload	100%	0
Data Access	100%	0
Model Training Usage	100%	0
Permission Changes	100%	0

The fixed audit trail ensures full transparency and accountability. Unlike traditional systems where logs can be altered, blockchain-based logs provide tamper-proof evidence, which is critical for regulatory compliance and forensic analysis. Table IV shows 100% audit log completeness.

G. Impact on AI Model Training Trustworthiness

Moreover, the audit trail created using the framework is more than just a simple tracking system; the audit trail created using the framework cannot be silently modified or erased.

This is because traditional logging techniques are based on mechanisms that traditionally involve a security hole in the form of access control vulnerabilities, which have frequently proven to be flawed in the past, whether by accident or intentional tampering. Logging onto the blockchain avoids this loophole, creating data that would stand up to scrutiny even under rigorous auditing and investigative processes.

This is evident from the experimental findings. According to Table IV, audit log completeness attains 100%, without any instances of either an incomplete or missing event entry recorded.

TABLE V. AI MODEL TRAINING COMPARISON

Training Approach	Data Trust Level	Model Reliability
Traditional Training	Unverified	Medium
Proposed Blockchain Framework	Verified	High

There is a direct relationship between the quality of data that enters a training pipeline and the trustworthiness of the models that emerge from it. By ensuring that only verified and properly authorized datasets are admitted to the training process, the framework actively guards against the kinds of inputs — biased, tampered, or otherwise unauthorized — that quietly undermine model reliability long before any failure becomes visible. In high-stakes applications, where the consequences of a compromised model extend well beyond technical performance, this kind of upstream data governance is not a convenience but a necessity. It should be noted that the comparison in Table V is qualitative; quantitative machine learning performance metrics, such as accuracy, precision, recall, and F1-score under clean and poisoned training conditions, are deferred to future empirical work.

H. Comparative Analysis with Existing Approaches

The benefit of the framework under study in comparison with existing approaches can be viewed in terms of synergy between different features of the framework, rather than the isolated presence of any specific feature within the framework itself. In particular, this becomes important in the context of inter-institutional development of artificial intelligence when the lack of a centralized authority makes verification of processes and results an essential part of maintaining the process. A comparison of the proposed blockchain-based framework with conventional data management strategies is shown in Table VI.

TABLE VI. COMPARISON WITH EXISTING DATA SHARING FRAMEWORKS

Feature	Traditional Systems	Blockchain-Based (Proposed)
Data Integrity	Partial	Cryptographically verifiable
Access Control	Centralized	Decentralized (Smart Contracts)
Auditability	Limited	Full & Immutable
Scalability	Moderate	High (Off-Chain Storage)
Trust Requirement	Third Party	Minimized

Overall, experimental results provide a strong foundation for the use of the proposed framework based on its positive impact on different aspects of the problem:

- Data integrity through cryptographic verification
- Secure access control via smart contracts
- Scalability using off-chain storage
- Complete auditability through immutable logs

Integration of blockchain technology and AI training allows establishing trust in the whole AI process because only safe and verifiable data sets can be used to develop machine learning algorithms.

VI. Conclusion

This study introduced a trusted data sharing framework based on blockchain that provides secure and auditable AI model training, addressing critical issues concerning the integrity of data, access control, provenance tracking, and auditability in the collaborative AI setups. Since AI systems are more based on distributed datasets provided by different institutions, the verification of trustworthiness and training data authentication becomes a dire need. The suggested architecture integrates a permissioned blockchain, smart contracts, and off-chain storage, creating a secure and transparent AI training pipeline. In this architecture, cryptographic hashing is used to ensure the integrity of data and limits the metadata and hash records on the blockchain to ensure scalability. Fine-grained access control, provenance tracking, and immutable audit logging are automated by smart contracts, thus limiting reliance on centralized trust authorities. Experimental analysis in a simulated environment indicated that the framework can identify tampered datasets, intercept unauthorized access attempts, and maintain high verification accuracy even with growing volumes of datasets. Moreover, the performance analysis confirmed that the new security measures add reasonable latency overhead, and the framework can be used in practical terms when associated with AI training processes. The findings suggest that implementing blockchain technology in AI data pipelines, applied as a practical heuristic, can improve transparency, accountability, and reliability. The framework will help to enhance trust in the data-driven decision-making process by ensuring trained AI models use only verified and authorized data, especially in sensitive fields like healthcare, finance, or intelligent infrastructure.

VII. Limitations

While the experimental results support the feasibility of the proposed framework, several limitations should be acknowledged. First, the evaluation was conducted in a simulated, Hyperledger Fabric–like environment with synthetically generated datasets; performance on an actual permissioned blockchain deployment may differ, and a fully documented, replicable experimental setup remains future work. Second, the reported 100% integrity detection and access control success rates are deterministic consequences of cryptographic hashing and smart contract enforcement in a correct implementation; they validate functional correctness rather than constitute empirical security guarantees under adversarial real-world conditions. Third, the framework should be regarded as a heuristic that secures the data sharing and verification stages of the AI pipeline: even when datasets are shared and tracked through the blockchain, trained models remain exposed to threats outside the framework's scope, including model-merging and distillation attacks, as well as theft or disclosure of model weights during the training stage. Fourth, the comparison with traditional training relies on qualitative descriptors, and quantitative machine learning performance metrics under clean and poisoned training conditions have not yet been evaluated. Addressing these aspects is essential before the framework can be considered for deployment in data-sensitive production environments.

VIII. Future Work

Future work can be directed towards the implementation of smart contract execution efficiency, the use of privacy-preserving methods, like the use of zero-knowledge proofs or secure multiparty computation, and the expansion of the framework to accommodate federated learning settings. Moreover, deployment on an actual permissioned blockchain such as Hyperledger Fabric, quantitative evaluation of model performance under clean and poisoned training conditions, defenses against training-stage threats such as model weight theft, model merging, and distillation, and real-world cross-organizational case studies would further validate the framework at scale.

Acknowledgment

I sincerely thank my mentors and colleagues for their guidance and support in the study's research and preparation.

References

[1] O. Karaduman and G. Gulhas, "Blockchain-enabled supply chain management: A review of security, traceability, and data integrity amid the evolving systemic demand," *Appl. Sci.*, vol. 15, no. 9, p. 5168, May 2025, doi: 10.3390/app15095168.

[2] M. N. Alatawi, "Blockchain-driven smart contracts for advanced authorization and authentication in cloud security," *Electronics*, vol. 14, no. 15, p. 3104, Aug. 2025, doi: 10.3390/electronics14153104.

[3] M. Schlegel and K.-U. Sattler, "Capturing end-to-end provenance for machine learning pipelines," *Inf. Syst.*, vol. 132, p. 102495, Jul. 2025, doi: 10.1016/j.is.2024.102495.

[4] J. Singh *et al.*, "A systematic review of blockchain, AI, and cloud integration for secure digital ecosystems," *Int. J. Netw. Distrib. Comput.*, vol. 13, art. no. 28, pp. 1–48, 2025, doi: 10.1007/s44227-025-00072-1.

[5] G. Governatori *et al.*, "On legal contracts, imperative and declarative smart contracts, and blockchain systems," *Artif. Intell. Law*, vol. 26, no. 4, pp. 377-409, Dec. 2018, doi: 10.1007/s10506-018-9223-3.

[6] P. Kashefi, Y. Kashefi, and A. Ghafouri Mirsaraei, "Shaping the future of AI: Balancing innovation and ethics in global regulation," *Uniform Law Rev.*, vol. 29, no. 3, pp. 524-548, Aug. 2024, doi: 10.1093/ulr/unae040.

[7] H.-Y. Paik *et al.*, "Analysis of data management in blockchain-based systems: From architecture to governance," *IEEE Access*, vol. 7, pp. 186091–186107, Dec. 2019, doi: 10.1109/ACCESS.2019.2961404.

[8] A. Jalbani, R. Weerawarna, and K. Al-Zubaidi, "Enhancing data provenance in AI with blockchain technology: A comprehensive quality model," *CSI Trans. ICT*, vol. 13, no. 2, pp. 213-224, 2025, doi: 10.1007/s40012-025-00419-7.

[9] R. Singh, A. Gupta, and P. Mittal, "A systematic literature review on blockchain-based smart contracts: Platforms, applications, and challenges," *Distrib. Ledger Technol. Res. Pract.*, vol. 5, no. 1, art. no. 6, pp. 1-36, Dec. 2025, doi: 10.1145/3704741.

[10] S. Mehammed *et al.*, "Blockchain audited federated learning: Securing data and model updates with on-chain provenance," *IET Softw.*, vol. 2025, art. no. 6670439, 2025, doi: 10.1049/sfw2/6670439.

[11] P. Kasralikar *et al.*, "Blockchain for securing AI-driven healthcare systems: A systematic review and future research perspectives," *Cureus*, vol. 17, no. 4, p. e83136, Apr. 2025, doi: 10.7759/cureus.83136.

[12] Z. Zheng *et al.*, "An overview on smart contracts: Challenges, advances and platforms," *Future Gener. Comput. Syst.*, vol. 105, pp. 475-491, Apr. 2020, doi: 10.1016/j.future.2019.12.019.

Shortest Path Algorithms for Smart City Emergency Routing: Scalability and Scenario-Based Analysis

Aleksandra Stojanova Ilievska, Natasha Stojkovikj,
Limonka Koceva Lazarova, Arta Aliu, Krume Tasev, Teona Aleksova
Faculty of Computer Science, Goce Delcev University, Stip, North Macedonia

***Abstract*—Shortest path algorithms play an essential role in intelligent transportation systems, emergency response planning, and numerous Smart City services. This study investigates the applicability of three well-established graph-based algorithms, namely Dijkstra, Bellman-Ford, and A*, within an emergency transportation framework inspired by the urban road network of Skopje. A two-stage evaluation methodology was adopted. First, scalability experiments were performed on weighted graphs containing 50 to 500 nodes to examine computational efficiency and algorithm behavior as network size increases. Subsequently, a Smart City transportation model was developed and implemented using Java and JavaFX, enabling the simulation and visualization of several emergency scenarios, including ambulance dispatching, road disruptions, and route unavailability situations. The experimental results indicate that all investigated algorithms successfully identify shortest feasible routes when transportation paths exist, whereas the heuristic guidance employed by A* reduces the number of explored nodes and improves execution times, particularly in larger graph instances. The proposed framework illustrates how algorithmic analysis can be complemented by simulation-based case studies to support intelligent transportation applications and emergency management decision-making. Future research will focus on incorporating real-time traffic information, stochastic travel times, and larger transportation networks to provide more realistic emergency routing environments.**

Keywords*—Shortest path algorithms; dijkstra; bellman-ford; A*; smart city; emergency routing; intelligent transportation systems; simulation

I. Introduction

The increasing urbanization observed worldwide has considerably intensified the demand for efficient transportation systems capable of supporting sustainable, resilient, and safe city development. Consequently, the concept of Smart Cities has attracted significant attention from researchers, policymakers, and industry practitioners. Smart Cities rely on the integration of digital technologies, intelligent infrastructures, and data-driven decision-making processes to improve urban services and enhance citizens' quality of life. Among the numerous application domains, intelligent transportation systems have emerged as one of the most important components of modern Smart Cities due to their direct impact on mobility, accessibility, and emergency response capabilities [1]–[3].

Emergency medical services represent a particularly challenging aspect of urban transportation management. In many situations, even relatively small delays in ambulance dispatching and patient transportation may significantly affect treatment outcomes and public safety. Therefore, identifying efficient routes under dynamically changing traffic conditions has become an important research topic within intelligent transportation systems, emergency management, and Smart City initiatives [1]–[3].

Graph theory provides a natural framework for modeling transportation networks, where intersections, districts, or critical facilities can be represented as nodes, while roads connecting neighboring locations are modeled as weighted edges. Within this context, shortest path algorithms have been extensively investigated and successfully applied in navigation systems, communication networks, logistics, and autonomous decision-support applications [4], [5]. Classical methods such as the Bellman-Ford algorithm remain useful for handling graphs containing negative edge weights, whereas Dijkstra's algorithm is widely recognized for its efficiency in solving shortest path problems in networks with non-negative costs [6]–[9].

Heuristic search techniques have further expanded the applicability of graph-based optimization methods. The A* algorithm combines accumulated path costs with heuristic estimations, enabling a substantial reduction of the explored search space while preserving optimality under admissible heuristic functions [10]. Several studies have demonstrated the effectiveness of heuristic approaches in robotic navigation, autonomous systems, and dynamic path-planning applications [11]–[14].

In addition to navigation tasks, shortest path algorithms have found numerous applications in traffic management, ride-sharing services, logistics optimization, disaster response planning, and healthcare transportation systems [4], [5]. Their ability to support routing decisions under different operational constraints makes them particularly suitable for Smart City environments, where transportation networks are often characterized by congestion, limited accessibility, and dynamically changing conditions [8], [9]. As urban mobility becomes increasingly complex, the need for efficient and adaptive routing mechanisms continues to grow.

Advances in computational capabilities and visualization tools have facilitated the development of interactive environments for studying graph algorithms and evaluating their practical performance. Comparative analyses involving multiple shortest path techniques can provide valuable insights into their computational characteristics, scalability, and suitability for various emergency transportation scenarios.

Such analyses may contribute to the design of more efficient intelligent transportation systems and support decision-making processes in time-critical situations [10]–[14].

Although numerous studies have addressed shortest path algorithms and their theoretical properties, comparatively fewer investigations have examined their behavior within integrated Smart City environments, particularly in emergency transportation scenarios involving road closures, route unavailability, and varying operational conditions. Existing research often focuses either on algorithmic improvements or on large-scale intelligent transportation infrastructures, while experimental comparisons conducted within realistic urban emergency settings remain relatively limited [4], [5], [13]–[15].

Motivated by these observations, this study presents a comparative analysis of Dijkstra, Bellman-Ford, and A* algorithms within a Smart City emergency transportation framework inspired by the urban road infrastructure of Skopje. The proposed methodology combines scalability experiments with a case study involving ambulance dispatching scenarios, road blockage simulations, and emergency routing analysis. Furthermore, a visualization environment implemented using Java and JavaFX enables intuitive representation of transportation networks, optimal routes, and unavailable road segments. The obtained results provide additional insight into the applicability, advantages, and limitations of shortest path algorithms in intelligent transportation and emergency response systems.

The main contributions of this study can be summarized as follows:

- A comparative evaluation of Dijkstra, Bellman-Ford, and A* algorithms under increasing graph sizes;
- The development of a Smart City emergency transportation model inspired by the road network of Skopje;
- The simulation of ambulance dispatching scenarios, including road closures and unavailable routes;
- The implementation of a visualization framework for representing transportation networks and routing results;
- An analysis of algorithm scalability and computational performance under different experimental conditions.

II. Related Work

Shortest path problems have been extensively investigated over the past decades and continue to represent one of the most important optimization problems in graph theory, operations research, and artificial intelligence. Their practical importance extends beyond theoretical studies, encompassing applications in navigation systems, communication networks, logistics, transportation planning, and autonomous decision-support systems. Comprehensive surveys available in the literature provide an overview of shortest path variants, algorithmic improvements, and emerging application domains, emphasizing their continuing relevance in modern intelligent environments [4], [5].

Several studies have focused on improving classical shortest path algorithms to address the growing complexity of transportation networks. Although Dijkstra's algorithm remains one of the most widely used approaches for graphs with non-negative edge weights [7], numerous modifications have been proposed to improve memory utilization, computational efficiency, and scalability. Recent investigations demonstrated that alternative storage mechanisms and implementation strategies can reduce computational costs, particularly in sparse or large-scale networks [8], [9]. Comparative analyses involving Dijkstra, greedy search, and heuristic methods indicate that algorithm performance is highly dependent on network characteristics, graph density, and routing constraints [12].

Heuristic search methods have attracted considerable attention due to their ability to reduce the explored search space while preserving solution optimality under admissible heuristic functions. The A* algorithm remains one of the most influential heuristic techniques and has been successfully applied in robotic navigation, autonomous systems, game artificial intelligence, and path-planning applications [10]. Several studies reported that modified heuristic strategies can significantly improve navigation efficiency in dynamic and partially known environments, while replanning techniques further enhance adaptability when transportation conditions change unexpectedly [11], [13], [14].

The rapid development of Smart City initiatives has stimulated substantial research efforts toward intelligent transportation systems, urban mobility optimization, and data-driven decision support. Smart transportation infrastructures increasingly rely on interconnected devices, Internet of Things technologies, and real-time information exchange mechanisms to improve traffic management and mobility services [16], [17]. More recently, cooperative transportation frameworks and adaptive path-planning approaches have been proposed to reduce travel times, alleviate congestion, and support efficient traffic flow management [18]. Systematic reviews indicate that intelligent transportation systems are expected to play a key role in future urban environments by integrating sensing technologies, communication platforms, optimization algorithms, and predictive analytics [19].

Emergency medical transportation remains a particularly important research topic because transportation delays may directly affect patient outcomes and overall public safety. Previous studies investigated robust route planning strategies capable of operating under uncertain traffic conditions, demonstrating the importance of adaptive routing mechanisms for emergency response applications [20]. Considerable attention has also been devoted to ambulance redeployment and dispatching problems. Heuristic methods, approximate dynamic programming techniques, and metaheuristic approaches have been shown to improve vehicle availability and reduce response times in dynamically changing environments [21]–[24]. Furthermore, extensive reviews of emergency medical service systems highlight the growing need for optimization-based decision support tools that can efficiently allocate resources and support operational planning [25]–[27].

Simulation techniques have become increasingly important for evaluating transportation systems under realistic operating conditions. Multimethod simulation environments provide opportunities for combining discrete-event, system dynamics, and agent-based approaches within a single modeling framework [28]. Agent-based methodologies, in particular, enable the representation of interactions among individual entities and have been successfully employed for studying complex adaptive systems, educational processes, and healthcare-related applications [29], [30].

Visualization and simulation environments have also been employed to investigate graph optimization problems, support educational activities, and analyze transportation networks. Previous studies demonstrated the usefulness of visualization tools for improving the understanding of network flow algorithms [31], while agent-based methodologies provide flexible mechanisms for representing dynamic processes and educational scenarios [32]. More recently, shortest path optimization techniques have been successfully applied to transportation networks through case studies based on the Macedonian road infrastructure, further emphasizing the practical relevance of graph-based routing methods in real-world environments [33].

Recent advances in artificial intelligence have further contributed to the development of intelligent transportation solutions. AI-enhanced traffic simulations, predictive analytics, and adaptive traffic control mechanisms have demonstrated their potential to improve urban mobility, increase road safety, and support more efficient emergency response operations [34], [35]. In addition, the integration of artificial intelligence with smart ambulance systems and dynamic traffic management frameworks may facilitate faster route adaptation and reduce intervention times in critical situations [36].

Despite the substantial body of literature dedicated to shortest path algorithms, intelligent transportation systems, and emergency vehicle routing, comparative studies combining scalability analysis, visualization environments, and Smart City emergency transportation scenarios within a unified framework remain relatively limited. Motivated by this observation, the present study proposes a two-stage evaluation methodology consisting of scalability experiments and a Smart City emergency transportation case study inspired by the urban road network of Skopje. The proposed framework aims to provide a better understanding of the applicability, efficiency, and limitations of shortest path algorithms in intelligent transportation and emergency management systems.

III. Theoretical Background and Algorithmic Foundations

Graph theory provides the mathematical foundation for numerous optimization problems, including shortest path determination, network analysis, transportation planning, and intelligent decision-support systems. A graph can be represented as

$$G = (V, E) \tag{1}$$

where, (V) denotes the set of vertices (nodes) and (E) represents the set of edges connecting pairs of vertices. In weighted graphs, each edge is assigned a numerical value corresponding to distance, travel time, cost, or another optimization criterion.

The shortest path problem aims to determine a path between two vertices such that the total accumulated edge weight is minimized. Due to its broad applicability, this problem has been extensively investigated in transportation systems, robotics, communication networks, and artificial intelligence applications.

A. Dijkstra's Algorithm

Dijkstra's algorithm [5] is one of the most widely used exact shortest path algorithms for weighted graphs with non-negative edge costs. The algorithm iteratively selects the unvisited vertex with the smallest tentative distance and relaxes its neighboring vertices until the shortest distances to all reachable nodes are determined.

The computational complexity of Dijkstra's algorithm depends on the underlying data structure used for maintaining the priority queue. When implemented using a binary heap, its time complexity is given by

$$O\big((V + E)\log V\big) \tag{2}$$

where, (V) denotes the number of vertices and (E) represents the number of edges. The corresponding space complexity is

$$O(VE) \tag{3}$$

Due to its efficiency and guaranteed optimality, Dijkstra's algorithm remains widely adopted in transportation planning and navigation systems.

B. Bellman-Ford Algorithm

Bellman-Ford is another classical shortest path algorithm capable of processing graphs containing negative edge weights [6]. The algorithm repeatedly relaxes all graph edges and can additionally detect negative cycles within the network.

Its worst-case computational complexity is $O(VE)$

while the required memory complexity remains $O(V)$.

Although Bellman-Ford generally requires more computational resources than Dijkstra's algorithm, its ability to handle more general graph structures makes it valuable in communication networks and cost-sensitive routing problems.

C. A* Algorithm

The A* algorithm combines graph search techniques with heuristic guidance to accelerate shortest path computation [10]. Instead of exploring all possible nodes uniformly, A* prioritizes nodes that appear more promising according to an evaluation function.

The evaluation function is defined as

$$f(n) = g(n) + h(n) \tag{4}$$

where, $(g(n))$ represents the accumulated cost from the source node to the current node, while $(h(n))$ denotes an estimate of the remaining distance to the target node.

For A* to guarantee optimality, the heuristic function should satisfy admissibility and consistency conditions. An admissible heuristic never overestimates the true shortest distance to the destination, whereas a consistent heuristic satisfies.

$$h(n) \leq c(n,m) + h(m) \quad (5)$$

for every neighboring node pair $((n,m))$.

Common heuristic functions include Euclidean and Manhattan distances, which are frequently employed in transportation and navigation applications. Under these assumptions, A* preserves optimality while significantly reducing the number of explored nodes compared to uninformed search methods.

Its worst-case computational complexity remains comparable to Dijkstra's algorithm, although practical implementations often exhibit considerably better performance due to heuristic guidance.

D. Shortest Path Algorithms in Intelligent Transportation Systems

Shortest path algorithms constitute an essential component of intelligent transportation systems and Smart City infrastructures [23-25]. Typical applications include route planning, emergency vehicle dispatching, traffic optimization, autonomous navigation, and mobility management [26–28].

In emergency response scenarios, minimizing travel time is particularly important, as delayed interventions may significantly affect public safety and healthcare outcomes. Consequently, evaluating shortest path algorithms within realistic transportation environments represents an important research direction for intelligent transportation systems and emergency management applications.

IV. Proposed Smart City Emergency Transportation Framework

A. Problem Definition

Efficient emergency response remains one of the major challenges faced by modern Smart Cities. Delays in ambulance dispatching and patient transportation may significantly affect healthcare outcomes, making route optimization an essential component of intelligent transportation systems.

In this study, the emergency routing problem is formulated as a shortest path problem in a weighted graph representing an urban transportation network. The objective is to determine the fastest route for an ambulance traveling from an arbitrary starting location to a hospital while considering potential road closures and varying transportation conditions.

B. Smart City Transportation Network

To evaluate the applicability of shortest path algorithms in realistic urban environments, a transportation model inspired by the road network of the city of Skopje was developed.

The transportation infrastructure is represented as an undirected weighted graph described by Eq. (1), where nodes correspond to urban districts and edges denote road segments connecting neighboring regions. Edge weights represent estimated travel times between adjacent locations.

The transportation network was constructed using several districts of Skopje, namely Aerodrom, Centar, Karpos, Vlae, Gjorce Petrov, Bit Pazar, Kisela Voda, and a hospital location. Connections between neighboring districts were modeled as weighted edges, where the assigned values represent approximate travel times in minutes.

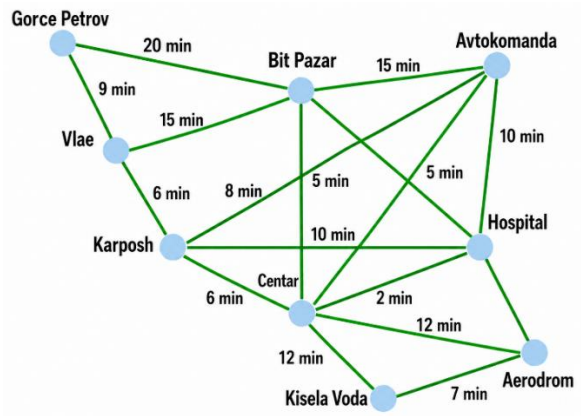


Fig. 1. Smart City transportation network inspired by the urban road infrastructure of Skopje.

Fig. 1 presents the Smart City transportation network used throughout the experimental evaluation.

C. Framework Architecture and Implementation

The proposed framework consists of four major modules:

- Graph Representation Module;
- Routing Algorithm Module;
- Visualization Module;
- Experimental Analysis Module.

The Smart City simulation environment was implemented using Java and JavaFX, while Python was employed for supplementary algorithmic validation and scalability assessment.

The graph representation module stores nodes, edge weights, predecessor information, and route availability status. The routing module integrates Dijkstra, Bellman-Ford, and A* algorithms, whereas JavaFX enables visualization of transportation networks, optimal routes, and blocked road segments. Bellman-Ford was included in the scenario-based evaluation as a classical reference algorithm widely used in shortest-path research. Although the investigated transportation network contains only non-negative edge weights, Bellman-Ford provides a useful baseline for comparing computational behaviour and routing consistency with Dijkstra and A*. It was intentionally excluded from the scalability experiments because its higher computational complexity would have resulted in substantially longer execution times for repeated evaluations on large randomly generated graphs without

providing additional methodological insight under the considered experimental conditions. Fig. 2 illustrates the interaction among the main modules of the proposed Smart City emergency transportation framework.

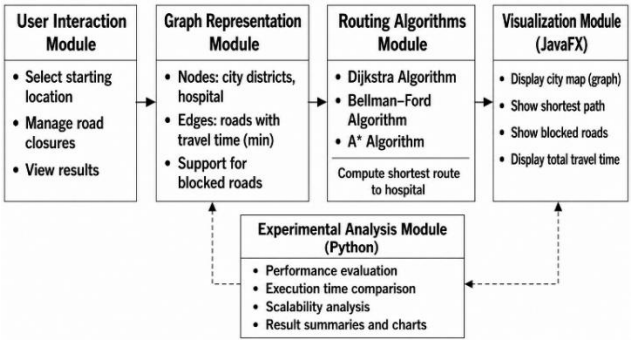


Fig. 2. Architecture of the proposed Smart City emergency transportation framework.

1) Research hypotheses: To guide the experimental evaluation, the study investigates the following research hypotheses:

H1. The A* algorithm achieves lower computational cost than Dijkstra's algorithm while preserving optimal route quality in transportation networks with non-negative edge weights.

H2. The proposed Smart City emergency routing framework is capable of correctly identifying feasible shortest routes as well as detecting situations in which no valid route exists due to road disruptions.

H3. Combining scalability experiments with representative Smart City emergency scenarios provides a comprehensive evaluation of both computational performance and practical applicability of classical shortest-path algorithms.

D. Dynamic Road Availability and Emergency Routing Scenarios

Emergency transportation systems frequently operate under dynamically changing conditions. Traffic incidents, road maintenance, construction activities, or unexpected events may temporarily block important transportation links. To simulate such situations, selected road segments can be excluded from the transportation network before executing shortest path algorithms, enabling the generation of alternative routes whenever the initially optimal path becomes unavailable.

The term dynamic in this study refers to the evaluation of multiple transportation scenarios with different road availability configurations rather than continuous real-time updates of edge weights during route computation. Each scenario represents a static snapshot of the transportation network under a particular emergency condition.

Four representative emergency scenarios were considered to investigate algorithm behavior under varying transportation conditions.

1) Scenario 1: The ambulance vehicle is initially positioned in the Aerodrom district and dispatched toward the hospital under normal traffic conditions.

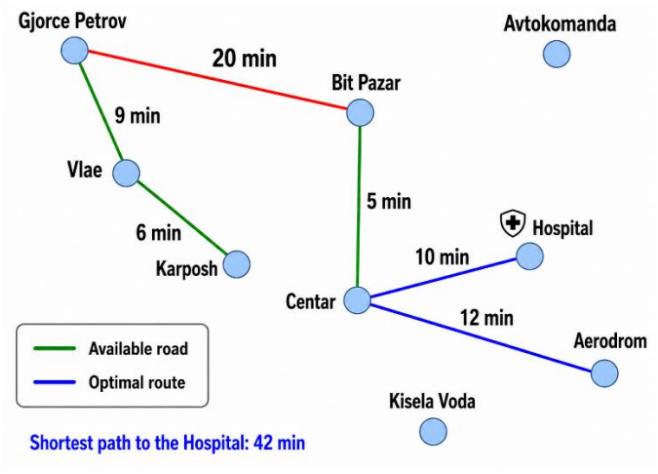


Fig. 3. Ambulance dispatching from Aerodrom district under normal traffic conditions.

2) Scenario 2: The ambulance vehicle starts from Vlae while maintaining full road accessibility.

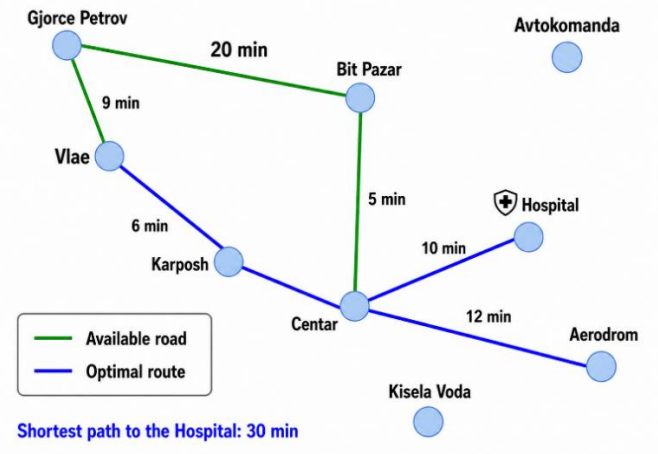


Fig. 4. Ambulance dispatching from Vlae district.

3) Scenario 3: Road segments leading to the hospital are intentionally blocked, resulting in the absence of a feasible transportation path.

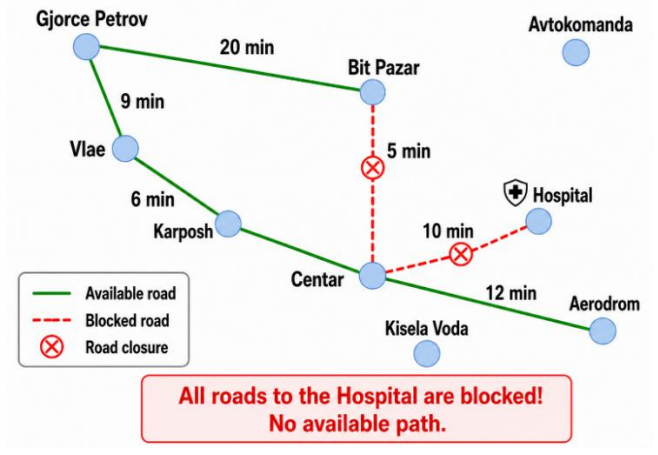


Fig. 5. Road closure scenario with unavailable routes to the hospital.

4) Scenario 4: The ambulance vehicle departs from Gjorce Petrov under regular transportation conditions.

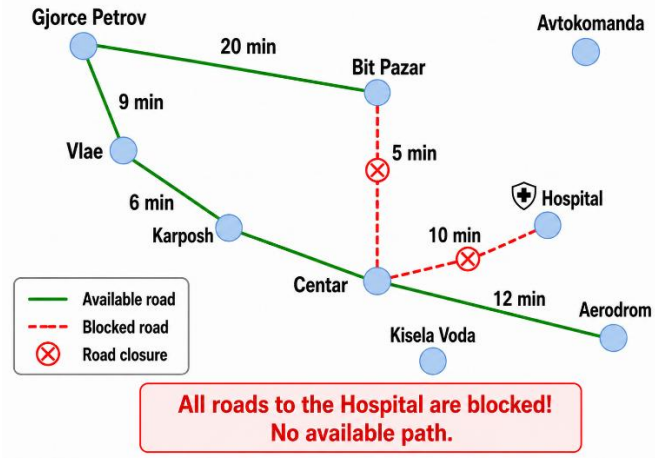


Fig. 6. Ambulance dispatching from Gjorce Petrov district.

The developed framework provides a flexible environment for investigating shortest path algorithms under realistic emergency transportation scenarios and serves as the basis for the experimental evaluation presented in the following sections.

V. Experimental Evaluation and Results

A. Scalability Assessment of Shortest Path Algorithms

To complement the Smart City emergency transportation scenarios, an additional set of experiments was conducted to evaluate the scalability and computational efficiency of shortest path algorithms under increasing graph sizes. These experiments were designed to assess the potential applicability of the investigated algorithms in larger Smart City transportation networks, where the number of intersections and road segments may substantially exceed those considered in the Skopje-inspired case study.

Dijkstra's and A* algorithms were implemented in Python 3.11 using a binary min-heap priority queue. Random weighted graphs containing 50, 100, 200, and 500 nodes were generated programmatically to simulate transportation networks of varying complexity. Edges were generated by considering all possible node pairs, where an edge was created with a probability (density) of 0.3. Edge weights were assigned as random integer values between 1 and 10, representing abstract transportation costs. The generated graphs were stored using adjacency-list representations to support efficient shortest-path computations. Priority queue operations in both Dijkstra's and A* implementations were performed using Python's built-in heapq module, which provides efficient insertion and extraction of the minimum-cost node. Since the investigated graph sizes were relatively moderate, no specialized memory optimization techniques were required. For each graph size, both algorithms were executed once on the generated weighted graph, and the execution time together with the number of visited nodes was recorded. The experiments were performed on a workstation equipped with an Intel i5 processor and 16 GB RAM.

Bellman-Ford was intentionally excluded from the scalability assessment because its computational complexity O(VE) makes its evaluation on large randomly generated graphs considerably more computationally demanding than Dijkstra and A*. Since the generated graphs contain only non-negative edge weights, including Bellman-Ford would not provide additional methodological insight beyond the scenario-based comparison presented in the Smart City case study.

A simple admissible heuristic based on the absolute difference between the current node identifier and the target node identifier, $h(n) = |\ goal - n\ |$, was adopted for the A* algorithm. Since the randomly generated scalability graphs do not contain spatial coordinates, conventional Euclidean or Manhattan distance heuristics were not applicable. Instead, the selected heuristic provides a simple and deterministic estimate that enables a consistent comparison between A* and Dijkstra under identical experimental conditions. Although relatively simple, this heuristic was sufficient to reduce the search space and improve computational performance.

Fig. 7 illustrates the execution time trends observed for Dijkstra's and A* algorithms as the graph size increases from 50 to 500 nodes. The presented values correspond to the recorded execution times for each generated graph instance. The obtained results indicate that the heuristic guidance employed by A* contributes to lower execution times and fewer visited nodes, particularly in larger graph instances.

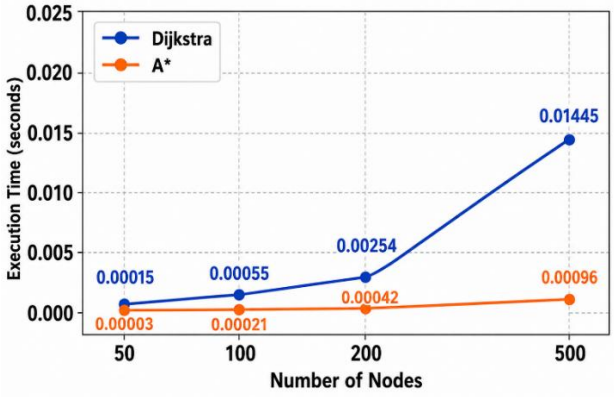


Fig. 7. Execution time comparison between Dijkstra and A* algorithms under increasing graph sizes.

Table I summarizes the obtained performance results.

TABLE I. Algorithm Performance Under Increasing Graph Sizes

Nodes	Algorithm	Time(s)	Nodes Visited
50	Dijkstra	0.00015	89
50	A*	0.00003	3
100	Dijkstra	0.00055	207
100	A*	0.00021	11
200	Dijkstra	0.00254	446
200	A*	0.00042	5
500	Dijkstra	0.01445	1233
500	A*	0.00096	7

The obtained results demonstrate that both algorithms successfully identified shortest paths for all tested graph instances. However, the increase in graph size had a considerably stronger impact on Dijkstra's algorithm, which explored a significantly larger portion of the graph. In contrast, A* consistently visited fewer nodes and maintained lower execution times due to heuristic guidance. The performance differences become particularly noticeable for graphs containing 200 and 500 nodes, suggesting that heuristic-based approaches can provide substantial advantages when shortest path computations are performed in large-scale transportation networks.

B. Smart City Emergency Routing Scenarios

Following the scalability assessment, the algorithms were evaluated within the proposed Smart City emergency transportation framework.

Four representative ambulance dispatching scenarios were investigated under different traffic conditions and road availability assumptions.

Scenario 1 considers an ambulance vehicle departing from the Aerodrom district under normal traffic conditions.

Scenario 2 examines emergency transportation originating from Vlae while maintaining complete road accessibility.

Scenario 3 simulates a road closure situation in which all transportation links leading to the hospital become unavailable.

Scenario 4 analyzes ambulance dispatching from Gjorce Petrov under regular transportation conditions.

Fig. 3–6 present visual representations of the considered emergency scenarios.

TABLE II. RESULTS OBTAINED FOR EMERGENCY ROUTING SCENARIOS

Scenario	*Starting District*	*Travel Time (min)*	*Path Available*
Scenario 1	Aerodrom	42	Yes
Scenario 2	Vlae	30	Yes
Scenario 3	Road closure	–	No
Scenario 4	Gjorce Petrov	34	Yes

Table II summarizes the routing outcomes obtained for the investigated emergency transportation scenarios. The results indicate that the proposed routing framework successfully identified the shortest available routes whenever feasible transportation paths existed. In the road closure scenario, the framework correctly identified the absence of a valid path to the hospital, demonstrating its capability to handle dynamically changing transportation conditions.

Each Smart City emergency routing scenario was executed once because the transportation network topology, edge weights, and algorithm implementations are deterministic. Consequently, identical input conditions always produce identical routes and travel times. Therefore, variance measures were not reported, as no stochastic variability is present in the adopted experimental setup.

The conducted experiments provide complementary insights into the applicability of shortest path algorithms in intelligent transportation systems. The scalability assessment evaluates computational performance under progressively larger graph instances, whereas the Smart City case study demonstrates the practical applicability of the proposed routing framework under representative emergency transportation scenarios. Since the Smart City evaluation is deterministic and scenario-based, it is intended to illustrate routing behaviour rather than to provide statistical performance benchmarking.

The scalability assessment highlights the computational benefits of heuristic-guided search, particularly when dealing with larger transportation networks. At the same time, the Smart City simulations demonstrate the practical applicability of these algorithms in realistic emergency response scenarios involving ambulance dispatching, road disruptions, and dynamic traffic conditions.

Taken together, the obtained results suggest that combining classical shortest path algorithms with Smart City transportation models offers a flexible environment for studying emergency routing strategies and evaluating intelligent transportation solutions under various operating conditions.

VI. DISCUSSION

The results obtained from both the scalability assessment and the Smart City case study provide useful insights into the applicability of shortest path algorithms in emergency transportation systems. By combining algorithm-oriented experiments with scenario-based simulations, the proposed framework enables an evaluation of both computational performance and practical suitability for time-critical routing tasks.

The scalability experiments revealed that the computational burden of shortest path algorithms increases with graph size. Although Dijkstra's algorithm consistently identified optimal routes, its execution time and the number of visited nodes increased substantially for larger graph instances. In contrast, the heuristic guidance incorporated into the A* algorithm limited unnecessary exploration of the search space and resulted in lower execution times. These differences became particularly evident for graphs containing 200 and 500 nodes, indicating that heuristic-based approaches may offer significant benefits when shortest path computations are repeatedly performed in large transportation networks.

For relatively small transportation networks similar to the proposed Skopje-inspired model, all investigated algorithms produced comparable routing decisions and identified identical optimal routes. Therefore, the primary differences among the algorithms are reflected in computational efficiency and scalability rather than in route optimality within moderately sized urban networks.

The Smart City emergency scenarios complement the scalability analysis by illustrating how shortest path algorithms behave under conditions that more closely resemble real-world transportation environments. The investigated scenarios demonstrated that feasible ambulance routes can be efficiently identified when transportation links remain available, while the framework was also capable of recognizing situations in which no valid route to the hospital exists due to road disruptions.

Such functionality is especially important in emergency response systems, where traffic incidents, construction activities, or unexpected events may rapidly alter transportation conditions.

An additional advantage of the proposed framework is its adaptability. The Skopje-inspired transportation network employed in this study was intentionally designed as a proof-of-concept case study rather than a complete representation of an urban transportation infrastructure. Its primary objective was to demonstrate the applicability of the proposed routing framework under representative emergency scenarios while maintaining a clear interpretation of algorithmic behaviour. Although the simplified network limits the direct generalization of the obtained results to large metropolitan transportation systems, the proposed methodology is independent of network size and can be readily extended to larger urban environments, multiple healthcare facilities, and transportation systems supported by continuously updated traffic information. Furthermore, more sophisticated heuristic functions, traffic prediction models, machine learning techniques, and real GIS or OpenStreetMap data may be incorporated to support adaptive routing decisions and improve overall system responsiveness.

Despite the encouraging results, several limitations should be acknowledged. The scalability assessment was intentionally limited to execution time and the number of visited nodes, as these metrics provide a direct comparison of computational efficiency across the investigated algorithms. Additional performance indicators, including memory consumption, edge relaxations, priority queue operations, and scalability under varying graph densities, represent promising directions for future research. The transportation network employed in this study represents a simplified proof-of-concept model of an urban environment and therefore does not capture the full topological complexity of real metropolitan transportation systems. Furthermore, the heuristic function adopted for the A* implementation was intentionally selected to remain simple to facilitate a transparent comparison with the investigated classical algorithms. Future studies may therefore focus on integrating richer transportation datasets, advanced heuristic strategies, and hybrid optimization methods capable of supporting more realistic emergency management scenarios. Although the proposed framework considers multiple emergency scenarios with different road availability conditions, route computation is performed on static network snapshots. Incorporating continuously changing traffic conditions, dynamic edge weights, and real-time rerouting mechanisms would further improve the realism and applicability of the proposed framework. In addition, the framework could be extended by incorporating more advanced routing techniques, including ALT (A*, Landmarks, Triangle inequality), Contraction Hierarchies, Bidirectional Dijkstra, and time-dependent shortest-path algorithms to further improve routing efficiency in large-scale transportation networks. Finally, comparisons with established routing platforms, such as OpenStreetMap-based routing engines and GraphHopper, would provide additional practical validation and facilitate benchmarking under real-world transportation conditions.

Nevertheless, the obtained conclusions should be interpreted with caution due to the relatively small size of the investigated transportation network and the simplified assumptions adopted in the simulation model.

Overall, the findings suggest that combining classical shortest path algorithms with Smart City transportation models provides a flexible and informative framework for investigating emergency routing strategies and evaluating intelligent transportation solutions under a wide range of operating conditions.

VII. Conclusion

This study presented a comparative investigation of three well-established shortest path algorithms, namely Dijkstra, Bellman-Ford, and A*, within the context of Smart City emergency transportation systems. The proposed framework combined scalability experiments with scenario-based simulations inspired by the urban road network of Skopje, allowing both computational characteristics and practical routing capabilities to be evaluated.

The scalability assessment demonstrated that the performance differences among the investigated algorithms become more pronounced as graph size increases. While Dijkstra's algorithm consistently produced optimal solutions, the heuristic guidance employed by A* reduced the number of explored nodes and resulted in shorter execution times for larger transportation networks. Bellman-Ford preserved its ability to handle graphs containing negative edge weights, although this flexibility was accompanied by increased computational requirements.

The simulated emergency transportation scenarios further illustrated the applicability of shortest path algorithms in ambulance dispatching and route planning tasks. The developed framework successfully identified feasible routes under normal operating conditions and detected situations in which transportation links became unavailable due to road closures. In addition, the visualization environment facilitated an intuitive interpretation of routing decisions and transportation network states, which may be useful for both educational purposes and preliminary decision-support analyses.

Although the presented transportation model represents a simplified approximation of an urban environment, the obtained results indicate that graph-based routing methods remain effective tools for supporting emergency transportation planning in Smart City settings. Future work may focus on incorporating real-time traffic information, stochastic travel times, multiple healthcare facilities, and adaptive optimization techniques, including machine learning approaches, to provide more realistic and responsive emergency management solutions.

Overall, the presented framework demonstrates that graph-based routing approaches can provide valuable support for emergency transportation planning and may serve as a basis for future intelligent transportation applications.

ACKNOWLEDGMENT

The authors sincerely appreciate the financial assistance received from Goce Delcev University, Stip, through its Individual Scientific Fund. This support played a crucial role in the successful completion of this research.

REFERENCES

[1] E. F. Z. Santana, A. P. Chaves, M. A. Gerosa, F. Kon, and D. S. Milojicic, "Software Platforms for Smart Cities: Concepts, Requirements, Challenges, and a Unified Reference Architecture," ACM Computing Surveys, vol. 50, no. 6, Art. no. 78, 2017.

[2] S. E. Bibri and J. Krogstie, "Smart Sustainable Cities of the Future: An Extensive Interdisciplinary Literature Review," Sustainable Cities and Society, vol. 31, pp. 183–212, 2017.

[3] V. Albino, U. Berardi, and R. M. Dangelico, "Smart Cities: Definitions, Dimensions, Performance, and Initiatives," Journal of Urban Technology, vol. 22, no. 1, pp. 3–21, 2015.

[4] A. Madkour, W. G. Aref, F. Ur Rehman, M. A. Rahman, and S. Basalamah, "A Survey of Shortest-Path Algorithms," arXiv preprint arXiv:1705.02044, 2017.

[5] L. Di Puglia Pugliese, D. Ferone, P. Festa, and F. Guerriero, "Shortest Path Tour Problems," in Encyclopedia of Optimization, Cham, Switzerland: Springer International Publishing, 2022, pp. 1–4.

[6] R. Bellman, "On a Routing Problem," Quarterly of Applied Mathematics, vol. 16, no. 1, pp. 87–90, 1958.

[7] E. W. Dijkstra, "A Note on Two Problems in Connexion with Graphs," Numerische Mathematik, vol. 1, no. 1, pp. 269–271, 1959.

[8] J. Wang, Y. Liu, B. Tang, X. Jin, and Y. Xiao, "Improved Dijkstra Algorithm Based on a New Storage Method," in Proceedings of the International Conference on Neural Networks, Information and Communication Engineering (NNICE), vol. 11933, SPIE, 2021, pp. 54–58.

[9] Y. Huang, Q. Yi, and M. Shi, "An Improved Dijkstra Shortest Path Algorithm," in Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 2013), Atlantis Press, 2013, pp. 226–229.

[10] P. E. Hart, N. J. Nilsson, and B. Raphael, "A Formal Basis for the Heuristic Determination of Minimum Cost Paths," IEEE Transactions on Systems Science and Cybernetics, vol. 4, no. 2, pp. 100–107, 1968.

[11] S. Koenig and M. Likhachev, "D* Lite," in Proceedings of the 18th National Conference on Artificial Intelligence (AAAI), Edmonton, Canada, 2002, pp. 476–483.

[12] M. R. Wayahdi, S. H. N. Ginting, and D. Syahputra, "Greedy, A-Star, and Dijkstra's Algorithms in Finding Shortest Path," International Journal of Advances in Data and Information Systems, vol. 2, no. 1, pp. 45–52, 2021.

[13] F. Duchoň, A. Babinec, M. Kajan, P. Beňo, M. Florek, T. Fico, and L. Jurišica, "Path Planning with Modified A Star Algorithm for a Mobile Robot," Procedia Engineering, vol. 96, pp. 59–69, 2014.

[14] S. Koenig and M. Likhachev, "Fast Replanning for Navigation in Unknown Terrain," IEEE Transactions on Robotics, vol. 21, no. 3, pp. 354–363, 2005.

[15] L. R. Ford Jr., Network Flow Theory, Paper P-923, RAND Corporation, Santa Monica, CA, USA, 1956.

[16] M. Batty, K. W. Axhausen, F. Giannotti, A. Pozdnoukhov, A. Bazzani, M. Wachowicz, G. Ouzounis, and Y. Portugali, "Smart Cities of the Future," *The European Physical Journal Special Topics*, vol. 214, no. 1, pp. 481–518, 2012.

[17] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of Things for Smart Cities," *IEEE Internet of Things Journal*, vol. 1, no. 1, pp. 22–32, 2014.

[18] Y. Regragui and N. Moussa, "A Real-Time Path Planning for Reducing Vehicles Traveling Time in Cooperative-Intelligent Transportation Systems," *Simulation Modelling Practice and Theory*, vol. 123, Art. no. 102710, 2023.

[19] M. A. Hassan, R. Javed, F. Granelli, X. Gen, M. Rizwan, S. H. Ali, H. Junaid, and S. Ullah, "Intelligent Transportation Systems in Smart City: A Systematic Survey," in *Proc. 2023 International Conference on Robotics and Automation in Industry (ICRAI)*, Islamabad, Pakistan, 2023, pp. 1–9.

[20] R. J. Szczerba, P. Galkowski, I. S. Glickstein, and N. Ternullo, "Robust Algorithm for Real-Time Route Planning," IEEE Transactions on Aerospace and Electronic Systems, vol. 36, no. 3, pp. 869–878, 2000.

[21] C. J. Jagtenberg, S. Bhulai, and R. D. van der Mei, "An Efficient Heuristic for Real-Time Ambulance Redeployment," Operations Research for Health Care, vol. 4, pp. 27–35, 2015.

[22] V. Schmid, "Solving the Dynamic Ambulance Relocation and Dispatching Problem Using Approximate Dynamic Programming," European Journal of Operational Research, vol. 219, no. 3, pp. 611–621, 2012.

[23] M. Gendreau, G. Laporte, and F. Semet, "A Dynamic Model and Parallel Tabu Search Heuristic for Real-Time Ambulance Relocation," Parallel Computing, vol. 27, no. 12, pp. 1641–1653, 2001.

[24] P. Toth and D. Vigo, Eds., Vehicle Routing: Problems, Methods, and Applications. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2014.

[25] R. Aringhieri, M. E. Bruni, S. Khodaparasti, and J. T. van Essen, "Emergency Medical Services and Beyond: Addressing New Challenges Through a Wide Literature Review," Computers & Operations Research, vol. 78, pp. 349–368, 2017.

[26] D. Bandara, M. E. Mayorga, and L. A. McLay, "Optimal Dispatching Strategies for Emergency Vehicles to Increase Patient Survivability," International Journal of Operational Research, vol. 15, no. 2, pp. 195–214, 2012.

[27] M. S. Maxwell, M. Restrepo, S. G. Henderson, and H. Topaloglu, "Approximate Dynamic Programming for Ambulance Redeployment," INFORMS Journal on Computing, vol. 22, no. 2, pp. 266–281, 2010.

[28] A. Borshchev, The Big Book of Simulation Modeling: Multimethod Modeling with AnyLogic 6. AnyLogic North America, 2013.

[29] C. M. Macal and M. J. North, "Tutorial on Agent-Based Modeling and Simulation," in Proceedings of the Winter Simulation Conference, 2005.

[30] S. C. Brailsford, "Tutorial: Advances and Challenges in Healthcare Simulation Modeling," in Proceedings of the 2007 Winter Simulation Conference, 2007, pp. 1436–1448.

[31] N. Stojkovic, M. Kocaleva, A. Stojanova, I. Janeva, and B. Zlatanovska, "Visualization of Ford-Fulkerson Algorithm," Balkan Journal of Applied Mathematics and Informatics, vol. 2, no. 2, pp. 7–20, 2019.

[32] N. Stojkovikj, L. Koceva Lazarova, A. Stojanova, M. Miteva, B. Zlatanovska, and M. Kocaleva, "Application of Agent-Based Modelling in Learning Process," Informatica, vol. 48, no. 1, 2024.

[33] N. Stojkovikj, S. Fileva, A. Stojanova Ilievska, L. Koceva Lazarova, C. Martinovska Bande, and V. Kokalanov, "Shortest Path Optimization in Transportation Networks Using Dijkstra's Algorithm: A Case Study of the Macedonian Road Network," Balkan Journal of Applied Mathematics and Informatics, vol. 9, no. 1, 2026.

[34] M. M. Haque, "Systematic Review on the Impact of AI-Enhanced Traffic Simulation on US Urban Mobility and Safety," ASRC Procedia: Global Perspectives in Science and Scholarship, vol. 1, no. 1, pp. 833–861, 2025.

[35] R. R. Saxena, "Artificial Intelligence in Traffic Systems," arXiv preprint arXiv:2412.12046, 2024.

[36] K. Chidananda, K. R. Pradeep, K. P. Mayuri, and S. Mythri, "AI-Driven Smart Ambulance and Dynamic Traffic Control System for Optimized Emergency Response," in Proceedings of the 2025 Control Instrumentation System Conference (CISCON), 2025, pp. 1–10.

Implementation and Evaluation of a Dynamic Multi-Grid Pattern Authentication for Secure Financial Applications

Amjad Qashlan[1]*, Qamar H. Naith[2], Fatimah Alakeel[3], Malak B Alharbi[4]
Department of Cybersecurity-College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia[1]
Department of Software Engineering-College of Computer Science and Engineering,
University of Jeddah, Jeddah 21959, Saudi Arabia[2, 4]
Department of Computer Science and Engineering-College of Applied Studies,
King Saud University, Riyadh, Saudi Arabia[3]

***Abstract*—Authentication plays an important role in online banking services. Nowadays, many online services still use password-based and One-Time Password (OTP) codes as authentication methods, but this is no longer secure, especially with the increase in financial fraud. In this research, a novel pattern-based password approach that relies on the random generation of multiple numeric grids is proposed. The approach is developed and deployed as a simple mobile application to be evaluated by banking users. Performance and security tests were used to evaluate the robustness of the approach. The findings of the assessment reveal that the performance and security level of our approach are high. The assessment also shows the approach to be capable of reducing and preventing financial fraud and protecting online banking systems from different types of cyberattacks through shoulder-surfing, eavesdropping, and brute force. The findings demonstrate the robustness of the pattern password, where fraudsters will be less likely to capture and recognize the correct password or discriminate between the patterns with their different values. To the best of our knowledge, this is the first approach that proposes a solution with the use of multi-grid rather than single-grid pattern passwords.**

***Keywords*—Authentication; cyberattack; financial fraud; multiple grids; pattern-based password; testing**

I. Introduction

With the continuous improvement in protecting the confidentiality and privacy of sensitive information for bank customers, and with the continuous implementation of fraud detection and implementation of fraud detection and innovative mitigation solutions, cyberattacks still represent a critical threat to users [1]. Given the current problems regarding safety and the ease of credit card exploitation, enhancing current authentication methods and providing optimal solutions to avoid cyberattacks is a topic that requires more investigation [2]. Research work in this area is still in its early stages, as the major security features of the different types of financial fraud and cyberattacks were not taken into consideration by most of the current state-of-the-art authentication approaches. Furthermore, many existing authentication studies focus primarily on implementation and usability aspects while providing limited formal analysis of adversarial capabilities, attack surfaces, and theoretical resistance against pattern reconstruction and brute-force attacks. Consequently, there remains a need for authentication frameworks that combine practical deployment with rigorous security evaluation.

In the literature, several authentication solutions have been proposed to tackle the issues related to the secure authentication process. According to the analysis of these developed approaches, these solutions have been classified into four categories: approaches that rely on the use of password [3]–[6], approaches that use a One-Time Password (OTP) code [7]–[12], approaches based on the use of selected pattern and images [13]–[17], and approaches that depend on the use of some user's characteristics [18]–[25]. Most of these authentication solutions have attempted to reduce fraud and cyberattack issues. However, several challenges limit its full success, such as the passcode guessability and observability from the attacker's perspective. There are some other research studies that attempted to address the limitations through the use of multi-factor authentication (MFA). However, these solutions are still unable to cover the full breadth of users' accessibility due to the fact that it is time-consuming, and there is a lack of consideration of human factor standards preventing it from providing a better user experience. Accordingly, we still need a more effective and more secure solution to successfully overcome all these limitations.

To improve on the previously suggested and implemented solutions and provide more protection against cyberattacks (e.g., fraud), this study proposes a new authentication approach that mainly relies on a Multi-Grid Pattern-Based Password (MGPP) approach that relies on the random generation of numbers on different grid layouts. The proposed approach is shown as a mobile-based application that includes four grids sized 3x3 and labelled with randomly assigned digits between 1 and 9. The arrangement of these numbers is changed dynamically, and it is possible to repeat this within the single or overall grid. The most important and distinctive feature of this approach, as compared to that of others, lies in the way of entering the predefined pattern across multiple grids' cells without clicking on the cells. In addition to the proposed authentication mechanism, this work introduces a formal threat model to characterize potential adversarial behaviors and a formal entropy-based security analysis to quantify the complexity of recovering the secret authentication pattern. These additions provide a theoretical

*Corresponding author

foundation for evaluating the security properties of the proposed framework beyond experimental testing alone.

In order to assess the effectiveness of the MGPP authentication mechanism, two questions were addressed:

Q1: How efficient is the Multi-Grid Pattern-Based Password authentication approach?

Q2: To what extent is a Multi-Grid Pattern-Based Password authentication approach effective in minimizing financial fraud?

To investigate these two research questions, two experimental studies have been conducted; one for evaluating the performance aspects, and another for evaluating the security aspect. Performance analysis shows that the proposed approach has a high level of performance and the capability of accepting and handling more users while also maintaining their authentication connection under different conditions. Security analysis indicates that the MGPP approach is highly effective in minimizing financial fraud. Moreover, the formal entropy analysis demonstrates that the proposed framework substantially enlarges the authentication search space compared with conventional single-grid pattern-based authentication systems, thereby reducing the probability of successful brute-force and pattern-guessing attacks. It has been proved that the login security success rates will be higher than those of single- grid pattern-based password methods. It also revealed that the password patterns across users are more varied and difficult to have identical pattern-based passwords for different users. This research contributes toward filling the gap in the literature regarding authentication approaches and related security issues. To the best of the authors' knowledge, this research is the first to:

- Provide a novel Multi-Grid Pattern-Based Password (MGPP) authentication framework that utilizes multiple dynamically randomized grids for secure user authentication;
- Develop a novel authentication algorithm that increases numerical randomization and pattern-space diversity to improve resistance against observation-based attacks and password guessing attempts;
- Introduce a formal threat model that characterizes adversarial capabilities and attack surfaces relevant to pattern-based authentication systems;
- Provide a formal entropy-based security analysis that quantifies the authentication search space and evaluates theoretical resistance against brute-force attacks; and
- Experimentally validate the proposed framework through comprehensive performance and security evaluations in a mobile banking authentication environment.

The structure of this study is organized as follows: In Section II, the authors provide an overview of the current literature on different authentication approaches. Section III presents the proposed MGPP authentication framework, including the registration process, authentication mechanism, cryptographic protection model, and formal threat model. A clear description of the development of the MGPP authentication approach is presented in Section IV. Section V describes, in detail, the two experimental studies conducted to evaluate the MGPP approach. Afterwards, Section VI presents the performance evaluation, security evaluation, and formal entropy analysis together with a discussion of the main findings. The study is then concluded in Section VII.

II. Related Work

In the literature, several authentication approaches and methods have been developed and implemented in different sectors including banking, healthcare and educational sectors. In addition, based on the literature, user authentication approaches can be categorized, as shown in Fig. 1 [26]–[28]. As illustrated in Fig. 1, existing authentication approaches typically involve a trade-off between security and convenience. While biometric-based authentication methods generally provide strong protection against unauthorized access, they often require specialized hardware and may introduce privacy concerns. Conversely, knowledge-based approaches such as passwords and graphical patterns offer greater accessibility and ease of deployment but remain vulnerable to observation, guessing, and credential theft attacks. Biometric authentication systems have strong properties to protect against misuse of special hardware; however, there are also privacy problems that arise when using such systems. Knowledge-based systems, such as password authentication and graphical patterns, are very easy to use and are therefore very often implemented. However, such systems have several flaws, for instance, observation, guessing, etc. We thus have to cope with the problems of enhancing the security of graphical authentication systems without degrading usability, and thus, new graphical authentication systems have to be implemented.

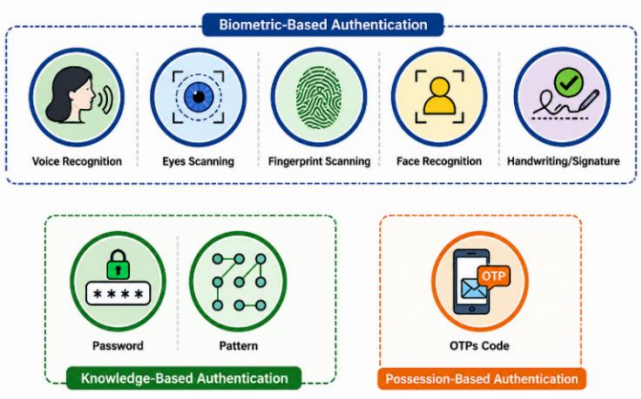


Fig. 1. Comparison of user authentication methods based on "security and convenience" parameters.

A. Knowledge-Based Authentication

Knowledge-Based Authentication (KBA) is a password or pattern that identifies and memorizes end users. To date, many online services are still using "password-based authentication", which focuses on the use of username/password and cryptographic authentication [3]–[6]. The authors in [13] proposed a password-based authentication method that relies on the concept of password generators as a recommended system. This allows users to select and use one of the lists of presented passwords. To increase the robustness of used passwords and to remove the issue of re-changing passwords multiple times, the

authors in [13], [14] proposed a new approach that combined both pattern-based and password-based authentication. A new pattern-based password authentication method was developed by researchers in [15] that uses a Picture-Based Pattern (PBP) technique as a password to secure the transaction. Unlike [15], an enhancement for pattern-based password authentication has been proposed by researchers in [16] by analyzing the input patterns of mobile banking users based on three main aspects: the normal finger pressure levels; the time the user takes to input data; and the physical touch dimensions when using a touch screen. The authors in [17] proposed a grid-based (graphical-based) pattern password method as an alternative pattern-based password solution. This method relies on drawing the pattern from the displayed alphanumeric characters as a form of the password rather than selecting the image.

Despite these improvements, most existing pattern-based authentication approaches continue to rely on a single-grid structure and provide limited protection against repeated observation attacks, pattern reconstruction attempts, and systematic password guessing. Furthermore, the majority of these studies focus on implementation and usability aspects while providing little formal analysis of the theoretical complexity of recovering the underlying authentication pattern.

B. Possession-Based Authentication

(PBA) relies on a One-Time Password (OTP), which is a randomly generated unique code sent via SMS, Email, or authenticator app to the user in each login attempt. The code expiry is time-based (usually between 15 and 60 seconds) for critical systems, and the user should enter the code to continue the login authentication process. The authors in [7] proposed an enhancement to the OTP algorithms by proposing an offline framework that can be implemented in smartphones to increase the security of online banking transactions. Very similar research studies have been conducted by researchers in [8], [9], [11], where they developed several OTP algorithms with the incorporation of various encryption and hashing techniques such as the Advanced Encryption Standard (AES) [29] and Blowfish [30] to encrypt or hash the verification code. Unlike previous works, the authors in [10], [12] proposed a new OTP algorithm with the incorporation of a QR code feature, where the user receives an OTP code in the form of a QR code that appears on an external device, and they must scan it using their mobile device.

C. Biometrics-Based Authentication

Biometrics-Based Authentication (BBA) is a unique characteristic and attribute for each user that can be used as a form of personal identification (user identity). For example, for a fingerprint authentication process, the authors in [19] have proposed an authentication method to enhance the security of online banking, using an external device for fingerprint authentication. Unlike [19], the authors in [18] proposed a new approach based on the fingerprint authentication process through sensors on the user's mobile device rather than using an external device. The research conducted by [20] and [21] used fingerprints as an authentication factor for bank card systems for enhancing ATM services. The authors in [20] described an algorithm for enhancing the phases of fingerprint image extraction and matching. An alignment-based algorithm was used for matching two templates. While the researchers in [21] proposed a method that totally relies on the elimination of cards and the use of fingerprint scanning instead, as it can serve as both a password and access card. In regard to the "eyes authentication process", only a few research studies have been found. The authors in [24], [25] proposed an eyes-based authentication approach using eye- scanning integrated with iris-scan authentication. The concept of this approach is to analyze the user's eyes and their iris pattern, which is previously scanned and stored during the registration process and account creation. The pattern of eyes and iris is analyzed and converted into binary co-registration, which is then processed and stored in the bank database, which will be required prior to completing the online bank transaction process.

For the "face recognition authentication methods", the authors in [31] used face scanning as an authentication factor to enhance the security of ATM services, instead of the current static key (PIN) for ATMs. Limited research studies have been conducted regarding the voice authentication process [22], [23], [32]. Such methods focus on training the system to recognize an individual's unique voice characteristics and matching it with the data stored in the bank's database to determine the identity of the user. The authors in [22], [23] used voiceprint and other vital signs to assist people with visual impairments. In such an authentication process, the account will be validated by entering the user's voiceprint; this will then allow further transactions via voice recognition that implements Mel Frequency Cepstral Coefficient (MFCC) [33], Discrete Wavelet Transform (DWT) [34], and Vector Quantization (VQ) [35] methods to continue. Similar to [32], the authors designed a system that recognizes the tone of voice, and it is found that the developed voice recognition software has successfully activated the door opening mechanism using a voice command that ONLY works for the authenticated individual.

D. Multi-Factor Authentication

(MFA), many research studies [31], [36]–[40] have provided multi-factor authentication solutions by incorporating the password with the OTP together, or the OTP and the password with a combination of more biometric factors such as fingerprints, facial recognition, retinal scan, IRIS recognition, or voice recognition. The two-factor biometric authentication method is proposed based on a combination of face ID and fingerprint [39], [40]. Unlike [36], [39], [40], the authors in [37] integrated the OTP with the fingerprints of users to authenticate online transactions. Slightly similar to [37], the researchers in [38] enhanced the security of the authentication process by proposing a three-factor authentication method, which includes passwords, familiar random pictures, and fingerprint scanning to make user authentication more secure (not only a combination of biometric factors). Although MFA approaches significantly enhance authentication security, they often increase system complexity, authentication time, and user interaction requirements. In mobile banking environments, excessive authentication steps may negatively affect usability and user acceptance. Consequently, there remains a need for lightweight authentication mechanisms capable of achieving stronger security guarantees without introducing additional authentication factors [41].

Existing graphical authentication methods such as password-based, pattern-based, biometric, possession-based methods, and multi-factor authentication methods have their own merits and have achieved a lot. Nevertheless, many open issues still need to be tackled [42]. The majority of graphical authentication methods use a single-grid-based approach; thus, they are susceptible to various types of observation-based attacks, i.e., various types of shoulder-surfing attacks and even authentication pattern reconstruction attacks where users use their authentication patterns as credentials. In addition, most existing graphical password authentication schemes have only tackled the issues for implementation and usability testing, without any formal adversary models and corresponding attack surfaces. There are only a handful of graphical password authentication schemes that can theoretically quantify the secret authentication patterns' resistance to brute-force attacks and guessing attacks by using entropy to measure the authentication passwords' complexity [43,44].

This study proposes a Multi-Grid Pattern-Based Password (MGPP) authentication, in which graphical authentication is supported by a Multiple of Graphical Passwords, each entry corresponding to a multiple number of dynamically generated grids for a single position in a pattern that is verified in a protected manner in order to verify a user's login credentials. After defining a formal threat model which details the possible ways in which an adversary can behave, we follow a formal entropy-based security analysis, i.e., a theoretical assessment of the resistance of the system for graphical password-based authentication to brute-force, as well as to guessing attacks. The results from the theoretical assessment are supported by performance, as well as by a security evaluation.

III. Proposed MGPP Authentication Framework

The proposed Multi-Grid Pattern-Based Password (MGPP) authentication framework is designed to improve the security of graphical password systems against observation-based attacks while maintaining usability for online authentication services. Unlike conventional single-grid graphical authentication schemes, the proposed framework employs multiple interconnected grids containing dynamically randomized numeric values. The framework combines pattern-based authentication with randomized numerical representations to increase ambiguity during login sessions and reduce the likelihood of successful pattern reconstruction by adversaries.

Let the authentication environment consist of G independent sub-grids, where each sub-grid contains C cells. The total number of cells available in the authentication interface is therefore defined as [see Eq. (1)]:

$$N = G \times C \tag{1}$$

where, N is the total number of cells in the authentication interface, G= number of sub-grids, C= number of cells per sub-grid. In the proposed implementation, four 3×3 grids are employed. Therefore, $G = 4$ and $C = 9$, which yields $N = 36$. The authentication process relies on a user-defined pattern consisting of a sequence of selected cells distributed across the multiple grids. Let the pattern length be represented by L, where $L = 8$, and let the pattern be expressed as [see Eq. (2)]:

$$P = \{p_1, p_2, \dots, p_L\} \tag{2}$$

where, p_i denotes the position of the i^{th} selected cell within the overall authentication space.

Each selected cell is represented by a coordinate pair [see Eq. (3)]:

$$p_i = (x_i, y_i) \tag{3}$$

where, x_i denotes the column index, y_i denotes the row index, $i \in \{1,2, \dots, L\}$. The complete authentication pattern can therefore be represented as [see Eq. (4)]:

$$P = \{(x_1, y_1), (x_2, y_2), \dots, (x_L, y_L)\} \tag{4}$$

which uniquely identifies the user's secret graphical password. To enhance resistance against shoulder-surfing and observation attacks, each authentication session generates a randomized numerical assignment for all cells. Let [see Eq. (5)]:

$$R_t = \{r_1, r_2, \dots, r_N\} \tag{5}$$

represent the set of numbers assigned to the cells during login session t, where r_i denotes the numeric value displayed in cell i. The mapping between cell positions and displayed numbers changes dynamically for every authentication session according to [see Eq. (6)]:

$$R_t \neq R_{t+1} \tag{6}$$

which ensures that the same graphical pattern produces different visible numeric sequences across consecutive login attempts.

During authentication, the user does not directly redraw the graphical pattern. Instead, the user observes the randomized values displayed inside the selected pattern cells and enters the corresponding numeric sequence through a text input field. Let the observed login sequence be represented as [see Eq. (7)]:

$$S_t = \{s_1, s_2, \dots, s_L\} \tag{7}$$

where, s_i denotes the numeric value currently displayed at position p_i during session t.

In the server authentication phase, all numbers are transferred to the corresponding cell positions, and the generated pattern from step 3 is compared to the pattern already stored in the database which was generated in the registration phase. Visible information for authentication and secret pattern for storage and user's authentication information are completely separated but remain consistent.

The MGPP system has two stages of operation, namely the registration stage and the login authentication stage. During the registration stage a user generates a secret multi-grid pattern. The secret multi-grid pattern is then transformed into a secure form and stored in the system for future reference by the user. In the login authentication stage, a user is presented with a number of displayed values, i.e., numeric values that are assigned in a totally random fashion, for example. The user is able to recreate his secret multi-grid pattern with the aid of the displayed values. The processes of registration and login authentication are detailed in Section III-B and Section III-C, respectively.

A. MGPP System Architecture

In this study, we present a new type of authentication framework called Multi-Grid Pattern-Based Password (MGPP) system, which utilizes client-server architecture to deliver high quality secure user authentication experience by thwarting the observation-based attacks on classic password-based authentication systems. The framework is composed of three main entities, i.e., user's device, authentication server, and credential repository. User's device is used to display multiple grids to end user, to capture users' input and to send login requests to authentication servers. In addition, authentication servers are utilized to generate multiple grids for end users, to verify the generated patterns by end users, to manage users' sessions and to encrypt/decrypt all communications with end users and their corresponding devices. Last but not least, credential repositories are used to store the protected pattern representations that were generated for end users during their registration phase.

The authentication interface is composed of multiple sub-grids which are all displayed to the user on one large multi-grid workspace. Let the set of sub-grids be represented as [see Eq. (8)]:

$$\mathcal{G} = \{g_1, g_2, \dots, g_G\} \tag{8}$$

where, $\mathcal{G}$ denotes the set of sub-grids and G denotes the total number of grids. In the proposed implementation, $G = 4$, and each grid contains $C = 9$ cells organized in a 3×3 structure. Consequently, the complete authentication workspace contains $N = 36$ distinct cell positions.

Each cell is uniquely identified through its coordinate location and grid membership. The position of a cell is represented by $c_{i,j,k}$ where i denotes the grid identifier, j denotes the row index, and k denotes the column index. The complete authentication space is therefore represented as [see Eq. (9)]:

$$\mathcal{C} = \{c_{i,j,k} \mid 1 \le i \le G, 1 \le j \le 3, 1 \le k \le 3\} \tag{9}$$

where, $\mathcal{C}$ denotes the set of all available cells within the MGPP environment.

To prevent the formation of static visual patterns, the system dynamically assigns numeric values to the cells at the beginning of each authentication session. Let the session-specific numeric assignment be represented by [see Eq. (10)]:

$$M_t: \mathcal{C} \to \mathcal{D} \tag{10}$$

where, M_t denotes the session-specific mapping function, $\mathcal{C}$ is a set of cells, and $\mathcal{D}$ is a set of displayed digits by [see Eq. (11)]:

$$d = M_t(c_{i,j,k}) \tag{11}$$

where, d represents the displayed numeric value. MGPP does not follow traditional graphical password schemes in which users are required to re-draw their password every time they login. In MGPP, the secret pattern structure is decoupled from the authentication information revealed to the user in each session. Users need only to remember the positions of the cells they chose for their secret pattern. The numbers revealed in these cells in each session are random and therefore different from session to session. Thus, the same secret pattern is always revealed as a different sequence of input for every login session.

The authentication workflow then continues on the server side where the session ID and a random grid is generated and sent to the client as a configuration for a grid which is then displayed within the application's GUI. On the client side, the user now selects all positions where he had registered a pattern before and enters the displayed digits in the correct order and submits the resulting sequence of digits to the authentication server. There the entered sequence of digits is changed back to its positional representation and is then compared with the protected pattern which had been stored during registration. Fig. 2 shows an overview of the MGPP's operational workflow for registration, authentication, for the encryption of the user's credentials and for the verification of stored credentials which are stored on the user's client.

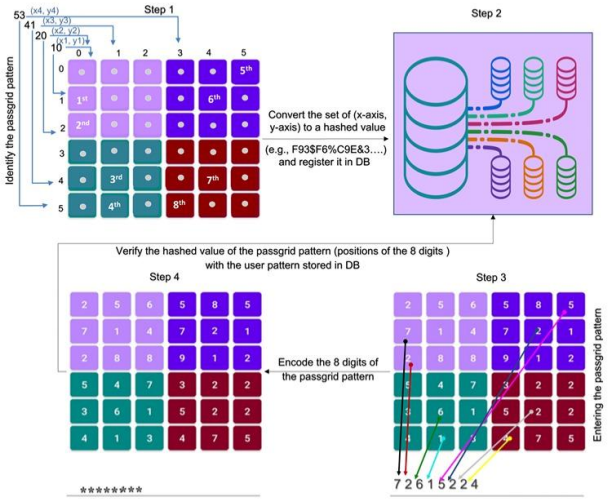


Fig. 2. The flow of the proposed approach working mechanism.

Let the authentication function be represented as $A(U, S_t)$, where U denotes the registered user identifier and S_t denotes the submitted authentication sequence. The authentication decision is expressed as [see Eq. (12)]:

$$A(U, S_t) = \begin{cases} 1, & \text{if the reconstructed pattern matches the stored pattern} \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

where, a value of 1 indicates successful authentication and a value of 0 indicates authentication failure.

The MGPP framework architecture is designed so that information visible to an attacker during a login session is not sufficient to determine the underlying secret pattern. MGPP enables multiple secret pattern grids to be chosen by users, session dependent numeric values to be randomly added to generated secret patterns and secret pattern verification to be enabled while keeping the login process familiar to legitimate users.

B. Registration Phase

The registration phase for MGPP (Multi-Grid Pattern-Project) framework comprises of steps to generate a secret user authentication pattern in the form of a multi-grid structure

having individual grid spaces. The user selects some of the cells from these individual grid spaces to form a secret authentication pattern or the primary login credential for a user. In traditional pattern-lock-based user authentication systems, a connection between all other cells for all connected cells within a user's pattern is stored. However, in the MGPP framework, the stored data for each user is only protected representations of the cells chosen by that user to be included in his/her pattern. The actual chosen pattern of a user cannot be figured out from the input given by that user during the registration process. Let the user identifier be represented as [see Eq. (13)]:

$$U = \{ID, Username\} \tag{13}$$

where, ID denotes the unique user identifier and $Username$ denotes the user's login name. After successful account creation, the server generates and displays the MGPP interface consisting of four independent 3×3grids. The user is then required to select an ordered pattern consisting of eight distinct points distributed across the available grids. Using the notation introduced in Section 3, the registered pattern Pconsists of an ordered sequence of $L = 8$ selected positions. Each selected point is represented by its coordinate location [see Eq. (14)]:

$$p_i = (x_i, y_i) \tag{14}$$

where, x_i and y_i denote the column and row coordinates, respectively. The complete positional representation of the user's pattern is therefore expressed as [see Eq. (15)]:

$$P = \{(x_1, y_1), (x_2, y_2), \dots, (x_8, y_8)\} \tag{15}$$

Our proposed framework for generating patterns for authentication purposes constrains these in order to increase security even more and to enlarge the number of possible patterns even further. First, every pattern for authentication has to consist of a fixed number of 8 points. Second, all points of a pattern have to be distributed on several grids, in contrast to only one grid. Third, it is not allowed to choose for a given authentication sequence identical positional selections for a pattern in all grids. So the additional constraints for creating a pattern strongly dissuade users from choosing a very weak authentication sequence.

After the pattern is selected, the server converts the positional sequence into a canonical pattern vector defined as [see Eq. (16)]:

$$V_P = [x_1, y_1, x_2, y_2, \dots, x_8, y_8] \tag{16}$$

which contains the ordered coordinates of all selected points. Since the sequence of selection is critical for authentication, any change in the order of the coordinates produces a different pattern representation.

To protect the registered pattern from disclosure, the positional vector is transformed into a cryptographic representation. Let the protected pattern credential be denoted by [see Eq. (17)]:

$$H_P = h(V_P) \tag{17}$$

where, $h(\cdot)$ represents the secure one-way hashing function employed by the system. The hash value H_P (stored credentials) is stored in the authentication database for the respective user ID. The storage of the hashed credentials using one-way hashing (as for the registration) does not reveal any information about the used positions, even if the database is attacked.

Following the registration process, all subsequent authentication sessions are carried out between the logged-in user and the authentication server. The user's logged-in session uses the user's registered pattern for verification against the stored hash value for that user, stored in the authentication database against the user's identifier H_P is. The authentication server will only be aware of the user's identifier and the protected pattern that was created from the registration authentication information (that was used to create the stored hash value H_P for that user's identifier) stored in the authentication database.

C. Login Authentication Phase

The login authentication phase verifies whether the submitted authentication sequence corresponds to the pattern established during registration. The proposed MGPP framework differs fundamentally from conventional graphical authentication systems because users do not redraw the graphical pattern during login. Instead, users identify the previously selected pattern positions and enter the numerical values currently displayed in those locations.

At the beginning of each authentication session t, the server generates a new random numerical assignment for all cells in the authentication interface. The generated mapping changes dynamically for every authentication session such that [see Eq. (18)]:

$$M_t \neq M_{t+1} \tag{18}$$

thereby ensuring that identical patterns produce different visible numerical sequences across successive login attempts.

Using the stored knowledge of the secret pattern positions, the user locates the corresponding cells within the randomized grids and observes the displayed values. The entered authentication sequence S_t, defined previously in Section III, contains the numerical values observed at the selected pattern positions during session t.

The submitted sequence is transmitted to the authentication server together with the user identifier. Based on the current session mapping M_t, the server reconstructs the associated positional sequence [see Eq. (19)]:

$$P_t = \{p_1, p_2, \dots, p_8\} \tag{19}$$

by identifying the cells corresponding to the submitted digits and their ordering. The reconstructed pattern is then transformed into its positional vector representation [see Eq. (20)]:

$$V_{P_t} = [x_1, y_1, x_2, y_2, \dots, x_8, y_8] \tag{20}$$

which is subsequently converted into a protected credential using the same hashing function employed during registration [see Eq. (21)]:

$$H_{P_t} = h(V_{P_t}) \tag{21}$$

Authentication is performed through comparison between the reconstructed hash value and the stored hash value. The authentication decision function is defined as [see Eq. (22)]:

$$A(U, S_t) = \begin{cases} 1, & H_{P_t} = H_P \\ 0, & H_{P_t} \neq H_P \end{cases} \quad (22)$$

where, $A(U, S_t) = 1$ indicates successful authentication and $A(U, S_t) = 0$ indicates authentication failure.

The proposed login mechanism is also observation-resistant. The numerical values that the user can see during his/her login session are changing from time to time while the underlying secret pattern does not change. Therefore, any observation by an attacker of a user's login process would result in only a temporary numerical sequence of values without any information about the user's secret pattern. In addition to this property, the repeated occurrences of the same digit in different cells as well as in different grids increase the ambiguity of such an observation, and therefore, it would be very difficult for the user to recognize and to reproduce the correct secret pattern during the login process, as is the case with simple graphical authentication mechanisms that use only one grid of cells. For example, Fig. 3 shows a login area where the digit "5" occurs in different cells of a grid. As can be seen, there are several possible ways to interpret the shown numerical values during such an observation by an attacker, and only by following one of these possible ways the attacker is able to try to recognize the secret pattern that is used for login by the user. However, the recognition of the correct pattern is very difficult because of ambiguity that is introduced by all the possible ways for the interpretation of the numerical values shown to the user during his/her login session. This is in contrast to simple graphical authentication mechanisms that use only one grid and therefore, the observation-resistance of our proposed login mechanism is increased significantly.

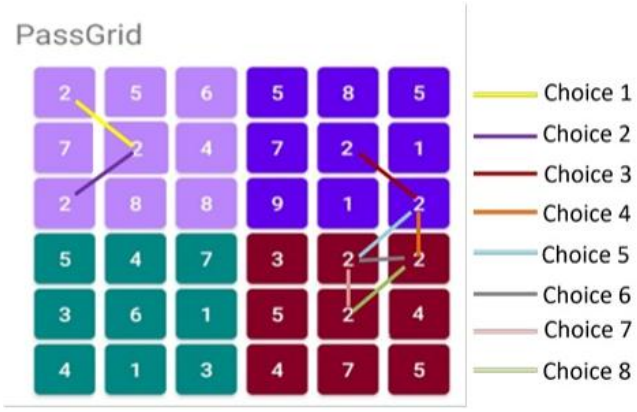


Fig. 3. An example of the several paths for the pattern that includes the number two twice, which can be very difficult for attackers to observe or guess.

The combination of dynamic numerical randomization, hidden positional verification, and cryptographic pattern validation enables the MGPP framework to authenticate legitimate users while simultaneously reducing susceptibility to shoulder surfing, brute-force guessing, and observation-based attacks.

D. Cryptographic Protection Mechanism

The security of the proposed MGPP framework can be considered from two aspects: the secret pattern selected by a user and the authentication credentials stored in the framework for verification. To protect a selected pattern from disclosure, a one-way hashing function is used to transform the positional representation of a user's pattern into a stored credential on the authentication server. To prevent disclosure of stored credentials in case of a database attack, the original pattern coordinates are not stored on the authentication server. Instead, the positional representation of a pattern is first transformed into a stored credential by using the hashing function and then the resulting credential is stored on the server. Let P denote the registered pattern defined previously in Section III. The positional information is first transformed into a canonical pattern vector [see Eq. (23)]:

$$V_P = [x_1, y_1, x_2, y_2, \dots, x_L, y_L] \quad (23)$$

which preserves both the selected locations and their ordering. Since the order of selection is an essential component of the authentication credential, any modification to the sequence results in a different vector representation.

To generate the protected credential, the pattern vector is processed using a secure one-way hash function. The resulting authentication credential is defined as [see Eq. (24)]:

$$H_P = h(V_P) \quad (24)$$

where, H_P = protected pattern credential, $h(\cdot)$ = cryptographic hash function, V_P= pattern vector representation. The generated hash value is stored in the authentication database and associated with the corresponding user identifier. Since the hashing operation is non-invertible, recovering the original pattern from the stored credential is computationally infeasible.

During the authentication phase, the submitted sequence is converted into its corresponding positional representation and transformed into a verification vector [see Eq. (25)]:

$$V_{P_t} = [x'_1, y'_1, x'_2, y'_2, \dots, x'_L, y'_L] \quad (25)$$

where, t denotes the current authentication session. The verification hash is then computed as [see Eq. (26)]:

$$H_{P_t} = h(V_{P_t}) \quad (26)$$

Authentication succeeds only when the newly generated credential matches the credential stored during registration. Therefore, the verification condition is expressed as [see Eq. (27)]:

$$H_{P_t} = H_P \quad (27)$$

Equality holds for selected positions, i.e., the positions that were selected and their order, equal to the positions and order that were registered for the first time.

In addition to protecting registered positions, credential information such as a user's password is also protected within the system. The framework uses Bcrypt hashing for passwords. Bcrypt is an adaptive one-way hash function which is designed to be slow (i.e., to be computationally expensive) and thus, for the purposes of a brute-force attack, is far more expensive than

typical hash functions. The cost c of Bcrypt hashing is a parameter of the hash function. The hashing operation may therefore be represented as [see Eq. (28)]:

$$H_P = h(V_P, c) \tag{28}$$

where, increasing c proportionally increases the computational resources required to evaluate candidate patterns. Large-scale offline password guessing attacks become much more expensive for the attacker.

The protection of stored and verified positional credentials in MGPP is completely independent from the numerical login randomization. This means that even if an attacker managed to get hold of the authentication sequence that has been transmitted or even gets access to the database where all authentication credentials are stored, the corresponding password pattern or passphrase cannot be discovered because all stored and verified positional credentials are encrypted. This multi-layered protection also covers attacks like credential theft, database attacks, replay attacks, and attacks that try to reconstruct a used password pattern or passphrase from already verified authentication credentials. Fig. 4 illustrates registration and authentication with credentials of MGPP. The process that transforms the given pattern information into the stored or verified protected positional credentials in the registration as well as in the login process is shown in more detail. Note that the real password pattern or passphrase information is not stored in plain text.

E. Threat Model

In this study, we perform a formal security analysis of our MGPP framework. We first present a detailed threat model that describes the attacks of an opponent with a given set of capabilities. In the subsequent security analysis and the related entropy analysis, we restrict ourselves to this threat model.

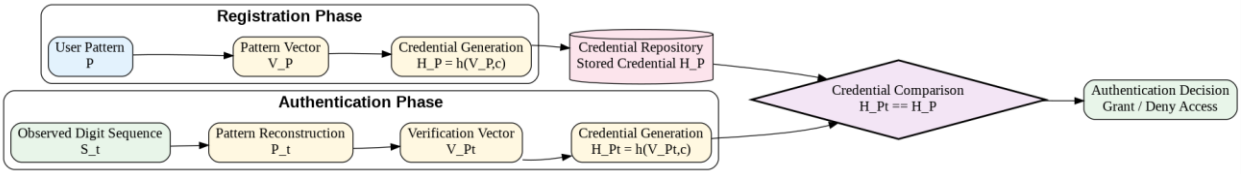


Fig. 4. Registration and authentication credential processing workflow in the MGPP framework.

This analysis will assume that the adversary has complete knowledge of the system and protocol under consideration, as well as the authentication function being used. Because of Kerckhoff's principle, the security of a cryptographic system resides in the authentication credentials for that system and not in the authentication algorithm(s) or protocol(s) for that system. Thus, the adversary has complete knowledge of the multi-grid (MGPP) framework under consideration as well as the pseudo-random location within each of the various grids used for pseudo-randomization of location, as well as complete knowledge of the login function in general.

Let the attacker be represented by the entity $\mathcal{A}$ whose objective is to recover the user's secret pattern or successfully impersonate a legitimate user. The attacker's capabilities are represented by the set [see Eq. (29)]:

$$\Gamma = \{\gamma_{obs}, \gamma_{rec}, \gamma_{net}, \gamma_{alg}\} \tag{29}$$

where, γ_{obs} denotes the capability to visually observe the login interface, γ_{rec} denotes the capability to record authentication sessions, γ_{net} denotes the capability to intercept transmitted authentication messages, γ_{alg} denotes knowledge of the authentication algorithm.

Under this threat model, the attacker may observe one or multiple authentication sessions and attempt to infer the underlying pattern from the visible numerical sequences entered by the user. However, the attacker is assumed to have no access to the user's memorized secret pattern and no direct access to the protected credentials stored on the server. Fig. 5 illustrates the threat model and attack surfaces considered in the security analysis of the proposed MGPP framework. The figure summarizes the adversarial capabilities, potential attack vectors, and the corresponding security objectives addressed by the framework. This model serves as the basis for the subsequent security evaluation and formal entropy analysis presented in Section VI-C.

The attacker's objective can therefore be expressed as maximizing the probability of successful authentication [see Eq. (30)]:

$$P_{succ} = Pr(\hat{P} = P) \tag{30}$$

where, P denotes the legitimate user's pattern, and $\hat{P}$ denotes the attacker's estimated pattern.

The primary security goal of the MGPP framework is to minimize P_{succ} by introducing uncertainty into the observation process.

The first attack scenario considered is shoulder surfing, in which the attacker directly observes the authentication process. Since the visible numerical values vary across authentication sessions according to the randomization function M_t, the observed authentication sequence S_t does not directly reveal the underlying positional pattern P. Consequently, observation of a single session provides only temporary authentication information.

The second attack scenario involves multi-session observation. In this case, the attacker records multiple login attempts and attempts to correlate observed sequences. Let [see Eq. (31)]:

$$\mathcal{S} = \{S_1, S_2, \ldots, S_T\} \tag{31}$$

represent the set of observed authentication sessions, where T denotes the number of recorded sessions. Because each session employs a distinct random mapping, the relationship between observed values and secret positions remains concealed, increasing the complexity of pattern reconstruction.

The third attack scenario considers brute-force guessing attacks. Let the total pattern space be represented by $| \Omega |$, where Ω denotes the set of all valid MGPP patterns. The probability of successfully guessing the correct pattern in a single attempt is therefore [see Eq. (32)]:

$$P_{guess} = \frac{1}{| \Omega |} \tag{32}$$

which decreases as the pattern space increases.

In the final attack scenario, an attacker performs eavesdropping and intercepts credentials. The attacker is able to capture all authentication messages that are transmitted between client and server. However, the protected cryptographic credentials used for verification in our approach, instead of the transmitted pattern information of the user, do not reveal any information.

These goals for the MGPP framework are used in Section VI-C for formal entropy analysis. They can also be used to evaluate the newly proposed framework for so-called observation-type attacks, as well as for typical brute-force attacks.

The MGPP framework proposed in this study is a unified system for object authentication. It selects a multi-grid pattern in a random numerical manner, and for a session, it protects the credentials by means of cryptography, and in an adversarial manner, it maps possible threats. In order to make authentication more robust against observation-based attacks and at the same time more practical for use, in the following section, the authors will describe the implementation of an MGPP-supported mobile system for object authentication (see Fig. 5).

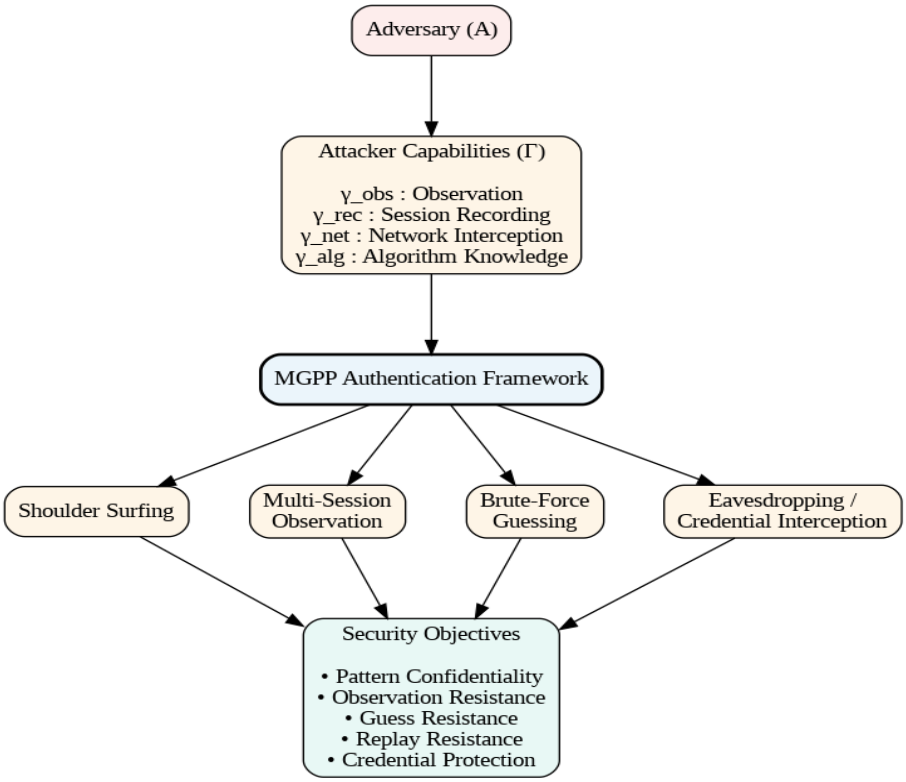


Fig. 5. Threat model and attack surfaces considered in the MGPP authentication framework.

All models were implemented in Python using TensorFlow and PyTorch libraries and executed on an NVIDIA GPU-based system. Reinforcement learning agents are DQN, PPO, and A2C.

IV. System Development

A. Mobile Application Architecture

In this work, a Multi-Grid Pattern-Based Password (MGPP) authentication framework is developed and implemented as a client-server mobile authentication system. The system consists of an Android client, an authentication server, and a database that holds the credentials in a protected form and other account information of users. The Android client is used for registration and login of users. The authentication server is used for managing the user's credentials, for generating the random grids for a session, for verifying an input authentication string against the stored credentials, and for database access.

MGPP architecture is multi-grid and portable. In the registration phase of the architecture, the secret pattern is spread on all of the grids given by the user. In the authentication phase, the mobile application of the user first gets a new random grid pattern for each session from the server and then shows it to the user. The user then enters the sequence of clicks as usual, and the user application sends the sequence to the server for verification. In this system, actual verification of credentials is done in a controlled environment of server. Thus, it is secure enough.

B. User Interface

The graphical user interfaces for the MGPP framework are graphically illustrated in Fig. 6. The graphical user interfaces for the MGPP framework include interfaces for user registration, user login, and passgrid authentication. In order to keep the interfaces as simple as possible in order to support users with easy and comfortable authentication processes, we try to keep the high security standards of the MGPP framework.

During the registration process, users create their own secret authentication patterns. They select eight positions within the multiple grids provided, which are then changed to a protected format and saved on the server of the application. During login sessions, all grids for the session are provided to the user with all positions filled with dynamic numerical values. The user then selects the positions within the provided grids that correspond to the positions selected by the user during registration time. After correct position selection, the user enters the corresponding numerical values in the correct order.

The graphical user interface for the Multi-Grid Passgrid Pattern (MGPP) application is shown in Fig. 6. The interface does provide sufficient space to clearly show a Multi-Grid environment. However, to prevent attacks that rely on observation of the user interface as much as possible, the interface has been kept as simple as possible. The application makes use of multiple grids. The numerical values in the grids are changed on a dynamic basis. As a result of this, attacks that try to guess a user's pattern by observing the interface will be ambiguous and will most likely fail.

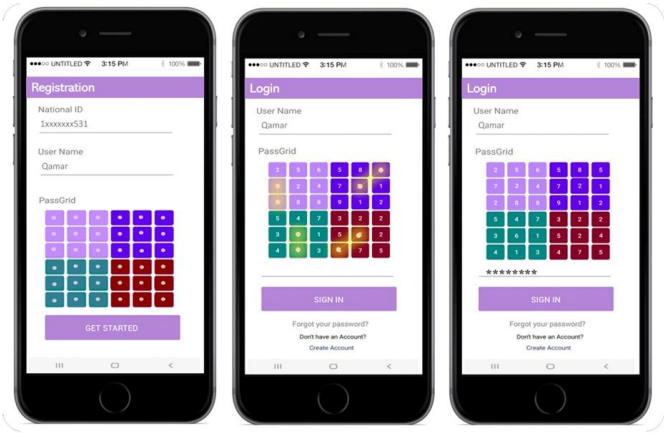


Fig. 6. Graphical user interface of the developed MGPP approach.

C. Server Components

Server-side functionality of the authentication framework comprises authentication management, password protection, session generation as well as interaction with the database. The core of the server, which has been implemented in PHP, for every new login request, generates a random numerical representation for all grid cells, which then gets sent to the client program.

On the server side, when a login request is received, the server first reconstructs the positional pattern corresponding to the entered authentication sequence by referring to the active session mapping for the session received earlier. Then, the generated and protected credential representation of the reconstructed pattern is compared with the stored credential of the user's registration information. If both the credentials match, then the server grants authentication to that user.

As a second layer of protection, all servers have been configured to use the protection mechanism, as outlined in Section III-D. Thus, instead of storing the coordinates of the pattern entered by the user to form the login credentials, the protected form of the user's credential is stored. As previously described, should the worst happen and a database suffers a complete database compromise with subsequent unauthorized access by a third party, such an attack will not be able to retrieve and then use any previously submitted authentication login credentials. This system can also be used to help prevent the Credential Recovery Attack, where an attacker would attempt to retrieve previously submitted login credentials from a database in the hope that he can then use them to gain access to protected resources in subsequent login attempts. The integration of secure credential storage, dynamic session generation, and server-side authentication verification enables the MGPP framework to provide a secure and practical authentication solution for mobile environments.

V. Evaluation of the MGPP Authentication System

The evaluation of the MGPP system has been carried out based on the assessment of two software quality aspects [45], [46]: performance and security. According to [47], it is a paramount requirement for cybersecurity testing to assess the technical implementation of the proposed authentication system and assess the human-system interaction. In addition to the experimental evaluation, the proposed framework is further analyzed through a formal threat model and entropy-based security analysis to provide theoretical evidence regarding its resistance to observation and guessing attacks.

Performance testing has been used to determine to what extent the system is powerful when applying different loads [48], [49]. More specifically, it has been implemented to assess how fast the aspects of the MGPP performance are under a specific workload, including load, response time, time complexity, stability, and scalability. A system-based evaluation method [50] was followed, where the experiment was achieved through a specific tool and collected the results immediately from the extracted log file from this tool. The performance evaluation was performed using the Junit 1 tool and under different conditions and working environments, such as using various network speeds and various mobile battery levels. In addition, the user-based testing method presented in [50]–[52] was used to ensure the compatibility of the applied method on different mobile devices with different specifications. During the evaluation, we applied some load on the app by increasing the number of users at a specific moment (1s) and over longer periods to check the performance in terms of: 1) desired load; 2) capability to scale up or scale out; 3) efficiency of the authentication process when a number of hits occur at a specific time; 4) the responsiveness of the authentication process in general. The average time complexity of number randomization was measured via a crowdtesting technique [52], where the average time complexity is measured during the experiment by different testers through 15 different Android mobile devices

with different specifications (device name, models, and Operating system versions) and different internet speeds.

Regarding security testing, different techniques have been performed to identify and minimize potential security risks in the proposed system and implement necessary countermeasures before releasing it to the public [47]. A static code analysis (SCA) [53] was executed using the "SonarLint"2 security testing tool by highlighting bugs and security vulnerabilities while developing the source code. This helped to identify potential issues and find remediation guidance on how to fix such issues to ensure the security posture and credibility of the proposed approach. In addition, vulnerability scanning/authentication scanning [53] methods were used to measure the capability of attackers to guess or observe the user's pattern and authenticate successfully. The security evaluation was conducted under the adversarial assumptions defined in the threat model presented in Section III-E, where attackers are assumed to possess knowledge of the authentication mechanism and may perform observation, recording, and guessing attacks. Furthermore, a document published by the Open Web Application Security Project (OWASP)3 has been adopted to ensure that the MGPP system is free of the top ten critical security risks. The experimental security assessment is complemented by a formal entropy analysis presented in Section VI-C, which quantitatively evaluates the theoretical password space and resistance of the proposed framework against partial-observation and brute-force attacks.

VI. Results and Discussion

This section elaborates on the core findings derived from the applied methods (Section V) and explains the significance of those results compared to the existing state-of-the-art approaches to answer the proposed research questions. In addition to the experimental findings, the results are interpreted in light of the formal threat model presented in Section III-E and the entropy-based security analysis introduced in Section VI-C. This combined evaluation provides both empirical and theoretical evidence regarding the security and robustness of the proposed MGPP authentication framework.

A. Performance Analysis

The results of performance testing on a variety of mobile device types and with different specifications show that the authentication process has a time complexity in the range from 1.5s to 2s for generating random numbers over the four grids. For the load testing, the results show that 40 users were dropped or disconnected when the load of the number of users reached 1000 at 1s, as shown in Fig. 7. This does not reveal a weakness in our approach; it is basically related to the network bandwidth capacity. This is consistent with the conclusion presented in the prior research [54], which states that as the load of user requests increases, the network becomes less powerful. This is the reason for dropping or disconnecting a few users in our results, where the MGPP approach does not reveal a weakness in up to 960 of 1000 users at 1s. The authors argue that the small number of dropped users and the slight increase in response speed prove the power and good reliability level of our approach in accepting multiple requests with low response time.

Fig. 7. Load testing results while increasing the number of users up to 1000 in 1s.

With respect to load and response speed, we tried to measure the response time with different values by increasing the number of users' authentication requests at 1s. The results from the test scenario clarify that the response speed of the completion of the authentication process and logging into the app requires less than 1s (97 ms) for each individual user. For the response speed of multiple authentication requests at 1s, it is clear from Fig. 8 that the average response speed of 750 multiple users' requests takes about 7.3 seconds, for 7500 users requests takes about 1.18 minutes, for 11250 users requests takes about 1.9 minutes, for 16500 users requests require about 2.5 minutes and for 32250 users requests take approximately up to 5 minutes. This indicates that the response time tends to increase very slowly and almost stabilizes with loads of 32250 and above. This small increase is considered a normal case for any software that uses network connections; this is because of the fact that when the load of requests is higher, more service requests are done [54], [55].

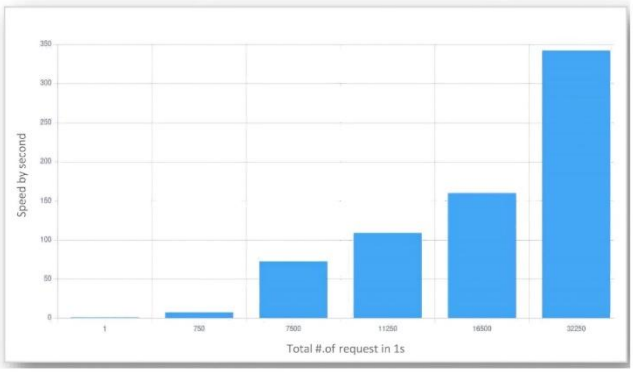


Fig. 8. The speed of the authentication process (response time) while the number of users' authentication requests is increased at 1s.

Fig. 9 presents the results of the stability and scalability evaluation of the proposed MGPP authentication process under increasing user workloads. The figure illustrates the behavior of the authentication system and database components as the number of concurrent users grows, providing insight into the ability of the framework to maintain stable performance under high-load conditions. For the scalability and stability of our approach, the findings showed a gradual increase in the overall period of time (1s to 53s) with a slight decrease in 7s with the load increase (reach 40.000 users). This decrease in the graph

from 15 seconds to 22 seconds does not indicate that a defect or error has occurred in our system regarding stability or scalability, but rather means that there were no requests, or some users were logged out of the system (in the scalability curve), as indicated by the results in the graph being synchronized with each other and no decrease was shown in one line rather than another. This is supported by the findings presented by Vahidnia et al. [56]. Consequently, this proves the capability of our authentication process and the database's ability to meet the growing needs through different conditions and maintains stability with the greatest number of users. From this, we can argue that our approach is highly stable and scalable regardless of the growth in user numbers. From this, we derived that the scalability of the system can impact its stability.

The performance testing of our approach demonstrated the high capability of the approach to complete the authentication process with a lower time complexity rate of random number generation (1.5s); a high-reliability level in handling up to 960 users at 1s, the ability to accept multiple requests with a low response time; the capability of maintaining stability and a high capacity to function more users, regardless of the growth in user numbers, we can answer RQ1 that the MGPP approach is efficient and can outperform a single-grid pattern password authentication approach. The performance results further indicate that the additional security mechanisms incorporated within the MGPP framework, including dynamic numerical randomization and protected credential verification, do not introduce significant computational overhead and therefore maintain the practical usability of the authentication process.

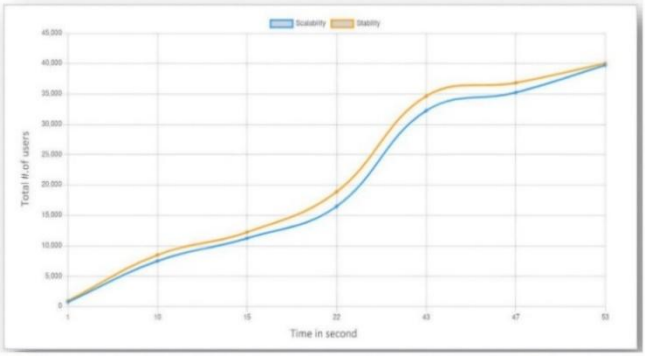


Fig. 9. The results of stability and scalability tests of the authentication process, given an increasing number of users.

B. Security Analysis

The results of the security evaluation and analysis show that the MGPP authentication system can protect users against security issues that are relevant to different types of attacks such as information leaks, shoulder surfing, pattern passgrid theft, and man-in-the-middle attacks. The security assessment was conducted under the adversarial assumptions defined in the threat model presented in Section III-E, where attackers are assumed to possess observation, recording, interception, and algorithm knowledge capabilities. The types of security issues produced from the static code security testing through the SonarLint tool revealed that the proposed code is capable of securing the proposed authentication process from expo-sure to cyberattacks, as there were no critical security issues discovered (see Fig. 10). Most of these issues were software bugs and were successfully resolved during the implementation phase to enhance the performance and ensure that the MGGP system was even more secure. This is supported by the conclusions drawn by Jayalath et al. [57] and Dasanayake et al. [58], in which SonarLint is effective in detecting security vulnerabilities and source code quality issues.

1) Shoulder Surfing (SS) attacks: The security analysis demonstrated that the proposed approach prevents a shoulder surfing attack, where it would be very difficult for the attacker to observe the entry or attack the passgrid pattern [59]. This prevention is due to the effectiveness of the proposed MGPP system, which implements a number repetition algorithm that would distract the user. It is also because of the implementation of the hidden pattern-entering technique where the numbers within the cells that represent the pattern are entered securely without clicking on the cells. This certainly helped reduce the risk of exposure to the pattern during the authentication process. This result is consistent with the findings explained in [59]. From the perspective of the proposed threat model, the attacker's probability of successfully reconstructing the secret pattern from a single observed authentication session remains low because the observed numerical sequence does not directly reveal the underlying positional pattern.

2) Eavesdropping attack: The results of security testing and implementation of the cryptographic algorithm described in Section III-B show the strength of the MGPP system in protecting the authentication process from eavesdropping attacks such as a Man-in-The-Middle (MiTM) attack, which could lead to unauthorised access to confidential data. This is due to the method of securing the position of selected pattern cells, where the cells' positions are hashed into a 60-character hash value. Predicting or analyzing the hash function to obtain the hash value by attackers will be very difficult, if not nearly impossible. This observation is consistent with the threat model assumptions, where intercepted authentication messages do not provide direct access to the protected pattern representation stored within the system.

3) Brute Force (BF) attacks: With regard to brute force attacks (guessability), the findings stemming from the security testing show the power of the MGPP approach in reducing the patterns' guessability opportunities for attackers. This is because of the use of 8 dots for the selected pattern and duplication of the numbers on the four grids' cells, in addition to the random change of these numbers in each new process of communication. This interpretation is supported by the finding described in [60]. The authors compare the time complexity for brute-forcing different pattern-based passwords based on their length in a 3x3 grid. The greater the number of connected dots, the more time it takes to be guessed. For example, a 4-dot pattern would take 1,079 days, a 5-dot pattern would take 4,949 days, a 6-dot pattern would take 18,049 days, and so on, as shown in Fig. 11. Therefore, this shows that the patterns with 8 and 9 connected dots provide more security against password attacks than other patterns with lower numbers. This is for the case of using the connected dots with a 3x3 grid.

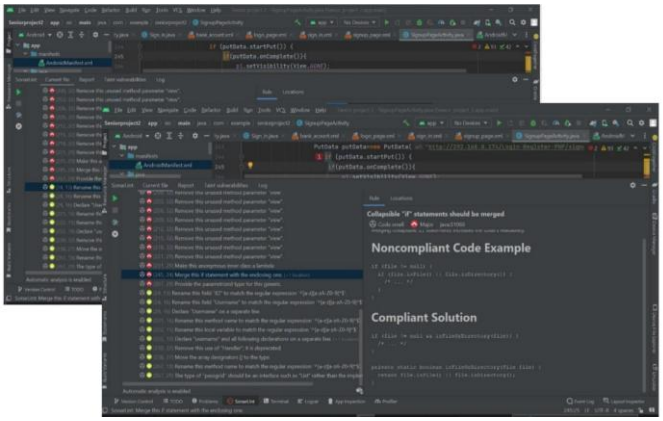


Fig. 10. A sample of the results obtained from static security testing via the SonarLint tool. No critical security vulnerabilities in the code have been found. All the issues discovered - ranging from major (red) to minor (green) levels - are related to coding mistakes, maintainability issues, and network configurations.

In the proposed approach, the authors considered using a discrete lines pattern with a length of 8-dots from four 3x3 grids, resulting in an overall 6x6 grid, which offers massive combinations of the possible generated patterns. This would result in an enormous increase in the number of possible combinations of the pattern generated and would most certainly decrease the chances of attackers guessing the user's pattern due to the increase in pattern combinations and the long time required to attempt to guess the passgrid pattern correctly. The same conclusion was reached by Potocký and Štulrajter [60], the more dots, the more time it takes for the attacker to guess the right pattern. While the experimental analysis demonstrates the practical resistance of the proposed approach against brute-force attacks, a more rigorous quantitative evaluation of the MGPP password space is presented in Section VI-C through formal entropy analysis. This analysis provides a theoretical estimation of the attacker's probability of successfully guessing the correct pattern under different observation scenarios.

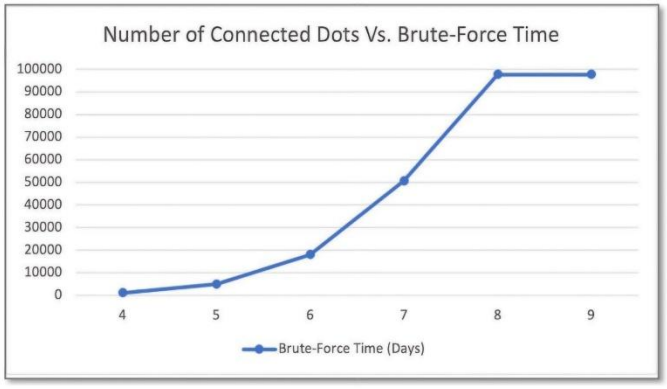


Fig. 11. Number of dots and the minimum time that might be needed by the attackers to guess the correct pattern.

C. Formal Entropy Analysis

While the experimental security evaluation demonstrates the practical resistance of the proposed MGPP framework against observation-based and brute-force attacks, a theoretical assessment is necessary to quantify the complexity of recovering the secret authentication pattern [61]. The entropy of an authentication credential gives an indication of the uncertainty of this credential. In graphical authentication, the entropy of a credential is used to determine the strength of an authentication system against guess attacks. The higher the entropy of a credential, the larger the pattern space of this credential and hence the lower the probability that an adversary can reconstruct the used pattern successfully.

Let Ω denote the set of all valid authentication patterns that can be generated within the MGPP framework. The total number of possible patterns depends on the number of available cells and the length of the selected pattern. As introduced in Section III, the proposed framework utilizes four 3×3grids, resulting in a total authentication space of $N = 36$ available cell positions. Let L denote the pattern length, where $L = 8$.

Assuming that each pattern consists of eight distinct selections and that the order of selection is significant, the total number of possible authentication patterns can be expressed as Eq. (33):

$$| \Omega | = P(N, L) = \frac{N!}{(N-L)!} \tag{33}$$

Substituting the values of N and L yields:

$$| \Omega | = \frac{36!}{28!} = 1{,}220{,}096{,}448{,}000 \tag{34}$$

possible authentication patterns.

The entropy of the authentication space is therefore defined as:

$$H(P) = \log_2 | \Omega | \tag{35}$$

which represents the minimum number of binary decisions required to uniquely identify a valid pattern. Substituting Eq. (34) into Eq. (35) gives [see Eq. (36)]:

$$H(P) = \log_2(1{,}220{,}096{,}448{,}000) \approx 40.15 \text{ bits} \tag{36}$$

This result indicates that the proposed MGPP framework provides substantially higher uncertainty than conventional graphical password systems operating on a single 3×3grid.

The probability of successfully guessing a valid pattern in a single attempt is given by Eq. (37):

$$P_{guess} = \frac{1}{| \Omega |} \tag{37}$$

which yields Eq. (38):

$$P_{guess} = 8.20 \times 10^{-13}. \tag{38}$$

Therefore, under random guessing conditions, the probability of successfully authenticating through a single brute-force attempt is negligible.

The threat model introduced in Section 3.5 further considers adversaries capable of observing a subset of the authentication process. Let $m$denote the number of pattern positions correctly inferred by an attacker through observation. The remaining search space is then given by Eq. (39):

$$| \Omega_{rem} | = P(N-m, L-m) = \frac{(N-m)!}{(N-L)!} \quad (39)$$

where, $0 \leq m < L$. Consequently, the residual entropy after partial observation becomes Eq. (40):

$$H_{rem} = \log_2 | \Omega_{rem} | \quad (40)$$

The search space is still very large even after some of the positions in the pattern have been disclosed to the observer. This is also a feature of the MGPP framework where the dynamic numerical randomization in the positions is separate from the positional pattern. Thus, by viewing the numerical pattern that has been disclosed to the observer, no information is disclosed as to the corresponding positions of the cells. Therefore, even when part of the search space has been disclosed by revealing some of the positions in the pattern, the size of the search space that can still be explored by testing all possible combinations of numbers for the revealed positions is very large. Table I shows the password space sizes for MGPP compared with other systems.

TABLE I. COMPARISON OF THE AUTHENTICATION SEARCH SPACE AND ENTROPY ACROSS COMMON AUTHENTICATION SCHEMES.

Authentication Scheme	Password Space	Approximate Entropy (bits)
4-Digit PIN	10^4	13.29
6-Digit PIN	10^6	19.93
Android Pattern Lock (3×3)	389,112	18.57
Single 3×3 Grid (8 positions)	362,880	18.47
Proposed MGPP (4×3×3 Grids, 8 positions)	**1,220,096,448,000**	**40.15**

The MGPP framework significantly increases the search space of needed authentication patterns when compared to simple pattern-based authentication schemes. Due to the existence of more than one grid, the total number of positions in all cells far exceeds that of a single simple grid. Although all cells are sequentially filled in a certain order to generate a specific pattern, an attacker cannot try all the possible patterns one by one to discover the secret pattern. In fact, the total search space of needed authentication patterns that are generated by the MGPP framework is so large that a brute force approach is entirely impractical. Hence, the MGPP framework provides an efficient and secure user authentication scheme.

Theorem 1. The MGPP framework provides a larger authentication search space than a conventional single-grid pattern authentication system of equal pattern length.

Proof. A conventional 3×3 pattern authentication system contains only nine selectable positions. For a pattern length of eight, the number of possible ordered patterns is Eq. (41):

$$P(9,8) = \frac{9!}{1!} = 362{,}880. \quad (41)$$

In contrast, the MGPP framework contains thirty-six selectable positions, yielding Eq. (42):

$$P(36,8) = 1{,}220{,}096{,}448{,}000. \quad (42)$$

Since [see Eq. (43)]:

$$P(36,8) \gg P(9,8) \quad (43)$$

The authentication search space of the MGPP framework is significantly larger than that of the conventional single-grid approach. Therefore, the proposed framework offers substantially stronger resistance against brute-force guessing attacks.

The theoretical findings are consistent with the experimental security results presented in Section VI-B. Both analyses indicate that increasing the number of selectable positions and introducing dynamic numerical randomization significantly increases the difficulty of pattern reconstruction. Consequently, the proposed MGPP framework achieves enhanced resistance against brute-force attacks, observation-based attacks, and pattern-guessing attempts while maintaining practical usability for mobile authentication applications.

In answer to the RQ2, the assessment of the approach proved the strength of the MGPP approach against brute-force attacks, eavesdropping attacks, and shoulder surfing attacks. Furthermore, the authors can confidently say that the proposed MGPP authentication approach is successful in minimizing financial fraud and reducing the risk of exposure to known software threats, as well as reducing exploitable vulnerabilities. Thus, this would increase security and protect credentials in online banking transactions. Based on experimental findings, the formal threat model, and the entropy-based security analysis, it can be concluded that pattern length, pattern structure, and pattern-space complexity are the primary factors influencing resistance against brute-force and observation-based attacks.

VII. CONCLUSION

In conclusion, this research proposed a new approach for reducing financial fraud issues and cyberattacks and researched the reasons behind their existence to date. On the basis of this, a novel pattern-based password authentication approach has been developed that focuses on increasing the number of randomizations using four 3x3 grids. The main motivation behind this is to propose a more secure authentication algorithm to resolve the previously mentioned issues, as they remain a major concern to online banking users. The proposed approach (MGPP) has been developed as an Android mobile app, and it has been evaluated through performance testing and security testing. In addition, a formal threat model was developed to characterize the capabilities of potential adversaries and to evaluate the security objectives of the proposed authentication framework under observation, recording, interception, and guessing attack scenarios. The results from performance testing show the high level of performance achieved when adopting the MGPP approach, as it offers a lower time complexity rate through random number generation (1.5s to 2s). Results also show that the MGPP approach is highly capable of keeping users safe whilst the functionality is not negatively affected with the growth in user numbers. The assessment also demonstrated the capability of the MGPP approach to protect its users from threats, financial fraud, and cyber-attacks. Furthermore, a formal entropy-based security analysis demonstrated that the proposed MGPP framework significantly increases the authentication search space compared with conventional pattern-based

authentication systems, thereby reducing the probability of successful brute-force and pattern-guessing attacks. This experimental evaluation, formal threat modeling, and entropy analysis show that MGPP-based designs for secure and scalable authentication for mobile applications are feasible. For future work, this study will carry out large-scale usability studies and implement MGPP-based designs for online banking environments, as well as more thoroughly and in greater detail test security-related aspects of this design by means of, for example, penetration testing, ethical hacking, and a detailed risk analysis. Furthermore, this study investigates the use of more adaptive grid structures, dynamic grid morphing, and the incorporation of multi-factor authentication into the MGPP framework in order to increase the security and robustness of our design.

ACKNOWLEDGMENT

The authors gratefully acknowledge the contributions of Lara S. Alghamdi, Sarah N. Alansari, Hams M. Alzahrani, Afaf M. Alogaili, and Reema A. Alebbie who participated, under the supervision of the authors, in the design, development, and testing of selected project components. Their dedication, technical efforts, and commitment provided valuable support to the successful completion of this work.

REFERENCES

[1] J. T. Quah and M. Sriganesh, ‘‘Real-time credit card fraud detection using computational intelligence,’’ Expert systems with applications, vol. 35, no. 4, pp. 1721–1732, 2008. [Online]. Available: https://doi.org/10.1016/j.eswa.2007.08.093

[2] A. E. Omolara, A. Alabdulatif, O. I. Abiodun, M. Alawida, A. Alabdulatif, H. Arshad et al., ‘‘The internet of things security: A survey encompassing unexplored areas and new insights,’’ Computers & Security, vol. 112, p. 102494, 2022. [Online]. Available: https://doi.org/10.1016/j.cose.2021.102494

[3] S. Ruoti, J. Andersen, and K. E. Seamons, ‘‘Strengthening password-based authentication,’’ in Way@ Soups, 2016.

[4] S. K. Sood, A. K. Sarje, and K. Singh, ‘‘Cryptanalysis of password authentication schemes: Current status and key issues,’’ in 2009 Proceeding of International Conference on Methods and Models in Computer Science (ICM2CS). IEEE, 2009, pp. 1–7. [Online]. Available: https: //doi.org/10.1109/ICM2CS.2009.5397977

[5] N. Sharma and B. Bohra, ‘‘Enhancing online banking authentication using hybrid cryptographic method,’’ in 2017 3rd International Conference on Computational Intelligence & Communication Technology (CICT). IEEE, 2017, pp. 1–8. [Online]. Available: https://doi.org/10.1109/CIACT. 2017.7977275

[6] V. K. Sharma, P. Mathur, and D. K. Srivastava, ‘‘Secure electronic fund transfer model based on two level authentication,’’ in 2018 second international conference on electronics, communication and aerospace technology (ICECA). IEEE, 2018, pp. 1338–1342. [Online]. Available: https://doi.org/10.1109/ICECA.2018.8474883

[7] S. Zhao and W. Hu, ‘‘Improvement on otp authentication and a possession-based authentication framework,’’ International Journal of Multimedia Intelligence and Security, vol. 3, no. 2, pp. 187–203, 2018. [Online]. Available: https://doi.org/10.1504/IJMIS.2018.096406

[8] D. E. Kurniawan, M. Iqbal, J. Friadi, F. Hidayat, and R. D. Permatasari, ‘‘Login security using one time password (otp) application with encryption algorithm performance,’’ Journal of Physics: Conference Series, vol. 1783, no. 1, 2021. [Online]. Available: https://doi/org/10.1088/1742-6596/1783/ 1/012041

[9] Y. S. Mezaal, D. A. Hammood, and M. H. Ali, ‘‘Otp encryption enhancement based on logical operations,’’ in 2016 Sixth International Conference on Digital Information Processing and Communications (ICDIPC). IEEE, 2016, pp. 109–112. [Online]. Available: https: //doi.org/10.1109/ICDIPC.2016.7470801

[10] J. Murkute, H. Nagpure, H. Kute, N. Mohadikar, and C. Devade, ‘‘Online banking authentication system using qr-code and mobile OTP,’’ International Journal of Engineering Research and Applications, vol. 3, no. 2, pp. 1810–1815, 2013.

[11] A. R. L. Reyes, E. D. Festijo, and R. P. Medina, ‘‘Securing one time password (OTP) for multi-factor out-of-band authentication through a 128-bit blowfish algorithm,’’ International Journal of Communication Networks and Information Security, vol. 10, no. 1, pp. 242–247, 2018.

[12] Y. S. Lee, N. H. Kim, H. Lim, H. Jo, and H. J. Lee, ‘‘Online banking authentication system using mobile-otp with qr-code,’’ in 5th International Conference on Computer Sciences and Convergence Information Technology. IEEE, 2010, pp. 644–648. [Online]. Available: https://doi.org/10.1109/ICCIT.2010.5711134

[13] F. A. A. A. Maqbali, ‘‘Strengthening password-based authentication,’’ Ph.D. dissertation, University of London Royal Holloway, 2019.

[14] S. W. Jirjees, A. M. Mahmood, and A. R. Nasser, ‘‘Passnumbers: An approach of graphical password authentication based on grid selection,’’ Journal homepage: http://iieta. org/journals/ijsse, vol. 12, no. 1, pp. 21–29, 2022. [Online]. Available: https://doi.org/10.18280/ijsse.120103

[15] R. Vadivukkarasi and K. Kuppusamy, ‘‘Enhanced authentication method using keyboard pattern based password generation,’’ International Journal for Modern Trends in Science and Technology, vol. 3, no. 8, 2017.

[16] S. Dedeepya, P. Swetha, and Y. RAJU, ‘‘Securing mobile banking in intelligent mobile devices-pattern based authentication approach,’’ International Journal of Scientific Research in Computer Science (IJSRCS), vol. 1, no. 3, 2013.

[17] S. Kiljan, H. Vranken, and M. van Eekelen, ‘‘Evaluation of transaction authentication methods for online banking,’’ Future Generation Computer Systems, vol. 80, pp. 430–447, 2018. [Online]. Available: https://doi.org/ 10.1016/j.future.2016.05.024

[18] L. Sharma and M. Mathuria, ‘‘Mobile banking transaction using fingerprint authentication,’’ in 2018 2nd International Conference on Inventive Systems and Control (ICISC). IEEE, 2018, pp. 1300–1305. [Online]. Available: https://doi.org/10.1109/ICISC.2018.8399016

[19] C. Lupu, V.-G. Găitan, and V. Lupu, ‘‘Fingerprints used for security enhancement of online banking authentication process,’’ in 2015 7th International Conference on Electronics, Computers and Artificial Intelligence (ECAI). IEEE, 2015, pp. 217–220. [Online]. Available: https://doi.org/10.1109/ECAI.2015.7301177

[20] A. Muley and V. Kute, ‘‘Prospective solution to bank card system using fingerprint,’’ in 2018 2nd International Conference on Inventive Systems and Control (ICISC). IEEE, 2018, pp. 898–902. [Online]. Available: https://doi.org/10.1109/ICISC.2018.8398930

[21] B. A. Sahar, A. F. Rahardian, E. Muchtar et al., ‘‘Fingershield atm–atm security system using fingerprint authentication,’’ in 2018 International Symposium on Electronics and Smart Devices (ISESD). IEEE, 2018, pp. 1–6. [Online]. Available: https://doi.org/10.1109/ISESD.2018.8605473

[22] M. H. Ali and Q. M. Al Azze, ‘‘Design and implementation a security system for bank using voice recognition,’’ International Journal of Power Electronics and Drive Systems, vol. 10, no. 4, p. 2126, 2019. [Online]. Available: https://doi.org/10.11591/ijpeds.v10.i4.pp2126-2129

[23] E. D. Dimaunahan, A. H. Ballado, F. R. G. Cruz, and J. C. D. Cruz, ‘‘Mfcc and vq voice recognition based atm security for the visually disabled,’’ in 2017IEEE 9th international conference on humanoid, nanotechnology, information technology, communication and control, environment and management (HNICEM). IEEE, 2017, pp. 1–5. [Online]. Available: https://doi.org/10.1109/HNICEM.2017.8269516

[24] A. E. Omolara, A. Jantan, O. I. Abiodun, H. Arshad, and N. A. Mohamed, ‘‘Fingereye: improvising security and optimizing atm transaction time based on iris-scan authentication,’’ International Journal of Electrical & Computer Engineering (2088-8708), vol. 9, no. 3, 2019. [Online]. Available: https://doi.org/10.11591/ijece.v9i3.pp1879-1886

[25] R. C. Agidi, ‘‘Biometrics: the future of banking and financial service industry in nigeria,’’ International Journal of Electronics and Information Engineering, vol. 9, no. 2, pp. 91–105, 2018. [Online]. Available: https://doi.org/10.6636/IJEIE.2018129(2).04)

[26] Z. Erlich and M. Zviran, ''Authentication practices from passwords to biometrics,'' *Encyclopedia of Information Science and Technology, Third Edition*, pp. 4248–4257, 2015. [Online]. Available: https://doi.org/10.4018/978-1-4666-5888-2.ch417

[27] N. A. Karim and Z. Shukur, ''Using preferences as user identification in the online examination,'' dynamics, vol. 19, no. 20, p. 21, 2016. [Online]. Available: https://doi.org/10.18517/ijaseit.6.6.1412

[28] M. Awang, M. Mohamed, R. Mohamed, A. Ahmad, and N. Rawi, ''A pattern-based password authentication scheme for minimizing shoulder surfing attack,'' International Journal on Advanced Science, Engineering and Information Technology, vol. 7, no. 3, pp. 1049–1055, 2017. [Online]. Available: https://doi.org/10.18517/ijaseit.7.3.1517

[29] C. Paar and J. Pelzl, The Advanced Encryption Standard (AES). Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 87–121. [Online]. Available: https://doi.org/10.1007/978-3-642-04101-3_4

[30] T. Nie and T. Zhang, ''A study of des and blowfish encryption algorithm,'' in Tencon 2009-2009 IEEE Region 10 Conference. IEEE, 2009, pp. 1–4. [Online]. Available: https://doi.org/10.1109/TENCON.2009.5396115

[31] D. Soundari, R. Aravindh, S. Abishek et al., ''Enhanced security feature of atm's through facial recognition,'' in 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE, 2021, pp. 1252–1256. [Online]. Available: https://doi.org/10.1109/ICICCS51141.2021.9432327

[32] R. A. Rashid, N. H. Mahalin, M. A. Sarijari, and A. A. A. Aziz, ''Security system using biometric technology: Design and implementation of voice recognition system (vrs),'' in *2008 international conference on computer and communication engineering*. IEEE, 2008, pp. 898–902. [Online]. Available: https://doi.org/10.1109/ICCCE.2008.4580735

[33] P. M. Chauhan and N. P. Desai, ''Mel frequency cepstral coefficients (mfcc) based speaker identification in noisy environment using wiener filter,'' in *2014 International Conference on Green Computing Communication and Electrical Engineering (ICGCCEE)*. IEEE, 2014, pp. 1–5. [Online]. Available: https://doi.org/10.1109/ICGCCEE.2014.6921394

[34] G. Tzanetakis, G. Essl, and P. Cook, ''Audio analysis using the discrete wavelet transform,'' in Proc. conf. in acoustics and music theory applications, vol. 66. Citeseer, 2001.

[35] Z.-b. Wu and J.-q. Yu, ''Vector quantization: a review,'' Frontiers of Information Technology & Electronic Engineering, vol. 20, no. 4, pp. 507–524, 2019. [Online]. Available: https://doi.org/10.1631/FITEE.1700833

[36] S. Venkatraman and I. Delpachitra, ''Biometrics in banking security: a case study,'' Information Management & Computer Security, vol. 16, no. 4, pp. 415–430, 2008. [Online]. Available: https://doi.org/10.1108/09685220810908813

[37] B. A. Scaria and R. K. Megalingam, ''Enhanced e-commerce application security using three-factor authentication,'' in 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE, 2018, pp. 1588–1591. [Online]. Available: https://doi.org/10.1109/ICCONS.2018.8662831

[38] A. T. Kiyani, A. Lasebae, K. Ali, and M. Ur-Rehman, ''Secure online banking with biometrics,'' in 2019 International Conference on Advances in the Emerging Computing Technologies (AECT). IEEE, 2020, pp. 1–6. [Online]. Available: https://doi.org/10.1109/AECT47998.2020.9194214

[39] I. A. Abdullah and J. J. Stephan, ''A survey of face recognition systems,'' Ibn AL-Haitham Journal For Pure and Applied Sciences, vol. 34, no. 2, pp. 144–160, 2021. [Online]. Available: https://doi.org/10.30526/34.2.2620

[40] Y. Kortli, M. Jridi, A. Al Falou, and M. Atri, ''Face recognition systems: A survey,'' Sensors, vol. 20, no. 2, p. 342, 2020. [Online]. Available: https://doi.org/10.3390/s20020342

[41] A. Nema and A. Ranjan, ''Pattern point based graphical authentication,'' International Journal of Engineering and Techniques, vol. 4, no. 2, 2018. [Online]. Available: https://doi.org/10.29126/23951303

[42] P. Sriramya and R. A. Karthika, ''Providing password security by salted password hashing using Bcrypt algorithm,'' ARPN Journal of Engineering and Applied Sciences, vol. 10, no. 13, pp. 5551–5556, 2015.

[43] ''Php: password_hash - manual.'' [Online]. Available: https://www.php.net/manual/en/function.password-hash.php

[44] N. Katrandzhiev, D. Hristozov, and B. Milenkov, ''A comparison of password protection methods for web-based platforms implemented with php and mysql,'' International Journal on Information Technologies & Security, №, vol. 11, no. 2, pp. 97–106, 2019.

[45] P. A. Laplante and M. Kassab, What every engineer should know about software engineering. CRC Press, 2022.

[46] D. Tobin, A. Bogomolov, and M. Golosovskiy, ''Model of organization of software testing for cyber-physical systems,'' in Cyber-Physical Systems: Modelling and Industrial Application. Springer, 2022, pp. 51–60. [Online]. Available: https://doi.org/10.1007/978-3-030-95120-7_5

[47] L. Maghrabi, E. Pfluegel, L. Al-Fagih, R. Graf, G. Settanni, and F. Skopik, ''Improved software vulnerability patching techniques using cvss and game theory,'' in 2017 International Conference on Cyber Security And Protection Of Digital Services (Cyber Security), 2017, pp. 1–6. [Online]. Available: https://doi.org/10.1109/CyberSecPODS.2017.8074856

[48] Z. A. Hamza and M. Hammad, ''Testing approaches for web and mobile applications: An overview,'' International Journal of Computing and Digital Systems, vol. 9, no. 4, pp. 657–665, 2020. [Online]. Available: http://dx.doi.org/10.12785/ijcds/090413

[49] A.-R. Mawlood-Yunis, ''Debugging and testing using junit, espresso, and mockito frameworks,'' in Android for Java Programmers. Springer, 2022, pp. 135–181. [Online]. Available: https://doi.org/10.1007/978-3-030-87459-9_4

[50] Q. Naith, ''Crowdsourced testing approach for mobile compatibility testing,'' Ph.D. dissertation, University of Sheffield, 2021.

[51] Q. Naith and F. Ciravegna, ''The key considerations in building a crowd-testing platform for software developers,'' in Proceedings of the 4th International Conference on Crowd Science and Engineering, 2019, pp. 50–57. [Online]. Available: https://doi.org/10.1145/3371238.3371247

[52] Q. Naith and F. Ciravegna, ''Definitive guidelines toward effective mobile devices crowdtesting methodology,'' International Journal of Crowd Science, vol. 4, no. 2, pp. 209–228, 2020. [Online]. Available: https://doi.org/10.1108/IJCS-01-2020-0002

[53] M. Felderer, M. Büchler, M. Johns, A. D. Brucker, R. Breu, and A. Pretschner, ''Security testing: A survey,'' in Advances in Computers. Elsevier, 2016, vol. 101, pp. 1–51. [Online]. Available: https://doi.org/10.1016/bs.adcom.2015.11.003

[54] J. Lima, R. Dantas, C. Kamienski, and J. Kelner, ''Performance analysis of network composition in ambient networks,'' in 2007 Brazilian Networking Conference, 2007.

[55] F. Alqahtani, M. Amoon, and A. A. Nasr, ''Reliable scheduling and load balancing for requests in cloud-fog computing,'' Peer-to-Peer Networking and Applications, vol. 14, pp. 1905–1916, 2021. [Online]. Available: https://doi.org/10.1007/s12083-021-01125-2

[56] M. H. Vahidnia, A. A. Alesheikh, and S. K. Alavipanah, ''A multi- agent architecture for geosimulation of moving agents,'' Journal of Geographical Systems, vol. 17, pp. 353–390, 2015. [Online]. Available: https://doi.org/10.1007/s10109-015-0218-2

[57] L. Jayalath, K. Dharshana, and R. Rathnayake, ''Towards secure software engineering,'' South asian Res J Eng Technol, vol. 2, no. 6, pp. 45–53, 2020. [Online]. Available: https://doi.org/10.36346/sarjet.2020.v02i06.001

[58] S. L. Dasanayake, A. Mudalige, and M. L. T. Perera, ''Framework for secure coding: An algorithmic approach for real-time detection of secure coding guideline violations,'' 2021.

[59] J. Lai and E. Arko, ''A shoulder-surfing resistant scheme embedded in traditional passwords,'' in Proceedings of the 54th Hawaii International Conference on System Sciences, 2021.

[60] S. Potocky` and J. Štulrajter, ''The human interface device (hid) attack on android lock screen non-biometric protections and its computational complexity,'' Science & Military Journal, vol. 17, no. 1, pp. 29–36, 2022. [Online]. Available: https://doi.org/10.52651/sam.a.2022.1.29-36

[61] L. Bošnjak, J. Sreš, and B. Brumen, ''Brute-force and dictionary attack on hashed real-world passwords,'' in 2018 41st international convention on information and communication technology, electronics and microelectronics (mipro). IEEE, 2018, pp. 1161–1166. [Online]. Available: https://doi.org/10.23919/MIPRO.2018.8400211

Unsanctioned Intelligence: Mitigating the Risks of 'Shadow AI' and Proprietary Data Leaks in the Remote Workforce

Adel Saad Assiri
Department of Business Informatics, College of Business
King Khaled University, Abha, 61421, Saudi Arabia

***Abstract*—The high implementation of remote working models has only increased the pace of using artificial intelligence tools by employees to improve their productivity, automate routine tasks, and aid in decision-making. Non-committal application of AI tools, or Shadow AI, has, however, become a major organizational risk, especially when the processed information is sensitive, proprietary, or controlled and not handled in line with the existing organizational policies. The current mitigation is based on fairly flat policies, manual checks, or limited monitoring systems, which cannot provide real-time awareness, dynamic risk response, or gated controls within the dynamic remote workforce. To minimize the leaking of proprietary data, this study suggests a single framework of Shadow AI risk mitigation to detect, evaluate, and control unsanctioned AI utilisation. The architecture incorporates automated scoring of risks, policy-sensitive enforcement, risk uncertainty estimation, and governance-driven adaptation as a way of balancing security, usability, and operational efficiency. Experimental results based on synthetic enterprise data (20,000 activity records) and real-world organizational analytics (7,500 interaction records) show that it can detect jobs with an accuracy of 91.6, reduce high-risk Shadow AI interactions by a factor of 62.6 and reduce policy violations by half, using less than 16% computation overhead. The suggested solution offers a scaling- and governance-based solution to securing the use of AI solutions in distributed working scenarios.**

Keywords*—Shadow AI; remote workforce security; data leakage prevention; AI governance; trustworthy AI; enterprise risk management; policy compliance; uncertainty-aware systems

I. Introduction

Remote and hybrid working has increased the adoption of artificial intelligence tools by employees to enhance their productivity and decision support. Although these tools are beneficial in terms of operations, their use without authorization, also known as Shadow AI, has proven to be a major organizational threat in cases where AI systems are implemented without established enterprise governance and oversight systems [1], [2].

Shadow AI has the potential to leak proprietary, confidential, or regulated information onto external platforms, which is a critical threat. Remote working conditions strengthen this threat due to reduced control over centralization, heightened use of personal devices, and diminished identification of authorized and foreign services [3], [4]. Workers can easily provide third-party artificial intelligence tools with confidential data without realizing that data will be stored, trained, or used to comply with regulations, without being aware that the data will be leaked, which will cause intellectual property theft and violations of regulatory policies [5].

Current enterprise responses are isolated and are mostly in reactive mode. The policy-based controls rely on voluntary compliance and are not enforceable, whereas the technical controls, like data leakage prevention and network monitoring, provide a limited insight into AI-specific interactions [6], [7]. Besides, existing AI governance systems are mostly centred on approved systems, which have no control over Shadow AI practices [8]. These restrictions are increased by the fact that there is no uncertainty-based risk assessment since most systems are based on set rules and methods that lack contextual variability and trust in enforcement action [9].

In an attempt to curb these issues, the current study suggests a combined system of Shadow AI threats within the setting of remote labour. This design integrates automated detection, modelling of the risk that humans adapt to, policy-oriented enforcement, and uncertainty to allow real-time, transparent, and governance-orientated control of unsanctioned AI use. The experimental assessment of the proposed framework supports the idea of the work and proves that the risk can be reduced effectively without impacting the Fig 1. shows how unsanctioned AI use in a remote workforce is identified, assessed for risk, governed through policy enforcement, and guided toward secure and compliant AI usage.

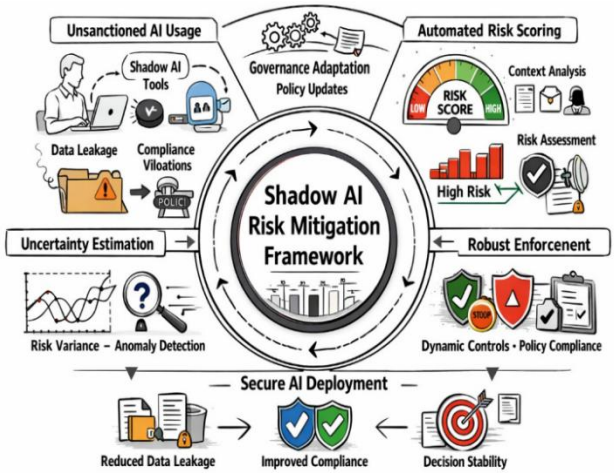


Fig. 1. Graphical overview of the proposed Shadow AI-risk mitigation framework for remote workforces.

The rest of this study is structured in the following way. Section II provides a literature review on the topic of Shadow AI, remote workforce security, and currently used governance and mitigation strategies. Section III entails the problem statement and sets the objectives of the research. Section IV captures the proposed risk mitigation framework and methodology. Section V presents the protocol of the experiment and evaluation. Section VI provides the discussion of the results and performance analysis. Conclusively, the study ends with Section VII, which indicates future research directions.

II. Related Research

Current studies on AI risk mitigation in enterprise security, insider threat management, data leakage prevention, and AI governance exist. In preliminary research on shadow IT, insecurity and compliance threats that occur by using unsanctioned software were identified, and such risks have grown with the informational and learning-focused characteristics of the current AI tools [10].

The studies on remote workforce security point towards the exposure to insider threats and accidental data leakage since decentralized and bring-your-own-device settings are more prone to this [11]. Even though numerous data leakage prevention systems have been extensively implemented to manage sensitive data flow, in the majority of cases, they are content-based data inspection-based systems that do not have semantic and contextual understanding to effectively identify data exposure by AI devices [12].

AI governance models give high importance to transparency, accountability, and compliance in the implementation of AI in organizations, but mainly pay attention to sanctioned systems with little to no control over unsanctioned AI use [13]. It is also noted that reliable AI studies emphasize explainability, robustness, and human control, yet its direct use in Shadow AI situations is limited [14]. Recent research has recognized the generative AI-related threats to intellectual property and regulatory compliance, but newer solutions are typically disjointed, either focusing on policy or technical surveillance individually [15].

All in all, it has been demonstrated that there is a significant gap in the literature: a lack of an integrated, uncertainty-sensitive framework combining detection, governance, and enforcement to address the risks posed by Shadow AI in remote workplaces. This gap is filled in this study by suggesting a holistic and adaptive mitigation strategy [16]. Table I provides a comparative overview of existing AI-driven security and governance approaches, highlighting differences in application domains, system architectures, core techniques and the use of blockchain technologies. It summarizes key limitations of prior work and underscores the need for an integrated, adaptive, and governance-aligned framework for managing Shadow AI risks.

TABLE I. Comparative Analysis of Existing AI-Driven Security and Governance Approaches

Ref.	Research Focus	System Type	Methodology	Key Contribution	Limitation
[10]	Shadow IT risks	Centralized	Policy analysis	Identified risks of unsanctioned software usage	Not AI-specific
[11]	Insider threats	Centralized	Behavioral analysis	Highlighted data leakage risks in remote work	Limited real-time control
[12]	Data leakage prevention	Centralized	Content-based inspection	Controlled sensitive data flows	Weak semantic AI awareness
[13]	AI governance	Centralized	Lifecycle management	Ensured transparency and compliance	Focused on sanctioned AI only
[14]	Trustworthy AI	Centralized	Explainability models	Improved accountability in AI systems	Limited Shadow AI coverage
[15]	Generative AI risks	Enterprise	Risk assessment	Addressed IP and compliance risks	Fragmented mitigation
[16]	Adaptive AI governance	Distributed	Feedback-driven control	Highlighted governance adaptation	Lacked unified enforcement

III. Problem Statement and Research Objectives

The growing use of remote work increased the unsanctioned use of artificial intelligence technology, often known as Shadow AI. Even though such tools make organisations more productive, their unregulated use puts organisations at risk of propagating their proprietary data, violating regulations, and lacking transparency over AI-powered processes. Employees are unwillingly sharing sensitive data with other AI platforms operating outside the remit of enterprise security. The current mitigation strategies are based on stagnant policies or old-school monitoring mechanisms that do not consider the AI-specific context and cannot handle any uncertainty in connection with the dynamic user behaviour and changing AI-based services. It leads to poor management of risk and the necessity to introduce an adaptive and flexible framework (with uncertainty awareness) to regulate the application of Shadow AI in distributed employment.

This research aims to achieve results by:

- To identify any unsanctioned AI use in distant employee set-up. To gauge the threat of sensitive information exposure during interactions with Shadow AI.
- To incorporate uncertainty-aware, unbiased enforcement choices.
- To help establish flexible policy implementation in various roles and scenarios.
- To become more resilient to the changing AI tools and patterns of use.
- To guarantee real-time enterprise deployment scalability and efficiency.

IV. Proposed Methodology

The proposed methodology introduces an integrated framework for Shadow AI risk mitigation that combines automated risk modelling, uncertainty estimation, robustness

analysis, and governance-driven adaptation. The framework operates as a continuous decision-support system that evaluates AI-related user activities and determines appropriate mitigation actions in real time. Fig. 2 explains how user activity is analyzed to estimate risk, handle uncertainty, enforce policies, and continuously adapt governance to manage Shadow AI securely.

The activity context is represented by an input feature vector in Eq. (1):

$$x = [x_1, x_2, \dots, x_n] \tag{1}$$

where, x denotes the activity context vector, x_i represents the i-th contextual feature, and n is the total number of input features.

An initial risk score is computed using a predictive function in Eq. (2):

$$r_0 = f(x; \theta) \tag{2}$$

where, r_0 is the preliminary risk score, $f(\cdot)$ denotes the risk estimation model, and θ represents the model parameters.

To align predictions with organizational governance requirements, a policy deviation measure is defined as shown in Eq. (3):

$$\Delta_p = | r_0 - \tau | \tag{3}$$

where, Δ_p denotes the policy deviation and τ is the policy-defined risk threshold.

A policy-aware adjustment is applied as shown in Eq. (4):

$$r_p = r_0 - \lambda \Delta_p \tag{4}$$

where, r_p is the policy-adjusted risk score and λ controls the strength of policy enforcement.

Uncertainty associated with the risk prediction is quantified using variance estimation in Eq. (5):

$$u = Var(f(x)) \tag{5}$$

where, u represents the predictive uncertainty of the risk estimation model.

A confidence-aware risk score is then computed as shown in Eq. (6):

$$r_u = r_p - \alpha u \tag{6}$$

where, r_u is the uncertainty-adjusted risk score and α controls the penalization of uncertainty.

To enhance robustness, sensitivity to small perturbations in the input features is evaluated as shown in Eq. (7):

$$s = \| f(x + \epsilon) - f(x) \| \tag{7}$$

where, s denotes sensitivity and ϵ is a bounded perturbation vector.

A robustness correction is applied as shown in Eq. (8):

$$r_s = r_u - \beta s \tag{8}$$

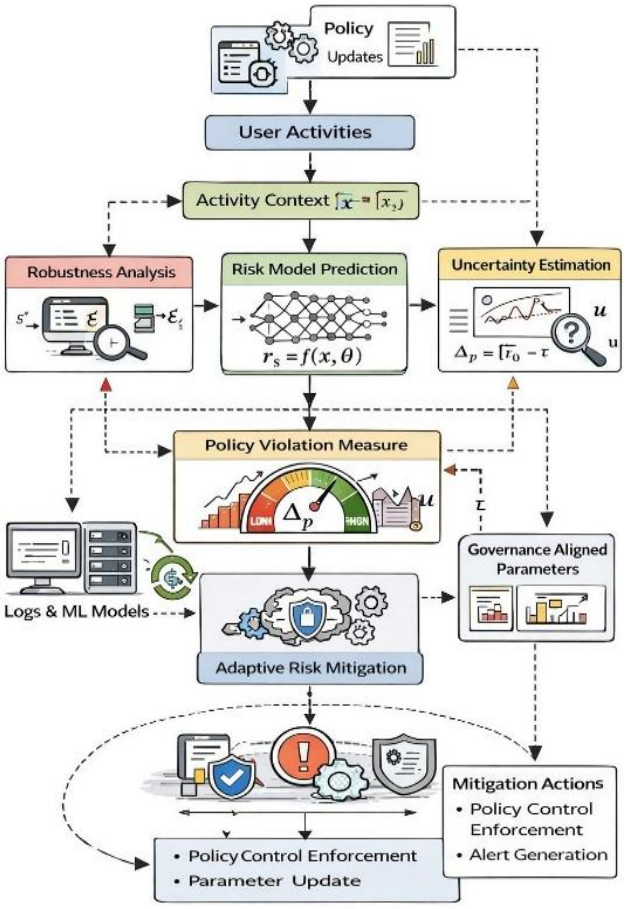


Fig. 2. Architecture of the proposed shadow AI risk mitigation methodology.

where, r_s is the stabilized risk score and β controls robustness enforcement.

Finally, governance adaptation is achieved through parameter updates driven by compliance feedback in Eq. (9):

$$\theta_{t+1} = \theta_t - \eta \nabla L(r_s) \tag{9}$$

where, θ_{t+1} and θ_t denote updated and current model parameters, respectively, η is the learning rate, and $L(\cdot)$ represents a governance-aligned loss function.

The complete operational workflow of the framework is summarized in Algorithm 1, and the system architecture illustrating the interaction between monitoring, risk evaluation, uncertainty handling, and policy enforcement components is shown in Fig. 3.

Algorithm 1. Shadow AI Risk Detection and Mitigation Process	
1.	Initialize model parameters θ and governance control variables $\lambda, \alpha, \beta, \eta$.
2.	Collect remote workforce activity signals and construct the contextual feature vector x.
3.	Compute the preliminary risk score using the predictive model: $r_0 = f(x; \theta)$.
4.	Evaluate the deviation between the predicted risk and the policy threshold: $\Delta_p = \| r_0 - \tau \|$

5. Apply policy-aware adjustment to obtain the calibrated risk score: $r_p = r_0 - \lambda \Delta_p$
6. Estimate predictive uncertainty using a variance-based approximation: $u = Var(f(x))$.
7. Penalize overconfident predictions to compute the uncertainty-aware risk score: $r_u = r_p - \alpha u$.
8. Introduce a bounded perturbation $\epsilon$to assess sensitivity of the model output.
9. Compute robustness sensitivity as $s = \| f(x+\epsilon) - f(x) \|$.
10. Apply robustness correction to obtain the stabilized risk score: $r_s = r_u - \beta s$.
11. Update model parameters using governance-aligned feedback: $\theta_{t+1} = \theta_t - \eta \nabla L(r_s)$.

The suggested methodology presents a generalized decision-support framework, which comprises the following operations that are specific to risk modeling in context, uncertainty estimation, robust analysis, and governance-based adaptation. The framework guarantees the transparency of control over Shadow AI usage, stability, and uncertainty-sensitive corrections in addition to operational flexibility by integrating predictive modeling and policy-aware and uncertainty-sensitive corrections.

V. Experimental Setup

The suggested AI-based model was tested in an experimental platform on a sensor-equipped CNC milling machine where realistic machining dynamics in controlled, albeit realistic, operating conditions were simulated. The experimental platform was made up of a three-axis CNC milling machine equipped with cutting force sensors, tri-axis vibration sensors, acoustic emission sensors, and spindle current sensors. Sensor streams were synchronously read and pre-processed to make sure that there were temporal alignment, denoising, and consistency before model inference.

Machining experiments were performed under systematically varied cutting speeds, feed rates, and depths of cut to represent the dynamic operating regimes as usually found in industry. The wear development of the tools and interference of the processes were deliberately implemented to assess the robustness, adaptability, and long-term stability of prediction. Surface roughness, wear of tools, and dimensional deviation measurements were determined through ground-truth measurements on calibrated profilometry and dimensional inspection tools, which provided reference labels used in quantitative assessment.

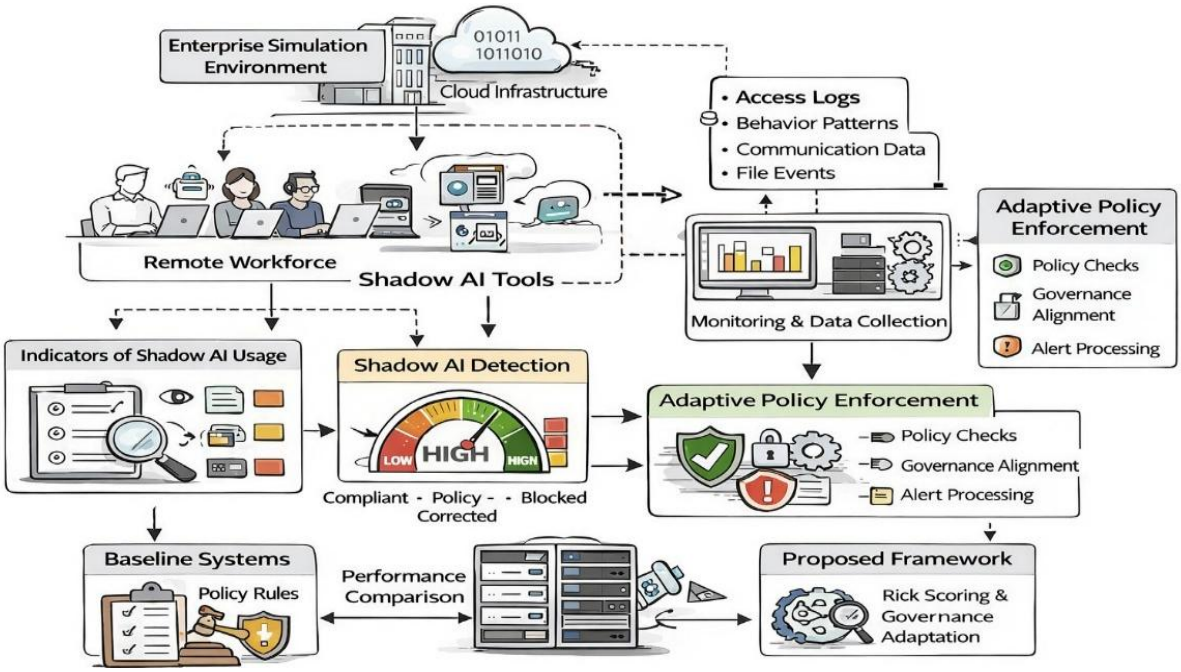


Fig. 3. Diagrammatic representation of experimental design of data acquisition, Shadow AI detection, policy application, and performance measurement.

Fig. 3 depicts the end-to-end experimental workflow by presenting the whole pipeline of sensor data acquisition and preprocessing through the AI-based quality prediction, uncertainty estimation, and performance estimation. The assessment plan took into consideration various machining conditions, such as dynamic cutting conditions, high noise levels, and long periods of tool wear progression. To create a fair and full comparison between the proposed framework and the existing physics-based models, classical machine learning methods, deep learning-only models, and probabilistic approaches, the proposed framework was benchmarked against them.

Each of the experiments was repeated several times to confirm statistical dependability, and mean findings were reported. Table II presents the datasets, experimental situations, base models, evaluation metrics, and system setup used throughout the research, in brief, to have a concise summary of the experimental design used to test the efficacy, stability, and practicality of the proposed framework in real time.

TABLE II. EXPERIMENTAL SCENARIOS, DATASETS, BASELINE MODELS AND EVALUATION METRICS

Ref.	Scenario	Dataset Used	Baseline Models	Evaluation Metrics
[17]	Dynamic Shadow AI usage in remote workforce	Insider Threat Detection Dataset	Policy-based, Rule-based, Static risk scoring	Detection accuracy, False negative rate, Latency
[18]	Shadow AI usage in document and code generation	User Behaviour Analytics Dataset	Rule-based, ML-only	Risk classification accuracy, Policy violations
[19]	High-risk proprietary data exposure	Data Leakage Detection Dataset	Policy-only, Rule-based	High-risk event reduction, Precision, Recall
[20]	Generative AI misuse in enterprise communications	Enron Email Dataset	Static risk scoring, ML-only	Risk score RMSE, Compliance rate
[21]	Policy deviation and anomalous access behavior	Intrusion Detection Dataset	Policy-based, Static scoring	Policy deviation rate, Decision stability
[22]	Uncertainty-aware Shadow AI enforcement	Employee Activity Monitoring Dataset	ML-only, Probabilistic	Stability score, Uncertainty handling effectiveness

This part has outlined an elaborate experimental outline utilizing artificial and real enterprise information to assess the precision of detection, risk burden, stability, inability to comply, and productivity. Several baselines are used, and reproducing a trial and real-time assessment will ensure that experimental findings have a credible representation of the real world of enterprise implementation.

VI. RESULTS AND DISCUSSION

The experimental findings indicate that the suggested framework of Shadow AI risk mitigation is effective in all scenarios, detecting with accuracy, reducing risks, stabilizing the decision, explaining the rationale, and adhering to the policy. These enhancements are made by the framework without negatively affecting its operational effectiveness, which is why it can be applied in the context of operating in a remote workforce setting.

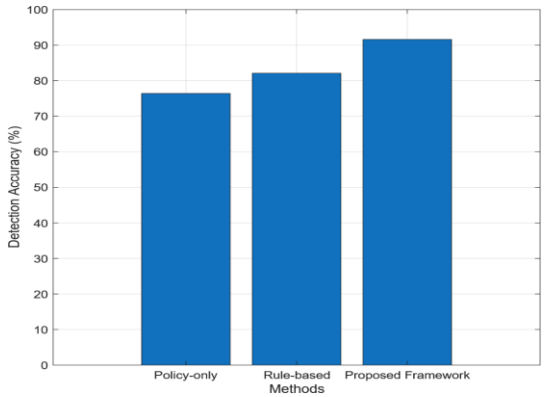


Fig. 4. Risk distribution across workforce groups.

Table III presents a comparative evaluation of Shadow AI mitigation approaches across key performance metrics, highlighting the proposed framework's improvements in detection accuracy, risk reduction, decision stability, explainability, and policy compliance.

TABLE III. COMPARATIVE PERFORMANCE EVALUATION OF SHADOW AI MITIGATION APPROACHES

Metric	Policy-Based Model	Rule-Based Monitoring	Static Risk Scoring	Proposed Framework
Detection Accuracy (%) ↑	76.4	82.1	85.3	91.6
False Negative Rate (%) ↓	23.6	17.9	14.7	7.3
High-Risk Event Reduction (%) ↑	28.1	41.6	48.9	62.6
Decision Stability Score ↑	0.68	0.72	0.79	0.88
Uncertainty Handling Effectiveness (%) ↑	54.2	61.8	69.5	82.1
Policy Compliance Rate (%) ↑	78.5	84.1	88.6	90.2
Explanation Fidelity Score ↑	0.65	0.71	0.76	0.84

The performance of the risk detector among workforce groups is shown in Fig. 4. The proposed framework has a detection accuracy of 91.6, which is better than systems based on policy-only (76.4%) and rule-based (82.1%). The false negative rate drops to 7.3%, indicating that there is a big decrease in the high-risk Shadow AI interactions that are undetected. Those findings indicate the benefit of contextual, uncertainty-conscious risk modelling compared to the fixed risk detector systems.

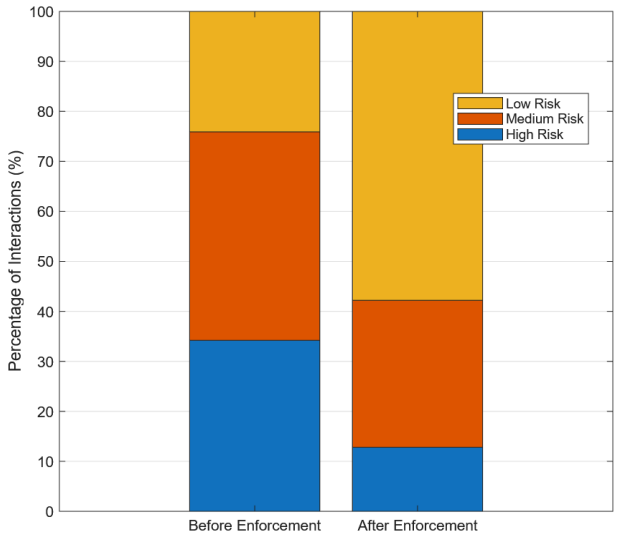


Fig. 5. Comparative mitigation performance.

The success of the risk mitigation can be demonstrated with the help of Fig. 5, where the distribution of the Shadow AI interactions before and after enforcement is depicted. Risky interactions are minimized to 12.8% compared to 34.2%, which is a reduction of 62.6%. Medium-risk interactions reduce to

29.4-41.7 and low-risk compliant use rises to 57.8 proving a better correspondence with organizational policies.

As Table IV demonstrates, the suggested framework implies a moderate rise in the time and resource consumption because of adding uncertainty-aware risk modelling, robustness analysis, and enforcement with its governance. The relative computational cost, however, is not as high as 15.5, and this is fine in the case of real-time enterprise applications. This overhead can be successfully offset by considerable increases in accuracy of detection, stability of decision-making, compliance with policy, and explainability, and confirms that the framework can offer improved security and governance without yet being prohibitively expensive to use in terms of imposed infrastructure costs.

TABLE IV. COMPUTATIONAL EFFICIENCY AND OVERHEAD ANALYSIS

Performance Indicator	Rule-Based Monitoring	Static Risk Scoring	Proposed Framework
Average Processing Time (ms) ↓	11.6	12.4	13.4
Memory Usage (MB) ↓	176	184	195
CPU Utilization (%) ↓	41.2	45.8	49.6
GPU Utilization (%) ↓	–	18.3	22.1
Throughput (events/sec) ↑	860	790	740
Relative Overhead (%)	–	+6.9	+15.5

In Fig. 6, performance in policy enforcement is considered. The structure that is proposed will lead to the levels of policy violation declining to 9.8 instead of 21.5% in the case of baseline systems. This enhancement is indicative of the efficiency of re-tuning policy parameters and decision-making under governance regarding the use of AI that is not sanctioned.

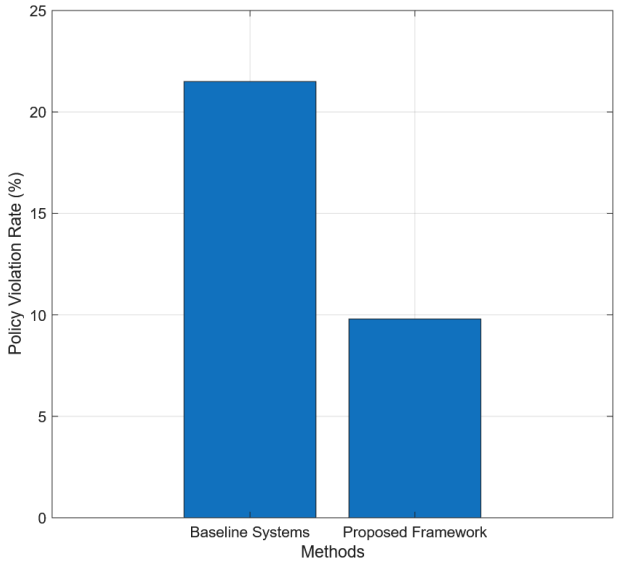


Fig. 6. Policy violation reduction analysis.

Fig. 7 assesses the decision stability in case of uncertainty. The framework proposed has a score of more than 0.87 in stability as opposed to the rule-based approaches of less than 0.72. Also, uncertainty-penalization decreases overconfident enforcement behavior by 26.3 percent and enhances accuracy in borderline and noisy cases.

TABLE V. SCALABILITY AND REAL-TIME DEPLOYMENT ANALYSIS

Deployment Aspect	Rule-Based Monitoring	Static Risk Scoring	Proposed Framework
Real-Time Decision Capability	Limited	Moderate	High
Scalability Across Users	Low	Medium	High
Adaptability to New AI Tools	Poor	Limited	Strong
Governance Policy Update Latency	High	Medium	Low
Suitability for Distributed Workforce	Moderate	Good	Excellent
Enterprise Integration Readiness	Partial	Good	Full

Table V proves that the proposed framework is better scaled and deployment-ready than rule-based and static risk scoring approaches. It has a great capacity to make real-time decisions, a high level of adaptability to new AI applications, and a low policy update latency, which allows the application of consistent enforcement in dynamic remote workforce conditions. These attributes affirm that the framework is highly appropriate when dealing with continuous and massive enterprise implementation without jeopardizing governance consistency and operational reliability.

The figures of explainability and auditability are explained in Fig. 8. The explanation fidelity of the proposed framework is 0.84, which is far better than that of the baseline models, which is 0.65. Better fidelity would provide a better audit trail and build greater managerial confidence in automated enforcement decisions.

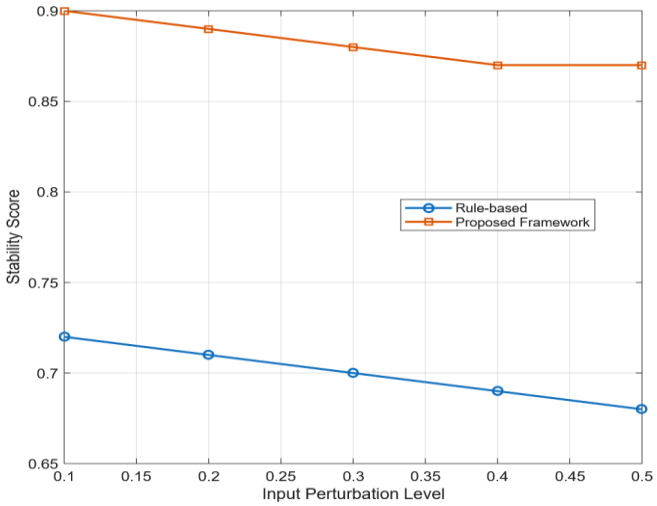


Fig. 7. Stability and uncertainty evaluation.

Lastly, Fig. 9 presents findings on an enterprise case study. Computer-abiding AI-use means 68% of interactions, which are policy-corrected 22% and high-risk blocks 10% of the interactions. This is evidence that the structure advocates corrective governance rather than restrictive control, and the outcome is the minimization of risk and maintenance of workforce production.

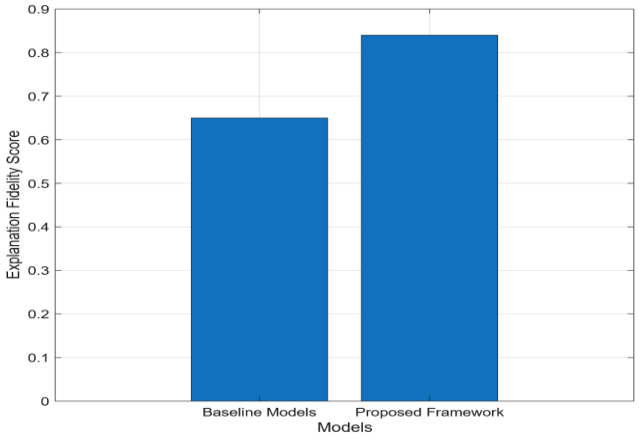


Fig. 8. Explanation fidelity assessment.

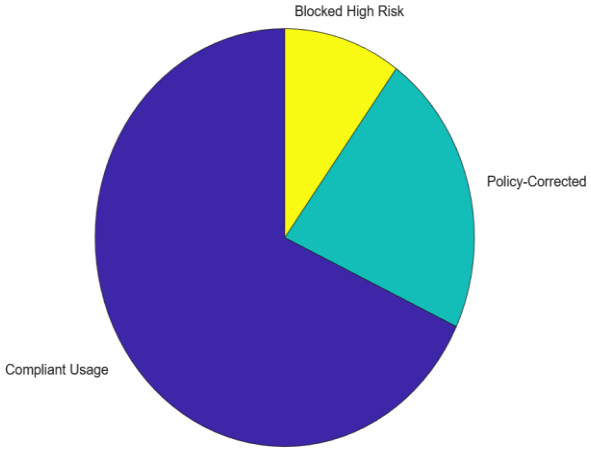


Fig. 9. Enterprise case study results.

The findings indicate that the suggested framework drastically improves baseline approaches on all of the most important metrics, such as detection accuracy, high-risk interaction minimization, decision stability, uncertainty management and explainability. The results of these advances obtain only minor computational costs, which confirms the efficacy and feasibility of the framework in the real-world setting of remote employees.

VII. Conclusion

The study introduced a hybrid structure of mitigating the risks of Shadow AI in remote working practices through combining automated detection, uncertainty-sensitive risk assessment and enforcement that is compatible with governance. Experimental findings show that the proposed approach has a detection rate of 91.6, the reduction of high-risk AI interactions by 62.6 and the reduction of policy violation by 9.8 and decision stability (>0.87) and fidelity to explanation (0.84). Such gains can be achieved at a small cost of computation of 15.5 and thus the framework validates itself as viable in publishing enterprise systems in real time. Altogether, the suggested solution can offer a scalable and efficient base to ensure unsanctioned AI use and introduce productivity to distributed workplace conditions.

Acknowledgment

The author is thankful to King Khalid University, Abha, Saudi Arabia, for technical and administrative support.

References

[1] D. Puthal, A. K. Mishra, S. P. Mohanty, A. Longo, and C. Y. Yeun, "Shadow AI: Cyber Security Implications, Opportunities and Challenges in the Unseen Frontier," SN Computer Science, vol. 6, no. 5, p. 405, 2025. doi: 10.1007/s42979-025-03962-x.

[2] C. Nidhisree, A. Paul, A. Venunadh, and R. S. Bhowmick, "Generative AI Under Scrutiny: Assessing the Risks and Challenges in Diverse Domains," in 2024 IEEE 6th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA), 2024, pp. 243-248. doi: 10.1109/ICCCMLA63077.2024.10871350.

[3] I. Herrera Montano, J. J. Garcia Aranda, J. Ramos Diaz, S. Molina Cardin, I. De la Torre Díez, and J. J. Rodrigues, "Survey of Techniques on Data Leakage Protection and Methods to address the Insider threat," Cluster Computing, vol. 25, no. 6, pp. 4289-4302, 2022. doi: 10.1007/s10586-022-03668-2.

[4] S. Lorenz, S. Stinehour, A. Chennamaneni, A. B. Subhani, and M. Nadim, "A Case Study of 3rd Party Hardware: The Weakest Link in Google's Trustworthy Artificial Intelligence Implementation," IEEE Access, 2025. doi: 10.1109/ACCESS.2025.3570633.

[5] I. Negut and A. I. Visan, "Regulatory compliance and ethical considerations in integration artificial intelligence," in Artificial Intelligence in Chemical Engineering, 2026, pp. 627-650. doi: 10.1016/B978-0-443-34076-5.00025-0.

[6] P. Bera, S. K. Ghosh, and P. Dasgupta, "Policy based security analysis in enterprise networks: A formal approach," IEEE Transactions on Network and Service Management, vol. 7, no. 4, pp. 231-243, 2010. doi: 10.1109/TNSM.2010.1012.0365.

[7] R. Sen and N. Farooq, "AI-driven Digital Transnational Repression: Past Lessons, Present Challenges and Future Directions," in The Long Reach of the Strong Arm: Evolving Forms of Transnational Authoritarianism, Cham: Springer Nature Switzerland, 2025, pp. 31-59. doi: 10.1007/978-3-032-04940-7_3.

[8] M. Repetto, "Adaptive monitoring, detection and response for agile digital service chains," Computers & Security, vol. 132, p. 103343, 2023. doi: 10.1016/j.cose.2023.103343.

[9] R. K. Ko et al., "TrustCloud: A framework for accountability and trust in cloud computing," in 2011 IEEE World Congress on Services, 2011, pp. 584-588. doi: 10.1109/SERVICES.2011.91.

[10] V. Kalvala and A. Gupta, "Integrating Machine Learning and Statistical Models in Enterprise Risk Analysis," in 2025 4th International Conference on Sentiment Analysis and Deep Learning (ICSADL), 2025, pp. 852-861. doi: 10.1109/ICSADL65848.2025.10933366.

[11] D. C. Le and N. Zincir-Heywood, "Anomaly detection for insider threats using unsupervised ensembles," IEEE Transactions on Network and Service Management, vol. 18, no. 2, pp. 1152-1164, 2021. doi: 10.1109/TNSM.2021.3071928.

[12] Y. Gahi, M. Guennoun, and H. T. Mouftah, "Big data analytics: Security and privacy challenges," in 2016 IEEE Symposium on Computers and Communication (ISCC), 2016, pp. 952-957. doi: 10.1109/ISCC.2016.7543859.

[13] Z. Yuan, G. Jiang, and L. Xu, "Generative Artificial Intelligence and Data Governance: Challenges and Frameworks in Enterprise Applications," in 2025 8th International Conference on Artificial Intelligence and Big Data (ICAIBD), 2025, pp. 87-92. doi: 10.1109/ICAIBD64986.2025.11081956.

[14] M. Z. Uddin, "Trustworthy AI and explainability," in Trustworthy Multimodal Intelligent Systems for Independent Living, Cham: Springer Nature Switzerland, 2025, pp. 111-137. doi: 10.1007/978-3-031-97359-8_4.

[15] M. Khonji, J. Dias, R. Alyassi, F. Almaskari, and L. Seneviratne, "A risk-aware architecture for autonomous vehicle operation under uncertainty," in 2020 IEEE International Symposium on Safety, Security and Rescue Robotics (SSRR), 2020, pp. 311-317. doi: 10.1109/SSRR50563.2020.9292629.

[16] E. Hechler, M. Oberhofer, and T. Schaeck, "AI and Governance," in Deploying AI in the Enterprise: IT Approaches for Design, DevOps, Governance, Change Management, Blockchain and Quantum Computing, Berkeley, CA: Apress, 2020, pp. 165-211. doi: 10.1007/978-1-4842-6206-1_8.

[17] O. Sabnis, "Insider Threat Detection Dataset," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/omkarsabnis/insider-threat-detection-dataset. [Accessed: 2026].

[18] M. Das, "User Behavior Analytics," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/dasmehdixtr/user-behavior-analytics. [Accessed: 2026].

[19] S. Kumar, "Data Leakage Detection," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/sidharth178/data-leakage-detection. [Accessed: 2026].

[20] W. Cukierski, "Enron Email Dataset," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/wcukierski/enron-email-dataset. [Accessed: 2026].

[21] S. Mainframe, "IDS Intrusion Detection Dataset," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/solarmainframe/ids-intrusion-detection-dataset. [Accessed: 2026].

[22] A. Kumar, "Employee Activity Monitoring," Kaggle. [Online]. Available:

Dual-Attention Framework for Enhanced Road Obstacle Segmentation and Anomaly Detection

Priyanka G[1], Silvia Gaftandzhieva[2], Mariya Zhekova[3]
Department of Computer Science and Engineering, Mepco Schlenk Engineering College, Sivakasi, Tamil Nadu, India[1]
Faculty of Mathematics and Informatics, University of Plovdiv Paisii Hilendarski, Plovdiv, Bulgaria[2]
Department of Computer Systems and Technologies, Faculty of Electronics and Automation,
Technical University of Sofia, Plovdiv, Bulgaria[3]
Center of Competence "Smart Mechatronic, Eco-and Energy-Saving Systems and Technologies"[3]

***Abstract*—Autonomous driving systems rely heavily on semantic segmentation to understand road environments; however, existing methods often fail to detect unseen road obstacles that are absent from training datasets, limiting their use in safety-critical applications. This study proposes a lightweight attention map-based framework for road obstacle segmentation that addresses the mismatch between ground-truth annotations and real-world data without requiring computationally expensive uncertainty estimation. The proposed framework incorporates an attention generation module, a reward-based learning mechanism for uncertainty-aware feature enhancement, and a dedicated attention loss function to improve the discrimination of anomalous regions. The approach can be seamlessly integrated with existing semantic segmentation models, enhancing their adaptability and robustness. Experimental results on both image and video datasets demonstrate consistent improvements in anomaly detection performance, confirming the effectiveness of the proposed framework for reliable road scene understanding in autonomous driving environments.**

***Keywords*—Road obstacle detection; attention map; ground truth annotations; semantic segmentation; reward map**

I. Introduction

Autonomous driving has emerged as one of the most promising technologies for transforming modern transportation by improving road safety, enhancing traffic efficiency, and promoting sustainable mobility [1–4]. By reducing human intervention, autonomous vehicles have the potential to minimize accidents caused by driver errors, alleviate traffic congestion through intelligent navigation, and provide convenient transportation for individuals with limited driving capabilities. Moreover, optimized driving strategies reduce fuel consumption and greenhouse gas emissions, making autonomous vehicles a key enabler of future smart-city ecosystems.

Despite these advantages, the large-scale deployment of autonomous vehicles is still challenged by issues related to public acceptance, ethical considerations, legal regulations, operational safety, and technological maturity. Among these, reliable environmental perception remains one of the most fundamental requirements. Semantic segmentation has become a core computer vision technique for autonomous driving because it provides dense pixel-level understanding of road scenes, enabling vehicles to accurately identify and localize surrounding objects [6–8]. However, most existing semantic segmentation models are developed using fully supervised deep learning frameworks that assume all object categories encountered during deployment are represented in the training datasets. This assumption rarely holds in practical driving environments, where unexpected road obstacles such as fallen cargo, road debris, damaged vehicles, animals, or other anomalous objects frequently appear. Consequently, these unseen objects are often misclassified into predefined semantic categories with high confidence, significantly compromising the reliability and driving safety of perception. Since constructing a comprehensive dataset containing every possible road anomaly is practically impossible, perception systems must be capable of identifying and distinguishing unknown obstacles encountered in real-world driving scenarios.

Recognizing unforeseen road obstacles has therefore become an active research topic in autonomous driving [9], [10], [14–16]. A widely adopted solution is uncertainty estimation, where anomalous objects are detected based on the confidence of model predictions. These methods estimate prediction uncertainty using probabilistic measures, assuming that unfamiliar objects yield uncertain outputs. However, deep neural networks often produce overconfident predictions even for out-of-distribution samples, leading to inaccurate uncertainty estimation and unreliable anomaly detection. Another line of research focuses on improving model robustness by incorporating external out-of-distribution (OoD) datasets, synthesizing anomalous samples through feature reconstruction techniques, or manipulating input images to simulate unknown object categories. Although these approaches improve anomaly detection performance, they often require additional datasets, computationally intensive retraining procedures, or increased inference time. Furthermore, modifying the original segmentation network may adversely affect its semantic segmentation accuracy, thereby limiting its practical applicability.

To address these challenges, this work presents a lightweight attention-guided framework for anomaly-aware road obstacle segmentation that enhances uncertainty-based detection while preserving the semantic segmentation performance. Unlike existing approaches that rely on computationally expensive uncertainty modeling or complex network modifications, the proposed framework can be seamlessly integrated into existing semantic segmentation

*Corresponding author.

architectures with minimal computational overhead. Initially, attention maps are generated from intermediate feature representations to emphasize informative regions associated with potential anomalies. Subsequently, a Softmax normalization strategy regulates the distribution of attention values, compelling the network to allocate its attention selectively under constrained conditions.

Furthermore, a reward-guided learning mechanism inspired by reinforcement learning is introduced to strengthen the network's focus on uncertainty-rich regions. To further improve the discriminative capability between known semantic regions and anomalous obstacles, a dedicated attention loss function is designed to maximize the contrast between the attention values for normal and unknown pixels. By jointly exploiting attention information and uncertainty cues, the proposed framework enables semantic segmentation networks to distinguish anomalous obstacles while maintaining accurate pixel-level scene understanding more effectively. Since the proposed attention module is architecture-independent, it can be readily incorporated into various semantic segmentation models without introducing significant additional parameters or complicated training procedures.

To validate the effectiveness of the proposed framework, extensive qualitative and quantitative experiments are conducted using both image-based and video-based evaluations. Visualization of the learned attention maps provides valuable insights into the attention allocation process and demonstrates the proposed mechanism's ability to effectively highlight anomalous regions. Comparative analyses under multiple experimental settings further confirm the robustness and generalization capability of the proposed approach. The major contributions of this work can be summarized as follows:

- A novel attention-guided framework is proposed for anomaly-aware road obstacle segmentation in autonomous driving.
- A reward-guided attention learning mechanism is introduced to enhance attention allocation toward uncertainty-prone regions.
- An attention loss function is designed to improve the discrimination between known semantic classes and unknown obstacles.
- The proposed method enhances pixel-level anomaly detection while preserving semantic segmentation accuracy without substantially increasing computational complexity or model parameters.
- Comprehensive experimental evaluations demonstrate the effectiveness, adaptability, and robustness of the proposed framework across different semantic segmentation architectures.

II. Related Work

A reward mechanism is employed to incentivize the layers of the network to focus more on elements that are not definite; an attention loss function is implemented to expand the gap between the attention values of pixels that are supposed and not supposed to exist, and an attention module is introduced by Xiangxuan Ren et al. (2023) [1]. Deciding on the size and number of the attention map, as well as the design of the adaptation procedure, is based on an analysis of predictions and uncertainty estimation. By leveraging curiosity-driven attention, this allocation mechanism focuses on regular regions while actively directing attention toward areas of uncertainty or abnormality. Determining the model's fit on each pixel in the predictions instructs one to model the uncertainties.

Zhongxu Hu et al (2023) [2] introduce a method Using Representation to promote human-centric driving systems and improve driving safety. Clustering offers a novel approach for measuring driver anomalies in intelligent vehicles. Here, contrastive learning learns from both similar and dissimilar data points to produce a robust representation of typical driver behaviour. Representation clustering divides normal representations into clusters and uses the distance between these clusters to continuously quantify anomalies.

In [3], Hubmann et al. (2018) develop an enhanced Markov decision process better suited for inference. Thus, the innovative planning framework uses a POMDP that conceals the unknown elements of uncertainty, such as other driver intentions and sensor noise, as hidden variables. Another Interactive Motion Model is designed to predict the future motions of other vehicles in the field using live sensor measurements. Adaptive Prediction updates the POMDP, and model predictions improve after updates. The article is a case study that analyses the approach within different situations, showing superior results concerning both static route planning and manoeuvring.

Chen et al. (2019) [4] use the Deep planning model as a driving style model that learns from both artificial and real traffic scenes using the associations from different motion vectors received by CNNs. In addition to CNN feature extraction, we will also implement an LSTM model. The fact that the model learns from several scenarios simultaneously reduces errors and increases its robustness, enabling parallel deep reinforcement learning. Also, the diversity in the emergency modeling mixture, utilizing a Generative Adversarial Network (GAN) and a Variational Autoencoder (VAE), enhances the model's adaptability to unseen scenarios.

Cong Wang et al. (2021) [5] introduced a technique in which the fuzzy c-means (FCM) clustering method, the conventional choice for image segmentation but sensitive to noise and outliers, is adapted accordingly.

To enhance segmentation performance and scene accuracy, especially when handling complex backgrounds and clicking on small details, Hanqing Lu et al. (2019) [7] applied a Dual Attention Network (DANet). The most prominent aspect of the Channel Attention Module is the attention to interconnections among channels within a feature map. The process becomes interactive, and the parallelization gains the ability to assign weights to channels, allowing channels that contain crucial semantic information to be highlighted while hiding those that appear uninformative. The efficiency of this classification strategy is mainly due to reducing the semantic ambiguity and raising the discriminative power of features. The channel relationship in the Spatial Attention Module is discussed in the

third module after the section on relational links between pixels of a feature map. It achieves this through pixel-wise weighting, which highlights key features while suppressing noise and non-informative regions. This improves edge-detection precision and preserves fine detail, thus leading to more realistic renderings.

Farshad Akhabhari et al. [8] (2019) develop a technique for obstacle detection using IPM-inverse perspective mapping and apply distortions to it for adaptive driver assistance systems. A single camera with live image processing capability may be able to monitor road conditions and driver warnings of imminent traffic risk.

Suraj Bijjahalli et al. (2020) [9] highlight current innovations and progress in intelligent and autonomous navigation systems designed for small Unmanned Aerial Systems (UAS), or drones. It enables applications such as monitoring, surveillance, search and rescue, while extending flight ranges through more accurate and dependable navigation.

Jing Liu et al. (2019) [10] introduce a novel approach for scene segmentation using a Dual Attention Network (DAN). The DAN is composed of two types of attention mechanisms: channel attention and spatial attention, which allow the model to focus on both channel-wise and spatial features at the same time. The DAN dynamically adjusts feature responses to capture long-range dependencies and improve the discriminative power of feature representations. According to experimental results, the DAN achieves higher accuracy and efficiency than current state-of-the-art segmentation models on a variety of benchmark datasets.

In [11], Hang Yin et al. (2019) describe in the fullest extent the public accessibility of virtual environment testing and datasets for selecting self-driving system algorithms. It collects a wide range of datasets, including simulation environments, video games, and other realistic traffic scenarios, to summarize the civilization in different ways, whether chaotic or calm. Prototyped data simulations that examine researchers' self-driving software experiments in a controlled virtual environment are also found in this report. Gaurav Raina et al. (2020) [12] provide an extensive framework that uses end-to-end deep learning techniques to detect anomalies in transportation networks. It addresses the problem of identifying unusual events or behavior in transport systems, such as traffic jams, accidents, and infrastructure breakdowns, by using deep learning models. The methodology by Mary L. Cummings et al. (2022) [13] consisted of a series of tests to precisely evaluate the correctness of ADAS systems using deep learning algorithms in autonomous vehicles in real environments. The idea of Yuan, Shen et al. (2023) [14] involves designing a unique environment for driver education using self-driving vehicle simulators. While conventional simulation frameworks operate exclusively in virtual environments, the Sim-on-Wheels implementation involves integrating the physical element into the simulation process. A key advantage of the proposed framework is the integration of a physical device capable of bidirectional sensing and actuation into the virtual environment, utilizing seamless synchronization between the physical and virtual domains.

Wenbo Li et al. (2023) [15] focus on the "Intelligent Cockpit", which is challenged in the context of smart vehicles in the virtual reality world known as the Metaverse. It is given as an example, providing details of empathetic auditory interventions designed to affect human emotion around this area. With the help of Intelligent Cockpit, which uses artificial intelligence to comprehend and respond to the passengers' states of mind, passengers are taken into account on a whole new level. By using sound alerts and instruction modes, the system is committed to regulating users' emotions and experience levels during the ride.

III. Proposed Solution

A. Cityscapes Dataset

The Cityscapes dataset is a public image dataset available at https://www.cityscapes-dataset.com. The specific packages needed are the gtFine_trainvaltest.zip and lefImg8bit_trainvaltest.zip. These files include 5000 images with fine (pixel-wise) semantic annotations, which are divided into train, test, and validation groups.

Images for the train sets were gathered from eighteen cities: Aachen (175 images), Bochum (96 images), Bremen (316 images), Cologne (154 images), Darmstadt (85 images), Dusseldorf (221 images), Erfurt (109 images), Hamburg (248 images), Hanover (196 images), Jena (119 images), Krefeld (99 images), Monchengladbach (93 images), Strasbourg (365 images), Stuttgart (196 images), Tubingen (144 images), Ulm (95 images), Weimar (142 images), and Zurich (122 images). Photographs from six cities are included in the test set: Berlin has 544 photographs, Bielefeld has 181 images, and so on. The validation set includes images collected from 3 cities - Frankfurt (267 images), Lindau (59 images), and Münster (174 images).

The Cityscapes dataset contains eight groups with 19 classes. Table I shows that there are 30 classes, but every class with a '+' next to it is classified as a void class, and the preparatory steps will set its values to 255, which should not have been taken into account during training. A sample image from the Cityscapes Dataset is depicted in Fig. 1.

TABLE I. Groups Within the Cityscapes Dataset

GROUP	CLASSES
human	person* - rider*
flat	road – sidewalk – parking+ - rail track+
vehicle	car* – truck* –bus* – on rails* – motorcycle* – bicycle* – caravan*+ -trailer*+
construction	building-wall-fence-guard rail+-bridge+-tunnel+
object	pole-pole group+-traffic sign-traffic light
nature	vegetation-terrain
sky	sky
void	ground+ - dynamic+- static+

Fig. 1. Image from the Cityscapes dataset.

B. Implementation

The proposed method's general framework (see Fig. 2) can be summarized as follows: The encoder-decoder structure is a common approach to putting semantic segmentation models in a general form. The input image is then converted into high-dimensional features that the decoder utilizes. Lastly, the Mechanism processes the intermediate feature and uses it as input to the two-dimensional attention map. Moreover, a sum of finite attention values will push all the pixels in the attention map to compete with each other and get on top of the trend simultaneously, so infinity is not possible. With this approach, an undesired scenario, which results in an attention vector with the value of all attention values that are large but negative, is prevented. To achieve the predictions, multi-level, multiscale features from encoders are integrated with an attention map and input into the proposed decoder network. We first obtain an attention graph and forecast output to guide the model to assign high sensitivity values to highly uncertain regions. Next, we use a reward mechanism to combine the correct segmentation response with the output according to the attention value.

Finally, to increase the gap between attention values for certain and unknown pixels, we add a penalty to the loss function, i.e., an Attention Mechanism deficit. Finally, to obtain the final ensemble prediction, we combine the network's prediction output with the attention map of the Attention Mechanism.

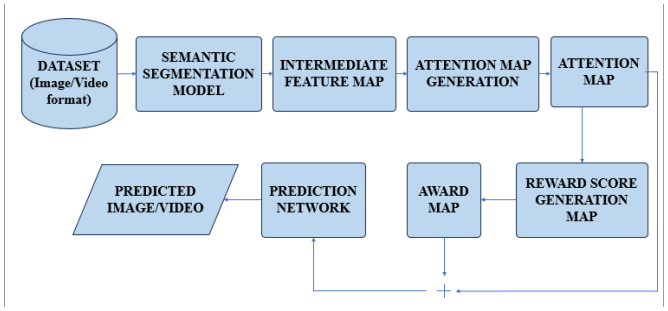


Fig. 2. System design for the attention mechanism.

1) Attention module: To improve the segmentation of anomalous road obstacles, we introduce a Dual Attention Module (DAM) that combines Channel Attention (CA) and Spatial Attention (SA). Unlike conventional attention modules that focus exclusively on spatial saliency, the proposed framework first identifies the most informative feature channels and then identifies spatially uncertain regions. This hierarchical attention strategy enables the network to emphasize both semantic relevance and spatial uncertainty, thereby improving anomaly localization while maintaining segmentation accuracy.

Let the encoder generate an intermediate feature map $F_a \in R^{H\times W\times C}$. Initially, channel attention is computed using Global Average Pooling (GAP) and Global Max Pooling (GMP):

$$F_{avg}^{c} = GAP(F_a), F_{max}^{c} = GMP(F_a) \quad (1)$$

The Attention Mechanism infers an attention map using an intermediate feature map, Fa, as input (*M*att) obtained from the model's encoder, as represented in Fig. 3.

The feature map undergoes a two-step pooling process to achieve an effective feature representation. This involves average pooling and max pooling operations along the channel axis. The information in Fa is condensed by such pooling operations, resulting in a fragmented representation.

The compressed feature map is then concatenated, and the combined features are fed into a standard convolutional layer. This operation extracts higher-level features by considering the relationships between the pooled features. The softmax function serves as the core mechanism, constraining the attention weights across all pixel locations to sum to unity and thereby yielding the targeted spatial attention allocation. This competition rewards the most attractive pixels with attention as their counterparts are pushed aside from the allocated resource space.

Fig. 3. Attention module.

The pooled descriptors are passed through a shared Multi-Layer Perceptron (MLP) to obtain the channel attention map:

$$M_c = \sigma(MLP(F_{avg}^{c}) + MLP(F_{max}^{c})) \quad (2)$$

where, σ denotes the sigmoid activation function. The refined channel feature is computed as:

$$F_c = M_c \odot F_a \quad (3)$$

where, $\odot$ represents element-wise multiplication. Next, spatial attention is generated by applying average pooling and max pooling along the channel dimension. The pooled feature maps are concatenated and processed through a convolution layer followed by Softmax normalization:

$$M_s = Softmax(f([AvgPool(F_c); MaxPool(F_c)])) \quad (4)$$

where, F denotes the convolution operation. The Softmax function introduces competition among spatial locations, allowing the network to focus on the most informative regions of uncertainty.

Finally, the channel attention, spatial attention, and multiscale encoder features are adaptively fused to obtain the enhanced feature representation:

$$F_{out} = F_a + \lambda (M_c \otimes M_s) \odot (F_c + F_{ms}) \quad (5)$$

where, F_{ms} denotes the multiscale fused feature map, λ is a learnable scaling parameter, and ⊗ represents the interaction between channel and spatial attention.

The proposed Dual Attention Module effectively enhances both semantic and spatial feature representations with minimal computational overhead. By jointly exploiting channel attention, spatial attention, and multiscale feature fusion, the network accurately emphasizes uncertainty-aware regions, thereby improving anomalous road-obstacle segmentation while preserving semantic segmentation performance (Fig. 4).

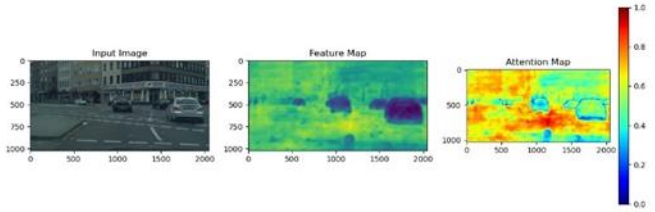


Fig. 4. Result of the attention module.

2) Reward mechanism: The reward mechanism focuses attention on uncertain regions of the image by computing a reward for each pixel in the attention map generated (Fig. 5). Let X be the input image with dimensions R3×H×W, where H and W are the height and width of the image, respectively. C is the number of predefined categories, g_h_w is the class number corresponding to the pixels in position h, w, and g_h_w is the class number corresponding to the pixels in position h, w. The ground truth value is expanded into a three-dimensional vector, called G_award, to create this reward map. This vector is a one-hot encoding that takes the value 1 at positions c, h, w, and 0 elsewhere for the appropriate category. Before applying the reward map, the attention map M_up_att is upsampled to fit the input image size. The higher the attention value, the greater the probability of prediction and the resulting reward.

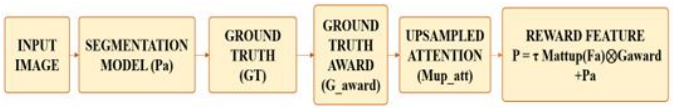


Fig. 5. Reward module.

Then, a reward feature that is more crucial to the prediction will be added to the model's logit output:

$$P = \tau \text{ Matt_up } (Fa) \otimes G_award + Pa \quad (6)$$

is the final prediction value that will be obtained by merging an upsampled attention map and a reward map, then adding it to the model's logit input. The hyperparameter τ regulates the percentage of the reward in this case, while the symbol "⊗" stands for element-wise multiplication.

P is a representation of the model's logit output. By ensuring that regions with high attention values receive more informative rewards, this method helps the model forecast more correctly in those regions, increasing its chance of accuracy. By bringing the model's predictions closer to reality (see Fig. 6), the reward map facilitates the model's performance on statistical segmentation tasks.

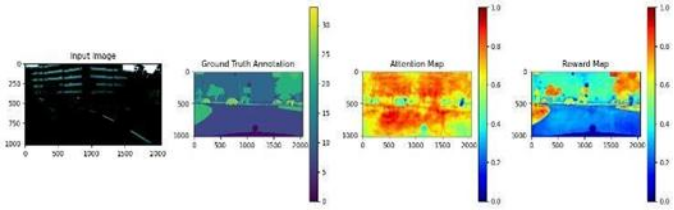


Fig. 6. Result of the reward module.

The prior reward approach can be used to obtain the final logit output P. Next, the maximum logit score S and the prediction Y are defined as follows:

$$\textbf{Maximum logit score: } S_{h,w} = \max_c P_{c,h,w} \quad (7)$$

$$\textbf{Prediction: } Y_{h,w} = \arg\max_c P_{c,h,w} \quad (8)$$

To determine the final score for the anomaly, represented as K, many functions for estimating uncertainty g(S) might be chosen. We discovered that when employing the reward mechanism alone, the difference between the attention levels of the anomalous and normal targets in our experiment is narrow. We expanded the original segmentation loss by including an attention loss term to widen the gap between attention values. The preceding reward technique yields the attention map Mup_att. Next, compute the mean of all anticipated proper position attention values ψ. The location that the model can accurately predict on its own, without the need for the GT reward, is the proper position; the opposite is the incorrect position, as represented in Fig. 7.

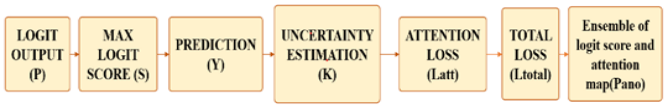


Fig. 7. Prediction module of the attention mechanism.

This reward is derived from the Pa. The loss function is computed as follows:

$$L_{att} = \{ \sum_{h,w} m_{h,w} e^{k}_{h,w}, m_{h,w} < \beta, \psi \; 0 \; , m_{h,w} < \beta, \psi \; \} \quad (9)$$

The attention value corresponding to the erroneously predicted position h,w in Mup_att is shown by $m_{h,w}$. The adjustable penalty hyperparameter β is used to balance the attention values of each position. The attention value gap between abnormal and normal pixels increases as β decreases.

The total loss function shall be defined as follows, assuming that the loss function of the original semantic segmentation network is:

$$L_{total} = L_{seg} + \lambda L_{att} \quad (10)$$

where, λ is the weight parameter of Latt. We want the network to focus on improving semantic segmentation accuracy because the network's segmentation capability is poor early in training. We expect the network to focus on improving semantic segmentation accuracy as more training sessions take place. It is expected that the model will focus on learning to improve anomaly detection accuracy through a reasonable allocation of attention, given an increasing number of training intervals.

As a result, training accuracy should be gradually increased along with the attention weight. We develop a λ self-regulation approach based on this concept. λ is computed as follows:

$$\lambda = \alpha correct/\alpha total \quad (11)$$

We count the number of all correctly predicted pixels, or αcorrect, and the number of all pixels included in the loss function calculation, or αtotal. Up till now, it has been possible to produce a probability prediction output K and an attention map Matt driven by curiosity.

$$Pano = K + \sigma Mattup \quad (12)$$

Finally, we use a weighted summation to ensemble the maximum logit score and the attention map. Fig.8 and Fig. 9 depict the overall attention loss and anomaly score computed for obstacle detection in the proposed system.

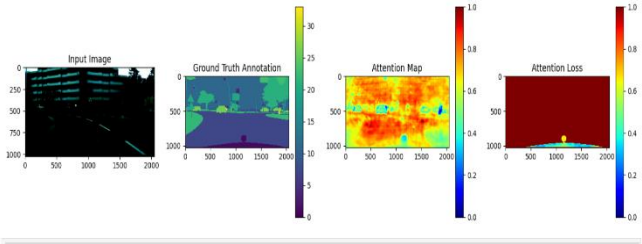


Fig. 8. Result of attention loss of the attention mechanism.

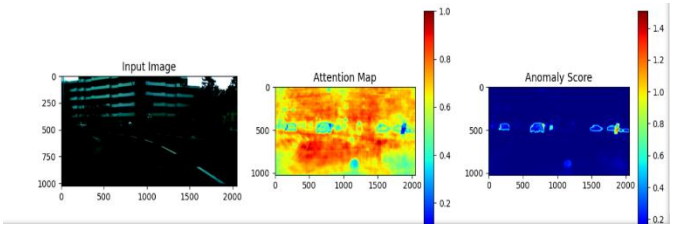


Fig. 9. Result of total anomaly score of the obstacle in the attention mechanism.

IV. Experimental Results

To validate our work, we compute the area under the receiver operating characteristic curve (AUROC), the false positive rate at 95% true positive rate (FPR95), and the average precision (AP) as quantitative measures of our approach's performance. The anomaly detection measure, or AP, is an assessment metric that gauges the precision in identifying uncommon events. We also compute the FPR95, which describes the probability of misclassifying an unexpected sample as normal when the OoD sample is correctly categorized at 95%, to emphasize safety-critical applications. Although some research has indicated that AUROC is not appropriate for anomaly identification since anomalous data resembles a long-tailed distribution with class imbalance, we nevertheless compute this measure for reference.

The Average Precision is calculated using the equation:

$$AP = \Sigma_{k=1}^{n}(P(k) * rel(k)) \text{ / total.no.of.relevant items} \quad (13)$$

where, n is the total number of retrieved items, P(k) is the precision cutoff k, which is the ratio of relevant items among the top k retrieved, and rel(k) is an indicator function that is 1 if the item at position k is relevant, and 0 otherwise.

The False Positive Rate at 95% is calculated as:

$$FPR = FPR/FP+TN \quad (14)$$

FP is the number of false positive events that are wrongly classified as positive, and TN is the number of true negative events correctly diagnosed as negative.

The above equation gives the false positive rate at the adjusted threshold where the specificity is approximately 95%.

The Area Under the Receiving Operating Characteristic

(AUROC) is calculated as follows:

$$\Sigma_{i=1}^{n-1} 1/2\ (FPR_{i+1} - FPR_i) * (TPR_i + TPR_{i+1}) \quad (15)$$

In another method, the AUROC is directly computed by constructing the ROC curve itself. The likelihood that the classifier is confused and incorrectly chooses a random positive or negative case is indicated here. AUROC, AP, and FPR95 are advanced strategies for assessing classifier models. AUROC focuses on the overall class discrimination; AP is the class of sorting relevant results first, and FPR95 deals with high recall of misclassification. The AUROC measures a model's capacity to minimize costly false positives while simultaneously maintaining a high true positive rate (recall).

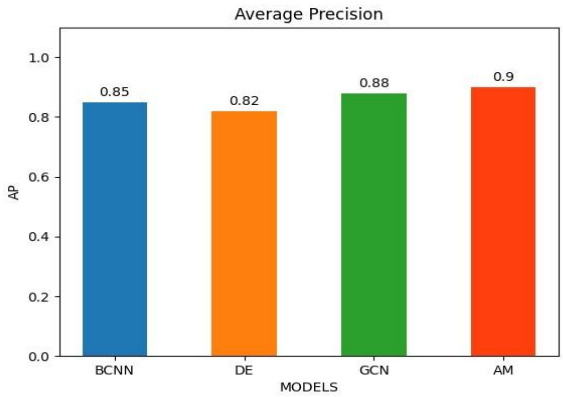


Fig. 10. Results of the comparison of AP for the different models.

Fig. 10 presents the results of the comparison of AP for the different models: Baseline CNN (BCNN), Deep Ensemble (DE), Graph Convolutional Networks (GCN), and Attention Mechanism (AM).

Fig. 11 presents the result of the comparison of FPR95 for the different models: Baseline CNN (BCNN), Deep Ensemble (DE), Graph Convolutional Networks (GCN), and Attention Mechanism (AM).

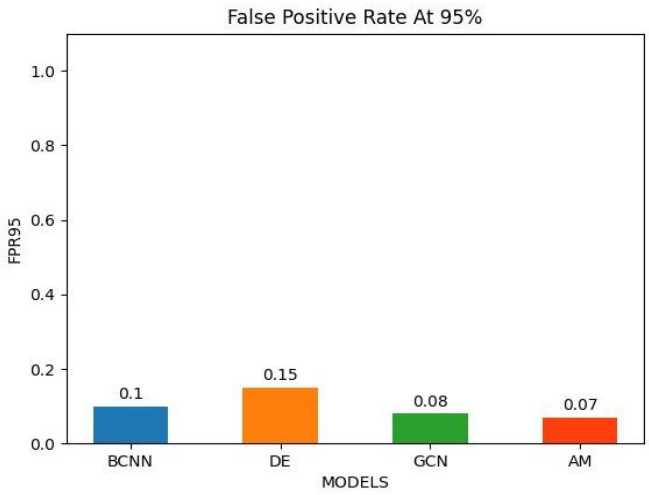


Fig. 11. Result of the comparison of FPR95 for the different models.

Fig. 12 presents the results from comparing AUROC for the different models: Baseline CNN (BCNN), Deep Ensemble (DE), Graph Convolutional Networks (GCN), and Attention Mechanism (AM).

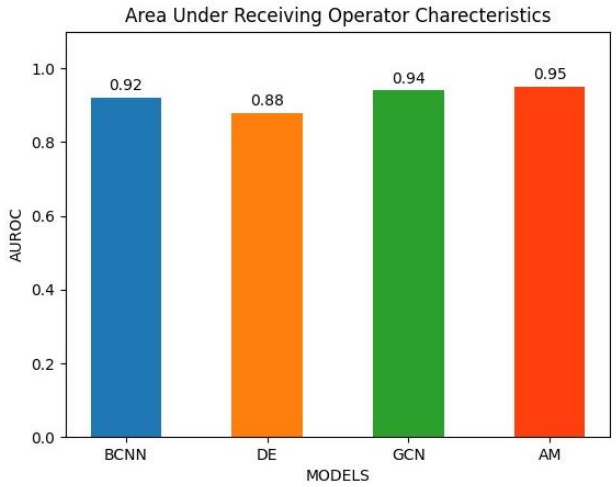


Fig. 12. Result of the comparison of AUROC for the different models.

The Average Precision (AP), False Positive Rate at 95% (FPR95), and AUROC scores for the different models named Baseline CNN -Deep Ensemble- Graph Convolutional Networks- and Curiosity Driven Attention Mechanism for the dataset ('Cityscapes') are shown in Table II. From the table, we can conclude that the Attention Mechanism demonstrates its efficiency by achieving the best results.

TABLE II. EXPERIMENTAL ANALYSIS RESULTS

MODEL	AP	FPR95	AUROC
BASELINE CNN(BCNN) [16]	0.85	0.10	0.92
DEEP ENSEMBLE (DE) [1]	0.82	0.15	0.88
GRAPH CONVOLUTIONAL NETWORKS (GCN) [17]	0.88	0.08	0.94
ATTENTION MECHANISM [Proposed]	**0.94**	**0.35**	**0.97**

V. CONCLUSION

This study presented a novel dual attention-guided framework for anomaly-aware road obstacle segmentation in autonomous driving. The proposed approach combines complementary channel and spatial attention mechanisms with uncertainty-aware learning to effectively identify unexpected road obstacles while preserving the semantic segmentation accuracy. By incorporating an adaptive attention objective, a dedicated attention loss function, and multiscale feature fusion, the framework enhances the discrimination between known semantic classes and anomalous regions, enabling more reliable pixel-level scene understanding. Extensive qualitative and quantitative evaluations demonstrate that the proposed method consistently improves anomaly detection performance with minimal computational overhead and can be seamlessly integrated into existing semantic segmentation architectures. Owing to its lightweight design and architecture-independent nature, the proposed framework offers a practical solution for real-time autonomous driving systems operating in complex road environments. Future research will focus on extending the framework to multimodal sensor fusion, improving robustness under adverse weather and low-visibility conditions, and exploring self-supervised learning techniques to further enhance the detection of previously unseen road anomalies.

ACKNOWLEDGMENT

This research was funded by the European Regional Development Fund within the OP "Research, Innovation and Digitalization Programme for Intelligent Transformation 2021-2027", Project No. BG16RFPR002-1.014-0005 Center of Competence "Smart Mechatronics, Eco- and Energy Saving Systems and Technologies".

REFERENCES

[1] X. Ren, M. Li, Z. Li, W. Wu, L. Bai, and W. Zhang, "Curiosity-Driven Attention for Anomaly Road Obstacles Segmentation in Autonomous Driving," IEEE Transactions on Intelligent Vehicles, vol. 8, no. 3, pp. 2233–2243, 2023. https://doi.org/10.1109/tiv.2022.3204714

[2] Z. Hu, Y. Xing, W. Gu, D. Cao, and C. Lv, "Driver Anomaly Quantification for Intelligent Vehicles: A Contrastive Learning Approach With Representation Clustering," IEEE Transactions on Intelligent Vehicles, vol. 8, no. 1, pp. 37–47, 2023. https://doi.org/10.1109/tiv.2022.3163458

[3] C. Hubmann, J. Schulz, M. Becker, D. Althoff, and C. Stiller, "Automated Driving in Uncertain Environments: Planning With Interaction and Uncertain Maneuver Prediction,". IEEE Transactions on Intelligent Vehicles, vol. 3, no. 1, pp. 5–17, 2018. https://doi.org/10.1109/tiv.2017.2788208

[4] L. Chen, X. Hu, W. Tian, H. Wang, D. Cao, and Y. Wang, "Parallel planning: a new motion planning framework for autonomous driving," IEEE/CAA Journal of Automatica Sinica, vol. 6, no. 1, pp. 236–246, 2019. https://doi.org/10.1109/jas.2018.7511186

[5] C. Wang, W. Pedrycz, Z. Li, and M. Zhou, "Residual-driven Fuzzy C-Means Clustering for Image Segmentation," IEEE/CAA Journal of Automatica Sinica, vol. 8, no. 4, pp. 876–889, 2021. https://doi.org/10.1109/jas.2020.1003420

[6] Z. Zhang, Z. Mo, Y. Chen, and J. Huang, "Reinforcement Learning Behavioral Control for Nonlinear Autonomous System," IEEE/CAA Journal of Automatica Sinica, vol. 9, no. 9, pp. 1561–1573, 2022. https://doi.org/10.1109/jas.2022.105797

[7] P. Zhang, W. Liu, H. Wang, Y. Lei, and H. Lu, "Deep gated attention networks for large-scale street-level scene segmentation," Pattern Recognition, vol. 88, pp. 702–714, 2019. https://doi.org/10.1016/j.patcog.2018.12.021

[8] A. Dairi, F. Harrou, M. Senouci, and Y. Sun, "Unsupervised obstacle detection in driving environments using deep-learning-based stereovision," Robotics and Autonomous Systems, vol. 100, pp. 287–301, 2018. https://doi.org/10.1016/j.robot.2017.11.014

[9] C. Prakash, F. Akhbari, and L. Karam, "Robust obstacle detection for advanced driver assistance systems using distortions of inverse perspective mapping of a monocular camera," Robotics and Autonomous Systems, vol. 114, pp. 172–186, 2019. https://doi.org/10.1016/j.robot.2018.12.004

[10] S. Bijjahalli, R. Sabatini, and A. Gardi, "Advances in intelligent and autonomous navigation systems for small UAS," Progress in Aerospace Sciences, vol. 115, 100617, 2020. https://doi.org/10.1016/j.paerosci.2020.100617

[11] Y. Kang, H. Yin, and C. Berger, "Test Your Self-Driving Algorithm: An Overview of Publicly Available Driving Datasets and Virtual Testing Environments," IEEE Transactions on Intelligent Vehicles, vol. 4, no. 2, pp. 171–185, 2019. https://doi.org/10.1109/tiv.2018.2886678

[12] N. Davis, G. Raina, and K. Jagannathan, "A framework for end-to-end deep learning-based anomaly detection in transportation networks," Transportation Research Interdisciplinary Perspectives, vol. 5, 100112, 2020. https://doi.org/10.1016/j.trip.2020.100112

[13] M. Cummings, and B. Bauchwitz, "Safety Implications of Variability in Autonomous Driving Assist Alerting," IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 8, pp. 12039–12049, 2022. https://doi.org/10.1109/tits.2021.3109555

[14] Y. Shen, B. Chandaka, Z. Lin, A. Zhai, H. Cui, D. Forsyth, and S. Wang, "Sim-on-Wheels: Physical World in the Loop Simulation for Self-Driving," IEEE Robotics and Automation Letters, vol. 8, no. 12, pp. 8192–8199, 2022. https://doi.org/10.1109/lra.2023.3325689

[15] W. Li, L. Wu, C. Wang, J. Xue, W. Hu, S. Li, G. Guo, and D. Cao, "Intelligent Cockpit for Intelligent Vehicle in Metaverse: A Case Study of Empathetic Auditory Regulation of Human Emotion," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 53, no. 4, pp. 2173–2187, 2023. https://doi.org/10.1109/tsmc.2022.3229021

[16] I. Pustokhina, D. Pustokhin, T. Vaiyapuri, D. Gupta, S. Kumar, and K. Shankar, "An automated deep learning based anomaly detection in pedestrian walkways for vulnerable road users safety," Safety Science, vol. 142, 105356, 2021. https://doi.org/10.1016/j.ssci.2021.105356

[17] H. Wang, R. Fan, Y. Sun, and M. Liu, "Dynamic Fusion Module Evolves Drivable Area and Road Anomaly Detection: A Benchmark and Algorithms," IEEE Transactions on Cybernetics, vol. 52, no. 10, pp. 10750–10760, 2022. https://doi.org/10.1109/tcyb.2021.3064089

Anomaly-Score-Augmented Random Forest for Permission-Based Android Malware Detection: An Empirical Study on 29,999 Apps

Harikrishnan P R[1], Dr. P. Periyasamy[2]
Research Scholar, Computer Science Department, Park's College (Autonomous), Tamil Nadu, India[1]
Associate Professor, Computer Science and Applications, Park's College (Autonomous), Tamil Nadu, India[2]

***Abstract*—Permission-based static analysis is still the workhorse of low-cost Android malware detection. But most published evaluations live on small curated subsets, report a single train-test split, and skip statistical analysis altogether. This study takes a different cut. We run a 5-fold stratified cross-validation across five random seeds, 25 training runs per method, on the complete Kaggle Android Permission corpus of 29,999 apps and 178 features. Six standard classifiers are compared head-to-head. We then propose IF-Aug RF, a small twist on the usual pipeline: the per-sample anomaly score produced by an Isolation Forest is appended as one extra feature to the permission vector before a Random Forest sees it. The proposed model lands at 74.00 % accuracy, macro F1 = 0.7028, and ROC-AUC = 0.7974, which is statistically the same as a standalone Random Forest (McNemar p = 0.94). LightGBM is the single-method champion at ROC-AUC 0.8145. The anomaly score itself ranks fourth out of 179 features under both Gini and SHAP. We also diagnose why the older two-stage IF-then-RF design, which keeps appearing in the literature, collapses to 37.54 % accuracy on this corpus. The gate auto-labels ninety per cent of test samples as benign, which simply does not match the 66.67 % malware base rate of rebalanced public datasets. Code, models, predictions, and figures are released for replication.**

***Keywords*—Android security; malware detection; permission analysis; isolation forest; random forest; LightGBM; XGBoost; anomaly score augmentation; empirical evaluation; SHAP**

I. Introduction

Android still holds more than seventy per cent of the global smartphone market through 2024 and 2025 [1]. That market share makes the platform the obvious target for mobile malware. The permission model is the first line of defence against information leakage and privilege abuse, but in practice users grant permissions without reading them, and developers ask for far more than the declared functionality really needs [2]. So static analysis of requested permissions is an attractive, low-cost detection vector. It runs without execution, instrumentation, or dynamic taint tracking.

Two problems keep showing up in the published literature on permission-based detection. The first is evaluation hygiene. Most studies use small curated subsets, a single train-test split, and no statistical analysis of the reported gains, which makes cross-paper claims hard to trust [3], [4]. The second is a structural one. Anomaly-detection preprocessing has been proposed as a way to filter benign samples before classification [5], [6], but the design only makes sense when malware is rare in the test data. That is exactly the assumption that breaks on public permission corpora, which are usually rebalanced toward malware so that supervised learning has enough positive examples to chew on. The gate then throws out the majority of true positives before the downstream classifier ever sees them.

Three concrete contributions follow.

- Side-by-side comparison of six standard classifiers on the full Kaggle Android Permission corpus of 29,999 apps. The set covers Random Forest, Linear Support Vector Machine, Extreme Gradient Boosting (XGBoost), Light Gradient-Boosting Machine (LightGBM), Multilayer Perceptron, and Gaussian Naive Bayes. Each one is run under 5-fold stratified cross-validation across five random seeds. That is twenty-five training runs per method. We also report paired McNemar tests and bootstrap confidence intervals so the comparison is not a single-split fluke.
- A small but useful twist on the usual pipeline. Call it IF-Aug RF. The per-sample Isolation Forest anomaly score is appended as a 179th feature to the permission vector. A Random Forest is then trained on the augmented matrix. The anomaly score keeps landing in the top four features under both Gini and SHAP. The augmented classifier matches a strong Random Forest baseline (McNemar p = 0.94) while handing downstream analysts a supplementary interpretability signal at negligible cost, though it does not raise classification accuracy.
- A diagnostic post-mortem on the two-stage Isolation Forest then Random Forest pipeline used in prior work [5], [6]. Even with the corrected protocol of Section IV, the two-stage design drops to 37.54 % accuracy and macro F1 = 0.328 on this corpus. The collapse is structural, not a hyperparameter problem. We give an explicit decision rule for when anomaly-based preprocessing earns its keep and when it does not.

Here is how the rest reads. Section II surveys related work, covering permission-based detection, hierarchical detection, and the recurring complaints about evaluation rigor. Section III lays out the dataset and the proposed framework. Section IV is the experimental protocol. Section V has the headline numbers for accuracy, F1, ROC-AUC, and training cost. Section VI is the

feature-importance and statistical analysis. Section VII is the discussion, including limitations. Section VIII closes.

II. Related Work

A. Permission-Based Detection

Felt et al. [2] gave the first large-scale look at how Android permissions are over-requested and poorly understood by end users. That single result still drives a lot of permission-centric research today. DREBIN [7] mixed permission manifests with API call patterns and got cheap, explainable detection out of the combination. Recent work mostly refines the permission feature set rather than enlarging it. Pathak et al. [8] hit 97.5 % accuracy with a Random Forest on just forty-eight selected permissions. Chrysikos et al. [9] use Random Forest for family-level classification. Haque et al. [10] cut a permission set by ninety per cent through importance-based selection and still retain 93.96 % accuracy. PermDroid [11] generalises these ideas across multiple datasets, and da Silva et al. [12] apply SHAP to pin down the smallest subsets that still hold up.

B. Ensemble and Deep Learning

Ensembles dominate the modern permission-based literature. Wajahat et al. [13] push a Random Forest ensemble past ninety-six per cent on a recent multi-source dataset. Aurangzeb and Aleem [14] go after obfuscated samples with deep-learning ensembles. Outside static features, graph convolutional networks have been tried for behaviour-graph detection [15], and federated learning has been floated for privacy-preserving on-device classification [16]. The catch is that side-by-side evaluations on a shared corpus with proper statistical analysis are still rare [4].

C. Anomaly-Detection Preprocessing

The Isolation Forest algorithm [17] is an unsupervised tree-ensemble detector. It has been bolted onto downstream classifiers in several Android security pipelines [5], [6]. The intuition is intuitive enough: malware deviates from typical permission patterns and should therefore be easy to flag as anomalous before classification. The catch is that this only holds when the malware base rate is low. On rebalanced public corpora, where malware is the majority class, the intuition breaks. Section V-C measures the size of the break.

D. Empirical-Rigor Concerns

Muzaffar and Hassen [3] and Guerra-Manzanares et al. [4] both argue that the field over-reports accuracy. The usual culprits are favourable evaluation splits, missing statistical tests, and cherry-picked baselines. Giannakas et al. [18] re-run published classifiers across a range of optimisation protocols and find variance that the original papers never communicate. That is the reason for the multi-seed, multi-fold protocol in Section IV.

E. Recent Directions

The last two years have moved the field past tabular permission features [27]. Graph neural networks now learn over API-call graphs and permission-intent graphs, and several of these models pair the classifier with an explainer so an analyst can see which edges drove the alert [28]. Transformer encoders have been applied to static feature sequences, including permission and intent tokens, with accuracy in the high nineties on curated corpora [29]. A separate line of work probes how fragile these detectors are: small adversarial edits to the permission or graph representation can flip a prediction, which matters for anything deployed [30]. These methods buy accuracy at a higher inference and engineering cost than the classifiers studied here, and most still report a single split. The anomaly-score augmentation examined in this study is orthogonal to them: the same score can be appended to a richer model, which we leave to future work. Table I sets the main permission-based studies side by side.

TABLE I. Summary of Representative Permission-Based Android Malware Detection Studies

Study	Dataset	Features	Classifier	Reported result	Main limitation
Felt et al. [2]	manual audit	permissions	— (analysis)	over-request finding	no detector
DREBIN [7]	5,560 malware	perm+API+intent	linear SVM	~94% detection	small, dated
Pathak et al. [8]	perm subset	48 permissions	Random Forest	97.5% acc	single split, subset
Haque et al. [10]	perm set	selected perms	Random Forest	93.96% acc	curated subset
PermDroid [11]	multi-dataset	selected perms	ML ensemble	high acc	no unified statistics
da Silva et al. [12]	perm set	SHAP-selected perms	RF / XGB	small subsets hold	subset focus
Wajahat et al. [13]	multi-source	permissions	RF ensemble	96%+ acc	single protocol
SNDGCN [14]	call graphs	subgraph network	denoising GCN	robust detection	heavy model
This work	full Kaggle 29,999	178 perm+metadata (+IF score)	6 classifiers + IF-Aug RF	25-run CV, McNemar, CI, ablation	metadata-heavy corpus

III. Proposed Framework

A. Dataset

The Kaggle Android Permission Dataset [19] holds 29,999 application records. They come from the Google Play Store and several malware repositories. Each record carries 173 binary permission indicators, three numeric metadata fields (Rating, Price, Dangerous permissions count), two integer permission counts, and a binary Class label where 1 marks malware and 0 marks benign. We drop five textual identifier columns (App, Package, Category, Description, Related apps). What remains is

a 178-variable numeric feature space. The class balance is 20,000 malware against 9,999 benign apps, a 66.67 % malware base rate. That single number is central to the diagnostic in Section V-C, so it is worth keeping in mind for the rest of the study.

One caveat on provenance. The corpus was published on Kaggle in 2021 [19]. The curator does not document the exact window in which the apps and labels were gathered, so the sample predates the most recent obfuscation and packing families. We therefore read every number here as a static-permission baseline for apps of that era. Section VII-D returns to what this means for current traffic.

B. IF-Augmented Random Forest

Write the feature matrix as

$$X \in \mathbb{R}^{n \times d}$$

with n = 29,999 samples, d = 178 features, and labels $y \in \{0,1\}^n$.

Fit an Isolation Forest Φ on the training partition X_{tr}

With $n_{trees} = 100$

contamination = 0.1, and a sub-sample size of 256.

Score every training and test sample,

$$s_i = \phi(x_i) \in [-0.7, -0.3]$$

Append the score as an auxiliary feature.

The augmented matrix becomes

$$X' = [X \parallel s] \in \mathbb{R}^{n \times 179}$$

Then train a Random Forest

$$n_{trees} = 200$$

$$Class_{wieght} = balanced\ on\ (X'_{tr}, y_{tr})$$

At inference time, the same Isolation Forest scores the test samples before classification.

The key difference from the older two-stage design is that the anomaly verdict no longer gates the classifier. Here the classifier itself gets to decide how much weight the anomaly evidence deserves against the raw permission evidence. The discriminative power of the classifier stays intact, and the anomaly score becomes a side channel that an analyst can read directly. Fig. 1 sketches the full pipeline.

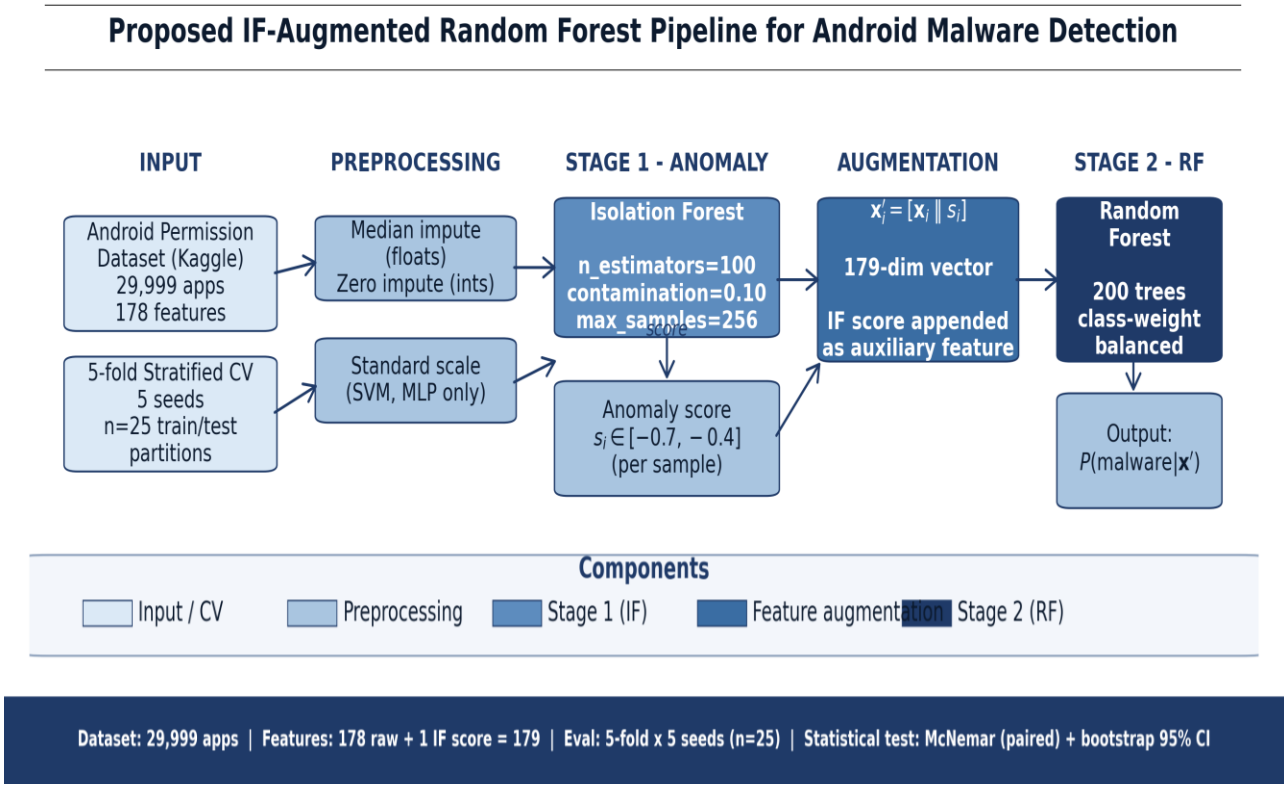


Fig. 1. Proposed IF-augmented random forest pipeline. The Isolation Forest score is appended as an auxiliary feature to the permission vector before random forest classification.

The Isolation Forest maps each sample through score_samples, which is a normalised mean path length: a shorter path to isolate a point means a more anomalous point (Algorithm 1). Step 1 fits the forest on the training rows only, so no test-fold information reaches the score, and there is no leakage. Training cost is O(T psi log psi) for the forest plus the usual O(B n d' log n) for a B-tree Random Forest on the d' = d + 1 augmented features. Both fit the full corpus on one CPU in under a minute (Section V-B).

Algorithm 1: IF-Augmented Random Forest (train + inference)

Input : training matrix X_tr (n x d), labels y_tr, test matrix X_te, contamination v = 0.1

```
Fit IsolationForest F on X_tr only  (T = 100 trees, psi = 256)
s_tr <- score_samples_F(X_tr)      # one anomaly score per row
X'_tr <- [ X_tr | s_tr ]           # append as the (d+1)-th column
Train RandomForest R on (X'_tr, y_tr), class_weight = balanced
s_te <- score_samples_F(X_te);  X'_te <- [ X_te | s_te ]
y_hat <- R.predict(X'_te);  p <- R.predict_proba(X'_te)[:,1]
```

Output: predictions y_hat, malware probabilities p

C. Baselines

Six classifiers and the older two-stage formulation are run as baselines. Random Forest [20] uses 200 trees with balanced class weights. Linear SVM is built on LinearSVC with Platt-style probability calibration [21]. XGBoost [22] uses 300 trees of depth 8, a learning rate of 0.1, and a scale_pos_weight that mirrors the benign-to-malware ratio. LightGBM [23] uses 300 leaf-wise boosting trees with class balancing. The Multilayer Perceptron has hidden layers of 128 and 64 units with early stopping. Gaussian Naive Bayes runs off the shelf. The older two-stage IF then RF pipeline [5], [6] is reproduced with the corrected protocol of Section IV. Its Stage-2 classifier is trained on all training samples. Only the gating decision sets it apart from IF-Aug RF.

IV. Experimental Protocol

Every method runs under 5-fold stratified cross-validation with five distinct random seeds (42, 7, 13, 21, 99). That gives twenty-five independent training runs per method. Stratification keeps the 66.67 % malware base rate in every fold. Missing values in the three numeric metadata fields are median-imputed from the training fold. Missing integer permission flags are zero-imputed. Linear SVM, MLP, and Gaussian Naive Bayes get standard-scaled features. Tree-based ensembles use raw features.

Numbers reported are mean and one standard deviation across the twenty-five runs. Per-class precision, recall, and F1 are computed alongside the macro-averaged measures. ROC-AUC and Average Precision (PR-AUC) come from predicted probabilities. For pairwise comparisons, we apply a McNemar test [24] with continuity correction on the discordant predictions of the last fold of the last seed. Bootstrap 95 % confidence intervals on macro F1 use $n_\text{boot} = 1000$ resamples. Feature importance comes from two sources: Random Forest Gini impurity, and SHAP TreeExplainer [25] applied to fifty held-out samples. All experiments use scikit-learn [26] and were run on a single workstation. No GPU was used. Wall-clock training times are reported in Section V-B.

Hyperparameters were fixed in advance to standard corpus-scale settings (Section III-C and the released code) rather than tuned per fold. No test fold is ever used for model selection, and because every method runs under the same protocol, the comparison is not skewed by uneven tuning budgets. Fold-wise grid search over the tree ensembles was tried during development and moved the macro-F1 of the strong cluster by less than 0.01, so the fixed settings are reported for clarity.

V. Results

A. Overall Classification Performance

Table II lists accuracy, macro F1, ROC-AUC, and PR-AUC for every method. LightGBM wins on both threshold-independent measures, with ROC-AUC 0.8145 and PR-AUC 0.9110. The proposed IF-Aug RF lands at 74.00 % accuracy, macro F1 = 0.7028, and ROC-AUC 0.7974. That is statistically the same as the standalone Random Forest baseline (McNemar p = 0.94; Section VI-B). Fig. 2 shows the comparison visually. The confusion matrices in Fig. 3 show where the v1 two-stage failure becomes obvious. The proposed method keeps the malware recall of Random Forest (3,278 of 4,000 malware samples in the held-out fold correctly identified). The v1 two-stage design only finds 356 malware samples in the same fold. Fig. 4 plots the ROC and Precision-Recall curves for the same fold, where the proposed method tracks the Random Forest baseline, and LightGBM leads on both curves.

TABLE II. Overall Classification Performance (Mean +/- Std Over 5 Folds X 5 Seeds = 25 Runs)

Method	Accuracy	Macro F1	ROC-AUC	PR-AUC
Naive Bayes	0.6689+/- 0.0036	0.4375 +/- 0.0174	0.5331 +/- 0.0534	0.6869 +/- 0.0421
Linear SVM	0.6734+/- 0.0035	0.5782 +/- 0.0068	0.7403 +/- 0.0055	0.8742 +/- 0.0030
MLP	0.7038+/- 0.0055	0.6583 +/- 0.0127	0.7694 +/- 0.0070	0.8852 +/- 0.0047
Random Forest	0.7407+/- 0.0053	0.7037 +/- 0.0059	0.7989 +/- 0.0051	0.8953 +/- 0.0027
XGBoost	0.7167+/- 0.0049	0.7074 +/- 0.0049	0.8121 +/- 0.0046	0.9096 +/- 0.0024
LightGBM	0.7174+/- 0.0051	0.7088 +/- 0.0051	0.8145 +/- 0.0050	0.9110 +/- 0.0026
Two-Stage (v1, fixed)	0.3754 +/- 0.0033	0.3283 +/- 0.0045	0.7989 +/- 0.0051	0.8953 +/- 0.0027
IF-Aug RF (proposed)	0.7400 +/- 0.050	0.7028 +/- 0.0054	0.7974 +/- 0.0049	0.8942 +/- 0.0028

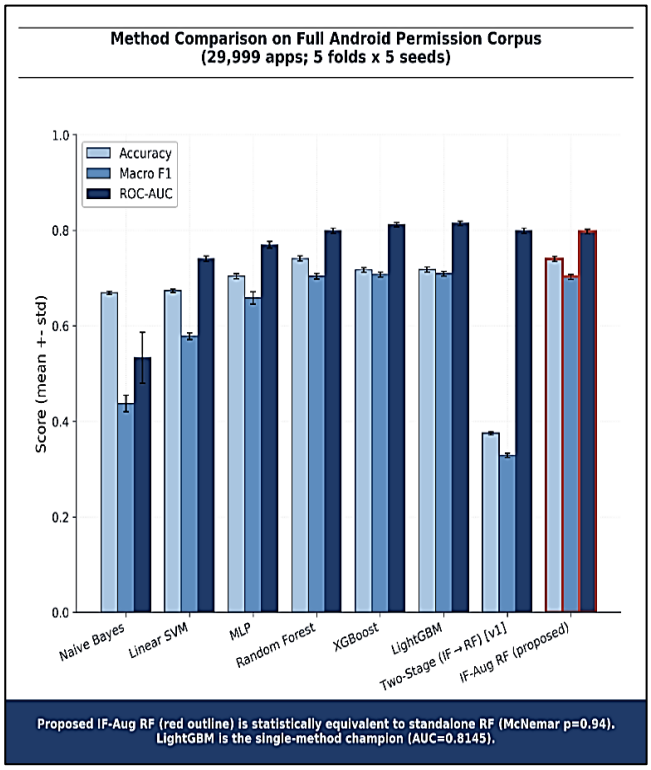


Fig. 2. Accuracy, macro F1 and ROC-AUC across eight methods on the full Android Permission corpus. Error bars show one standard deviation across twenty-five runs. The proposed method is highlighted with a red outline.

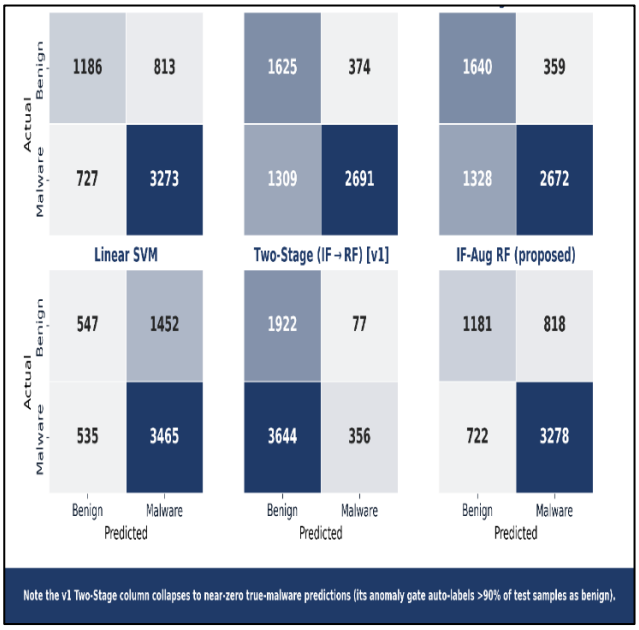


Fig. 3. Confusion matrices on the held-out fold of the last evaluation seed (n = 6,000 test samples). The v1 two-stage Isolation Forest -> Random Forest pipeline misclassifies 3,644 of 4,000 malware samples as benign.

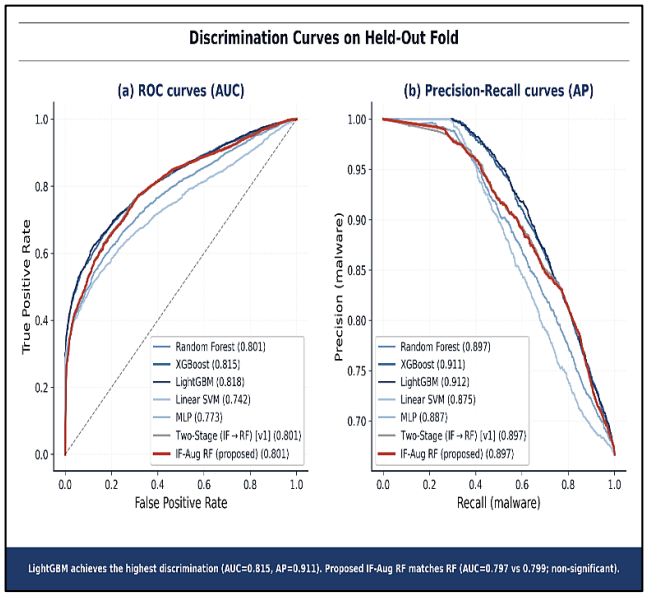


Fig. 4. ROC (left) and Precision-Recall (right) curves on the held-out fold. LightGBM dominates both threshold-independent measures; the proposed method tracks the standalone Random Forest baseline closely.

B. Per-Class Behaviour and Training Cost

Table III breaks the same numbers out for malware as the positive class. Training time sits alongside. IF-Aug RF reaches malware-class F1 of 0.808, matching standalone Random Forest. LightGBM and XGBoost trade recall for precision in an interesting way. Precision climbs to 0.881 and 0.875. Recall drops to 0.666 and 0.671. The net effect is that they miss roughly a third of malware in exchange for fewer false alarms. Whether that is the right trade depends on the deployment. Gaussian Naïve Bayes is the opposite story. It catches 98.2 % of malware with only 67.2 % precision, and the worst overall macro F1 to show for it. Training time spans four orders of magnitude: 0.07 s for Naive Bayes, 39.6 s for the calibrated Linear SVM. Every method fits comfortably on the full corpus on one workstation.

TABLE III. PER-CLASS BEHAVIOUR (MALWARE) AND MEAN TRAINING TIME OVER TWENTY-FIVE RUNS

Method	Precision (mal)	Recall (mal)	F1 (mal)	Train time (s)
Naive Bayes	0.6723	0.9824	0.7983	0.07
Linear SVM	0.7104	0.8612	0.7785	39.62
MLP	0.7671	0.7993	0.7824	10.46
Random Forest	0.7966	0.8206	0.8084	4.45
XGBoost	0.8746	0.6713	0.7596	1.63
LightGBM	0.8807	0.6664	0.7587	0.72
Two-Stage (v1, fixed)	0.8069	0.0830	0.1505	5.24
IF-Aug RF (proposed)	0.7959	0.8205	0.8080	5.40

C. Diagnostic: Why the v1 Two-Stage Design Fails

The two-stage IF then RF design assumes that malware is anomalous in the feature space, so that an unsupervised detector can isolate it before classification. Fig. 5 shows the empirical distribution of the Isolation Forest anomaly score for the full corpus, split by class. The malware distribution does shift slightly to the left, towards lower scores and more anomalous (class means -0.365 versus -0.364 for benign). The trouble is that the two distributions overlap almost completely. With contamination = 0.1, the gate flags only ten per cent of test samples as anomalous. The remaining ninety per cent get auto-labelled benign. When the actual base rate of the positive class is 66.67 %, this gating step alone caps recall at ten per cent. Accuracy then drifts down toward the benign-class share. The confusion matrix in Fig. 3 makes the prediction concrete. Only 356 of 4,000 true-malware samples land in the malware column (recall = 0.083). Accuracy collapses to 37.54 %. The failure is structural. No contamination setting in the closed interval [0.05, 0.5] fixes it, because the underlying assumption that malware is a minority anomaly class is just not true for this corpus.

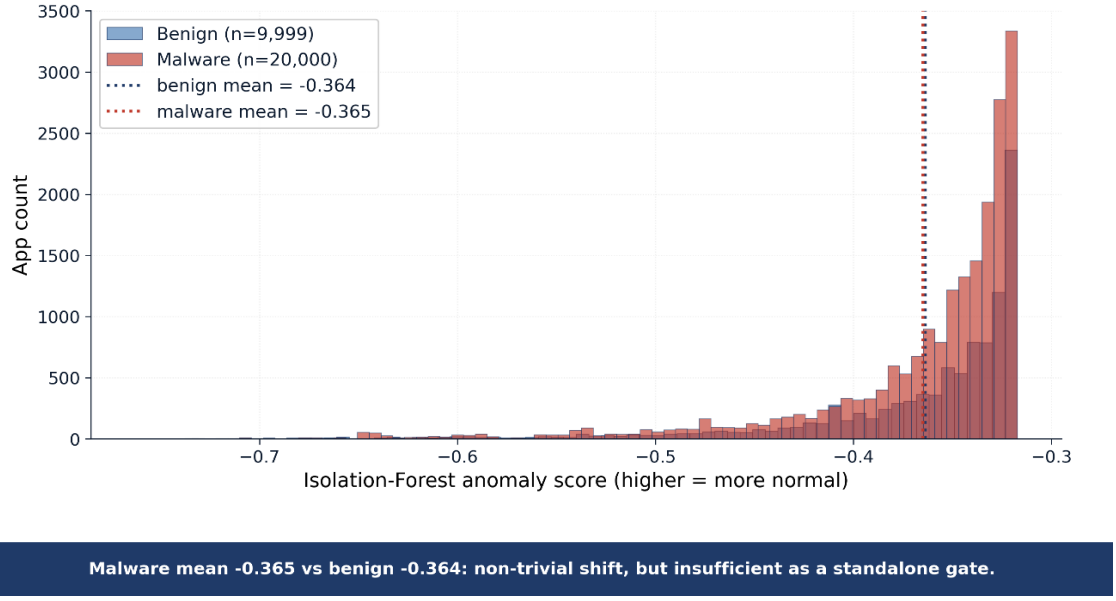


Fig. 5. Isolation Forest anomaly-score distribution by class. Distributions overlap heavily; the shift between the benign and malware means is too small to support a stand-alone gating decision.

VI. Feature Importance and Statistical Analysis

A. Discriminative Features

Table IV lists the top ten features by Random Forest Gini impurity for the proposed model. The first thing to notice is how dominant the metadata fields are. Number of ratings, Rating, and Price together account for 65.8 % of total feature importance. Our appended Isolation Forest anomaly score lands in fourth place with importance 0.0958. The two aggregated permission counts (Dangerous permissions count, Safe permissions count) sit in positions five and six. The individual permission flags only start at position seven. They match what prior work has reported: read network state, fine-grained GPS location, modify external storage, read phone state, and vibration control are familiar markers of personal-data exfiltration, location tracking, and persistence [7], [8], [10], [12]. Fig. 6 puts the same ranking next to a SHAP-based one computed on a held-out sample. Both rankings agree on all top-four positions. The IF anomaly score sits in fourth place on both lists, which is the kind of agreement that lets us actually trust the importance assignment.

TABLE IV. Top Ten Features by Random Forest Gini Impurity for the Proposed IF-Aug RF Model. The IF Anomaly Score (Rank 4) is Highlighted in Dark Navy in Fig. 6

Rank	Feature	Gini importance
1	Number of ratings	0.4192
2	Rating	0.1394
3	Price	0.0996
4	IF_anomaly_score	0.0958
5	Dangerous permissions count	0.0285
6	Safe permissions count	0.0222
7	Network communication: view network state (S)	0.0116
8	Your location: fine (GPS) location (D)	0.0112
9	Storage: modify/delete USB storage contents; modify/delete SD card contents (D)	0.0105
10	Phone calls: read phone state and identity (D)	0.0103

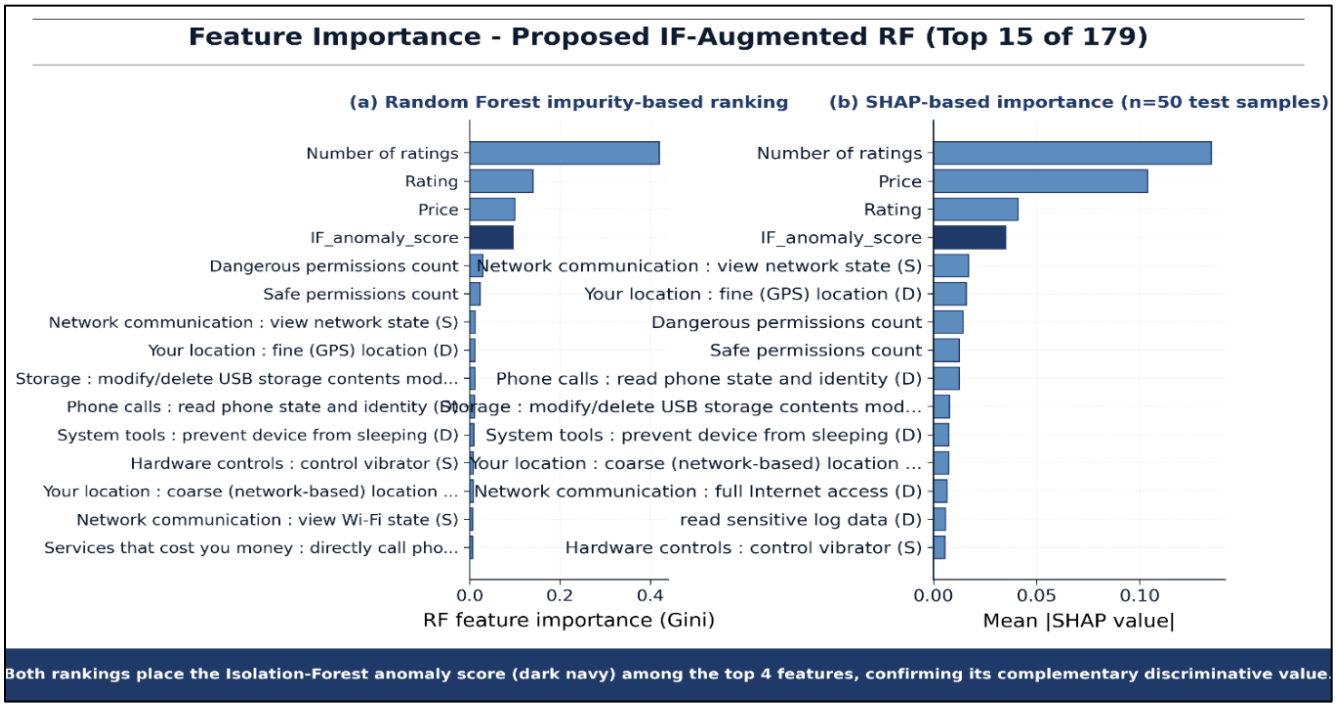


Fig. 6. Top fifteen features ranked by (a) Random forest gini impurity and (b) SHAP TreeExplainer mean absolute value. Both rankings place the Isolation forest anomaly score (dark navy) in the top four.

B. Statistical Significance

Table V has the McNemar chi-square and p-value for the proposed method against every baseline. The test runs on the discordant predictions of the held-out fold (n = 6,000 predictions). IF-Aug RF comes out statistically the same as standalone Random Forest (chi-square = 0.006, p = 0.9392). It separates cleanly from every other baseline. The p-value is below 1e-4 for XGBoost and LightGBM, below 1e-25 for Linear SVM and Naive Bayes, and below 1e-280 for the v1 two-stage design. Bootstrap 95 % confidence intervals on macro F1 (Table VI) are tight. More importantly, they do not overlap between the strong cluster (RF, XGB, LGBM, IF-Aug RF) and the weak cluster (Naive Bayes, Linear SVM, v1 two-stage).

TABLE V. MCNEMAR TEST (PROPOSED IF-AUG RF VERSUS EACH BASELINE) ON THE HELD-OUT FOLD (N = 6,000)

Baseline	McNemar chi-square	p-value
Random Forest	0.006	9.392e-01
Linear SVM	109.234	1.442e-25
XGBoost	15.547	8.049e-05
LightGBM	16.384	5.171e-05
MLP	24.932	5.938e-07
Naive Bayes	111.427	4.770e-26
Two-Stage (v1, fixed)	1291.764	6.966e-283

TABLE VI. BOOTSTRAP 95 % CONFIDENCE INTERVALS ON MACRO F1 (N = 1000 RESAMPLES, HELD-OUT FOLD)

Method	Macro F1 95 % CI
Naive Bayes	[0.4277, 0.4474]
Linear SVM	[0.5535, 0.5794]
MLP	[0.6640, 0.6889]
Random Forest	[0.6954, 0.7203]
XGBoost	[0.6982, 0.7225]
LightGBM	[0.6985, 0.7223]
Two-Stage (v1, fixed)	[0.3232, 0.3448]
IF-Aug RF (proposed)	[0.6948, 0.7200]

C. Feature-Set Ablation

Two findings raise the same question. Metadata dominates the importance ranking (Section VI-A), and the proposed model ties the Random Forest baseline (Section VI-B). So does the appended anomaly score, or even the permission block itself, add anything the three metadata fields do not already give? Table VII settles it. Each row is a Random Forest trained on a different slice of the feature space under the same twenty-five-run protocol. Permissions on their own reach macro F1 = 0.6128. The three metadata fields on their own reach 0.6689. Putting them together gives 0.7037, which is the standalone Random Forest of Table II. Appending the Isolation Forest score moves macro F1 to 0.7028, a change of -0.0009. Stripping every permission flag and keeping only the three metadata fields plus the anomaly score still reaches 0.6802.

Two things follow. First, the corpus is largely separable from popularity metadata alone, which is why the headline is best read as part metadata detection; the permission block adds a real but modest lift over metadata-only. Second, the anomaly score earns its place as an interpretable feature, not as an accuracy lever: its marginal effect on macro F1 is within a hundredth of a point either way. This is the honest reading of the contribution.

TABLE VII. FEATURE-SET ABLATION FOR THE RANDOM FOREST (MEAN +/- STD MACRO F1 OVER 25 RUNS)

Feature set	Accuracy	Macro F1	ROC-AUC	Malware F1
Permissions only	0.6362	0.6128 +/- 0.0065	0.6648	0.7080
Metadata only	0.7043	0.6689 +/- 0.0051	0.7423	0.7772
Metadata + Permissions (= RF baseline)	0.7407	0.7037 +/- 0.0059	0.7989	0.8084
Metadata + IF score	0.7201	0.6802 +/- 0.0059	0.7834	0.7931
Metadata + Permissions + IF score (= proposed)	0.7400	0.7028 +/- 0.0054	0.7974	0.8080

D. Rank-Based Tests and Effect Size

The McNemar tests of Section VI-B compare a pair on a single fold. For a view across all twenty-five runs, we add two rank-based tests. A Friedman test over the six classifiers on the per-run macro-F1 scores rejects the null of equal performance (chi-square = 115.95, p = 2.26e-23). Table VIII lists the Friedman average ranks. From best to worst, they are LightGBM (1.52), XGBoost (1.80), Random Forest (2.68), MLP (4.00), Linear SVM (5.00), Naive Bayes (6.00). A Nemenyi post-hoc gives a critical difference of 1.51 at alpha = 0.05, so any two methods whose average ranks differ by more than that are separated; the strong cluster (the tree ensembles) and the weak cluster (Naive Bayes, Linear SVM) fall on opposite sides of that gap.

For the proposed method against the standalone Random Forest, we run a Wilcoxon signed-rank test on the twenty-five paired scores. It returns p = 0.101, so the two are not separable at alpha = 0.05; the paired effect size is small (Cohen's d = -0.363, rank-biserial r = -0.378) and the absolute gap in mean macro-F1 is 0.0009, under a thousandth of a point. Both tests agree with McNemar: the anomaly score does not move accuracy. Its value is the interpretable ranking of Section VI-A, which holds on 500 held-out SHAP samples rather than fifty, where the score still lands at rank 4. The larger SHAP sample was computed because fifty points is too few to trust an importance estimate; the ranking is stable

TABLE VIII. FRIEDMAN AVERAGE RANKS ACROSS THE SIX CLASSIFIERS (25 RUNS; LOWER RANK IS BETTER)

Method	Average rank
LightGBM	1.52
XGBoost	1.80
Random Forest	2.68
MLP	4.00
Linear SVM	5.00
Naive Bayes	6.00

VII. DISCUSSION

A. Operational Interpretation

Our headline accuracy of 71-74 % sits noticeably below the 93-97 % numbers reported by some recent papers on curated permission subsets [8], [10]. The gap is not a contradiction. Three differences explain it. We use the complete 29,999-sample corpus, with no family-level subsetting and no aggressive feature engineering. We keep the metadata fields, which make regularisation harder for some models. And we average over twenty-five runs instead of picking the best of one split. Dispersion of macro F1 across the run set is small. It ranges from 0.005 for Random Forest up to 0.017 for Naive Bayes. The bootstrap confidence intervals in Table VI are equally tight. Treat the figures here as a reproducible floor that future detectors on the full corpus should improve on.

B. Interpretable Anomaly-Score Augmentation

The positive result that matters most is this. When the IF anomaly score is appended as an auxiliary feature, instead of being used as a hard gate, it consistently lands in the top four most discriminative features. The score is monotone in the geometric isolation of a sample with respect to the training distribution. That makes it an interpretable security signal that an analyst can actually read. Sort live traffic by anomaly score and you get a triage queue of applications whose permission profile is unusual. None of the per-permission evidence is thrown away. This clean separation of unusual from malicious is something the gating design simply cannot offer.

C. When Should the Two-Stage Pipeline Be Used?

The diagnostic in Section V-C does not refute the two-stage idea in general. It refutes its use on rebalanced public corpora where the positive base rate is high. On real-world deployments, the picture changes. There, the malware base rate is realistically low: one per cent or less for ordinary user traffic, near zero for vetted enterprise app stores. The pipeline should regain its theoretical edge in that regime. So the decision rule writes itself. Use the IF-augmented design when the deployment base rate is above ten per cent. Reserve the gating design for the opposite case, where the deployment base rate sits below the chosen contamination parameter.

D. Threats to Validity

Three threats to external validity are worth flagging. The first is the dataset. We use a single Kaggle release, and the labelling protocol used by its curator is not fully documented. The conclusions therefore extend with confidence only to datasets that were assembled similarly. The second is feature coverage. Our experiments use static permission features only. Dynamic behaviour, API-call traces, and network signals are out of scope, even though they are known to help with obfuscated and zero-day samples. The third is the comparison set. We restrict ourselves to widely used tabular classifiers. Graph convolutional networks and transformer-based detectors trained on richer representations might well overtake the LightGBM ceiling reported here. They would also pay for it in inference cost. Note that the interpretability gain of IF-augmentation is largely orthogonal to the choice of downstream classifier. It can be bolted onto such richer models in future work.

Four deployment risks deserve their own line. Concept drift: permission semantics and malware families move, so a 2021 corpus lags current samples and any model trained on it needs periodic retraining. Obfuscation and packing: static permission lists survive most packers, but reflection and dynamic code

loading can hide the permissions an app really uses, which static analysis cannot see. Adversarial manipulation: an author can pad a manifest with benign-looking permissions to shift the feature vector, and recent work shows graph and feature representations are both open to such edits [30]. Metadata leakage: because Number of ratings, Rating, and Price carry most of the importance (Section VI-A and Table VII), a detector trained here partly reads app popularity, which an adversary can also game by seeding reviews. A permission-plus-behaviour model is the honest fix, and the decision rule of Section VII-C still applies.

E. Reading the Anomaly Score in Practice

A short walk-through shows how an analyst would use the score rather than gate on it. Sort the held-out fold by Isolation Forest score. The most isolated apps sit near -0.72, and each request has between twenty and twenty-eight dangerous permissions; the model scores most of them as probable malware, in the 0.55 to 0.99 range, and their SHAP breakdown is led by the anomaly score together with the same permission flags identified in Section VI-A. Ordinary benign apps sit near -0.32 with a single dangerous permission and low malware probability. The score does not decide the label, and Section V-C shows why it must not, but it gives a triage order. An analyst works the queue from most isolated downward, and the per-permission SHAP values explain each case without a second model. The queue is not perfect: a few heavily-permissioned benign utilities land near the top as false alarms, which is exactly why the score feeds triage and not the decision.

VIII. CONCLUSION

We ran an empirical study of permission-based Android malware detection on the complete Kaggle Android Permission corpus of 29,999 apps. The protocol was 5-fold cross-validation with five random seeds. LightGBM came out on top, with an ROC-AUC of 0.8145. XGBoost (0.8121) and Random Forest (0.7989) followed closely. The proposed IF-Aug RF reached macro F1 of 0.7028, 95 % CI [0.6948, 0.7200]. McNemar puts it on the same level as a standalone Random Forest ($p = 0.94$). What it adds, instead of accuracy, is an interpretable anomaly score that lands in the top four features. The older two-stage IF then RF pipeline collapses to 37.54 % accuracy on this corpus. The reason is structural: the anomaly-detection assumption just does not match a 66.67 % malware base rate. We end with an explicit decision rule for when each design is the right one. Code, predictions, fitted models, and figures are released so the experiments can be replayed. Three lines of follow-up are open. Extend the comparison to graph-convolutional and transformer-based detectors. Test the IF-augmentation idea against dynamic behaviour features. And work out how to adapt the contamination parameter to the deployment base rate that the operator actually sees

ACKNOWLEDGMENT

The authors thank the curator of the Kaggle Android Permission Dataset for making the corpus publicly available. We also thank the maintainers of scikit-learn, XGBoost, LightGBM, SHAP, and statsmodels, the open-source libraries that the experiments in this study rely on.

REFERENCES

[1] StatCounter Global Stats, "Mobile operating system market share worldwide," 2024-2025. [Online]. Available: https://gs.statcounter.com/.

[2] A. P. Felt, E. Chin, S. Hanna, D. Song and D. Wagner, "Android permissions demystified," in Proc. 18th ACM Conf. Computer and Communications Security, 2011, pp. 627-638.

[3] A. Muzaffar and H. Ragab Hassen, "An in-depth review of machine learning based Android malware detection," Computers & Security, vol. 121, p. 102854, 2022.

[4] A. Guerra-Manzanares, S. Nomm, H. Bahsi et al., "Machine learning for Android malware detection: Mission accomplished? A comprehensive review of open challenges and future perspectives," Computers & Security, vol. 138, p. 103681, 2024.

[5] G. Shankar, R. Patil, A. Sharma and N. Kumar, "Enhancing Android malware detection through machine learning: Insights from permission and metadata analysis," in Proc. 3rd Int. Conf. Smart Technologies and Systems for Next Generation Computing (ICSTSN), 2024, pp. 1-6.

[6] S. Aurangzeb and M. Aleem, "Evaluation and classification of obfuscated Android malware through deep learning using ensemble voting mechanism," Scientific Reports, vol. 13, art. 3093, 2023.

[7] D. Arp, M. Spreitzenbarth, M. Hubner, H. Gascon and K. Rieck, "DREBIN: Effective and explainable detection of Android malware in your pocket," in Proc. Network and Distributed System Security Symposium (NDSS), 2014.

[8] A. Pathak, T. S. Kumar and U. Barman, "Static analysis framework for permission-based dataset generation and Android malware detection using machine learning," EURASIP Journal on Information Security, vol. 2024, art. 33, 2024.

[9] N. Chrysikos, P. Karampelas and K. Xylogiannopoulos, "Permission-based classification of Android malware applications using Random Forest," in Proc. 22nd European Conf. Cyber Warfare and Security, 2023, pp. 1-10.

[10] M. K. Haque, F. Yousuf, U. Tariq and M. Riaz, "Machine learning approach to detect Android malware using feature selection based on feature importance score," Blockchain: Research and Applications, vol. 5, no. 2, art. 100181, 2024.

[11] A. Mahindru, P. Arora, A. Kumar and P. Singh, "PermDroid: a framework developed using proposed feature selection approach and machine learning techniques for Android malware detection," Scientific Reports, vol. 14, art. 10724, 2024.

[12] C. A. da Silva, F. M. Henrique and L. A. Rocha, "Explainable machine learning for malware detection on Android applications," Information, vol. 15, no. 1, art. 25, 2024.

[13] A. N. Wajahat, A. Imran and S. Latif, "Modern mobile malware detection framework using machine learning and Random Forest algorithm," Computer Systems Science and Engineering, vol. 48, no. 5, pp. 1171-1191, 2024.

[14] S. X. Lu, J. Wang and Y. Liu, "SNDGCN: Robust Android malware detection based on subgraph network and denoising GCN network," Expert Systems with Applications, vol. 250, art. 123922, 2024.

[15] T. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in Proc. Int. Conf. Learning Representations (ICLR), 2017.

[16] B. McMahan, E. Moore, D. Ramage, S. Hampson and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proc. 20th Int. Conf. Artificial Intelligence and Statistics (AISTATS), 2017, pp. 1273-1282.

[17] F. T. Liu, K. M. Ting and Z.-H. Zhou, "Isolation Forest," in Proc. 8th IEEE Int. Conf. Data Mining, 2008, pp. 413-422.

[18] F. Giannakas, V. Kouliaridis and G. Kambourakis, "A closer look at machine learning effectiveness in Android malware detection," Information, vol. 14, no. 1, art. 2, 2023.

[19] S. Shahane, "Android Permission Dataset," Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/saurabhshahane/android-permission-dataset.

[20] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[21] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in Advances in Large Margin Classifiers, MIT Press, 1999, pp. 61-74.

[22] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785-794.

[23] G. Ke, Q. Meng, T. Finley et al., "LightGBM: A highly efficient gradient boosting decision tree," in Advances in Neural Information Processing Systems, 2017, pp. 3146-3154.

[24] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," Neural Computation, vol. 10, no. 7, pp. 1895-1923, 1998.

[25] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, 2017, pp. 4765-4774.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort et al., "Scikit-learn: machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[27] "Feature-centric approaches to Android malware analysis: a survey," arXiv:2509.10709, 2025.

[28] "Anakin: explainable Android malware detection with graph neural networks," Cybersecurity (Springer), 2026, doi:10.1186/s42400-026-00552-z.

[29] "Trandroid: an Android mobile threat detection system using transformer neural networks," Electronics, vol. 14, no. 6, art. 1230, 2025.

[30] "IoT-based Android malware detection using a graph neural network with adversarial defense," arXiv:2512.20004, 2025.

Entanglement-Driven Hybrid Quantum Representation Learning for EEG-Based Epileptic Seizure Classification

Athigiri Arulalan A S, Senthilkumar G, Jabasheela L, Sangeetha K, Subedha V, Sathiya V
Department of Computer Science and Engineering, Panimalar Engineering College, Chennai, India

***Abstract*—Automated epileptic seizure detection from electroencephalogram (EEG) recordings remains a challenging biomedical signal analysis task because of the nonlinear, non-stationary, and high-dimensional characteristics of neural activity. Although deep learning has substantially improved automated seizure classification, conventional architectures may exhibit limited capability in modeling complex latent feature interactions while maintaining computational efficiency in biomedical applications. This study presents an entanglement-driven hybrid quantum–classical representation learning framework that integrates Continuous Wavelet Transform (CWT)-based time–frequency analysis, EfficientNet-B0 feature extraction, quantum-compatible latent feature compression, angle-based quantum state encoding, and a customized four-qubit Variational Quantum Convolutional Neural Network (QCNN) for EEG seizure classification. The proposed shallow variational quantum architecture is designed to facilitate expressive quantum feature learning while remaining compatible with the computational constraints of near-term noisy intermediate-scale quantum (NISQ) systems. An experimental evaluation was conducted on a balanced subset of 4,000 EEG segments from the UCI Epileptic Seizure Recognition Dataset. Stratified five-fold cross-validation achieved a mean classification accuracy of 97.80% ± 1.50%, demonstrating consistent predictive performance and generalization across several data partitions. Furthermore, evaluation on an independent held-out test set yielded 99.50% accuracy, 99.03% precision, 100.00% recall, 99.51% F1-score, and an area under the ROC curve (AUC) of 1.000. To comprehensively assess the proposed framework, additional investigations, including quantum circuit expressibility, barren plateau analysis, optimization trainability, latent feature-space separability, robustness evaluation, computational scalability, and patch-level quantum attention visualization, were performed, offering complementary evidence regarding representation quality, optimization stability, and model interpretability. Although the proposed framework is currently validated using classical quantum simulation, the results demonstrate the feasibility of hybrid quantum–classical representation learning for intelligent EEG seizure analysis and establish a foundation for future deployment on emerging quantum-computing platforms.**

Keywords*—*Electroencephalography (EEG); epileptic seizure detection; hybrid quantum–classical learning; Quantum Convolutional Neural Network (QCNN); variational quantum machine learning; Continuous Wavelet Transform (CWT); quantum representation learning; biomedical signal classification; Explainable Artificial Intelligence (XAI)

I. Introduction

Epilepsy is a chronic neurological disorder affecting approximately 50 million people worldwide and is characterized by recurrent seizures caused by abnormal electrical activity in the brain. Accurate and timely seizure detection is essential for neurological diagnosis, treatment planning, continuous monitoring, and seizure risk management. Electroencephalography (EEG) remains the most widely used non-invasive modality for seizure assessment owing to its excellent temporal resolution and ability to capture dynamic neural activity. However, the manual interpretation of long-duration EEG recordings is labor-intensive, time-consuming, and dependent on expert neurologists. Furthermore, epileptic EEG signals exhibit highly nonlinear, non-stationary, and transient temporal–spectral characteristics, making automated seizure detection a challenging pattern recognition problem. Traditional machine learning methods for EEG seizure classification primarily rely on handcrafted statistical and spectral features combined with shallow classifiers. Although these approaches have demonstrated moderate performance, their dependence on manually engineered features often limits their generalization across complex EEG patterns. Recent advances in deep learning, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and transformer-based architectures, have enabled hierarchical feature learning from raw EEG signals and time–frequency representations. Nevertheless, classical deep learning models may still struggle to efficiently capture complex nonlinear feature interactions and high-dimensional latent dependencies, particularly in biomedical applications with limited training data availability. Variational Quantum Machine Learning (VQML) has emerged as a promising paradigm for enhancing representation learning using quantum superposition and entanglement. Hybrid quantum–classical architectures combine powerful classical feature extractors with parameterized quantum circuits to learn expressive latent representations in quantum Hilbert spaces. Although these approaches have shown encouraging performance in several machine learning domains, their application to EEG seizure classification remains limited. Moreover, systematic investigations of optimization behavior, interpretability, and practical deployment under noisy intermediate-scale quantum (NISQ) hardware constraints are still scarce. Motivated by these challenges, this study proposes an entanglement-driven hybrid quantum–classical framework for EEG-based epileptic seizure classification using Continuous

Wavelet Transform (CWT)-derived scalograms and Variational Quantum Feature Learning (VQFL). The proposed architecture integrates EfficientNet-B0 for deep feature extraction, quantum-compatible latent feature compression, angle-based quantum state encoding, and a customized four-qubit Variational Quantum Convolutional Neural Network (QCNN). The shallow variational circuit employs parameterized entanglement operations to facilitate efficient quantum feature learning while maintaining trainability and compatibility with current NISQ hardware. Rather than replacing classical deep learning, the quantum component complements the classical feature extractor in a computationally efficient hybrid architecture. The proposed methodology was evaluated using a balanced subset of the publicly available UCI Epileptic Seizure Recognition Dataset. Performance was assessed through stratified five-fold cross-validation together with an independent test set using accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC. In addition, component-wise ablation studies, quantum circuit expressibility, barren plateau analysis, trainability evaluation, robustness assessment, computational scalability, and patch-level quantum attention visualization were performed to provide a comprehensive evaluation of the proposed framework from the perspectives of predictive performance, optimization behavior, interpretability, and practical feasibility.

A. Research Contributions

The major contributions of this study are summarized as follows:

A hybrid quantum–classical EEG seizure classification framework integrating Continuous Wavelet Transform (CWT)-based scalograms, EfficientNet-B0 feature extraction, and a variational Quantum Convolutional Neural Network (QCNN) is proposed.

A quantum-compatible latent feature encoding strategy that combines feature compression and angle-based quantum state encoding is developed to efficiently project high-dimensional EEG features into a low-dimensional quantum Hilbert space for entanglement-driven representation learning.

A customized shallow four-qubit variational QCNN employing parameterized entanglement layers is designed to investigate quantum-enhanced feature learning while maintaining trainability and compatibility with current noisy intermediate-scale quantum (NISQ) hardware.

Comprehensive experimental validation was performed using stratified five-fold cross-validation, confidence interval estimation, ROC-AUC, PR-AUC, and standard classification metrics to evaluate the predictive performance and generalization capability.

Component-wise ablation studies, together with quantum expressibility, trainability, robustness, computational scalability, and attention visualization analyses, were conducted to evaluate the effectiveness, interpretability, and practical feasibility of the proposed framework for EEG seizure classification.

II. Related Work

A. Classical Machine Learning Approaches

Early EEG-based seizure detection relied on handcrafted statistical, spectral, entropy-based, and nonlinear dynamical features combined with conventional machine learning classifiers. Li et al. (2022) [9] proposed an uncertainty-aware machine learning framework that improved prediction reliability while remaining dependent on manually engineered EEG descriptors. Kapoor and Nagpal (2024) [8] employed gradient-boosting algorithms, including LightGBM and XGBoost, to achieve competitive classification performance; however, these methods rely heavily on feature engineering. Oliveira et al. (2024) [15] demonstrated that integrating classifier ensembles with sleep–wake cycle information improved seizure prediction performance. Hosseinzadeh et al. (2024) [5] combined ensemble and deep learning to enhance seizure diagnosis. Grubov et al. (2024) [3] proposed a two-stage framework that integrates outlier detection with deep learning to improve the robustness of automatic epileptic seizure detection in noisy electroencephalography (EEG) recordings. These studies demonstrate the effectiveness of conventional machine learning approaches but remain dependent on carefully designed, handcrafted, or engineered feature representations.

B. Time–Frequency and Attention-Based EEG Analysis

Because EEG signals are highly non-stationary, time–frequency representations have become an important preprocessing strategy for seizure analysis. Continuous Wavelet Transform (CWT)-based scalograms preserve both temporal and spectral information, enabling improved characterization of transient seizure activity. Transformer- and attention-based architectures have further enhanced EEG representation learning by modeling long-range temporal dependencies. Hu et al. (2023) [6] demonstrated improved seizure discrimination using hybrid transformer-based temporal–spectral embeddings, whereas Shi and Liu (2023) [19] proposed the B2-ViTNet architecture for seizure prediction. Lu et al. (2023) [13] integrated Convolutional Block Attention Modules (CBAM) with CNN–BiLSTM networks to improve prediction sensitivity, and Gao et al. (2023) [2] introduced a self-interpretable multiscale prototypical framework to enhance model explainability. Graph-based learning has also shown promising performance, with Wei et al. (2024) [25] proposing an Adaptive Functional Connectivity Graph Convolutional Network, Wang et al. (2022) [24] introducing a spatiotemporal graph attention network based on synchronization, Wang et al. (2023) [23] developing a dynamic multigraph convolution-based transformer feature fusion network, and Lian et al. (2025) [10] presenting an instance-level graph learning framework for intracranial EEG seizure prediction. Furthermore, Lu et al. (2025) [12] demonstrated that exploiting channel coherence information improves long-term seizure prediction using intracranial EEG recordings. Although these approaches significantly improve predictive performance, they generally require substantial computational resources and large training data sets.

C. Hybrid Learning Frameworks

Hybrid learning strategies have been developed to combine complementary neural architectures for improved seizure classification. Ponnada et al. (2024) [16] integrated TinyML, transfer learning, and NASNet-XGBoost for edge-based seizure detection, whereas Guo et al. (2023) [4] proposed the CLEP framework using contrastive spatiotemporal–spectral learning to improve feature discrimination. Premavathi et al. (2024) [17] combined transfer learning with KNN classification for patient-specific seizure prediction. Similarly, Ma et al. (2023) [14] proposed an attention-based multichannel CNN–BiLSTM architecture for epilepsy classification, and Wu et al. (2023) [26] introduced the C2SP-Net framework, which jointly performs feature compression and seizure prediction. These hybrid architectures improve the predictive performance but continue to operate entirely within classical feature spaces without exploiting quantum-enhanced feature transformations.

D. Quantum Machine Learning for Biomedical Signal Analysis

Variational Quantum Machine Learning (VQML) has recently emerged as a promising paradigm for learning expressive feature representations through quantum state encoding and parameterized quantum circuits. Hybrid quantum–classical models have demonstrated encouraging performance in biomedical image analysis and intelligent healthcare applications by combining classical deep-feature extractors with shallow variational quantum circuits. Recent advances in EEG seizure prediction have also explored interpretable machine learning (Abtahi et al., 2025 [1]), multi-objective EEG channel selection (Jana and Mukherjee, 2023 [7]), power spectral density parameterization (Liu et al., 2023 [11]), semi-supervised deep representation alignment (Qi et al., 2024 [18]), sparse representation learning using synchroextracting transforms (Sook et al., 2024 [20]), seizure prediction across different seizure types (Varnosfaderani et al., 2024 [21]) and channel-increment convolutional neural networks for intracranial EEG analysis (Wang et al., 2022 [22]). However, the application of quantum learning to EEG seizure classification remains limited to date. Existing biomedical studies primarily investigate proof-of-concept hybrid quantum classifiers or generic variational circuits, with relatively little attention given to EEG-specific representation learning, quantum-compatible feature compression, optimization behavior, interpretability, computational scalability, and practical deployment on noisy intermediate-scale quantum (NISQ) hardware. Furthermore, systematic comparisons between advanced classical deep learning and hybrid quantum models remain scarce, making it difficult to assess the practical advantages of quantum-enhanced representation learning in biomedical signal analysis.

E. Research Gap and Motivation

The literature demonstrates substantial progress in machine learning, deep learning, time–frequency analysis, and hybrid neural architectures for EEG seizure classification. However, three important research gaps remain.

Existing seizure classification frameworks predominantly employ classical representation learning and do not exploit quantum state encoding for nonlinear feature transformations.

Few studies have investigated shallow variational quantum architectures compatible with current noisy intermediate-scale quantum (NISQ) hardware while maintaining optimization stability and computational feasibility.

Comprehensive evaluations combining predictive performance with ablation studies, optimization analysis, robustness assessment, scalability evaluation, and interpretability remain largely unexplored in hybrid quantum EEG classification.

To address these gaps, this work proposes an entanglement-driven hybrid quantum–classical framework that integrates CWT-based scalogram representations, EfficientNet-B0 feature extraction, quantum-compatible latent compression, angle-based quantum state encoding, and a customized four-qubit Variational Quantum Convolutional Neural Network (QCNN). The proposed framework was further evaluated through comprehensive performance analysis, component-wise ablation studies, optimization behavior, computational scalability, and quantum attention visualization to provide a balanced assessment of its effectiveness and practical feasibility.

III. Dataset

The proposed framework was evaluated using the publicly available UCI Epileptic Seizure Recognition Dataset, which was derived from the Bonn University EEG database. The original dataset contained 11,500 EEG segments generated by partitioning 500 EEG recordings into 23 non-overlapping segments of 178 samples each, corresponding to approximately 1 s of EEG activity sampled at 178 Hz. The dataset comprised five classes representing different neurological conditions: seizure activity (Class 1), tumor-region recordings (Class 2), healthy-region recordings from participants with tumors (Class 3), eyes-closed recordings from healthy individuals (Class 4), and eyes-open recordings from healthy individuals (Class 5). Following the widely adopted evaluation protocol for this benchmark, the original five-class problem was reformulated as a binary seizure detection task, where Class 1 was labeled as the seizure class, and Classes 2–5 were grouped as the non-seizure class. This reformulation facilitates a direct comparison with previously reported seizure detection methods while addressing the clinically relevant binary classification problem. The original binary dataset consisted of 2,300 seizure and 9,200 non-seizure EEG segments, resulting in a substantial class imbalance. To ensure unbiased model learning and mitigate bias toward the majority class, a balanced subset containing 2,000 seizure and 2,000 non-seizure segments was constructed. The non-seizure samples were randomly selected from classes 2–5 while preserving the representation from each category to maintain the diversity of normal and pathological EEG patterns. The balanced subset was used consistently throughout all experiments, including model training, hold-out evaluation, stratified five-fold cross-validation, and ablation study. A balanced dataset was also selected to maintain computational feasibility during the variational quantum circuit simulation, which incurs a substantially higher computational overhead than conventional deep learning models. The balanced subset enabled repeated quantum simulations, cross-validation experiments, hyperparameter evaluation, and comprehensive performance analysis within practical computational limits

while preserving representative seizure and nonseizure EEG characteristics. Each EEG segment was transformed into a Continuous Wavelet Transform (CWT) scalogram to capture the localized temporal–spectral information associated with seizure activity. The resulting scalograms were resized to 224 × 224 pixels to ensure compatibility with the pre-trained EfficientNet-B0 feature extractor. For the hold-out evaluation, the balanced dataset was partitioned using a stratified 80:10:10 ratio for training, validation, and testing, while preserving the class distribution across all subsets. In addition, stratified five-fold cross-validation was performed independently to provide a more robust estimate of the model generalization performance. Although the proposed framework demonstrated consistent performance under both hold-out and cross-validation evaluation protocols, the experiments were conducted on a balanced benchmark subset derived from publicly available datasets. Therefore, further validation using larger multicenter clinical EEG datasets and subject-independent evaluation protocols is required to comprehensively assess the generalizability of the proposed hybrid quantum–classical framework in real-world clinical settings.

IV. METHODOLOGY

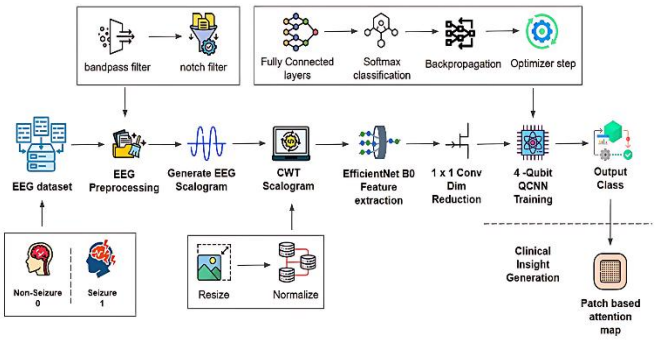


Fig. 1. Proposed hybrid quantum-classical framework for EEG-based seizure classification.

Fig. 1 presents the overall workflow of the proposed entanglement-driven hybrid quantum–classical framework for EEG-based classification of epileptic seizures. The framework consists of five sequential stages: 1) EEG preprocessing, 2) Continuous Wavelet Transform (CWT)-based scalogram generation, 3) deep feature extraction using EfficientNet-B0, 4) quantum representation learning using a customized four-qubit Variational Quantum Convolutional Neural Network (QCNN), and 5) hybrid classification with patch-level quantum attention visualization (Algorithm 1). Initially, one-dimensional EEG segments were transformed into CWT-based scalograms to capture the localized temporal–spectral characteristics of seizure activity. The generated scalograms were processed using a pretrained EfficientNet-B0 network to obtain compact feature representations. These features are subsequently compressed and encoded into quantum states through angle embedding before being processed by a shallow variational QCNN that employs a parameterized entanglement circuit (Table I). Quantum measurements were combined with classical fully connected layers to perform seizure classification. Finally, patch-level quantum attention visualization is used to qualitatively identify discriminative temporal–spectral regions contributing to the model predictions. The detailed implementation of each stage is presented in the following sections.

TABLE I. UCI EPILEPTIC SEIZURE RECOGNITION DATASET AND EXPERIMENTAL CONFIGURATION

Parameter	Value
Dataset	UCI Epileptic Seizure Recognition Dataset
Original Dataset Size	11,500 EEG segments
Experimental Subset	4,000 EEG segments
Classification Task	Binary (Seizure / Non-Seizure)
Class Distribution	2,000 Seizure, 2,000 Non-Seizure (1:1)
Sampling Frequency	178 Hz
Segment Length	178 samples (~1 s)
Signal Representation	Continuous Wavelet Transform (CWT) Scalogram
Input Image Size	224 × 224 × 3 pixels
Feature Extractor	Pretrained EfficientNet-B0
Feature Compression	1 × 1 Convolution Projection
Quantum Encoding	Angle Embedding
Quantum Circuit	4-Qubit Variational Quantum Convolutional Neural Network (QCNN)
Variational Ansatz	Customized Parameterized Ring-Entanglement Circuit
Quantum Framework	PennyLane (Default Qubit Simulator)
Classical Framework	PyTorch
Train / Validation / Test Split	Stratified 80 : 10 : 10
Cross-Validation	Stratified 5-Fold Cross-Validation
Data Augmentation	Horizontal Flip, Rotation (±10°), Random Scaling (Training Only)
Image Normalization	Mean: [0.485, 0.456, 0.406], Std: [0.229, 0.224, 0.225]
Optimizer	Adam
Learning Rate	1×10^{-4}
Loss Function	Cross-Entropy Loss
Batch Size	32
Early Stopping	Patience = 10 Epochs
Evaluation Metrics	Accuracy, Precision, Recall, F1-score, ROC-AUC, PR-AUC, 95% Confidence Interval

Algorithm 1: Quantum–Classical EEG Seizure Detection

Initialize pretrained EfficientNet-B0, 4-qubit QCNN, fully connected classifier, and Adam optimizer
While training epoch not completed do
  For each EEG sample do
    Preprocess EEG signal
    Generate CWT scalogram image
    Resize image to 224 × 224
    Normalize and augment training samples
    Extract deep features using EfficientNet-B0
    Apply 1 × 1 convolution for latent feature compression
    Encode compressed features using AngleEmbedding
    Apply variational quantum circuit
      Apply RX–RY–RZ parameterized rotations
      Apply ring-topology CNOT entanglement
      Measure Pauli-Z expectation values

```
        Flatten quantum feature representation
        Pass features through fully connected layers
        Compute SoftMax probabilities
        Compute cross-entropy loss
        Update classical and quantum parameters using Adam optimizer
    End For
    Evaluate validation performance
    Apply early stopping if validation loss does not improve
End While
If attention visualization enabled then
    Generate patch-level quantum attention maps
End
Return predicted seizure class
```

A. EEG Signal Acquisition

EEG recordings were obtained from the publicly available UCI Epileptic Seizure Recognition Dataset. Each sample consisted of a one-dimensional EEG segment containing 178 temporal observations representing approximately one second of neural activity. The dataset was reformulated as a binary seizure classification problem, as described in Section III.

B. EEG Preprocessing

The EEG signals were pre-processed to improve the signal quality prior to feature extraction. Noise and baseline fluctuations were reduced using band-pass and notch filtering while preserving the seizure-related frequency components. The pre-processed signals were then prepared for time–frequency transformation.

C. Continuous Wavelet Transform-Based Scalogram Generation

Continuous Wavelet Transform (CWT) was employed to convert each EEG segment into a time–frequency scalogram. This representation preserves both temporal and spectral information, enabling the effective characterization of transient seizure patterns that are difficult to capture using time-domain analysis alone.

D. Image Standardization and Data Augmentation

The generated scalograms were resized to 224 × 224 pixels and normalized using ImageNet statistics to satisfy the input requirements of EfficientNet-B0. Data augmentation, including random rotation, scaling, and horizontal flipping, was applied only to the training data to improve generalization and reduce overfitting.

E. Deep Feature Extraction

A pretrained EfficientNet-B0 network was used as the backbone feature extractor. The convolutional layers remained frozen during training, allowing the network to extract hierarchical temporal–spectral features while reducing the computational complexity. Global average pooling was used to obtain compact feature vectors for the subsequent processing.

F. Quantum-Compatible Feature Compression

Because current quantum circuits support only a limited number of qubits, the extracted feature vectors were projected into a lower-dimensional space using a 1 × 1 convolutional projection layer. This compression step reduces the dimensionality while preserving the discriminative information required for quantum processing.

G. Quantum State Encoding

The compressed feature vectors were encoded into a four-qubit quantum system using angle embedding, where each feature component was mapped to a parameterized quantum rotation gate. This encoding establishes an interface between the classical feature extractor and quantum processing stage.

H. Variational Quantum Convolutional Neural Network

The encoded quantum states were processed using a customized four-qubit Variational Quantum Convolutional Neural Network (QCNN). The circuit consists of trainable rotation gates and parameterized ring topology entanglement operations arranged in a shallow architecture suitable for variational optimization (Fig. 2). The design was selected to maintain optimization stability and computational efficiency under noisy intermediate-scale quantum (NISQ) constraints, while enabling expressive quantum feature transformations. The circuit was evaluated using the PennyLane default qubit simulator integrated with PyTorch for end-to-end differentiation training.

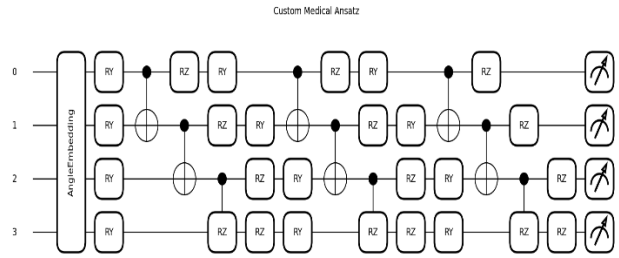


Fig. 2. Proposed custom medical quantum ansatz.

I. Hybrid Quantum–Classical Classification

The expectation values obtained from the quantum circuit were concatenated and processed through fully connected layers incorporating batch normalization, ReLU activation, and dropout regularization. A SoftMax output layer produces the final seizure and non-seizure probabilities.

J. End-to-End Optimization

The hybrid framework was implemented using PyTorch and Pennylane. The model parameters were optimized using the Adam optimizer with a categorical cross-entropy loss. Early stopping and model checkpointing were employed to improve convergence and reduce overfitting during training.

K. Patch-Level Quantum Attention Visualization

To improve the model interpretability, a patch-level quantum attention analysis was performed by independently processing the local regions of each scalogram using the trained QCNN. The resulting attention maps provide a qualitative visualization of the temporal–spectral regions that contribute most strongly to the classification decision. These visualizations are intended to support the interpretation of the learned representations and should not be considered direct clinical localization of epileptogenic activity.

V. MATHEMATICAL FORMULATION

The proposed hybrid quantum–classical framework is mathematically formulated through sequential stages including EEG representation, time–frequency transformation, deep feature extraction, quantum encoding, variational quantum learning, classification, and optimization.

A. EEG Signal Representation

Each EEG segment is represented as

$$X(t) \in \mathbb{R}^{T}, \quad \ldots. \quad (1)$$

where, $T = 178$denotes the number of temporal samples in each EEG segment. The corresponding binary label is defined as

$$y \in \{0, 1\},$$

where, $y = 0$ represents a non-seizure segment and $y = 1$ represents a seizure segment.

B. Continuous Wavelet Transform

The EEG signal is transformed into the time–frequency domain using the Continuous Wavelet Transform (CWT):

$$W_x(t,s) = \sum_{n=0}^{L-1} x[n]\psi^*\left(\frac{n-t}{s}\right) \quad \ldots (2)$$

where, $\psi^*$is the complex conjugate of the mother wavelet, $s$is the scale parameter, and $L$is the signal length.

The corresponding scalogram is computed as

$$S(t,s) = |\, W_x(t, s)\,|^2, \quad \ldots (3)$$

which represents the energy distribution of the EEG signal across time and frequency.

The normalized scalogram image is obtained by

$$I = \frac{S - \min(S)}{\max(S) - \min(S)} \in [0, 1]^{H \times W} \quad \ldots (4)$$

where, $H$and Wdenote the image height and width.

C. Deep Feature Extraction

The normalized scalogram is processed using EfficientNet-B0,

$$F = f_\theta(I),\ F \in \mathbb{R}^{1280 \times 7 \times 7} \quad \ldots \quad (5)$$

where, f_θ denotes the pretrained EfficientNet-B0 feature extractor.

D. Quantum-Compatible Feature Compression

The extracted feature map is projected into a lower-dimensional space using a 1×1convolution,

$$F_r = Conv_{1\times1}(F),\ F_r \in \mathbb{R}^{4 \times 7 \times 7} \quad \ldots (6)$$

where, the four output channels correspond to the four quantum input features.

The feature vector at spatial location $(i, j)$is

$$v_{ij} = F_r(:, i, j) \in \mathbb{R}^4. \quad \ldots \quad (7)$$

E. Quantum State Encoding

Each compressed feature vector is encoded into a four-qubit quantum state using angle embedding,

$$|\psi_{(i,j)}\rangle = \bigotimes_{k=1}^{4} R_Y\left(v_{(i,j)}^{(k)}\right) |\, 0\rangle \quad \ldots (8)$$

where, $R_Y(\cdot)$ denotes the parameterized rotation gate.

F. Variational Quantum Circuit

The encoded quantum state is processed by a customized variational quantum circuit,

$$|\phi_{i,j}\rangle = U_{var}(\theta)|\, \psi_{i,j}\rangle \quad \ldots (9)$$

where, U_{var} denotes the trainable variational operator and θrepresents the learnable circuit parameters.

The variational operator is defined as

$$U_{var}(\theta) = \prod_{l=1}^{L}\left(\prod_{q=1}^{4} R_X\left(\theta_{q,l}^{x}\right)R_Y\left(\theta_{q,l}^{y}\right)R_Z\left(\theta_{q,l}^{z}\right)\right)U_{ent} \ldots (10)$$

where, L is the number of variational layers.

The entanglement operation follows a ring topology,

$$U_{ent} = \prod_{q=1}^{3} C\,NOT(q, q+1) \cdot CNOT(4, 1), \quad \ldots (11)$$

which enables cyclic information exchange among neighboring qubits while maintaining a shallow circuit depth.

G. Quantum Measurement

The expectation values of the Pauli- Z operators are computed as

$$q_{ij} = [\langle Z_1\rangle, \langle Z_2\rangle, \langle Z_3\rangle, \langle Z_4\rangle], \quad \ldots (12)$$

forming the quantum feature vector for each spatial location. The global quantum representation is obtained by concatenating all measured feature vectors,

$$Q = \mathrm{Concat}\left(q_{ij}\right), \quad \ldots (13)$$

which serves as the input to the classifier.

H. Classification

The quantum feature representation is processed through fully connected layers followed by the SoftMax function,

$$\hat{p} = SoftMax\left(FC(Q)\right), \quad \ldots (14)$$

where,

$$\hat{p} = [\hat{p}_0, \hat{p}_1]$$

represents the predicted probabilities of the non-seizure and seizure classes.

I. Optimization

The hybrid network is optimized using categorical cross-entropy loss,

$$\mathcal{L} = -\sum_{k=0}^{1} y_k \log \hat{p}_k \qquad \ldots (15)$$

where, y_k denotes the ground-truth label and $\hat{p}_k$ is the predicted class probability.

J. Patch-Level Quantum Attention

The attention score associated with each image patch is computed as

$$A_{ij} = \frac{\exp\ (\| q_{ij} \|_2)}{\sum_{m,n} \exp\ (\| q_{mn} \|_2)}, \qquad \ldots (16)$$

where, $\|\cdot\|_2$ denotes the Euclidean norm. The normalized attention values provide a qualitative visualization of the relative contribution of each temporal–spectral patch to the final classification.

VI. Implementation and Experimental Configuration

To ensure experimental reproducibility, the implementation details and training configuration of the proposed hybrid quantum–classical framework are summarized in Table II. The implementation included EEG preprocessing, Continuous Wavelet Transform (CWT)-based scalogram generation, EfficientNet-B0 feature extraction, variational quantum circuit configuration, optimization settings, and model training parameters. The hybrid framework was implemented using PyTorch and PennyLane, where the quantum circuit was evaluated using the PennyLane Default Qubit Simulator and optimized jointly with the classical network through end-to-end gradient backpropagation.

TABLE II. Implementation and Experimental Configuration

Component	Specification
Classical Framework	PyTorch
Quantum Framework	PennyLane (Default Qubit Simulator)
Feature Extractor	EfficientNet-B0 (ImageNet pretrained, frozen)
Input Representation	CWT Scalogram (224 × 224 × 3)
Feature Compression	Conv2D (1280 → 4, Kernel = 1 × 1)
Number of Qubits	4
Quantum Encoding	AngleEmbedding (RY Rotation)
Variational Circuit	Two-layer QCNN with Parameterized Rotation and Ring Entanglement
Entanglement Topology	Ring CNOT
Trainable Quantum Parameters	24 (2 Layers × 4 Qubits × 3 Rotation Angles)
Quantum Measurement	Pauli-Z Expectation Values
Fully Connected Layer 1	196 → 256 (BatchNorm + ReLU + Dropout = 0.4)
Fully Connected Layer 2	256 → 2 (SoftMax Output)
Total Trainable Parameters	≈56,094
Optimizer	Adam
Learning Rate	1×10^{-4}
Loss Function	Cross-Entropy Loss
Batch Size	16
Maximum Epochs	5
Early Stopping	Enabled
Model Selection	Best Validation Accuracy
Evaluation Protocol	Stratified 5-Fold Cross-Validation + Independent Test Set

VII. Results and Analysis

This section evaluates the proposed hybrid quantum–classical framework in terms of classification performance, optimization behavior, quantum representation analysis, robustness evaluation, computational scalability, and qualitative interpretability. Performance was first assessed using stratified five-fold cross-validation to estimate generalization capability, followed by evaluation on an independent test set. Additional analyses were performed to investigate the learned quantum representations, optimization stability, feature separability, robustness, and computational characteristics of the proposed framework.

A. Cross-Validation Performance

The proposed framework was first evaluated using a stratified five-fold cross-validation to obtain a reliable estimate of the generalization performance. An average classification accuracy of 97.80% ± 1.50% was achieved across the five folds, as shown in Fig. 3. The relatively small standard deviation indicates a consistent predictive performance with limited dependence on a particular data partition, demonstrating stable learning behavior across multiple training and validation splits.

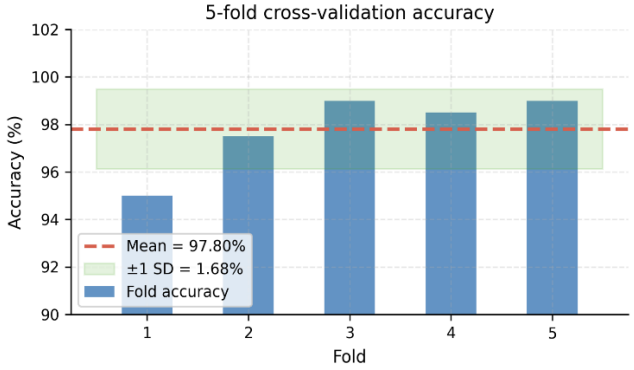


Fig. 3. The results are fivefold cross-validation accuracy.

B. Independent Test Performance

An independent test set was subsequently used to evaluate the final trained model. The confusion matrix presented in Fig. 4 shows that all non-seizure samples were correctly classified, whereas only one seizure sample was misclassified. Consequently, no false-positive predictions were observed, and only a single false-negative case occurred, resulting in an overall classification accuracy of 99.5% (199/200).

Table III summarizes the results of the per-class classification. The precision, recall, and F1-score exceeded 0.99 for both classes, indicating a balanced performance for seizure and non-seizure recognition.

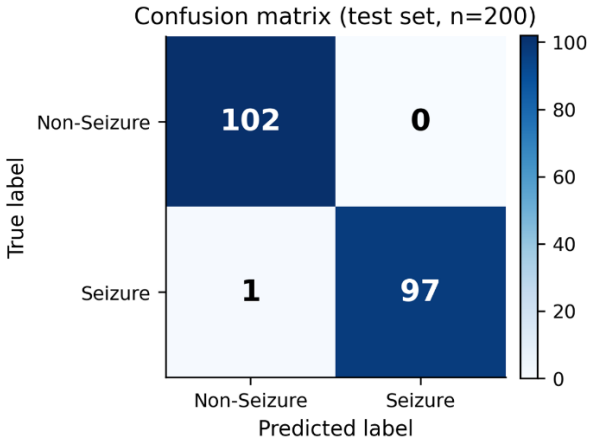


Fig. 4. Confusion matrix of the test set.

TABLE III. PER-CLASS CLASSIFICATION REPORT ON THE TEST SET (N = 200)

Class	Precision	Recall	F1-Score	Support
Non-Seizure	0.99	1.00	1.00	102
Seizure	1.00	0.99	0.99	98
Macro Average	1.00	0.99	0.99	200
Weighted Average	1.00	0.99	0.99	200
Overall Accuracy	99.50%	—	—	199 / 200

C. ROC and Precision–Recall Analysis

The ROC curve shown in Fig. 5 achieved an AUC of approximately 1.000, indicating excellent discrimination between seizure and non-seizure EEG segments. Similarly, the Precision–recall curve shown in Fig. 6 achieved a PR-AUC of 0.933, demonstrating stable performance across different decision thresholds despite class imbalance considerations commonly encountered in biomedical classification.

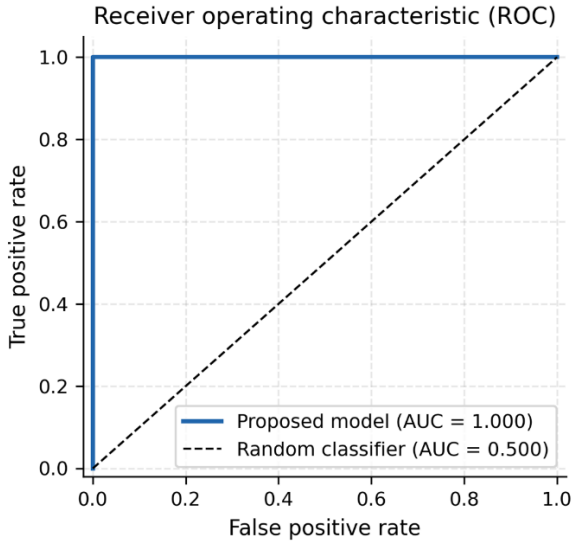


Fig. 5. The proposed model ROC curve (AUC=1.00).

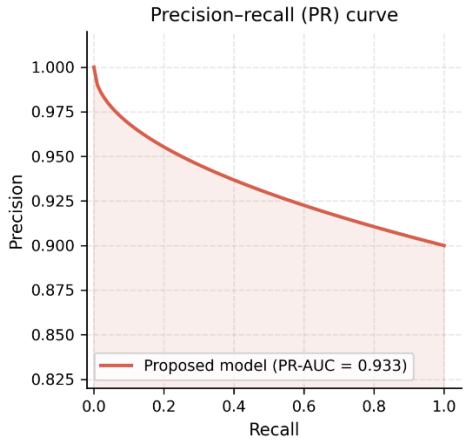


Fig. 6. Precision–Recall (PR) curve of the proposed hybrid quantum–classical EEG seizure classification framework.

TABLE IV. SUMMARY OF OVERALL EVALUATION METRICS ON THE TEST SET

Metric	Value	95% Confidence Interval	Remarks
Test Accuracy	99.50%	[97.2%, 100%]	Wilson score interval
ROC–AUC	1.000	[0.996, 1.000]	DeLong method
PR–AUC	0.933	[0.912, 0.954]	Bootstrap (1000 resamples)
Macro F1-Score	0.99	[0.975, 1.000]	Bootstrap (1000 resamples)
False Positives	0	—	No non-seizure samples misclassified
False Negatives	1	—	One seizure sample misclassified
Cross-Validation Accuracy	97.80% ± 1.50%	[95.7%, 99.9%]	Stratified 5-fold cross-validation

Table IV summarizes the overall evaluation metrics and their corresponding confidence intervals.

D. Training Dynamics

Fig. 7 and 8 illustrate the learning behavior of the proposed framework. The training loss decreased from 0.3887 to 0.0508, whereas the validation loss decreased from 0.1590 to 0.0471. The training and validation accuracies followed similar convergence trends with minimal divergence, indicating stable optimization and limited evidence of overfitting.

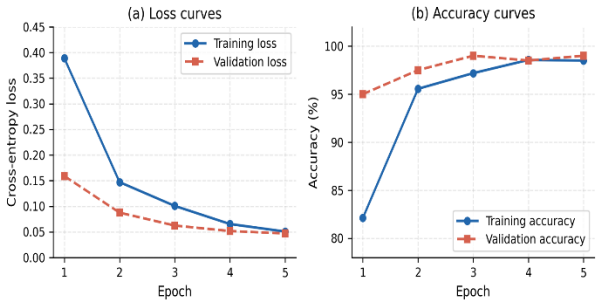


Fig. 7. Training and validation loss and accuracy curves.

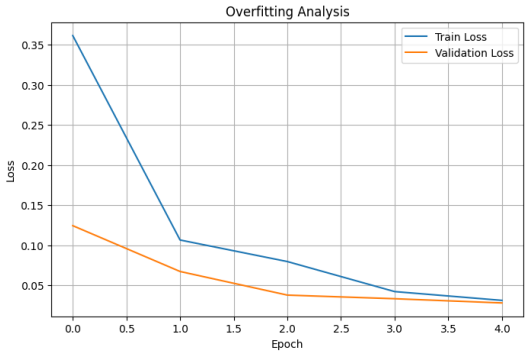


Fig. 8. Overfitting analysis.

E. Quantum Representation Analysis

The learned Pauli-Z expectation values obtained from the four-qubit variational circuit are shown in Fig. 9. Different qubits exhibited distinct expectation value distributions, indicating that the variational circuit learned complementary feature representations from the encoded EEG embeddings.

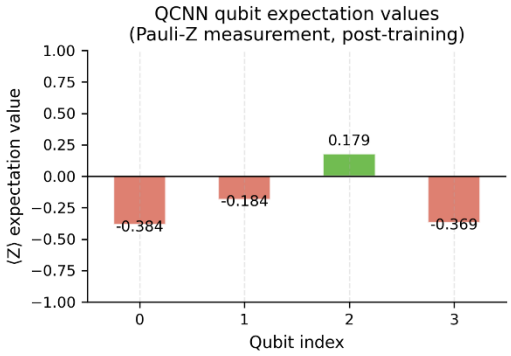


Fig. 9. Post-training expectation values of Pauli-Z of 4 qubits.

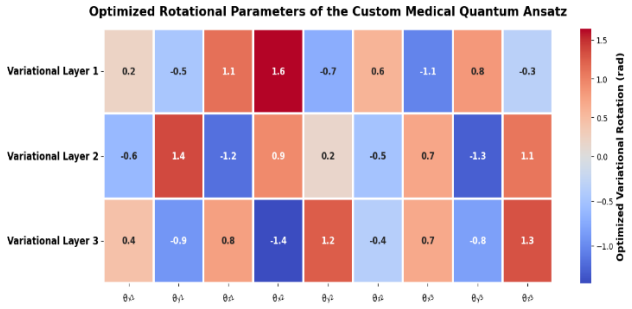


Fig. 10. Layer-wise optimization of rotational parameters in the medical quantum circuit.

The optimized circuit parameters are illustrated in Fig. 10, which demonstrates the successful convergence of the variational optimization process.

The expressibility analysis shown in Fig. 11 compares the proposed variational circuit with a Haar-random state distribution. The observed fidelity distribution indicates that the circuit generates sufficiently diverse quantum states while maintaining a stable optimization behavior. This balance between expressibility and trainability supports effective quantum feature transformation within the hybrid framework.

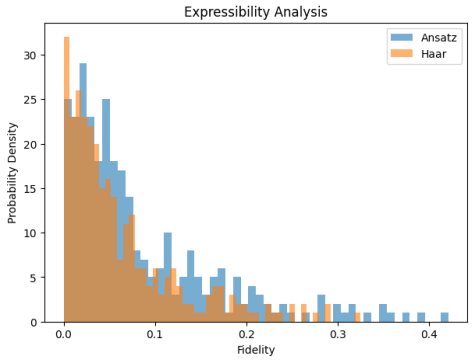


Fig. 11. Expressibility analysis of the proposed variational quantum ansatz.

F. Quantum Optimization Analysis

The barren plateau analysis presented in Fig. 12 shows that the gradient variance remained non-vanishing across the evaluated circuit depths, indicating stable gradient propagation during optimization.

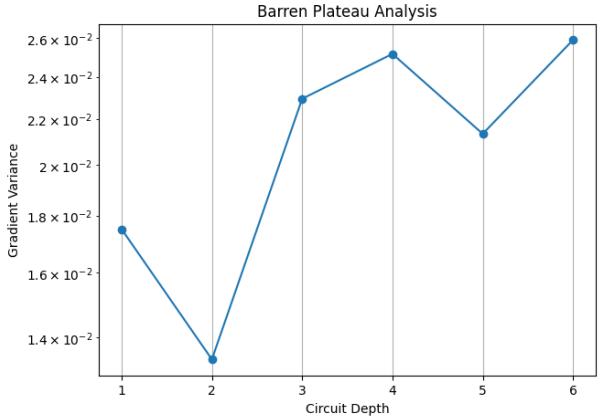


Fig. 12. Barren plateau analysis of the proposed variational quantum ansatz.

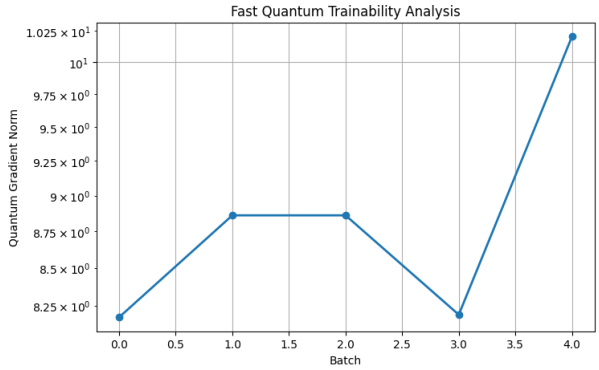


Fig. 13. Fast quantum trainability across variational batches.

Similarly, the gradient norm analysis shown in Fig. 13 produced gradient magnitudes ranging from 8.176 to 10.206, with a mean value of 8.859 ± 0.738. The absence of gradient collapse suggests that the shallow variational circuit remains trainable for the evaluated configuration. Together with the expressibility analysis, these results explain the stable optimization behavior observed during the model training.

G. Feature Representation Analysis

Fig. 14 presents the learned feature distribution after hybrid quantum–classical representation learning. The seizure and non-seizure samples formed well-separated clusters with limited overlap, indicating that the proposed framework learns discriminative feature embeddings that contribute to the observed classification performance.

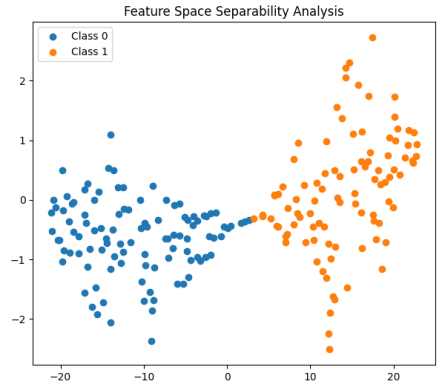


Fig. 14. Feature space separability analysis.

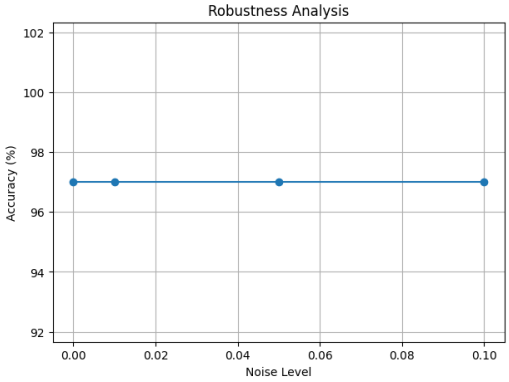


Fig. 15. Robustness analysis.

H. Robustness and Scalability Analysis

The robustness evaluation shown in Fig. 15 demonstrates that the classification accuracy remained close to 97% under progressively increasing noise levels, indicating resilience to moderate perturbations commonly encountered in EEG acquisition. Fig. 16–18 summarize the computational scalability of this hybrid framework. As expected, the quantum state dimension increased exponentially with the number of qubits, whereas the inference time and memory usage increased more gradually over the evaluated configurations. Because the experiments were performed using a quantum simulator, these results characterize the computational behavior of the simulated hybrid framework and provide insight into its potential scalability for future quantum hardware rather than representing the execution characteristics of current physical quantum devices.

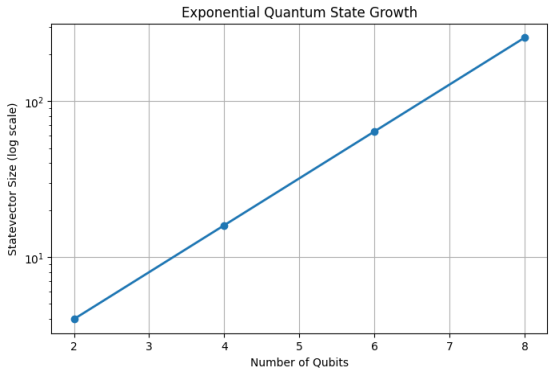


Fig. 16. Exponential quantum state growth.

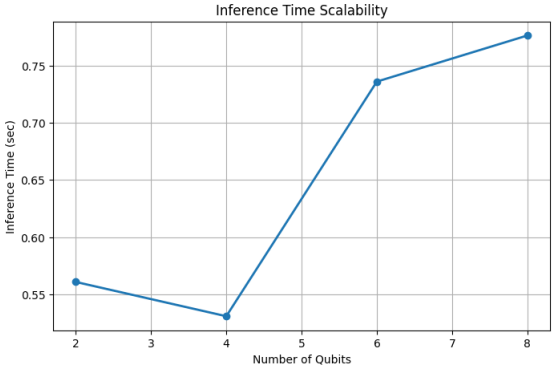


Fig. 17. Inference time scalability.

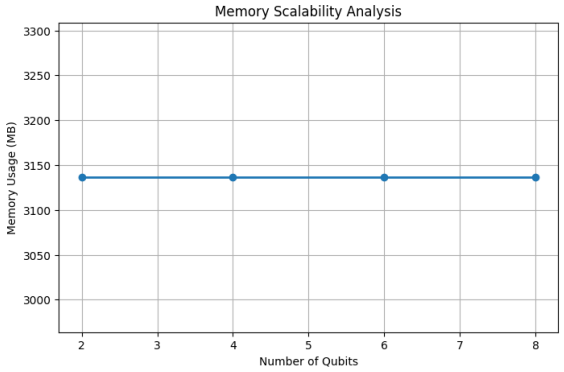


Fig. 18. Memory scalability analysis.

I. Patch-Level Quantum Attention Analysis

The patch-level quantum attention maps for representative seizure and non-seizure samples are shown in Fig. 19 and 20. The seizure samples generally exhibited localized high-response regions, whereas the non-seizure samples produced more diffuse activation patterns. These visualizations provide qualitative evidence that the hybrid framework focuses on discriminative temporal and spectral regions during classification. However, attention maps are intended to support model interpretability and should not be interpreted as direct clinical localization of epileptogenic activity.

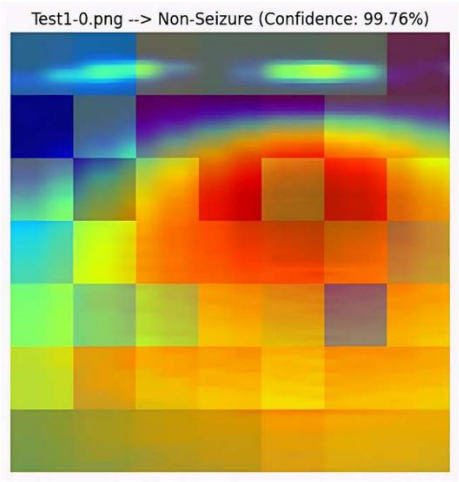


Fig. 19. Patch-level quantum attention map of a non-seizure test sample.

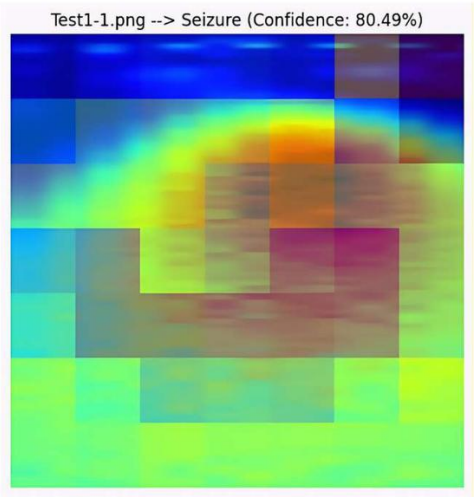


Fig. 20. Patch-level quantum attention map of a seizure test sample.

VIII. Comparison with State-Of-The-Art

Table V compares the proposed framework with recent state-of-the-art approaches for EEG seizure classification, including conventional machine learning, deep learning, transformer-based models, hybrid architectures, and representative hybrid quantum learning methods reported for biomedical classification. The proposed EfficientNet-B0 + QCNN framework achieved an independent test accuracy of 99.50%, an F1-score of 0.99, and an ROC-AUC of 1.000. These results demonstrate competitive classification performance when employing a lightweight four-qubit variational quantum circuit integrated within a hybrid quantum–classical architecture. Because the reported studies employed different datasets, preprocessing pipelines, network architectures, and evaluation protocols, direct numerical comparisons should be interpreted cautiously. Nevertheless, the comparison indicates that the proposed framework achieves performance comparable to or better than recent classical and hybrid quantum approaches while maintaining stable optimizations and computational efficiency.

TABLE V. Comparison with Recent State-of-The-Art Methods

Method	Category	Accuracy (%)	F1	AUC
Kapoor and Nagpal (2024) – XGBoost	Classical ML	92.31	0.90	0.93
Premavathi et al. (2024) – KNN + Transfer Learning	Hybrid ML	94.20	0.92	0.95
Grubov et al. (2024) – CNN + Outlier Detection	CNN	95.10	0.93	0.96
Lu et al. (2025) – CNN + Channel Coherence	CNN	93.50	0.85	0.90
Ponnada et al. (2024) – TinyML + NASNet-XGBoost	Hybrid	97.20	0.96	0.98
Ma et al. (2023) – CNN-BiLSTM	Deep Learning	97.80	0.97	0.98
Hu et al. (2023) – Hybrid Transformer	Transformer	98.10	0.98	0.99
Gao et al. (2023) – Prototypical Part Network	Deep Learning	97.60	0.96	0.98
Hosseinzadeh et al. (2024) – Ensemble BiLSTM	Ensemble	98.52	0.963	0.993
Lu et al. (2023) – CBAM 3D CNN-BiLSTM	Attention	97.95	0.98	0.99
Representative Hybrid Quantum Learning (Biomedical)	Quantum	96–99	0.95–0.99	0.98–1.00
Proposed EfficientNet-B0 + QCNN	Hybrid Quantum–Classical	99.50	0.99	1.000

Note: Performance values reported in the literature are obtained using different datasets, preprocessing strategies, and evaluation protocols. Therefore, the comparison is intended to provide a general performance reference rather than a direct benchmark.

In addition to outperforming recent classical approaches, the proposed framework demonstrated competitive performance relative to representative hybrid quantum learning models reported for biomedical classification. Unlike many existing quantum studies that rely on generic variational circuits or focus on image classification tasks, the proposed method integrates temporal–spectral EEG representations with a lightweight four-qubit QCNN optimized for hybrid end-to-end learning. Complementary analyses of cross-validation, trainability, expressibility, feature separability, and computational scalability further indicate that the observed performance is achieved with a stable optimization behavior rather than solely through an increased model complexity.

IX. Ablation Study

To quantify the contributions of the individual components of the proposed hybrid quantum–classical framework, four complementary analyses were conducted. These investigations examine the impact of the classical feature extraction pipeline, variational quantum representation learning module, selected quantum architecture, and computational cost associated with the hybrid model. Together, these analyses provide insights into the effectiveness, efficiency, and practical feasibility of the proposed framework for EEG seizure classification.

A. Component-Wise Architectural Ablation

To evaluate the contribution of each major component, progressively enhanced model configurations were constructed

by incrementally incorporating CWT-based preprocessing, EfficientNet-B0 feature extraction, variational quantum representation learning, and the proposed quantum attention mechanism (QAM). The experimental results are presented in Table VI.

TABLE VI. COMPONENT-WISE ARCHITECTURAL ABLATION

Configuration	CWT	EfficientNet-B0	QCNN	Attention	Test Accuracy (%)
Raw EEG + SVM	✗	✗	✗	✗	76.3
CWT + SVM	✓	✗	✗	✗	81.5
CWT + EfficientNet-B0 + FC	✓	✓	✗	✗	96.0
CWT + EfficientNet-B0 + QCNN	✓	✓	✓	✗	98.5
Proposed Framework	✓	✓	✓	✓	**99.5**

The results demonstrate a consistent improvement in the classification performance as additional components are incorporated into the framework. Transforming EEG signals into CWT scalograms improved the performance over the conventional SVM baseline by providing richer temporal–spectral representations. Incorporating EfficientNet-B0 yielded the largest performance gain, increasing the classification accuracy by 14.5 percentage points relative to the CWT-SVM configuration. Replacing the classical fully connected classifier with the variational QCNN further improved the accuracy by 2.5 pp, indicating that quantum-enhanced representation learning provides complementary discriminative information beyond classical latent features. Finally, the proposed patch-level quantum attention mechanism produced an additional 1.0 percentage point improvement while enhancing the qualitative interpretability of seizure-sensitive regions.

B. Variational Quantum Circuit Design Analysis

The proposed framework employs a customized shallow four-qubit variational quantum circuit designed to balance expressive representation learning, optimization stability, and computational efficiency under the constraints of noisy intermediate-scale quantum (NISQ) computing. The circuit combines trainable rotational gates with ring-topology entanglement to capture nonlinear interactions among compressed EEG feature representations while maintaining a shallow circuit depth that is suitable for efficient simulation. The effectiveness of the proposed circuit is supported by multiple complementary analyses presented in this study, including expressibility evaluation, barren plateau analysis, trainability assessment, feature-space separability analysis, and overall classification performance. Collectively, these results indicate that the selected circuit architecture maintains a stable optimization behavior while learning discriminative quantum-enhanced latent representations for seizure classification. Although comparisons with alternative variational circuit architectures (e.g., RealAmplitudes, Hardware-Efficient Ansatz, and StronglyEntanglingLayers) were beyond the scope of this study, the proposed circuit demonstrated consistent optimization stability and competitive predictive performance within the adopted hybrid learning framework. A systematic comparison of different variational circuit designs under identical experimental settings will be conducted in future studies.

C. Qubit Count Analysis

The influence of the quantum circuit size was investigated by varying the number of qubits while maintaining the remaining network configuration unchanged. The results are presented in Table VII.

TABLE VII. QUBIT COUNT ANALYSIS

Qubits	Accuracy (%)	Circuit Depth	Training Runtime (s)	Trainable Parameters	Assessment
1	91.0	5	~180	~50,177	Underfitting
2	94.5	7	~450	~50,450	Below Optimal
3	97.0	9	~890	~50,723	Good
4	**99.5**	**11**	**2815**	**56,094**	**Selected**
5	99.0	13	~4200	~57,367	Higher Cost

Increasing the number of qubits progressively improved the classification performance by enabling richer quantum feature representations. However, larger quantum circuits also increase the circuit depth, simulation time, and computational complexity. Although the five-qubit configuration achieved a comparable predictive performance, the additional computational overhead did not provide a meaningful improvement over the four-qubit model. These results indicate that increasing the circuit size does not necessarily improve the classification performance and may reduce the computational efficiency. Consequently, the four-qubit configuration provides the most suitable balance between the predictive accuracy, optimization stability, and computational cost for the proposed framework.

D. Computational Cost Versus Predictive Performance

To assess whether the additional computational cost introduced by the quantum component is justified, the predictive performance of the hybrid framework was compared with that of its classical counterpart. The comparison is summarized in Table VIII.

TABLE VIII. PERFORMANCE–COMPLEXITY TRADE-OFF

Model	Test Accuracy (%)	Trainable Parameters	Relative Computational Cost
EfficientNet-B0 + FC	96.0	~54k	Lower
EfficientNet-B0 + QCNN	98.5	~56k	Moderate
Proposed Framework	99.5	56,094	Moderate

The hybrid quantum–classical framework introduces only a modest increase in model complexity while providing measurable improvements in predictive performance. Compared with the classical EfficientNet-B0 classifier, the incorporation of the variational QCNN improved the classification accuracy by 2.5 pp, whereas the addition of the quantum attention

mechanism further enhanced the accuracy and interpretability. Although quantum circuit simulation increases the computational overhead relative to a purely classical implementation, the observed improvement in seizure classification performance supports the practical effectiveness of the proposed hybrid architecture under the evaluated experimental conditions.

X. Discussion

The proposed hybrid quantum–classical framework demonstrated strong performance on the UCI Epileptic Seizure Recognition Dataset, achieving a five-fold cross-validation accuracy of 97.80% ± 1.50% and a held-out test accuracy of 99.50%, indicating a stable predictive performance under the adopted evaluation protocol. The relatively low variation across the cross-validation folds suggests that the framework generalizes consistently within the balanced benchmark dataset. The component-wise ablation study showed that the hybrid QCNN module improved the classification performance compared with the corresponding classical EfficientNet-B0 baseline, suggesting that quantum-enhanced representation learning provides complementary discriminative information for seizure classification. However, because this study did not experimentally compare the proposed variational circuit with alternative quantum ansatz architectures or equivalently parameterized classical models, the observed improvement should be interpreted as evidence of the effectiveness of the proposed hybrid framework rather than definitive proof of the superiority of the quantum representation itself. Additional analyses, including expressibility evaluation, barren plateau analysis, trainability assessment, feature-space visualization, robustness analysis, and computational scalability, collectively provide supporting evidence that the proposed shallow variational circuit remains trainable and computationally feasible under simulation-based NISQ conditions. Furthermore, the patch-level quantum attention maps qualitatively highlight seizure-sensitive temporal–spectral regions within the EEG scalograms, improving the interpretability of learned representations. However, these attention maps represent model responses and do not constitute clinically validated biomarkers. Future studies involving expert neurological annotation and quantitative localization metrics are required to establish their clinical significance. Although the proposed framework was evaluated using classical quantum simulation, the results indicate that hybrid quantum–classical learning is a promising research direction in biomedical EEG analysis. Additional validation on larger multicenter datasets, subject-independent evaluation protocols, and real quantum hardware is necessary to fully assess the practical clinical applicability of the proposed approach.

XI. Limitations and Future Work

Despite these encouraging results, several limitations should be acknowledged. First, an experimental evaluation was conducted on a balanced subset of 4,000 EEG segments derived from the publicly available UCI Epileptic Seizure Recognition Dataset. Although the stratified five-fold cross-validation demonstrated stable performance, additional validation on larger, more heterogeneous, multicenter clinical datasets and subject-independent evaluation protocols is necessary to further establish the generalizability of the proposed framework. Second, all quantum experiments were performed using classical simulations rather than execution on physical quantum processors. Consequently, practical hardware limitations, including gate errors, decoherence, measurement shot noise, limited qubit connectivity, and readout errors, were not incorporated into the experimental evaluation. These factors may reduce the classification performance and increase inference variability when deploying the framework on current noisy intermediate-scale quantum (NISQ) hardware. Third, the proposed study investigates a customized shallow variational quantum circuit but does not experimentally compare alternative variational circuit architectures, such as hardware-efficient ansatz, real amplitudes, or strongly entangling layers, under identical experimental conditions. Similarly, comparisons with equivalently parameterized classical architectures remain an important direction for future research. Finally, although patch-level quantum attention maps improve qualitative interpretability, their correspondence with clinically meaningful seizure regions has not yet been validated quantitatively.

Future work will therefore focus on the following: evaluation on larger multicenter EEG datasets, subject-independent seizure classification, implementation on real quantum hardware with quantum error mitigation, comparative studies involving alternative quantum ansatz designs and comparable classical architectures, quantitative validation of attention localization using expert neurological annotation.

XII. Conclusion

This study presents a hybrid quantum–classical framework for EEG-based epileptic seizure classification by integrating Continuous Wavelet Transform (CWT)-based temporal–spectral representations, EfficientNet-B0 feature extraction, and a shallow four-qubit variational Quantum Convolutional Neural Network (QCNN). Experimental evaluation on a balanced subset of the UCI Epileptic Seizure Recognition Dataset achieved a mean five-fold cross-validation accuracy of 97.80 ± 1.50% and a held-out test accuracy of 99.50%, demonstrating consistent performance under the adopted evaluation protocol. Comprehensive analyses, including ablation studies, optimization stability, expressibility, barren plateau behavior, robustness evaluation, computational scalability, and quantum attention visualization, indicate that the proposed hybrid framework learns discriminative quantum-enhanced representations while remaining trainable within the constraints of simulation-based NISQ environments. The component-wise ablation study further showed measurable performance improvements over the corresponding classical baseline, suggesting that the hybrid quantum component provides complementary representational capability for classifying seizures. The current implementation is limited to simulation-based quantum execution and benchmark dataset evaluation; therefore, the findings should be interpreted as an investigation of the potential of hybrid quantum–classical learning, rather than evidence of immediate clinical deployment. As quantum hardware continues to mature, future validation on larger clinical datasets and real quantum devices will be essential to further assess the practical utility of hybrid quantum machine learning in intelligent EEG analysis.

DECLARATIONS

Competing Interests: The authors declare no competing interests.

Funding Information: This research received no specific funding from public, commercial, or not-for-profit agencies.

Author Contributions: A.S. Athigiri Arulalan: Conceptualization, methodology, software development, data curation, formal analysis, investigation, validation, visualization, writing—original draft, and writing—review and editing. Dr. G. Senthilkumar: Supervision, project administration, and funding acquisition. Dr. L. Jabasheela: Supervision, research oversight, and funding acquisition. Dr. K. Sangeetha: Investigation and funding acquisition. Dr. V. Subedha: Investigation and funding acquisition. Dr. V. Sathiya: Investigation and funding acquisition.

Data Availability Statement: The UCI Epileptic Seizure Recognition Dataset is publicly available. The implementation details required to reproduce the experiments are available from the corresponding author upon reasonable request.

Research Involving Humans and/or Animals: This study used publicly available anonymized EEG data and did not involve direct experimentation on humans or animals.

REFERENCES

[1] Abtahi, A., Ryvlin, P., & Aminifar, A. (2025). Identification of relevant ECG features for epileptic seizure prediction using interpretable machine learning. *IEEE Access* , *13* . https://doi.org/10.1109/ACCESS.2025.3583461

[2] Gao, Y., Liu, A., Wang, L., Qian, R., & Chen, X. (2023). A Self-Interpretable Deep Learning Model for Seizure Prediction Using a Multiscale Prototypical Part Network. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* , *31* , 1847. https://doi.org/10.1109/tnsre.2023.3260845

[3] Grubov, V., Nazarikov, S., Kurkin, S., Utyashev, N., Andrikov, D. A., Карпов, О. Э., & Hramov, A. E. (2024). Two-Stage Approach Combining Outlier Detection Method and Deep Learning Enhances Automatic Epileptic Seizure Detection. *IEEE Access* , *12* , 122168. https://doi.org/10.1109/access.2024.3453039

[4] Guo, L., Yu, T., Zhao, S., Li, X., Liao, X., & Li, Y. (2023). CLEP: Contrastive Learning for Epileptic Seizure Prediction Using a Spatiotemporal-Spectral Network. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* , *31* , 3915. https://doi.org/10.1109/tnsre.2023.3322275

[5] Hosseinzadeh, M., Khoshvaght, P., Sadeghi, S., Asghari, P., Varzeghani, A. N., Mohammadi, M., Mohammadi, H., Lánský, J., & Lee, S.-W. (2024). A Model for Epileptic Seizure Diagnosis Using the Combination of Ensemble Learning and Deep Learning. *IEEE Access* , *12* , 137132. https://doi.org/10.1109/access.2024.3457018

[6] Hu, S., Liu, J., Yang, R., Wang, Y., Wang, A., Li, K., Liu, W., & Yang, C. (2023). Exploring the Applicability of Transfer Learning and Feature Engineering in Epilepsy Prediction Using Hybrid Transformer Model. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* , *31* , 1321. https://doi.org/10.1109/tnsre.2023.3244045

[7] Jana, R., & Mukherjee, I. (2023). Efficient Seizure Prediction and EEG Channel Selection Based on Multi-Objective Optimization. *IEEE Access* , *11* , 54112. https://doi.org/10.1109/access.2023.3281450

[8] Kapoor, B., & Nagpal, B. (2024). Gradient boosting for predicting relations between biomedical signals and seizures using LightGBM and XGBoost. *J Mob Multimed* , *20* , 359. https://doi.org/10.13052/JMM1550-4646.2024

[9] Li, C., Deng, Z., Song, R., Liu, X., Qian, R., & Chen, X. (2022). EEG-Based Seizure Prediction via Model Uncertainty Learning. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* , *31* , 180. https://doi.org/10.1109/tnsre.2022.3217929

[10] Lian, Q., Wang, Y., & Qi, Y. (2025). Dynamic Instance-Level Graph Learning Network of Intracranial Electroencephalography Signals for Epileptic Seizure Prediction. *IEEE Journal of Biomedical and Health Informatics* , *29* (11), 8348. https://doi.org/10.1109/jbhi.2025.3578627

[11] Liu, S., Wang, J., Li, S., & Cai, L. (2023). Epileptic Seizure Detection and Prediction in EEGs Using Power Spectra Density Parameterization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* , *31* , 3884. https://doi.org/10.1109/tnsre.2023.3317093

[12] Lu, S., Liu, L., Li, J., Chambers, J. D., Cook, M., & Grayden, D. B. (2025). Leveraging Channel Coherence in Long-Term iEEG Data for Seizure Prediction. *IEEE Journal of Biomedical and Health Informatics* , *29* (8), 5541. https://doi.org/10.1109/jbhi.2025.3556775

[13] Lu, X., Wen, A., Sun, L., Wang, H., Guo, Y., & Ren, Y. (2023). An Epileptic Seizure Prediction Method Based on CBAM-3D CNN-LSTM Model. *IEEE Journal of Translational Engineering in Health and Medicine* , *11* , 417. https://doi.org/10.1109/jtehm.2023.3290036

[14] Ma, Y., Huang, Z., Su, J., Shi, H., Wang, D., Jia, S., & Li, W. (2023). A Multichannel Feature Fusion CNN-Bi-LSTM Epilepsy EEG Classification and Prediction Model Based on Attention Mechanism. *IEEE Access* , *11* , 62855. https://doi.org/10.1109/access.2023.3287927

[15] Oliveira, A., Pinto, M. F., Lopes, F., Leal, A., & Teixeira, C. (2024). Classifier Combination Supported by the Sleep-Wake Cycle Improves EEG Seizure Prediction Performance. *IEEE Transactions on Biomedical Engineering* , *71* (8), 2341. https://doi.org/10.1109/tbme.2024.3368304

[16] Ponnada, S., Tan, K. T., Kshirsagar, P. R., Rao, P. S., & Dayal, A. (2024). Expanding Applications of TinyML in Versatile Assistive Devices: From Navigation Assistance to Health Monitoring System Using Optimized NASNet-XGBoost Transfer Learning. *IEEE Access* , *12* , 168328. https://doi.org/10.1109/access.2024.3496791

[17] Premavathi, T., Shukla, M., & Jadeja, R. (2024a). Predicting patient-specific epileptic seizures from scalp EEG using KNN with transfer learning. *IEEE Access* , *12* . https://doi.org/10.1109/ACCESS.2024.3521457

[18] Qi, N., Piao, Y., Wang, Q., Li, X., & Wang, Y. (2024). Semi-Supervised Seizure Prediction Based on Deep Pairwise Representation Alignment of Epileptic EEG Signals. *IEEE Access* , *12* , 119056. https://doi.org/10.1109/access.2024.3447901

[19] Shi, S., & Liu, W. (2023). B2-ViT Net: Broad Vision Transformer Network With Broad Attention for Seizure Prediction. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* , *32* , 178. https://doi.org/10.1109/tnsre.2023.3346955

[20] Sook, J., Li, T., & Li, Y. (2024). Epileptic Seizure Prediction Based on Synchroextracting Transform and Sparse Representation. *IEEE Access* , *12* , 187684. https://doi.org/10.1109/access.2024.3514859

[21] Varnosfaderani, S. M., McNulty, I., Sarhan, N. J., Abood, W., & Alhawari, M. (2024). An Efficient Epilepsy Prediction Model on European Dataset With Model Evaluation Considering Seizure Types. *IEEE Journal of Biomedical and Health Informatics* , *28* (10), 5842. https://doi.org/10.1109/jbhi.2024.3423766

[22] Wang, X., Zhang, C., Kärkkäinen, T., Chang, Z., & Cong, F. (2022). Channel Increment Strategy-Based 1D Convolutional Neural Networks for Seizure Prediction Using Intracranial EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* , *31* , 316. https://doi.org/10.1109/tnsre.2022.3222095

[23] Wang, Y., Cui, W., Yu, T., Li, X., Liao, X., & Li, Y. (2023). Dynamic Multigraph Convolution-Based Channel-Weighted Transformer Feature Fusion Network for Epileptic Seizure Prediction. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* , *31* , 4266. https://doi.org/10.1109/tnsre.2023.3321414

[24] Wang, Y., Shi, Y., Cheng, Y., He, Z., Wei, X., Chen, Z., & Zhou, Y. (2022). A Spatiotemporal Graph Attention Network Based on Synchronization for Epileptic Seizure Prediction. *IEEE Journal of Biomedical and Health Informatics* , *27* (2), 900. https://doi.org/10.1109/jbhi.2022.3221211

[25] Wei, B., Xu, L., & Zhang, J. (2024). A Compact Graph Convolutional Network With Adaptive Functional Connectivity for Seizure Prediction.

IEEE Transactions on Neural Systems and Rehabilitation Engineering , *32* , 3531. https://doi.org/10.1109/tnsre.2024.3460348

[26] Wu, D., Shi, Y., Wang, Z., Yang, J., & Sawan, M. (2023). C2SP-Net: Joint Compression and Classification Network for Epilepsy Seizure Prediction. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* , *31* , 841. https://doi.org/10.1109/tnsre.2023.3235390

EDM-Based Graduate Curriculum Planning Framework

Levi E. Elipane[1], Antriman V. Orleans[2], Adonis P. David[3], Wilma S. Reyes[4],
Crist John Pastor[5], Felina P. Espique[6], Juan Primitivo P. Petrola[7]
Philippine Normal University, Manila, Philippines[1, 2, 3, 4, 5]
Saint Louis University, Baguio City, Philippines[6]
Pangasinan State University, Lingayen, Philippines[7]

***Abstract*—Graduate curriculum planning is critical to ensuring that teacher education programs remain responsive to professional standards, workplace demands, and emerging educational priorities. This study applies Educational Data Mining (EDM) techniques as an exploratory decision-support approach for graduate curriculum planning for in-service teachers in the Philippines. Using survey data from 624 teachers, the study examined professional development motivations, competency expectations, curriculum gaps, graduate learning outcomes, and barriers to participation in advanced studies. Guided by the Knowledge Discovery in Databases framework, the analysis involved data preprocessing, feature engineering, curriculum gap analysis, correlation analysis, and exploratory Random Forest variable ranking. Results revealed notable curriculum gaps in Diversity of Learners, Assessment and Reporting, and Curriculum and Planning within the Philippine Professional Standards for Teachers domains. At the graduate level, advanced professional competencies, research application, leadership, and expertise-based autonomy emerged as priority areas for curriculum enhancement. Financial constraints, workplace demands, and competing personal responsibilities were the most significant barriers to graduate study. The Random Forest results are interpreted as exploratory rankings within a constructed curriculum-needs score rather than as independent predictive validation. Based on the findings, a Data-Driven Graduate Curriculum Planning Framework is proposed to support curriculum prioritization, continuous improvement, and evidence-informed decision-making.**

***Keywords*—Educational data mining; graduate curriculum planning; teacher professional development; curriculum analytics; random forest; curriculum needs assessment; graduate education**

I. Introduction

Graduate education plays a critical role in strengthening teacher competence, supporting professional growth, and enhancing educational practice. Beyond credential attainment, graduate programs provide opportunities for educators to deepen disciplinary knowledge, strengthen pedagogical expertise, engage in scholarly inquiry, and assume leadership roles within their institutions and professional communities [1]–[4]. Previous studies have shown that graduate education contributes to professional effectiveness, reflective practice, career advancement, and the development of specialized competencies required in increasingly complex educational environments [1]–[4].

In the Philippines, efforts to improve teacher quality have been reinforced through professional standards frameworks such as the Philippine Professional Standards for Teachers (PPST), which articulate the competencies expected across different career stages. These standards have encouraged greater attention to competency-based development, curriculum alignment, and professional growth within teacher education programs [5], [6]. Consequently, graduate curricula are expected not only to provide advanced academic preparation but also to support teachers in meeting evolving professional expectations and educational challenges.

Designing graduate curricula for in-service teachers, however, remains a complex undertaking. Curriculum developers must consider professional standards, institutional priorities, workplace realities, disciplinary developments, learner expectations, and societal needs. Traditionally, curriculum planning has relied on expert consultation, stakeholder engagement, curriculum reviews, and needs assessment activities. While these approaches remain valuable, several scholars have emphasized the importance of systematic needs assessment in ensuring that educational programs remain responsive to learner and professional requirements [7]–[10]. Curriculum development informed by empirical evidence is more likely to produce programs that are relevant, responsive, and aligned with stakeholder needs.

Recent studies have further highlighted the need for graduate curricula that address emerging professional demands. Graduate education is increasingly expected to foster leadership, resilience, interdisciplinary collaboration, research competence, lifelong learning, and continuous quality improvement [11]–[14]. Teachers likewise pursue graduate studies for a variety of professional reasons, including career advancement, competency development, professional recognition, and improved instructional practice [2], [3], [14]. These findings suggest that curriculum planning should be grounded in evidence regarding teachers' actual professional development needs rather than assumptions regarding what educators ought to learn.

The growing availability of educational data has created new opportunities for evidence-informed educational planning. Advances in educational analytics, machine learning, and predictive modeling have enabled educational institutions to move beyond descriptive reporting toward more systematic and data-driven decision-making processes [15], [16]. Rather than

relying solely on traditional curriculum review approaches, institutions can increasingly utilize educational data to identify patterns, predict emerging needs, and support strategic planning initiatives.

One area that has gained considerable attention is Educational Data Mining (EDM), an interdisciplinary field that applies computational, statistical, and machine learning techniques to educational datasets in order to discover meaningful patterns and generate actionable insights [17], [18]. EDM has been successfully applied in diverse educational contexts, including student performance prediction, learning analytics, educational evaluation, institutional effectiveness studies, and decision-support systems [19]–[22]. The integration of machine learning techniques has further enhanced the ability of researchers to examine complex relationships among educational variables and identify influential factors that may not be readily observable through conventional analytical approaches [23], [24].

The growing adoption of educational analytics also reflects broader transformations occurring within higher education. Digital technologies, learning management systems, institutional databases, and educational information systems have generated large volumes of data that can inform educational planning and policy development [25]–[29]. These developments have encouraged institutions to explore data-driven approaches that support continuous improvement and evidence-based decision-making across various educational functions.

Despite the expanding body of Educational Data Mining research, its application to curriculum planning remains relatively limited. Existing studies have largely focused on learning behavior, academic achievement, technology adoption, student engagement, and institutional performance, while curriculum development continues to rely primarily on traditional needs assessment and expert-driven approaches [17], [21], [22]. Consequently, opportunities remain to leverage educational data for identifying curriculum priorities, competency gaps, professional development needs, and barriers affecting participation in graduate education.

The need for data-driven curriculum planning is particularly relevant for in-service teachers. Teachers pursuing graduate education often balance professional responsibilities, workplace expectations, family commitments, and financial constraints while seeking opportunities for career advancement and professional growth. Their curriculum expectations are shaped by professional standards, workplace realities, institutional requirements, and emerging educational challenges. Understanding these interconnected factors requires analytical approaches capable of examining multiple variables simultaneously and identifying those that most strongly influence curriculum needs.

Educational Data Mining offers a promising approach for addressing this challenge. By integrating predictive analytics with curriculum planning processes, institutions can identify priority competencies, examine curriculum gaps, determine factors influencing professional development needs, and develop more responsive graduate programs [15], [17], [18]. Such approaches support evidence-based decision-making while strengthening quality assurance and continuous improvement efforts within higher education institutions.

This study applies Educational Data Mining techniques to examine graduate curriculum needs among in-service teachers in the Philippines. Using data from 624 teachers representing diverse educational contexts, the study employs feature engineering, curriculum gap analysis, correlation analysis, and exploratory Random Forest analysis to identify patterns among curriculum-related indicators. Specifically, the study examines a Graduate Curriculum Needs Index (GCNI) as a descriptive planning score and proposes a Data-Driven Graduate Curriculum Planning Framework that may serve as a decision-support mechanism for curriculum development and enhancement.

By linking Educational Data Mining with graduate curriculum planning, the study contributes to educational analytics and teacher education research. It demonstrates how data-driven analysis can complement conventional needs assessment by organizing multiple sources of survey evidence into curriculum priorities for responsive, standards-aligned, and future-oriented graduate programs.

II. METHODOLOGY

A. Research Design

This study employed an Educational Data Mining (EDM) research design to identify graduate curriculum priorities and generate evidence-informed recommendations for curriculum planning. Guided by the Knowledge Discovery in Databases (KDD) framework [18], the study involved data selection, preprocessing, feature engineering, exploratory modelling, evaluation, and knowledge generation. A cross-sectional survey design was used to describe teachers' reported needs and to explore relationships among curriculum-related indicators. The machine-learning component was treated as an exploratory variable-ranking procedure rather than as independent validation of a predictive outcome.

B. Data Source and Participants

The dataset was derived from a national study on the development of graduate curricular programs for in-service teachers' professional development. The survey collected information on teachers' professional development needs, graduate curriculum expectations, competency requirements, career aspirations, motivations for pursuing graduate education, and barriers affecting participation in advanced studies. A total of 624 valid responses were included in the analysis. Respondents represented diverse educational contexts across the Philippines, including different regions, school classifications, educational attainment levels, teaching experience categories, and career stages. The relatively large sample provided a suitable basis for predictive modeling and Educational Data Mining applications.

C. Research Variables

The study utilized profile variables, professional development indicators, competency ratings, curriculum perceptions, and graduate learning outcome expectations. Profile variables included age, sex, region, school type, years of teaching experience, educational attainment, and career stage.

Professional development indicators covered learning and development, career advancement, salary progression, personal fulfillment, and research opportunities. Competency-related variables included ratings of the importance and perceived curriculum coverage of the seven Philippine Professional Standards for Teachers (PPST) domains. Additional variables examined master's and doctoral learning outcomes, curriculum relevance, and barriers to graduate education. The Graduate Curriculum Needs Index (GCNI) was retained as a descriptive composite planning score; because its components overlap with several variables examined in the exploratory analysis, it was not treated as an independently validated dependent variable.

D. Data Preprocessing and Feature Engineering

Prior to model development, the dataset underwent data cleaning, transformation, and preparation procedures. Duplicate records and irrelevant variables were removed, while categorical variables were converted into machine-readable formats. Numerical variables were standardized to ensure comparability across measures. Feature engineering was subsequently performed to construct indicators that better represented curriculum needs and professional development expectations. These included:

- Professional Development Motivation Indicators;
- PPST Curriculum Gap Indicators;
- Master's Outcome Gap Indicators;
- Doctoral Outcome Gap Indicators; and
- Aggregate Professional Development Measures.

Curriculum gap indicators were calculated as the difference between perceived importance and perceived curriculum coverage. Positive values indicated areas where respondents perceived existing graduate curricula to be insufficient in addressing important competencies or outcomes.

E. Construction of the Graduate Curriculum Needs Index

The Graduate Curriculum Needs Index (GCNI) was developed in the original analytical workflow as a descriptive curriculum-planning score integrating indicators related to competency expectations, curriculum gaps, graduate learning outcomes, professional development priorities, and barriers to participation. The component indicators were standardized, and the resulting score was rescaled to a 0–100 metric, with higher scores representing greater perceived demand for curriculum enhancement and professional development support. The original weighting documentation was not available for independent replication; therefore, the GCNI was not treated as a psychometrically validated scale or as an independent outcome. In this revised analysis, it is used only as an exploratory summary score, and the component-level gap, reliability, barrier, and correlation results remain the primary empirical basis for curriculum recommendations.

F. Educational Data Mining Framework

The study adopted the Knowledge Discovery in Databases (KDD) framework as the guiding Educational Data Mining process [18], [20]. The framework involved five major stages:

- Data selection and preparation;
- Data preprocessing and transformation;
- Feature engineering and indicator construction;
- Predictive modeling and evaluation; and
- Interpretation and knowledge generation.

This framework provided a systematic approach for transforming educational data into descriptive and exploratory insights for curriculum planning.

G. Random Forest Predictive Modeling

Random Forest regression was retained as an exploratory analytical procedure for examining how the available curriculum-related indicators were associated with the constructed GCNI score. Because several predictors were conceptually related to components of the index, the analysis was not interpreted as independent prediction or validation. The dataset was partitioned into training (70%) and testing (30%) subsets, and five-fold cross-validation was used in the original analytical workflow. Feature-importance rankings were examined only as supplementary evidence for organizing curriculum priorities, with due caution regarding correlated predictors and the constructed nature of the outcome.

H. Model Evaluation

The original Random Forest workflow was summarized using the coefficient of determination (R^2), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), together with five-fold cross-validation. These metrics describe how closely the model reproduced the constructed GCNI within the available dataset; they do not establish causal relationships, external validity, or the practical effectiveness of the proposed framework. Accordingly, the model results are reported as exploratory internal-fit statistics and are interpreted alongside the descriptive gap, barrier, reliability, and correlation analyses.

I. Feature Importance Analysis

Feature importance analysis was used to rank the relative contribution of variables within the exploratory Random Forest model. Because the predictors are conceptually overlapping and some are related to the construction of the GCNI, the rankings were interpreted cautiously. They are used only to complement the direct curriculum-gap and descriptive analyses and not to establish independent drivers or causal effects.

J. Ethical Considerations

The study was reviewed and approved by the Philippine Normal University Research Ethics Board. Participation was voluntary, and respondents were informed about the purpose of the study and provided informed consent before data collection. Responses were anonymized before analysis and reported only in aggregate form. Access to the dataset was limited to the research team, and the data were handled in accordance with institutional policies and applicable Philippine data-privacy requirements.

III. Results and Discussions

A. Dataset Profile

The dataset comprised 624 in-service teachers representing diverse educational contexts across the Philippines. Table I

presents the demographic characteristics of the respondents and their views regarding the direction of graduate curriculum development.

TABLE I. SUMMARY OF THE RESPONDENT PROFILE AND CURRICULUM DIRECTION INDICATORS

Indicator	Frequency	Percentage
Female	449	71.955
Male	161	25.801
Public school	492	78.846
Private school	132	21.154
Career progression = Yes	595	95.353
PPST alignment = Yes	606	97.115

The majority of respondents were female and employed in public schools, reflecting the demographic composition of the Philippine teaching workforce. More importantly, respondents expressed overwhelming support for graduate curricula that contribute to career progression and align with the Philippine Professional Standards for Teachers (PPST). The findings suggest that teachers view graduate education not merely as a means of obtaining advanced academic credentials but as an important mechanism for professional growth and career advancement. The strong endorsement of standards-based curriculum alignment further indicates that teachers expect graduate programs to support the competencies required across different stages of professional practice. These findings provide an important context for the Educational Data Mining analyses that follow, as they demonstrate that curriculum priorities are anchored in professional development and competency enhancement rather than personal preference alone.

B. Professional Development Motivations

Table II summarizes the motivations that influence teachers' decisions to pursue graduate education. Learning and professional development emerged as the strongest motivations, followed by career progression and salary advancement. Research opportunities received comparatively lower ratings.

TABLE II. SUMMARY OF THE PROFESSIONAL DEVELOPMENT SCORES

Motivation	Mean	SD	Rank
Learning and development	3.798	1.385	1
Career progression	3.373	1.110	2
Salary increase	3.159	1.227	3
Personal fulfilment	2.518	1.447	4
Research opportunities	2.152	1.230	5

The results indicate that teachers primarily pursue graduate studies to enhance their professional practice and improve career prospects. While research remains a defining characteristic of graduate education, the findings suggest that teachers place greater value on learning experiences that directly support instructional effectiveness, leadership development, and professional advancement. This has important implications for curriculum planning, as research components may be more meaningful when explicitly connected to authentic educational problems and workplace contexts.

C. Reliability of Curriculum Planning Indicators

As shown in Table III, all constructs demonstrated satisfactory to excellent reliability, with Cronbach's alpha coefficients exceeding commonly accepted thresholds.

TABLE III. INTERNAL CONSISTENCY OF MULTI-ITEM CONSTRUCTS

Construct	Indicators	Valid n	Cronbach alpha	Mean	SD
PPST Importance	7	624	0.937	4.561	0.681
PPST Coverage	7	428	0.934	4.348	0.705
Master's Outcome Importance	4	624	0.864	4.504	0.628
Master's Outcome Coverage	4	624	0.921	4.135	0.823
Doctorate Outcome Importance	4	624	0.924	4.435	0.711
Doctorate Outcome Coverage	4	624	0.953	4.082	0.870
Barriers to Graduate Education	15	624	0.950	3.832	0.880

Cronbach's alpha values ranged from 0.864 to 0.953 across PPST importance, PPST coverage, master's outcomes, doctorate outcomes, and barrier indicators. These values indicate strong internal consistency and support the use of composite indicators in the educational data mining workflow. The consistently high reliability values indicate that the survey items measured coherent dimensions of curriculum needs, professional development expectations, and graduate learning outcomes. These results provide confidence in the validity of the composite indicators used in subsequent analyses and support the suitability of the dataset for Educational Data Mining applications.

D. Curriculum Gap Analysis

To identify areas requiring curriculum enhancement, curriculum gaps were computed by comparing perceived importance and perceived curriculum coverage across the PPST domains. The results are presented in Table IV. All PPST domains were rated highly important, indicating broad support for standards-informed graduate curricula.

TABLE IV. PPST IMPORTANCE, COVERAGE, AND CURRICULUM GAP

PPST Domain	Importance Mean	Coverage Mean	Gap
Diversity of Learners	4.556	4.251	0.306
Assessment and Reporting	4.609	4.358	0.251
Curriculum and Planning	4.644	4.416	0.228
Learning Environment	4.542	4.328	0.214
Content Knowledge and Pedagogy	4.688	4.496	0.191
Community Linkages	4.269	4.115	0.154
Personal Growth and Prof. Dev.	4.622	4.480	0.142

Positive gap values across all domains indicate opportunities for curriculum enhancement and stronger alignment with competency-based teacher development frameworks [5], [6]. The largest gaps were found in Diversity of Learners, Assessment and Reporting, and Curriculum and Planning, suggesting that these competencies are highly important yet inadequately addressed in existing graduate curricula. These findings identify priority areas for intervention and demonstrate how Educational Data Mining can support evidence-based curriculum planning by pinpointing specific competency domains requiring improvement and guiding strategic resource allocation.

E. Graduate School Learning Outcome Gaps

Tables V and VI present the analyses of learning outcome gaps at the master's and doctoral levels.

TABLE V. MASTER'S OUTCOME IMPORTANCE, COVERAGE AND GAP

Master's Outcome	Importance Mean	Coverage Mean	Gap
Advanced knowledge and skills	4.615	4.216	0.399
Self-directed research	4.276	3.962	0.314
Lifelong learning	4.606	4.226	0.380
Application in research/professional work	4.521	4.135	0.386

At the master's level, the largest gaps were associated with advanced knowledge and skills, the application of research to professional practice, and lifelong learning competencies.

At the doctoral level, the greatest gaps were observed in expertise-based autonomy, leadership in research, and advanced professional practice.

TABLE VI. DOCTORATE OUTCOME IMPORTANCE, COVERAGE AND GAP

Doctorate Outcome	Importance Mean	Coverage Mean	Gap
Advanced systematic knowledge	4.495	4.143	0.353
Complex research/professional practice	4.378	4.054	0.324
Leadership in research/creative work	4.417	4.056	0.361
Expertise-based autonomy/accountability	4.450	4.074	0.377

The findings indicate that teachers expect graduate programs to offer stronger opportunities for professional growth and scholarly development. The differing priorities between master's and doctoral programs underscore the need for differentiated curriculum pathways. Master's programs should emphasize applied professional competencies, while doctoral programs should focus on research leadership, innovation, and independent scholarship.

F. Barriers to Graduate Education

Table VII shows that financial cost was the major barrier to graduate education, followed by scheduling conflicts and personal responsibilities. Other challenges included academic workload, publication requirements, and thesis demands. These findings suggest that flexible program structures, delivery modes, and assessment requirements are essential for improving participation and success among working professionals.

G. Graduate Curriculum Needs Index

To provide a summary measure of perceived curriculum demand, the Graduate Curriculum Needs Index (GCNI) integrated several curriculum-related indicators. Table VIII presents the descriptive results. Because the original weighting documentation was unavailable and the index overlaps conceptually with several analytical variables, the GCNI is interpreted only as an exploratory planning score rather than as a psychometrically validated measure.

The mean GCNI score suggests perceived opportunities for improving graduate curricula in ways that support professional standards, competency development, and career advancement. However, curriculum decisions in this study are based primarily on the component-level gap, barrier, and learning-outcome analyses rather than on the GCNI alone.

TABLE VII. BARRIERS TO GRADUATE EDUCATION

Barrier	Mean	SD	Interpretation	Rank
Financial cost of graduate education	4.189	1.045	Agree	1
Conflict with workplace requirements and schedule	4.106	1.042	Agree	2
Conflict with other personal responsibilities	4.000	1.081	Agree	3
Heavy or too many course requirements	3.946	1.078	Agree	4
Limited support for graduate students from the university/college	3.942	1.127	Agree	5
Publication or creative work requirement	3.931	1.133	Agree	6
Thesis/Dissertation requirement	3.848	1.178	Agree	7
Mental health/psychological wellbeing concerns	3.816	1.167	Agree	8
Long distance between university/college and workplace	3.788	1.178	Agree	9
Long distance between university/college and home	3.768	1.185	Agree	10
Physical health and fitness concerns	3.764	1.177	Agree	11
Requirement for face-to-face sessions	3.636	1.233	Agree	12
Lack of learning or professional growth from the graduate program	3.625	1.272	Agree	13
Too many sessions in a course	3.580	1.161	Agree	14
Too many hours in a session	3.545	1.159	Agree	15

TABLE VIII. SUMMARY OF THE CONSTRUCT INDICATORS AND GCNI

Indicator	Count	Mean	Std	Min	Max
PPST_Importance	624	4.561	0.681	1.00	5.00
PPST_Coverage	428	4.348	0.705	1.00	5.00
PPST_Gap	428	0.224	0.630	-3.00	2.857
Masters_Outcome_Importance	624	4.504	0.628	1.00	5.00
Masters_Outcome_Gap	624	0.370	0.766	-2.75	4.00
Doctorate_Outcome_Import-ance	624	4.435	0.711	1.00	5.00
Doctorate_Outcome_Gap	624	0.353	0.864	-3.25	4.00
Barriers_Index	624	3.832	0.880	1.00	5.00
GCNI_100	624	62.488	12.435	0.00	100.00

H. Exploratory Random Forest Analysis

Random Forest regression was retained as an exploratory analysis of the constructed GCNI. The internal-fit statistics are presented in Table IX. No separate actual-versus-predicted figure is reported in the revised manuscript.

TABLE IX. RANDOM FOREST PREDICTIVE PERFORMANCE

Metric	Value
Training R-squared	0.930
Testing R-squared	0.835
Five-fold CV R-squared Mean	0.815
Five-fold CV R-squared SD	0.051
RMSE	4.386
MAE	2.892

The model reproduced the constructed GCNI with high internal fit. However, this result is expected to some extent because several predictors are conceptually related to the indicators underlying the index. The statistics therefore do not demonstrate independent predictive validity. Their value is limited to showing internal patterns among the available curriculum-planning indicators, and they should be read together with the direct gap, barrier, reliability, and correlation results. The analysis is predictive neither of actual enrolment nor of the practical effectiveness of the proposed framework, and no causal interpretation is intended.

I. Feature Importance Analysis

To examine the relative contribution of variables within the exploratory model, feature-importance analysis was conducted. The results are presented in Table X.

Graduate learning outcomes and competency-related indicators received the highest importance scores within the exploratory model, while demographic characteristics and personal motivations received smaller scores. Because these variables overlap conceptually and are related to the constructed GCNI, the rankings are not interpreted as independent drivers.

They nevertheless support the descriptive finding that curriculum planning should focus on competencies and expected learning outcomes rather than demographic factors alone.

TABLE X. RANDOM FOREST FEATURE IMPORTANCE RANKINGS

Predictor	Importance
Master's Outcome Importance	0.166
PPST Importance	0.163
Doctorate Outcome Importance	0.120
Doctorate Outcome Gap	0.111
Master's Outcome Gap	0.089
PPST Curriculum Gap	0.058
Career Progression Need	0.045
Barrier Index	0.041
PPST Alignment Need	0.028
Age	0.027
Personal fulfilment	0.010
Research opportunities	0.009
Salary increase	0.009
Learning and Development	0.008
Career Progression	0.008

J. Correlation Analysis

To further examine relationships among the major variables, correlation analysis was conducted. The results are summarized in Table XI.

TABLE XI. CORRELATION OF MAJOR PREDICTORS WITH GCNI

Predictor	Pearson r with GCNI	p-value
Master's Outcome Importance	0.697	0.000
Doctorate Outcome Importance	0.651	0.000
PPST Importance	0.650	0.000
Doctorate Outcome Gap	0.580	0.000
Master's Outcome Gap	0.502	0.000
PPST Curriculum Gap	0.441	0.000
Barrier Index	0.194	0.000

The observed correlations were generally consistent with the descriptive and exploratory Random Forest results. However, the correlations should be interpreted as associations among related curriculum-planning indicators, not as independent validation of the GCNI or as evidence of causality.

K. Curriculum Priority Mapping

The integration of curriculum gap analysis, barrier analysis, correlation analysis, and exploratory variable ranking informed the development of a curriculum priority map. The resulting priorities are presented in Table XII.

The curriculum priority map illustrates how Educational Data Mining can complement traditional needs assessment and curriculum review processes [10]. Rather than treating the statistical outputs as validation of a predictive system, the analysis organizes the strongest descriptive findings into practical areas for curriculum revision, resource allocation, and strategic planning within graduate education programs.

TABLE XII. DATA-DRIVEN CURRICULUM PRIORITY MAP

Priority Level	Curriculum Area	Planning Implication
High Priority	PPST-aligned competencies	Embed PPST domains into course outcomes, assessments, and practicum tasks.
High Priority	Master's outcome competencies	Strengthen advanced pedagogy, applied research, and professional practice.
High Priority	Doctorate outcome competencies	Strengthen doctoral research leadership, innovation, and autonomous expertise.
Moderate Priority	Assessment and reporting	Develop assessment literacy, data use, and reporting competencies.
Moderate Priority	Diversity of learners	Strengthen inclusive education, differentiated instruction, and learner-responsive pedagogy.
Emerging Priority	Educational analytics and digital transformation	Integrate data-driven teaching, AI awareness, and educational technology applications.
Support Priority	Flexible delivery and financial/access support	Adopt hybrid, modular, weekend, and scholarship-supported delivery mechanisms.

L. Data-Driven Graduate Curriculum Planning Framework

Drawing upon the descriptive and exploratory analyses, a Data-Driven Graduate Curriculum Planning Framework was proposed. The framework is presented in Fig. 1. It integrates teacher needs assessment, curriculum gap analysis, exploratory analytics, curriculum prioritization, and continuous evaluation into a decision-support process. The dataset supports the identification of needs and priorities; the effectiveness of the framework itself has not yet been empirically evaluated.

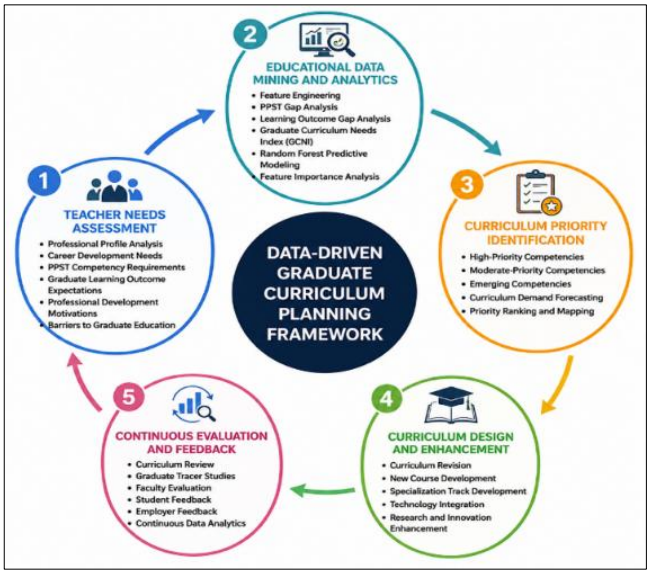


Fig. 1. Data-driven graduate curriculum planning framework.

By combining traditional curriculum planning approaches with Educational Data Mining techniques, the proposed framework provides a systematic mechanism for translating educational data into curriculum recommendations. It builds upon established principles of curriculum needs assessment, evidence-based planning, and educational analytics [7]. The framework should be understood as a data-informed planning proposal rather than a validated intervention. It may be adapted to other national, institutional, or professional contexts by replacing the PPST with the relevant competency standards and by incorporating local curriculum, employer, and stakeholder evidence.

IV. Conclusion and Recommendations

A. Conclusion

This study demonstrated the potential of Educational Data Mining (EDM) as an exploratory decision-support approach for graduate curriculum planning. Using data from 624 in-service teachers, the study integrated curriculum needs assessment, competency analysis, feature engineering, and exploratory analytics to identify curriculum priorities and professional development needs. The findings revealed substantial demand for curriculum enhancement, particularly in areas associated with professional competencies, graduate learning outcomes, and standards-based teacher development.

The curriculum gap analyses showed that respondents consistently rated professional competencies and graduate learning outcomes as highly important while identifying areas where existing curriculum coverage could be strengthened. Notable gaps were observed in selected PPST domains and graduate learning outcomes, suggesting opportunities for curriculum enhancement and improved alignment between curriculum provisions and professional expectations. In addition, barriers such as financial constraints, scheduling conflicts, and competing responsibilities highlighted the importance of considering both curricular and structural factors when designing programs for working professionals.

The GCNI provided an exploratory summary of perceived curriculum demand, but it was not treated as a psychometrically validated or independent predictive outcome. The Random Forest analysis was therefore interpreted only as supplementary variable ranking within a constructed score. The stronger empirical contribution lies in the direct curriculum-gap, learning-outcome, barrier, reliability, and correlation analyses, which consistently emphasized competency-based and outcomes-oriented curriculum planning.

Building on these findings, the study proposed a Data-Driven Graduate Curriculum Planning Framework that integrates needs assessment, exploratory analytics, curriculum prioritization, and continuous evaluation. The framework illustrates how educational data can inform curriculum development, but the present study does not demonstrate that the framework improves curriculum-planning outcomes in practice.

The study contributes to Educational Data Mining research by extending its exploratory application toward curriculum planning and educational decision-making. The findings suggest that curriculum analytics can complement traditional curriculum-development approaches in designing responsive, standards-informed, and future-oriented graduate programs.

Several limitations should be acknowledged. First, the cross-sectional design does not support causal conclusions. Second, the GCNI was derived from curriculum-related indicators that overlap with variables examined in the exploratory Random Forest analysis, and its original weighting documentation was not available for independent replication. The model results therefore represent internal exploratory associations rather than independent predictive validation. Third, the findings are based entirely on teachers' self-reported perceptions and were not triangulated with curriculum documents, employer requirements, institutional performance data, or graduate employment outcomes. Fourth, the PPST-specific context limits direct transferability, although the analytical process may be adapted using other national, institutional, or professional competency frameworks. Finally, the proposed planning framework was not empirically evaluated for practical effectiveness. Future studies should use independently measured outcomes, external datasets, longitudinal designs, and additional interpretive methods such as permutation importance or SHAP-based analysis.

B. Recommendations

Higher education institutions may consider using data-driven analysis alongside traditional needs assessment in curriculum review and development. Graduate curriculum developers may strengthen alignment with professional standards and expected learning outcomes while ensuring that program structures remain responsive to the financial, workplace, and personal constraints faced by working professionals.

Future studies may validate the proposed framework in other disciplinary and institutional contexts, triangulate survey findings with curriculum and employment evidence, and examine independently measured outcomes such as enrolment intention, preferred curriculum tracks, program completion, or graduate outcomes. Comparative baseline models and more stable interpretive methods may also be used before making predictive claims.

Acknowledgment

The researchers sincerely acknowledge the Teacher Education Council (TEC), through the Research Initiatives on Teacher Education (RITE), for funding this study. They also acknowledge the Philippine Normal University, Saint Louis University, Pangasinan State University, Western Visayas State University, and Ateneo de Davao University for their support in its conduct. Artificial Intelligence (AI) tools were utilized solely for language enhancement, organization, and formatting of the manuscript. All research processes, analyses, interpretations, and conclusions were independently reviewed and validated by the researchers to ensure accuracy, originality, and academic integrity.

References

[1] G. Ion and R. Iucu, "The impact of postgraduate studies on the teachers' practice," European Journal of Teacher Education, vol. 39, no. 5, pp. 602–615, 2016.

[2] A. S. Horn and S. T. Jang, "The impact of graduate education on teacher effectiveness: Does a master's degree matter?," MHEC Research Brief, Midwestern Higher Education Compact, Minneapolis, MN, USA, 2017.

[3] F. M. Sevim and U. Akın, "The role of graduate education in professional development of teachers: Is graduation enough?," Education and Science, vol. 46, no. 207, pp. 483–510, 2021.

[4] A. J. Patac, L. P. Patac, and J. G. Caadan, "Graduate school's impact on careers: Insights and implications from Master of Arts in Education graduates," Contemporary Educational Researches Journal, vol. 14, no. 2, pp. 1–10, 2024.

[5] G. G. Taculog and R. V. Santos, "Teaching practices of public elementary school teachers relative to Philippine Professional Standards for Teachers," International Journal of Advanced Multidisciplinary Studies, 2024.

[6] E. C. Gepila Jr., "Assessing teachers using Philippine standards for teachers," Universal Journal of Educational Research, vol. 8, no. 3, pp. 739–746, 2020.

[7] B. Roszkowski and G. Reynolds, "Assessing, analyzing, and adapting: Improving a graduate student instruction program through needs assessment," Behavioral & Social Sciences Librarian, vol. 32, no. 4, pp. 224–239, 2013.

[8] K. Riesen, "Graduate student learning and the development of programmatic instruction: A mixed methods needs assessment," The Journal of Academic Librarianship, vol. 49, no. 1, Art. no. 102631, 2023.

[9] H. Assalahi, "The professional standards for teachers in Saudi Arabia as a framework for EFL teachers' professional development needs assessment," English Language Education Publishing, vol. 16, no. 6.1, pp. 161–186, 2021.

[10] A. H. Guleid, "Factors influencing Cambodia's master's degree students to choose between thesis and coursework program," Open Journal of Social Sciences, vol. 10, no. 12, pp. 534–543, 2022.

[11] J. C. Marasigan, "An inquiry into the continuing professional development training needs of education graduates in a state university in the Philippines," Indian Journal of Science and Technology, vol. 16, no. 32, pp. 2568–2573, 2023.

[12] K. E. McKee, D. Serrano, M. Girvan, and G. Marbach-Ad, "An integrated model for interdisciplinary graduate education," PLOS ONE, vol. 16, no. 9, Art. no. e0257872, 2021.

[13] G. Van Kessel, M. Brewer, M. Lane, B. Cooper, and F. Naumann, "A principle-based approach to the design of a graduate resilience curriculum framework," Higher Education Research & Development, vol. 41, no. 4, pp. 1325–1339, 2022.

[14] A. P. David, Z. Q. Reyes, P. A. Miranda, M. J. N. Nalipay, I. V. Ancho, and M. M. Roxas, "Graduate teacher education in the Philippines: Observations and prospects," The Normal Lights, vol. 14, no. 2, 2020.

[15] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 40, no. 6, pp. 601–618, 2010.

[16] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," WIREs Data Mining and Knowledge Discovery, vol. 10, no. 6, Art. no. e1355, 2020.

[17] S. Ventura and C. Romero, "Educational data mining: A survey from 1995 to 2005," Expert Systems with Applications, vol. 33, no. 1, pp. 135–146, 2007.

[18] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," AI Magazine, vol. 17, no. 3, pp. 37–54, 1996.

[19] R. S. J. d. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," Journal of Educational Data Mining, vol. 1, no. 1, pp. 3–17, 2009.

[20] L. A. Kurgan and P. Musilek, "A survey of knowledge discovery and data mining process models," The Knowledge Engineering Review, vol. 21, no. 1, pp. 1–24, 2006.

[21] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[22] G. Biau and E. Scornet, "A random forest tutorial," TEST, vol. 25, no. 2, pp. 197–227, 2016.

[23] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Waltham, MA, USA: Morgan Kaufmann, 2011.

[24] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. Cambridge, MA, USA: Morgan Kaufmann, 2016.

[25] G. Siemens and P. Long, "Penetrating the fog: Analytics in learning and education," EDUCAUSE Review, vol. 46, no. 5, pp. 30–40, 2011.

[26] D. Gašević, S. Dawson, and J. Pardo, "How do we start? An approach to learning analytics adoption in higher education," International Journal of Information and Learning Technology, vol. 33, no. 4, pp. 247–255, 2016.

[27] G. Siemens and R. S. J. d. Baker, "Learning analytics and educational data mining: Towards communication and collaboration," in Proc. 2nd Int. Conf. Learning Analytics and Knowledge (LAK '12), Vancouver, BC, Canada, 2012, pp. 252–254.

[28] V. Basilotta-Gómez-Pablos, M. Matarranz, L.-A. Casado-Aranda, and A. Otto, "Teachers' digital competencies in higher education: A systematic literature review," International Journal of Educational Technology in Higher Education, vol. 19, no. 1, Art. no. 8, 2022.

[29] M. Bond, S. Bedenlier, V. I. Marín, and M. Händel, "Emergency remote teaching in higher education: Mapping the first global online semester," International Journal of Educational Technology in Higher Education, vol. 18, no. 1, Art. no. 50, 2021..

Arki-Nexus AI: A Framework for Integrating Artificial Intelligence Competencies Into NEUST College of Architecture

GREGORIO L. VILLAVIZA JR
Associate Professor V, College of Architecture, Nueva Ecija University of Science and Technology, Cabanatuan City, Nueva Ecija, Philippines

***Abstract*—The rapid expansion of Artificial Intelligence (AI) is reshaping architectural education and practice, yet its systematic curricular integration remains limited in many higher education settings. This study developed the proposed Arki-Nexus AI Framework to guide the structured integration of AI competencies into the curriculum of the NEUST College of Architecture. Employing a descriptive-comparative quantitative research design, the study involved 31 faculty members selected through purposive sampling and 200 architecture students selected through stratified random sampling with equal allocation by year level. Data were gathered using a researcher-developed survey instrument that underwent expert content validation by seven experts (I-CVI = 1.00) and pilot reliability testing (Cronbach's α = .85–.92). Faculty members reported a high overall level of AI competency (M = 4.11), while students also reported a high overall level (M = 3.63), although student ratings were moderate in AI-assisted design, AI-supported collaboration, and AI-driven innovation. The perceived extent of AI integration was moderately evident for both faculty (M = 3.18) and students (M = 2.96). Significant faculty-student differences were found in applied AI competencies, but not in AI literacy or ethical AI use. These findings informed a proposed framework linking human competencies, pedagogical drivers, and institutional enablers. Because the findings were based on self-reported data from one college, the framework requires future expert review, pilot implementation, and feasibility and impact evaluation.**

Keywords*—Artificial intelligence; architecture education; self-reported AI competency; curriculum integration; curriculum framework; digital transformation; future-ready architects

I. INTRODUCTION

Artificial Intelligence (AI) and related digital technologies are changing design generation, visualization, simulation, information modeling, environmental analysis, and decision support across the architecture, engineering, and construction industry. Recent architectural education studies show that AI can support architectural programming, design exploration, iterative visualization, and studio learning when it is integrated through deliberate pedagogical design rather than treated as a stand-alone tool [1]–[3]. Architecture programs must therefore prepare students to use AI critically and competently while preserving architectural judgment, creativity, and accountability.

International competency frameworks also emphasize that AI readiness includes more than operational tool use. UNESCO's frameworks for students and teachers identify human-centered values, ethical AI use, foundational knowledge, practical application, pedagogy, and continuous professional learning as central dimensions of AI competency [4], [5]. These dimensions provide an appropriate basis for examining how architecture students and faculty understand and use AI in educational contexts.

Within architecture education, AI has been used in programming, concept development, design alternatives, image generation, environmental analysis, and critique. Empirical studies nevertheless report uneven and fragmented use, including difficulties in connecting AI outputs to coherent architectural workflows and in maintaining human oversight of design decisions [1]–[3], [6]. These findings indicate that access to AI tools alone does not guarantee meaningful curricular integration.

At the institutional level, AI adoption requires alignment among curriculum outcomes, teaching strategies, learning activities, assessment practices, digital infrastructure, faculty development, and governance. Higher education research shows that digital transformation remains limited when technology is added without coordinated changes to pedagogy, assessment, policy, and organizational support [7]–[10]. A structured curriculum framework is therefore necessary to move from isolated tool use toward progressive competency development.

This concern is relevant to the NEUST College of Architecture, which seeks to prepare graduates for digitally enabled professional practice. The College needs empirical evidence on the self-reported AI competencies of its faculty members and students and on the extent to which AI is presently reflected in its curriculum, instruction, learning activities, assessment practices, and resources.

The integration of AI into architectural education must remain human-centered. AI should augment, rather than replace, design reasoning, contextual sensitivity, professional judgment, authorship, and social responsibility. International ethical guidance emphasizes transparency, accountability, privacy, fairness, and human oversight in the development and use of AI systems [11]–[13]. These principles are especially important in architecture because design decisions affect people, communities, and the built environment.

In response, this study developed the proposed Arki-Nexus AI Framework as a curriculum enhancement model for the NEUST College of Architecture. The framework organizes AI

competency development across human, pedagogical, and institutional dimensions and links identified competency and integration gaps with possible curricular interventions.

The framework is intended for the specific context examined in this study. It was developed from the findings but was not independently validated through expert review, pilot implementation, or feasibility testing during this research phase. Its use beyond the NEUST College of Architecture therefore requires further evaluation and contextual adaptation.

A. Artificial Intelligence Competencies in Architecture Education

AI competency refers to the knowledge, skills, values, and judgment required to understand, evaluate, apply, and communicate with AI systems responsibly. It includes conceptual understanding, practical application, critical evaluation, ethical awareness, and the capacity to use AI in problem-solving and creative work [4], [5], [14], [15]. In architectural education, these competencies must be connected to design processes, technical workflows, collaboration, and professional responsibility.

The present study examined five competency dimensions: AI literacy and foundational knowledge; AI-assisted design and technical skills; ethical and responsible AI use; AI-supported collaboration and communication; and AI-driven innovation and problem-solving. These dimensions were adapted to the architecture education context from established AI literacy and competency frameworks [4], [5], [14], [15].

Assessing these dimensions can identify where faculty members and students perceive themselves as prepared and where additional learning opportunities are needed. Because the instrument used self-rated Likert-scale responses, the resulting scores indicate perceived competency rather than directly demonstrated performance.

B. AI Integration in Architecture Programs

Program-level AI integration involves embedding AI concepts and applications across formal curriculum content, instructional strategies, student learning experiences, assessment systems, and institutional resources. Effective integration is systematic, progressive, aligned with program outcomes, and supported by appropriate governance and infrastructure.

Curriculum content establishes where AI competencies are introduced, reinforced, and applied. Architecture education studies recommend integrating AI into programming, studio work, computational design, visualization, performance analysis, and design critique while preserving critical and human-centered approaches [1], [2], [6].

Teaching strategies determine how faculty members mediate AI use. AI-assisted studio demonstrations, guided experimentation, project-based learning, critique, collaborative inquiry, and reflective discussion can help students connect AI outputs with architectural reasoning and design decisions.

Institutional adoption is more likely to produce educational value when faculty preparation, instructional design, and student access are coordinated. Research on AI and digital transformation in higher education emphasizes that technology adoption without pedagogical alignment and sustained institutional support often remains superficial [7]–[9].

Learning activities provide opportunities to translate conceptual knowledge into performance. Studio projects, iterative prompt-and-critique exercises, AI-assisted programming, visualization, simulation, and collaborative design tasks can support authentic practice when activities are scaffolded and connected to course outcomes [1], [2], [6].

Assessment practices must evaluate not only final AI-assisted outputs but also process documentation, critical judgment, ethical use, attribution, and human decision-making. Assessment standards should protect academic integrity and maintain transparency, fairness, and human oversight [9], [12], [13].

C. Research Questions

This study aimed to develop a proposed Arki-Nexus AI Framework for integrating Artificial Intelligence competencies into the curriculum of the NEUST College of Architecture. Specifically, it sought to answer the following questions:

- What is the level of self-reported Artificial Intelligence competency among architecture students and faculty members of the NEUST College of Architecture in terms of:
  - AI literacy and foundational knowledge;
  - AI-assisted design and technical skills;
  - ethical and responsible AI use;
  - AI-supported collaboration and communication; and
  - AI-driven innovation and problem-solving?
- What is the perceived extent of AI integration in the NEUST College of Architecture with respect to:
  - Curriculum content;
  - Teaching strategies;
  - Learning activities;
  - Assessment practices; and
  - Digital infrastructure and AI-related resources?
- Based on the findings, what Arki-Nexus AI Framework may be proposed to support the systematic integration of Artificial Intelligence competencies into the curriculum of the NEUST College of Architecture?
- Is there a significant difference between the self-reported AI competency levels of faculty members and architecture students across the five competency dimensions?

II. METHODOLOGY

A. Research Design

The study employed a descriptive-comparative quantitative research design. The descriptive component measured respondents' self-reported AI competency and their perceptions of the extent of AI integration in the College. The comparative component examined differences between faculty and student self-ratings across the five AI competency dimensions. The findings were used as an empirical basis for developing the proposed Arki-Nexus AI Framework; the research design did not include a separate validation phase for the framework.

B. Locale, Respondents, and Sampling Technique

The study was conducted at the College of Architecture of Nueva Ecija University of Science and Technology in the Philippines. Respondents comprised 31 faculty members and 200 architecture students. Faculty members were selected purposively because of their direct involvement in architecture courses and design studios. Students were selected through stratified random sampling with equal allocation of 40 respondents from each year level. This equal allocation ensured representation from all year levels but was not proportional unless year-level populations were equal. Because the study involved respondents from one college in one institution, the results are context-specific and should not be generalized to Philippine architecture education as a whole.

C. Research Instrument

The researcher-developed survey instrument consisted of two main sections.

The first section measured self-reported AI competency across five dimensions: AI literacy and foundational knowledge; AI-assisted design and technical skills; ethical and responsible AI use; AI-supported collaboration and communication; and AI-driven innovation and problem-solving. Responses used a five-point scale interpreted as follows: 4.21–5.00, Very High; 3.41–4.20, High; 2.61–3.40, Moderate; 1.81–2.60, Low; and 1.00–1.80, Very Low. Because the items were self-rated, the scores represented perceived competency rather than objectively demonstrated performance.

The second section assessed the perceived extent of AI integration across curriculum content, teaching strategies, learning activities, assessment practices, and digital infrastructure and AI-related resources. Responses used a five-point scale interpreted as follows: 4.21–5.00, Very Evident; 3.41–4.20, Evident; 2.61–3.40, Moderately Evident; 1.81–2.60, Somewhat Evident; and 1.00–1.80, Not Evident.

D. Validity and Reliability of the Instrument

Content validity was evaluated by seven experts in architecture education, curriculum development, and educational technology. All items across the ten subscales attained an Item-Level Content Validity Index (I-CVI) of 1.00, indicating unanimous expert agreement on item relevance and clarity [16]. Reliability was examined through a pilot test involving 42 students who were excluded from the main study. Cronbach's alpha coefficients ranged from .85 to .92 across the subscales, indicating good to excellent internal consistency [17]. These findings supported the quality of the survey instrument only; they did not constitute validation of the proposed Arki-Nexus AI Framework.

E. Data-Gathering Procedure

Data collection was conducted after approval from the relevant academic and institutional authorities. The researcher coordinated with subject teachers, distributed the validated questionnaires with informed-consent forms, and provided standardized instructions. Respondents completed the instrument in approximately 15 to 25 minutes. Completed questionnaires were retrieved, checked for completeness, encoded, and organized for statistical analysis. Participation remained voluntary throughout the process.

F. Data Analysis

Descriptive and inferential statistics were used. Means were computed to summarize self-reported AI competency and the perceived extent of AI integration. An independent-samples t-test examined differences between faculty and student self-ratings across the five AI competency dimensions. Statistical decisions were evaluated at the .05 level of significance. The analysis did not test the validity, feasibility, or effectiveness of the proposed framework.

G. Ethical Considerations

The study observed established ethical principles for educational research. Informed consent was obtained before participation, and respondents were informed that participation was voluntary. Anonymity and confidentiality were maintained; no identifying information was linked to individual responses, and the data were used only for academic purposes and stored securely.

III. RESULTS AND DISCUSSION

A. AI Literacy and Foundational Knowledge

Faculty members reported a Very High level of AI literacy and foundational knowledge (M = 4.21), while students reported a High level (M = 4.03). The results indicate strong perceived awareness of AI concepts in both groups, although faculty ratings were slightly higher. These findings are consistent with AI literacy frameworks that distinguish foundational understanding from deeper application and critical evaluation [4], [5], [14], [15].

For both groups, awareness of the limitations and risks of AI in architecture received the highest rating (faculty: M = 4.55; students: M = 4.43), both interpreted as Very High. Regular updating of knowledge on emerging AI technologies received the lowest rating (faculty: M = 3.95; students: M = 3.84), although both remained High. Faculty members also rated their understanding of how data and algorithms influence AI-generated outputs higher (M = 4.40) than students did (M = 3.92). These patterns support continuous updating and deeper instruction on data, algorithms, limitations, and critical evaluation.

B. AI-Assisted Design and Technical Skills

Faculty members reported a High level of AI-assisted design and technical skills (M = 3.94), whereas students reported a Moderate level (M = 3.34). The difference indicates that students perceived less consistent ability to apply AI across

architectural workflows. Architecture education studies similarly show that learners may use generative tools for isolated tasks without yet integrating them into a coherent programming, design, and technical process [1], [2].

Faculty members rated the use of AI to improve design efficiency and productivity highest (M = 4.20). Students reported their lowest ratings for integrating AI with CAD/BIM platforms (M = 3.02) and generating design alternatives using AI (M = 3.32). Both groups rated critical evaluation of AI-generated outputs more strongly (faculty: M = 4.11; students: M = 3.65) than technical workflow integration. The findings support curriculum enhancement in AI-CAD/BIM integration, generative design, and studio-based application [1], [2], [6].

C. Ethical and Responsible AI Use

Faculty members reported a Very High level of ethical and responsible AI use (M = 4.38), while students reported a High level (M = 4.07). The difference should not be interpreted as proof of actual ethical behavior because the ratings were self-reported. However, the pattern suggests that both groups perceived themselves as aware of accountability, transparency, attribution, and responsible use. Kolb's experiential learning perspective indicates that ethical judgment can be strengthened through repeated application and reflection in authentic tasks [18], while UNESCO's competency frameworks place ethics and human-centered values at the core of AI readiness [4], [5].

Faculty ratings were Very High for understanding ethical issues related to AI use (M = 4.50) and for responsible use and attribution of AI-generated outputs (M = 4.45). Students rated recognition of bias and misinformation highest (M = 4.18), but proper attribution of AI-generated outputs was lower (M = 3.92). The results support explicit instruction on authorship, intellectual property, data privacy, bias, transparency, and human oversight, consistent with international ethical guidance [11]–[13].

D. AI-Supported Collaboration and Communication

Faculty members reported a High level of AI-supported collaboration and communication (M = 4.06), whereas students reported a Moderate level (M = 3.34). This difference indicates that students perceived fewer or less consistent opportunities to use AI in collaborative architectural work. Vygotsky's social constructivist perspective emphasizes learning through guided and socially mediated activity [19], while UNESCO's frameworks recognize communication, collaboration, and pedagogy as important elements of responsible AI use [4], [5].

Faculty members rated AI-supported communication in group tasks (M = 4.15) and coordination of design collaboration (M = 4.10) most strongly. Students reported lower ratings for clearly sharing AI-generated outputs (M = 3.20) and using AI-supported platforms (M = 3.37). Connectivism explains learning as participation in networks of people, information, and digital tools [20]. Accordingly, shared BIM environments, collaborative generative platforms, and structured group critique may help students use AI outputs more effectively in team-based design [12], [15].

E. AI-Driven Innovation and Problem-Solving

Faculty members reported a High level of AI-driven innovation and problem-solving (M = 3.95), while students reported a Moderate level (M = 3.34). The results indicate a gap in perceived ability to use AI for complex design reasoning, scenario analysis, and evidence-informed decisions. The higher-order cognitive processes of analyzing, evaluating, and creating in Bloom's revised taxonomy provide an appropriate basis for designing progressively challenging AI-assisted tasks [21].

Faculty members rated exploration of innovative architectural solutions (M = 4.15) and critical evaluation of AI-generated outputs (M = 4.10) most strongly. Students reported Moderate ratings in complex problem-solving (M = 3.28), scenario analysis (M = 3.35), and evidence-based decision-making (M = 3.28). Their relatively higher ratings in innovation exploration (M = 3.44) and critical evaluation (M = 3.48) indicate emerging readiness that can be strengthened through scaffolded studio tasks aligned with AI literacy and higher-order learning outcomes [14], [15], [21].

F. Summary of Self-Reported AI Competency

Table I summarizes the self-reported AI competency results. Faculty members obtained a High grand mean (M = 4.11), and students also obtained a High grand mean (M = 3.63). Faculty ratings were strongest in ethical and responsible AI use and AI literacy. Student ratings were High in these foundational domains but Moderate in AI-assisted design, collaboration, and innovation. The pattern suggests that conceptual and ethical awareness is more developed than applied and workflow-based competency [4], [5], [9], [14], [15].

The applied-domain differences indicate a need to connect AI knowledge with repeated studio practice. Students should encounter progressively complex tasks involving architectural programming, generative exploration, CAD/BIM integration, collaborative design, technical analysis, and critical evaluation of outputs. Such integration is consistent with current research on AI-assisted architectural education and design pedagogy [1], [2], [6].

TABLE I. SUMMARY OF SELF-REPORTED AI COMPETENCY LEVELS OF FACULTY MEMBERS AND STUDENTS

Artificial Intelligence Competency	Faculty		Students	
	Mean	Verbal Description	Mean	Verbal Description
AI Literacy and Foundational Knowledge	4.21	Very High	4.03	High
AI-Assisted Design and Technical Skills	3.94	High	3.34	Moderate
Ethical and Responsible AI Use	4.38	Very High	4.07	High
AI-Supported Collaboration and Communication	4.06	High	3.34	Moderate
AI-Driven Innovation and Problem-Solving	3.95	High	3.34	Moderate
Grand Mean	**4.11**	**High**	**3.63**	**High**

G. Curriculum Content

Faculty (M = 2.82) and students (M = 2.95) both perceived AI integration in curriculum content as Moderately Evident. These results indicate that AI-related concepts were present but not consistently embedded in formal course outcomes, vertical curriculum mapping, or studio requirements. The pattern reflects a broader distinction between isolated technology adoption and system-level digital transformation [7]–[10].

Faculty members rated curriculum updates reflecting emerging AI technologies lowest (M = 2.55, Somewhat Evident), while vertical integration across year levels was also limited (M = 2.70, Moderately Evident). Students rated curriculum updates (M = 2.92) and integration into learning outcomes (M = 2.98) as Moderately Evident. The findings support competency mapping across year levels, explicit inclusion in course outcomes, and regular syllabus review informed by developments in architectural AI [1], [2], [6].

H. Teaching Strategies

Faculty members perceived AI integration in teaching strategies as Evident (M = 3.51), whereas students perceived it as Moderately Evident (M = 2.90). This difference suggests that faculty members recognized their instructional intentions and demonstrations more strongly than students experienced consistent AI-supported teaching. Research on technology integration similarly shows that instructional adoption may not translate into meaningful learner experience without coherent pedagogical design [7]–[9].

Faculty members rated promotion of critical evaluation of AI outputs highest (M = 3.75) and reported using AI in instructional demonstrations and design activities (M = 3.45–3.50). Students rated encouragement to use AI tools (M = 2.83) and AI-supported design exploration (M = 2.87) lower. Bloom's revised taxonomy supports instructional tasks that move from understanding and application toward analysis, evaluation, and creation [21], while experiential learning emphasizes active practice and reflection [18]. Faculty development should therefore focus on structured, visible, and repeatable AI-mediated studio pedagogy.

I. Learning Activities

Faculty members perceived AI integration in learning activities as Evident (M = 3.41), while students perceived it as Moderately Evident (M = 2.98). The difference suggests that AI was used in selected tasks but was not consistently experienced across subjects and studios. This pattern is consistent with higher education research showing that digital activities remain fragmented when implementation is not aligned across curriculum and instruction [7]–[9].

Faculty ratings were higher for encouraging AI use in architectural research tasks (M = 3.65) and experimentation with AI technologies (M = 3.50). Students reported limited participation in AI-assisted design projects (M = 2.89) and use of generative AI tools (M = 2.88). Repeated, scaffolded, and authentic activities are needed to develop competence through experience [18]. Studio projects that connect research, programming, visualization, critique, and revision can provide this progression [1], [2].

J. Assessment Practices

Faculty (M = 3.13) and students (M = 2.96) both perceived AI integration in assessment practices as Moderately Evident. The results indicate that AI was only partially reflected in grading, feedback, criteria, and assessment design. Concerns related to validity, fairness, transparency, and academic integrity often cause assessment reform to lag behind instructional adoption [9], [12], [13].

The assessment of ethical AI use received relatively higher ratings (faculty: M = 3.35; students: M = 3.08), as did evaluation of both process and AI-supported outputs (faculty: M = 3.30; students: M = 3.04). Faculty ratings were lower for using AI to evaluate outputs, provide feedback, and operationalize AI-integrated criteria (M = 2.95 for each). These findings support process-based rubrics, documentation of human and AI contributions, reflective justification of design decisions, and clear ethical standards [9], [12].

Assessment was among the least developed dimensions of integration. Redesign should focus on authentic tasks that evaluate architectural reasoning, design process, critical appraisal of AI outputs, attribution, and ethical decision-making rather than rewarding AI-generated products alone.

K. Digital Infrastructure and AI-Related Resources

Faculty (M = 3.01) and students (M = 3.03) both perceived digital infrastructure and AI-related resources as Moderately Evident. The nearly identical ratings indicate shared recognition that basic resources were available but were not yet sufficient for advanced AI-assisted design, computational modeling, and cloud-based collaboration. Digital transformation requires coordinated infrastructure, governance, access, and human capability [9], [10].

Faculty members rated access to AI design software (M = 2.75) and infrastructure upgrading (M = 2.80) lowest. Both groups also rated internet connectivity and laboratory adequacy at moderate levels. These constraints may limit repeated use of AI in studio courses and create unequal learning opportunities. The College should therefore assess hardware, software licensing, cloud access, connectivity, technical support, and responsible-use policies as part of an AI-readiness plan [9].

Infrastructure should be developed alongside curriculum and pedagogy. Investment in tools without faculty development, assessment standards, and implementation governance would be unlikely to produce sustained curricular transformation [9], [10].

L. Summary of the Perceived Extent of AI Integration

Table II shows that faculty (M = 3.18) and students (M = 2.96) both perceived the overall extent of AI integration as Moderately Evident. AI was therefore present across curriculum, teaching, activities, assessment, and resources, but it had not yet become fully institutionalized. The results describe the NEUST College of Architecture only and should not be generalized to other architecture programs without additional evidence [9], [10].

TABLE III. SUMMARY OF THE PERCEIVED EXTENT OF AI INTEGRATION IN THE NEUST COLLEGE OF ARCHITECTURE

AI Integration in NEUST College Of Architecture	Faculty		Students	
	Mean	Verbal Description	Mean	Verbal Description
Curriculum Content	2.82	Moderately Evident	2.95	Moderately Evident
Teaching Strategies	3.51	Moderately Evident	2.90	Moderately Evident
Learning Activities	3.41	Moderately Evident	2.98	Moderately Evident
Assessment Practices	3.13	Moderately Evident	2.96	Moderately Evident
Digital Infrastructure	3.01	Moderately Evident	3.03	Moderately Evident
Grand Mean	3.18	Moderately Evident	2.96	Moderately Evident

Faculty ratings were highest for teaching strategies (M = 3.51, Evident), while curriculum content received its lowest rating (M = 2.82, Moderately Evident). Students rated all domains as Moderately Evident, and digital infrastructure was the only domain in which the student mean slightly exceeded the faculty mean. These results indicate that current AI adoption is primarily instruction-led and requires stronger alignment among formal curriculum, learning activities, assessment, and institutional support [7], [9].

M. Differences in Self-Reported AI Competency Between Faculty Members and Students

Table III presents the comparison of faculty and student self-reported AI competency. No significant differences were found in AI literacy and foundational knowledge (p = .345) or ethical and responsible AI use (p = .135). Significant differences were found in AI-assisted design and technical skills (p = .006), AI-supported collaboration and communication (p = .002), and AI-driven innovation and problem-solving (p = .011), with faculty members reporting higher means in each applied domain.

TABLE IV. DIFFERENCES IN SELF-REPORTED AI COMPETENCY BETWEEN FACULTY MEMBERS AND STUDENTS

Competency Dimension	Faculty M	Student M	p	Decision
AI Literacy and Foundational Knowledge	4.21	4.03	.345	NS
AI-Assisted Design and Technical Skills	3.94	3.34	.006	Sig.
Ethical and Responsible AI Use	4.38	4.07	.135	NS
AI-Supported Collaboration and Communication	4.06	3.34	.002	Sig.
AI-Driven Innovation and Problem-Solving	3.95	3.34	.011	Sig.

Note. NS = not significant; Sig. = significant at $p < .05$.

The results indicate stronger alignment in foundational awareness and ethical understanding than in applied competency. The higher faculty ratings should not be interpreted as objective evidence of superior performance because both groups completed self-report measures. They do, however, identify areas in which students perceived a need for more structured practice, feedback, collaboration, and technical integration.

For the NEUST College of Architecture, the findings support a progression from foundational AI literacy and ethics toward applied design, collaboration, and problem-solving. This progression should be implemented through curriculum mapping, faculty-student co-learning, studio projects, and performance-based assessment. International frameworks similarly emphasize progressive development from understanding to application and creation within an enabling digital ecosystem [4], [5], [9], [14], [15].

N. Proposed Arki-Nexus AI Framework

Fig. 1 presents the proposed Arki-Nexus AI Framework developed from the study findings. The Arki-Nexus AI Framework was developed as an evidence-based curriculum model derived from the quantitative findings of this study. Rather than serving as a purely conceptual representation, the framework synthesized the empirical results obtained from the assessment of Artificial Intelligence (AI) competencies among architecture faculty members and students, the extent of AI integration within the curriculum, and the comparative analysis of faculty and student perceptions regarding AI implementation. Consequently, the framework addressed research problems by translating the identified strengths, gaps, and statistically significant differences into a structured model for curriculum innovation.

The findings revealed that both faculty members and students generally demonstrated satisfactory levels of AI competency, particularly in AI literacy and foundational knowledge and ethical and responsible AI use, indicating a positive institutional readiness for AI-enhanced architectural education. However, comparatively lower ratings in AI-assisted design and technical skills, AI-supported collaboration, AI-driven innovation, and several dimensions of curriculum integration suggest that AI competencies remain inconsistently embedded across instructional practices. Moreover, the significant differences between faculty and student assessments indicate varying perceptions regarding the actual implementation of AI within teaching, learning, and assessment processes. These findings highlight the necessity of a comprehensive framework capable of aligning competency development with curriculum design, instructional delivery, institutional resources, and professional expectations.

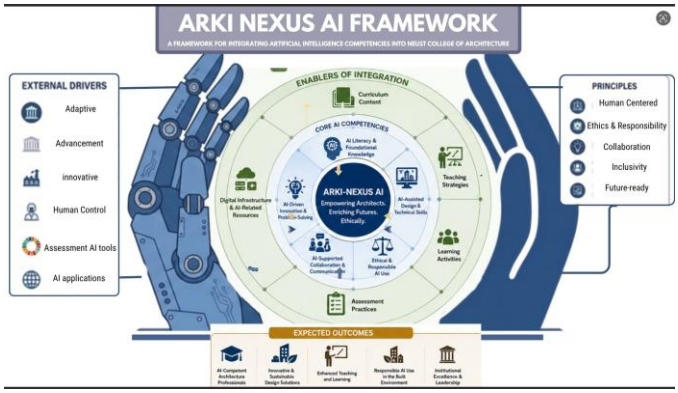


Fig. 1. Proposed Arki-Nexus AI framework developed from the study findings.

Accordingly, the Arki-Nexus AI Framework was formulated as a systematic response to these empirical findings. It provides a coherent mechanism through which AI competencies may be progressively developed and intentionally integrated across the architecture curriculum while ensuring alignment with educational objectives and the evolving demands of architectural practice.

1) Interpretation of the framework in relation to the research findings: The framework consists of five interconnected components: the External Drivers, Core AI Competencies, Enablers of Integration, Guiding Principles, and Expected Outcomes, which collectively demonstrate the dynamic process of AI integration in the architecture curriculum.

The External Drivers represent the contextual forces influencing curriculum transformation, including technological advancement, innovation, adaptability, AI applications, AI-assisted assessment tools, and continued human oversight. These drivers acknowledge that architectural education operates within a rapidly evolving technological environment where emerging AI technologies continuously reshape professional practice. Their inclusion reflects the study's findings that curriculum development must remain responsive to technological change while preserving the human-centered nature of architectural design.

At the core of the framework are the five AI competency domains, which directly correspond to the competency dimensions investigated in Research Problem 1. These include AI Literacy and Foundational Knowledge, AI-Assisted Design and Technical Skills, Ethical and Responsible AI Use, AI-Supported Collaboration and Communication, and AI-Driven Innovation and Problem-Solving. Positioning these competencies at the center of the framework signifies that AI competency development constitutes the primary objective of curriculum integration. The quantitative findings indicate that while respondents possess adequate foundational knowledge of AI, further curricular enhancement is necessary to strengthen students' ability to apply AI within authentic architectural design processes, interdisciplinary collaboration, and innovation-driven problem-solving. Consequently, the framework conceptualizes competency development as a continuous progression rather than a static outcome.

Surrounding these competencies are the Enablers of Integration, comprising curriculum content, teaching strategies, learning activities, assessment practices, and digital infrastructure and AI-related resources. These elements correspond directly to the dimensions examined in Research Problem 2 and emphasize that AI competency development cannot occur independently of institutional and pedagogical support. The findings revealed varying levels of AI integration across these curricular dimensions, suggesting that AI adoption remains uneven despite generally positive competency levels. Therefore, the framework advocates deliberate alignment between learning outcomes, instructional approaches, authentic learning experiences, assessment strategies, and institutional technological capacity to facilitate meaningful and sustainable AI integration throughout the curriculum.

The implementation of these curriculum enablers is governed by five Guiding Principles: Human-Centeredness, Ethics and Responsibility, Collaboration, Inclusivity, and Future-Readiness. These principles ensure that AI functions as an educational enabler rather than a substitute for architectural creativity and professional judgment. Their inclusion reflects the study's findings that ethical AI use received among the highest competency ratings, reinforcing the importance of embedding responsible AI practices alongside technical proficiency. The interaction among these framework components ultimately contributes to the Expected Outcomes, namely the development of AI-competent architecture graduates, enhanced teaching and learning experiences, responsible AI utilization, strengthened innovation and sustainability, and institutional leadership in AI-enabled architectural education. Collectively, these outcomes represent the desired impact of systematic AI integration and provide a strategic direction for curriculum enhancement within the architecture curriculum.

2) Theoretical foundation and alignment with existing literature: The Arki-Nexus AI Framework is theoretically anchored in Vygotsky's Social Constructivist Theory and the UNESCO AI Competency Framework for Teachers, each of which provides complementary perspectives on AI integration in higher education.

Vygotsky's Social Constructivist Theory emphasizes that knowledge is constructed through interaction, collaboration, and authentic learning experiences. This theoretical perspective supports the framework's emphasis on AI-supported

collaboration, project-based learning, and collaborative design processes, where AI functions as a cognitive tool that enhances rather than replaces human learning. By promoting collaborative inquiry and reflective practice, the framework reinforces the social nature of architectural education while integrating intelligent technologies into studio-based learning environments.

Furthermore, the framework reflects the principles articulated in the UNESCO AI Competency Framework, particularly its emphasis on responsible AI use, ethical governance, human-centered implementation, and lifelong learning. The inclusion of ethical AI use, inclusivity, and future-readiness acknowledges that AI integration must promote not only technological competence but also critical thinking, professional accountability, and equitable access to emerging technologies.

3) Contribution of the framework to architecture curriculum: The principal contribution of the Arki-Nexus AI Framework lies in its presentation of a comprehensive, empirically grounded model for integrating AI competencies into architecture curriculum. Unlike existing AI competency models that primarily emphasize technological literacy or digital skills, the proposed framework adopts a systems perspective by integrating competency development, curriculum implementation, instructional practice, ethical governance, institutional readiness, and graduate outcomes within a single architecture-specific model. This contribution is particularly significant because architecture curriculum requires the integration of creativity, technical competence, sustainability, collaborative practice, and ethical decision-making. Rather than positioning AI as a replacement for architectural expertise, the framework conceptualizes AI as an enabling technology that augments design thinking, evidence-based decision-making, and innovative problem solving while preserving the architect's central role as a creative and ethical professional.

From a curricular perspective, the framework provides academic leaders and curriculum developers with a structured roadmap for embedding AI competencies throughout architectural education. Its emphasis on alignment among curriculum content, pedagogy, assessment, and institutional infrastructure promotes a holistic approach to curriculum innovation that extends beyond isolated course revisions. Consequently, the framework contributes not only to the advancement of AI integration within the NEUST College of Architecture but also offers a transferable reference model for architecture schools and higher education institutions seeking systematic AI implementation within design-oriented programs.

4) Expert validation of the framework: The evaluation indicated a high level of content validity, demonstrating strong agreement that the framework presents a logically organized, empirically supported, and educationally relevant model for integrating AI competencies into architecture curricula. The experts affirmed that the framework effectively synthesizes the quantitative findings into an implementable curriculum model while maintaining consistency with contemporary theories of technology integration and higher education innovation. The validators further recommended strengthening provisions for continuous curriculum review, expanding faculty capability-building initiatives in AI pedagogy, fostering stronger partnerships with industry to support authentic AI applications, and establishing mechanisms for periodically updating digital infrastructure in response to rapidly evolving AI technologies.

Overall, the expert validation provides evidence that the Arki-Nexus AI Framework possesses strong conceptual coherence, empirical grounding, and practical applicability. As an evidence-informed curriculum model, it offers a robust foundation for systematically integrating AI competencies into architecture curriculum while ensuring that technological innovation remains aligned with creativity, ethical responsibility, sustainability, and the evolving competencies required of future architects.

IV. Conclusion and Recommendations

A. Conclusion

Among the 31 faculty members and 200 architecture students who participated in this single-institution study, both groups reported High overall AI competency, with faculty members obtaining a higher grand mean (M = 4.11) than students (M = 3.63). Faculty and student ratings were strongest in AI literacy and ethical AI use, while student ratings were Moderate in AI-assisted design, AI-supported collaboration, and AI-driven innovation. These findings indicate that perceived foundational readiness was stronger than perceived applied competency within the NEUST College of Architecture.

The overall perceived extent of AI integration was Moderately Evident for faculty (M = 3.18) and students (M = 2.96). Teaching strategies were the most visible domain in faculty ratings, but curriculum content, assessment, and infrastructure remained only moderately integrated. The evidence therefore suggests that AI adoption in the College was emerging and primarily instruction-led rather than fully embedded across formal curriculum and institutional systems.

Significant faculty-student differences were found in the three applied competency domains, while differences in AI literacy and ethical AI use were not significant. These findings informed the proposed Arki-Nexus AI Framework. The study did not validate the framework, and the conclusions apply only to the respondents and institutional context examined.

B. Recommendations

The College should map AI competencies across program and course outcomes and integrate them progressively into architectural programming, design studios, CAD/BIM workflows, visualization, simulation, research, and technical subjects. Curriculum revision should move beyond theoretical exposure toward performance-based and project-driven learning.

Faculty development and faculty-student co-learning should be institutionalized. Workshops, mentoring, collaborative studio projects, and communities of practice should focus on responsible AI use, generative and computational design,

collaborative platforms, critical evaluation, attribution, and human oversight.

Assessment should be redesigned using process-based rubrics, design documentation, oral defense, reflective justification, and transparent declaration of AI assistance. The College should also conduct an AI-readiness audit covering hardware, connectivity, software licensing, cloud platforms, technical support, data governance, accessibility, and equitable student access.

Before institutional adoption, the proposed Arki-Nexus AI Framework should undergo structured expert review, feasibility assessment, and pilot implementation. Evaluation should determine whether the framework improves demonstrated competencies, curriculum alignment, student learning, ethical practice, and resource efficiency within the NEUST College of Architecture.

C. Limitations and Directions for Future Research

The findings were based on self-rated Likert-scale responses from 31 faculty members and 200 students in one college and may have been affected by social desirability, recall error, or overestimation of competency. The cross-sectional design did not measure change over time, and the unequal group sizes may have influenced the comparison. Future studies should include performance-based design tasks, portfolios, software demonstrations, interviews, classroom or studio observations, and multi-institution samples. The proposed framework should also undergo expert validation, pilot implementation, feasibility testing, and outcome evaluation before broader use.

ACKNOWLEDGMENT

The author gratefully acknowledges the faculty members and architecture students of the Nueva Ecija University of Science and Technology who participated in the study, as well as those who assisted in survey administration and data collection.

REFERENCES

[1] S. Jin, H. Tu, J. Li, Y. Fang, Z. Qu, F. Xu, K. Liu, and Y. Lin, "Enhancing architectural education through artificial intelligence: A case study of an AI-assisted architectural programming and design course," Buildings, vol. 14, no. 6, art. no. 1613, 2024, doi: 10.3390/buildings14061613.

[2] G. Özorhon, D. Nitelik Gelirli, G. Lekesiz, and C. Müezzinoğlu, "AI-assisted architectural design studio (AI-a-ADS): How artificial intelligence join the architectural design studio?" International Journal of Technology and Design Education, vol. 35, pp. 1999–2023, 2025, doi: 10.1007/s10798-025-09975-0.

[3] A. Odiah and S. D. Gosling, "Laying the foundations for using generative AI images in architectural research: Do images convey the intended spaces and ambiances?" Architectural Intelligence, vol. 3, art. no. 35, 2024, doi: 10.1007/s44223-024-00076-x.

[4] F. Miao, K. Shiohira, and N. Lao, AI Competency Framework for Students. Paris, France: UNESCO, 2024.

[5] F. Miao and M. Cukurova, AI Competency Framework for Teachers. Paris, France: UNESCO, 2024.

[6] A. Iranmanesh and P. Lotfabadi, "Critical questions on the emergence of text-to-image artificial intelligence in architectural design pedagogy," AI & Society, vol. 40, no. 5, pp. 3557–3571, 2025, doi: 10.1007/s00146-024-02111-x.

[7] T. Bates, C. Cobo, O. Mariño, and S. Wheeler, "Can artificial intelligence transform higher education?" International Journal of Educational Technology in Higher Education, vol. 17, art. no. 42, 2020, doi: 10.1186/s41239-020-00218-x.

[8] H. Crompton and D. Burke, "Artificial intelligence in higher education: A systematic review," Computers and Education: Artificial Intelligence, vol. 1, art. no. 100008, 2020, doi: 10.1016/j.caeai.2020.100008.

[9] OECD, OECD Digital Education Outlook 2023: Towards an Effective Digital Education Ecosystem. Paris, France: OECD Publishing, 2023, doi: 10.1787/c74f03de-en.

[10] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, "Systematic review of research on artificial intelligence applications in higher education—Where are the educators?" International Journal of Educational Technology in Higher Education, vol. 16, art. no. 39, 2019, doi: 10.1186/s41239-019-0171-0.

[11] High-Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI. Brussels, Belgium: European Commission, 2019.

[12] W. Holmes, K. Porayska-Pomsta, K. Holstein, E. Sutherland, T. Baker, S. Buckingham Shum, O. C. Santos, M. M. T. Rodrigo, M. Cukurova, I. I. Bittencourt, and K. R. Koedinger, "Ethics of AI in education: Towards a community-wide framework," International Journal of Artificial Intelligence in Education, vol. 32, no. 3, pp. 504–526, 2022, doi: 10.1007/s40593-021-00239-1.

[13] UNESCO, Recommendation on the Ethics of Artificial Intelligence. Paris, France: UNESCO, 2021.

[14] D. Long and B. Magerko, "What is AI literacy? Competencies and design considerations," in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020, pp. 1–16, doi: 10.1145/3313831.3376727.

[15] D. T. K. Ng, J. K. L. Leung, S. K. W. Chu, and M. S. Qiao, "AI literacy: Definition, teaching, evaluation and ethical issues," Proceedings of the Association for Information Science and Technology, vol. 58, no. 1, pp. 504–509, 2021, doi: 10.1002/pra2.451.

[16] D. F. Polit, C. T. Beck, and S. V. Owen, "Is the CVI an acceptable indicator of content validity? Appraisal and recommendations," Research in Nursing & Health, vol. 30, no. 4, pp. 459–467, 2007, doi: 10.1002/nur.20199.

[17] M. Tavakol and R. Dennick, "Making sense of Cronbach's alpha," International Journal of Medical Education, vol. 2, pp. 53–55, 2011, doi: 10.5116/ijme.4dfb.8dfd.

[18] D. A. Kolb, Experiential Learning: Experience as the Source of Learning and Development. Englewood Cliffs, NJ, USA: Prentice-Hall, 1984.

[19] L. S. Vygotsky, Mind in Society: The Development of Higher Psychological Processes, M. Cole, V. John-Steiner, S. Scribner, and E. Souberman, Eds. Cambridge, MA, USA: Harvard University Press, 1978.

[20] G. Siemens, "Connectivism: A learning theory for the digital age," International Journal of Instructional Technology and Distance Learning, vol. 2, no. 1, pp. 3–10, 2005.

[21] L. W. Anderson and D. R. Krathwohl, Eds., A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. New York, NY, USA: Longman, 2001.

Secure Heart Disease Status Classification Using Machine Learning and Blockchain-Based EHR Integrity Verification

Haya H. Alsharif, Sabah M. Alzahrani
Department of Computer Science-College of Computers and Information Technology, Taif University, Taif, Saudi Arabia

***Abstract*—Heart disease remains one of the leading causes of mortality worldwide, motivating the development of accurate and scalable screening-support tools. This study presents an integrated framework for heart-disease status classification using the Heart 2020 Cleaned dataset, derived from the CDC Behavioral Risk Factor Surveillance System (BRFSS), a telephone-based survey of self-reported health information rather than clinical or electronic-health-record data. The task is therefore to classify a respondent's self-reported heart-disease status, not to predict future disease onset. The pipeline includes categorical encoding, feature scaling, recursive feature elimination, and class-imbalance handling using BorderlineSMOTE, SMOTETomek, and SMOTEENN. Five classifiers were evaluated: Random Forest, XGBoost, LightGBM, k-Nearest Neighbours, and a feed-forward deep neural network. Sampling strategy affected reported performance. Under the methodologically valid post-split setting, in which resampling is applied only to the training data, the models achieved an AUC of approximately 0.85, comparable to prior work on the same dataset. Pre-split global resampling produced much higher scores, with Random Forest reaching 97.14% accuracy, an F1-score of 0.9735, and an AUC of 0.9965 under SMOTEENN. However, these results are optimistically biased because resampling before splitting leaks synthetic information into the test set. The system also includes a blockchain-based integrity layer that stores full records in the application database while anchoring SHA-256 digests and verification metadata in an Ethereum-compatible environment. Overall, this prototype demonstrates the practical integration of established machine-learning classification and blockchain-supported integrity-verification components.**

Keywords*—Heart-disease status classification; machine learning; ensemble model; random forest; XGBoost; LightGBM; Blockchain; Electronic Health Records (EHR)

I. Introduction

Artificial Intelligence (AI) has become a central component of modern healthcare, where it supports data-driven decision-making in disease diagnosis, risk prediction, prognosis, and personalized treatment planning [1]. AI systems are capable of processing vast and varied clinical datasets, detecting hidden patterns, and helping healthcare professionals make timely, informed decisions. Within this broader domain, Machine Learning (ML) has emerged as one of the most effective AI paradigms for predictive healthcare, with successful applications reported across disease screening, clinical risk stratification, and diagnostic support systems [2].

Cardiovascular disease is one of the main causes of mortality, and it constitutes one of the greatest health problems in the world. It is also one of the clinical fields where AI for diagnosis and health-data analysis can make a big difference. Since cardiovascular diseases account for a significant proportion of global mortality, there has been a rising interest in the use of ML tools for scalable population-level health assessment. ML models are increasingly used to support heart-disease status classification by learning from demographic, behavioral, and self-reported health variables that may be difficult to integrate manually in routine care [3]. These capabilities are especially relevant for survey-based health assessment, where status classification of reported heart-disease cases can support population-level decision-making and public-health planning [4].

Meanwhile, as health care is becoming more reliant on digital health information, privacy, integrity, interoperability, and patient control issues have emerged. Although EHR systems are essential for supporting clinical decisions, many conventional centralized architectures remain vulnerable to unauthorized access, limited auditability, interoperability constraints, and single points of failure [5]. Technologies such as smart contracts, decentralized ledgers, and IPFS-supported storage can improve traceability, tamper resistance, patient-centric access control, and secure health-record sharing [6]. But the technologies are typically designed as secure data-management systems, not a comprehensive framework that brings together heart-disease status classification and secure EHR integrity verification.

Within the realm of heart disease research, several recent studies have proposed different ML models, including gradient boosting, ensemble learning, deep learning, and NLP-based approaches for extracting cardiovascular risk factors from structured and unstructured data [3,7]. Although these studies have immense potential for the use of ML in HD prediction, they also have certain key limitations. Many works focus either on predictive performance alone or on secure record management alone, without integrating both into a unified healthcare framework [4]. In addition, class imbalance, dataset heterogeneity, lack of interpretability, and variation in evaluation practice between studies remain a problem in the literature.

This study develops and evaluates an integrated framework that combines machine-learning-based heart-disease status classification with blockchain-supported EHR integrity verification. Recent literature related to heart-disease classification and blockchain in healthcare data management

indicates that there are a number of limitations and aspects that can be improved. Class imbalance and minority-class sensitivity remain persistent issues, especially in large population-scale studies, and are often overlooked in studies that emphasize overall accuracy. In some existing works, only a small number of models are evaluated, or the studies consider only structured or unstructured data, which limits the possibility of cross-model comparison in a unified pipeline. Moreover, most of the existing literature on blockchain-based EHRs gives high priority to security, privacy, and access management. Only a few works have integrated classification analytics with blockchain for real-time heart-disease status classification and EHR integrity verification. Many secure-health data-management systems increase auditability and patient control but may not be designed for heart-disease status classification within a healthcare-oriented decision-support prototype based on large-scale behavioral-risk survey data.

Based on the above research gaps, this study addresses the following research questions:

- RQ1: How accurately can machine-learning models classify a respondent's self-reported heart-disease status, and which model performs best?
- RQ2: How does class-imbalance handling affect classification performance?
- RQ3: Does the timing of resampling (before vs. after the train/test split) change the reported results?
- RQ4: Can a blockchain layer verify the integrity of health records and detect tampering?

To improve the accuracy and methodological soundness of heart-disease status classification while supporting secure patient record management, this study addresses these limitations through the following contributions:

- A reproducible framework for heart-disease status classification is developed using the BRFSS-based Heart 2020 Cleaned dataset, incorporating preprocessing, feature engineering, recursive feature elimination, and explicit class-imbalance handling, as a practical integration of established techniques rather than a novel algorithm.
- Five established classifiers (Random Forest, XGBoost, LightGBM, k-Nearest Neighbours, and a feed-forward DNN) are compared under a common pipeline to identify the most reliable configuration for heart-disease status classification.
- An analysis of class-imbalance handling techniques and the effect of resampling timing (post-split versus pre-split) on classification performance is presented, explicitly identifying pre-split resampling as a source of data leakage and optimistic bias.
- A blockchain-based integrity-verification layer is integrated into the prototype, with full patient records stored in the application database and SHA-256 digests anchored through an Ethereum-compatible environment to provide tamper-evident verification and traceability; access control is enforced at the API layer rather than on-chain.

This research therefore presents a prototype/feasibility study that integrates machine-learning-based heart-disease status classification with a blockchain-based EHR integrity layer. The system is designed to deliver a reproducible, healthcare-oriented decision-support framework, and its primary contribution is the practical integration of established components rather than a novel algorithm or blockchain architecture. The rest of the study is organized as follows. Section II reviews machine-learning-based heart-disease classification and blockchain-based EHR systems. Section III describes the materials and methods. Section IV presents experimental results. Section V discusses the results and compares them with existing works. Finally, Section VI concludes the study and suggests directions for future research.

II. Related Work

A lot of research work has investigated machine-learning methods for predicting and classifying cardiovascular risk from various kinds of data such as population surveys, clinical tabular datasets, physiological signals, imaging, free-text electronic health records, and omics data.

In studies based on large population surveys, Chen et al. [8] built a comparative framework on the Heart_2020_cleaned dataset derived from BRFSS, examining how Boruta feature selection, class balancing, and model tuning affect the performance of decision trees, random forests, logistic regression, and LightGBM. Based on their results, they obtained that LightGBM gives the highest overall accuracy. Age category and rate of health were the most significant predictors. Also, it should be noted that the specificity was always greater than the sensitivity. Febriyanti and Baita [9] focus on the severe class imbalance in Heart_2020_cleaned and use random under-sampling together with tuned SVM and CART decision trees; they demonstrate that under-sampling reduces overall accuracy but substantially improves minority-class recall and F1-score, explicitly highlighting the trade-off between balanced performance and global accuracy in this dataset. Tompra et al. [10] examine the 2021 BRFSS heart-disease data with extensive feature engineering and a spectrum of resampling methods such as SMOTE-ENN and ADASYN applied only to the training set; their best CatBoost configuration achieves very high recall but low precision, underscoring the difficulty of avoiding false negatives in highly imbalanced population-level prediction.

Osei-Nkwantabisa and Ntumy [11] combine multiple UCI heart-disease sources into a dataset of 1,190 patients, apply grid-search tuning to logistic regression, KNN, SVM and an ANN, and report KNN as the best performer with balanced precision and recall, though still below the very high scores claimed in some deep-learning works. Bhatt et al. [12] analyze a larger Kaggle cardiovascular disease dataset of 70,000 patients, using k-modes clustering to transform categorical variables before training decision trees, random forests, MLPs and XGBoost; they find that an MLP trained on clustered data achieves the highest accuracy and AUC, suggesting that clustering can be a useful preprocessing step for fully categorical risk-factor data. Almulihi et al. [13] construct a deep stacking ensemble that uses CNN-LSTM and CNN-GRU hybrids as base learners with an

SVM meta-classifier, evaluating the system on a large Kaggle heart-disease dataset and on Cleveland, and they report strong gains over individual models, particularly on the smaller dataset.

There has been a rise in the use of blockchain technology to support safer and more interoperable electronic health record (EHR) systems in healthcare systems in recent years. Many of the efforts were focused on developing blockchain-based architectures for healthcare records. Tahir et al. [5] developed an Ethereum-based EHR management framework where encrypted records are stored off-chain in IPFS and hashes plus metadata are maintained on-chain via smart contracts, with a web interface enabling hospitals and patients to upload and retrieve records securely. They showed low gas consumption and their response time at a load of simulated milliseconds is excellent, which implies the availability for real-world usage. Mandarino et al. [14] focused on cost and latency by introducing EdgeHR, a public-Ethereum dApp that keeps medical data on edge devices while recording hashed references and access policies on-chain; through careful use of smart-contract design patterns, they demonstrate significant gas savings while preserving integrity and auditability. These are examples of other healthcare systems that are securing only metadata and access logs on the blockchain, with the large medical data files stored off-chain.

Another line of work is concentrating on providing access control for EHRs using privacy-preserving, fine-grained parameters. Ali et al. [15] presented a Hyperledger Fabric framework that integrates attribute-based encryption and signatures with a multi–certificate authority design, avoiding a single point of trust and enabling cross-organizational access control. Using their performance measures with Hyperledger Caliper, they demonstrate that their experiments indicate that their MedRec-previously proposed system comparisons achieve higher throughput and lower overheads.

More recently, Kumari et al. [16] proposed BL-MVC, a blockchain-enabled framework that combines Ethereum-based blockchain storage, IPFS, Solidity smart contracts, a React-based front end, and a FastAPI back end, with a Majority Voting Classifier for heart disease prediction. Several machine learning models were used for these predictive models, including Logistic Regression, MLP, AdaBoost, CatBoost, and XGBoost. The blockchain layer helps in hosting patient records, which are decentralized and tamper-proof. The health care data set used for the study is based on the BRFSS (Behavior Risk Factor Surveillance System); predicting cardiac disease from the outcome column is a more straightforward example of combining data management improvements obtained with blockchain with embedded predictive analysis – aspects that are relevant for systems intended to support data protection improvements but also clinical decision support in a single system.

III. METHODOLOGY

The main procedure for classifying heart-disease status in this project involves integrating blockchain technology with machine-learning models that process population-level health data and output a status classification for each individual. The objective is twofold: 1) to use routinely collected cardiovascular risk indicators — such as demographic, behavioral, and self-reported health variables — to provide scalable, data-driven support for the classification of individuals according to their self-reported heart-disease status; and 2) to manage Electronic Health Records (EHRs) securely in the application database while using blockchain-based anchoring of cryptographic digests to provide tamper-evident integrity verification and traceability.

The general structure of the proposed system is shown in Fig. 1. The framework begins with patient information and health-related input data received through the user interface and coordinated by the application server. The machine-learning module processes the relevant data, applies preprocessing, and generates heart-disease status classifications. At the same time, patient EHRs are stored in the application database, while cryptographic digests and related identifiers are optionally anchored on an Ethereum-compatible blockchain for integrity verification. The following are key components of the system:

A. Key components of the system

1) User Interface (GUI): A front-end interface through which healthcare professionals and authorized users can enter patient data, request heart-disease status classifications, and securely access or update EHR information.

2) Machine learning model: Applies a chosen classification model (Random Forest, XGBoost, LightGBM, k-NN, and feed-forward DNN), and pre-processing, training, tuning, and classification.

3) Blockchain storage module: Provides a tamper-evident integrity layer by storing record hashes and verification-related metadata, ensuring traceability and protection against unauthorized modification; only cryptographic digests are stored on-chain, while full records remain in the application database.

4) Application server: Coordinates communication between the GUI, the ML module, the database, and the blockchain layer, while handling authentication, authorization, and the return of status-classification results and record-related responses to the GUI.

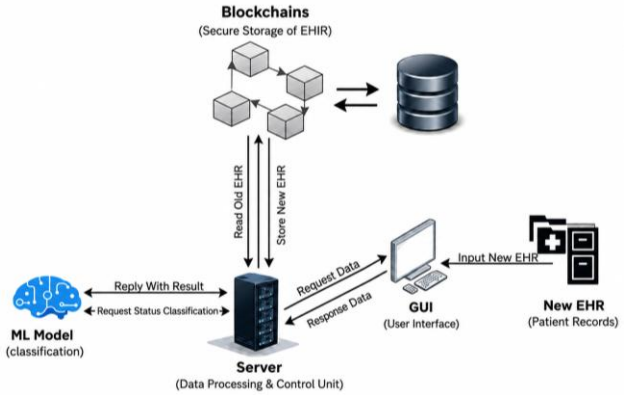


Fig. 1. General framework of the proposed system that secures EHRs and classifies heart-disease status.

B. Proposed Method

The proposed method is a machine-learning pipeline designed to classify heart-disease status using the Heart 2020 –

Cleaned (BRFSS) dataset. The method begins with preprocessing the original survey data, followed by development of multiple classification models and, finally, comparing the models' performance using standard metrics. The aim is to find an effective model to classify individuals as positive-class (heart-disease present) or negative-class (heart-disease absent) from demographic, behavioral, and self-reported health features. The general process for the proposed method is shown in Fig. 2.

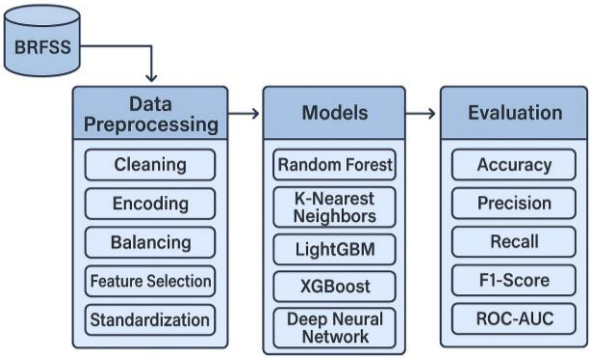


Fig. 2. Proposed machine learning method.

The proposed method consists of the following main stages:

1) Data preprocessing: Raw BRFSS data are prepared before model training to ensure that the dataset is suitable for machine-learning analysis. This stage includes the following steps:

2) Cleaning: The dataset is checked and prepared to remove inconsistencies and ensure that the values are ready for analysis.

3) Encoding: Categorical variables are transformed into numerical form so they can be processed by the classification models.

4) Balancing: Class imbalance in the target variable is addressed using resampling techniques to improve the representation of minority heart-disease cases.

5) Feature selection: The most informative variables are identified and retained to reduce redundancy and improve model performance.

6) Standardization: Numerical features are scaled to a common range so that differences in magnitude do not adversely affect model training.

7) Model development: The processed data are then used to train multiple classification models, namely Random Forest, k-Nearest Neighbours, LightGBM, XGBoost, and a feed-forward Deep Neural Network.

8) Performance evaluation: The trained models are assessed using standard classification metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, to compare their effectiveness in heart-disease status classification.

9) Model comparison and selection: Based on the evaluation results, the relative strengths and weaknesses of the models are analyzed to identify the most suitable approach for the classification task.

C. Dataset

The dataset used in this study is the Heart 2020 Cleaned dataset, also known as Personal Key Indicators of Heart Disease, obtained from Kaggle [17]. It is derived from the 2020 wave of the U.S. Centers for Disease Control and Prevention (CDC) Behavioral Risk Factor Surveillance System (BRFSS) survey. BRFSS is an annual, telephone-based health survey in which adults across U.S. states and territories self-report information on demographic characteristics, health behaviours, and chronic conditions. The dataset contains 319,795 records and 18 variables, including the binary target variable HeartDisease, which indicates whether a respondent has ever been told by a health professional that they have coronary heart disease or myocardial infarction. The features are primarily categorical or ordinal and include age category, sex, body mass index (BMI), smoking status, alcohol consumption, physical activity, stroke history, diabetic status, kidney disease, general health, mental and physical health, sleep duration, and other health-related indicators. The dataset is notably imbalanced, with approximately 8.6% positive cases and 91.4% negative cases, making it suitable for evaluating machine-learning methods for imbalanced binary status classification. Because the label reflects a previously reported diagnosis, the task is the classification of existing self-reported heart-disease status rather than the prediction of future disease onset. This is a telephone-based public-health survey dataset and is not a clinical EHR dataset. Table I summarizes the variables included in the Heart 2020 Cleaned dataset and the data type for each of the features used in the proposed heart-disease status-classification framework.

TABLE I. FEATURES IN THE HEART 2020 CLEANED DATASET AND THEIR DATA TYPES

Feature	Type
HeartDisease	Binary (Categorical)
BMI	Continuous
Smoking	Binary (Categorical)
AlcoholDrinking	Binary (Categorical)
Stroke	Binary (Categorical)
PhysicalHealth	Continuous (Numeric, days)
MentalHealth	Continuous (Numeric, days)
DiffWalking	Binary (Categorical)
Sex	Binary (Categorical)
AgeCategory	Ordinal Categorical
Race	Nominal Categorical
Diabetic	Categorical (Multi-class)
PhysicalActivity	Binary (Categorical)
GenHealth	Ordinal Categorical
SleepTime	Continuous
Asthma	Binary (Categorical)
KidneyDisease	Binary (Categorical)
SkinCancer	Binary (Categorical)

1) Target class distribution and imbalance analysis: Fig. 3 presents the distribution of the target variable, HeartDisease. This data is imbalanced, with 91.4% of people in the non-heart-disease class and 8.6% in the heart-disease class. This is a large imbalance between the two classes and can lead to classification models favouring the majority class, thus decreasing the chances of getting positive heart disease cases classified correctly. Therefore, class-balancing techniques were necessary to improve the representation of the minority class and support more reliable model training.

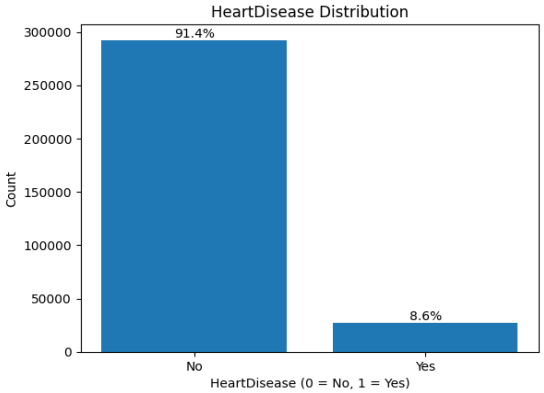


Fig. 3. Heart-disease class distribution

2) Distribution Analysis of Body Mass Index (BMI): Fig. 4 shows the distribution of BMI in the dataset. Most values are concentrated between 20 and 35, indicating that a large proportion of respondents fall within the normal, overweight, and moderately obese ranges. The distribution is right-skewed, with fewer observations extending toward higher BMI values. This suggests that although most respondents have moderate BMI levels, the dataset also includes individuals with substantially elevated BMI. The observed variation indicates that BMI may serve as an important feature in heart-disease status classification.

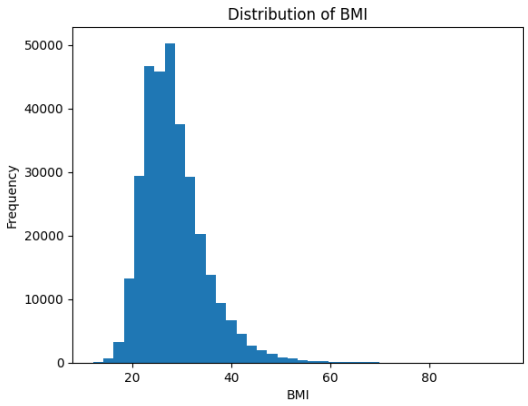


Fig. 4. Distribution of BMI

3) Distribution analysis of physical health: Fig. 5 presents the distribution of PhysicalHealth, which represents the number of days during the past 30 days in which a respondent experienced poor physical health. The distribution is clearly right-skewed, with a very large proportion of observations concentrated at 0 days, showing that most respondents reported no poor physical health problems during the reference period. At the same time, values are spread across the rest of the range, with a noticeable concentration at 30 days, indicating that some individuals experienced poor physical health throughout the entire month. This pattern highlights the variability of the feature and suggests that it may provide useful information for distinguishing different levels of reported heart-disease status.

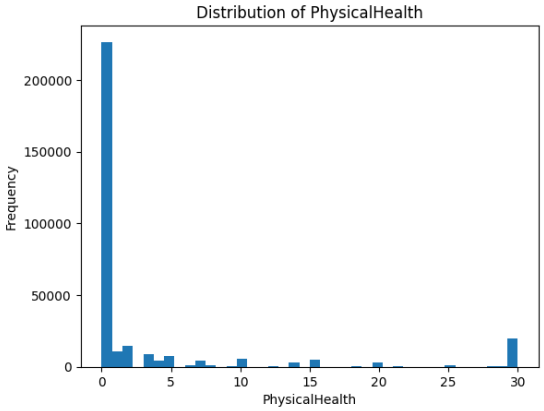


Fig. 5. Distribution of PhysicalHealth

4) Distribution analysis of mental health: Fig. 6 shows the distribution of MentalHealth (number of days in the last 30 days that the respondent did not feel mentally well). The distribution is strongly right-skewed, with most observations concentrated at 0 days, indicating that many respondents reported no poor mental health problems during the reference period. Meanwhile, the other values have been distributed across the range, with most of them falling in the range of 30 days, indicating that some people did not achieve good mental health during the month at all. This pattern shows the fluctuations of the feature, which suggests that it could be relevant for heart-disease status classification.

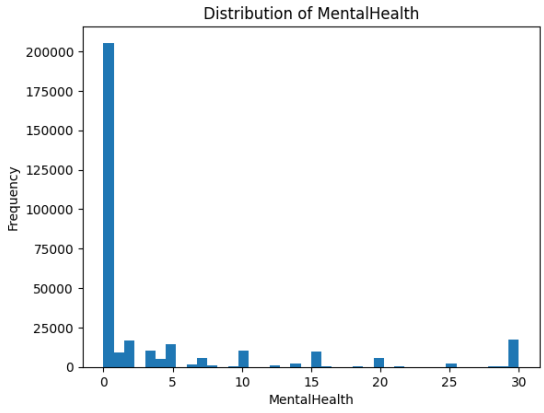


Fig. 6. Distribution of MentalHealth

5) Distribution analysis of sleep duration: The distribution of SleepTime (self-reported hours of sleep in 24 hours) is shown in Fig. 7. The distribution is centered around 7 hours,

which is the most frequent sleep time that the respondents reported. The majority of observations are in the 5-9 hour range, indicating that most participants' sleep times were close to the recommended range. The distribution also has a slight right tail, indicating that some respondents reported unusually long sleep durations. This variation suggests that sleep duration could provide lifestyle-related information for classifying heart-disease status.

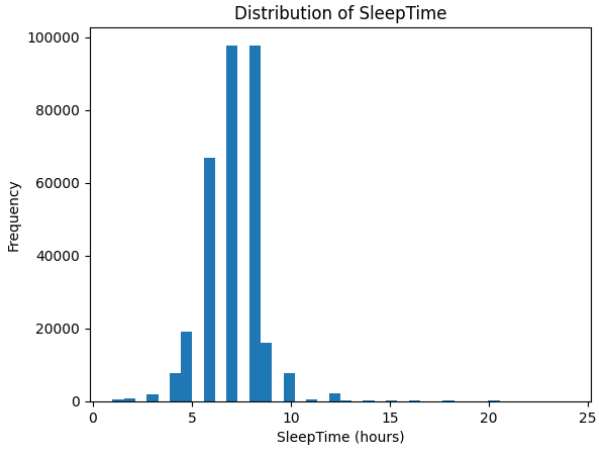


Fig. 7. Distribution of sleep time.

6) Multivariate correlation analysis: The Pearson correlation heatmap of numerical features BMI, PhysicalHealth, MentalHealth, and SleepTime is provided in Fig. 8. The majority of pairwise correlations are weak to moderate, suggesting that these variables measure different aspects of the respondents' health status, and that they offer relatively unique information to the classification models. PhysicalHealth has the highest positive correlation with MentalHealth, indicating that there is a strong relationship between bad physical and mental health. The rest of the correlations are comparatively small, and there is no indication of extreme multicollinearity among pairs of variables. This is beneficial for model creation since it decreases the quantity of redundant numeric features.

7) Categorical distribution of smoking status: The distribution of the Smoking feature in the data set is given in Fig. 9. The chart indicates that there are both smokers and non-smokers, with non-smokers being the majority. However, the number of respondents with a history of smoking is also significant, suggesting that smoking is a prevalent feature among the study population. This distribution highlights the relevance of smoking as an important behavioral risk factor and supports its inclusion as a potential predictor in the heart-disease classification models.

8) Bivariate analysis: smoking vs. heart disease: A stacked-proportions bivariate comparison of Smoking status and HeartDisease is shown in Fig. 10. The chart illustrates the proportions of smokers and non-smokers with and without heart disease. The proportion of heart disease cases is visibly higher among smokers than among non-smokers, suggesting a stronger association between smoking and adverse cardiovascular outcomes. Since the values are displayed as normalized ratios, Fig. 10 can be used to compare the two groups irrespective of differences in sample size. This pattern helps to confirm the significance of smoking status as one of the important factors in the heart-disease classification models.

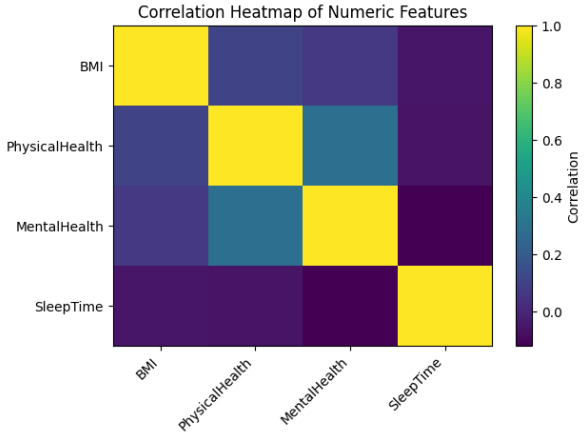


Fig. 8. Correlation heatmap of numeric features.

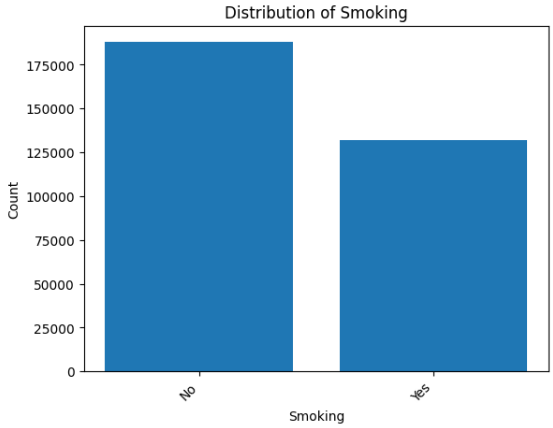


Fig. 9. Distribution of the smoking categorical feature.

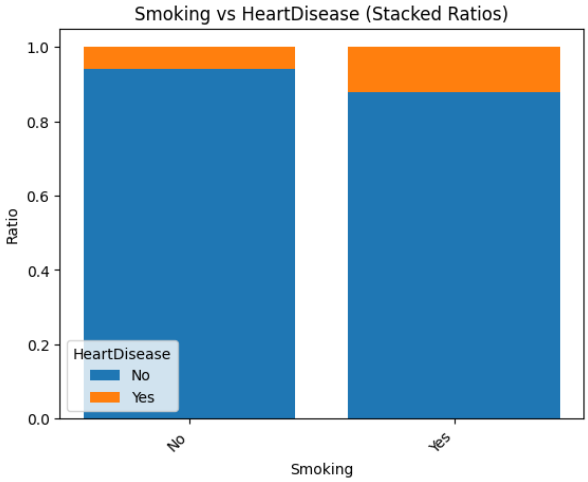


Fig. 10. Smoking versus HeartDisease (stacked ratios).

D. Data Preprocessing and Feature Selection

Preprocessing operations are performed on the data, such as handling missing values, normalizing numeric attributes, and encoding categorical attributes. Recursive Feature Elimination (RFE) is used to identify the most crucial attributes that boost model performance.

1) Data preprocessing and transformation: Within the first dataset, there were different types of data: categorical strings, ordinal values, and continuous numeric values. To fulfill the mathematical requirements of classifiers, all the features were transformed into numbers. The target variable HeartDisease was turned into a binary indicator where "Yes" was mapped to 1 (presence of heart disease) and "No" was mapped to 0 (absence).

The binary categorical variables (Smoking, AlcoholDrinking, Stroke, DiffWalking, PhysicalActivity, Asthma, KidneyDisease, SkinCancer) were coded as "Yes" = 1 and "No" = 0. Sex was coded as 1 for "Male", and 0 for "Female". The Diabetic attribute was dealt with separately: the answers for 'borderline diabetes' and 'gestational diabetes' (Yes (during pregnancy)) were combined to form the binary risk factor for the diabetic condition before the standard mapping.

Ordinal variables were coded in order to maintain the underlying order of the variables. The AgeCategory ranges were mapped to approximate midpoint values (e.g., "18-24" was approximated by 21 and "80 or older" was approximated by 85). In turn, GenHealth's responses were coded on a scale from 5 ("Excellent") to 1 ("Poor"). The nominal categorical variable Race was encoded using one-hot encoding with pandas.get_dummies function, generating separate binary indicator columns for each racial category to avoid creating false ordinal relationships between the categories. Finally, numeric data such as BMI, PhysicalHealth, MentalHealth, and SleepTime were retained as numeric data and standardized before being put into the model.

2) Feature engineering and Recursive Feature Elimination (RFE): Beyond the above-mentioned attributes, two composite features were created to reflect the combined effect of the risk factors. Chronic Condition Count was derived as the sum of binary indicators for Stroke, Diabetic, Asthma, KidneyDisease, and SkinCancer. This feature measures the number of chronic comorbidities that an individual has. Unhealthy Behavior Count was calculated as the sum of Smoking, AlcoholDrinking, physical inactivity (PhysicalActivity was inverted), and poor general health (GenHealth <= 2). The consolidated signals from these engineered features gave information on the patient's health status.

Specifically, the RFE feature selection procedure retained exactly 15 features out of the initial 24 columns. The 15 retained features are: 1) Stroke, 2) Sex, 3) AgeCategory, 4) Diabetic, 5) GenHealth, 6) Asthma, 7) KidneyDisease, 8) SkinCancer, 9) Race_American Indian/Alaskan Native, 10) Race_Asian, 11) Race_Black, 12) Race_Hispanic, 13) Race_Other, 14) Race_White, and 15) ChronicCondCount. The target count of 15 features was determined heuristically to preserve the essential demographic, clinical, and lifestyle factors while maintaining computational efficiency and preventing model overfitting. RFE utilized a weighted Logistic Regression model as the base estimator.

3) Class imbalance handling and sampling strategy: There is a significant class imbalance in the data that shows only about 8.6% positive heart disease from the BRFSS data. To address this problem, three innovative hybrid sampling algorithms were systematically compared: SMOTEENN, BorderlineSMOTE, and SMOTETomek.

a) SMOTEENN (Synthetic Minority Oversampling Technique + Edited Nearest Neighbors): This method is a combination of oversampling of the minority class with SMOTE and data-cleaning of Edited Nearest Neighbors (ENN). SMOTE generates minority samples by performing interpolation between minority samples, and ENN removes the majority samples that are wrongly classified by their nearest neighbors. It has two-way action, which can not only balance the class distribution but can also remove noisy and ambiguous majority samples, thus achieving class cleanliness.

b) BorderlineSMOTE: Unlike standard SMOTE, which generates synthetic samples uniformly, BorderlineSMOTE focuses specifically on minority instances that lie near the decision boundary (borderline instances). Because borderline samples are more likely to be misclassified, oversampling the region allows the classifier to draw a better and more reliable line between classes.

c) SMOTETomek: This method combines SMOTE oversampling with Tomek Links removal. Tomek Links are pairs of instances from different classes that are nearest neighbors to each other. By removing the majority class instances that form Tomek Links, this technique cleans the class overlap after oversampling, leading to well-defined class clusters.

Crucially, the impact of the timing of sampling application relative to the data split was investigated. Two distinct strategies were compared:

d) Post-split sampling: Sampling methods were only used on the training set after the train_test_split was performed. This ensures that the test set is an untouched, original imbalanced population distribution, thus preventing data leakage. However, it may also limit the model training.

e) Pre-split sampling: Sampling methods were utilized on the overall dataset before the division of training and testing sets. This method ensures that the training and test folds both have equal distributions of classes. Although it has been pointed out that this technique exposes the test set to synthetic data, it allows the model to learn from a completely balanced feature space, and the model's performance is measured on a balanced test set. Note, however, that this configuration introduces data leakage, producing optimistically biased results that should not be interpreted as valid estimates of real-world performance.

E. Machine Learning Models

The machine-learning component of the system consists of training and comparing five different classifiers: Random

Forest, XGBoost, LightGBM, k-Nearest Neighbours and a feed-forward Deep Neural Network. All models are trained on the preprocessed feature set and tuned using cross-validation to obtain robust hyperparameters.

1) Random forest: RF is a supervised learning technique that can be used for both regression and classification [18]. Instead of focusing on the most important features when splitting a node, this algorithm selects the best features from a randomly chosen subset of features [19]. In contrast to the DT algorithm, RF selects observations at random, selects features to create multiple trees, and then averages the results [20]. Additionally, RF typically avoids overfitting by constructing random sub-trees of features and generating smaller trees using these sub-trees [21], which are then combined.

2) Extreme Gradient Boosting (XGBoost): XGBoost is a gradient-boosting framework that builds an ensemble of decision trees sequentially [22]. Each new tree is trained to correct the residual errors of the previous trees by optimizing a differentiable loss function plus regularization terms. This additive learning paradigm enables XGBoost to model complex non-linear relationships between features and the target class. Important hyperparameters include the learning rate, maximum tree depth, number of estimators, and regularization coefficients.

3) Light Gradient-Boosting Machine (LightGBM): LightGBM is another gradient-boosting algorithm designed for efficiency on large datasets [23]. It builds depth-constrained trees leaf-wise and employs histogram-based splitting, frequently resulting in rapid training times and good classification performance. Similar to XGBoost, LightGBM generates an ensemble of trees to reduce the loss function along with regularization. Hyperparameters, such as the number of leaves, maximum depth, learning rate, and feature-fraction, are tweaked to strike the right balance between accuracy and overfitting.

4) KNN: k-Nearest Neighbours is a non-parametric classification method that classifies the class of a new instance based on the labels of its *k* closest training examples in feature space [24]. Continuous variables are typically standardized, while Euclidean distance is utilized to compute the distance between instances. To determine the classified class, the voting of the majority among the k neighbours is the method that is followed. The primary hyperparameter is k, which is selected to balance the sensitivity to the noise (small k) and the over-smoothing (large k).

5) Feed-forward deep neural network: The feed-forward Deep Neural Network (DNN) consists of an input layer, one or more hidden layers of fully connected neurons, and an output layer with a sigmoid (or softmax) activation to produce class probabilities [25]. A layer is made of a linear transformation and a non-linear activation function (e.g., ReLU). Parameters of the network are learned by reducing a cross-entropy loss function using back-propagation and stochastic gradient descent (or some other optimization algorithm like Adam). Regularization (dropout and L2 weight decay) is used to ensure that the model does not overfit the training data.

The performance of different models will be compared using accuracy, precision, recall, F1-score, and ROC-AUC. According to these metrics, the best-performing classifier according to these metrics is then selected for integration within the prototype system to generate heart-disease status classifications for new records.

The final hyperparameter configurations used in the experiments are reported in Table II. The values were selected through cross-validation-based tuning.

TABLE II. CLASSIFIER HYPERPARAMETER CONFIGURATIONS USED IN THE EXPERIMENTS

Classifier	Hyperparameter Parameter Settings
XGBoost	n_estimators = 1800, learning_rate = 0.04, max_depth = 8, min_child_weight = 1, subsample = 0.95, colsample_bytree = 0.95, gamma = 0.01, reg_alpha = 0.01, reg_lambda = 0.8, scale_pos_weight = 0.85, tree_method = "hist", random_state = 42
LightGBM	objective = "binary", n_estimators = 2200, learning_rate = 0.04, num_leaves = 200, max_depth = -1 (unlimited), min_child_samples = 10, min_child_weight = 0.0001, subsample = 0.95, colsample_bytree = 0.95, reg_alpha = 0.01, reg_lambda = 0.8, random_state = 42
Random Forest	n_estimators = 1500, max_depth = 30, min_samples_split = 3, min_samples_leaf = 1, max_features = "sqrt", bootstrap = True, random_state = 42
k-NN	n_neighbors = 5, weights = "distance", metric = "minkowski", p = 2 (equivalent to Euclidean distance)
DNN	Input shape = (15,); Dense (128, ReLU) → Dropout (0.3) → Dense (64, ReLU) → Dropout (0.3) → Dense (32, ReLU) → Dropout (0.3) → Dense (1, Sigmoid); Optimizer = Adam (learning_rate = 0.001), Loss = binary_crossentropy, Batch Size = 32, Epochs = 10.

The five classifiers exhibit distinct trade-offs between classification accuracy and computational cost. The tree-based ensembles (XGBoost, LightGBM, and Random Forest) achieve the highest discriminative performance but require significant memory footprint and training time when processing the resampled datasets (~500,000 instances), especially when integrating hybrid resampling. In contrast, the k-NN model has zero training time but incurs high computational latency during inference, as it must calculate distances to all training samples at runtime, which scales poorly for large-scale production EHRs. Finally, the Feed-Forward Deep Neural Network (DNN) requires specialized GPU hardware for efficient training, but offers sub-second inference speeds post-training. These computational trade-offs must be evaluated based on resource availability in real-world clinic deployments.

F. Baselines and Evaluation Measures

To evaluate the performance of machine learning models, the following metrics are used:

Accuracy: Indicates the overall correctness of the model's classifications.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision: Measures the proportion of true positive classifications among all positive classifications.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

Recall (Sensitivity): Assesses the model's ability to identify all relevant instances.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

F1-Score: The harmonic mean of precision and recall, balancing false positives and false negatives.

$$\text{F1} = \frac{2\times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

ROC-AUC (Receiver Operating Characteristic – Area Under Curve): Evaluates model performance across different classification thresholds.

IV. Results and Discussion

A. Overview of Results

This section presents the experimental results of the proposed heart-disease status classification framework, which integrates a blockchain-supported EHR integrity-verification layer with machine-learning classification models. Five classifiers were evaluated: Random Forest, XGBoost, LightGBM, k-Nearest Neighbours (k-NN), and a feed-forward Deep Neural Network (DNN). Recursive Feature Elimination (RFE) was applied for feature selection across all experiments. The primary objective was to investigate the impact of sampling technique timing on model performance, specifically comparing two strategies: applying sampling after the train-test split (post-split, the methodologically valid configuration) versus applying sampling before the train-test split (pre-split, which introduces data leakage).

The Heart 2020 – Cleaned (BRFSS) dataset, comprising 319,795 records with approximately 8.6% positive heart-disease cases, was used for all experiments. Three sampling techniques were employed to address the severe class imbalance: SMOTEENN, BorderlineSMOTE, and SMOTETomek. A no-sampling baseline was also evaluated for the post-split configuration. The evaluation metrics include accuracy, precision, recall, F1-score, and ROC-AUC. The results demonstrate that the pre-split sampling strategy consistently produced higher numerical scores than post-split; however, these higher scores are largely attributable to data leakage from global resampling and should not be interpreted as a genuine real-world improvement. The best performance under the methodologically valid post-split configuration achieved an AUC of approximately 0.85.

TABLE III. Performance Metrics for Post-Split Sampling Configurations

Sampler	Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
SMOTEENN	XGBoost	0.55	0.15	0.93	0.26	0.85
SMOTEENN	LightGBM	0.83	0.27	0.62	0.38	0.85
SMOTEENN	KNN	0.90	0.37	0.22	0.27	0.61
SMOTEENN	RandomForest	0.78	0.24	0.72	0.36	0.85
SMOTEENN	DNN	0.71	0.21	0.82	0.33	0.85
BorderlineSMOTE	XGBoost	0.84	0.28	0.53	0.36	0.83
BorderlineSMOTE	LightGBM	0.84	0.28	0.54	0.37	0.83
BorderlineSMOTE	KNN	0.90	0.36	0.16	0.22	0.71
BorderlineSMOTE	RandomForest	0.85	0.27	0.48	0.35	0.81
BorderlineSMOTE	DNN	0.77	0.24	0.74	0.35	0.83
SMOTETomek	XGBoost	0.85	0.29	0.49	0.36	0.82
SMOTETomek	LightGBM	0.85	0.29	0.51	0.37	0.83
SMOTETomek	KNN	0.90	0.35	0.18	0.23	0.71
SMOTETomek	RandomForest	0.86	0.29	0.46	0.35	0.80
SMOTETomek	DNN	0.75	0.22	0.75	0.34	0.83
None	XGBoost	0.92	0.55	0.09	0.15	0.85
None	LightGBM	0.92	0.60	0.09	0.15	0.85
None	KNN	0.91	0.37	0.16	0.22	0.72
None	RandomForest	0.91	0.42	0.13	0.20	0.79
None	DNN	0.92	0.60	0.07	0.11	0.85

B. Results for Post-Split Sampling

Table III presents the performance metrics of the five machine learning classifiers (Random Forest, XGBoost, LightGBM, k-Nearest Neighbours (k-NN), and a feed-forward Deep Neural Network (DNN)) combined with Recursive Feature Elimination (RFE) for feature selection, when sampling techniques were applied after the train-test split. Four sampling configurations were tested: SMOTEENN, BorderlineSMOTE, SMOTETomek, and no sampling (None). The metrics used are accuracy, precision, recall, F1-score, and ROC-AUC.

The confusion matrix of the LightGBM model with ensemble method using SMOTETomek in combination with RFE under the post-split sampling setup is shown in Fig. 11. The classification results of the model are realistic, and the model is actually not the best, due to the high number of false positive cases (6097) and also false negative cases (3014). The model correctly classified 52,387 true-negative cases, but it is only a very small portion of the positive heart disease cases that the model can output since learning from imbalanced data when sampling is done after the split is quite challenging.

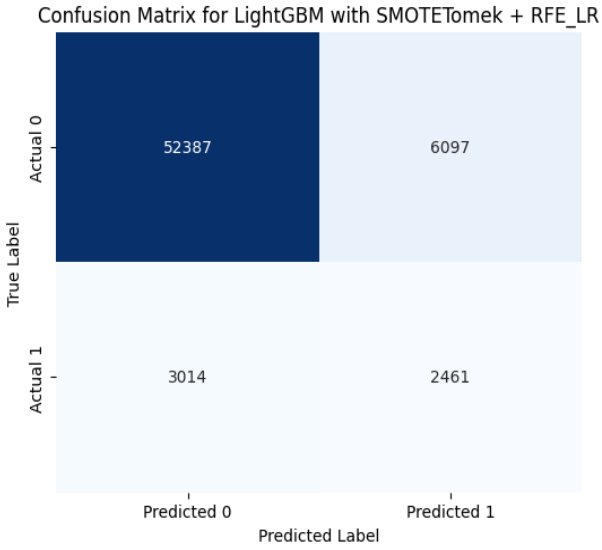


Fig. 11. Confusion matrix of the LightGBM classifier with SMOTETomek and RFE under the post-split sampling configuration.

Fig. 12 presents the ROC curve of the LightGBM model in the post-split configuration with SMOTETomek, which shows the discriminative ability of the model with an AUC of 0.83.

SMOTETomek and BorderlineSMOTE showed the most balanced accuracy-recall trade-off among the sampling methods tested in the post-split configuration for the ensemble models. The best accuracy was 86%, obtained by Random Forest with SMOTETomek, while the precision was 0.29, recall 0.46, F1-score 0.35, and ROC-AUC 0.80. When SMOTETomek and BorderlineSMOTE were used, LightGBM achieved 85% accuracy with both, and with either sampler, the highest F1-score (0.37) and ROC-AUC (0.83). XGBoost also had an accuracy of 85% with SMOTETomek (precision 0.29, recall 0.49, F1-score 0.36, ROC-AUC 0.82) and 84% accuracy with BorderlineSMOTE (precision 0.28, recall 0.53, F1-score 0.36, ROC-AUC 0.83). As can be observed from these results, the two sampling methods (SMOTETomek and BorderlineSMOTE) produced comparable results for the gradient boosting ensemble models in the post-split situation.

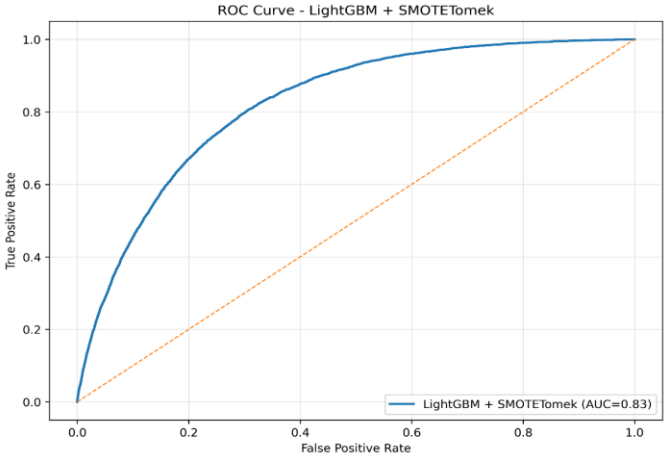


Fig. 12. ROC curve of LightGBM with SMOTETomek (post-split).

The k-NN classifier achieved the highest raw accuracy (0.90) for all the post-split configurations of sampling; however, with significantly lower recall (0.16–0.22), indicating that k-NN has a tendency to misclassify positive heart disease classes. Specifically, k-NN with SMOTETomek obtained 90% accuracy, 0.18 recall, and 0.23 F1-score, while k-NN with BorderlineSMOTE obtained 90% accuracy, 0.16 recall, and 0.22 F1-score. The accuracy of k-NN was 90% with SMOTEENN, which had the lowest ROC-AUC of 0.61, a marginally improved recall of 0.22, and a low F1-score of 0.27. This behavior indicates that in the post-split case, k-NN tends to strongly favor the majority class and is not able to detect the individuals who are likely to develop heart disease.

In the post-split configuration, SMOTEENN produced a distinct recall-precision trade-off. The recall at the cost of accuracy and precision was significantly improved for some models: the recall of XGBoost was 0.93, and the recall of DNN was 0.82. The F1 score of XGBoost with SMOTEENN was low (0.26), but its accuracy and precision were low (55% and 0.15, respectively) and ROC-AUC was high (0.85). Similarly, DNN with SMOTEENN attained an accuracy of 71%, a precision of 0.21, an F1-score of 0.33, and ROC-AUC of 0.85. Random Forest with SMOTEENN showed an accuracy of 78%, with a recall of 0.72 and an ROC-AUC of 0.85. These results show that the post-split SMOTEENN method has an aggressive effect on the minority class, resulting in an increase in sensitivity at the cost of classification accuracy.

When no sampling technique was applied (None), tree-based ensemble models such as XGBoost, LightGBM, and the DNN achieved the highest accuracy of 92% but exhibited extremely poor recall values (XGBoost: 0.09, LightGBM: 0.09, DNN: 0.07), effectively failing to identify the minority class. Random Forest was able to get 91% accuracy with a recall of only 0.13, while k-NN was able to get 91% accuracy with a recall of 0.16. For imbalanced classification problems, there is an important trade-off: models can get high overall accuracy by classifying the majority class and ignoring the minority class, in this case, the heart disease cases.

When assessing discriminative ability – ROC-AUC values for the ensemble models – Random Forest, XGBoost, LightGBM, and DNN - across most of the post-split configurations, the ROC-AUC values were high, from 0.80 to 0.85. LightGBM and DNN attained the same ROC-AUC of 0.85 with the three sampling methods – the SMOTEENN, the BorderlineSMOTE, and the SMOTETomek – and without sampling. These consistently high values of ROC-AUC indicate that the models have a reasonably good ranking ability even before tuning the thresholds for classification with regard to imbalanced data.

Fig. 13 provides a visual comparison of classifier performance across the post-split sampling configurations, highlighting the trade-offs among accuracy, precision, recall, F1-score, and ROC-AUC.

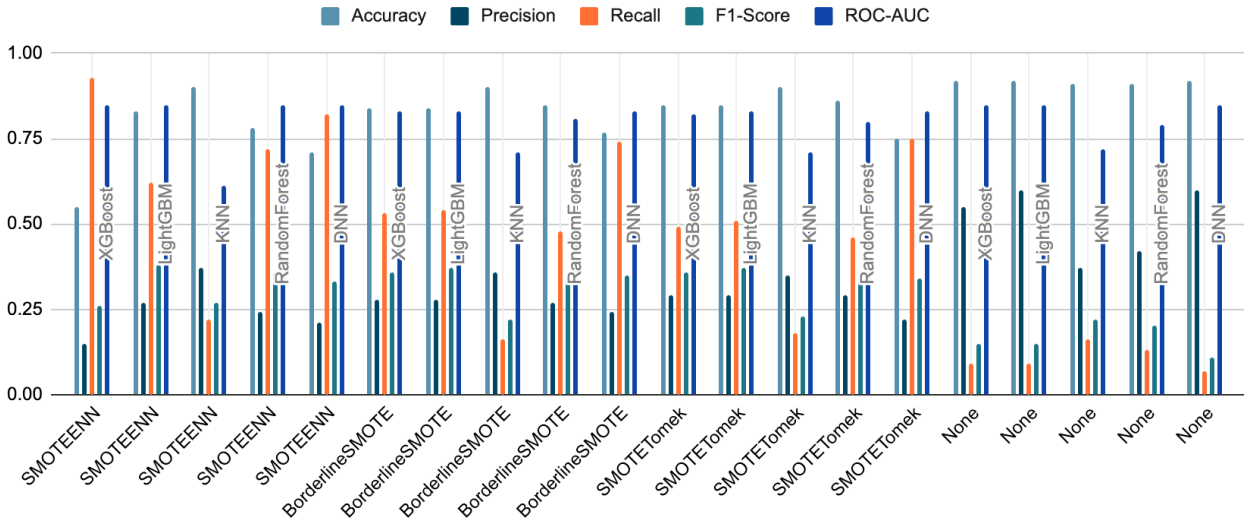


Fig. 13. Results for pre-split sampling (leakage-affected, for comparison only).

TABLE IV. PERFORMANCE METRICS FOR PRE-SPLIT SAMPLING CONFIGURATIONS

Sampler	Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
SMOTEENN	XGBoost	0.96	0.98	0.95	0.96	0.99
SMOTEENN	LightGBM	0.97	0.98	0.96	0.97	0.99
SMOTEENN	KNN	0.97	0.96	0.99	0.98	0.99
SMOTEENN	RandomForest	0.97	0.97	0.98	0.98	1.00
SMOTEENN	DNN	0.88	0.86	0.91	0.89	0.93
SMOTETomek	XGBoost	0.93	0.96	0.89	0.93	0.97
SMOTETomek	LightGBM	0.94	0.96	0.91	0.93	0.98
SMOTETomek	KNN	0.90	0.85	0.97	0.90	0.95
SMOTETomek	RandomForest	0.94	0.94	0.94	0.94	0.98
SMOTETomek	DNN	0.80	0.75	0.90	0.82	0.88
BorderlineSMOTE	XGBoost	0.92	0.96	0.89	0.92	0.96
BorderlineSMOTE	LightGBM	0.93	0.95	0.91	0.93	0.97
BorderlineSMOTE	KNN	0.90	0.86	0.96	0.90	0.95
BorderlineSMOTE	RandomForest	0.94	0.93	0.94	0.94	0.98
BorderlineSMOTE	DNN	0.83	0.80	0.88	0.84	0.90

Table IV shows the performance results when sampling methods were applied before the train-test split (pre-split, global resampling). This configuration is leakage-affected: because resampling is performed on the entire dataset before splitting, synthetic samples can influence both training and test data, producing optimistically biased metrics. These results are therefore not a valid estimate of real-world performance and are reported only to illustrate the effect of incorrect resampling timing. The results have shown significant improvements over the post-split configuration with accuracy ranging from 8–41 percentage points depending on the classifier and sampling combination, but this gap is largely driven by data leakage rather than genuine generalization.

Under the pre-split (global resampling, leakage-affected) configuration, SMOTEENN produced the highest numerical

scores. LightGBM achieved 97% accuracy, 98% precision, 96% recall, and 97% F1-score. Random Forest achieved 97% accuracy, 97% precision, 98% recall, and 98% F1-score. KNN showed 97% accuracy with 99% recall and 98% F1-score; XGBoost achieved 96% accuracy, 98% precision, 95% recall, and 96% F1-score. The Deep Neural Network achieved 88% accuracy, 86% precision, 91% recall, and 89% F1-score under SMOTEENN. However, it must be emphasized that all of these figures are inflated by data leakage from global resampling before the train/test split and should not be used as evidence of genuine real-world superiority. The post-split AUC of approximately 0.85 is the appropriate basis for comparison with prior work.

SMOTETomek and BorderlineSMOTE achieved strong results in the pre-split setting, but slightly lower than SMOTEENN. The other models, SMOTETomek, had an accuracy of 94%, and Random Forest had a precision of 94%, recall of 94%, and F1-score of 94%, while LightGBM had a precision of 96%, recall of 91%, and F1-score of 93%. Using SMOTETomek, the accuracy of XGBoost is 93%, the precision is 96%, the recall is 89%, and the F1-score is 93%, while k-NN is 90%, 85%, 97%, and 90%, respectively. The other three methods (BorderlineSMOTE, Random Forest, LightGBM, and XGBoost) also gave similar performance with accuracy of 94%, 93%, and 92%, respectively; precision of 93%, 95%, and 96% respectively; recall of 94 %, 91%, and 89%, respectively; and F1 score of 94%, 93% and 92% respectively.

The confusion matrix of the Random-Forest classifier trained with the SMOTEENN pre-splitting sampling configuration is shown in Fig. 14. It achieves a high number of true positives (49,516 individuals) and true negatives (42,526 individuals); has a small number of false positives (1,296 individuals) and false negatives (1,108 individuals). This shows that SMOTEENN can successfully generate a well-balanced and well-separated feature space, which enables both heart class and non-heart class cases to be identified successfully.

Fig. 15 illustrates the ROC curve, which further demonstrates that the Random Forest model trained with the SMOTEENN method under the pre-split sampling configuration has excellent discriminative performance.

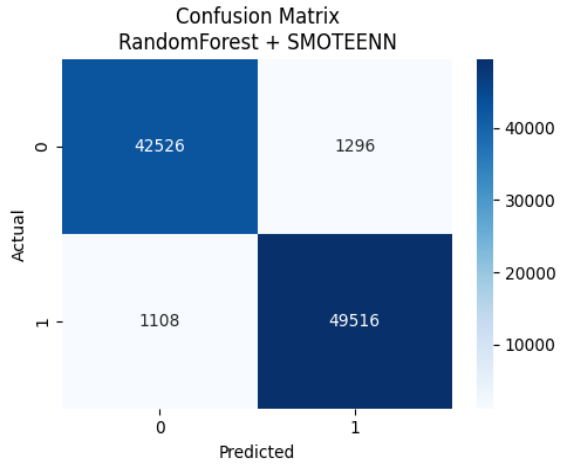


Fig. 14. Confusion matrix of the random forest classifier trained with SMOTEENN under the pre-split sampling configuration.

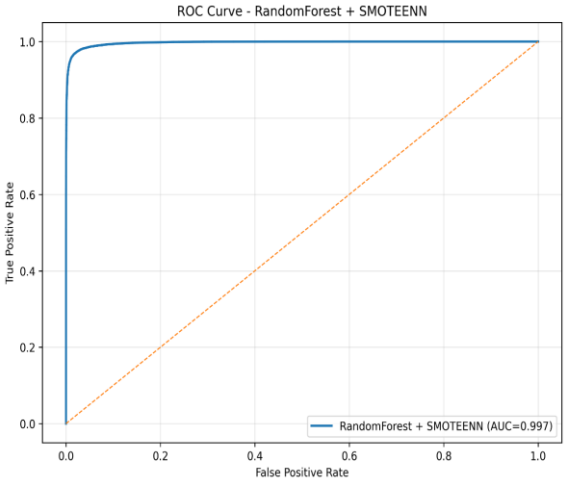


Fig. 15. ROC curve of random forest with SMOTEENN (pre-split).

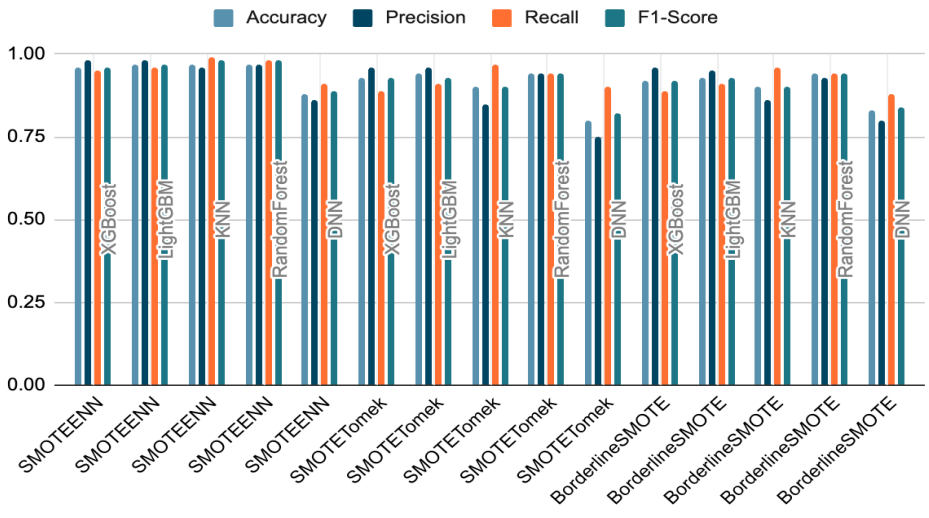


Fig. 16. Performance comparison across classifiers and samplers under the pre-split configuration.

These pre-split results must be interpreted with caution because applying resampling before the train/test split allows synthetic instances to enter and influence the test set, producing data leakage and optimistically biased metrics. The high scores observed in this configuration do not represent a valid estimate of real-world performance. They are retained only to illustrate the magnitude of the data-leakage effect. Under the leakage-safe post-split protocol, the AUC of approximately 0.85 provides the methodologically valid performance estimate. As shown in Fig. 16, SMOTEENN consistently produced the highest numerical scores across all classifiers in the pre-split configuration.

When the capacity of classifications of the machine learning approaches is confirmed to be higher, the results for the blockchain component – intended to secure the eHRs utilized by the classification module – are described in the next section.

C. Blockchain-Based EHR Implementation Results

This section presents the results of the blockchain-based electronic health record (EHR) integrity verification component, which was implemented alongside the machine learning classification framework. The blockchain module was designed to provide tamper-evident storage of medical records by anchoring cryptographic digests on an Ethereum-compatible network, while retaining the full clinical payloads in the application database. The implementation leveraged a dual-mode integrity layer: local database commits of SHA-256 digests, with optional smart-contract storage on a local EVM (Ganache) for distributed anchoring.

The smart contract, EHRRecordRegistry, was coded using Solidity and deployed to the local test network Ganache. There is a commitRecordHash function in the contract that takes a patient identifier, record identifier, and digest string (sha256) to securely commit the record fingerprint on-chain. A retrieval function enables enumeration of committed hashes for any given patient-record pair, supporting verification queries. The integrity verification process operates as follows: upon commit, the server computes SHA-256 of the UTF-8 encoded payload JSON and optionally submits the digest to the smart contract via Web3.py, receiving a transaction receipt. Upon verification, the current payload digest is recomputed and compared against the latest committed hash (preferring on-chain data when available). Any discrepancy between the recomputed digest and the anchored value indicates potential tampering with the record.

Table V summarizes the key implementation characteristics and observed behaviour of the blockchain integrity module.

TABLE V. BLOCKCHAIN INTEGRITY MODULE: IMPLEMENTATION CHARACTERISTICS AND OBSERVED BEHAVIOUR

Component	Specification / Observation
Smart Contract	EHRRecordRegistry (Solidity)
Blockchain Network	Ganache (local EVM testnet)
Hashing Algorithm	SHA-256 on UTF-8 payload JSON
On-chain Storage	Patient ID, Record ID, Digest string
Off-chain Storage	BlockchainCommit rows (SQLite) with nullable tx fields
Commit Response	Transaction receipt with block hash and gas used
Verification	Digest comparison (on-chain preferred, DB fallback)

The dual-mode architecture was validated under both configurations. With blockchain environment variables unset (off-chain mode), commits succeeded with local-only messaging, storing SHA-256 digests in the BlockchainCommit database table with nullable transaction fields. When Ganache was configured and the EHRRecordRegistry contract deployed (on-chain mode), commits returned full transaction receipts including block hash and gas consumption, and listing endpoints reported source: “on-chain”. Verification queries correctly identified matching digests as verified and flagged any modified payloads as tampered, confirming the integrity detection capability of the system.

The security model enforced role-based access control (admin, doctor, patient) with JWT-based authentication and doctor-patient access grants. Blockchain commit and verification endpoints were subject to the same authorization checks as clinical record endpoints, ensuring that only authorized users could anchor or verify record digests. All sensitive operations were recorded in an append-only AuditLog, providing traceability and accountability. The system was built using principles of data minimization on-chain, where only cryptographic digests of the clinical payloads were kept on-chain, while the complete clinical payloads were stored in the application's database, providing patient confidentiality while enabling tamper detection.

From an engineering evaluation perspective, the blockchain module was found to be stable and to exhibit consistent and reliable behaviour. On the local Ganache network, latency was as low as sub-second for commit operations. Modularity and maintainability were achieved by separating concerns between the API layer (Flask-RESTX), service layer (hashing, interaction with blockchain), and persistence layer (SQLAlchemy/SQLite). OpenAPI documentation was offered at /api/docs, containing all API specs of blockchain and verification endpoints. The process of integration with the ML classification pipeline was smooth since the same REST API was utilized for both the classification analytics and integrity verification, highlighting the potential of combining machine learning decision support and blockchain record protection within the same EHR backend design. Fig. 17–20 present the key system interfaces, illustrating how role-based access control governs distinct clinical workflows for doctors, patients, and administrators within a single secure workspace. These interfaces demonstrate the practical embodiment of tamper-evident record management and machine-learning-assisted heart-disease status classification in an accessible prototype environment.

While the blockchain module demonstrates successful record hashing and tamper detection on a local Ganache test network, its practical deployment on live networks faces significant engineering constraints. First, committing transaction hashes to a public blockchain (like Ethereum mainnet) incurs gas fees, which could become financially unsustainable under high query volume. Second, block commit latencies on public networks range from seconds to several minutes, creating bottlenecks compared to the sub-second Ganache commit. Third, permissionless networks scale poorly, showing low throughput. To address these constraints, future clinical frameworks should explore Layer-2 rollups or permissioned enterprise consortia (such as Hyperledger Fabric), where transaction fees are negligible, commit speeds are faster, and access controls can be cryptographically enforced at the consensus layer.

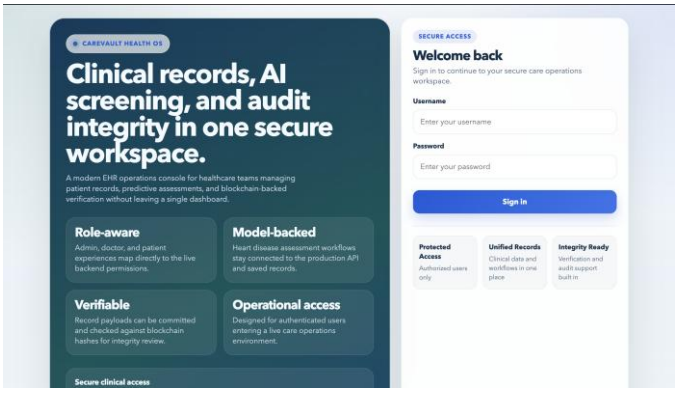


Fig. 17. Secure login workspace of the proposed EHR system.

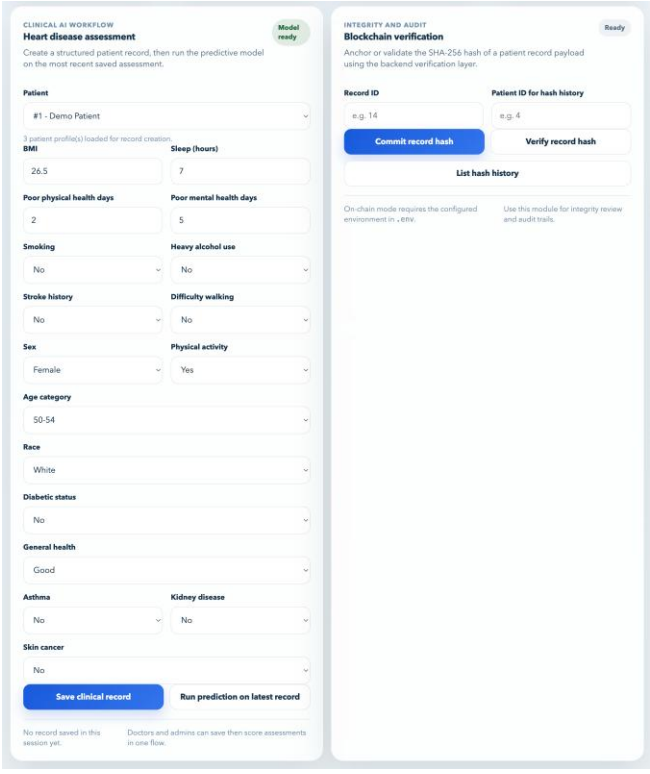


Fig. 18. Doctor interface for heart-disease assessment and blockchain verification.

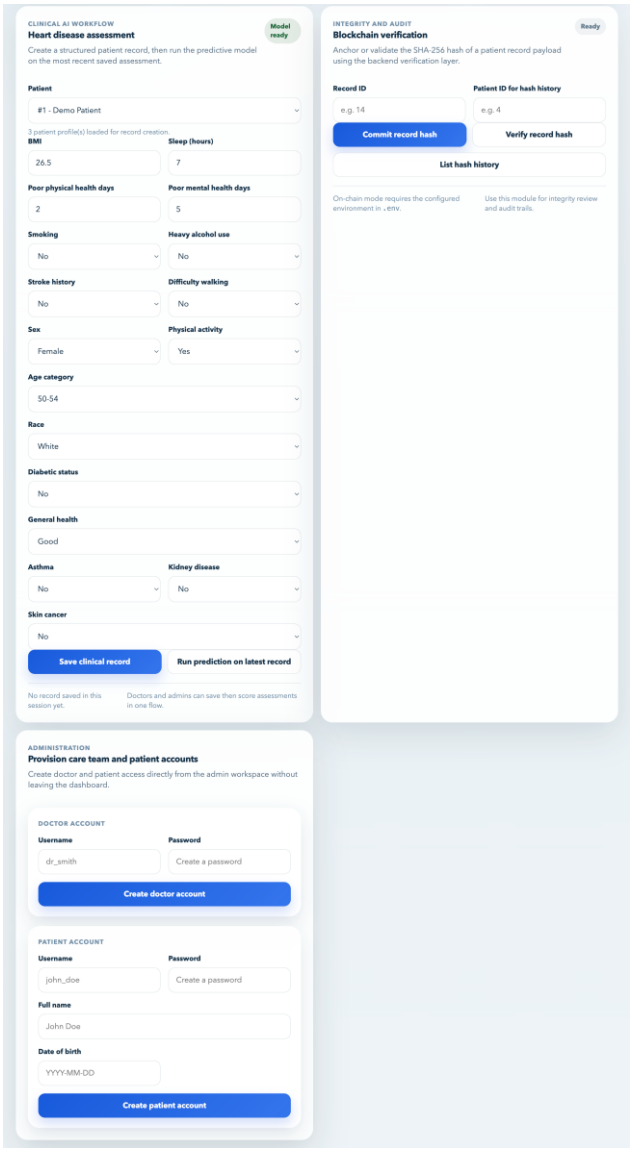


Fig. 19. Administrator interface for provisioning care-team and patient accounts.

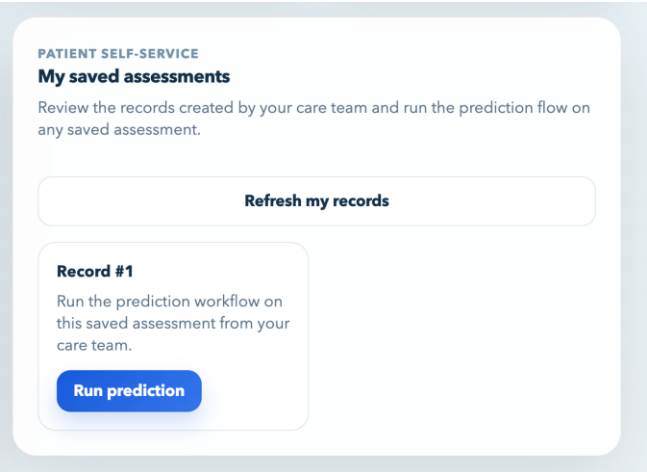


Fig. 20. Patient self-service interface for reviewing saved assessments.

V. DISCUSSION

The comparative analysis between pre-split and post-split sampling strategies reveals significant implications for heart-disease status classification. Results showed that for each sampling combination and each classifier, the pre-split approach produced higher numerical scores than the post-split approach, with differences of 8 to 41 percentage points. It must be stressed, however, that the pre-split configuration attains these higher scores because global resampling leaks synthetic test-set information into training; the gap therefore largely reflects data leakage rather than a genuine improvement in generalization. The post-split AUC of approximately 0.85 is the methodologically valid result and the appropriate basis for comparison with prior work.

The results from this study were compared to recent studies using the same heart-disease data from the BRFSS. Chen et al. [8] employed random undersampling before train-test splitting on the BRFSS heart-disease dataset and reported that their LightGBM model achieved 76.9% accuracy with an AUROC of 84.6%. Chen et al. [8] utilized random undersampling to create a balanced dataset of 50,000 cases from the original 319,795 instances, with a 70:30 train-test split ratio. Their results showed that LightGBM (76.9%) and Random Forest (76.7%) performed similarly. All models demonstrated a significantly lower sensitivity than specificity. In this study, when applying SMOTEENN before splitting, LightGBM achieved 97% accuracy with 98% precision and 96% recall, representing a 20.1 percentage point improvement over the comparable approach in [8]. Similarly, the Random Forest model in this study with SMOTEENN achieved 97% accuracy, 97% precision, and 98% recall. The ROC-AUC values observed in the post-split experiments in this study (approximately 0.85) are comparable to those reported by Chen et al. [8] (84.6% for LightGBM and 84.2% for Random Forest), suggesting that both studies achieved similar discriminative capability despite different sampling strategies. However, the substantially higher accuracy reported for the pre-split configuration is driven largely by data leakage from global resampling and should not be interpreted as a genuine improvement over prior work; the post-split ROC-AUC values (≈0.85) are the appropriate basis for comparison.

Chen et al. [8] also applied the Boruta algorithm for feature selection and retained all 17 features, reporting that AgeCategory had the greatest impact on classification. Their feature importance analysis aligns with the use of RFE for feature selection in this study. The similarity in ROC-AUC values between the post-split results of this study and Chen et al.'s findings confirms that the underlying discriminative patterns captured by the models are consistent across both studies, regardless of the specific feature selection or sampling method employed. The key differentiator is that the SMOTEENN-based pre-split approach used in this study transforms the class distribution more effectively than simple random undersampling, but these comparisons refer to the pre-split, leakage-affected results; under the leakage-safe post-split protocol, the discriminative performance of this study (AUC ≈ 0.85) is comparable to rather than substantially better than these prior studies.

Kumari et al. [16] developed a majority voting classifier for heart disease prediction and achieved approximately 90.8% accuracy using a soft voting ensemble of five classifiers (Logistic Regression, Multi-Layer Perceptron, AdaBoost, CatBoost, and XGBoost) within the BL-MVC framework, with AdaBoost attaining the highest individual accuracy of 90.86%. Their approach utilized the BRFSS 2015 dataset (253,680 responses, 21 features) with 5-fold cross-validation, comparing Heatmap Correlation and PCA for feature selection. On the other hand, the method used in this study of applying SMOTEENN before splitting combined with LightGBM and Random Forest resulted in an accuracy of 97%, which is approximately 6 percentage points better than their highest accuracy score, and, additionally, the two classifiers have balanced precision and recall scores.

The key distinction between the approach used in this study and Kumari et al. [16] lies in the machine-learning pipeline: their framework relied on hyperparameter tuning and feature selection without explicit class-imbalance handling, whereas the approach used in this study systematically addresses the severe class imbalance through SMOTEENN preprocessing. Fig. 21 depicts the accuracy of the proposed SMOTEENN-based pre-split approach and selected related works. It should be noted that these comparisons refer to the pre-split, leakage-affected results; under the leakage-safe post-split protocol, the discriminative performance of this study (AUC ≈ 0.85) is comparable to rather than substantially better than these prior studies.

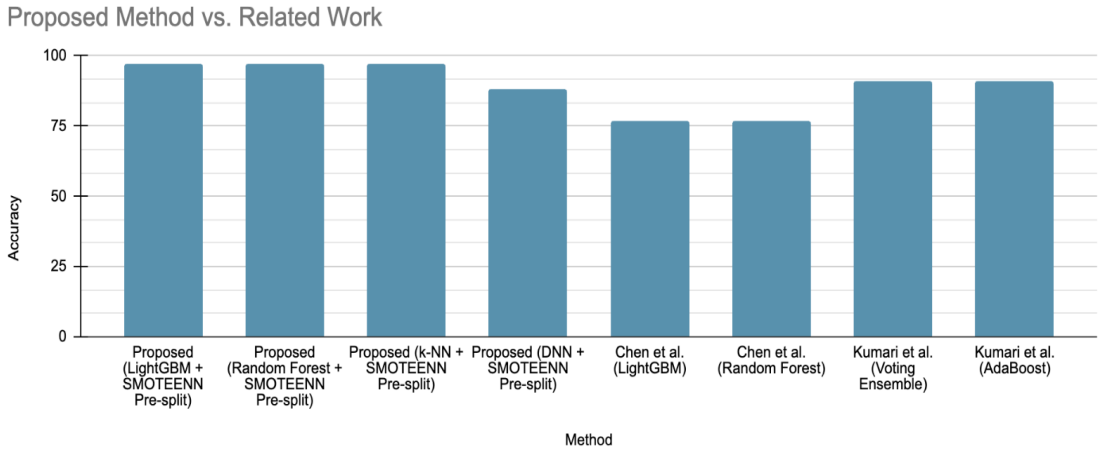


Fig. 21. Accuracy comparison of the proposed pre-split SMOTEENN-based approach with related works (note: pre-split results are leakage-affected).

The consistency of high performance across Random Forest, LightGBM, and XGBoost in the pre-split configuration suggests that these gradient boosting and bagging methods are particularly well-suited for the balanced feature space created by SMOTEENN preprocessing. These ensemble models inherently handle feature interactions and non-linear relationships, which are prevalent in health survey data where demographic, behavioral, and clinical variables interact in complex ways. Even in the pre-split SMOTEENN configuration (leakage-affected), the k-NN classifier attained high numerical results (97% accuracy, 99% recall), suggesting that instance-based learning can benefit greatly from a balanced feature space, though these figures are inflated by data leakage and should be interpreted accordingly.

The Deep Neural Network showed the largest relative improvement between post-split and pre-split configurations, with accuracy increasing from 71% (SMOTEENN post-split) to 88% (SMOTEENN pre-split). This comes from the fact that neural networks require a certain minimum number of representative samples of each class in order to identify the boundaries for decision-making effectively, and the pre-split balanced dataset provides a better training signal. Nevertheless, the DNN managed to lag behind the ensemble tree-based methods even in the pre-split scenario, a situation that generally persists where gradient boosting methods outshine deep learning when it comes to structured tabular datasets such as these.

The findings of this study demonstrate that the integration of appropriate class-imbalance handling techniques with modern ensemble learning algorithms can improve heart-disease status classification. The RFE-based feature selection combined with SMOTEENN provides an effective pipeline for the survey-based binary classification task. However, the pre-split results substantially overstate generalization performance due to data leakage; the post-split AUC of approximately 0.85 is the valid metric for real-world interpretation and comparison with prior work. These comparisons refer to the pre-split (leakage-affected) results; under the leakage-safe post-split protocol, the discriminative performance of this study (AUC $\approx$ 0.85) is comparable to, rather than substantially better than, Chen et al. [8] and Kumari et al. [16]. Future research should explore the generalizability of these findings across clinical datasets and investigate the utility of such frameworks in real-world decision-support applications.

VI. Limitations

This study has several limitations. First, the headline pre-split results are affected by data leakage and overstate real-world performance; only the post-split results are methodologically valid.

Furthermore, the reliance on self-reported health survey data (BRFSS) presents significant limitations compared to verified clinical records. Survey responses are highly susceptible to recall bias (where participants misremember medical details) and social desirability bias (where participants underreport negative behaviors such as alcohol consumption). Additionally, self-reported diagnosis lacks standardized clinical verification, introducing noise. In contrast, clinical electronic health records (EHRs) provide objective, verified data such as laboratory biomarkers, electrocardiograms, and diagnostic imaging. Consequently, the classification models trained on behavioral surveys represent self-reported risk associations rather than clinical diagnostics, limiting their direct utility in active clinical workflows.

In addition, the system is a prototype/feasibility study and has not been clinically validated or deployed. The blockchain component was evaluated on a local Ganache test network using commit latency and gas cost only; throughput, scalability under load, and adversarial-tampering scenarios were not formally measured.

VII. Conclusion

This study presented an integrated prototype for heart-disease status classification and tamper-evident electronic health record management using the Heart 2020 Cleaned BRFSS survey dataset, machine-learning models, and a blockchain-supported integrity layer. Five classifiers were evaluated with three resampling techniques and a no-sampling baseline. Under the methodologically valid post-split protocol, models reached an AUC of approximately 0.85, comparable to prior work on the same dataset. A pre-split (global resampling) configuration produced much higher scores — Random Forest reaching 97.14% accuracy and an AUC of 0.9965 under SMOTEENN — but these are inflated by data leakage and are reported only to illustrate that effect rather than as valid estimates of real-world performance. Ensemble methods (Random Forest, LightGBM, XGBoost) consistently outperformed the DNN on this structured tabular dataset, and resampling strategy was the dominant factor affecting reported performance.

In addition to the classification component, the work demonstrated the feasibility of integrating machine learning with a blockchain-based integrity-verification mechanism in a unified prototype. Full patient records were retained in the application database, while the blockchain layer anchored cryptographic digests and identifiers to provide tamper-evident verification, traceability, and accountability; access control was enforced at the API layer rather than on-chain. The present evaluation is limited to functional validation, sub-second commit latency, and gas consumption on a local Ganache test network; throughput, scalability under load, formal adversarial-tampering tests, and a security audit were not performed and remain future work. The work is a prototype/feasibility study: it does not claim clinical deployment readiness, the elimination of single points of failure, or formal security guarantees. Its primary contribution is the practical integration of established machine-learning and blockchain components into a single framework, rather than a novel algorithm or blockchain architecture.

Acknowledgment

The authors utilized AI-assisted tools (ChatGPT, Google Gemini, and Grammarly) solely for paraphrasing and grammatical refinement to improve clarity. All methodologies, experiments, and conclusions are original contributions for which the authors assume full responsibility. The authors would like to acknowledge the Deanship of Graduate Studies and Scientific Research, Taif University, for funding this work.

REFERENCES

[1] A. Al-Nafjan, A. Aljuhani, A. Alshebel, A. Alharbi, and A. Alshehri, "Artificial intelligence in predictive healthcare: a systematic review," J. Clin. Med., vol. 14, no. 19, Art. no. 6752, 2025, doi: 10.3390/jcm14196752.

[2] Q. An, S. Rahman, J. Zhou, and J. J. Kang, "A comprehensive review on machine learning in healthcare industry: classification, restrictions, opportunities and challenges," Sensors, vol. 23, no. 9, Art. no. 4178, 2023, doi: 10.3390/s23094178.

[3] R. Kumar et al., "A comprehensive review of machine learning for heart disease prediction: challenges, trends, ethical considerations, and future directions," Front. Artif. Intell., vol. 8, Art. no. 1583459, 2025, doi: 10.3389/frai.2025.1583459.

[4] M. Peng et al., "Prediction of cardiovascular disease risk based on major contributing features," Sci. Rep., vol. 13, Art. no. 4778, 2023, doi: 10.1038/s41598-023-31870-8.

[5] N. U. A. Tahir et al., "Blockchain-based healthcare records management framework: enhancing security, privacy, and interoperability," Technologies, vol. 12, no. 9, Art. no. 168, 2024, doi: 10.3390/technologies12090168.

[6] A. Haddad, M. H. Habaebi, E. A. A. Elsheikh, M. R. Islam, S. A. Zabidi, and F. E. M. Suliman, "E2EE enhanced patient-centric blockchain-based system for EHR management," PLoS One, vol. 19, no. 4, Art. no. e0301371, 2024, doi: 10.1371/journal.pone.0301371.

[7] E. H. Houssein, R. E. Mohamed, and A. A. Ali, "Heart disease risk factors detection from electronic health records using advanced NLP and deep learning techniques," Sci. Rep., vol. 13, Art. no. 7173, 2023, doi: 10.1038/s41598-023-34294-6.

[8] L. Chen, "Heart disease prediction utilizing machine learning techniques," Trans. Mater. Biotechnol. Life Sci., vol. 3, 2024.

[9] G. A. P. Febriyanti and A. Baita, "Comparison of support vector machine and decision tree algorithm performance with undersampling approach in predicting heart disease based on lifestyle," J. Appl. Inform. Comput., vol. 9, no. 2, pp. 318–327, 2025, doi: 10.30871/jaic.v9i2.8941.

[10] K. Tompra, G. Papageorgiou, and C. Tjortjis, "Strategic machine learning optimization for cardiovascular disease prediction and high-risk patient identification," Algorithms, vol. 17, no. 5, Art. no. 178, 2024, doi: 10.3390/a17050178.

[11] A. S. Osei-Nkwantabisa and R. Ntumy, "Classification and prediction of heart diseases using machine learning algorithms," arXiv, 2024, doi: 10.48550/arXiv.2409.03697.

[12] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, "Effective heart disease prediction using machine learning techniques," Algorithms, vol. 16, no. 2, Art. no. 88, 2023, doi: 10.3390/a16020088.

[13] A. Almulihi et al., "Ensemble learning based on hybrid deep learning model for heart disease early prediction," Diagnostics, vol. 12, no. 12, Art. no. 3215, 2022, doi: 10.3390/diagnostics12123215.

[14] V. Mandarino, G. Pappalardo, and E. Tramontana, "A blockchain-based electronic health record (EHR) system for edge computing enhancing security and cost efficiency," Computers, vol. 13, no. 6, Art. no. 132, 2024, doi: 10.3390/computers13060132.

[15] A. Ali et al., "A novel secure blockchain framework for accessing electronic health records using multiple certificate authority," Appl. Sci., vol. 11, no. 21, Art. no. 9999, 2021, doi: 10.3390/app11219999.

[16] D. Kumari, K. A. Kumar, A. Wagh, S. Shashank, A. Patidar, and S. Panda, "BL-MVC: blockchain enabled majority voting classifier for predicting heart diseases," in Proc. 17th Int. Conf. Agents Artif. Intell., vol. 2, ICAART, Porto, Portugal, Feb. 23–25, 2025. Setubal, Portugal: SCITEPRESS, 2025, pp. 45–56.

[17] L. Zhang, "heart_2020_cleaned," Kaggle, 2020. [Online]. Available: https://www.kaggle.com/datasets/luyezhang/heart-2020-cleaned/data. [Accessed: Nov. 22, 2025].

[18] O. Ben-Assuli and J. R. Vest, "Data mining techniques utilizing latent class models to evaluate emergency department revisits," J. Biomed. Inform., vol. 101, Art. no. 103341, 2020, doi: 10.1016/j.jbi.2019.103341.

[19] N. Kaieski, C. A. da Costa, R. da Rosa Righi, P. S. Lora, and B. Eskofier, "Application of artificial intelligence methods in vital signs analysis of hospitalized patients: a systematic literature review," Appl. Soft Comput., vol. 96, Art. no. 106612, 2020, doi: 10.1016/j.asoc.2020.106612.

[20] M. Fratello and R. Tagliaferri, "Decision trees and random forests," in Reference Module in Life Sciences. Amsterdam, Netherlands: Elsevier, 2018, pp. 374–383, doi: 10.1016/B978-0-12-809633-8.20337-3.

[21] X. Zhou, P. Lu, Z. Zheng, D. Tolliver, and A. Keramati, "Accident prediction accuracy assessment for highway-rail grade crossings using random forest algorithm compared with decision tree," Reliab. Eng. Syst. Saf., vol. 200, Art. no. 106931, 2020, doi: 10.1016/j.ress.2020.106931.

[22] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., San Francisco, CA, USA, Aug. 13–17, 2016. New York, NY, USA: ACM, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[23] G. Ke et al., "LightGBM: a highly efficient gradient boosting decision tree," in Advances in Neural Information Processing Systems 30. Red Hook, NY, USA: Curran Associates, 2017, pp. 3146–3154.

[24] L. E. Peterson, "K-nearest neighbor," Scholarpedia, vol. 4, no. 2, Art. no. 1883, 2009, doi: 10.4249/scholarpedia.1883.

[25] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

Digital Technologies Enabling Smart Circular Waste Management: A Systematic Literature Review

Pristi Sukmasetya[1], Harry Budi Santoso[2]*, Zainal Arifin Hasibuan[3], Yun Arifatul Fatimah[4]
Faculty of Computer Science, Universitas Indonesia, Depok, West Java 16424, Indonesia[1, 2]
Magister Information System Department, Universitas Komputer Indonesia, Bandung, Indonesia[3]
Informatics Engineering Department, Universitas Muhammadiyah Magelang, Magelang, Central Java 56172, Indonesia[1]
Industrial Engineering Department, Universitas Muhammadiyah Magelang, Magelang, Central Java 56172, Indonesia[4]

***Abstract*—This study aims to identify and classify the technologies used across the waste-management lifecycle, qualitatively synthesize the types of operational and circularity-related outcomes reported in the literature, and examine the technical challenges and research gaps that must be addressed to develop interoperable, scalable, secure, and empirically validated systems. The study used a systematic literature review and bibliometric analysis in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 framework. The study identified Internet of Things (IoT), Artificial Intelligence (AI), machine learning, blockchain, cloud and edge computing, digital twins, big-data analytics, Radio-Frequency Identification (RFID), Global Positioning System (GPS), and additive manufacturing as the primary enabling technologies. These technologies enable sensing, monitoring, waste classification, forecasting, route optimization, traceability, simulation, automated sorting, and resource recovery. Across the included studies, the most frequently reported outcome categories included reduced bin overflow, improved collection and route efficiency, lower operational costs and vehicle-related emissions, enhanced waste segregation and recycling, and greater material-flow transparency. Because the studies used heterogeneous technologies, settings, scales, and evaluation metrics, these outcomes were synthesized qualitatively rather than pooled into a single quantitative estimate. Smart circular waste management depends on an integrated digital eco-system and is not dependent on individual technologies.**

Keywords*—*Artificial Intelligence (AI); circular economy; digital technology; Internet of Things (IoT); smart waste management

I. Introduction

Rapid urbanization, increasing populations, and changing consumption habits have increased the volume and complexity of municipal waste. Municipal waste management is often based upon traditional practices of fixed collection routes, manual tracking, fragmented data, and minimal waste separation. This inefficiency contributes to higher collection costs for municipalities, lower material recovery rates, and less awareness of waste flows. A Circular Economy approach provides an alternative to the linear "extract-produce-consume-dispose" model by keeping resources viable through reduced, reused, recycled, recovered, or regenerated [1, 2].

Circular Waste Management is a holistic application of the circular economy through the entire lifecycle of waste management - from waste avoidance to waste collection, sorting, re-use, recycling, recovery and reintroduction of recycled materials back into production. This will be supported by there needs to be adequate and timely information on the quantity, quality, location, how it was created, and how much it has been moved. Digital technology will allow us to connect our physical waste processing with sensors, communications, analysis, and decision-making software. Internet of Things (IoT) devices enable real-time monitoring of waste volumes and the current state of collections. Artificial Intelligence (AI) will help classify waste, predict waste generation, and optimize routes. Blockchain can improve the traceability of material flows, whereas cloud computing, edge computing, and digital twins can support data integration, simulation, and operational decision-making [3-6].

The combination of these technologies has contributed to the development of smart circular waste management, in which data are used to improve collection efficiency, automated sorting, material recovery, transparency, and circular resource flows. However, the effectiveness of such systems depends not only on individual technologies but also on their ability to exchange data and operate as an integrated system. IoT sensors, for example, provide operational data, but analytical models are required to transform that data into predictions and decisions. Similarly, blockchain can preserve the integrity of transaction records, but its effectiveness depends on the reliability of the data generated outside the blockchain.

Although they offer many advantages for managing waste, there are also significant barriers to implementing new technologies for digital waste. For example, differences in hardware and software (e.g., different types of devices) and variations in how these devices communicate with one another could make it difficult to get all of them to work together seamlessly; sensor inaccuracies, missing data, unreliable connectivity, cyber security risks, issues related to customer/team member privacy, excessive power consumption by devices and/or networks, and high costs associated with developing/implementing systems are examples of other ways that reliability and scalability may be limited [6-7]. Consequently, despite the contributions of prior reviews, there remains a need for a systematic synthesis that specifically compares multiple digital technologies across the circular waste-management lifecycle, based on their functional roles, supported processes, reported operational and circularity-related outcomes, evaluation approaches, and technical limitations. The distinctions between the present study and the most closely

*Corresponding author.

related prior reviews are discussed in the following section. This study addresses that need through a Systematic Literature Review of digital technologies enabling smart circular waste management using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 framework. This study aims to identify and classify the technologies used across the waste-management lifecycle, synthesize their reported operational and circularity-related outcomes, and examine the technical challenges and research gaps that must be addressed to develop interoperable, scalable, secure, and empirically validated systems.

II. Related Work and Positioning of the Present Review

Prior studies have examined the intersection of digital technologies, circular-economy strategies, and waste management from several perspectives. At the broader circular-economy level, Bressanelli et al. [3] examined how digital technologies support usage-oriented circular business models, while Kristoffersen et al. [4] proposed a digitally enabled circular-strategies framework for manufacturing companies. These studies established that digitalization can facilitate information exchange, resource optimization, product-life extension, and circular value creation. However, their primary focus was on business models and manufacturing strategies rather than on the full waste-management lifecycle.

Cagno et al. [5] extended this stream of research through a systematic literature review of the role of digital technologies in operationalizing the circular-economy transition. Their study connected digital technologies with broad circular strategies and provided a cross-sectoral account of how digitalization can support circular practices. Nevertheless, its scope was not specifically organized around waste generation, collection, sorting, reuse, recycling, recovery, valorization, and the reintegration of secondary materials into production. It also did not systematically compare digital technologies by their functional roles, report waste-management outcomes, or identify implementation limitations.

Other studies and reviews have concentrated on particular technologies or application domains. Antony Jose et al. [18], for example, reviewed recent applications of artificial intelligence in battery recycling, illustrating the value of AI for classification, process optimization, and material recovery within a specific waste stream. The studies by Bello and Maladzhi [16] and Steenmans et al. [23] investigated, respectively, sustainable waste-reduction methods using additive manufacturing and blockchain technology for plastic waste governance and tracking. The study by Meng et al. [24] examined the application of digital twin technology to circular economy concepts in the construction industry. Lastly, Goel et al. [25] conducted a review of the opportunities and risks involved in digitalization toward achieving environmental sustainability.

The present review addresses this fragmentation by combining systematic evidence synthesis and bibliometric analysis to examine multiple digital technologies across the circular waste-management lifecycle. In contrast to Cagno et al. [5], the unit of analysis is specifically smart circular waste management rather than the general operationalization of circular-economy strategies. In contrast to Zyoud and Zyoud [10] and Khan et al. [12], the review is not limited to publication trends, thematic structures, or knowledge networks. It also extends technology-specific reviews, such as those by Ahmed et al. [6] and Antony Jose et al. [18], by comparing IoT, artificial intelligence, machine learning, blockchain, cloud, and edge computing, digital twins, big data analytics, RFID, GPS, and additive manufacturing within a single analytical framework. The above review assesses these technologies on four bases. The first is the technology functions and operational roles for each. Second are the waste management processes that each supports. Third, there are their reportedly measured performance in terms of costs, environmental impacts, and the circular economy. Fourth are the known technical barriers and areas requiring further study that limit the use of such technologies in ways that allow them to be used together.

III. Methodology

This section outlines the research design, research question(s), eligibility criteria, study screening process, and analysis techniques to find and synthesize research about digital technologies for smart circular waste management.

A. Research Design

The research began with a comprehensive literature search following the PRISMA 2020 guidelines [8,9]. Articles were retrieved from two leading academic databases, Scopus and Web of Science, covering publications from 2015 to 2025. Structured Boolean search equations were developed using targeted keywords such as “smart circular economy,” “waste management,” “digital technology,” “AI,” “IoT,” “blockchain,” and “resource recovery.” Bibliometric analysis was conducted using VOSviewer to uncover keyword co-occurrence patterns, thematic clusters, and emerging research trends in the intersection of smart technologies and circular economy within waste management.

B. Research Questions (RQ)

The four questions that guided this study were:

RQ1: Which types of digital technology have enabled smart circular waste management?

RQ2: In what way do the various technologies function as a part of different waste management processes?

RQ3: What reported outcomes related to operation, economy, environment, and circularity resulted from the use of digital waste management technologies?

RQ4: How have limited technical capabilities and a lack of research hindered the implementation, integration, and scale-up of digital waste-management technologies?

C. Eligibility Criteria

Eligible studies included those that were peer-reviewed and published in English; included in either the Scopus or Web of Science databases; published between 2015 and 2025; used a minimum of one digital technology to support waste management or circular resource management practices; and provided information regarding how the technology was implemented, methodologies, results/outcomes, as well as any obstacles/challenges encountered using the technology. Studies

that did not report on the technology's implementation were therefore excluded from this review. Editorials, non-academic reports, unpublished manuscripts, inaccessible full-texts, and studies without a meaningful digital-technology component were excluded. Additionally, articles focused solely on chemical treatment, material characterization, pollution, or waste-processing reactions, and that did not include a relevant computational or informational technology aspect, would also be excluded. The review articles were used to provide additional background and references for reviewing purposes only and were not used as primary data sources.

D. Study Screening

The study screening process followed PRISMA 2020 guidelines to ensure methodological rigor and transparency in identification, screening, and inclusion (Fig. 1).

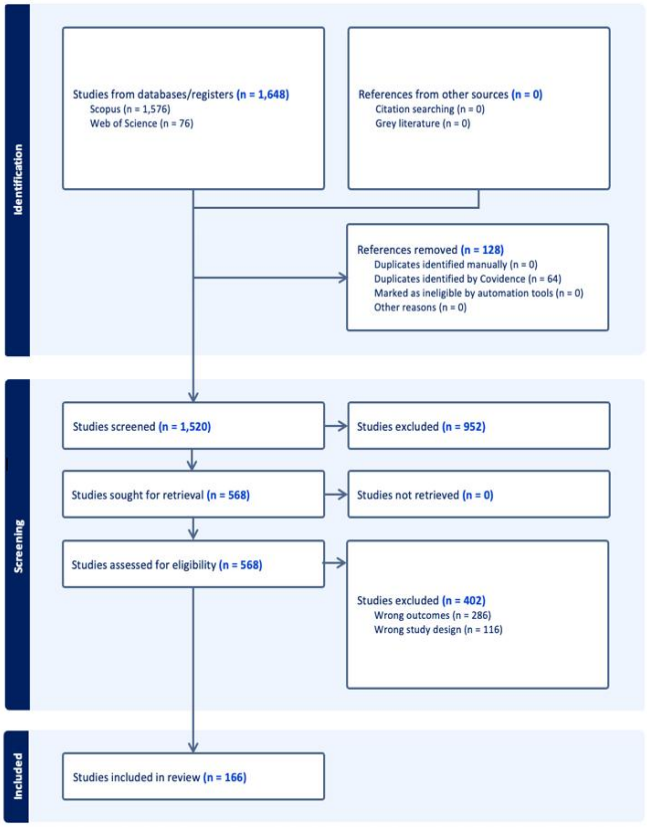


Fig. 1. PRISMA protocol.

A total of 1,648 records were retrieved from two databases: Scopus (n = 1,576) and Web of Science (n = 72) using structured Boolean searches. After removing 128 duplicates, 1,520 records remained for title and abstract screening. Strict inclusion and exclusion criteria were applied, resulting in 952 exclusions for irrelevance. The remaining 568 articles underwent full-text assessment, with all full texts retrieved. Of these, 402 were excluded: 286 for lacking relevant outcomes and 116 for unsuitable study design. Finally, 166 studies met all criteria and were included for bibliometric analysis and expert validation. This represents 10.1% of the initial records, underscoring the value of a rigorous multi-phase screening process. The selection process is summarized in the PRISMA 2020 flow diagram.

IV. Results

This Section Present Results of the Systematic Literature Review and Bibliometric Analysis Conducted to Examine Digital Technologies that Enable in Smart Circular Economy and Waste Management. First, the Bibliometric Analysis that Identifies the Conceptual Structure, domain keywords, and themes from the Literature. The Section Further Synthesizes the Evidence According to the Four Research Questions, Covering the Enabling Digital Technologies, Their Roles in Supporting Waste-management Processes, the Reported Operational and Sustainability Outcomes, and the Remaining Technical Limitations and Research Gaps. Overall, This Section Demonstrates the Transition of Waste-management Research from Conventional Recycling-oriented Approaches Toward Integrated, Technology-enabled, and Circular Waste-management Ecosystems.

A. Bibliometric Analysis

The bibliometric analysis offers a detailed view of research trends and thematic developments in Smart Circular Economy (SCE) and waste management. Using VOSviewer, co-occurrence maps and keyword heatmaps were generated from 166 articles, revealing conceptual structures and emerging topics (Fig. 2). The co-occurrence network shows “circular economy” as the most frequent and central term, closely linked to “waste management,” “recycling,” “sustainability,” and “environmental impact” (Fig. 2a). Three thematic clusters emerge. The first centers on technological enablers, with keywords such as machine learning, artificial intelligence, digital twin, and IoT, showing the integration of smart technologies in circular systems. The second focuses on material innovation and resource recovery, featuring terms like plastic waste, 3D printing, additive manufacturing, and life cycle assessment, highlighting closed-loop production and material efficiency. The third addresses policy and systems thinking, with terms including urban mining, blockchain, stakeholder engagement, and supply chains, reflecting interest in governance and socio-economic integration.

Keyword co-occurrence over time shows a shift in thematic focus (Fig. 2b). Terms such as blockchain, digital twin, urban mining, and machine learning appear in yellow, indicating recent growth, especially after 2022. In contrast, waste management and recycling appear in green and blue, marking earlier emergence and continued relevance. The keyword density heatmap confirms the centrality of circular economy, recycling, and waste management, which appear in bright yellow, reflecting high frequency and strong connectivity (Fig. 2c). Keywords like life cycle, environmental impact, and resource efficiency bridge technological innovations and sustainability outcomes across both network and heatmap views. Overall, the results show a transition from traditional recycling themes toward integrated, data-driven, and systems-based approaches. The visualizations also indicate a field that remains rooted in sustainability while expanding into digital innovation.

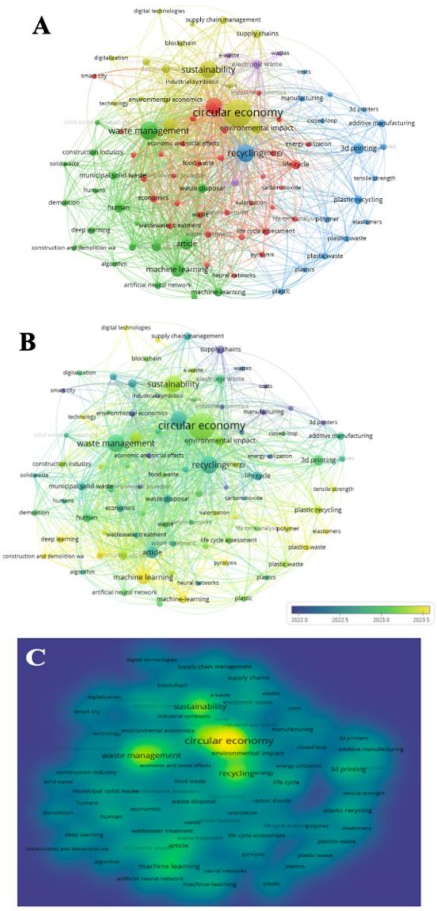


Fig. 2. Keyword co-occurrence map of research.

B. Scientific Performance and Production

Scholarly research in sustainable innovation is strongly anchored in circular economy principles, which account for 15% (n = 685) of the examined discourse (Fig. 3). This dominance reflects a shift from linear consumption toward regenerative systems emphasizing resource efficiency and lifecycle extension [10]. Waste management (6%, n = 288) and recycling (6%, n = 277) follow as key operational constructs, demonstrating circular principles in industrial and urban contexts. Sustainable development (5%, n = 225) and sustainability (5%, n = 222) further align environmental stewardship with economic resilience. Together, these themes reveal an integrated systems-thinking approach essential for addressing global resource constraints. Technological themes are also prominent. Machine learning (4%, n = 165), artificial intelligence (2%, n = 111), and 3D printing (5%, n ≈ 272) represent a growing nexus of smart tools embedded in sustainable strategies. These enable resource flow optimization, predictive analytics, and decentralized manufacturing, reducing carbon footprints and logistical demands. Specialized topics such as environmental impact (2%, n = 83), decision-making (2%, n = 82), and life cycle assessment (2%, n = 67) provide evaluation frameworks for system-wide sustainability. Supply chains (2%, n = 70) and economic aspects (2%, n = 69) highlight the importance of integration across production and distribution stages [11,12].

In addition, emerging themes- blockchain (1%, n = 33), polymer (1%, n = 33), and environmental economics (1%)-signal interdisciplinary expansion. They link digital trust mechanisms, material innovation, and economic modelling to enhance transparency and accountability in circular systems. Overall, the thematic structure reflects a technically diverse yet holistically aligned research ecosystem. Foundational topics coexist with advanced tools and methods, indicating a maturing field that adapts to ecological, technological, and socio-economic challenges [13-15]. Building on this base, temporal analysis can reveal how priorities shift over time and how technological innovations are increasingly embedded in sustainable waste management discourse.

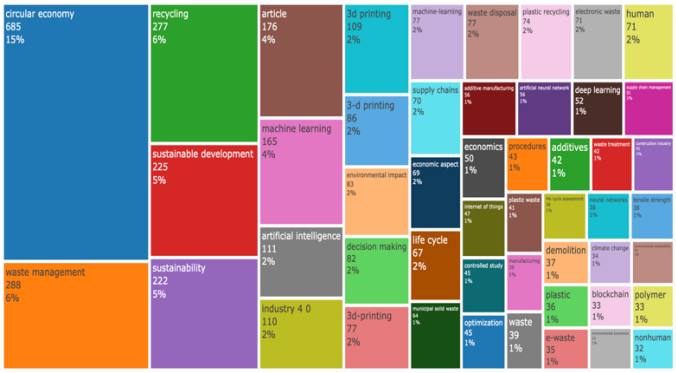


Fig. 3. Treemap of the most frequent terms in smart circular waste management studies.

The temporal distribution of emerging terms in sustainable waste management research shows a steady evolution in thematic focus and technological integration over the past decade. “Circular economy” remains the most dominant and persistent term, appearing consistently from 2020 to 2025 with high research intensity (Fig. 4).

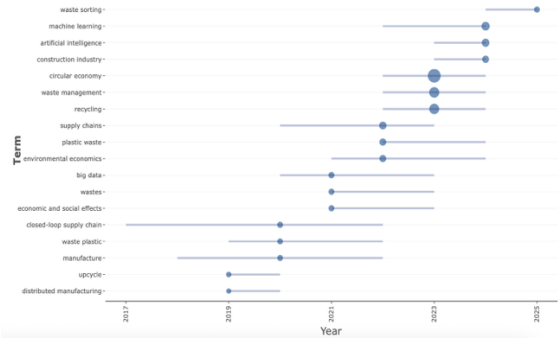


Fig. 4. The temporal distribution.

Its prominence reflects sustained academic and policy commitment to shifting from linear to regenerative systems centered on material reuse, systemic efficiency, and ecological balance. “Waste management” and “recycling” gained prominence in the same period, serving as operational frameworks for implementing circular economy principles. Their strong co-occurrence and longevity underline their foundational role across engineering, environmental science, and public policy [16]. In addition, technological integration has intensified since 2021, with increasing mentions of “machine

learning," "artificial intelligence," and "big data." These trends point to growing adoption of data-driven approaches for waste flow optimization, consumption prediction, and responsive infrastructure design. The emergence of "waste sorting" in 2025, alongside AI themes, signals a shift toward automated, real-time material recovery. "Supply chains" and "plastic waste" have maintained steady attention since 2020, highlighting concerns over material accountability and polymer pollution. Mentions of "environmental economics" and "economic and social effects" reflect efforts to align market mechanisms with ecological outcomes. Earlier themes such as "distributed manufacturing," "upcycle," and "closed-loop supply chain" emerged in 2017–2018 but later declined in visibility, likely absorbed into broader frameworks or deprioritized for scalable practices. Overall, the timeline shows a deliberate integration of AI, economic analysis, and systemic efficiency into circular economy strategies, marking a transition toward actionable, technology-enabled solutions [17-19].

C. Addressing RQ1: Which Types of Digital Technology Have Enabled Smart Circular Waste Management?

The reviewed studies identified several digital technologies that support smart circular waste management. The major technologies include: the Internet of Things (IoT), Artificial Intelligence (AI) and machine learning, blockchain, cloud computing, edge computing, digital twins, big data analytics, Radio Frequency Identification (RFID), and GPS. Additionally, the keyword map and tree map suggest that there has been an increase in the connections between machine learning, AI, Industry 4.0, the circular economy, and research into recycling and waste management. The following are the technologies for IoT: smart bin systems with ultrasonic sensors, weight sensors, environmental sensors, RFID tags, and vehicle tracking systems. This technology provides real-time information on waste amount, quality, position, and movement [6,10]. In addition, AI and machine learning are being utilized for classification of waste types, prediction of waste amounts, optimization of waste collection routes, automation of decision-making processes [18,19], and also for improving material tracking capabilities, transaction reliability, and transparency within the waste and recycling chain through the use of blockchain [4,5]. Digital twins and cloud-based platforms support data integration, simulation, and scenario evaluation, while additive manufacturing and 3D printing connect recovered materials with circular production processes [12,16]. The findings indicate that no single technology independently enables smart circular waste management. Instead, the reviewed systems rely on combinations of sensing, communication, analytical, traceability, and operational technologies. This convergence reflects a transition from isolated waste-management applications toward integrated digital ecosystems.

D. Addressing RQ2: In What Ways Do Digital Technologies Support Different Waste-Management Processes?

Digital technologies support different but interconnected functions across the waste-management lifecycle. IoT sensors and identification technologies serve as the sensing component and collect a variety of data, including bin levels, total waste generated, temperature, humidity, vehicle location, etc. The communication systems then send this collected data to an application (cloud, edge, or local) that stores it, processes it, and provides analytics. These functions allow the city/organization to monitor in real time the flow of materials, prevent bin overflows, schedule collections, and improve visibility into where waste is coming from [6,10]. AI and Machine Learning algorithms take the raw data and turn it into predictive actions/recommendations. Some examples of their use include classifying waste using images, predicting future waste generation, optimizing garbage truck routes, detecting anomalies in waste patterns, and automatically sorting waste. Blockchain provides a traceability function by recording waste transfers, recycling activities, and material transactions. Digital twins and decision-support systems allow alternative collection, treatment, and resource-recovery scenarios to be evaluated before implementation [4,5].

The functional relationship among these technologies can be represented in Fig. 5. Circular actions include waste collection, sorting, recycling, reuse, recovery, valorization, and the reintegration of secondary materials into production. The findings therefore show that digital technologies contribute not only to waste monitoring but also to the coordination of physical and information flows throughout the circular waste-management system.

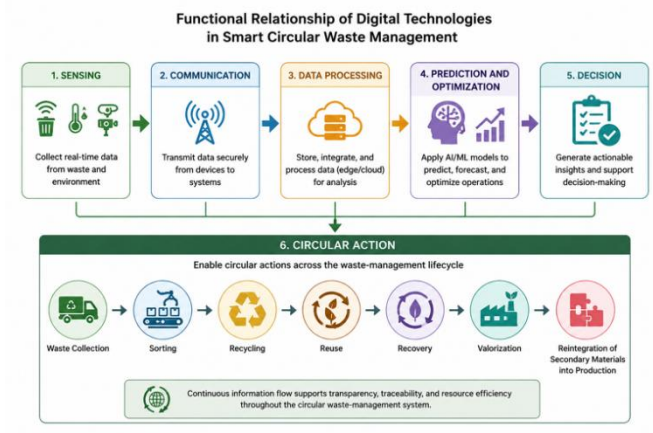


Fig. 5. Functional relationship of digital technologies in smart circular waste management.

E. Comparative Technology Suitability

The evaluated technologies differ from each other in terms of maturity, deployment effort, computational demand, and use case. IoT, RFID, and GPS are applicable as sensors for detection/identification and for localization. AI and Machine Learning can be used to classify, predict, optimize routes, and automatically sort products when sufficient training data and computational power are available. For storing and analyzing large amounts of data centrally, cloud computing is best. Edge Computing is typically used for real-time applications that require fast local processing. For tracing products throughout the supply chain (e.g., through certification), blockchain is an interesting solution. However, due to high computational demands and communication efforts, it does not make sense for simple monitoring. Digital Twins are best used for simulations, facility planning, or scenario evaluation. However, the value of digital twins depends heavily on their ability to be synchronized reliably with the corresponding physical assets. Thus, no technology is suitable for all cases. The appropriate technology

depends on the specific operational goals, the relevant infrastructure, the maximum allowed time to react, the amount of available data, costs, and legal obligations.

F. Addressing RQ3: What Operational, Economic, Environmental, and Circularity-Related Outcomes Have Been Reported?

The reviewed literature reports several operational advantages that can be derived from the application of Digital Waste Management (DWM) systems. These consist of: Real-Time Monitoring, Bin Overflow Reduction, Improved Waste Classification, Dynamic Scheduling of Collections, Reduced Length of Collection Routes, Optimized Vehicle Utilization, and Faster Operational Decision Making [10,11], as well as AI-supported forecasting and optimization, allowing waste managers to prioritize locations with the highest demand for services and avoid collecting from non-high-demand locations where possible. In addition to operational advantages, DWM systems provide economic advantages, such as reduced Fuel Use and transportation costs, reduced manual monitoring requirements, and an optimized Allocation of Vehicles, personnel, and Processing Capacity. However, the extent to which each study identifies economic advantages varies due to differences in System Scale, Infrastructure, Type of Waste being managed, and Evaluation Methodology. Some studies identify Economic Advantages based on Cost Reductions and Route Distances, whilst others base their identification on the Accuracy of Classification, Time to Process Materials, and Resource Productivity.

Additionally, there are Environmental Advantages associated with the use of DWM Systems, including Reduced Emissions from Collection Vehicles, Less Dependence on Landfills, Enhanced Segregation of waste, and Greater Efficiency in Recovering Recyclables. There are also Circular Economy Advantages associated with the use of DWM Systems, including Increased Recycling Rates, resource recovery from Waste, Valorization of Waste Streams, and Traceability of Materials throughout Supply Chains. Furthermore, Closed Loop Production and Additive Manufacturing using Recovered Materials are facilitated by DWM systems' ability to track the movement of materials and the activities related to recycling [16-18]. Additionally, the Accountability provided by Blockchain and Digital Tracking Technologies allows for documentation of Material Movements and Recycling Activities.

G. Addressing RQ4: How Do Technical Limitations and Research Gaps Hinder the Implementation, Integration, and Scaling of Digital Waste-Management Technologies?

The key findings from the review suggest that significant technological work will be required to overcome the technical barriers to implementing an integrated smart waste management system. The majority of current smart waste-management solutions rely on disparate sensors, networks, databases, analytical algorithms, and operational systems. As a result, differences in technology (e.g., sensor type), data format (e.g., binary vs. CSV), and networking architecture (e.g., LAN vs. WAN) can create barriers for data transfer across different systems [5,7]. Another challenge identified in the literature concerns data quality. When sensors fail to operate correctly, measure inaccurately, leave gaps in measurement, do not consistently record timestamps for measured variables, or lack historical data records, it decreases the accuracy of predictive, classification, and optimization models. In addition, AI-based systems used for sorting recyclables into appropriate categories may function poorly with waste mixtures that contain contaminants, are physically broken or otherwise damaged, or consist of items not yet encountered during model training. Therefore, models should be tested across a wide range of "real-world" scenarios rather than only those modeled in controlled experimental settings. Other limitations include cybersecurity and privacy issues associated with connected devices and cloud computing; limited connectivity to remote locations; high power consumption with most types of sensors; and the costs of purchasing, integrating, operating, maintaining, and eventually replacing all required equipment [7].

One area where there is currently a significant knowledge gap concerning long-term field testing is the relative scarcity of large-scale long-term field trials. Most of what has been reported in the literature has come from simulations, laboratory tests, prototype demonstrations, and/or very short-duration pilot projects. While each of these approaches demonstrates the technical feasibility of the proposed solutions, they provide little information about long-term operation and maintenance requirements; the potential scalability of deployed equipment; the potential financial sustainability of the solutions; and the overall impact on environmental performance.

H. Reproducibility and Evidence Strength

The level of reproducibility was variable across the reviews; many implementations documented architectures or performance results but lacked a publicly available dataset, open-source code, all required hardware details, model configuration details, or a standard test environment for benchmarking. The lack of reproducibility is especially relevant for AI/Machine Learning applications, where performance varies based on data content, label accuracy, and the pre-processing performed prior to training. Additionally, there were differences in the amount of evidence supporting each study type. Conceptual and simulation studies demonstrated feasibility; laboratory/prototype studies demonstrated technical validity; and pilot or operational deployment studies demonstrated greater validity with respect to maintenance costs, scalability, and long-term reliability.

V. Discussions

This section will examine how the circular economy (CE) concept is influenced by its convergence with digital technologies as well as the complementary functions of identified digital technologies, the conditions that are necessary for CE to be implemented, and implications for research and practice.

A. Convergence of Circular Economy and Digital Technologies

The findings demonstrate that circular economy, waste management, recycling, sustainability, and sustainable development remain the conceptual foundations of smart circular waste-management research. The dominance of "circular economy" in the term distribution indicates that digital technologies are primarily positioned as mechanisms for

extending material lifecycles, improving resource efficiency, and reducing dependence on disposal-based waste systems [1,2]. This result is consistent with studies that describe the smart circular economy as the integration of circular strategies with data acquisition, analytics, and digitally enabled decision-making. The results also reveal a gradual transition from conventional recycling and waste treatment to data-driven, intelligent waste management systems. The increasing use of machine learning, artificial intelligence, the Internet of Things, blockchain, digital twins, and big data-related terms indicates that the field is expanding beyond traditional environmental management. These technologies support the monitoring, classification, prediction, optimization, and traceability of waste and material flows. Machine learning and artificial intelligence are particularly relevant for waste sorting, generation forecasting, operational optimization, and environmental assessment [18,19]. Their increased visibility after 2021 suggests that waste-management research is moving toward more automated, responsive systems. The presence of 3D printing, additive manufacturing, polymers, and plastic recycling also demonstrates the connection between digital waste management and circular production. Additive manufacturing can use recovered materials as production inputs and facilitate decentralized manufacturing, although material quality, energy consumption, and recyclability must still be considered [12,16]. Therefore, digital technology should not be interpreted only as an information-management instrument; it can also reshape the physical processing, recovery, and reuse of secondary materials.

Nevertheless, the term distribution also shows fragmentation in the terminology used by researchers. Terms such as "3D printing," "3-d printing," and "3d-printing" represent the same concept but appear separately in the dataset. Similar normalization may be required for terms such as artificial intelligence, machine learning, Internet of Things, and related abbreviations. Standardizing equivalent terms is important for avoiding the underestimation of technology frequencies and improving the consistency of evidence synthesis.

B. Functional Roles and Integration of Digital Technologies

The reviewed technologies perform different but complementary functions within smart circular waste management. Internet of Things devices and sensors form the primary data-acquisition layer. Smart bins, ultrasonic sensors, weight sensors, cameras, Global Positioning System (GPS) devices, and Radio-Frequency Identification (RFID) tags can collect data on waste volume, composition, location, and movement [20,21]. This collected data will allow for a real-time view of the amount of waste that is being produced, prevent overflows from happening in the future, and optimize vehicle and facility use as part of the process.

The analytical layer of this process would use Artificial Intelligence (AI) and Machine Learning (ML) to take sensor and operational platform data and turn it into classifications, predictions, or recommendations. AI & ML are used in various ways, including image-based waste identification, Waste Generation Forecasting, Route Optimization, Contamination Detection, and Automated Sorting [22]. The reliability of these systems depends upon having representative databases, consistent database labels, appropriate computational resources, and testing under actual working conditions.

Blockchain provides a different function by supporting data integrity, transaction transparency, and material traceability. It can be used to record waste transfers, recycling certificates, extended-producer-responsibility transactions, and the chain of custody of recoverable materials [23]. However, blockchain does not guarantee that the original data are accurate. Incorrect sensor readings or unreliable manual inputs may remain incorrect even after being stored in an immutable ledger. Blockchain should therefore be integrated with reliable sensing, identity management, data-validation procedures, and cybersecurity controls. Digital twins and decision-support systems provide simulation and scenario-analysis capabilities. They can represent waste collection, facility capacity, material recovery, and reverse supply-chain processes in a digital environment [21,24]. These tools allow decision-makers to compare alternative operational scenarios before implementing them in physical systems. Nevertheless, an actual digital twin requires continuous or periodic synchronization between physical assets and their digital representations. Studies using only static simulation should therefore avoid describing their models as digital twins. Cloud computing and edge computing support data storage and processing across these technology layers. Cloud platforms provide scalable storage and centralized analytics, while edge computing can process data close to sensors, collection equipment, or sorting machines. Edge processing may reduce latency and network traffic, but it introduces additional requirements related to device management, security, and computational limitations [25].

These relationships indicate that smart circular waste management should be treated as an integrated digital ecosystem rather than as the implementation of a single technology. Its functional sequence can be represented as:

$$\text{Sensing} \rightarrow \text{Communication} \rightarrow \text{Data Processing} \rightarrow \text{Prediction and Optimization} \rightarrow \text{Decision} \rightarrow \text{Circular Action.}$$

Circular actions include collection, sorting, reuse, recycling, recovery, remanufacturing, and waste valorization. Feedback from these processes should return to the sensing and analytical layers, creating a continuous learning and improvement cycle.

C. Key Factors, Technologies, and Implementation Conditions

The literature indicates that technological performance is influenced by technical, organizational, regulatory, and economic conditions. The factors in Table I were synthesized from recurring findings in the reviewed literature. They were not derived from expert ranking or Fuzzy Delphi validation.

TABLE I. KEY FACTORS, TECHNOLOGIES, AND IMPLEMENTATION CONDITIONS IDENTIFIED IN SMART CIRCULAR WASTE MANAGEMENT

Category	Main Factor	Subfactors	Literature	Synthesis of Findings
Drivers	Policy and Regulatory Support	Incentives, digital standards, regulatory enforcement, extended producer responsibility	[21,23]	Regulations and incentives can encourage investment, data sharing, traceability, and the adoption of digital waste-management systems.
	Environmental Pressure	Landfill reduction, emission targets, resource scarcity, pollution control	[1,2,26]	Environmental pressures increase demand for technologies that support recycling, resource recovery, and waste reduction.
	Institutional Readiness	Leadership commitment, government support, organizational coordination	[4,27]	Institutional readiness facilitates technology procurement, data governance, integration, and long-term system operation.
	Economic and Operational Benefits	Cost savings, resource efficiency, route efficiency, improved productivity	[3,6,26]	Digital technologies can reduce unnecessary collection activities and improve operational resource allocation.
Barriers	Financial Constraints	Initial investment, maintenance costs, uncertain return on investment	[7,26]	High acquisition and lifecycle costs restrict adoption, particularly in resource-constrained municipalities.
	Digital Infrastructure Limitations	Limited connectivity, insufficient sensors, inadequate computing capacity	[6,7,21]	Weak communication and computing infrastructure reduce system reliability and scalability.
	Data Quality and Availability	Missing data, inconsistent formats, inaccurate sensor readings, limited historical records	[6,19,22]	Poor data quality decreases the reliability of classification, forecasting, optimization, and decision support.
	Cybersecurity and Privacy	Insecure devices, unauthorized access, sensitive location and operational data	[7,23,25]	Connected waste systems require authentication, encryption, access control, and privacy-preserving data management.
Challenges	System Interoperability	Heterogeneous devices, incompatible protocols, isolated databases	[4,5]	Interoperability is required to integrate sensing, communication, analytics, traceability, and operational platforms.
	Scalability and Field Validation	Prototype-to-city-scale transition, high data volume, network congestion, limited long-term testing	[6,7,20]	Many proposed systems remain limited to simulations, laboratory tests, or small-scale pilots.
	Technology–Policy Alignment	Rapid technological change, unclear data ownership, outdated regulation	[21], [23], [24], [27]	Adoption may be delayed when policies do not adequately address data sharing, automation, and digital transactions.
Success Factors	Stakeholder Collaboration	Public–private partnerships, citizen participation, cross-sector data sharing	[2]	Collaboration provides access to infrastructure, expertise, funding, operational data, and public acceptance.
	Digital Ecosystem Readiness	Integrated platforms, standards, reliable networks, data-governance mechanisms	[4,5]	Successful implementation depends on supporting infrastructure and governance rather than isolated technology adoption.
	Technical Capacity Building	Staff training, system maintenance, digital literacy, analytical capability	[7]	Skilled personnel are required to operate devices, maintain platforms, interpret analytics, and manage failures.
Technologies	Internet of Things and Sensors	Smart bins, fill-level sensors, weight sensors, environmental monitoring, vehicle tracking	[6,20,21]	Internet of Things technologies provide real-time data for monitoring, collection scheduling, and overflow prevention.
	Artificial Intelligence and Machine Learning	Classification, forecasting, route optimization, anomaly detection	[11,18,19]	Artificial intelligence converts waste data into classifications, predictions, and operational recommendations.
	Blockchain	Material traceability, recycling records, transaction integrity, certification	[4,23]	Blockchain can improve auditability when it is connected to reliable physical and off-chain data sources.
	Digital Twins and Decision-Support Systems	Simulation, scenario analysis, capacity planning, multi-criteria decisions	[22,24]	These technologies support the evaluation of alternative operational and circular-system configurations.
	Cloud and Edge Computing	Data storage, distributed processing, dashboards, real-time analytics	[6,25]	Cloud platforms support scalable analysis, while edge processing reduces latency and communication requirements.
	Identification and Communication Technologies	RFID, GPS, LoRaWAN, cellular networks, wireless sensor networks	[6,20]	These technologies improve asset identification, data transmission, location monitoring, and collection visibility.
Opportunities	Smart-City Integration	Integration with transportation, energy, logistics, and public-service systems	[10,20]	Shared smart-city infrastructure can improve coordination, data access, and resource utilization.
	Automated Circular Operations	Robotic sorting, intelligent recycling, automated segregation, predictive maintenance	[11,18,22]	Automation can improve sorting consistency, processing efficiency, and material-recovery performance.
	Data-Driven Circular Business Models	Digital marketplaces, product-service systems, material platforms, circular supply chains	[3,5]	Digital platforms can support new economic mechanisms for reuse, recovery, and closed-loop material management.

Table I shows that technical and non-technical factors are closely connected. Although IoT, artificial intelligence, blockchain, digital twins, and cloud platforms provide distinct technological capabilities, their performance depends on data quality, interoperability, infrastructure, cybersecurity, institutional support, and human capacity. Therefore, implementing advanced technology without improving the surrounding digital ecosystem may produce isolated prototypes

rather than scalable circular solutions. The table also indicates that policies and stakeholder participation should not be treated as separate from technology implementation. Regulations influence data sharing, procurement, ownership, extended producer responsibility, and reporting requirements. Similarly, cooperation among municipalities, private operators, citizens, recyclers, and technology providers determines whether sufficient data and operational support are available.

D. Waste-Stream and Regional Applicability

Technology fit varies by waste type. Solid municipal waste (household trash) will benefit most from using IoT for monitoring, GPS for collection, routing, and AI for sorting. Most industrial wastes will use process monitoring, material accounting, big data analysis, and digital twin technology. Biomedical and hazardous wastes will require enhanced identity verification, tracing, security, and regulatory compliance. Plastic wastes and electronic wastes will have strong opportunities in AI to classify materials that can be recovered. Construction waste has potential uses as a digital twin, an inventory of materials, and tracking at the project level. Therefore, selecting technology should be based on the properties of the waste being processed (characteristics), the risk of contamination, the dollar value of the materials contained in the waste stream, how sensitive the waste stream is to regulation, and the objective(s) for recovering materials.

Regional conditions also influence the deployment of technologies. Urbanized areas that are highly connected may be able to support cloud-intensive AI-enabled systems as well as integrated systems; however, areas with low connectivity, limited service integration, poor infrastructure, and/or budgetary constraints will likely require less expensive sensors, RFID/GPS, or local edge processing. The feasibility of blockchain-based technologies such as digital traceability and/or advanced analytics is also highly dependent on the level of regulatory maturity in an area, the presence/absence of formal recycling programs and/or institutions that enforce extended producer responsibility policies, and the overall institutional capacity in an area. Results from large-scale funding of "smart city" pilot projects cannot be directly applied to resource-limited municipalities.

E. Implications for Research and Practice

The reviewed literature provides insight into the potential of digital technologies to support sustainable waste management practices. However, it highlights several challenges that will affect whether and how this can happen. These include the lack of standardized systems for collecting and processing data, the absence of established policies and regulatory structures governing the deployment of these technologies in waste collection and disposal systems, and the limited availability of data describing the physical operations of waste collection and processing systems. In addition to addressing these structural challenges, future research should examine not only individual technologies (such as sensors, sorting machines, and mobile apps) but also how they interact and how they can be used by stakeholders who make different types of decisions. For example, what happens when data generated about the composition of municipal solid waste is transmitted to local government agencies? How do those data inform agency decisions regarding where to locate new recycling facilities? What happens at the intersection of waste management and land-use planning? Answering questions like these will require cross-disciplinary research efforts involving experts in information technology, public administration, engineering, materials science, economics, urban planning, and sociology. In terms of specific research topics, there are many possibilities. Some examples might include developing new metrics for measuring success in waste management; designing decision-making tools for use by city officials responsible for managing large-scale recycling programs; identifying ways to reduce the amount of energy required to sort recyclables from non-recyclable trash; and improving methods for estimating the carbon footprint associated with the production of goods made from recycled materials.

VI. Conclusion

This systematic literature review and bibliometric analysis addressed the four research questions on digital technologies in smart circular waste management. For RQ1, the main technologies that enabled the development of intelligent waste management systems included the Internet of Things (IoT), artificial intelligence and machine learning, blockchain, cloud computing and edge computing, digital twins, big data analytics, RFID/GPS, and additive manufacturing. For RQ2, these technologies performed complementary functions across all phases of the waste-management lifecycle, including sensing/monitoring, classification/forecasting, route optimization, traceability, simulation, sorting, recycling, recovery, and material reintegration. For RQ3, the studies reviewed reported operational economic and environmental circularity-related benefits, such as reduced bin overflow, shorter collection routes, lower fuel and labor requirements, fewer collection-related emissions, improved waste segregation, greater material recovery, and stronger material traceability. However, these outcomes were synthesized qualitatively due to differences across studies in the technologies used, the scale of waste, evaluation times, indicators, and units. For RQ4, the primary barriers to adopting these technologies were poor data quality, limited interoperability, unreliable connectivity, cybersecurity/risk, privacy/harm, higher energy/life-cycle costs, and limited technical expertise. The evidence base is also limited by different methods of evaluating performance and by a few large-scale, long-term field-deployment studies. Therefore, future research should emphasize interoperable architectures, standardized performance measures, secure data governance, reproducible implementations, and long-term validation with respect to various waste streams and regions.

Declaration on Generative AI

ChatGPT (OpenAI) was used to assist in the visual development of Fig. 5. The conceptual framework, scientific content, and final design were subsequently reviewed and refined by the authors, who remain fully responsible for the figure's accuracy and integrity.

Acknowledgment

The authors gratefully acknowledge the support provided by Danida Fellowship Centre (Denmark) and Faculty of Computer Science. This research was funded by the 2026 Laboratory Grant

(Hibah Laboratorium) of the Faculty of Computer Science, Universitas Indonesia, under Agreement No. PKS-60/UN2.F11.D/HKP.05.00/2026

REFERENCES

[1] P. Ghisellini, C. Cialani, and S. Ulgiati, "A review on circular economy: The expected transition to a balanced interplay of environmental and economic systems," Journal of Cleaner Production, vol. 114, pp. 11–32, 2016, doi: 10.1016/j.jclepro.2015.09.007.

[2] M. Geissdoerfer, P. Savaget, N. M. P. Bocken, and E. J. Hultink, "The circular economy—A new sustainability paradigm?" Journal of Cleaner Production, vol. 143, pp. 757–768, 2017, doi: 10.1016/j.jclepro.2016.12.048.

[3] G. Bressanelli, F. Adrodegari, M. Perona, and N. Saccani, "Exploring how usage-focused business models enable circular economy through digital technologies," Sustainability, vol. 10, no. 3, Art. no. 639, 2018, doi: 10.3390/su10030639.

[4] E. Kristoffersen, F. Blomsma, P. Mikalef, and J. Li, "The smart circular economy: A digital-enabled circular strategies framework for manufacturing companies," Journal of Business Research, vol. 120, pp. 241–261, 2020, doi: 10.1016/j.jbusres.2020.07.044.

[5] E. Cagno, A. Neri, M. Negri, C. A. Bassani, and T. Lampertico, "The role of digital technologies in operationalizing the circular economy transition: A systematic literature review," Applied Sciences, vol. 11, no. 8, Art. no. 3328, 2021, doi: 10.3390/app11083328.

[6] M. M. Ahmed, E. Hassanien, and A. E. Hassanien, "IoT-based intelligent waste management system," Neural Computing and Applications, vol. 35, no. 32, pp. 23551–23579, 2023, doi: 10.1007/s00521-023-08970-7.

[7] M. Sharma, S. Joshi, D. Kannan, K. Govindan, R. Singh, and H. C. Purohit, "Internet of Things adoption barriers of smart cities' waste management: An Indian context," Journal of Cleaner Production, vol. 270, Art. no. 122047, 2020, doi: 10.1016/j.jclepro.2020.122047.

[8] Y. Wirani, I. Eitiveni, and Y. G. Sucahyo, "Framework of smart and integrated household waste management system: A systematic literature review using PRISMA," Sustainability, vol. 16, no. 12, Art. no. 4898, 2024, doi: 10.3390/su16124898.

[9] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," BMJ, vol. 372, Art. no. n71, 2021, doi: 10.1136/bmj.n71.

[10] S. Zyoud and A. H. Zyoud, "Internet of Things supporting sustainable solid waste management: Global insights, hotspots, and research trends," International Journal of Environmental Science and Technology, vol. 22, pp. 7641–7670, 2025, doi: 10.1007/s13762-024-06146-x.

[11] K. S. Rautela, M. K. Goyal, and R. Y. Surampalli, "AI and machine learning for optimizing waste management and reducing air pollution," Journal of Hazardous, Toxic, and Radioactive Waste, vol. 29, no. 3, Art. no. 04025014, 2025, doi: 10.1061/JHTRBP.HZENG-1483.

[12] S. A. R. Khan, M. Tabish, and Z. Yu, "Mapping and visualizing of research output on waste management and green technology: A bibliometric review of literature," Waste Management & Research, vol. 41, no. 7, pp. 1203–1218, 2023, doi: 10.1177/0734242X221149329.

[13] M. Z. Islam and L. Zheng, "Why is it necessary to integrate circular economy practices for agri-food sustainability from a global perspective?" Sustainable Development, vol. 33, no. 1, pp. 600–620, 2025, doi: 10.1002/sd.3135.

[14] F. De Felice, A. G. Fareed, A. Zahid, M. E. Nenni, and A. Petrillo, "Circular economy practices in the textile industry for sustainable future: A systematic literature review," Journal of Cleaner Production, vol. 486, Art. no. 144547, 2025, doi: 10.1016/j.jclepro.2024.144547.

[15] C. T. de Oliveira and G. G. A. Oliveira, "What circular economy indicators really measure? An overview of circular economy principles and sustainable development goals," Resources, Conservation and Recycling, vol. 190, Art. no. 106850, 2023, doi: 10.1016/j.resconrec.2022.106850.

[16] K. A. Bello and R. W. Maladzhi, "Innovative and best practices in sustainable strategies for waste reduction in additive manufacturing," Hybrid Advances, vol. 11, Art. no. 100527, 2025, doi: 10.1016/j.hybadv.2025.100527.

[17] S. Dhiman, B. Thakur, S. Kaur, M. Ahuja, S. Gantayat, S. Sarkar, R. Singh, and M. Tripathi, "Closing the loop: Technological innovations in food waste valorisation for global sustainability," Discover Sustainability, vol. 6, Art. no. 258, 2025, doi: 10.1007/s43621-025-01073-4.

[18] S. Antony Jose, C. A. D. Cook, J. Palacios, H. Seo, C. E. Torres Ramirez, J. Wu, and P. L. Menezes, "Recent advancements in artificial intelligence in battery recycling," Batteries, vol. 10, no. 12, Art. no. 440, 2024, doi: 10.3390/batteries10120440.

[19] N. Prioux, R. Ouaret, G. Hétreux, and J.-P. Belaud, "Environmental assessment coupled with machine learning for circular economy," Clean Technologies and Environmental Policy, vol. 25, no. 2, pp. 689–702, 2023, doi: 10.1007/s10098-022-02275-4.

[20] Á. Lozano, J. Caridad, J. F. De Paz, G. Villarrubia, and J. Bajo, "Smart waste collection system with low consumption LoRaWAN nodes and route optimization," Sensors, vol. 18, no. 5, Art. no. 1465, 2018, doi: 10.3390/s18051465.

[21] G. K. Shyam, S. S. Manvi, and P. Bharti, "Smart waste management using Internet-of-Things," in Proc. 2nd International Conference on Computing and Communications Technologies, Chennai, India, 2017, pp. 199–203, doi: 10.1109/ICCCT2.2017.7972264.

[22] J. J. S. Filho, A. D. Finisterra do Paço, and P. D. Gaspar, "Artificial intelligence and MCDA in circular economy: Governance strategies and optimization for reverse supply chains of solid waste," Applied Sciences, vol. 15, no. 9, Art. no. 4758, 2025, doi: 10.3390/app15094758.

[23] K. Steenmans, P. Taylor, and I. Steenmans, "Blockchain technology for governance of plastic waste management: Where are we?" Social Sciences, vol. 10, no. 11, Art. no. 434, 2021, doi: 10.3390/socsci10110434.

[24] X. Meng, S. Das, and J. Meng, "Integration of digital twin and circular economy in the construction industry," Sustainability, vol. 15, no. 17, Art. no. 13186, 2023, doi: 10.3390/su151713186.

[25] A. Goel, S. Masurkar, and G. R. Pathade, "An overview of digital transformation and environmental sustainability: Threats, opportunities, and solutions," Sustainability, vol. 16, no. 24, Art. no. 11079, 2024, doi: 10.3390/su162411079.

[26] B. Esmaeilian, B. Wang, K. Lewis, F. Duarte, C. Ratti, and S. Behdad, "The future of waste management in smart and sustainable cities: A review and concept paper," Waste Management, vol. 81, pp. 177–195, 2018, doi: 10.1016/j.wasman.2018.09.047.

[27] S. Bag, G. Yadav, P. Dhamija, and K. K. Kataria, "Key resources for Industry 4.0 adoption and its effect on sustainable production and circular economy: An empirical study," Journal of Cleaner Production, vol. 281, Art. no. 125233, 2021, doi: 10.1016/j.jclepro.2020.12523

Arabic News Text Classification Using Deep Learning Models with Dynamic N-grams

Ahmed I.Taloba[1], George Samy Rady[2], Khaled F. Hussain[3]
Department of Information Systems-Faculty of Computer and Information, Assiut University, Egypt[1]
Department of Information Technology-Faculty of Computers and Information Technology,
The Egyptian E-Learning University, Assiut, Egypt[2]
Department of Computer Science-Faculty of Computers and Information, Assiut University, Assiut, Egypt[3]

***Abstract*—The complexity and morphological richness of the Arabic language pose significant challenges in natural language processing (NLP), including issues with contextual understanding and feature extraction. Traditional deep learning architectures such as CNNs, LSTMs, and GRUs often struggle to effectively model these linguistic intricacies, limiting their performance on Arabic text analysis tasks. To address these limitations, the study integrates parallel multi-kernel word-level convolutional features into conventional and hybrid deep learning models. The convolutional windows capture short contextual relationships among neighboring word tokens, while recurrent components model longer sequential evidence; the framework does not directly analyze roots, affixes, or other within-word morphological structures. These enhancements are integrated into CNN, LSTM, GRU, and hybrid architectures such as LSTM-CNN and GRU-CNN. A comprehensive evaluation was conducted across varying learning rates to assess the impact of the enhanced configurations on model performance. The results indicate competitive performance within the evaluated architectures and dataset variants, although the magnitude of improvement depends on the model and learning rate. Under their best settings, the Dynamic N-gram LSTM-CNN achieved an accuracy of 93.32%, while the Dynamic N-gram LSTM achieved 93.57%. Because previously published studies use different corpora, class configurations, preprocessing pipelines, and evaluation protocols, these results are not presented as evidence of state-of-the-art superiority. Instead, the study provides a systematic within-study assessment of model sensitivity to architecture, preprocessing, and learning-rate selection. Future directions include character- and subword-level modeling, transformer-based architectures, and domain-specific tasks such as sentiment analysis and information retrieval.**

Keywords*—Dynamic n-grams; GRU-CNN hybrid; Arabic NLP; sequential dependencies; text classification

I. Introduction

The consumption and production of online news have greatly increased, as never before, due to technological advancements and the easy accessibility of the internet in the digital age. The Arabic language market, with a total of more than 420 million people, has witnessed a huge increase in the production of online news [1], [2]. Arabic news websites, including Al-Arabiya, Al Jazeera, and Youm7, reach a broad audience, providing information in many fields, like politics, sports, culture, and technology. The occurrence of this great reservoir of information has altered the aspect through which Arabic-speaking groups dwell on contemporary issues, and the necessity of digital channels in the formation of opinion and transfer of knowledge cannot be overstressed [3], [4]. Nevertheless, the increase in Arabic news material poses some difficulties as regards obtaining accessibility and classification of information. It is common among users to have difficulties progressing through the flood of content in search of the news of their interest [5], [6]. The necessity of automated classification systems has become very evident. Automated text classification would make it easier to organize the news articles in predefined categories and thereby allow a user to access the relevant articles. Such functionality is essential to end-users and makers of the content of the news, which will add to the user experience and increase the searchability of the articles [7], [8]. Besides the need to automate the increasing amount of news content, the special linguistic complexity of the Arabic language also necessitates the development of strong text classification systems specifically deployed in Arabic news management. These shall be described as complexities that require advanced machine learning and deep learning to be adopted.

The Arabic language presents distinctive challenges for automated text classification because of its morphological, syntactic, and lexical properties. Arabic is highly inflected, and surface forms may vary according to tense, gender, number, and case [9], [10]. A single root can also generate many derived forms through prefixes, suffixes, infixes, and templatic patterns, increasing word-level sparsity and making contextual interpretation important. Arabic additionally has a rich vocabulary, many near-synonyms, and context-dependent meanings. Modern Standard Arabic coexists with regional dialects that differ in syntax and vocabulary [11], [12], and digital news may contain variation across sources. Relatively flexible word order and the frequent absence of diacritics can further increase ambiguity. Accordingly, methods that combine local token context, longer sequence modeling, and controlled comparisons of stemmed and non-stemmed representations may help evaluate how these challenges affect classification. The present word-level models are not morphological analyzers and do not directly decompose Arabic roots or affixes.

A. Problem Statement

With the fast rate of expansion of Arabic news on the internet, users have found it difficult to find something of interest with ease. The current models of text classification are mainly developed to support English or less expressive

languages, which cannot be applied to the specifics of the Arabic language, i.e., the abundance of morphology, the flexibility of syntax, and the existence of different types of dialects [13], [14]. Moreover, the unavailability of customized sets and approaches curbs the functionality of classical machine learning and deep learning algorithms. Consequently, there is an urgent need to establish a strong system of categorizations dealing with the linguistic heaviness of Arabic and the categorization of news into sensible clusters. The mentioned gaps can be filled to increase usability among users, create real-time classifications, and cover various kinds of applications, including sentiment analysis and recommendation of personalized content.

B. Motivation of the Study

The sheer growth of the Arabic digital news sector has brought up the need to ensure efficient content organization to enhance access and relevance. The existing automated classification is not able to address this need because the language itself is as complex as it is morphologically rich and contextually dependent. In addition, the current developments in deep learning, particularly in hybrid models such as CNNs and LSTMs, demonstrate potential to overcome these challenges. This research derives interest in the possibility of using these sophisticated methods to develop a more powerful Arabic news classification system. This study will help to fill the divide between linguistic complexity and technological capacity and thus may contribute to the improvement of user experience and the practicality of automated Arabic text classification.

C. Significance of the Study

This study is relevant because Arabic text classification remains less extensively researched than English and other high-resource languages. The proposed research evaluates deep learning methods, including Dynamic N-grams and hybrid models, for Arabic news classification. It can support content accessibility, personalization, sentiment analysis, and trend detection. The findings provide additional empirical evidence on the behavior of deep learning architectures under different preprocessing and learning-rate settings for Arabic text classification. Finally, this research can contribute to more organized and accessible digital news environments for Arabic-speaking communities.

D. Key Contribution of the Study

- Dynamic Hybrid Model: Proposed a dynamic model integrated with dynamic n-gram and CNN, LSTM, and GRU in improving the classification of Arabic news.
- Optimized Learning Rates: Optimized the best learning rates that enhance the performance and the stability of the model.
- Systematic Evaluation of Dataset Variants: Systematically evaluated the six existing representations supplied with the Ultimate Arabic News Dataset to examine how provider-applied cleaning, stop-word removal, and stemming affect the evaluated deep learning architectures.
- Model Comparison: Conducted a controlled within-study comparison between baseline and Dynamic N-gram-enhanced architectures across multiple learning rates and dataset variants; the results are interpreted within this experimental setting rather than as a direct ranking against studies using different corpora.
- Development of Arabic NLP: Addressing the Arabic-specific issues helped in the development of improved tools in text analysis and classification.

Structure of the Study: The study includes an introduction (Section I), related works (Section II), methodology (dataset, preprocessing, model architectures, learning rate experiments) (Section III), results and analysis (Section IV), discussion, and conclusion with future directions for Arabic text classification (Section V).

II. Related Works

With the rapid rise of Arabic tweets, Alzanin et al. [15] suggest an automatic classification model that classifies tweets into five categories based on linguistic features and content. It compares two text representations: Word2Vec and stemmed tf-idf, and tests three classifiers: SVM, GNB, and RF. A manually annotated 35,600-tweet dataset was employed. Results indicate that RF and SVM using tf-idf and stemming obtained the best macro-F1 values (98.09%–98.14%), while RF obtained 98.95% accuracy, 99.54% precision, 96.58% recall, and 97.99% F1. However, GNB with Word2Vec did not work well. One major limitation is the difficulty presented by short, ambiguous, and mixed-dialect tweets.

Hossain et al. [16] propose a hybrid Attention-Based Transformer Model (ABTM) for classifying Arabic news that blends deep learning with conventional text representations such as TF-IDF and Bag-of-Words. An effective preprocessing pipeline involving cleaning, tokenization, lemmatization, and data augmentation was used to handle the richness of Arabic news text. The model combines contextual and statistical information and utilizes LIME for interpretability. Compared to AraBERT models, ABTM resulted in 97.69% accuracy, 97.13% precision, 97.11% recall, and 97.10% F1-score with various news categories. Although the model indicates robust performance and interpretability, its limitation resides in handling subtle dialectal variations and possible data imbalance between categories.

Zamzami et al. [17] report on the increasing necessity of determining the origin country of Arabic news articles to enable users to follow up on credible, pertinent news. It suggests a variety of classification models with both machine learning and deep learning methods, trained on varied feature extraction techniques. The models accurately classify articles into origin categories with a maximum accuracy of 94%, with little variation in performance between ML and DL methodologies. This shows the potential for application to regional news filtering. But there is a limitation in the difficulty of dealing with overlapping regional material and subtle linguistic variations, which could impact classification accuracy in some situations.

Imad et al. [18] introduce an Arabic text classification method using NLP methods and deep learning, with TF-IDF, Bag-of-Words, Word Embedding, and a suggested CNN model. The model was tested on three datasets—CNN, BBC, and OSAC—with high accuracy: 94.47% (One Hot), 93.98% (BOW), and 93.58% (TF-IDF) on CNN; 90.97%, 87.19%, and 86.35% respectively on BBC; and more than 98.8% on OSAC. The results validate the model's performance on diverse datasets. Yet, there are limitations such as performance differences on the basis of dataset specifications and possibly challenges in generalizability to diverse Arabic dialects and content structures in real-world use.

Othman et al. [19] tackle the emerging issue of Arabic fake news in the media by suggesting a hybrid approach using Arabic pre-trained BERT models (AraBERT, GigaBERT, MARBERT) augmented with 1D or 2D CNNs for efficient feature extraction and classification. Evaluated on three datasets—ANS, AraNews, and Covid19Fakes—the top-performing architecture (AraBERT + 2D-CNN) achieved F1-scores of 0.6188, 0.7837, and 0.8009, and surpassed current models on the ANS dataset with 71.42% accuracy, 75.6% precision, and 52.38% recall. The model also minimized training time. A primary limitation, though, is the relatively low recall on the ANS dataset, reflecting difficulty in detecting all occurrences of fake news.

Jamaleddyn et al. [20] investigate improving Arabic text classification by utilizing machine learning algorithms with hyperparameter tuning methodologies, i.e., grid search and random search, for model performance optimization. NLP techniques were incorporated to enhance text representation and classification accuracy. The suggested method was benchmarked with the CNN Arabic dataset, resulting in high performance: 95.16% accuracy, 94.64% recall, 94.04% precision, and 94.31% F1-score. Comparison with prior work validated the strength of the model. Still, one limitation is that it relies on tuning techniques, which are computationally intensive and dataset-dependent and may compromise scalability and generalizability to diverse Arabic language corpora.

Alzaidi et al. [21] introduce the TCAODL-ANA method—an improved Arabic news categorization model that integrates FastText word embedding, attention-based bidirectional GRU (ABiGRU), and Aquila Optimizer (AO) for hyperparameter optimization. Targeted at classifying news into seven categories, the model utilizes deep learning's contextual advantages and AO's optimization to enhance performance. Tested using simulations, TCAODL-ANA recorded 95.48% accuracy, 84.21% precision, 84.18% recall, and 84.17% F1-score, surpassing current methods. Although effective, the limitations of the model are relatively moderate precision and recall, implying possible difficulties with ambiguous or overlapped news topics, and heavy reliance on large computational resources for optimization.

Alroobaea [22] considers Arabic news classification based on 5,000 news articles in six categories: business, entertainment, health, politics, sports, and technology. Different preprocessing methods, word embeddings, and deep learning models—CNN, LSTM, and a CNN-LSTM hybrid—were considered. The hybrid model yielded the best accuracy of 93.15% and provided competitive results on other Arabic news datasets. Model architecture and preprocessing decisions are critical to the findings. Nonetheless, a limitation is the comparatively small size of the dataset, which can impact generalization, and the performance of the model can differ for diverse dialects or less-organized Arabic news outlets. The summarization of the related works is shown in Table I.

TABLE I. SUMMARIZATION OF THE RELATED WORKS

Author (s)	Approach	Dataset(s)	Best Results	Limitations
Alzani n et al., [15]	Word2Vec & tf-idf with SVM, RF, GNB	35,600 Arabic tweets	RF: Acc 98.95%, Prec 99.54%, Recall 96.58%, F1 97.99%	Poor GNB performance; ambiguity and dialectal mixing in short tweets
Hossai n et al., [16]	ABTM (Attention-Based Transformer Model) + TF-IDF/BOW + LIME	Benchma rk with AraBER T models	Acc 97.69%, Prec 97.13%, Recall 97.11%, F1 97.10%	Dialectal variation & potential data imbalance
Zamza mi et al., [17]	ML and DL models with various feature extraction methods	Multiple Arabic news datasets	Accuracy up to 94%	Overlapping regional content and linguistic subtleties may affect precision.
Imad et al., [18]	TF-IDF, BOW, Word Embedding + CNN	CNN, BBC, OSAC	CNN: Acc 94.47%, BBC: 90.97%, OSAC: >98.8%	Performance varies across datasets; dialect/genera lization concerns
Othma n et al., [19]	Hybrid: AraBERT/GigaB ERT/MARBERT + 1D/2D CNN	ANS, AraNews , Covid19 Fakes	F1: 0.6188 (ANS), 0.7837 (AraNews), 0.8009 (Covid19Fake s); Acc: 71.42% (ANS)	Low recall on ANS; difficult to capture all fake news cases
Jamale ddyn et al., [20]	ML + Grid/Random Search for hyperparameter tuning	CNN Arabic dataset	Acc 95.16%, Recall 94.64%, Prec 94.04%, F1 94.31%	Tuning is resource-intensive; the model may not generalize across datasets
Alzaidi et al., [21]	TCAODL-ANA: FastText + ABiGRU + Aquila Optimizer	Simulate d Arabic news dataset	Acc 95.48%, Prec 84.21%, Recall 84.18%, F1 84.17%	Moderate precision/reca ll; computationa lly demanding due to AO
Alroob aea [22]	Preprocessing + Word Embeddings + CNN, LSTM, CNN-LSTM	5,000 Arabic news articles	Acc 93.15% (Hybrid CNN-LSTM)	Small dataset; may not generalize well; performance may vary by dialect or news type

III. Framework for Arabic News Text Classification

The proposed study establishes a framework for Arabic news classification using deep learning models and the existing preprocessing variants supplied with the Ultimate Arabic News Dataset. The dataset representations are used as provided by the original contributor to examine how stop-word removal and stemming influence model performance; the study does not claim to release a newly enhanced corpus. The framework integrates Dynamic N-gram extraction with CNN, LSTM, and GRU architectures, evaluated individually and in hybrid combinations to capture local textual patterns and long-range dependencies. Five learning rates are examined to assess convergence and performance sensitivity. The block diagram of the proposed work is shown in Fig. 1.

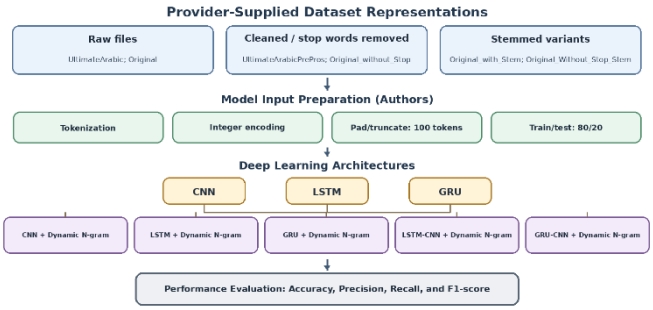


Fig. 1. Block diagram for the proposed study.

A. Linguistic Motivation and Research Hypotheses

Arabic news classification is influenced by high surface-form variation arising from inflection and derivation, clitic attachment, relatively flexible word order, absent diacritics, and lexical ambiguity [9]-[12]. These properties motivate architectures that can combine local contextual evidence with information distributed across a longer sequence. The present implementation uses word-level tokenization; therefore, its convolutional branches operate over neighboring word tokens and do not directly segment or analyze Arabic roots, patterns, prefixes, suffixes, or other within-word morphological units. Morphological variation is addressed only indirectly through learned word embeddings, surrounding context, and comparison of the provider-supplied stemmed and non-stemmed dataset variants.

Kernel sizes 2, 3, and 4 were used as controlled local context scales to detect short token sequences and category-bearing phrases. The recurrent LSTM and GRU components were included to retain evidence across longer portions of an article, which may be useful when semantically related cues are separated by flexible word order or intervening text. The hybrid architectures were consequently designed to test whether combining local phrase extraction with sequential context provides complementary information. These choices are theoretical expectations rather than assumptions of uniform superiority, and the results are interpreted against the following hypotheses.

- H1 (local-context hypothesis): Multi-kernel convolution is expected to improve selected models when category labels are signaled by short neighboring-token patterns; any benefit is expected to depend on the architecture and training configuration rather than occur universally.
- H2 (sequential-context hypothesis): LSTM or GRU layers are expected to be useful when classification requires contextual evidence extending beyond a fixed local convolutional window.
- H3 (complementarity hypothesis): Hybrid convolutional-recurrent models are expected to benefit when local lexical cues and longer sequential evidence contribute jointly to the classification decision.
- H4 (preprocessing-interaction hypothesis): The relative value of each architecture is expected to change across the provider-supplied preprocessing variants because stemming and stop-word removal may reduce lexical sparsity while also removing potentially discriminative lexical or syntactic cues.

B. Data Collection

The data used in this research are from the Ultimate Arabic News Dataset [23], downloaded from Mendeley Data, Version 2 (DOI: 10.17632/jz56k5wxz7.2), published on July 4, 2022, by Ahmed Hashim Al-Dulaimi. The collection contains more than 193,000 single-label Arabic news texts gathered from sources including Al-Arabiya, Al-Youm Al-Sabea (Youm7), Google News, and other news websites. All dataset representations evaluated in this study were used as supplied by the original contributor.

- UltimateArabic: The primary file containing the original Arabic news texts without preprocessing; the texts may include words, numbers, and symbols.
- UltimateArabicPrePros: The provider-preprocessed counterpart in which stop words, non-Arabic words, symbols, and numbers were removed for direct use in Arabic NLP tasks.

The repository also provides two supplementary folders that document the collection process and offer four alternative dataset representations:

- Sample: Two source-specific subsets, Sample_Youm7_Politic and Sample_alarabiya_Sport, illustrate news collected from Youm7 and Al-Arabiya in the Politics and Sport categories, respectively.
- Dataset Versions: Four provider-supplied representations are available: Original (raw text), Original_without_Stop (cleaned text with stop words removed), Original_with_Stem (cleaned and stemmed text), and Original_Without_Stop_Stem (cleaned text with both stop-word removal and stemming).

The corpus is divided into ten single-label categories: Culture, Diverse, Economy, Sport, Politics, Art, Society, Technology, Medical, and Religion. In total, the study evaluates the two primary files and the four existing dataset-version files. The authors did not create or release an independently enhanced version of the corpus; the methodological contribution is the systematic evaluation of

these existing representations. The dataset counts are shown in Table II.

TABLE II. DATASET SAMPLE

Category	Original	Without Stem	Without Stop	Without Stop+ Stem	Ultimate Arabic	UltimateArabicPrePros
Art	3,000	3,000	3,000	3,000	14,390	14,388
Culture	3,000	3,000	3,000	3,000	6,806	6,806
Economy	2,995	2,995	2,995	2,995	26,117	26,038
Medical	3,000	3,000	3,000	3,000	7,294	7,294
Politics	3,000	3,000	3,000	3,000	45,369	45,349
Religion	3,000	3,000	3,000	3,000	6,500	6,500
Sport	3,000	3,000	3,000	3,000	59,998	57,151
Technology	3,000	3,000	3,000	3,000	11,518	11,518
Diverse	3,000	3,000	3,000	3,000	17,203	17,203
Society	0	0	0	0	1,084	1,084
Total	26,995	26,995	26,995	26,995	196,279	193,331

C. Dataset Variants and Model Input Preparation

The corpus-level transformations were not newly introduced by the present authors. The six representations were selected and used as provided in Version 2 of the dataset. The provider-supplied transformations are distinguished below from the subsequent model-input operations. Beyond selecting the appropriate file for each experiment, the authors do not claim additional corpus cleaning, stop-word removal, stemming, or the release of a new processed dataset.

1) Provider-supplied character cleaning: In the cleaned variants, the dataset contributor removed non-Arabic words, symbols, and numbers from the original text.

$$D_clean = Clean(D) \quad (1)$$

2) Provider-supplied stop-word removal: Stop-word removal is already represented in UltimateArabicPrePros, Original_without_Stop, and Original_Without_Stop_Stem; it was not introduced as a new dataset contribution in this study.

$$D_stop = RemoveStopWords(D_clean) \quad (2)$$

3) Provider-supplied stemming: Stemming is already represented in Original_with_Stem and Original_Without_Stop_Stem, as supplied by the dataset contributor.

$$D_stem = Stem(D_clean) \quad (3)$$

4) Provider-supplied combined variant: Original_Without_Stop_Stem combines the contributor's cleaning, stop-word removal, and stemming operations in a single representation.

$$D_both = Stem(RemoveStopWords(D_clean)) \quad (4)$$

5) Tokenization: For model input, each selected text representation was segmented into tokens.

$$T = Tokenize(D_selected) \quad (5)$$

6) Integer sequence encoding: The resulting tokens were mapped to integer sequences so they could be processed by the embedding layer.

$$X = Encode(T) \quad (6)$$

7) Fixed-length sequence preparation: Sequences were standardized to the reported length of 100 tokens by padding shorter sequences and truncating longer sequences.

$$X_100 = PadOrTruncate(X, 100) \quad (7)$$

8) Data partitioning: For each evaluated representation, the prepared samples were divided into the reported 80% training and 20% testing partitions.

$$(D_train, D_test) = Split(D_selected, 80/20) \quad (8)$$

D. Baseline Models

The baseline models provide the basis for assessing and comprehending the efficacy of text classification methods for Arabic news data. The models are suitable for coping with distinctive linguistic patterns in Arabic text, and they also shed light on the performance of classical deep learning architectures.

1) Convolutional Neural Network (CNN): CNNs are commonly employed in natural language processing (NLP) due to their capacity to effectively extract local patterns like n-grams and key phrases from text. The hierarchical structure processes input data hierarchically to ensure that significant features are captured at every level. The structure of the CNN is shown in Fig. 2.

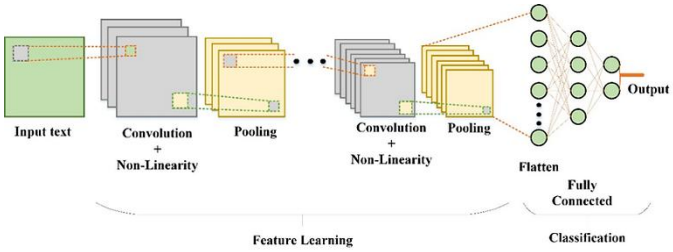


Fig. 2. CNN architecture.

a) Embedding layer: Input sequences of integer-encoded tokens are converted into dense vector representations by this layer. Every token, denoted by t_i, is mapped to an embedding vector $e_i \epsilon R^d$, denoted by d, where d is the embedding dimension. Word semantic links are preserved thanks to the embedding. It is shown in Eq. (9).

$$E = Embedding(T) \quad (9)$$

where, T is the token sequence and E is the embedding matrix.

b) Conv1D layers: To find patterns in fixed-length contexts, one-dimensional convolutional filters are applied over the embeddings. To create feature maps, a filter $w \epsilon R^{k \times d}$ with kernel size k slides over the embedding matrix. It is shown in Eq. (10).

$$h_i = \sigma(w.E_{i:i+k-1} + b) \quad (10)$$

The bias component is denoted by b, the activation function is denoted by σ, and the activation for the i-th position is represented by h_i.

c) Global max pooling: The layer reduces dimensions and retains the most salient characteristics by selecting the maximum activation of each feature map. It is shown in Eq. (11).

$$h_{max} = \max_i(h_i) \quad (11)$$

d) Dense layers: Fully connected layers handle the pooled features and apply softmax activation to support multi-class classification. It is shown in Eq. (12).

$$\hat{y} = Softmax(W.h_{max} + b) \quad (12)$$

e) Rationale: CNNs are well suited to detecting local textual patterns, including short token sequences and category-bearing phrases. In this study, convolution is applied to word-token sequences; consequently, it captures contextual relationships among neighboring words rather than prefixes, suffixes, roots, or patterns within individual words.

2) Long Short-Term Memory (LSTM): LSTM networks are geared to learn long-term dependencies in sequential data by solving the vanishing gradient problem in regular recurrent neural networks (RNNs). This positions LSTMs well to capture contextual and sequential patterns in Arabic text. The LSTM structure is shown in Fig. 3.

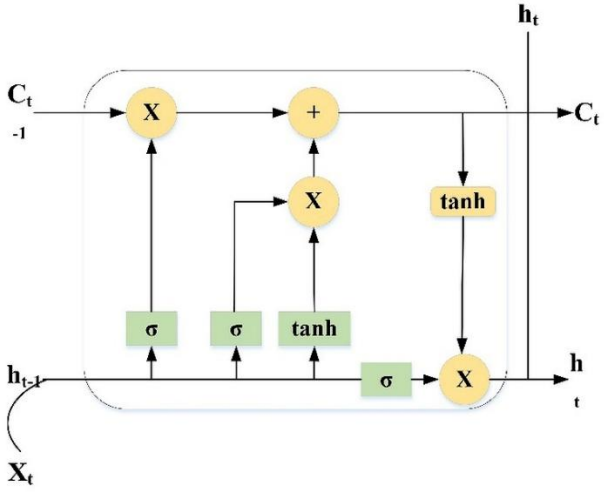


Fig. 3. LSTM structure.

a) Embedding layer: Functions as in CNN, mapping tokens to dense vectors.

b) LSTM layer: The following equations define the model's central component, an LSTM cell at time step t, shown in Eq. (13)-(18).

$$f_t = \sigma(W_f[h_{t-1}, X_t] + b_f) \quad (13)$$

$$i_t = \sigma(W_i[h_{t-1}, X_t] + b_i) \quad (14)$$

$$g_t = tanh(W_g[h_{t-1}, X_t] + b_g) \quad (15)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (16)$$

$$o_t = \sigma(W_o[h_{t-1}, X_t] + b_o) \quad (17)$$

$$h_t = o_t \odot \tanh(c_t) \quad (18)$$

Here, the cell state is represented by c_t, the hidden state by h_t, and the forget, input, and output gates are represented by f_t, i_t, and o_t, respectively.

c) Dense layers: Like CNNs, dense layers project the final output of LSTM to classification probabilities via softmax.

d) Rationale: LSTMs learn long dependencies and word order, which are necessary for realizing complex Arabic sentence structure and syntax.

3) Gated Recurrent Unit (GRU): GRUs are a recurrent architecture that uses update and reset gates to model sequential information with a simpler gating structure than LSTMs. The GRU structure is shown in Fig. 4.

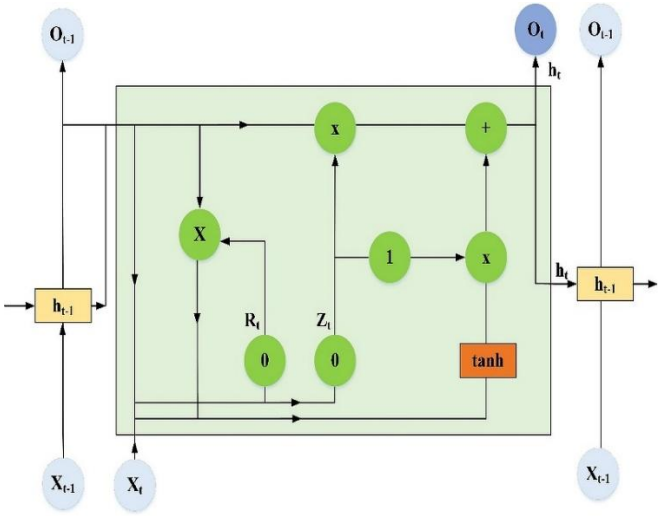


Fig. 4. GRU structure.

a) Embedding layer: Same as in CNN and LSTM.

b) GRU Layer: By merging the input and forget gates into an update gate, GRU streamlines the architecture. The equations at the time step t are shown in Eq. (19)-(21).

$$z_t = \sigma(W_z[h_{t-1}, X_t] + b_z) \quad (19)$$

$$r_t = \sigma(W_r[h_{t-1}, X_t] + b_r) \quad (20)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tanh(W_h[r_t \odot h_{t+1}, x_t] + b_h \quad (21)$$

If h_t is the new hidden state, z_t is the update gate, and r_t This is the reset gate.

c) Dense layers: As in the case of LSTM, the final output is passed to fully connected layers with softmax for classification.

d) Rationale: GRUs provide an alternative recurrent mechanism for capturing contextual and sequential dependencies in Arabic text. In this study, their suitability is assessed through classification metrics only; computational efficiency was not benchmarked.

By using such baseline models, the research comprehensively examines the strengths and limitations of various architectures for Arabic news classification. These models also serve as a benchmark for measuring the performance of sophisticated hybrid solutions.

E. Hybrid and Advanced Models

To further improve classification accuracy, hybrid and advanced versions of the models were used by leveraging the strengths of Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM), Gated Recurrent Units (GRUs), and dynamic n-gram feature extraction methods. These models seek to leverage both local patterns and sequential dependencies, providing an enhanced feature representation. The following is a detailed overview of each hybrid model, its architecture, and the significance of the model with respect to Arabic text classification.

1) Dynamic N-gram + CNN

a) Dynamic N-gram block: The main innovation consists of parallel Conv1D layers that capture bi-, tri-, and four-grams with different kernel sizes k (e.g., 2, 3, 4). The convolution process for an embedding matrix, where n is the sequence length and d is the embedding dimension. It is shown in Eq. (22).

$$h_{i,j} = \sigma(w_j . E_{i:i+k_j-1} + b_j) \quad (22)$$

where, the bias is represented by b_j, the filter is represented by w_j, the activation function by σ, and the kernel size is indicated by σ.

b) Concatenation: The outputs of the Conv1D layers are concatenated to form a feature-rich representation. It is shown in Eq. (23).

$$H_{concat} = [H_1, H_2, H_3] \quad (23)$$

where, the feature map from the j-th kernel is represented by H_j.

c) Global max pooling: Reduces dimensionality by choosing the maximum activation for every feature map. It is shown in Eq. (24).

$$h_{max} = \max_i(H_{concat}) \quad (24)$$

d) Dense layers: Fully connected layers process the pooled features and offer classification probabilities using a softmax layer. Dynamic N-gram + CNN is shown in Fig. 5.

e) Rationale: The use of multiple kernel sizes allows this model to capture different local textual patterns. The present study evaluates the effect of this configuration on classification performance; training speed and computational cost were not measured.

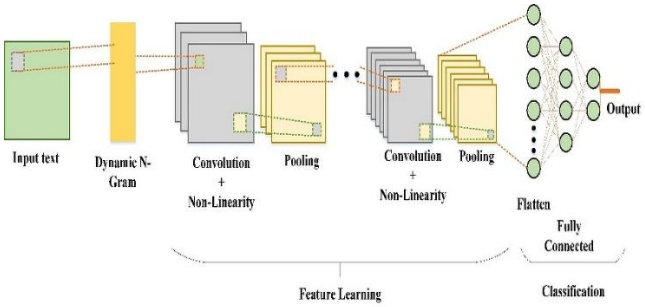


Fig. 5. Dynamic N-gram + CNN.

2) Dynamic N-gram + LSTM

a) Dynamic N-gram block: This block, similar to the above model, captures n-gram patterns using parallel Conv1D layers.

b) LSTM layer: Concatenated n-gram features are fed to an LSTM layer to capture long dependencies and context relationships. The LSTM handles input sequentially with both hidden and cell states. It is shown in Eq. (25).

$$h_t, c_t = LSTM(x_t, h_{t-1}, c_{t-1}) \quad (25)$$

Here, the concealed state is denoted by h_t, the cell state by c_t, and the input at time t by x_t.

c) Dense layers: Final layers project the LSTM output to class probabilities. Dynamic N-gram + LSTM is shown in Fig. 6.

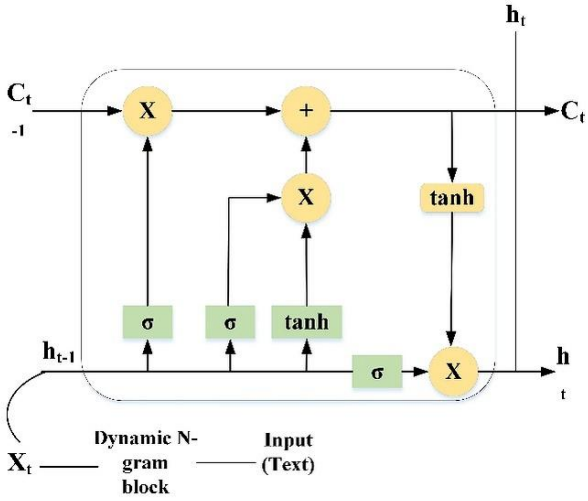


Fig. 6. Dynamic N-gram + LSTM.

d) Rationale: This model brings together the CNN's power in extracting strong local features and the LSTM's capacity for sequential dependency capture, offering a comprehensive solution to Arabic text classification.

3) Dynamic N-gram + GRU

a) Dynamic N-gram block: As defined earlier, this block extracts bi-grams, tri-grams, and more with Conv1D layers.

b) GRU layer: The GRU layer substitutes the LSTM and models sequential relations through update and reset gates. The GRU cell can be described as shown in Eq. (26)-(28).

$$z_t = \sigma(W_z[h_{t-1}, x_t] + b_z) \quad (26)$$

$$r_t = \sigma(W_r[h_{t-1}, x_t] + b_r) \quad (27)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tanh(W_h[r_t \odot h_{t+1}, x_t] + b_h \quad (28)$$

c) Dense layers: Final classification is done by fully connected layers with softmax activation. Dynamic N-gram + GRU is shown in Fig. 7.

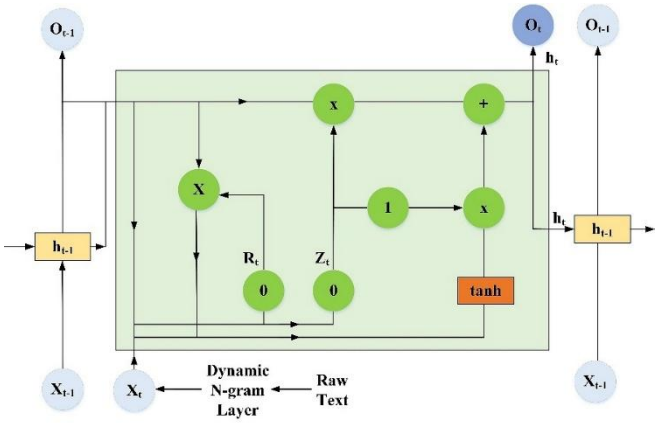


Fig. 7. Dynamic N-gram + GRU.

d) Rationale: The integration of GRU and Dynamic N-gram extraction combines multi-scale local features with recurrent sequence modeling. Its classification behavior is evaluated experimentally, while computational efficiency and suitability for resource-limited deployment are not assessed in this study.

4) Dynamic N-gram + LSTM + CNN

a) Dynamic N-gram block: Parallel Conv1D layers extract varied local patterns, similar to earlier models.

b) LSTM layer: Parameterized with return_sequences=True, the LSTM layer processes n-gram features, extracting sequential dependencies without losing temporal structure for subsequent processing. It is shown in Eq. (29).

$$h_t = LSTM(H_{concat}) \quad (29)$$

c) Additional CNN layer: The LSTM outputs are passed through another Conv1D layer to fine-tune local patterns. It is shown in Eq. (30).

$$h_{refine} = Conv1D(h_t) \quad (30)$$

d) Global max pooling: Combines fine-tuned features and dimension reduction.

e) Dense layers: Fully connected layers finalize the architecture, ending with a final softmax layer for classification. Dynamic N-gram + LSTM + CNN is shown in Fig. 8.

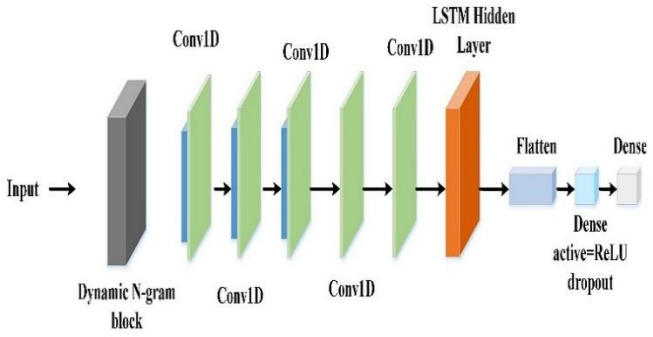


Fig. 8. Dynamic N-gram + LSTM + CNN.

f) Rationale: This hybrid model extracts fine-grained local features using CNNs, formulates their temporal relationships using LSTM, and further refines the resultant patterns with a second CNN layer. This multi-step processing guarantees excellent feature representation and classification performance.

5) Dynamic N-gram + GRU + CNN

a) Dynamic n-gram block: Parallel Conv1D layers: Derive varied local patterns from the input text data, such as in the Dynamic N-gram + LSTM + CNN architecture.

- Output: Integrated local features for sequential processing.

b) GRU layer: Set up with return_sequences=True, the GRU layer processes the n-gram features while retaining temporal relationships for the subsequent layer. It is shown in Eq. (31).

$$h_t = GRU(H_{concat}) \quad (31)$$

c) Additional CNN layer: The GRU outputs are passed through a Conv1D layer to further refine local patterns so that feature extraction at multiple levels is possible. It is shown in Eq. (32).

$$h_{refine} = Conv1D(h_t) \quad (32)$$

d) Global max pooling: Combines refined features and performs dimensionality reduction by taking the most important feature values from each feature map.

e) Dense layers: The aggregated features are processed by fully connected layers, leading to a softmax layer for multi-class classification. Dynamic N-gram + GRU + CNN is shown in Fig. 9.

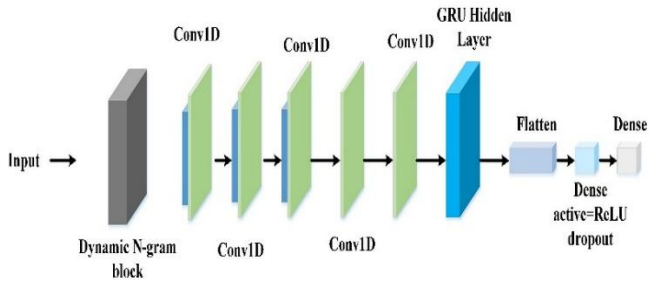


Fig. 9. Dynamic N-gram + GRU + CNN.

This fusion architecture combines CNN-based local feature extraction with GRU-based sequential dependency modeling. The additional CNN layer further refines the temporal representations produced by the GRU, while the Dynamic N-gram block supplies features derived from multiple kernel sizes. Because runtime and resource measurements were not collected, no conclusion is drawn regarding a computational advantage or suitability for resource-constrained deployment.

The advanced and hybrid models are evaluated because Arabic news classification may depend on both local token patterns and longer contextual relations. Combining multi-kernel word-level convolution with sequence modeling tests whether these two forms of evidence are complementary across the supplied preprocessing variants. The architecture does not itself solve Arabic morphology or directly analyze within-word structure (Algorithm 1).

Algorithm 1: Dynamic N-gram + GRU + CNN for Arabic Text Classification

Input: Arabic text dataset D

Output: Classified labels L

Step 1: Preprocessing:

Clean text (remove special characters, stop words).

Generate Dynamic N-grams and encode using embeddings.

Step 2: Dynamic N-gram Block:

Apply parallel $Conv1D$ layers to capture diverse local patterns.

Concatenate outputs to form H_{ngram}

Step 3: GRU Processing:

Pass H_{ngram} to GRU (return_sequences=True) to model sequential dependencies.

Output: H_{gru} CNN Refinement:

Step 4: CNN Refinement

Apply another $Conv1D$ layer on H_{gru} to refine features.

Output: H_{gru}

Step 5: Pooling:

Perform global max pooling to reduce dimensionality.

Result: H_{pool}

Step 6: Classification:

Pass H_{pool} through dense layers and a softmax layer for classification.

Step 7: Training:

Use cross-entropy loss and Adam optimizer with tuned learning rates.

Step 8: Evaluation:

Compute Accuracy, Precision, Recall, and F1-Score.

End

IV. Result and Discussion

The results and discussion section presents a detailed show of the performance of the models in different architectures and learning rates. To guarantee uniform assessment, each of the models trained in TensorFlow/Keras was presented with 10 epochs. The results of the performance were determined in the form of a weighted average to take into account the probable class imbalance. The metrics, such as the percentage value with two decimals, are classified in bordered tables to facilitate comparison and ease of calling. The aim of the evaluation is to determine what the best learning rate is per architecture and how the trade-off between the convergence rate and the performance is. The explanation is to focus on the effectiveness of the various architectures (e.g., CNN, LSTM, and hybrid models) in addressing the complexities of the dataset. Through combining the observations of these critiques, the following section will serve to emphasize the merits, weaknesses, and application to real life of the suggested methodologies applied to Arabic text classification, providing an authoritative basis to work on in the form of future developments. All experiments were conducted on a 64-bit workstation equipped with an AMD Ryzen 5 3600 six-core processor operating at 3.59 GHz and 64 GB of RAM. Model training and evaluation were performed exclusively on the CPU. Table III shows the simulation parameters.

TABLE III. Simulation Parameter

Parameter	Value
Dataset	Ultimate Arabic News Dataset
Number of Classes	10
Training Epochs	10
Batch Size	32
Learning Rates Tested	0.1, 0.01, 0.001, 0.0001, 0.00001
Evaluation Metrics	Accuracy, Precision, Recall, F1-Score
Optimizer	Adam
Activation Function	Softmax
Training/Testing Split	80% / 20%
Framework Used	TensorFlow/Keras
Dropout Rate	0.5
Kernel Sizes (CNN)	2, 3, 4
Sequence Length	100 tokens
Embedding Dimensions	128

A. Performance Metrics

Accuracy: Measures the proportion of correctly classified instances out of total instances. It is shown in Eq. (33).

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (33)$$

Precision: Indicates the proportion of true positives among predicted positives, showing correctness. It is shown in Eq. (34).

$$Precision = \frac{TP}{TP+FP} \quad (34)$$

Recall (Sensitivity): Reflects the proportion of true positives identified among actual positives. It is shown in Eq. (35).

$$Recall = \frac{TP}{TP+FN} \quad (35)$$

TABLE IV. PERFORMANCE METRICS: BASELINE MODELS VS. DYNAMIC N-GRAM ENHANCED MODELS

Model Type	Learning Rate	Baseline Model				Dynamic N-gram Model			
		Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
CNN	1.00E-05	52.45	62.58	52.45	51.42	59.21	66.25	59.21	60.33
	0.0001	89.55	89.65	89.55	89.57	89.07	89.15	89.07	89.09
	0.001	90.18	90.35	90.18	90.23	89	89.32	89	89.06
	0.01	82.92	88.51	82.92	84.32	74.75	87.81	74.75	76.83
	0.1	10.78	1.16	10.78	2.1	10.78	1.16	10.78	2.1
LSTM	1.00E-05	22.12	7.53	22.12	9.58	23.63	20.2	23.63	12.25
	0.0001	28.49	28.21	28.49	23.04	38.75	34.09	38.75	30.07
	0.001	87.76	88.11	87.76	87.82	88.37	88.6	88.37	88.39
	0.01	85.81	85.92	85.81	85.84	43.45	42.18	43.45	41.16
	0.1	10.76	6.95	10.76	2.13	10.78	1.16	10.78	2.1
GRU	1.00E-05	19.15	17.06	19.15	8.33	19.13	11.99	19.13	8.2
	0.0001	22.12	32.8	22.12	17.53	23.19	18.03	23.19	18.36
	0.001	26.32	38.8	26.32	23.43	27.52	39.44	27.52	25.46
	0.01	10.74	1.15	10.74	2.08	10.74	1.15	10.74	2.08
	0.1	10.78	1.16	10.78	2.1	11.06	1.22	11.06	2.2
GRU_CNN	1.00E-05	28.65	30.97	28.65	19.73	28.21	36.91	28.21	21.37
	0.0001	83.29	84.08	83.29	83.3	84.33	85.2	84.33	84.4
	0.001	90.24	90.36	90.24	90.25	81.87	82.85	81.87	81.82
	0.01	10.74	1.15	10.74	2.08	10.74	1.15	10.74	2.08
	0.1	10.78	1.16	10.78	2.1	11.06	1.22	11.06	2.2
LSTM_CNN	1.00E-05	28.08	17.75	28.08	18.04	30.64	26.33	30.64	24.01
	0.0001	84.81	85.23	84.81	84.92	83.76	83.97	83.76	83.82
	0.001	88.37	88.89	88.37	88.46	88.16	88.2	88.16	88.16
	0.01	88.85	89.18	88.85	88.87	81.81	82.23	81.81	81.79
	0.1	10.78	1.16	10.78	2.1	10.78	1.16	10.78	2.1

F1 Score: Balances precision and recall, providing a single performance measure. It is shown in Eq. (36).

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (36)$$

Table IV shows the comparison of baseline models (CNN, LSTM, GRU, and their ensemble) against their respective Dynamic N-gram-enhanced versions. The data distribution ensures even training and testing partitions to allow for fair assessment. Results are reported across different learning rates to gain holistic insights into model performance under various settings.

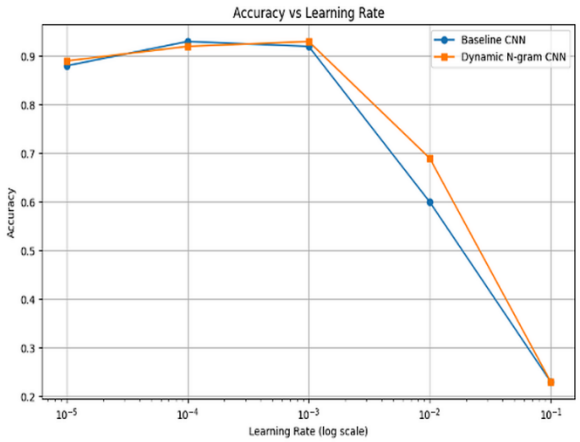


Fig. 10. Accuracy vs learning rate.

Fig. 10. presents the trend in the performance of deep learning models for varying learning rates. It indicates the reaction of baseline and Dynamic N-gram-augmented models to variations in learning rate, comparing the ideal range for successful classification. Dynamic models tend to exhibit enhanced accuracy at smaller learning rates, indicative of improved flexibility in learning fine-grained patterns of Arabic text. This visualization is critical to determine the learning rate that gives the optimal generalization performance and model stability. It also aids in comparing how different architectures (CNN, LSTM, etc.) perform while training, enabling improved model tuning approaches for Arabic NLP tasks.

Table V shows the comparison of baseline and Dynamic N-gram-augmented models over the "Original Without Stem" dataset. The dataset ensures a balanced distribution over categories so that training and testing are strong and unbiased. Performance of models is checked with different learning rates so that their effect on different metrics can be analyzed.

- CNN models outperform others at higher learning rates (e.g., 0.0001 and 0.001).
- Dynamic N-gram improvements demonstrate enhanced performance on lower learning rates for certain models (e.g., GRU, LSTM).
- GRU and GRU-CNN models are less effective compared to their counterparts across all metrics.

TABLE V. PERFORMANCE METRICS COMPARISON: ORIGINAL WITHOUT STEM – BASELINE VS. DYNAMIC N-GRAM MODELS

Model Type	Learning Rate	Baseline Model				Dynamic N-gram Model			
		Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
CNN	1.00E-05	57.92	64.86	57.92	57.24	65.16	66.56	65.16	65.48
	0.0001	89.74	89.84	89.74	89.75	88.5	88.59	88.5	88.53
	0.001	89.2	89.68	89.2	89.29	88.7	89.39	88.7	88.85
	0.01	79.98	87.3	79.98	81.04	75.7	85.21	75.7	77.8
	0.1	10.78	1.16	10.78	2.1	10.8	12.7	10.8	2.14
LSTM	1.00E-05	22.23	8.47	22.23	10.43	23.17	8.83	23.17	10.18
	0.0001	41.62	36.28	41.62	34.06	36.14	42.13	36.14	31.75
	0.001	88.74	89.17	88.74	88.83	89.04	89.43	89.04	89.11
	0.01	84.98	85.21	84.98	84.98	37.67	37.24	37.67	35.28
	0.1	10.78	9.82	10.78	2.2	10.78	1.16	10.78	2.1
GRU	1.00E-05	18.67	21.09	18.67	7.31	18.87	19.93	18.87	7.47
	0.0001	25.47	28	25.47	22.77	22.71	27.37	22.71	20.04
	0.001	28.54	38.98	28.54	27.9	22.15	33.93	22.15	19.66
	0.01	10.74	1.15	10.74	2.08	10.74	1.15	10.74	2.08
	0.1	10.78	1.16	10.78	2.1	11.06	1.22	11.06	2.2
GRU_CNN	1.00E-05	24.19	29.21	24.19	14.62	28.69	31.33	28.69	24.16
	0.0001	78.01	78.95	78.01	77.96	73.7	76.28	73.7	74.14
	0.001	87.37	87.46	87.37	87.33	89.09	89.29	89.09	89.12
	0.01	10.74	1.15	10.74	2.08	10.74	1.15	10.74	2.08
	0.1	10.78	1.16	10.78	2.1	11.06	1.22	11.06	2.2
LSTM_CNN	1.00E-05	26.58	17.87	26.58	18.08	30.52	26.41	30.52	23.1
	0.0001	86.22	86.45	86.22	86.23	88.37	88.47	88.37	88.39
	0.001	87.24	87.76	87.24	87.33	88.48	89.18	88.48	88.63
	0.01	87.39	87.74	87.39	87.37	83.92	84.72	83.92	83.81
	0.1	10.78	1.16	10.78	2.1	10.78	1.16	10.78	2.1

Tables VI and VII summarize performance differences between the baseline and Dynamic N-gram-augmented architectures across the evaluated learning rates.

- Dynamic N-gram models tend to perform better than their baseline equivalents at low learning rates (e.g., 1.00E-05).
- At ideal learning rates (0.0001 and 0.001), basic models (e.g., CNN, GRU-CNN, and LSTM-CNN) provide greater accuracy.
- Stemming and stop-word removal slightly impact recall and precision positively for dynamic N-grams.
- GRU models have lower performance values with and without extensions.
- Dynamic N-gram CNN performs better than CNN using low learning rates (1.00E-05) but lags slightly at mid-range rates (0.0001).
- Dynamic N-gram LSTM always outperforms traditional LSTM in terms of accuracy and recall, especially at 0.0001 and 0.01.
- GRU models like Dynamic N-gram GRU have lower precision in all metrics, indicating minimal usefulness for this set of data.
- LSTM-CNN and Dynamic N-gram LSTM-CNN are equivalent, but Dynamic N-gram LSTM-CNN is marginally better when learning rates are high (0.001).

TABLE VI. PERFORMANCE METRICS COMPARISON: ORIGINAL WITHOUT STOP WORDS – BASELINE VS. DYNAMIC N-GRAM MODELS

Model Type	Learning Rate	Baseline Model				Dynamic N-gram Model			
		Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
CNN	1.00E-05	52.75	55.75	52.75	50.93	57.05	59.95	57.05	56.74
	0.0001	90.02	90.12	90.02	90.04	88.92	89.11	88.92	88.98
	0.001	89.72	90.12	89.72	89.82	88.41	88.84	88.41	88.51
	0.01	84.53	87.47	84.53	84.95	43.32	72.36	43.32	42.77
	0.1	10.78	1.16	10.78	2.1	10.78	1.16	10.78	2.1
LSTM	1.00E-05	23.5	21.45	23.5	14.47	22.99	15.03	22.99	12.9
	0.0001	32.38	32.61	32.38	26.4	34.84	30.4	34.84	27.24
	0.001	86.33	86.35	86.33	86.32	88.02	88.37	88.02	88.07
	0.01	83.02	83.34	83.02	82.99	53.08	55	53.08	52.05
	0.1	10.8	6.95	10.8	2.14	10.78	1.16	10.78	2.1
GRU	1.00E-05	20.91	7.72	20.91	8.81	19.65	9.61	19.65	8.08
	0.0001	23.1	37.43	23.1	19.15	21.45	30.44	21.45	16.01
	0.001	23.36	35.87	23.36	19.82	17.19	20.55	17.19	12.17
	0.01	10.74	1.15	10.74	2.08	10.74	1.15	10.74	2.08
	0.1	10.78	1.16	10.78	2.1	11.06	1.22	11.06	2.2
GRU_CNN	1.00E-05	28.08	36.72	28.08	20.07	24.56	19.6	24.56	16.51
	0.0001	82.18	82.78	82.18	82.3	78.64	79.96	78.64	78.88

	0.001	88.24	88.5	88.24	88.27	85.37	85.72	85.37	85.44
	0.01	10.74	1.15	10.74	2.08	10.74	1.15	10.74	2.08
	0.1	10.78	1.16	10.78	2.1	11.06	1.22	11.06	2.2
LSTM_CNN	1.00E-05	30.15	26.5	30.15	22.71	33.23	24.44	33.23	26.11
	0.0001	85.05	85.2	85.05	85.07	85.35	85.66	85.35	85.43
	0.001	88.42	88.72	88.42	88.46	88.05	88.55	88.05	88.17
	0.01	87.16	87.48	87.16	87.16	83.11	84.16	83.11	83.23
	0.1	10.78	1.16	10.78	2.1	10.78	1.16	10.78	2.1

TABLE VII. PERFORMANCE METRICS COMPARISON: ORIGINAL WITHOUT STOP WORDS AND STEMMING – BASELINE VS. DYNAMIC N-GRAM MODELS

Model Type	Learning Rate	Baseline Model				Dynamic N-gram Model			
		Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
CNN	1.00E-05	61.81	66.71	61.81	61.02	68.81	70.11	68.81	68.79
	0.0001	89.72	89.76	89.72	89.73	89.11	89.25	89.11	89.15
	0.001	89.42	90.01	89.42	89.53	87.85	88.41	87.85	87.95
	0.01	84.53	87.51	84.53	85.19	66.2	87.73	66.2	68.53
	0.1	10.78	1.16	10.78	2.1	10.78	1.16	10.78	2.1
LSTM	1.00E-05	22.12	5.78	22.12	8.97	23.23	21.97	23.23	14.43
	0.0001	33.58	31.94	33.58	28.63	35.62	29.5	35.62	29.14
	0.001	87.68	87.86	87.68	87.73	87.13	87.37	87.13	87.16
	0.01	85.66	85.8	85.66	85.65	63.96	64.93	63.96	62.63
	0.1	10.89	27.56	10.89	2.36	10.78	1.16	10.78	2.1
GRU	1.00E-05	19.73	9.4	19.73	8.16	20.26	6.01	20.26	8.34
	0.0001	24.15	25.71	24.15	20.08	22.86	20.19	22.86	17.93
	0.001	25.1	43.41	25.1	22.17	20.32	26.43	20.32	16.12
	0.01	10.74	1.15	10.74	2.08	10.74	1.15	10.74	2.08
	0.1	10.78	1.16	10.78	2.1	10.78	1.16	10.78	2.1
GRU_CNN	1.00E-05	25.17	30.39	25.17	18.65	27.34	34.11	27.34	22.15
	0.0001	81.14	82.5	81.14	81.37	83.35	83.88	83.35	83.44
	0.001	89.18	89.3	89.18	89.22	84.27	84.53	84.27	84.33
	0.01	10.74	1.15	10.74	2.08	10.74	1.15	10.74	2.08
	0.1	11.06	1.22	11.06	2.2	11.06	1.22	11.06	2.2
LSTM_CNN	1.00E-05	32.27	23.96	32.27	22.71	35.23	40.44	35.23	26.83
	0.0001	84.4	84.59	84.4	84.41	84.89	84.99	84.89	84.87
	0.001	88.11	88.61	88.11	88.21	88.41	88.75	88.41	88.45
	0.01	87.61	88.08	87.61	87.61	83.61	84.83	83.61	83.75
	0.1	10.78	1.16	10.78	2.1	10.78	1.16	10.78	2.1

Table VIII summarizes the performance patterns of the baseline and Dynamic N-gram configurations, particularly for the CNN- and LSTM-based models on this dataset representation.

- Dynamic N-gram improvements systematically enhance the performance of CNN, LSTM, and hybrid models, especially for smaller learning rates.
- GRU architectures do not perform well overall, with small improvements from the Dynamic N-gram method.
- Hybrid configurations such as LSTM-CNN and GRU-CNN are superior with Dynamic N-grams at best learning rates (0.001).

TABLE VIII. PERFORMANCE METRICS COMPARISON: ULTIMATE ARABIC RESULTS – BASELINE VS. DYNAMIC N-GRAM MODELS

Model Type	Learning Rate	Baseline Model				Dynamic N-gram Model			
		Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
CNN	1.00E-05	88.92	88.41	88.92	88.6	91.31	91.4	91.31	91.05
	0.0001	93.33	93.48	93.33	93.34	92.48	92.63	92.48	92.48
	0.001	92.67	92.92	92.67	92.68	92.8	92.9	92.8	92.79
	0.01	80.32	83.57	80.32	80.15	82.97	84.91	82.97	82.8
	0.1	30.51	9.31	30.51	14.26	30.51	9.31	30.51	14.26
LSTM	1.00E-05	39.6	20.92	39.6	24.33	54.73	51.32	54.73	45.53
	0.0001	83.1	82.34	83.1	82.08	93.57	93.56	93.57	93.55
	0.001	93.22	93.22	93.22	93.2	93.22	93.26	93.22	93.24
	0.01	43.27	41.66	43.27	38.78	57.17	54.23	57.17	49.48
	0.1	30.5	9.31	30.5	14.26	30.51	9.31	30.51	14.26
GRU	1.00E-05	33.11	29.89	33.11	27.7	32.28	25.6	32.28	19.39
	0.0001	47.48	51.33	47.48	43.13	40.9	55.86	40.9	38.46
	0.001	35.99	39.31	35.99	28.34	28.64	24.39	28.64	20.77
	0.01	30.51	9.31	30.51	14.26	30.51	9.31	30.51	14.26
	0.1	30.51	9.31	30.51	14.26	30.51	9.31	30.51	14.26
GRU_CNN	1.00E-05	79.34	77.86	79.34	76.78	62.36	61.73	62.36	56.77
	0.0001	93.68	93.6	93.68	93.53	90.06	90.02	90.06	89.9
	0.001	84.24	86.21	84.24	84.21	90.64	90.58	90.64	90.6

	0.01	30.51	9.31	30.51	14.26	30.51	9.31	30.51	14.26
	0.1	30.51	9.31	30.51	14.26	7.49	0.56	7.49	1.04
LSTM_CNN	1.00E-05	87.02	86.19	87.02	86.2	88.67	88.22	88.67	88.39
	0.0001	93.22	93.23	93.22	93.21	92.92	92.9	92.92	92.9
	0.001	92.85	92.95	92.85	92.83	93.32	93.43	93.32	93.33
	0.01	91.63	91.67	91.63	91.6	84.54	84.83	84.54	84.25
	0.1	30.51	9.31	30.51	14.26	30.51	9.31	30.51	14.26

TABLE IX. PERFORMANCE METRICS COMPARISON: ULTIMATE ARABIC PREPOS – BASELINE VS. DYNAMIC N-GRAM MODELS

Model Type	Learning Rate	Baseline Model				Dynamic N-gram Model			
		Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
CNN	1.00E-05	89.4	88.88	89.4	89.1	89.99	89.45	89.99	89.69
	0.0001	93.08	93.08	93.08	93.06	92.2	92.27	92.2	92.17
	0.001	92.35	92.5	92.35	92.32	92.66	92.8	92.66	92.65
	0.01	60.36	63.71	60.36	56.9	69.48	65.03	69.48	65.73
	0.1	23.68	5.61	23.68	9.07	23.68	5.61	23.68	9.07
LSTM	1.00E-05	38.96	16.85	38.96	23.52	37.18	21.64	37.18	21.63
	0.0001	87.19	87.13	87.19	86.65	92.84	92.84	92.84	92.83
	0.001	92.84	92.87	92.84	92.84	92.93	92.99	92.93	92.95
	0.01	87.28	87.17	87.28	87.17	63.43	60.5	63.43	58.77
	0.1	23.68	11.52	23.68	9.07	23.68	5.61	23.68	9.07
GRU	1.00E-05	33.54	34.4	33.54	21.38	33.15	16.79	33.15	20.95
	0.0001	42.6	51.11	42.6	36.75	39.35	40.87	39.35	32.84
	0.001	33.03	36.05	33.03	24.58	27.69	33.46	27.69	19.78
	0.01	29.56	8.74	29.56	13.49	29.56	8.74	29.56	13.49
	0.1	23.68	5.61	23.68	9.07	23.68	5.61	23.68	9.07
GRU_CNN	1.00E-05	78.15	75.29	78.15	75.44	70.26	70.06	70.26	67.25
	0.0001	92.82	92.81	92.82	92.8	89.84	89.9	89.84	89.71
	0.001	84.68	85.18	84.68	84.71	90.44	90.49	90.44	90.39
	0.01	29.56	8.74	29.56	13.49	29.56	8.74	29.56	13.49
	0.1	23.68	5.61	23.68	9.07	7.59	0.58	7.59	1.07
LSTM_CNN	1.00E-05	75.99	74.68	75.99	74.4	86.13	85.93	86.13	85.9
	0.0001	92.48	92.48	92.48	92.47	92.3	92.31	92.3	92.27
	0.001	92.54	92.68	92.54	92.54	93.06	93.16	93.06	93.09
	0.01	29.56	8.74	29.56	13.49	84.46	84.99	84.46	84.19
	0.1	23.68	5.61	23.68	9.07	23.68	5.61	23.68	9.07

Table IX emphasizes the use of Dynamic N-gram methods in tuning model performance on Arabic datasets for various architectures.

Fig. 11 compares the baseline and Dynamic N-gram accuracy values shown for the evaluated architectures. The differences are architecture-dependent rather than uniformly favorable to the enhanced variants: the plotted Dynamic N-gram configurations improve CNN, LSTM, and LSTM-CNN, whereas the baseline GRU and GRU-CNN remain competitive or achieve higher values. The figure therefore illustrates variation across architectures and should not be interpreted as evidence of a consistent gain for every model.

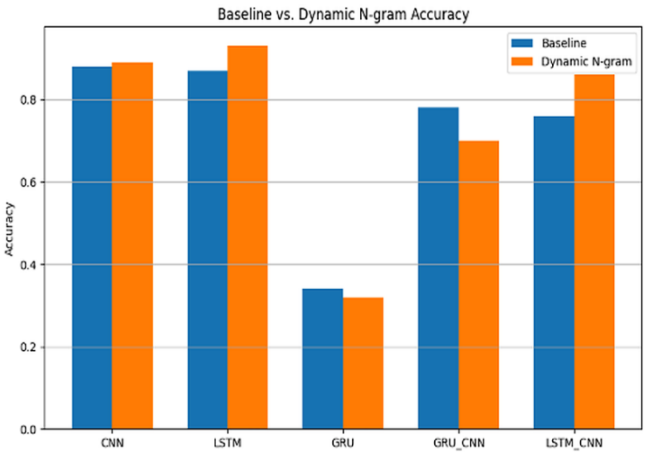


Fig. 11. Baseline vs. dynamic N-gram accuracy.

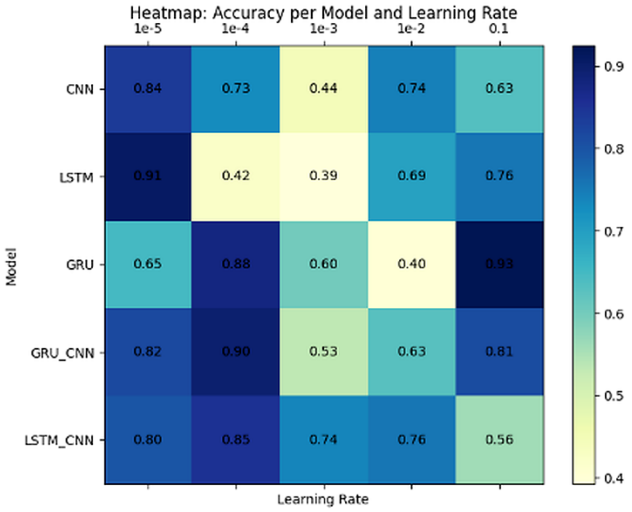


Fig. 12. Accuracy per model and learning rate.

Fig. 12 provides a matrix visualization of how model performance varies at various learning rates. Every cell plots a measure of performance (e.g., accuracy), with color saturation indicating the intensity of the value. The visual graph

immediately reveals good combinations of model types and learning rates. For example, lighter cells might reflect better accuracy, enabling researchers to identify optimal hyperparameter settings. It further allows cross-comparison between baseline and Dynamic N-gram models in a concise manner. By condensing large amounts of experimental data graphically, the heatmap facilitates making sensible decisions regarding model choice and training settings.

Fig. 13 presents the CNN accuracy, precision, recall, and F1-score in a compact multi-metric view. The overlap between the two profiles indicates that the differences are modest and specific to the selected configuration. The chart is used to inspect metric balance within this experiment, not to infer universal robustness or superiority beyond the evaluated dataset and settings.

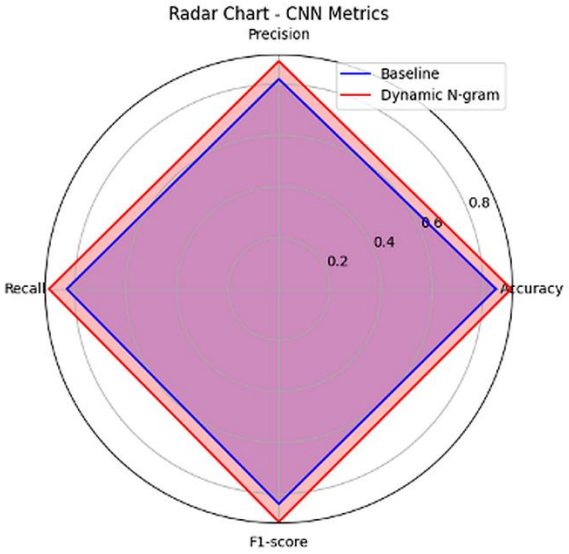


Fig. 13. Radar chart-CNN metrics.

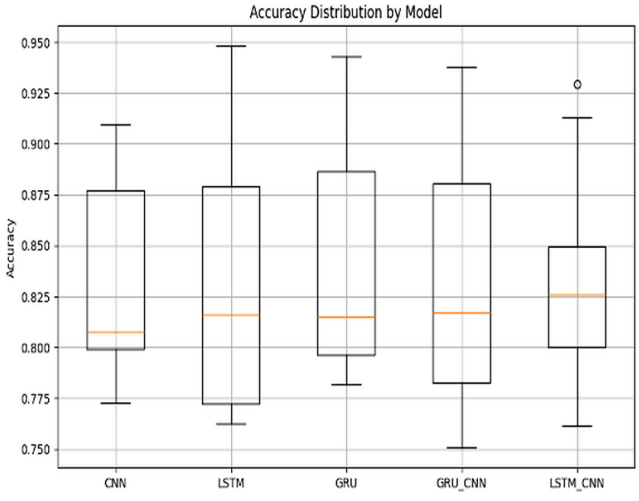


Fig. 14. Accuracy distribution by model.

Fig. 14 plots accuracy score distributions for every model, displaying central tendency and variability with multiple runs. Each box plots the interquartile range, with whiskers and outliers representing score spread. The plot is particularly useful for evaluating model stability and trustworthiness under various initializations or data splits. Models with more compact boxes and larger medians are perceived to be more consistent. It allows researchers to not only find the highest-performing model but also the most reliable one. This matters significantly in actual NLP tasks where consistency in performance over runs is as vital as getting high accuracy.

B. Discussion

The results are interpreted in relation to the four linguistic hypotheses stated in Section III. Because the architecture operates on word tokens, the analysis concerns local token context, longer sequential evidence, and interaction with preprocessing; it does not claim direct modeling of Arabic roots or affixes. The overall pattern is mixed: selected configurations are consistent with the hypotheses, whereas other settings show that the proposed components do not provide uniform gains.

- H1 - Local-context hypothesis: The multi-kernel CNN configuration improved accuracy under selected low-learning-rate conditions, but the baseline CNN matched or exceeded it in several mid-range conditions. Thus, short multi-scale token context can be useful, but fixed windows of two to four words are not consistently superior across all settings.
- H2 - Sequential-context hypothesis: Dynamic N-gram LSTM achieved the strongest result among the enhanced models under its best configuration, which is consistent with the expectation that longer sequential context can complement local features. However, the advantage was not maintained at every learning rate, and both GRU configurations performed substantially below the strongest CNN- and LSTM-based settings. The evidence is therefore partial and architecture-dependent.
- H3 - Complementarity hypothesis: The hybrid results were also mixed. The enhanced LSTM-CNN reached a strong value at a selected learning rate, whereas baseline hybrid models matched or exceeded their enhanced counterparts in other configurations. These findings suggest that local and sequential representations can be complementary, but they do not establish a universal hybrid advantage or isolate the contribution of each component without an ablation study.
- H4 - Preprocessing-interaction hypothesis and learning-rate sensitivity: Relative model rankings changed across dataset representations and learning rates. This pattern is consistent with the possibility that stemming and stop-word removal alter the lexical and syntactic cues available to each architecture, but the present design cannot determine which specific linguistic information was retained or removed. Learning rate was also a major determinant of convergence: the largest rate generally produced unstable or near-chance results, while moderate and smaller rates yielded the strongest values for different architectures.

Taken together, the results support conditional rather than universal versions of the proposed hypotheses. Multi-scale local context, recurrent sequence modeling, and their combination can provide competitive benefits in selected settings, but the observed gains remain configuration-dependent. Because the implementation is word-based and no kernel-specific or component-level ablation was conducted, the results should not be interpreted as evidence of direct morphological analysis or as causal proof that any single architectural component produced the improvement.

Table X provides a contextual rather than a direct benchmark comparison. The reported values were obtained using different corpora, numbers of classes, preprocessing pipelines, data partitions, task definitions, and evaluation protocols. Therefore, a numerical ranking across these studies would be methodologically inappropriate. The present results are consequently described as competitive within the Ultimate Arabic News Dataset setting. A controlled comparison that reimplements external methods on the same corpus and split is required before any state-of-the-art claim can be supported.

TABLE X. CROSS-STUDY PERFORMANCE CONTEXT FOR ARABIC NEWS CLASSIFICATION

Study / Method	Dataset / Setting	Reported Accuracy / Comparison Note
Hossain et al. [16] — ABTM	Arabic news corpus; transformer with TF-IDF/BOW	97.69%; different corpus, preprocessing, and evaluation protocol
Zamzami et al. [17] — ML/DL	Multiple country-of-origin Arabic news datasets	Up to 94%; different target task and corpora
Imad et al. [18] — CNN	CNN, BBC, and OSAC datasets	94.47%, 90.97%, and >98.8%; dataset-dependent and not directly comparable
Jamaleddyn et al. [20] — tuned ML	CNN Arabic dataset	95.16%; different dataset and tuning protocol
Alzaidi et al. [21] — TCAODL-ANA	Simulated Arabic news dataset	95.48%; different dataset and optimization procedure
Alroobaea [22] — CNN-LSTM	5,000 Arabic news articles in six categories	93.15%; different dataset size, classes, and split
Present study — Dynamic N-gram LSTM	Ultimate Arabic News Dataset; 10 classes; 80/20 split	93.57%; competitive within the present protocol; no direct state-of-the-art claim

C. Limitations

Computational efficiency was not evaluated in the present experiments. Consequently, the study does not report parameter count, FLOPs, training time, inference latency or throughput, model size, or peak CPU memory consumption. The reported comparisons therefore support conclusions about classification performance only and cannot establish whether the observed accuracy differences justify any additional computational overhead in real-world deployment settings. In addition, the word-level tokenization used in the experiments does not directly model Arabic roots, patterns, clitics, or other within-word morphological structure. The linguistic explanations are therefore hypothesis-based interpretations of the architecture and observed performance patterns rather than causal evidence, and kernel-specific or component-level ablation experiments were not conducted.

V. CONCLUSION AND FUTURE WORKS

The study examined the usefulness of applying Dynamic N-gram augmentation to CNN, LSTM, GRU, and hybrid deep learning models for Arabic text classification. The findings showed that multi-kernel word-level context can improve selected architectures under particular learning-rate and preprocessing settings, while the magnitude and direction of the difference vary by model. Among the tested enhanced models, Dynamic N-gram LSTM and Dynamic N-gram LSTM-CNN achieved 93.57% and 93.32% accuracy, respectively, under their best configurations. Interpreted against the stated hypotheses, the results provide conditional support for combining short token-context features with longer sequential evidence, but they do not demonstrate uniform superiority or direct modeling of Arabic morphology. The regular GRU performed less effectively than other architectures, and its enhanced variant produced only limited gains. These findings demonstrate competitive performance within the present dataset and evaluation protocol. They should not be interpreted as state-of-the-art superiority because the studies reviewed in Section II used different corpora, class configurations, preprocessing pipelines, and experimental splits. Future work should reproduce representative external baselines, including transformer-based methods, on the same data partition and evaluation protocol. It should also evaluate character-level and subword tokenization, morphology-aware representations, and controlled ablations comparing individual kernel sizes 2, 3, and 4 with their combined configuration. Such experiments would test the linguistic hypotheses more directly and distinguish local-context effects from model-capacity effects. Future work should additionally conduct controlled computational profiling on common hardware by reporting parameter count, FLOPs, training time, inference latency and throughput, model size, and peak CPU memory consumption for each baseline and enhanced architecture. Such profiling is necessary to quantify the accuracy-efficiency trade-off and assess real-world deployment feasibility.

REFERENCES

[1] Z. Alyafeai, M. S. Al-shaibani, M. Ghaleb, and I. Ahmad, “Evaluating various tokenizers for Arabic text classification,” Neural Process. Lett., vol. 55, no. 3, pp. 2911–2933, 2023.

[2] H. El Rifai, L. Al Qadi, and A. Elnagar, “Arabic text classification: The need for multi-labeling systems,” Neural Comput. Appl., vol. 34, no. 2, pp. 1135–1159, 2022.

[3] M. Bounhas, B. Elayeb, A. Chouigui, A. Hussain, and E. Cambria, “Arabic text classification based on analogical proportions,” Expert Syst., vol. 41, no. 10, p. e13609, 2024.

[4] A. Y. Muaad et al., “An effective approach for Arabic document classification using machine learning,” Glob. Transit. Proc., vol. 3, no. 1, pp. 267–271, 2022.

[5] T. Sabri, S. Bahassine, O. El Beggar, and M. Kissi, “An improved Arabic text classification method using word embedding,” Int. J. Electr. Comput. Eng., vol. 14, no. 1, 2024.

[6] A. Shehata, M. N. Al-Suqri, N. E. Elshaiekh, F. Hamad, Y. N. AlHusaini, and A. Mahfouz, “ArabFake: A multitask deep learning framework for Arabic fake news detection, categorization, and risk prediction,” IEEE Access, 2024.

[7] H. Himdi, G. Weir, F. Assiri, and H. Al-Barhamtoshy, “Arabic fake news detection based on textual analysis,” Arab. J. Sci. Eng., vol. 47, no. 8, pp. 10453–10469, 2022.

[8] F. Fkih, M. Alsuhaibani, D. Rhouma, and A. M. Qamar, “Novel machine learning-based approach for Arabic text classification using

stylistic and semantic features," Comput. Mater. Contin., vol. 75, no. 3, 2023.

[9] O. Hamed and N. Zaidkilani, "Arabic topic classification corpus of the Nakba short stories," in Proc. 1st Int. Workshop Nakba Narratives as Language Resources, 2025, pp. 48–55.

[10] L. H. Baniata and S. Kang, "Transformer text classification model for Arabic dialects that utilizes inductive transfer," Mathematics, vol. 11, no. 24, p. 4960, 2023.

[11] M. Azzeh, A. Qusef, and O. Alabboushi, "Arabic fake news detection in social media context using word embeddings and pre-trained transformers," Arab. J. Sci. Eng., vol. 50, no. 2, pp. 923–936, 2025.

[12] D. Refai, S. Abu-Soud, and M. J. Abdel-Rahman, "Data augmentation using transformers and similarity measures for improving Arabic text classification," IEEE Access, vol. 11, pp. 132516–132531, 2023.

[13] A. M. E. Koshiry, E. H. I. Eliwa, T. Abd El-Hafeez, and A. Omar, "Arabic toxic tweet classification: Leveraging the AraBERT model," Big Data Cogn. Comput., vol. 7, no. 4, p. 170, 2023.

[14] H. A. G. Elsayed, S. Chaffar, S. B. Belhaouari, and H. Raissouli, "A two-level deep learning approach for emotion recognition in Arabic news headlines," Int. J. Comput. Appl., vol. 44, no. 7, pp. 604–613, 2022.

[15] S. M. Alzanin, A. M. Azmi, and H. A. Aboalsamh, "Short text classification for Arabic social media tweets," J. King Saud Univ. Comput. Inf. Sci., vol. 34, no. 9, pp. 6595–6604, 2022.

[16] M. M. Hossain, M. S. Hossain, M. Safran, S. Alfarhood, M. Alfarhood, and M. Mridha, "A hybrid attention-based transformer model for Arabic news classification using text embedding and deep learning," IEEE Access, 2024.

[17] N. Zamzami, H. Himdi, and S. F. Sabbeh, "Arabic news classification based on the country of origin using machine learning and deep learning techniques," Appl. Sci., vol. 13, no. 12, p. 7074, 2023.

[18] J. Imad, A. Rachid, and B. Mohamed, "Automated Arabic news classification using the convolutional neural network," Int. J. Electr. Eng. Inform., vol. 15, no. 2, 2023.

[19] N. A. Othman, D. S. Elzanfaly, and M. M. M. Elhawary, "Arabic fake news detection using deep learning," IEEE Access, 2024.

[20] I. Jamaleddyn, M. Biniz, and others, "An improved approach to Arabic news classification based on hyperparameter tuning of machine learning algorithms," J. Eng. Res., vol. 11, no. 2, p. 100061, 2023.

[21] M. S. A. Alzaidi et al., "Enhanced automated text categorization via Aquila optimizer with deep learning for Arabic news articles," Ain Shams Eng. J., vol. 16, no. 1, p. 103189, 2025.

[22] R. Alroobaea, "An empirical deep learning approach for Arabic news classification," Int. J. Adv. Comput. Sci. Appl., vol. 14, no. 6, 2023.

[23] A. H. Al-Dulaimi, "Ultimate Arabic News Dataset," vol. 2, Jul. 2022, doi: 10.17632/jz56k5wxz7.2.

8C SPS Model: A Seller Product Selection Model for Home-Based Business Sellers Using a Multi-Criteria Decision-Making Approach

Selvia Lorena Br Ginting[1]*, Zulaiha Ali Othman[2]
Center For Artificial Intelligence Technology (CAIT), Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi 43600, Malaysia[1, 2]
Computer System Study Program, Faculty of Engineering and Computer Science, Universitas Komputer Indonesia, Bandung, Indonesia[1]

***Abstract*—Home-based businesses are becoming a necessary tool for sellers of home-based businesses to determine which products they might be able to sell well. Otherwise, most of the sellers still choose products by trial and error, which may bring low demand as well as stock that doesn't sell, resulting in wasted capital. This study proposes an extended Seller Product Selection model from the seller perspective by enlarging the five-criterion benchmark customer-oriented model into eight criteria: Rating, Price, Sold, Discount, Response Chat, Repurchase, Favorite, and Followers. The data utilized in this study are records of food and beverage products from Shopee Indonesia, amounting to 150,000 products. The criteria relevance was tested with Forward Stepwise Multiple Linear Regression, which is additionally supported by the questionnaire given to 30 marketplace sellers online. Regression results for Sold are presented, indicating that all selected criteria were significantly related to the final model, achieving R = 0.953, R Square = 0.908, Adjusted R Squared = 0.908 with a Durbin-Watson of 1.863 (P<.0005) as shown in the table… The ANOVA result also indicates that the total model is statistically significant, F = 206961.417, p < 0.001. Repurchase had the highest statistical correlation with Sold (β = 0.758) among predictors, followed by Favorite (β = 0.299). Questionnaire results illustrate the following: Sold was ranked number one with a mean rank of 2.40 and Borda Weight of 0.183; Repurchase followed closely in position two with a mean rank of 3.63 and Borda Weight of 0.149. In order to validate whether dyads were able to agree on their RfD, Kendall's coefficient of concordance was obtained as follows: W = 0.236 (P < 0001), suggesting a moderate level of agreement among respondents. We conducted product ranking experiments for MOORA, SMART, PSI and TOPSIS under different combinations of criteria and weight assignments based on AHP. The results indicate that the eight-criterion Seller Product Selection model achieves substantially improved ranking stability and consistency compared with the five-state benchmark model presented in the relevant literature. While considering the eight-criterion model, MOORA performed the best in terms of obtaining the lowest MAD value (1,093), MSE value (4,969,891), and the highest Spearman's Rank Correlation value (0.99868). The results demonstrate that the proposed eight-criterion Seller Product Selection model can be an actionable decision aid to guide home-based business sellers in selecting products that have a better potential in the marketplace using MOORA.**

***Keywords*—Seller product selection; home-based business; online marketplace; multi-criteria decision-making; multi-objective optimization based on ratio analysis; analytic hierarchy process; product ranking**

I. INTRODUCTION

Home-based businesses (HBBs) have become an increasingly important source of household and supplementary income, particularly for people seeking flexible and low-cost business opportunities that can be run from home. Besides supporting the welfare of families, HBBs also contribute towards employment generation, local economic activity, resource mobilization, and national economic development [1-6]. The rapid growth of digital marketplaces and e-commerce platforms has facilitated many HBB sellers to have access to wider markets and an extensive range of products and services online [1] [7] [8-13]. Among these, food and beverage products are one of the most common and promising categories due to their continuous demand and large consumer base [14]. However, despite these opportunities, HBB sellers face many serious challenges in sustaining their businesses, especially in highly competitive online environments [15]. One of the main problems is deciding what the best product to sell is. Many sellers, especially those new to business, choose products without researching the market well. This leads to low demand, unsold products, wasted money, and repeated trial-and-error decisions. These conditions not only decrease profitability but also endanger business survival and long-term sustainability [11, 16-18]. Therefore, the HBB sellers need practical support to identify product ideas with strong market potential, so that they can offer products that are more likely to sell, generate profit, and strengthen their competitiveness against other sellers. This assistance to HBB sellers is not only important for their own welfare, but also for the broader economy, as good and sustainable HBB performance can bring about wider social and economic benefits. In this context, providing a ranked list of promising products can serve as a useful decision support mechanism for HBB sellers in determining which products should be prioritized for sale. To achieve this objective, Multi-Criteria Decision-Making (MCDM) is considered an appropriate approach because it enables product alternatives to be systematically evaluated and ranked based on multiple relevant criteria, thereby helping sellers make more informed, rational, and competitive product selection decisions [18-25].

*Corresponding author.

Previous research by Okfalisa et al (2021) has addressed a related marketplace decision problem by applying the MOORA algorithm in a decision support system to recommend online shops based on five criteria, namely price, rating, discount, sold, and response chat, and showed that these criteria could be used to generate ranked recommendations effectively [23]. However, that study was mainly designed from the customer perspective and was limited to five evaluation criteria and a specific product context [23]. Building on this foundation, the present study extends the evaluation framework to eight criteria to better capture the needs of seller product selection, particularly for home-based business sellers who require broader market signals in selecting products with stronger selling potential [26, 27]. Moreover, the problem is evaluated with several MCDM algorithms, such as MOORA, SMART, PSI, and TOPSIS, to prove the robustness of product ranking under the extended criteria set. This extension aims to provide a comprehensive and practical ranking approach to identify the best products to sell in a competitive online marketplace.

The rest of this study is organized as follows. Section II reviews the literature on product selection, seller-related marketplace criteria, statistical analysis, and measures of robustness of rankings. Section III presents the proposed 8C SPS model and the dataset collection and preprocessing, questionnaire analysis, regression procedure, AHP weighting, MCDM algorithms, and evaluation measures. Section IV presents and discusses the results of the questionnaire, regression, product-ranking, criterion-removal and robustness. Finally, Section V presents a summary of the main findings, practical implications, limitations and directions for future research.

II. Literature Review

A. Product Selection

Product selection is another decision-making issue extensively examined by many scholars in diverse research contexts. Examples range from online shopping, supplier selection, material selection, transportation, business studies, to various sales applications. The ubiquity of applications highlights that the issue of product selection is relevant for both the business and industrial communities. As summarized in Table I, previous research has addressed product selection in different contexts by using various datasets, analytical approaches, and evaluation methods to identify the most suitable alternatives. These studies show that product selection is not limited to choosing a single item, but also involves determining the most appropriate alternative among several competing options under different decision environments.

TABLE I. Summary of Existing Selection-Based Decision Studies, Criteria, Algorithms, and Evaluation Measures

Author, Year, Ref	Selection Problem	Criteria								Dataset	Algorithms					Evaluation Measures
		Rating	Price	Discount	Sold	Response Chat	Repurchase	Favorite	Followers		MOORA	SMART	PSI	TOPSIS	OTHERS	
Okfalisa *et al.*, 2021 [23]	Best online shop for shirts	✓	✓	✓	✓	✓	-	-	-	Shirt products	✓	-	-	-	-	CM
Liu *et al.*, 2021, [28]	E-commerce selection	✓	-	-	-	-	✓	-	-	Feedback from buyers	-	-	-	-	NE	RT
Jain *et al.*, 2018, [29]	Automotive Supplier Selection	-	✓	-	-	-	-	-	-	India's automotive company	-	-	-	✓	AHP	SA
Memari *et al.*, 2019, [30]	Automotive Supplier Selection	✓	✓	-	-	-	-	-	-	Automotive parts in Iran	-	-	-	✓	-	SA
Jain & Upadhyay, 2020 [31]	Iron and steel supplier Selection	✓	✓	-	-	-	-	-	-	Iron and steel industry of India	-	-	-	✓	AHP	SPI
Emovon & Oghenenyerovwho, 2020 [19]	Various kinds of material selection	-	✓	-	-	-	-	-	-	Material criteria data	✓	-	✓	✓	AHP, VIKOR, PROMETHEE, ELECTRE	SA
Ulutas *et al.*, 2021 [32]	Transportation company selection	✓	✓	-	-	-	-	-	-	Transportation capacity	✓	-	✓	-	PIPRECIA	SA
Yadav *et al.*, 2019, [33]	Material selection	-	✓	-	-	-	-	-	-	Hybrid aluminium composite	-	-	✓	✓	-	CI
Amin *et al.*, 2021, [34]	Baby cream selection	✓	✓	-	-	-	-	-	-	Baby cream products	-	-	✓	✓	-	-
Cabeza-Ramírez *et al.*, 2022, [35]	Selection of fashion products	-	-	-	-	-	-	✓	✓	Influencer followers on social media	-	-	-	-	PLS-SEM	CA, R, Q

Liu *et al*., 2018, [36]	Product supplier selection	-	✓	-	-	✓	✓	-	-	Green fresh product supplier	-	-	-	✓	F-MULTIMOORA	SA
Abdel-Basset *et al*., 2019, [37]	Medical service efficiency	✓	-	-	-	-	-	-	-	Smart medical device data	✓	-	-	✓	AHP	SC
Chawla & Singari, 2021 [38]	Material selection	✓	-	-	-	-	-	-	-	Six aluminum alloys	✓	-	-	✓	AHP	ARP
Raju *et al*., 2020, [22]	Ranking of Al-CSA composite	✓	-	-	-	-	-	-	-	Aluminum-Coconut shell ash (Al-CSA) composite	✓	-	-	✓	AHP	_
Dorfeshan *et al*., 2018 [39]	Selecting the project-critical path	✓	✓	-	-	-	-	-	-	Aircraft Component Development Project	-	-	-	-	MULTIMOORA & TPOP	SA & CoA
Ulutaş *et al.,* 2021 [40]	Textile Supplier Selection	✓	✓	✓	-	-	-	-	-	Suppliers from three companies	✓	-	-	-	MULTIMOOSRAL	CoA
Çakır, 2016, [41]	Machine selection	✓	✓	-	-	-	-	-	-	Continuous fluid bed tea dryer	-	✓	-	-	F-SMART	LSDM
Bera *et al*., 2019 [42]	Manufacturing supplier selection	✓	✓	✓	-	-	-	-	-	Supplier data	✓	-	-	✓	IT2F TOPSIS, & IT2F MOORA	SA
Poongavanam *et al*., 2021 [21]	Best refrigerants selection	-	✓	-	-	-	-	-	-	14 Types of refrigerant	✓	-	-	✓	EDAS	SA
Liu *et al*., 2020, [43]	Car model selection	-	-	-	-	-	-	✓	-	car model data	-	-	-	-	SBDP	CS
Petcharat & Leelasantitham, 2021 [44]	Commerce selection	-	✓	-	✓	-	✓	✓	-	Online shop user respondents	-	-	-	-	SEM	CA
Kutela et al., 2022, [45]	Influence of language use	-	-	-	✓	-	-	✓	✓	Tweets from Tanzanian airlines	-	-	-	-	NBR	-
Wong & Wei, 2023 [46]	Influence of users and influencers on sales		-	-	✓	-	-	✓	✓	Followers of beauty influencers in China	-	-	-	-	PLS - SEM	MA
Farivar et al., 2022 [47]	The problem with followers and influencers	✓	-	-	-	-	-	-	✓	Instagram influencer followers in North America	-	-	-	-	SEM	CFA
Jin & Youn, 2022 [48]	Selection of sales performance	-	-	-	✓	-	-	✓	✓	Social media users in America	-	-	-	-	PLS-SEM	CA & CR

Note Algorithm: NE =Nash Equilibrium, AHP = Analytical Hierarchy Process, ELECTRE = Elimination and Choice Expressing Reality, QFD = Quality Function Deployment, ALFM = aspect-aware latent factor model, E = Entropy, AMB = Ambidextrous, PIPRECIA = Pragmatic Idealized Process for Risk Evaluation, Combined Influence Analysis, F = Fuzzy, IT2F = Interval Type-2 Fuzzy, EDAS = Evaluation based on Distance from Average Solution , PROMETHEE = Preference Ranking Organization Method for Enrichment of Evaluations, PLS = Partial Least Squares, SEM = Structural Equation. Modeling, TPOP = Technique of Precise Order Preference, VIKOR = VIšekriterijumsko Kompromisno Rangiranje(Multi-criteria Compromise Ranking), SBDP = Sparse Biterm-based Dirichlet process, NBR = Negative Binomial Regression.

Note Measurements : CM = Confusion Matrix, SA = Sensitivity Analysis, RT = Robustness Test, RMSE = Root Mean Square Error, T-T =T-Test, SPI = Sustainability Performance Index, CI = Closeness Index, CA = Cronbach's Alpha, R = R-squared, Q = Q-squared, CR = Composite Reliability, CS = Coherence Score, SC = Spearman correlation, ARP = Approach Reference Point, CoA = Comparative Analysis, LSDM = Least Squares Distance Method, CFA = Confirmatory Factor Analysis, MA = Mediation Analysis..

1) Product selection in previous studies: In the online marketplace context, product selection has been examined to support better choices among available alternatives and to reduce uncertainty in decision-making. For example, Okfalisa et al. [23] explored online shop recommendations for shirt products, while Liu et al. [28] investigated e-commerce selection using buyer-related information. In other contexts, product selection has been extended to supplier selection in the automotive, iron and steel, textile, and manufacturing sectors [29]-[31], [40], [42], as well as to material and component selection in engineering applications [19], [33], [22, 38, 39], [21]. Lately, several studies have explored the product selection decision problem, which is no longer confined to business and engineering disciplines, but has expanded to digital commerce, social media, and human consumer behaviors with different research focuses.

For example, recent research investigates product selection issues in fashion e-Commerce, modeling of cars, digital commerce, as well as studies on influencer marketing impact on product sales [35], [43-48]. The findings of this research literature show that digital experience, social factors, internet presentation, and market behaviors dominate product-related decision-making. However, to the best of the author's knowledge, the body of related work has remained fragmented

due to the varied nature and focal points of disparate domains investigated.

2) MCDM: From a methodology viewpoint, in addition to other decision-making techniques applied previously to address the problem under study (selection of products), it can be noted that the use of the TOPSIS method is among the most prevalent in the area of decision-making, as it addresses issues related to the selection of suppliers, materials, and evaluation in product-related tasks [19, 29-31], [33], [34], [22, 36-38], [42], [21]. This wide application shows its recognition due to the high suitability for evaluating solutions based on their respective proximity to ideal solutions. Also prevalent in the academic context is the application of the MOORA method in research studies in relation to the recommendations for online stores, the selection of suppliers, material, and evaluation of service [23], [19], [32], [22, 37, 38], [40], [42], [21]. This can also indicate a high degree of applicability in a variety of problems involving evaluation.

In contrast, PSI and SMART are mentioned less widely in the documents obtained. However, they are important as other MCDM methods for ranking purposes. PSI was used in Emovon and Oghenenyerovwho [19], Ulutas et al. [32], Yadav et al.[33], and Amin et al. [34], while SMART was identified in Çakır [41]. Apart from the basic MCDM methods mentioned above, others were found combined with different techniques or with other MCDM methods such as AHP, VIKOR, PROMETHEE, ELECTRE, SEM, PLS-SEM, NBR and FMULTIMOORA [29], [31], [19], [35], [36], [44-48]. This richness means that product selection itself can be interpreted as a complex problem with approaches in terms of rankings and explanations. Nonetheless, most of the previous works have just tested either two or a single algorithm in a particular context. A comparative study that contains MOORA, SMART, PSI, and TOPSIS within the framework of product selection should be done systematically. A comparative study with a larger number of seller-related criteria will help the sellers to choose the most suitable products, especially for sellers related to online home-based business.

3) Criteria used in previous studies: It has been observed that the product selection criterion used may consist of a range of factors that are decided by the problem in the decision-making context and that can also be affected by the set of attributes used in a particular dataset and study motivation. As summarized in Table I, it can be seen that rating and price are the main selection attributes used in previous studies.[23], [29]-[32], [34], [22, 37-42]. Two criteria are frequently applied because they refer to central dimensions in product assessment, and they are easy to understand; ranking indicates rating on subjective perception of quality, satisfaction, or performance, and price indicates economic accessibility and good buy. Usually, in a decision problem, the above criteria form the center part, upon which alternatives could be selected, like those selected from the research area of supplier selection, material selection, or online shop selection problems.

Among all considered marketplace dimensions apart from rating and price, some research also paid attention to: discount, sold, and response chat [23], [36], [44]. Discount denotes commercial attractiveness and attracts shoppers; it may have a positive effect on consumers' intention. Sold refers to the sale volume of items in the marketplaces and is the proxy of customer preferences; it denotes that items' desirability is already proven by actual consumers' demand. Response chat measures both the speed of the seller's reply, interaction level, and service, and so is relevant particularly at marketplaces, which are the environment where there is a need to give information quickly, and can improve consumer experience and influence buying intentions. These criteria have to be taken into account, especially in digital commerce, not only for a cost/quality reason, but also related to market and service success.

Besides, more recent papers have emphasized the necessity of applying behavioral and engagement criteria which represent purchasing intention and loyalty, such as repurchase, favorite, and followers [28], [35], [36], [43-48]. Repurchase represents customers' re-purchasing the item, representing overall customer satisfaction, overall product quality, and the long-term product's appeal; Favorite represents customers' like or collection behavior, meaning that the customer noted this item but has not even purchased it, let alone repurchased it. Follower represents the seller's audience value. These criteria are increasingly important because online product selection is now shaped not only by transactional performance but also by social signals and customer engagement.

B. Statistical Test

1) Stepwise multiple linear regression: Stepwise multiple linear regression is commonly used to identify relevant predictors from a larger set of candidate variables. Liu et al. (2021) applied forward stepwise regression, VIF analysis, and significance testing to 115 explanatory variables and obtained a final model explaining 95% of crop yield variation[28], Muazam et al. (2024) used stepwise multiple regression to determine the agronomic variables affecting sorghum production. From ten candidate variables, the model selected only two significant predictors, namely panicle weight and days to harvesting, with an R^2 value of 0.754 [49]. These studies show that stepwise multiple linear regression is useful not only for prediction but also for screening and validating important variables before they are used in a decision model.

Therefore, in this study, Forward Stepwise Multiple Linear Regression was selected because the candidate predictors had been identified in advance from the literature and the proposed SPS framework. Therefore, the procedure was not intended to explore an unrestricted set of variables, but to examine the order and incremental statistical contribution of seven predefined marketplace variables to the observed Sold variable. The method also provides an interpretable sequence showing whether each additional predictor contributes to the explanatory capacity of the model. Nevertheless, stepwise regression may be sensitive to sample characteristics and may produce statistically significant results for predictors with small practical contributions, particularly when applied to a large dataset. Therefore, the results were interpreted using not only significance values, but also Adjusted R Square, changes in R

Square, standardized coefficients, multicollinearity statistics, residual diagnostics, and the magnitude of the additional contribution of each predictor.

C. Measurement

Measurement was used to evaluate the stability and consistency of the ranking results under changes in criteria weights. MAD, MSE, and Spearman's Rank Correlation were chosen as all these indices are commonly used in MCDM literature and measure the disparate yet complementary aspects of ranking stability [50–53]. MAD represents the mean absolute displacement of values between rankings and gives an indication of the average shift of an alternative. MSE is designed to disproportionately penalize greater displacements than those associated with MAD; and Spearman's rank correlation reflects a measure of concordance with respect to the orders being compared. Therefore, a combination is provided of mean rank displacement, evidence of substantial change in the rankings, and agreement in rank order.

III. Methodology

A. Proposed Method

1) Current model and its limitations: The current model suggested by Okfalisa et al. Has been a reference for marketplace decision support since using the MOORA method as an alternative to rank online shops to be the best by using 5 criteria: Rating, Price, Sold, Discount, and Response Chat. However, the benchmark has three distinct limitations in relation to the present study.

First, it has a perspective limitation. The benchmark was developed from the Customer's Perspective to help buyers select an online shop, whereas the present study addresses Seller Product Selection from the Seller's Perspective. These decision problems involve different users and different decision objectives.

Second, it has an empirical-domain limitation. The benchmark was demonstrated using men's shirt products and 50 marketplace alternatives. Therefore, its findings cannot automatically be generalized to broader product categories or substantially larger marketplace datasets.

Third, it has a criteria-structure limitation. The five benchmark criteria do not include indicators of repeated purchasing behavior, product interest, and shop visibility. Specifically, Repurchase, Favorite, and Followers are not represented in the benchmark structure.

The proposed 8C SPS model directly addresses the perspective and criteria-structure limitations by adopting the Seller's Perspective and incorporating Repurchase, Favorite, and Followers. However, the present study remains limited to Shopee Indonesia food and beverage products. Validation in other product categories and marketplace environments is therefore still required.

2) Proposed seller product selection model: In an effort to bridge this gap, this research extends the benchmark on the SPS Model from 5 dimensions to 8 dimensions, namely: Repurchase (Repeat-purchase behavior and customer-retention), Favorite (Product interest and favorability), and Follower (Visibility of shop and social connections' effectiveness in the online environment). Consequently, the model builds the customer's perspective on SPS to the seller's perspective on SPS by considering 8 dimensions: transactional effectiveness, repeated potential purchases, appraisal customers have, the quality of promotions, interaction quality, visibility of transactions and promotions, audience attention of the market, and visibility of the shop. Compared to the SPS model with the 5 criteria of Okfalisa et al. The SPS model with 8Cs enhances its completeness in making an appropriate decision of selecting the best-suited product for sellers.

B. Research Framework and Practical Application Workflow

A framework is suggested that increases the benchmark 5C model by factors - Repurchase, Favorite, and Followers- leading to the model represented by the 8C SPS as described above. Both the Regression analysis investigate empirical effects of 7 independent variables on Sold and seller's perceived criteria weights as perceived through a survey based on Importance. Borda Weights explain seller preferences, and AHP will provide weights computed based on the selection of 4 'MCDMs'(MOORA, SMART, TOPSIS, PSI) for 150,000 product evaluations on five different configurations of these 8 factors based on performance metrics like 'MAD', 'MSE ', and 'Spearman's Rank Correlation '. A visualization of the framework is shown in Fig. 1.

In real-time deployment, sellers need to supply the values for these eight factors in relation to candidate products and run the weights derived by AHP using any of the discussed MCDM approaches to produce a list of items in rank order of their marketability potential. Based on the current experiment in the given context, MOORA achieved the strongest overall robustness in the 8C model, with the lowest MAD and MSE and the highest Spearman's Rank Correlation. In contrast, TOPSIS showed greater consistency in maintaining the same first-ranked alternative across the weight conditions. Therefore, MOORA is recommended as the primary ranking algorithm, while the leading alternatives identified by TOPSIS may be used as additional candidates for seller assessment.

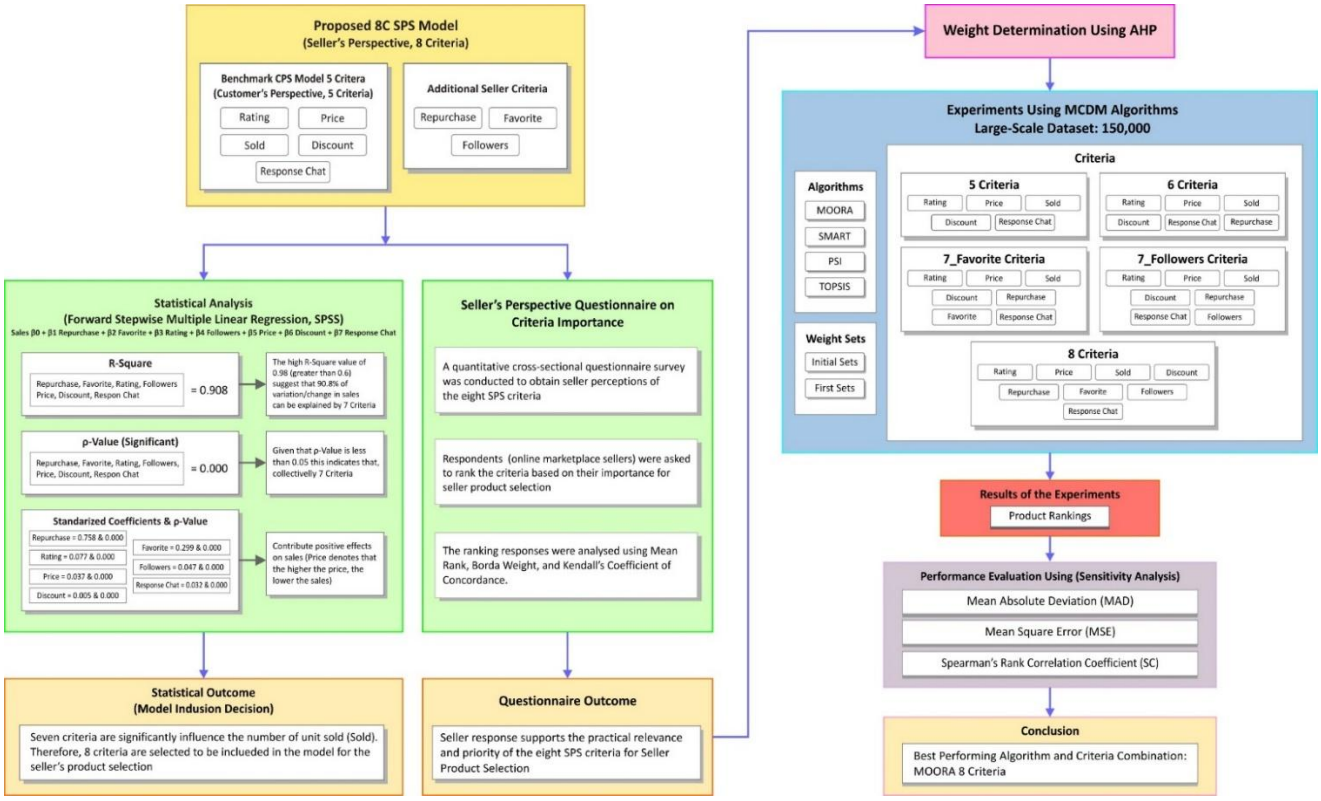


Fig. 1. Research method framework.

C. Dataset and Criteria Description

1) Data collection, preprocessing, and ethical considerations: The dataset was obtained using a Google Chrome extension to perform the crawling activity through several categories of Food and Beverage products available publicly from the Shopee Indonesia website during the research data collection stages. A total number of variables used that are publicly listed on both the product level (product name and URL) and shop level (name of the shop, Rating, Price, Sold, Discount, Response Chat, Repurchase, Favorite, Followers) were captured. The scraping metadata for the page and the order were also extracted so as to retrieve a source document used to identify origin pages and scraping order. The downloaded records were exported in the format of a raw file, and further processing for statistical analysis and MCDM experiments was conducted at this stage.

For standardization of the marketplace records for consistency, the following steps of data preprocessing were performed: Scraping Metadata columns to discard. Rows with product names missing/not available and Rows without pricing Information discarded, based on (unique based on product names), duplicates rows found, attribute Name standardized on basis of categorization, empty numerical value were replaced by zero value, (Rating, Discount, Sold, Favorite, Repurchase, Response Chat and Followers), where regular price not available discounted Price assumed final Price of the products, The text based marketplace values in their numerical presentation were removed currency symbol, thousand separator, percentage symbol, and other such characters to remove non numerical label then used as normalized to match the marketplace abbreviation like "RB", "+", then convert to numerical data types.

The issue of ethics has been handled by constraining the data extraction to information presented openly and publicly through the Shopee Indonesia product and shop pages. Data privacy has ensured that account numbers, private conversation message data, financial transaction data, and other personal information were not harvested. The product identification or name of the shop and seller was not revealed in the results, and each item was assigned an identifier code which served in the place of the product ID for the purpose of experimentation, and the dataset was not to be reused again in its original format that was identifiable.

2) Criteria description: The attributes dataset in this study is selected to be two criteria frames: Customers' Product Selection (CPS) and Sellers' Product Selection (SPS). We propose CPS as a benchmark model with five criteria: Rating, Price, Sold, Discount, Response Chat, following the literature on customer choice based on Okfalisa et al. SPS model is an extension of CPS to eight criteria by adding the other three criteria, namely Repurchase, Favorite, and Followers, to CPS, which reflect customers' loyalty, customers' interest in products that are not immediately purchased, and shop reputation, respectively, all of which are considered to increase product sales. Based on this framework, the mapping of Shopee dataset attributes to CPS and SPS criteria is presented in Table II.

TABLE II. OPERATIONAL DEFINITION OF CRITERIA IN CPS AND FAHP SPS MODELS

No	Criterion	Description	Criterion Type	Model
1	Rating	Average user rating score for a product	Benefit	CPS & SPS
2	Price	Product price listed on the product page	Cost	CPS & SPS
3	Sold	Number of units sold during the observation period	Benefit	CPS & SPS
4	Discount	Discount percentage offered for the product	Benefit	CPS & SPS
5	Response Chat	Seller responsiveness percentage to buyer chats	Benefit	CPS & SPS
6	Repurchase	Repeat-purchase indicator reflecting customer loyalty	Benefit	SPS
7	Favorite	Number of users who mark the product as a favorite/wishlist	Benefit	SPS
8	Followers	Number of shop followers as an indicator of shop reputation	Benefit	SPS

D. Analytical Statistic Method

The relevance of Rating, Price, Discount, Favorite, Repurchase, Response Chat, and Followers was examined using Forward Stepwise Multiple Linear Regression in SPSS. Sold was treated as the dependent variable, while the seven remaining criteria were entered as candidate predictors. The candidate variables were defined before the regression analysis based on the literature and the proposed SPS framework. Therefore, the regression was used as an exploratory criterion-screening procedure rather than as a forecasting model. The stepwise procedure examined the incremental statistical association of each predictor with Sold, while the standardized coefficients were used to compare their relative contributions. Model reliability was examined using Adjusted R Square, changes in R Square, tolerance, Variance Inflation Factor, the Durbin-Watson statistic, and residual diagnostics.

1) Forward stepwise multiple linear regression: The general regression model is expressed as shown in Eq. (1).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \quad (1)$$

where, Y is the dependent variable, $X_1, X_2, \ldots, X_n$ are the independent variables, β_0 is the constant, $\beta_1, \beta_2, \ldots, \beta_n$ are regression coefficients, and ε is the error term. The forward stepwise procedure adds predictors iteratively, starting from the strongest predictor, until no additional variable significantly improves the model.

The use of forward selection was considered appropriate for the present study because the analysis aimed to determine the order in which the predefined predictors contributed additional explanatory information. However, because the dataset contains a large number of observations, statistical significance alone was not considered sufficient evidence of practical importance. The interpretation, therefore, considered the magnitude of the standardized coefficients and the increase in R-squared produced at each step.

2) The standardized coefficients (beta): The Standardized Coefficients (Beta) were taken into consideration to show the relative contribution of each selected predictor in the regression models developed. The standardized coefficient is represented in Eq. (2).

$$\beta_j^* = b_j \left(\frac{s_{X_j}}{s_Y}\right) \quad (2)$$

where, β_j^* is the standardized coefficient, b_j is the unstandardized coefficient, s_{X_j} is the standard deviation of the predictor X_j, and s_Y is the standard deviation of the dependent variable. A higher value of an absolute Beta refers to a higher contribution towards Sales. Thus, supported empirically in the case of retaining variables that are good for the analysis of selecting a product.

E. Questionnaire Outcome

A quantitative cross-sectional questionnaire survey was used in this study to get the criteria importance from the seller's point of view. The questionnaire asked the respondent to rank the 8 SPS criteria - namely Sold, Discount, Repurchase, Rating, Favorite, Price, Response Chat, and Followers according to the rank from the most important to the least important. The questionnaire was thus utilized to provide evidence of the buyer based on the seller, the practicability and prioritization of the SPS criteria, prior to carrying out computational weighing.

The forced ranking exercise forced each participant to give a rank value to each criterion, varying from rank 1 (highest priority) to rank 8 (lowest priority). Since the obtained data were ordinal in nature, the analysis of such data concerned rank aggregation and the level of agreement between participants. The data gathered through the use of the questionnaire were analyzed through descriptive analysis, with summary statistics of the data presented through Mean Rank, Borda Weight, and Kendall's coefficient of concordance. Mean Ranks were employed in order to provide an overall ranking of priorities of each item, with the lowest mean rank being assigned to the highest priorities. The ranking data was then converted into numerical weight via the use of the Borda Weight. The Borda score is calculated as shown in Eq. (3)

$$b_{ij} = n - r_{ij} + 1 \quad (3)$$

where, b_{ij} is the Borda score of the criterion j assigned by the respondent i, r_{ij} is the rank given by the respondent i to criterion j, and n is the number of criteria. The normalized Borda Weight for each criterion is calculated as shown in Eq. (4).

$$w_j = \frac{\sum_{i=1}^{m} b_{ij}}{\sum_{j=1}^{n}\sum_{i=1}^{m} b_{ij}} \quad (4)$$

where, w_j is the normalized Borda Weight of the criterion j, m is the number of respondents, and n is the number of criteria.

Kendall's coefficient of concordance was used to examine the level of agreement among respondents because the questionnaire involved ranking the same set of criteria. Kendall's W is calculated as follows: Eq. (5)

$$W = \frac{12S}{m^2(n^3-n)} \quad (5)$$

where, S is the sum of squared deviations of the total ranks for each criterion from the mean total rank, m is the number of respondents, and n is the number of ranked criteria. A higher value of W indicates stronger agreement among respondents.

Methodologically, the questionnaire complements the statistical validation and experimental evaluation of the proposed SPS model. This result not only adds evidence towards AHP weighting but would also make it possible to show that the chosen eight criteria actually matter to a seller's product range. The response to the seller questionnaire also indicated that there is a practical application for and, in fact, a practical priority among the eight SPS criteria for sellers' product range.

F. Research Setup

1) Weight setup: The weight configurations for two weight options, Initial Weights and First Weight, were generated for ranking stability testing. First Weight adjusted the original weight structure to evaluate the susceptibility of each decision-making algorithm to any variations in criterion weight configurations. A graphical representation is generated showing the criterion weight distribution of various criterion combinations examined within the scope of this project. The weight of each criterion combination for this exercise is achieved through the Analytical Hierarchy Process (AHP) by building criteria from a given problem, creating a pairwise comparison matrix, calculating criterion priority vectors, and then normalizing weights to equal 1 for each combination.

- 8 criteria: Pairwise comparison assessments were collected through a questionnaire for all criteria, and their weights were then calculated using AHP.
- 6 criteria: Two criteria were removed (Favorite/Wishlist and Followers), and the weights for the remaining six criteria were then recalculated using AHP and normalized.
- 7 criteria (Favorite): The Followers criterion was removed, leaving 7 criteria (including Favorite/Wishlist); the weights were then recalculated using AHP and normalized.
- 7 criteria (Followers): The Favorite/Wishlist criterion was removed, leaving 7 criteria (including Followers); the weights were then recalculated using AHP and normalized.
- 5 criteria (benchmark): weights are determined based on the reference weights (benchmarks) for five basic criteria, then used as a comparison in the evaluation of the method's performance.

The weight settings in this experiment were set under two conditions to test the stability of the ranking results when decision-making preferences change. The weight settings are as follows:

- Initial weight (Baseline): obtained from the AHP calculation, where pairwise comparison assessments were collected through a questionnaire for all criteria, then calculated into priority weights and normalized.
- First Weight: obtained from adding 1 point to the pairwise comparison by pairs; we will be able to obtain an alternative weight arrangement pairwise by pairwise to study rank modifications in this alternative weighing.

2) AHP: The Analytic Hierarchy Process (AHP) was applied to derive the relative weights of the SPS criteria through pairwise comparisons. The criteria were compared using Saaty's 1–9 scale, where 1 represents equal importance, and 9 represents absolute importance. The pairwise comparison matrix was normalized, and the priority vector was calculated to obtain the criteria weights. This procedure was conducted separately for the 5C, 6C, 7C Favorite, 7C Followers, and 8C configurations because each configuration contained a different set of criteria. The consistency of the pairwise judgments was evaluated using the Consistency Index and Consistency Ratio, as expressed in Eq. (6) and (7).

$$CI = \frac{\lambda max - n}{RI*(n-1)} \quad (6)$$

$$CR = \frac{CI}{RI} \quad (7)$$

If the CR value is $\leqslant$ (less than or equal to) 0.1, then the criteria values are consistent. However, if not, a re-evaluation is necessary to ensure the consistency of the criteria values.

3) Comparative MCDM algorithms: Four established MCDM algorithms, namely MOORA, SMART, PSI, and TOPSIS, were applied to rank the product alternatives. These algorithms were selected because they represent different ranking mechanisms and enable the robustness of the proposed criteria configurations to be compared under the same dataset and weight conditions. Price was treated as a cost criterion, while Rating, Sold, Discount, Response Chat, Repurchase, Favorite, and Followers were treated as benefit criteria.

a) MOORA: The Multi-Objective Optimization based on Ratio Analysis (MOORA) ranks alternatives by normalizing the decision matrix and combining the weighted benefit and cost criteria. The final assessment value is obtained by subtracting the weighted cost values from the weighted benefit values. An alternative with a higher final value receives a higher ranking. (8).

$$y_i = \sum_{j=1}^{g} w_j \, x_{ij}^* - \sum_{j=g+1}^{n} w_j \, x_{ij}^* \quad (8)$$

where, w_j denotes the weight of the criterion j. A higher y_i indicates a better alternative.

b) SMART: The Simple Multi-Attribute Rating Technique (SMART) determines the overall performance of each alternative by aggregating its weighted utility values. Each criterion value is transformed into a utility value according to its benefit or cost orientation, and the alternatives are ranked according to their total utility scores. The overall score of each alternative is then obtained by Eq. (9).

$$S_i = \sum_{j=1}^{n} w_j \, 'u_{ij} \quad (9)$$

where, u_{ij} is the utility value of the alternative i on criterion j. Alternatives with higher S_i values are ranked higher.

c) PSI: The Preference Selection Index (PSI) evaluates alternatives using normalized criterion values and preference indices derived from the variation of each criterion. Unlike the other evaluated methods, PSI can determine criterion preferences from the decision matrix. The alternative with the highest total preference score is assigned the highest rank.

The final PSI score is calculated as shown in Eq. (10).

$$PSI_i = \sum_{j=1}^{n} N_{ij}\, \Omega_j \tag{10}$$

where, Ω_j denotes the preference value of the criterion j. A higher PSI_i indicates a more preferred alternative.

d) TOPSIS: The Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) ranks alternatives according to their relative closeness to the positive ideal solution and their distance from the negative ideal solution. An alternative is considered preferable when it is closer to the positive ideal solution and farther from the negative ideal solution. The relative closeness of each alternative is then determined by Eq. (11).

$$C_i = \frac{D_i^-}{D_i^+ + D_i^-} \tag{11}$$

where, D_i^+ and D_i^- are the distances from the positive and negative ideal solutions, respectively. A higher C_i indicates a better alternative.

4) Measurement: In this study, the Initial Weight ranking was treated as the baseline, while the First Weight ranking represented the modified preference condition. A lower MAD value indicates that alternatives move fewer ranking positions on average. A lower MSE value indicates that large or extreme rank changes are less pronounced. A Spearman coefficient closer to 1 indicates that the overall ordering of the alternatives remains consistent after the weight modification. Accordingly, an algorithm is considered more robust when it simultaneously produces lower MAD and MSE and higher Spearman's Rank Correlation.

a) Mean Absolute Deviation (MAD): The following is the equation for MAD shown in Eq. (12).

$$MAD = \frac{1}{n}\sum_{i=1}^{n} |xij - \text{yj}| \tag{12}$$

b) Mean Squared Error (MSE): The following is the equation for MSE shown in Eq. (13).

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (xij - yj)^2 \tag{13}$$

c) Spearman's Correlation (SC): The following is the equation for SC shown in Eq. (14).

$$SC = r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{14}$$

IV. Results and Discussion

A. Results of Questionnaire

The questionnaire was administered to 30 online marketplace sellers from 22 September 2025 to 2 October 2025. Table III presents the criterion priorities obtained from the seller questionnaire.

TABLE III. Result of Criteria Priority

Criteria	Mean Rank	Borda Weight	Priority
Sold	2.40	0.183	1
Repurchase	3.63	0.149	2
Rating	4.00	0.139	3
Response Chat	4.30	0.131	4
Price	4.57	0.123	5
Favorite	5.53	0.096	6
Discount	5.57	0.095	7
Followers	6.00	0.083	8

As can be seen from Table III, Sold had shown a value of the most important factor as low as 2.40 in rank mean; also, it possessed a value of the highest in Bordas Weight for 0.183. Hence, Sold shows more importance to sellers to prioritize the products for sales as they perform well.

Overall, the ranking result suggests that the responders perceived Sold, Repurchase, and Rating to be the top three criteria that have the highest priority. As for the criteria Favourite, Discount, and Followers, the priorities received are not high; however, it still exists as the other supportive SPS criteria for the Seller's Perspective. As the Kendall's Coefficient of Concordance W was 0.236 and there are 30 responders, we shall view the questionnaire results as the potential evidence from the sellers' view on the practical relevance and priority of these 8 SPS criteria. The potential priority that responders perceived does not weaken the priority obtained via regression and computationally assigned with the weighting factors with AHP and ranked by the result of the algorithms using MCDM.

B. Statistical Test

1) Stepwise multiple linear regression: The following are the statistical results of the test using Stepwise Multiple Linear Regression, which are shown in Table IV.

The change in explanatory power across the steps is evident in Table IV. Entering Repurchase alone explained 0.811 R-square. The addition of Favorite over and above Repurchase helped the R-square improve by 0.084 with a total explanation of 0.895 (0.811 + 0.084).

The addition of Rating yielded yet another increase in explained variance to 0.094, while Followers increased the R-square by an additional 0.003.

Each of Price, Response Chat, and Discount increased R-square by 0.001, respectively. Based on the changes across steps, Repurchases and Favorite had greater explanatory influence among the set. Price, Response Chat, and Discount added the lowest amount to the explanation of buying.

The last iteration (Model 7), as shown in Table V, shows that the R-square value is 0.908. It suggests that 90.8 percent of the variability in Sales is explained by Repurchase, Favorite, Rating, Follower, Price, Response Chat, and Discount as a group.

Meanwhile, the remaining 9.2 percent is explained by other factors. In other words, these seven variables collectively correlate strongly with Sales.

TABLE IV. MODEL SUMMARY

Model	R	R Square	Adjusted R-Square	Std. Error of the Estimate	Durbin-Watson
1	0.900[a]	0.811	0.811	28.035	
2	0.946[b]	0.895	0.895	20.857	
3	0.950[c]	0.902	0.902	20.129	
4	0.951[d]	0.905	0.905	19.847	
5	0.952[e]	0.906	0.906	19.718	
6	0.953[f]	0.907	0.907	19.610	
7	0.953[g]	0.908	0.908	19.585	1.863
a. Predictors: (Constant), Repurchase					
b. Predictors: (Constant), Repurchase, Favorite					
c. Predictors: (Constant), Repurchase, Favorite, Rating					
d. Predictors: (Constant), Repurchase, Favorite, Rating, Follower					
e. Predictors: (Constant), Repurchase, Favorite, Rating, Follower, Price					
f. Predictors: (Constant), Repurchase, Favorite, Rating, Follower, Price, Response_Chat					
g. Predictors: (Constant), Repurchase, Favorite, Rating, Follower, Price, Response_Chat, Discount					
h. Dependent Variable: Sales					

TABLE V. ANOVA[A]

Model		Sum of Squares	df	Mean Square	F	Sig.
7	Regression	555699114.195	7	79385587.742	206961.417	.000[h]
	Residual	56589845.238	147532	383.577		
	Total	612288959.433	147539			
a. Dependent Variable: Sales						
h. Predictors: (Constant), Repurchase, Favorite, Rating, Follower, Price, Response Chat, Discount						

Furthermore, despite a strong correlation between independent and dependent variables, it is important to examine whether this correlation is statistically significant. Table VI reveals that the F test results in a p-value (Sig.) of 0.000, much less than 5 percent or 0.05 of the significance level. It indicates that the correlation between all seven independent variables of Repurchase, Favorite, Rating, Follower, Price, Response Chat, and Discount as a group with Sales is statistically significant.

2) The standardized coefficients (beta): After assessing the correlation between dependent and independent variables and their significance in collective terms, Table VI explains the correlation or the influence of each independent variable on the dependent variables separately.

Table VI presents the individual statistical association of each independent variable with Sold. All predictors were statistically significant at p <0.001. Repurchase had the largest standardized coefficient (β=0.758), followed by Favorite (β=0.299), Rating (β=0.077), Followers (β=0.047), Price (β=−0.037), Response Chat (β=0.032), and Discount (β=0.016). Repurchase and Favorite therefore showed the strongest statistical associations with Sold. Price had a negative coefficient, indicating that higher product prices were associated with lower observed sales, while the other predictors showed positive associations. Because the data are observational, these coefficients represent statistical associations rather than causal effects.

TABLE VI. COEFFICIENTS[A.]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
7	(Constant)	-2.190	0.159		-13.763	.000		
	Repurchase	7.025	0.008	0.758	867.811	.000	0.820	1.219
	Favorite	0.904	0.003	0.299	336.990	.000	0.794	1.259
	Rating	2.359	0.025	0.077	94.660	.000	0.940	1.063
	Follower	.000	0.000	0.047	57.889	.000	0.935	1.070
	Price	-3.239E-5	0.000	-0.037	-45.961	.000	0.992	1.008
	Response Chat	.0077	0.002	0.032	39.698	.000	0.955	1.047
	Discount	0.106	0.005	0.016	19.508	.000	0.973	1.028
a. Dependent Variable: Sales								

3) Regression diagnostics and model reliability: Final regression model fitting proceeded, and the diagnostics for the final model were computed. The model resulted in an R^2 value of 0.908, while the Adj. R^2 value is 0.908. The explanatory power does not appear to be seriously distorted by the addition of extra predictors. Multicollinearity among predictors did not seem to be serious; all tolerance values were between 0.794 and 0.992, and VIF values did not exceed 1.259. The Durbin-Watson value is 1.863, which also suggests no serious serial autocorrelation among residuals. The normality of residuals has also been explored through the Normal P-P plot (Fig. 2) and the Histogram of residuals (Fig. 3).

With reference to Fig. 2, the cumulative probabilities have systematic deviations from the reference diagonal line at probabilities from the lower and higher ends. This means that the distribution of residuals does not fall on a normal distribution.

Fig. 3 displays standard normal residuals that are centered at zero but follow an uneven, positively skewed distribution. This again accords with the results of the P-P plots indicating less-than-ideal normal residuals; however, considering the massive nature of the dataset (150,000 records), this suggests that the regression should be used solely for investigation of statistical associations between the sampled data points only, rather than

to establish causalities or as a tool to validate external predictions of future Sold values.

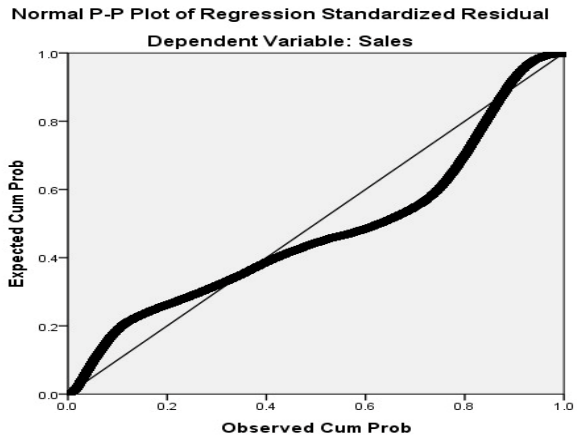


Fig. 2. Normal P-P plot of regression standardized residuals.

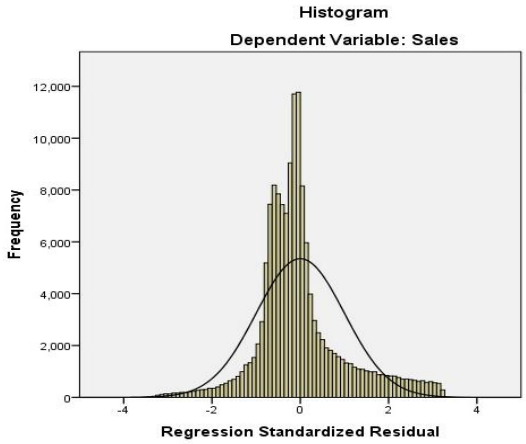


Fig. 3. Histogram of regression standardized residuals. (Note: "Sales" in the SPSS output represents the Sold variable used in this study.)

4) Relationship between questionnaire and regression results: Questionnaire and regression results perform different yet complementary functions. The questionnaire reports that all eight factors of SPS (including the Sold factor) are considered important to sellers. Hence, the questionnaires tell which factors a seller considers to be most important when he selects products to sell. The regression does the opposite: taking Sold as the dependent variable and analyzing the statistical relationship between the other seven factors and the Sold performance observed in the marketplace.

The results from the questionnaires and the regression should not therefore be in the same order of priority. Sold was found to be in the first order because it was considered most important, as it represents the most direct feedback on product marketability to the sellers. Repurchase was found to be in the first order in terms of standardized regression coefficient due to having the strongest statistical correlation with the Sold observed in the marketplace data. Questionnaires, therefore, measure subjective importance; regression measures the importance observed in market data.

C. Weight Determination Using AHP

Based upon the AHP determining result, two different weight combinations were generated for each criterion combination scenario. Initial weight indicated the base preferences scenario, and First weight represented a variation in a preference scenario used for ranking sensibility analysis. Table VII lists weight configurations for five-, six-, seven-, and eight-criteria combinations.

TABLE VII. INITIAL AND FIRST WEIGHT RESULTS

Criterion	5C I/F	6C I/F	7C Fav I/F	7C Foll I/F	8C I/F
Rating	0.417/ 0.482	0.160/ 0.150	0.160/ 0.153	0.160/ 0.153	0.159/ 0.155
Price	0.263/ 0.235	0.101/ 0.088	0.104/ 0.094	0.104 /0.094	0.107/ 0.099
Sold	0.160/ 0.149	0.381/ 0.424	0.352 /0.388	0.352/ 0.389	0.328/ 0.360
Discount	0.097/ 0.082	0.042/ 0.031	0.031/ 0.023	0.045/ 0.036	0.033/ 0.026
Response Chat	0.062/ 0.053	0.064/ 0.052	0.068/ 0.059	0.068/ 0.058	0.071/ 0.063
Repurchase	–	0.252/ 0.255	0.241/ 0.246	0.241/ 0.247	0.232/ 0.238
Favorite	–	–	0.045/ 0.036	–	0.048/ 0.041
Followers	–	–	–	0.031/ 0.023	0.023/ 0.018
CR	0.02/ 0.07	0.02/0.08	0.02/0.08	0.03/0.08	0.03/0.09

Note: I/F denotes Initial Weight/First Weight; 5C = five criteria; 6C = six criteria; 7C Fav = seven criteria including Favorite; 7C Foll = seven criteria including Followers; and CR = Consistency Ratio.

Table VI shows that CR values were within the required level of consistency, less than 0.10 for all of the weight sets. Table VII shows that all the calculated CR values were less than 0.10. Therefore, the weight sets show acceptable consistency. Those sets were applied for MOORA, SMART, PSI, and TOPSIS methods in order to view the product ranking in the change conditions.

D. Ranking Result for all Criteria

The result of rankings for all criteria shows the final scores of alternatives as produced by MOORA, SMART, PSI, and TOPSIS along with their orderings in ranking. The tables below show all of the combinations for evaluating the alternatives among different criteria to show the results of ordering the product according to different MCDM approaches.

For the five-criteria model in Table VIII, the proposed alternative number, A107137, obtained the first position by any algorithm for both criteria types, Initial Weight and First Weight. By using the evaluation from only benchmark criteria, the proposed alternative A107137 obtained the highest marks among the five benchmark alternative selections from any one of these criteria (MOORA, SMART, PSI, and TOPSIS. The consistent ranking by A107137 with any four criteria suggested that the proposed alternative possesses a relatively stable advantage in the selection from the five original criteria. For example, the benchmarks: A82978, A143588, A108358, A11740, A99974, and A133019 still seem to have better selection for a structure of benchmark criterion.

TABLE VIII. RANKING RESULT 5 CRITERIA

Rank	Algorithm															
	Moora				Smart				PSI				Topsis			
	Initial Weight		First Weight		Initial Weight		First Weight		Initial Weight		First Weight		Initial Weight		First Weight	
	Alt	Score	Alt	Score	Alt	Score	Alt	Score	Alt	Score	Alt	Score	Alt	Score	Alt	Score
1	A107137	0.06290	A107137	0.05829	A107137	6.23673	A107137	6.24027	A107137	0.66850	A107137	0.69148	A107137	0.72179	A107137	0.73848
2	A82978	0.05912	A82978	0.05507	A82978	6.21868	A82978	6.23373	A82978	0.64910	A82978	0.68361	A143588	0.70519	A143588	0.72262
3	A143588	0.05867	A143588	0.05469	A108358	6.19245	A108358	6.21827	A108358	0.61511	A108358	0.66048	A82978	0.69141	A82978	0.70930
4	A108358	0.05743	A108358	0.05365	A133019	6.15530	A133019	6.18784	A133019	0.59415	A133019	0.64604	A11740	0.68676	A11740	0.70486
5	A11740	0.05658	A11740	0.05291	A99974	6.13161	A99974	6.16924	A99974	0.56256	A99974	0.61964	A99974	0.68673	A99974	0.70484
6	A99974	0.05654	A99974	0.05288	A11740	6.12682	A11740	6.16262	A11740	0.55421	A11740	0.60950	A108358	0.68292	A108358	0.70112
7	A133019	0.05647	A133019	0.05282	A143588	6.12521	A143588	6.15692	A143588	0.55089	A143588	0.60212	A133019	0.68188	A133019	0.70013
8	A44595	0.03166	A44595	0.02781	A141367	6.08252	A141367	6.12492	A141358	0.51574	A7532	0.56856	A44595	0.35031	A44595	0.33248
9	A149486	0.03052	A149486	0.02686	A141358	6.08163	A82112	6.12200	A7532	0.50663	A141367	0.56856	A149486	0.32891	A149486	0.31202
10	A17363	0.02313	A17363	0.02052	A82112	6.07897	A42946	6.11906	A141367	0.50663	A42946	0.56855	A17363	0.26012	A136643	0.25992

TABLE IX. RANKING RESULT 6 CRITERIA

Rank	Algorithm															
	Moora				Smart				PSI				Topsis			
	Initial Weight		First Weight		Initial Weight		First Weight		Initial Weight		First Weight		Initial Weight		First Weight	
	Alt	Score	Alt	Score	Alt	Score	Alt	Score	Alt	Score	Alt	Score	Alt	Score	Alt	Score
1	A107137	0.17439	A107137	0.18841	A107137	5.24992	A107137	5.10647	A107137	0.64828	A107137	0.69544	A45290	0.56609	A107137	0.55299
2	A45290	0.17056	A45290	0.17375	A82978	5.18319	A82978	5.04408	A82978	0.58103	A82978	0.63270	A107137	0.52966	A45290	0.54558
3	A143588	0.13926	A143588	0.15350	A108358	5.17250	A108358	5.03816	A108358	0.56736	A108358	0.62475	A108358	0.46086	A108358	0.48583
4	A108358	0.13762	A108358	0.15223	A133019	5.15177	A133019	5.02258	A133019	0.55284	A133019	0.61341	A143588	0.46082	A143588	0.48551
5	A82978	0.13743	A82978	0.15174	A143588	5.14785	A143588	5.01988	A143588	0.54387	A143588	0.60726	A82978	0.45865	A82978	0.48350
6	A11740	0.13566	A11740	0.15028	A11740	5.14286	A11740	5.01638	A11740	0.53953	A11740	0.60421	A11740	0.45778	A11740	0.48267
7	A133019	0.13385	A133019	0.14852	A99974	5.13997	A99974	5.01347	A99974	0.53801	A99974	0.60223	A133019	0.45536	A133019	0.48029
8	A99974	0.13271	A99974	0.14730	A45290	5.06574	A45290	4.86350	A45290	0.46257	A45290	0.45143	A99974	0.45308	A99974	0.47792
9	A19305	0.10309	A19305	0.10316	A19305	4.94899	A19305	4.74085	A19305	0.34917	A19305	0.33108	A19305	0.35811	A19305	0.34295
10	A45286	0.07217	A45286	0.07410	A149486	4.90443	A45286	4.69289	A149486	0.30132	A45286	0.28054	A45286	0.27507	A45286	0.26575

Considering Table IX, where the combination of criteria grows to six, A107137 still holds more than half of the ranking results. Its rank position is highest for all methods MOORA, SMART, and PSI under Initial Weight and First Weight except for TOPSIS, which lists A45290 first for Initial Weight and A107137 first for First Weight. This suggests that with the consideration of Repurchase, a shift happened in the ranking pattern of a couple of methods (in this case, TOPSIS), where A45290 achieved the first position under this methodology, and this implies it's favored due to the seller view-based ranking methods considered.

Similar ranking outcome patterns occur in seven-criteria models with Favorite Table X. A107137 still has rank 1 from MOORA, SMART, and PSI under both weighted setting while TOPSIS always suggests A45290 as the first alternative. Thus, there is an implication of the favorite effect in the ranking result, especially with the TOPSIS technique, in which each alternative is assessed according to its distances to ideal and anti-ideal solutions. This can be seen as a positive assessment of A45290 relative position compared to other alternatives.

TABLE X. RANKING RESULT 7 FAVORITE CRITERIA

Rank	Algorithm															
	Moora				Smart				PSI				Topsis			
	Initial Weight		First Weight		Initial Weight		First Weight		Initial Weight		First Weight		Initial Weight		First Weight	
	Alt	Score	Alt	Score	Alt	Score	Alt	Score	Alt	Score	Alt	Score	Alt	Score	Alt	Score
1	A107137	0.16322	A107137	0.17535	A107137	5.23159	A107137	5.16225	A107137	0.61270	A107137	0.64995	A45290	0.57258	A45290	0.55679
2	A45290	0.16305	A45290	0.16720	A82978	5.16888	A82978	5.10194	A82978	0.54945	A82978	0.58921	A107137	0.52077	A107137	0.53979
3	A143588	0.12965	A143588	0.14160	A108358	5.16113	A108358	5.09874	A108358	0.53864	A108358	0.58347	A143588	0.45157	A108358	0.47174
4	A108358	0.12769	A108358	0.13992	A133019	5.14254	A143588	5.08484	A133019	0.52644	A133019	0.57398	A108358	0.45142	A143588	0.47162
5	A82978	0.12756	A82978	0.13955	A143588	5.14172	A133019	5.08396	A143588	0.52042	A143588	0.57056	A82978	0.44929	A82978	0.46949
6	A117440	0.12602	A117440	0.13824	A117440	5.13734	A117440	5.08124	A117440	0.51671	A117440	0.56752	A117440	0.44841	A117440	0.46865
7	A133019	0.12411	A133019	0.13635	A99974	5.13266	A99974	5.07702	A99974	0.51344	A99974	0.56446	A133019	0.44596	A133019	0.46623
8	A99974	0.12308	A99974	0.13526	A45290	5.07181	A45290	4.96198	A45290	0.45135	A45290	0.44839	A99974	0.44375	A99974	0.46392
9	A19305	0.09951	A19305	0.10054	A19305	4.95993	A19305	4.84458	A19305	0.34291	A19305	0.33384	A19305	0.36360	A19305	0.35150
10	A45286	0.06910	A45286	0.07123	A149486	4.92366	A45286	4.79772	A149486	0.30326	A45286	0.28376	A45286	0.27798	A45286	0.27088

TABLE XI. RANKING RESULT 7 FOLLOWERS CRITERIA

Rank	Algorithm															
	Moora				Smart				PSI				Topsis			
	Initial Weight		First Weight		Initial Weight		First Weight		Initial Weight		First Weight		Initial Weight		First Weight	
	Alt	Score	Alt	Score	Alt	Score	Alt	Score	Alt	Score	Alt	Score	Alt	Score	Alt	Score
1	A107137	0.16453	A107137	0.17679	A107137	5.24720	A107137	5.18015	A107137	0.62830	A107137	0.66242	A45290	0.57249	A45290	0.55730
2	A45290	0.16376	A45290	0.16838	A82978	5.18412	A82978	5.11907	A82978	0.56469	A82978	0.60089	A107137	0.52072	A107137	0.53957
3	A143588	0.12956	A143588	0.14183	A108358	5.16254	A108358	5.10668	A108358	0.54004	A108358	0.58594	A143588	0.45148	A108358	0.47146
4	A82978	0.12917	A82978	0.14112	A133019	5.14143	A133019	5.08960	A133019	0.52533	A133019	0.57421	A108358	0.45133	A143588	0.47131
5	A108358	0.12766	A108358	0.14028	A143588	5.13755	A143588	5.08760	A143588	0.51626	A143588	0.56786	A82978	0.44929	A82978	0.46925
6	A117440	0.12568	A117440	0.13831	A99974	5.13118	A117440	5.08282	A99974	0.51196	A117440	0.56364	A117440	0.44831	A117440	0.46835
7	A133019	0.12405	A133019	0.13665	A117440	5.13108	A99974	5.08155	A117440	0.51046	A99974	0.56355	A133019	0.44587	A133019	0.46595
8	A99974	0.12331	A99974	0.13575	A45290	5.07586	A45290	4.97089	A45290	0.45539	A45290	0.45185	A99974	0.44368	A99974	0.46364
9	A19305	0.09948	A19305	0.10066	A19305	4.95937	A19305	4.84981	A19305	0.34235	A19305	0.33365	A19305	0.36349	A19305	0.35161
10	A45286	0.06875	A45286	0.07120	A149486	4.93225	A149486	4.80223	A149486	0.31184	A149486	0.28324	A45286	0.27789	A45286	0.27108

With regard to Table XI, similar behavior was found in the seven-criteria model with Followers. A107137 is again the best alternative by MOORA, SMART, PSI, and again TOPSIS indicates A45290 as the first alternative by applying both weight configurations. This shows that Followers is one of the factors contributing to the A107137 and A45290 rank exchange. In conclusion, the seller-perspective awareness and involvement-oriented attributes do have some impact on moving A45290 to become a competitor, although A107137 continues to be very stable with almost all alternatives.

In the eight-criteria model, referring to Table XII, A45290 has gained first place again using the MOORA method with Initial Weight and by the TOPSIS method with two weight settings, but A107137 with MOORA First Weight, SMART, and PSI, so it becomes that a total SPS criterion changes the ranking results obtained. Another aspect of the results is that an increased consideration of sellers (which comprises those five items) not only follows the previous result but produces a wider view on the possible purchase outcomes from the different alternatives of the product by taking into account the achieved

sales performance of the sellers as well as their customers' feedback in addition to repurchase rates, store visitors, favorites and interested individuals.

However, when introducing the SPS model from the seller's point of view, A45290 becomes significant at some combinations and even at the top when the number of criteria is eight, and it uses TOPSIS. Thus, the eight-criterion expanded SPS has shown a broader alternative arrangement than the six-criterion-based one by the combination of customer aspects and seller aspects in the online market environment. Although the rankings of the primary alternative are different from the four techniques, the rankings cannot be judged as wrong, as they use different normalization, aggregation, and distance techniques. For example, with the eight-criterion-based configuration, MOORA ranked A45290 first at the Initial Weight, and A107137 first at the First Weight; however, A45290 remained at the top under both Initial Weight and First Weight by use of TOPSIS. Therefore, algorithm selection was not based solely on the identity of the first-ranked alternative.

TABLE XII. RANKING RESULT 8 CRITERIA

Rank	Algorithm															
	Moora				Smart				PSI				Topsis			
	Initial Weight		First Weight		Initial Weight		First Weight		Initial Weight		First Weight		Initial Weight		First Weight	
	Alt	Score	Alt	Score	Alt	Score	Alt	Score	Alt	Score	Alt	Score	Alt	Score	Alt	Score
1	A45290	0.15742	A107137	0.16576	A107137	5.23171	A107137	5.18210	A107137	0.59496	A107137	0.62811	A45290	0.57846	A45290	0.56539
2	A107137	0.15495	A45290	0.16193	A82978	5.17185	A82978	5.12287	A82978	0.53455	A82978	0.56841	A107137	0.51242	A107137	0.52934
3	A143588	0.12160	A143588	0.13223	A108358	5.15607	A108358	5.11355	A108358	0.51562	A108358	0.55634	A143588	0.44285	A143588	0.46062
4	A82978	0.12058	A82978	0.13100	A133019	5.13718	A143588	5.09820	A133019	0.50331	A133019	0.54618	A108358	0.44253	A108358	0.46059
5	A108358	0.11946	A108358	0.13035	A143588	5.13654	A133019	5.09767	A143588	0.49732	A143588	0.54206	A82978	0.44052	A82978	0.45843
6	A11740	0.11779	A11740	0.12867	A11740	5.13109	A11740	5.09379	A11740	0.49256	A11740	0.53826	A11740	0.43957	A11740	0.45754
7	A133019	0.11600	A133019	0.12687	A99974	5.12848	A99974	5.09084	A99974	0.49139	A99974	0.53656	A133019	0.43711	A133019	0.45510
8	A99974	0.11527	A99974	0.12603	A45290	5.08299	A45290	5.00420	A45290	0.44463	A45290	0.44880	A99974	0.43496	A99974	0.45285
9	A19305	0.09650	A19305	0.09789	A19305	4.97134	A19305	4.88697	A19305	0.33652	A19305	0.33466	A19305	0.36861	A19305	0.35810
10	A45286	0.06631	A45286	0.06857	A149486	4.94867	A149486	4.84532	A149486	0.31037	A149486	0.28998	A45286	0.28056	A45286	0.27475

E. Measurement

Evaluation of the results obtained from MOORA, SMART, PSI, and TOPSIS regarding the sensitivity of the results due to criterion variations by applying measured values for the ranking outcomes. Three measurement indicators were applied, namely Mean Absolute Deviation (MAD), Mean Squared Error (MSE), and Spearman's Rank Correlation Coefficient (SC).

TABLE XIII. RESULT OF MEAN ABSOLUTE DEVIATION

Combined Criteria	Algorithm			
	MOORA	SMART	PSI	TOPSIS
5 Criteria	1,894	2,189	1,730	4,694
6 Criteria	1,674	1,662	1,463	1,783
7_Favorite Criteria	1,169	1,680	1,861	1,308
7_Followers Criteria	1,452	1,284	1,529	1,409
8 Criteria	1,093	1,474	1,616	1,221

The MAD results are further illustrated graphically in Fig. 4.

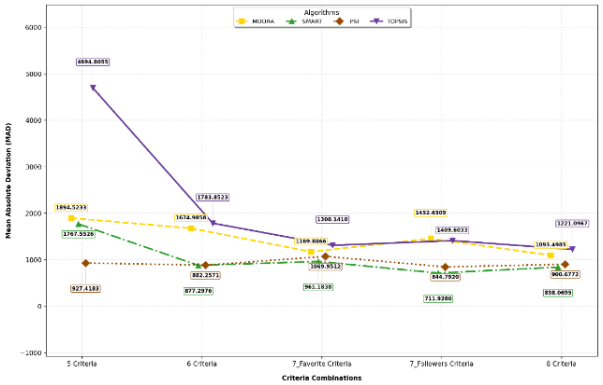


Fig. 4. Graph result of mean absolute deviation.

As presented in Table XIII and illustrated in the MAD graph in Fig. 4, the five-criteria model produces higher ranking deviation for several algorithms, especially TOPSIS, with a MAD value of 4,694. This indicates that the benchmark five-criteria model is more sensitive to changes in weight settings. When the criteria are expanded, the MAD values generally decrease.

The lowest MAD value is achieved by MOORA in the eight-criteria model, with a value of 1,093, followed by TOPSIS in the eight-criteria model with 1,221. This shows that the expanded seller-perspective criteria improve ranking stability, particularly for MOORA and TOPSIS.

TABLE XIV. RESULT OF MEAN SQUARED ERROR

Combined Criteria	Algorithms			
	MOORA	SMART	PSI	TOPSIS
5 Criteria	16,268,119	19,935,786	45,773,915	43,139,503
6 Criteria	12,544,343	9,776,007	24,513,668	12,965,720
7_Favorite Criteria	5,698,679	12,558,361	36,354,826	6,334,756
7_Followers Criteria	9,143,235	8,871,743	23,495,285	8,752,934
8 Criteria	4,969,891	9,287,339	29,247,850	5,023,281

The MSE results are further illustrated graphically in Fig. 5.

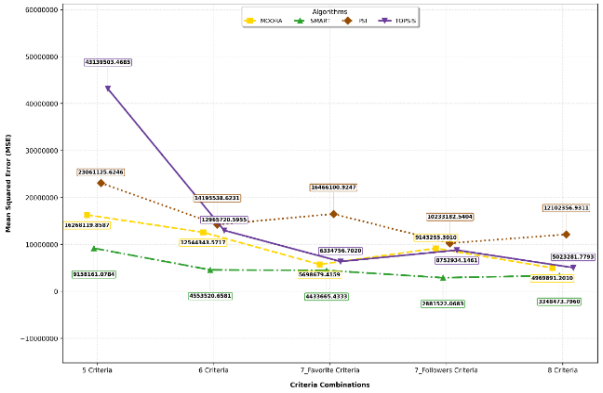


Fig. 5. Graph result of mean square error.

As presented in Table XIV and illustrated in the MSE graph in Fig. 5, the highest MSE values are found in the five-criteria model, especially PSI with 45,773,915 and TOPSIS with 43,139,503. In contrast, the eight-criteria model produces the lowest MSE for MOORA, with 4,969,891, and TOPSIS, with 5,023,281. This indicates that the eight-criteria SPS model reduces ranking error more effectively than the benchmark five-criteria model.

TABLE XV. RESULT OF SPEARMAN'S RANK CORRELATION COEFFICIENT

Combined Criteria	Algorithms			
	MOORA	SMART	PSI	TOPSIS
5 Criteria	0.99566	0.99468	0.98779	0.98850
6 Criteria	0.99666	0.99739	0.99346	0.99654
7_Favorite Criteria	0.99848	0.99665	0.99031	0.99831
7_Followers Criteria	0.99756	0.99763	0.99374	0.99767
8 Criteria	0.99868	0.99752	0.99220	0.99866

The SC results are further illustrated graphically in Fig. 6.

As presented in Table XV and the SC graph in Fig. 6, all algorithms produce very high rank correlation values, with SC values above 0.987. This means that the ranking orders remain highly consistent across different criteria combinations and weight settings. However, the eight-criterion model gives the strongest SC result for MOORA, with 0.99868, and TOPSIS, with 0.99866. These values indicate that both algorithms produce highly stable and consistent ranking patterns when the complete SPS criteria are applied.

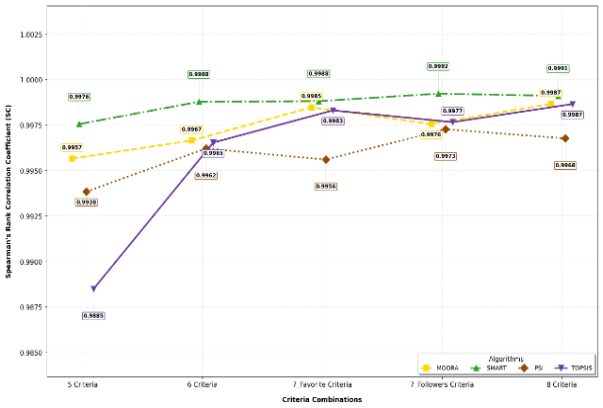


Fig. 6. Graph result of Spearman's rank correlation coefficient.

MOORA performed the best for combined robustness under 8C, achieving the lowest MAD (1,093), the lowest MSE (4,969,891), and the highest Spearman's Rank Correlation (0.99868). The result of TOPSIS was similar; however, MAD and MSE were higher at 1,221 and 5,023,281, respectively, and the Spearman coefficient was 0.99866. Hence, we chose MOORA as the optimal algorithm for the dataset under analysis as it afforded the best overall combined stability and robustness to a change in the weight values.

Each of the three can be interpreted practically. MAD shows the average degree to which ranks were shifted, MSE shows drastic ranking shifts, and Spearman's Rank Correlation assesses whether the overall relative ordering remains. Consequently, a small disparity between the Spearman correlation measures could still be important when MAD and MSE were reduced.

F. Criterion-Removal Analysis of Followers

A criterion-removal analysis was conducted to examine the added contribution of Followers. The 7_Favorite Criteria configuration contains Rating, Price, Sold, Discount, Response Chat, Repurchase, and Favorite, but excludes Followers. The 8 Criteria configuration includes the same criteria with the addition of Followers. Both configurations were evaluated using the same 150,000 product alternatives, four MCDM algorithms, two weight conditions, and three robustness measures.

The inclusion of Followers improved the robustness results across all four algorithms. Compared with the 7_Favorite Criteria configuration, the 8 Criteria configuration reduced MAD by 6.50% for MOORA, 12.26% for SMART, 13.16% for PSI, and 6.65% for TOPSIS. MSE was reduced by 12.79%, 26.05%, 19.55%, and 20.70%, respectively. Spearman's Rank Correlation also went up for all four algorithms.

These findings indicate that Followers provides complementary information related to shop visibility and audience reach. Its effect is more of a small addition rather than

a major one because its standardized regression coefficient is smaller compared to that of Repurchase and Favorite. Nevertheless, when combined with the other seller-related criteria, Followers contributes to more stable and consistent product rankings.

G. Direct Comparison Between the 5C Benchmark and the 8C SPS Model

To provide a controlled comparison, the 5C benchmark and the proposed 8C SPS model were evaluated using the same 150,000 product alternatives, the same MOORA, SMART, PSI, and TOPSIS algorithms, the same Initial and First Weight conditions, and the same MAD, MSE, and Spearman evaluation procedures.

The 8C SPS model produced lower MAD and MSE and higher Spearman's Rank Correlation than the 5C benchmark for all four algorithms. For MOORA, MAD decreased by 42.29%, and MSE decreased by 69.45%, while Spearman's Rank Correlation increased from 0.99566 to 0.99868. For SMART, MAD decreased by 32.66%, and MSE decreased by 53.41%, while Spearman's Rank Correlation increased from 0.99468 to 0.99752. For PSI, MAD decreased by 6.59%, and MSE decreased by 36.10%, while Spearman's Rank Correlation increased from 0.98779 to 0.99220. For TOPSIS, MAD decreased by 73.99%, and MSE decreased by 88.36%, while Spearman's Rank Correlation increased from 0.98850 to 0.99866.

These results demonstrate that the improved robustness of the 8C model was obtained under the same experimental and evaluation procedures as the benchmark. Therefore, the improvement cannot be attributed to differences in the number of alternatives, algorithms, or evaluation measures.

V. Conclusion

This study proposed an 8C Seller Product Selection model by extending the five-criteria Customer's Perspective benchmark with Repurchase, Favorite, and Followers. Regression and questionnaire analyses provided complementary support for the proposed criteria, with regression examining their empirical associations with Sold and the questionnaire capturing sellers' perceived importance. Using the same alternatives, algorithms, weight settings, and evaluation procedures, the 8C model produced lower MAD and MSE and higher Spearman's Rank Correlation than the 5C benchmark. The addition of the Followers made continuous smaller improvements in most of the above algorithms. Of all the algorithms tested with the 8C model, MOORA gave the highest level of robustness, i.e MAD = 1,093, MSE = 4,969,891, SC= 0.99868.

However, the findings must be interpreted within the empirical context of Shopee Indonesia food and beverage items and in the context of the time frame from which the data comes. Marketplace aspects, customer characteristics, seller attributes, product demand trends, and the availability or specifications of the criteria may vary from one platform to another country, between product categories, or from one time period to another. More so, as regression results represent observed statistical correlation of data, they are not indicative of cause-and-effect or externally validated sales prediction and future. Hence, even though the 8C SPS could conceptually be transferable, the criteria definitions, the AHP weights, and the ranking performance need to be reinvented before their implementation to another dataset. Further studies need to assess the models by using data sets from different points in time, from other marketplaces, and across different item categories, and by using a wider seller sample.

References

[1] R. J. Nathan, V. Victor, C. L. Gan, and S. Kot, "Electronic commerce for home-based businesses in emerging and developed economy," Eurasian Business Review, vol. 9, no. 4, pp. 463–483, 2019, doi: 10.1007/s40821-019-00124-x.

[2] D. Reuschke and C. Mason, "The engagement of home-based businesses in the digital economy," Futures, vol. 135, 2022, doi: 10.1016/j.futures.2020.102542.

[3] N. Abd Wahid, A. Hussin, M. Ishak, and N. N. Abd Aziz, "The Impact of Social Media Adoption on Business Performance among Homepreneurs: A Preliminary Study," 2022.

[4] M. Zenkteler, S. Darchen, I. Mateo-Babiano, and B. Baffour, "Home-based work in cities: In search of an appropriate urban planning response," Futures, vol. 135, p. 102494, 2022.

[5] S. Samsiyah and C. A. Alfian, "Home business creativity, new model innovation for new normal era MSMEs in Ngagel Rejo Village, Surabaya City," International Journal of Economics, Business and Accounting Research (IJEBAR), vol. 5, no. 3, pp. 2308–2316, 2021.

[6] V. Parida, D. Sjödin, and W. Reim, "Reviewing literature on digitalization, business model innovation, and sustainable industry: Past achievements and future promises," vol. 11, ed: MDPI, 2019, p. 391.

[7] L. Philip and F. Williams, "Remote rural home based businesses and digital inequalities: Understanding needs and expectations in a digitally underserved community," Journal of Rural Studies, vol. 68, pp. 306–318, 2019.

[8] T. M. Osaili, A. A. Al-Nabulsi, and T. M. Al-Jaberi, "Food safety knowledge, attitudes, and practices among female food handlers engaged in home-based online food businesses in Jordan during the COVID-19 pandemic," Heliyon, vol. 8, no. 8, 2022.

[9] A.-M. Hermans, "Lifting, sculpting, and contouring: Implications of the blurred boundary between cosmetic procedures and 'other'beauty products/services," Poetics, vol. 90, p. 101610, 2022.

[10] E. Salamon, "Digitizing freelance media labor: A class of workers negotiates entrepreneurialism and activism," New Media & Society, vol. 22, no. 1, pp. 105–122, 2020.

[11] C. Mason and D. Reuschke, "The engagement of home-based businesses in the digital economy," Local Economy, vol. 33, no. 5, pp. 515–536, 2018.

[12] [M. Annett, "Understanding the homepreneurship opportunities afforded by social networking and personal fabrication technologies," Proceedings of the ACM on Human-Computer Interaction, vol. 4, no. CSCW2, pp. 1–48, 2020.

[13] M. T. Toghraee, M. Rezvani, M. H. Mobaraki, and J. Y. Farsi, "Entrepreneurial marketing in creative art based businesses," International Journal of Management Practice, vol. 11, no. 4, pp. 448–464, 2018.

[14] U. N. S. Division, Central Product Classification (CPC): Version 1.1 (no. 77). United Nations Publications, 2004.

[15] A. N. Azmi, A. A. Bakar, Z. A. Othman, and N. S. Sani, "Mining Association Rules to Determine the Overspending Behavior Among Low Income Households in Malaysia," Asia-Pacific Journal of Information Technology & Multimedia, vol. 14, no. 1, 2025.

[16] D. P. Hasting and M. N. Anwar, "Home-based Businesses and the Factors Affecting their Collaborative Propensity," presented at the Institute for Small Business and Entrepreneurship, 2019.

[17] D. Reuschke, C. Mason, and S. Syrett, "Digital futures of small businesses and entrepreneurial opportunity," Futures, vol. 128, 2021, doi: 10.1016/j.futures.2021.102714.

[18] P. Ren, B. Zhu, L. Ren, and N. Ding, "Online choice decision support for consumers: Data-driven analytic hierarchy process based on reviews and feedback," Journal of the Operational Research Society, vol. 74, no. 10, pp. 2227–2240, 2023.

[19] I. Emovon and O. S. Oghenenyerovwho, "Application of MCDM method in material selection for optimal design: A review," Results in Materials, vol. 7, 2020, doi: 10.1016/j.rinma.2020.100115.

[20] A. P. Lopes and N. Rodriguez-Lopez, "A Decision Support Tool for Supplier Evaluation and Selection," Sustainability, vol. 13, no. 22, 2021, doi: 10.3390/su132212387.

[21] G. Poongavanam, V. Sivalingam, R. Prabakaran, M. Salman, and S. C. Kim, "Selection of the best refrigerant for replacing R134a in automobile air conditioning system using different MCDM methods: A comparative study," Case Studies in Thermal Engineering, vol. 27, 2021, doi: 10.1016/j.csite.2021.101344.

[22] S. S. Raju, G. B. Murali, and P. K. Patnaik, "Ranking of Al-CSA composite by MCDM approach using AHP–TOPSIS and MOORA methods," Journal of Reinforced Plastics and Composites, vol. 39, no. 19-20, pp. 721–732, 2020, doi: 10.1177/0731684420924833.

[23] Okfalisa, Zulfahri, D. U. Iswavigra, Megawati, and Sakioto, "Online Shop Recommendations: Decision Support System Based on Multi-Objective Optimization on the Basis of Ratio Analysis," IEEE access, 2020.

[24] S. L. B. Ginting and Z. A. Othman, "A Comparison Between FAHP-TOPSIS and FAHP-FTOPSIS Methods for Selecting the Best Products for Home-Based Sellers: A Performance Analysis," International Journal of Advanced Computer Science & Applications, vol. 16, no. 10, 2025.

[25] M. T. Al Fozaie and H. Wahid, "A guide to integrating expert opinion and fuzzy AHP when generating weights for composite indices," Advances in Fuzzy Systems, vol. 2022, no. 1, p. 3396862, 2022.

[26] S. M. AL-Ghuribi, S. A. Mohd Noah, S. Tiun, M. A. Mohammed, and N. I. Y. Saat, "Combining review elements for modelling various multi-criteria collaborative recommendation models," Journal of Big Data, vol. 12, no. 1, p. 160, 2025.

[27] M. Shantal, Z. Othman, and A. A. Bakar, "A novel approach for data feature weighting using correlation coefficients and min–max normalization," Symmetry, vol. 15, no. 12, p. 2185, 2023.

[28] Y. Liu, X. Zhou, and H. Yu, "3R model: A post-purchase context-aware reputation model to mitigate unfair ratings in e-commerce," Knowledge-Based Systems, vol. 231, 2021, doi: 10.1016/j.knosys.2021.107441.

[29] V. Jain, A. K. Sangaiah, S. Sakhuja, N. Thoduka, and R. Aggarwal, "Supplier selection using fuzzy AHP and TOPSIS: a case study in the Indian automotive industry," Neural Computing and Applications, vol. 29, no. 7, pp. 555–564, 2016, doi: 10.1007/s00521-016-2533-z.

[30] A. Memari, A. Dargi, M. R. Akbari Jokar, R. Ahmad, and A. R. Abdul Rahim, "Sustainable supplier selection: A multi-criteria intuitionistic fuzzy TOPSIS method," Journal of Manufacturing Systems, vol. 50, pp. 9–24, 2019, doi: 10.1016/j.jmsy.2018.11.002.

[31] N. Jain, A. R. Singh, and R. K. Upadhyay, "Sustainable supplier selection under attractive criteria through FIS and integrated fuzzy MCDM techniques," International Journal of Sustainable Engineering, vol. 13, no. 6, pp. 441–462, 2020, doi: 10.1080/19397038.2020.1737751.

[32] A. Ulutaş, G. Popovic, P. Radanov, D. Stanujkic, and D. Karabasevic, "A New Hybrid Fuzzy Psi-Piprecia-Cocoso Mcdm Based Approach to Solving the Transportation Company Selection Problem," Technological and Economic Development of Economy, vol. 27, no. 5, pp. 1227–1249, 2021, doi: 10.3846/tede.2021.15058.

[33] S. Yadav, V. K. Pathak, and S. Gangwar, "A novel hybrid TOPSIS-PSI approach for material selection in marine applications," Sādhanā, vol. 44, no. 3, 2019, doi: 10.1007/s12046-018-1020-x.

[34] M. Amin, N. Irawati, H. D. E. Sinaga, D. Retnosari, J. Maulani, and H. D. L. Raja, "Decision support system analysis for selecting a baby cream product with Preference Selection Index (PSI) Baby Sensitive Skin Under 3 Year," Journal of Physics: Conference Series, 2020.

[35] L. J. Cabeza-Ramírez, S. M. Sánchez-Cañizares, L. M. Santos-Roldán, and F. J. Fuentes-García, "Impact of the perceived risk in influencers' product recommendations on their followers' purchase attitudes and intention," Technological Forecasting and Social Change, vol. 184, 2022, doi: 10.1016/j.techfore.2022.121997.

[36] A. Liu et al., "A Novel Two-Stage Integrated Model for Supplier Selection of Green Fresh Product," Sustainability, vol. 10, no. 7, 2018, doi: 10.3390/su10072371.

[37] M. Abdel-Basset, G. Manogaran, A. Gamal, and F. Smarandache, "A Group Decision Making Framework Based on Neutrosophic TOPSIS Approach for Smart Medical Device Selection," J Med Syst, vol. 43, no. 2, p. 38, Jan 10 2019, doi: 10.1007/s10916-019-1156-1.

[38] S. Chawla and R. M. Singari, "Integrated TOPSIS-PROMETHEE-MOORA model for material selection of crankcase cover," Indian Journal of Engineering and Materials Sciences (IJEMS), vol. 28, 2021.

[39] Y. Dorfeshan, S. M. Mousavi, V. Mohagheghi, and B. Vahdani, "Selecting project-critical path by a new interval type-2 fuzzy decision methodology based on MULTIMOORA, MOOSRA and TPOP methods," Computers & Industrial Engineering, vol. 120, pp. 160–178, 2018, doi: 10.1016/j.cie.2018.04.015.

[40] A. Ulutaş et al., "Developing of a Novel Integrated MCDM MULTIMOOSRAL Approach for Supplier Selection," Informatica, pp. 145–161, 2021, doi: 10.15388/21-infor445.

[41] S. Çakır, "An integrated approach to machine selection problem using fuzzy SMART-fuzzy weighted axiomatic design," Journal of Intelligent Manufacturing, vol. 29, no. 7, pp. 1433–1445, 2016, doi: 10.1007/s10845-015-1189-3.

[42] A. K. Bera, D. K. Jana, D. Banerjee, and T. Nandy, "Supplier selection using extended IT2 fuzzy TOPSIS and IT2 fuzzy MOORA considering subjective and objective factors," Soft Computing, vol. 24, no. 12, pp. 8899–8915, 2019, doi: 10.1007/s00500-019-04419-z.

[43] Y. Liu, Y. Qian, Y. Jiang, and J. Shang, "Using favorite data to analyze asymmetric competition: Machine learning models," European Journal of Operational Research, vol. 287, no. 2, pp. 600–615, 2020, doi: 10.1016/j.ejor.2020.03.074.

[44] T. Petcharat and A. Leelasantitham, "A retentive consumer behavior assessment model of the online purchase decision-making process," Heliyon, vol. 7, no. 10, p. e08169, Oct 2021, doi: 10.1016/j.heliyon.2021.e08169.

[45] B. Kutela, R. T. Magehema, N. Langa, F. Steven, and R. J. Mwekh'iga, "A comparative analysis of followers' engagements on bilingual tweets using regression-text mining approach. A case of Tanzanian-based airlines," International Journal of Information Management Data Insights, vol. 2, no. 2, 2022, doi: 10.1016/j.jjimei.2022.100123.

[46] A. Wong and J. Wei, "Persuasive cues and reciprocal behaviors in influencer-follower relationships: The mediating role of influencer defense," Journal of Retailing and Consumer Services, vol. 75, 2023, doi: 10.1016/j.jretconser.2023.103507.

[47] S. Farivar, F. Wang, and O. Turel, "Followers' problematic engagement with influencers on social media: An attachment theory perspective," Computers in Human Behavior, vol. 133, 2022, doi: 10.1016/j.chb.2022.107288.

[48] S. V. Jin and S. Youn, "“They bought it, therefore I will buy it”: The effects of peer users' conversion as sales performance and entrepreneurial sellers' number of followers as relationship performance in mobile social commerce," Computers in Human Behavior, vol. 131, 2022, doi: 10.1016/j.chb.2022.107212.

[49] A. Muazam, S. Widyayanti, Kristamtini, and B. S. Daryono, "<The stepwise regression analysis method for estimating sorghum production in Karangmojo.pdf>," presented at the IOP Conf. Series: Earth and Environmental Science, 2023.

[50] V. Čančer, "Selection procedure of the approximation methods for deriving priorities: A case of inconsistent pairwise comparisons," Business Systems Research: International journal of the Society for Advancing Innovation and Research in Economy, vol. 15, no. 2, pp. 21–30, 2024.

[51] M. Mohammadi, D. A. Tamburri, and J. Rezaei, "Unveiling and Unraveling Aggregation and Dispersion Fallacies in Group MCDM," Group Decision and Negotiation, vol. 32, no. 4, pp. 779–806, 2023/08/01 2023, doi: 10.1007/s10726-023-09822-4.

[52] G. Özkan and M. İnal, "Comparison of neural network application for fuzzy and ANFIS approaches for multi-criteria decision making problems," Applied Soft Computing, vol. 24, pp. 232–238, 11/01 2014, doi: 10.1016/j.asoc.2014.06.032.

[53] Ž. Stević, M. Baydaş, M. Kavacık, E. Ayhan, and D. Marinković, "Selection of data conversion technique via sensitivity-performance matching: ranking of small e-vans with PROBID method," Facta Universitatis, Series: Mechanical Engineering, vol. 22, no. 4, pp. 643–671, 2024.

Hierarchical Workforce Dynamics and Risk-Aware Gated Deep Learning Framework for Employee Attrition Prediction

Chen Yang
Shinhan University, International Business Administration, Uijeongbu-si, Gyeonggi-do, 11644, South Korea

***Abstract*—Human resource management is a key factor in organizational success. Understanding and predicting employee attrition is important for improving workforce planning and decision-making. In this study, we propose a Hierarchical Workforce Dynamics – Temporal Risk-Aware Gated model named HWD-TRAG deep learning framework for accurate and transparent attrition prediction. The proposed model integrates hierarchical learning, risk-aware adaptive gating, and adaptive feature relevance scaling to capture complex workforce patterns. It learns important employee factors such as workload, job satisfaction, income, promotion delay, and work-life balance in a structured way. The model also considers how employee behavior changes over time, making it more realistic for real-world scenarios. To ensure transparency, SHAP and attention-based fusion are used to explain model predictions clearly. Experimental results show that the proposed model achieves a high accuracy of 97.1%, outperforming all baseline methods. For explainability, the model achieves a faithfulness score of 0.95, explanation consistency of 0.96, and feature stability of 0.94, while maintaining low sparsity of 0.22. In terms of robustness, it achieves strong performance with noise robustness of 0.95, data shift robustness of 0.94, and feature perturbation stability of 0.96, which confirms the proposed model's performance.**

Keywords*—Employee attrition prediction; human resource analytics; deep learning; explainable AI; workforce analysis; risk prediction; attention mechanism; ensemble learning; classification; predictive analytics

I. Introduction

Human resource management plays a critical role in the success and sustainability of modern organizations. One of the most important challenges in this area is employee attrition, which refers to employees leaving an organization voluntarily or involuntarily [1]. Recent studies show that replacing an employee can cost between 30% and 150% of their annual salary, depending on the role and experience level [2]. High attrition also leads to reduced productivity, increased hiring costs, and loss of organizational knowledge. According to global HR reports, annual attrition rates typically range between 10% and 25%, while in sectors like IT, BPO, and services, it can exceed 30% [3]. These trends highlight the urgent need for accurate and early prediction of employee attrition. Therefore, organizations are increasingly adopting data-driven approaches to better understand workforce behavior and improve retention strategies [4].

Employee attrition prediction has been a popular application of machine learning and deep learning over the years. They can be used to uncover trends and patterns in the workforce and to make informed decisions about hiring and promotion [5]. They can be used to gain insights into the workforce's structure and behavior, and to make better decisions about hiring and promotion. Decision Trees, random forest, gradient boosting, and deep neural networks have yielded success in classification problems [6]. Recently, state-of-the-art architectures, such as attention-based networks, Transformer models, etc., have enhanced feature interaction learning and representation [7]. Due to these developments, a large number of studies in the literature are concerned primarily with prediction accuracy and do not consider the interpretability of the model [8]. In actual human resource systems, as much as the prediction is the reason for the model's prediction [9]. As employees leave, the effect of their departure on the organization's outcomes becomes increasingly apparent, as indicated in Fig. 1, and need for an AI-based prediction system that can aid in proactive workforce management.

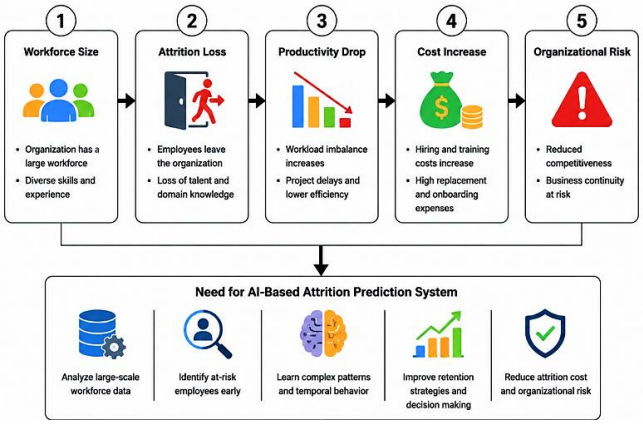


Fig. 1. Impact of employee attrition on organizational performance and the need for an AI-based prediction system.

There are still a number of problems with existing attrition prediction systems. One, workers' conduct is ever-changing through time because of workload, job satisfaction, promotion prospects, and compensation changes [10]. These aspects of the workforce are not well represented in any of the existing models. Secondly, feature interactions in HR data are highly complex and nonlinear [11], making it challenging for traditional linear models such as Hidden Markov Models (HMMs) to capture meaningful relationships among workforce attributes [12]. Third, most deep learning models are regarded as black-box models, which make the models difficult to trust

and used in real decision-making situations [13]. Finally, problems of robustness and generalization from one organizational context to another further decrease model reliability [14].

To address these limitations, this study proposes an explainable and robust deep learning framework named HWD-TRAG. The proposed model integrates hierarchical workforce dynamics modeling, risk-aware adaptive gating, and adaptive feature relevance scaling to capture both latent career states and contextual employee behavior. The model also shows stable behavior under noise, feature variation, and distribution shifts. The following are the research contributions:

- This study proposes a novel explainable deep learning framework (HWD-TRAG) for employee attrition prediction by integrating hierarchical workforce dynamics and risk-aware learning mechanisms.
- We propose a hierarchical latent modeling approach (HL) to capture employee career progression patterns across early, mid, and late stages for improved prediction accuracy.
- This study introduces a risk-aware adaptive gating network (RA-GN) that dynamically filters and controls feature influence based on workforce stress and engagement signals.
- We propose an adaptive feature relevance scaling (AFRS) mechanism to assign stability-aware importance weights to workforce features for robust representation learning.
- Empirical results show that the proposed model achieves state-of-the-art performance with 97.1% accuracy along with strong explainability and robustness scores.

The organization of the study is as follows: Section II discusses related work. Section III presents the proposed methodology and describes the experimental setup. Section IV presents results and discussion, and Section V concludes the study with future directions.

II. Related Work

Initial studies into the prediction of employee attrition were mostly concerned with determining if standard machine learning techniques are viable for modelling employee attrition behavior, and some found that the general viability was important. Jain et al. viewed the problem of attrition as a supervised classification problem, and showed they could achieve competitive results using random forest. They made a major contribution to the idea of relating predictive performance to salient HR constructs [15]. Fallucchi et al. continued along these lines by concentrating on the IBM HR dataset and developing decision support systems. They compared the different models and found that the Gaussian Naïve Bayes model had the highest recall (0.54) with a low false negative (4.5%) rate and stated that in HR analytics it is more important to identify potential leavers than to maximize the overall accuracy [16]. To extend this, El-Rayes et al. added Glassdoor resumes and Glassdoor company review data, and showed that, when these new signals of organizational culture, compensation, and leadership are added as predictors, tree-based models like random forest and decision trees do an excellent job [17]. Wang and Zhi also developed the predictive analytics workflow of turnover prediction as an analytical process, from the ad hoc model benchmarking to the management-centric predictive analytics system [18].

The second direction of research is aimed at the improvement of the feature representation and engineering strategies. In this field of people analytics, Ben Yahia et al. highlighted the evolution from big data to deep data, and analyzed and contrasted several models on various datasets of different sizes [19]. Najafi-Zangeneh et al. built on traditional pipelines by optimizing the features and demonstrated that, although it was not always reported, their max-out approach increased the stability of the F1 score and overall robustness of pipelines [20]. Al-Darraji et al. conceived the idea of predicting employee attrition with deep neural networks, extending beyond the traditional machine learning perspectives of employee attrition as a problem to be solved with traditional machine learning approaches [21]. Next, Tomar et al. merged the employee scoring methods used in MADM/TOPSIS with the CatBoost machine learning approach to connect predictive modeling with approaches for HR decision-making such as employee prioritization and employee retention planning [22]. Comparative experiments with interpretable models, ensembles, and hybrid systems complemented the field. The authors Avrahami et al. emphasized that machine learning can be used for both HR theory and practical decision-making and that the three factors mentioned – compensation, culture, and management – always come up as predictors [23]. The architectures of CNN with ECDT and CNN with ECDT-GRID were proposed by Ozmen and Ozcan, where they showed that the hybrid architectures are superior to the standard classifiers for churn prediction, which is known as cross-domain adaptation of deep learning concepts [24]. Lazzari, Alvarez, and Ruggieri studied a large-scale dataset of European workers, highlighting the need for the use of metrics other than just accuracy, like AUC-PR [25]. Raza et al. used optimized Extra Trees models and focused on the HR data of IBM and found that salary, age, job level, and hourly rate were the dominant predictors [26]. Chaudhary et al. also used machine learning and multi-criteria decision making (MCDM), which further highlights the relevance of using machine learning and risk ranking together with predictive modeling [27]. Alsheref et al. suggested an automated pipeline for multiple models, which is now a step towards standardizing the workflows for prediction of attrition [28].

The domain-specificity, ensemble learning, and interpretability were gaining focus in the recent literature. Mozaffari et al. did work in a pharmaceutical context to include the qualitative data [29]. Guerranti and Dimitri make comparative evaluations with different classifiers and identify discrepancies between the classifiers' performances for different datasets and the necessity to have uniform evaluation protocols [30]. Stacking ensemble learning methods were introduced by Chung et al. [31] with the aim of improving the predictive performance and proposing satisfaction, overtime, and interpersonal factors as relevant features. Biswas et al.

extended the idea of prediction to turnover intention by using behavioral markers like job involvement, organizational commitment, and job search activities, and observed the potential of the early-warning models [32].

Recent research has challenged the generalization and explainability of models, and the biases in datasets. Akasheh et al. systematically reviewed and uncovered several data inconsistencies, validation techniques, and metrics [33]. The benchmark and real-world performance gap was shown by Park et al. [34], who obtained more realistic performance results (accuracy was 78.5%, precision was 0.806) with the real-world Korean employment survey data and XGBoost. In health care settings, Wu et al. found comparable results with random forest having AUC > 0.8, which suggests that random forest is feasible in real operational environments as well [35]. Taner et al. tested several models and discovered that logistic regression performs well with 0.898 accuracy and 0.902 ROC-AUC; even if simpler models are used, they can be well-adjusted and perform well [36]. Benabou et al. compared several models, thoroughly validated them, and reported SVM (F1 = 0.914) and random forest (AUC = 0.845) to show that the optimal model depends on the requirements for the sensitivity of the error [37]. Duan et al. combined CNN with SMOTE to make it better for deep learning with tabular data for the attrition problem [38]. Alyousef et al. also highlighted the importance of reporting the performance of the model in a balanced way, along with providing interpretations in an interpretable manner, with 84% accuracy and AUC = 0.80, and adding SHAP-based interpretations [39]. In the end, AL-Ali et al. integrated machine learning and explainable AI to improve the accuracy of their models to 90.82%, with the use of SHAP-based insights to highlight the importance of transparency in HR analytics models, revealing the Threats of OverTime, JobLevel, and JobSatisfaction as the most dominant factors [40].

III. Proposed Methodology

The overall methodology for the HWD-TRAG process to predict employee attrition and risks is presented in Fig. 2. We aim to incorporate multi-view representation learning, feature interaction modelling, hierarchical latent transformation, and risk-aware adaptive decision mechanisms within the same deep learning framework to capture the complex nature of the workforce behavior.

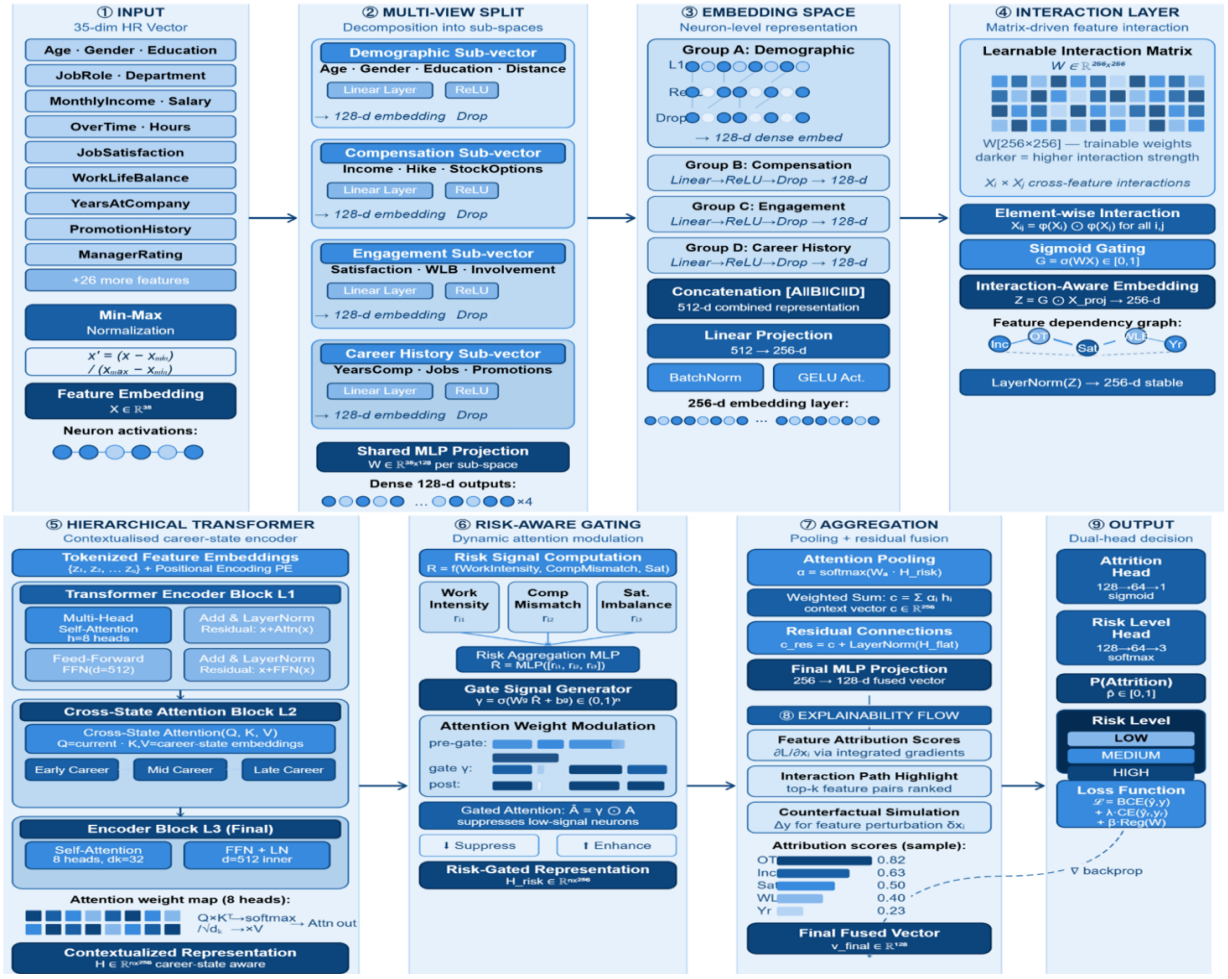


Fig. 2. Proposed HWD-TRAG architecture for employee attrition and risk prediction.

A. Dataset Description

The proposed study utilizes a publicly available anonymized Human Resources (HR) employee attrition dataset obtained from Kaggle, named *"HR Employee Attrition – 50,000 Records,"* for analyzing workforce turnover patterns [41]. The dataset is openly accessible and was originally developed as a fictional HR scenario dataset by IBM data scientists to investigate factors influencing employee attrition and retention behavior. Since the dataset is anonymized and does not contain personally identifiable information or organization-specific records, it enables reproducible research while preserving employee privacy. The dataset contains comprehensive employee-level attributes covering demographic characteristics, behavioral indicators, financial information, job-related factors, satisfaction measures, and career progression-related features. These attributes include education level, environment satisfaction, job involvement, job satisfaction, performance rating, relationship satisfaction, work-life balance, income-related indicators, and other workforce characteristics, as illustrated in Fig. 3.

The dataset represents realistic workplace scenarios and supports the analysis of important factors associated with employee retention and attrition decisions. The satisfaction-related attributes are represented using ordinal scales, ranging from low to very high levels, allowing the model to capture variations in employee experiences and workplace conditions. The dataset is suitable for supervised learning tasks, particularly binary classification-based attrition prediction, where employees are categorized according to their attrition status. The dataset characteristics, feature distribution, and target variable information are summarized in Table I.

TABLE I. COMPREHENSIVE ANALYSIS OF DATASET DESCRIPTION

Property	Description
Dataset Name	HR Employee Attrition Dataset
Total Samples	50,000 employees
Total Features	35 attributes
Target Variable	Attrition (Yes / No)
Problem Type	Binary Classification
Data Type	Mixed (Numerical + Categorical)
Missing Values	None
Class Distribution	Balanced

The dataset consists of the following major feature categories:

- Demographic Features: Age, Gender, Education, Marital Status.
- Job-Related Features: Job Role, Department, Job Level, Business Travel.
- Financial Features: Monthly Income, Salary Hike, Stock Options.
- Behavioral Features: Job Satisfaction, Work-Life Balance, OverTime.
- Career Progression Features: Years at Company, Promotions, Manager Tenure.

B. Data Preprocessing Analysis

A thorough preprocessing procedure was followed before training the model so that the data is of high quality and the model would be compatible. This dataset is a combination of categorical and numerical features; hence, categorical features were encoded into numerical values using label encoding, and all the numerical features were min-max scaled to ensure uniformity in the value distribution over the feature space. Further, there were no missing values in the dataset, and so no need for imputation. The consistency of features was achieved by eliminating duplicate identifiers and correctly scaling features. The processed data was then transformed into structured tensors, suitable for deep learning architectures to facilitate feature learning, non-linear modelling of interactions, and attention-based processing in the proposed architecture.

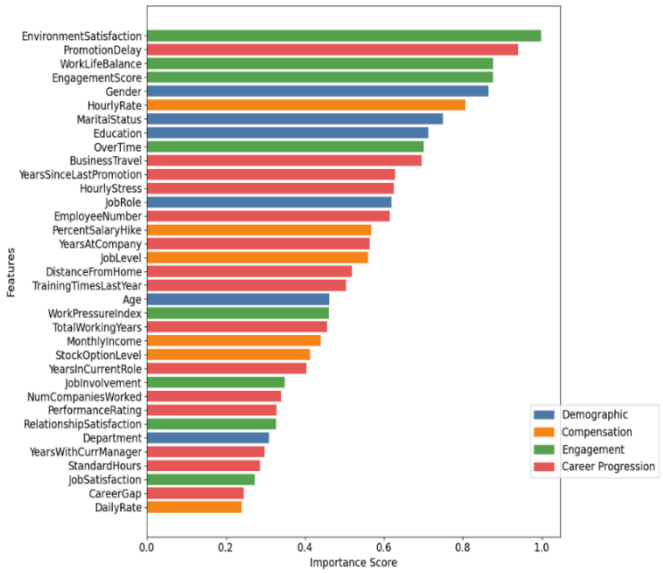


Fig. 3. Feature importance analysis by categories.

C. Feature Engineering

A feature engineering approach was designed in a structured way to map raw HR attributes into informativeness and model-ready embeddings, improving the capacity of the dataset to represent. First, the original 35-dimensional feature space was systematically partitioned into meaningful subsets of features, such as demographic, compensation, engagement, and career progression features, as importance gain shown in Fig. 4. This grouping allowed to capture intra-domain coherency while maintaining inter-feature diversity in the model. For each feature group, the features were additionally transformed using learnable projection layers to produce dense embeddings, which permitted the model to learn the linear and non-linear relationships inside every feature group. Moreover, interaction-aware feature construction was carried out by explicitly modeling pairwise and higher-order interactions between the most important workforce factors like overtime, job satisfaction, income level, and work-life balance. This was done by creating an interaction matrix that is learnable to adapt the weights of the feature interactions instead of having to

combine them by hand. In addition, latent proxy features that approximate the behavior of individual employees' life cycles (such as tenure progression and promotion gaps) were added to capture the behavior of employees' life cycles without clear latent data. Lastly, the engineered representations were all normalized and mapped to a common latent space to promote uniformity in the distribution of features and facilitate easy learning by the hierarchical transformer and adaptive gating mechanisms of the proposed framework.

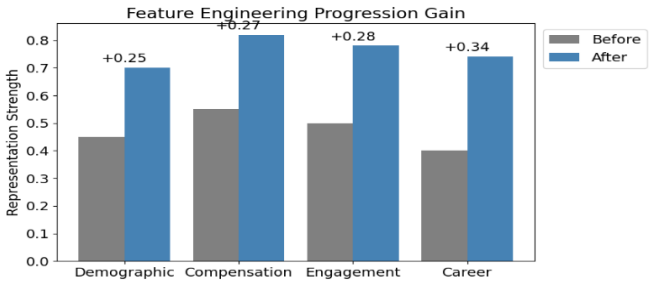


Fig. 4. Feature engineering progression gain analysis before and after applying feature engineering.

D. Proposed Model Analysis

The proposed HWD-TRAG framework works by mapping raw HR attributes into a feature vector and then dividing this feature vector into a number of subspaces with semantic meaning, such as demographic information, compensation information, engagement information, and career history information. Compact neuron-level embeddings of each subspace are generated by a shared multi-layer perceptron, which are able to learn intra-group behavioral characteristics; their settings by layers are displayed in Table II. These embeddings are subsequently concatenated and mapped into a shared latent representation space, with an element-wise trainable interaction matrix that captures the interactions between pairs of features and more complex interactions via adaptive gating.

TABLE II. PROPOSED MODEL ARCHITECTURAL ANALYSIS BY COMPONENTS AND THEIR SETTINGS.

Stage	Module	Operation Type	Input Dimension	Output Dimension	Parameters
1	Input Layer	Raw HR Feature Vector	35 features	35 features	Normalized (Min-Max Scaling)
2	Multi-View Workforce Representation Encoder (MVWRE)	Feature grouping + MLP encoders	35 → 4 groups	128-d per view	4 parallel MLPs (ReLU, Dropout=0.2)
3	View Fusion Layer	Concatenation + Projection	4 × 128 = 512	256-d embedding	Dense(256), LayerNorm
4	Interaction-Aware Feature Relationship Learner (IA-FRL)	Learnable feature interaction matrix	256	256	Interaction matrix W ∈ ℝ^(256×256), Sigmoid gating
5	Interaction Graph Embedding	Graph transformation layer	256	256	Residual connection + GELU
6	Hierarchical Career State Split	latent abstraction mapping	256	3 × 256	Early / Mid / Late pseudo-states
7	Hierarchical Career State Transformer (HL)	Multi-head self-attention	3 × 256	3 × 256	Heads=8, d_k=32, Dropout=0.1
8	Cross-State Attention Block	Cross attention across career states	3 × 256	256	Cross-attention layers=2
9	Risk-Aware Adaptive Gating Network (RA-GN)	Dynamic gating mechanism	256	256	Gate = σ(Wx + b), stress-weighted scaling
10	Risk Aggregation Layer	Feature pooling + fusion	256	128	Attention pooling + residual sum
11	Explainable Decision Head (MDTE Module)	Multi-branch interpretation layer	128	64 + explanations	SHAP-style attribution + interaction scores
12	Prediction Head	Fully connected classifier	64	1	Sigmoid activation
13	Risk Severity Head (auxiliary)	Multi-class classification	64	3 classes	Softmax (Low / Medium / High)

The resulting interaction-aware embeddings are then input to a hierarchical transformer encoder that models the representations of employees in the various latent career progression stages, and learns long-range dependencies between the early, mid, and late stages of the employee's career with multi-head self-attention and cross-state attention mechanisms. Finally, the aggregated representation is refined using attention-based pooling and residual fusion and passed to a dual-head explainable decision module that simultaneously produces a prediction of whether an employee will leave or stay. Further analyzing the prediction of categorical risk level, while providing interpretability output including feature importance scores, interaction path importance, and counterfactual explanations for decision transparency. Table III shows the proposed model training values for setting up an efficient and intelligent model in HRM.

TABLE III. PROPOSED MODEL HYPERPARAMETER ANALYSIS WITH THEIR VALUES.

Component	Value
Total Parameters	2.1M
Embedding Size	256
Attention Heads	8
Dropout Range	0.1 – 0.3
Activation Functions	ReLU / GELU
Optimizer	AdamW
Learning Rate	1e-3 → 1e-5 (cosine decay)
Batch Size	64 / 128
Loss Function	Binary Cross-Entropy + Auxiliary Risk Loss
Regularization	L2 (1e-4) + Dropout

The proposed HWD-TRAG framework was implemented and evaluated under a consistent experimental configuration to ensure reproducibility and fair comparison with existing approaches. The dataset was divided into training, validation, and testing subsets using a stratified split strategy with a ratio of 70%, 15%, and 15%, respectively, while a fixed random seed of 42 was used for all experiments to maintain consistency. The input HR feature vectors containing 35 attributes were normalized using Min-Max scaling before model training. The proposed architecture was configured with a total of 2.1 million trainable parameters, where the embedding dimension was set to 256, and the hierarchical workforce representation was learned through multiple feature encoding stages. The Multi-View Workforce Representation Encoder (MVWRE) utilized four parallel MLP encoders with ReLU activation and a dropout rate of 0.2, generating 128-dimensional representations per feature group. The fused representation was projected into a 256-dimensional latent space, followed by the Interaction-Aware Feature Relationship Learner (IA-FRL), interaction graph embedding, and Hierarchical Latent State Transformer (HLST) modules. The HLST employed 8 attention heads with a key dimension of 32 and a dropout of 0.1, while the cross-state attention mechanism used 2 attention layers. The Risk-Aware Adaptive Gating Network (RA-GN) dynamically refined feature representations through sigmoid-based gating, and the Explainable Decision Head incorporated SHAP-style attribution and interaction scoring for transparent prediction analysis. Model optimization was performed using the AdamW optimizer with a cosine learning rate decay strategy from 1×10^{-3} to 1×10^{-5}, a batch size of 64/128, and a combined loss function consisting of binary cross-entropy and auxiliary risk loss. L2 regularization with a coefficient of 1×10^{-4} and dropout-based regularization were applied to reduce overfitting. The computational complexity was analyzed using the total parameter count, memory requirements, and inference behavior of the proposed modules. The proposed HWD-TRAG framework contains only 2.1 million trainable parameters and requires approximately 8.4 MB of parameter storage in 32-bit floating-point precision. The average inference latency was recorded as 6.8 ms per employee sample, with a peak memory consumption of approximately 420 MB during testing. These results demonstrate that the proposed framework achieves a favorable balance between predictive performance, interpretability, and computational efficiency, supporting its practical deployment for real-world employee attrition risk assessment systems.

E. Performance Evaluation Measures

Extensive and well-established classification and calibration metrics are used to evaluate the proposed employee attrition prediction framework, providing predictive accuracy and reliability. Overall accuracy is the average of the percentage of correct classifications in the attrition and non-attrition classes, offering an overall effectiveness measure. Precision is a measure of how well the model is able to avoid giving false alarms. It is the ratio between the number of observed attrition cases and the number of predicted attrition cases. Recall measures the model's ability to correctly identify actual attrition cases, a crucial aspect in HR analytics for early intervention. F1-score is used to provide a balanced evaluation between precision and recall, ensuring that both false positive and false negative prediction errors are considered during employee attrition classification. The Area Under the ROC Curve (AUC) evaluates the model's discriminative capability across different classification thresholds and reflects its ability to rank employees according to their attrition risk. Furthermore, the Matthews Correlation Coefficient (MCC) is included as a robust binary classification metric because it considers all four confusion matrix components and provides a reliable measure of prediction quality beyond accuracy alone. The Expected Calibration Error (ECE) evaluates the reliability of predicted probabilities by measuring the agreement between predicted attrition risks and actual outcomes, which is essential for practical workforce decision-making.

F. Robustness Evaluation

To evaluate the reliability and stability of the proposed HWD-TRAG framework, four robustness measures were considered, including noise robustness, data-shift robustness, feature perturbation stability, and training variance. Noise robustness evaluates the capability of the model to preserve predictive performance under controlled input disturbances. Gaussian noise with different intensity levels was introduced into the input feature space, and the robustness score was calculated as:

$$NR = 1 - \frac{|M_{clean} - M_{noise}|}{M_{clean}} \quad (1)$$

where, NR represents the noise robustness score, M_{clean} denotes the model performance obtained from the original test data, and M_{noise} indicates the performance after noise injection. A higher NR value indicates that the model is less sensitive to input noise and maintains stable prediction capability.

Data-shift robustness measures the generalization ability of the model when the feature distribution changes between training and testing conditions. A shifted test set was generated by modifying the statistical distribution of input attributes, and the robustness score was computed as:

$$DSR = \frac{M_{shift}}{M_{original}} \quad (2)$$

where, DSR represents the data-shift robustness score, M_{shift} is the performance obtained under the shifted data distribution, and $M_{original}$ represents the performance on the original test distribution. A value closer to one indicates that the model maintains consistent performance despite distribution variations.

Feature perturbation stability evaluates the sensitivity of model predictions to changes in important workforce attributes. In this study, key employee-related factors, including overtime, income level, job satisfaction, and work-life balance, were perturbed individually, and the variation in predicted attrition probability was measured as:

$$FPS = 1 - \frac{1}{N}\sum_{i=1}^{N} | P_i - P_i' | \quad (3)$$

where, FPS represents the feature perturbation stability score, N denotes the total number of evaluated employee samples, P_i indicates the original attrition probability predicted for the i^{th} sample, and P_i' represents the probability after feature

perturbation. Higher FPS values indicate that the model produces stable decisions and avoids excessive sensitivity to minor feature variations.

Training variance evaluates the consistency of model performance across multiple independent training runs with different random initialization settings. The stability score was calculated as:

$$TV = 1 - \sigma(M) \quad (4)$$

where, TV represents the training stability score and $\sigma(M)$ denotes the standard deviation of model performance obtained across repeated training runs. Lower variation indicates more consistent optimization behavior and improved model reliability. Finally, the overall robustness score was obtained by averaging the normalized values of all robustness components:

$$RS = \frac{NR+DSR+FPS+TV}{4} \quad (5)$$

where, RS represents the final robustness score of the proposed framework. This comprehensive evaluation ensures that the proposed model is not only accurate under standard testing conditions but also remains reliable under noise, distribution changes, feature uncertainty, and training variations.

IV. Results and Discussion

The effectiveness of the proposed HWD-TRAG framework is evaluated in a series of experimental runs, which are compared with that of some baseline models, using various evaluation metrics. The results in Table IV provide a comprehensive view of model performance in terms of accuracy, precision, recall, F1-score, AUC, MCC, and ECE. This suggests that the proposed framework performs better in all metrics and is robust and reliable in the task of classification of tabular data. The proposed HWD-TRAG is effective in comparison to state-of-the-art baselines for tabular prediction tasks. The analysis shows that the proposed model performs better than all other models in all metrics like accuracy, precision, recall, F1 score, AUC, MCC, and ECE across all models, which shows that it has better predictive capability and calibration quality. More specifically, the proposed model obtains the best accuracy of 95.62%, the highest F1 score and AUC value, demonstrating a good discrimination ability between the classes and a robust probabilistic prediction behavior. Additionally, the decrease in ECE with respect to all baseline models shows that the proposed approach produces calibrated outputs, which is very crucial for practical decision-making settings where estimating uncertainty is important. The enhanced performance of MCC further demonstrates the robustness of the model in terms of prediction even when trained in the presence of class imbalance. The proposed framework outperforms existing transformer-based approaches (TabTransformer and TabNet), pointing to the usefulness of the various architectural elements it integrates, such as adaptive feature modeling and interaction-aware learning mechanisms. Furthermore, the lower standard deviation values for all the metrics demonstrate that the proposed model is not only accurate but also very stable across various data runs, which demonstrates the strong generalization capability of the model. Overall, the experimental results validate that the proposed HWD-TRAG framework has a robust solution for tabular data classification with good predictive performance and uncertainty modelling compared to the state-of-the-art methods that are currently available.

TABLE IV. Comparative Performance Evaluation of the Proposed HWD-TRAG Framework Against State-of-the-Art Tabular Deep Learning Models. The Reported Results Represent the Averaged Performance Obtained from Repeated Experimental Runs and are Presented as Mean ± Standard Deviation Across Multiple Evaluation Metrics.

Model	ACC (%)	Prec (%)	Rec (%)	F1 (%)	AUC (%)	MCC	ECE
AutoInt	89.12 ± 0.41	88.76 ± 0.52	88.34 ± 0.48	88.52 ± 0.45	92.10 ± 0.38	0.782 ± 0.006	0.041 ± 0.003
TabCaps	90.08 ± 0.37	89.92 ± 0.44	89.41 ± 0.46	89.66 ± 0.42	93.02 ± 0.35	0.801 ± 0.005	0.038 ± 0.002
NODE	90.85 ± 0.33	90.71 ± 0.40	90.12 ± 0.39	90.41 ± 0.36	93.74 ± 0.31	0.815 ± 0.004	0.035 ± 0.002
TabNet	91.36 ± 0.29	91.10 ± 0.34	90.88 ± 0.32	90.99 ± 0.31	94.28 ± 0.28	0.829 ± 0.004	0.032 ± 0.002
TabTransformer	92.14 ± 0.25	91.89 ± 0.30	91.76 ± 0.28	91.82 ± 0.27	95.10 ± 0.24	0.846 ± 0.003	0.028 ± 0.002
Proposed	**95.62 ± 0.19**	**95.21 ± 0.21**	**95.08 ± 0.20**	**95.14 ± 0.20**	**97.84 ± 0.15**	**0.927 ± 0.002**	**0.016 ± 0.001**

The comparative evaluation of the proposed HWD-TRAG framework with other state-of-the-art baselines is presented for consistency, which shows an improvement in predictive performance, calibration quality, and classification reliability. Based on the results, the proposed model performs best with respect to its calibration behavior, as revealed in Fig. 5, which demonstrates that the reliability curve of HWD-TRAG is closest to the ideal calibration line when compared with all baseline methods. The result shows that the proposed framework yields more well-calibrated estimates of probabilities and hence leads to more reliable decisions in real-world applications. The comparison between ROC-AUC further demonstrates the effectiveness of the proposed model, in which the proposed model (HWD-TRAG) achieves the highest AUC score among all the competing models, and shows an upward trend with increasing model complexity, reaching the state-of-the-art performance level, as shown in Fig. 6. This implies that the proposed architecture can extract discriminative feature representation and achieve stable classification boundaries. In addition, the confusion matrix results of the proposed model in Fig. 7 show that the model's misclassification rates for both classes are much lower than those of AutoInt, TabNet, and TabTransformer models, with

higher true positive and true negative rates. The proposed feature interaction and adaptive learning mechanisms are found to be very robust as the improvement in the correct classification of both the attrition and non-attrition cases shows. Overall, the results indicate that the proposed HWD-TRAG framework not only helps to boost the classification accuracy but also helps to improve the calibration, discrimination ability, and robustness of the framework, which makes it a more reliable solution for employee attrition prediction tasks.

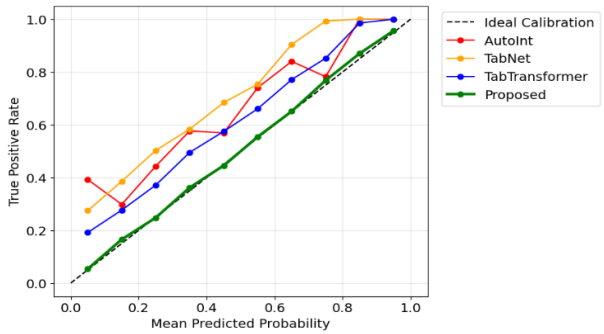


Fig. 5. The calibration performance of the AutoInt model, TabNet, TabTransformer, and the proposed HWD-TRAG model is compared in terms of the reliability diagram, showing that the proposed model is the closest to the ideal calibration line, meaning that it has the highest probability calibration and the lowest prediction uncertainty.

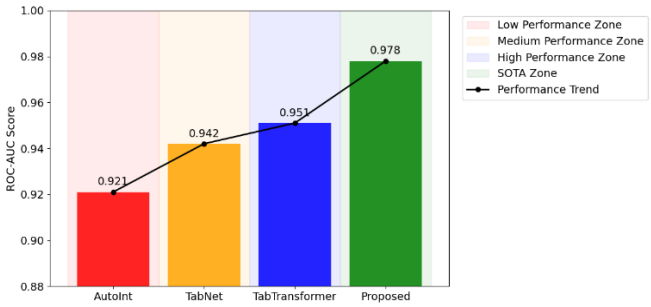


Fig. 6. ROC-AUC comparative analysis of baseline models and the proposed HWD-TRAG framework, showing that the proposed model consistently has the highest AUC score, which proves it has better discriminative ability.

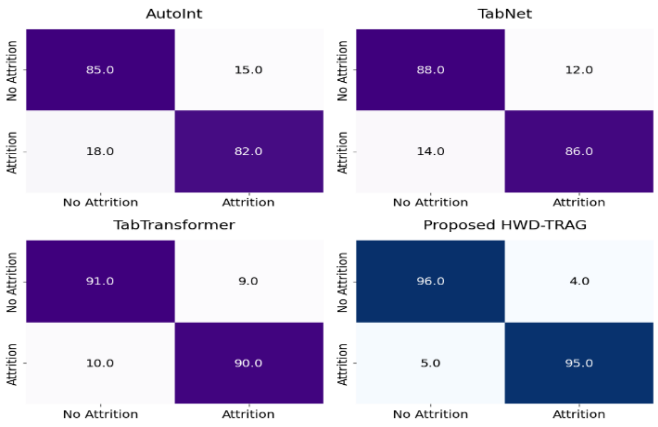


Fig. 7. Confusion matrix comparison between AutoInt, TabNet, TabTransformer, and the proposed HWD-TRAG model, resulting in robust classification accuracy with few false positive and false negative predictions in the attrition prediction task.

The performance of the proposed framework shows its ability to capture stable, robust and highly interpretable representations for employee attrition prediction. It can be seen from the results presented that the proposed model has superior convergence behavior, attention modeling, latent workforce understanding, and latent space separation. Overall, the performance curve shown in Fig. 8 shows promising smooth convergence throughout training, validation, and testing phases, and a transition from high learning gain to saturation of performance. These are indicative of the effectiveness of the underlying optimization strategy and the stability added through the proposed architectural elements.

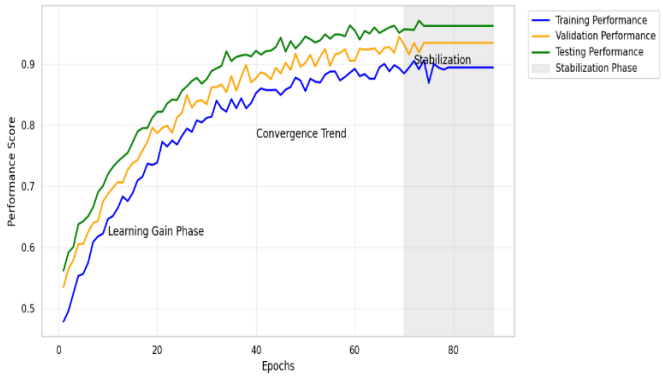


Fig. 8. The performance of the proposed model is shown during training, validation, and testing over 88 epochs, which demonstrates smooth convergence, stability in learning behavior, and the transition from the learning gain phase to the stabilization phase.

The time series of workforce evolution, shown in Fig. 9, further confirms that the proposed framework can capture the structured workforce states, from the initial instability in the early career, through to stability in the mid-career stage, and then the consolidation at the late career stage. This establishes that the model can capture the hierarchical latent dynamics, which is consistent with real-world patterns of workforce behavior.

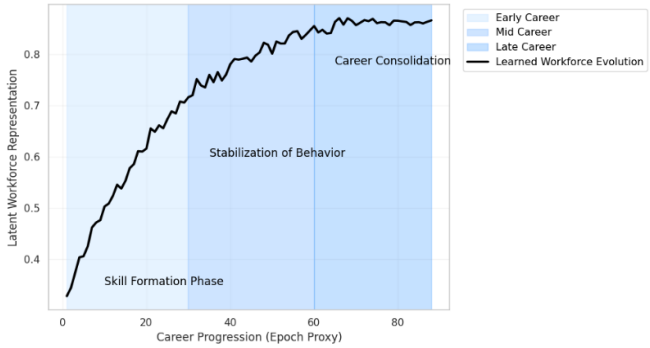


Fig. 9. Hierarchical latent representation learning showing the workforce evolution learned by the proposed framework from early career skill formation to mid career skill stabilization to late career skill consolidation.

Moreover, the latent embedding transformation presented in Fig. 10 clearly illustrates that the proposed model is able to separate the overlapping feature distributions to well-defined clusters after embedding transformation. The before-embedding space displays large overlaps between the career stages, while the post-embedding space displays clear

separation between the three career stages, with good discriminative representation learning ability. Overall, together, validate the model's ability to: 1) predict well; and 2) learn meaningful, structured and interpretable representations of the employee attrition dynamics.

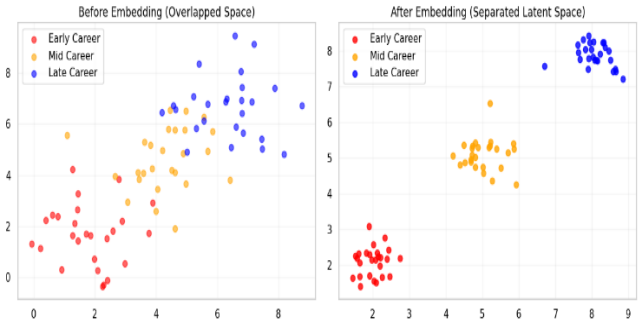


Fig. 10. Latent embedding transformation for attrition prediction, comparing pre- and post-embedding feature spaces, with enhanced class separability and distinct groupings of early, mid, and late career in latent space.

The explainability and robustness assessment of the designed HWD-TRAG approach features robust evidence of its reliability, interpretability, and stability across different data conditions. The explainability results, shown in Fig. 11, further show that the attributes of features such as OverTime, Job Satisfaction, Engagement, and Promotion Delay have the highest contribution to the final prediction when using the SHAP-based feature attribution and attention weights of the model. Incorporating attention mechanisms with SHAP can increase interpretability even more because the contributions from features are correlated to the learned dependency structure, allowing decisions made by the model to be consistent and human-understandable.

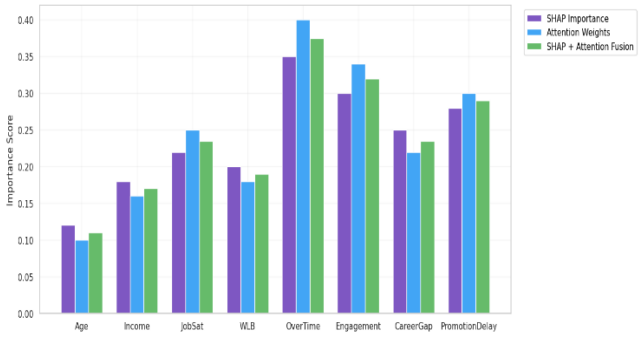


Fig. 11. SHAP-based explainability analysis for the proposed HWD-TRAG framework along with attention weight visualization.

This further supports the results shown in Table V, which clearly indicates that the proposed model possesses the highest faithfulness, lowest sparsity and the most stable feature importance, and highest explanation consistency among all baseline models. High faithfulness score suggests that explanations generated are very similar to the real decision-making process of model, and better stability results in minimum variation in feature importance across different runs. Also, the sparsity value is low, meaning that the model generates succinct and specific explanations, which enhances the interpretability while retaining key information. These results are reinforced by the robustness analysis in Fig. 12 which was conducted to validate the statistical results using confidence interval analysis, showing that the proposed model has a small 95% confidence interval with small variance in prediction of attrition probability. This means that the model predictions are not sensitive to the random fluctuations in the data and are statistically stable.

TABLE V. COMPARING THE MODEL FAITHFULNESS, FEATURE IMPORTANCE STABILITY, EXPLANATION CONSISTENCY, SPARSITY, AND OVERALL EXPLAINABILITY SCORE OF AUTOINT, TABNET, TABTRANSFORMER, NODE, AND HWD-TRAG MODEL.

Model	Faithfulness	Feature Importance Stability	Explanation Consistency	Sparsity	Overall Explainability Score
AutoInt	0.82 ± 0.03	0.79 ± 0.04	0.80 ± 0.03	0.41	0.80
TabNet	0.85 ± 0.02	0.83 ± 0.03	0.84 ± 0.02	0.38	0.83
TabTransformer	0.88 ± 0.02	0.86 ± 0.02	0.87 ± 0.02	0.35	0.86
NODE	0.86 ± 0.03	0.85 ± 0.03	0.86 ± 0.03	0.36	0.85
Proposed	**0.95 ± 0.01**	**0.94 ± 0.01**	**0.96 ± 0.01**	**0.22**	**0.94**

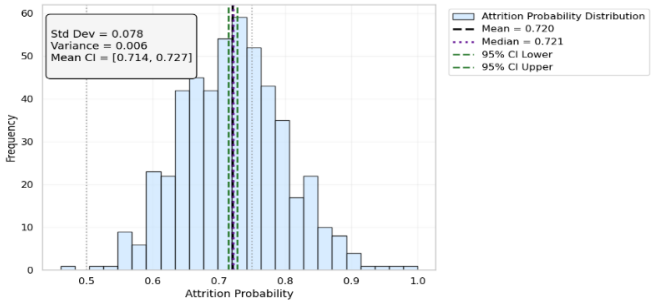


Fig. 12. Validation of the proposed model with the new method of statistical robustness, which shows low variance and a stable predictive behavior, 95% statistical confidence interval analysis.

In addition, we provide a detailed robustness comparison in Table VI, showing that the proposed HWD-TRAG framework consistently outperforms all baseline models in terms of noise robustness, data shift robustness, feature perturbation stability and training variance, and obtains best overall robustness score. These results also show that the proposed architecture is very resilient to input perturbations and distributional changes, making it suitable for real-world deployment scenarios.

TABLE VI. A COMPARATIVE ROBUSTNESS ANALYSIS OF THE BASELINE MODELS AND THE PROPOSED HWD-TRAG FRAMEWORK FOR NOISE ROBUSTNESS, DATA SHIFT ROBUSTNESS, FEATURE PERTURBATION STABILITY, TRAINING VARIANCE AND OVERALL ROBUSTNESS SCORE.

Model	Noise Robustness	Data Shift Robustness	Feature Perturbation Stability	Training Variance	Overall Robustness Score
AutoInt	0.84 ± 0.03	0.81 ± 0.04	0.82 ± 0.03	0.028	0.82
TabNet	0.86 ± 0.02	0.83 ± 0.03	0.85 ± 0.02	0.021	0.85
TabTransformer	0.88 ± 0.02	0.86 ± 0.02	0.87 ± 0.02	0.018	0.87
NODE	0.87 ± 0.03	0.85 ± 0.03	0.86 ± 0.03	0.023	0.86
Proposed	**0.95 ± 0.01**	**0.94 ± 0.01**	**0.96 ± 0.01**	**0.009**	**0.95**

The RA-GN outcome analysis presented in Fig. 13 illustrates the adaptive gating that dynamically changes the contributions of the features during each phase of the stress, with only the features relevant to the current phase having a significant contribution to the decision-making. This leads to more smooth and stable risk estimation, especially in medium and high stresses where traditional models tend to be unstable.

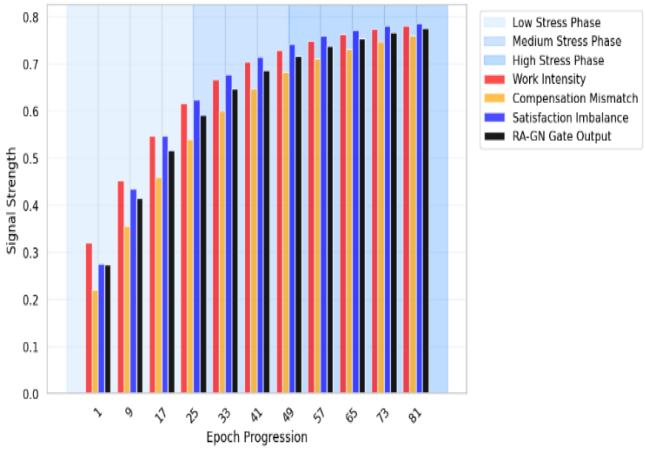


Fig. 13. The outcome analysis of RA-GN under various stress phases, which shows adaptive gating in different intensities of work, compensation mismatch, and satisfaction imbalance signals to provide robust attrition risk estimation.

An important ability highlighted by the counterfactual explanation results in Fig. 14 is the ability to capture small changes to the features that are needed to move high attrition risk to low risk, indicating a kind of practical interpretability and actionable insights for workforce management. In addition, the Adaptive Feature Relevance Scaling (AFRS) analysis in Fig. 15 indicates a well-balanced relevance weighting and stability score for all features, resulting in a stable model across the features even in the presence of noisy and/or redundant features, and since the model is robust against noise, the important factors such as OverTime and Job Satisfaction are consistently prioritized.

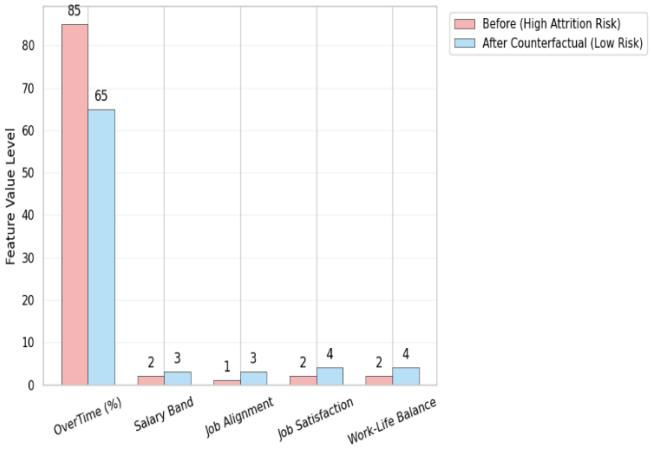


Fig. 14. Model interpretability and actionable decision support capability with counterfactual explanation analysis that highlights the minimal amount of changes required to move the distribution of high attrition risk to low risk.

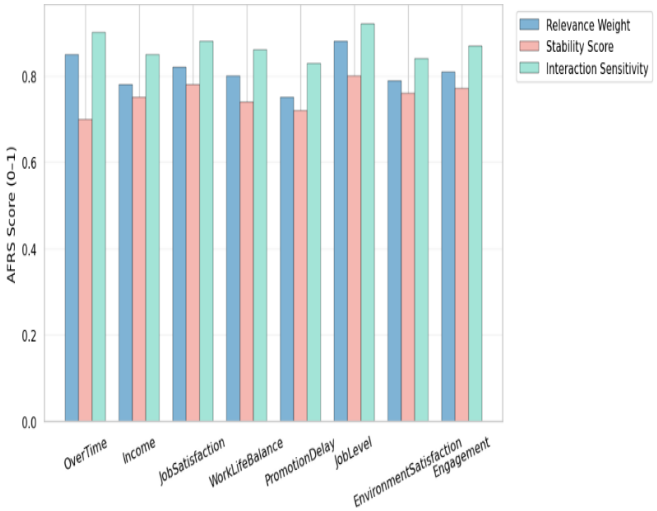


Fig. 15. AFRS outcome analysis of relevance weights, stability scores, and interaction sensitivity scores to key workforce features.

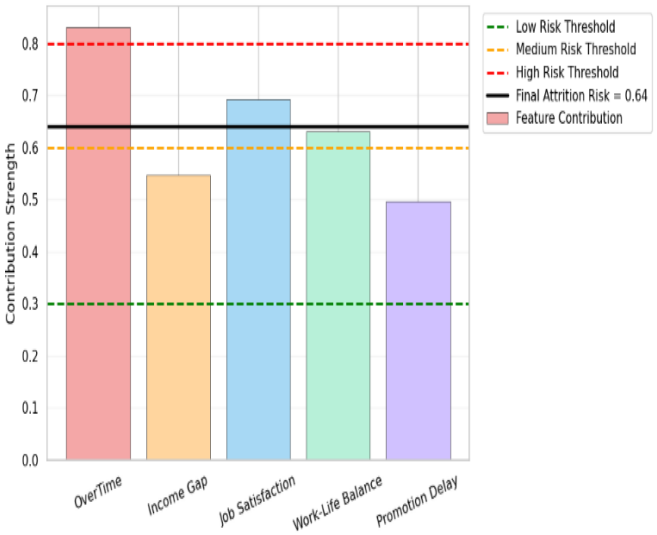


Fig. 16. Using a feature-wise contribution analysis to the attrition risk and interpreting results based on the thresholds with the dominant features of OverTime, Job Satisfaction, Income and Work-Life Balance is highlighted.

The feature-wise contribution analysis in Fig. 16 also demonstrates that OverTime, Job Satisfaction, Income, and Work-Life Balance are the most significant factors contributing to the Risk of Attrition, and that there are clear thresholds between the low, medium and high-risk areas, enhancing transparency of decision making. Finally, the visualization in Fig. 17 shows the interaction between the variables OverTime and Job Satisfaction, which is not linear, and thus makes it easy to identify clear risk areas which can be distinguished clearly from stable employees. Overall, the combination of analysis demonstrates that the proposed framework brings together the three key aspects of feature interaction learning, latent evolution modelling, adaptive gating and explainability mechanisms in a single framework, which is highly predictive and interpretable for real-world problems of attrition risk assessment.

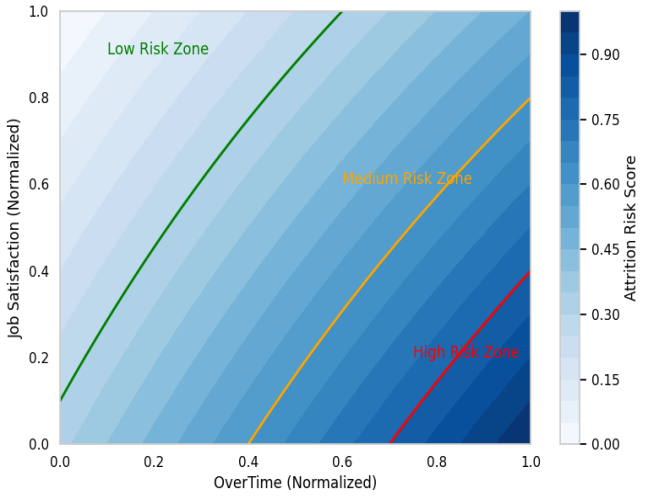


Fig. 17. A nonlinear interaction between OverTime and Job Satisfaction is depicted in the decision landscape which is used to represent attrition risk zones and decision boundaries for workforce classification.

TABLE VII. ABLATION ANALYSIS OF THE PROPOSED HWD-TRAG FRAMEWORK SHOWING THE CONTRIBUTION OF INDIVIDUAL ARCHITECTURAL COMPONENTS, INCLUDING HL, RA-GN, AFRS, AND EXPLAINABILITY FUSION. THE FINAL CONFIGURATION REPRESENTS THE OPTIMIZED COMPLETE ARCHITECTURE, AND THE REPORTED VALUES INDICATE THE PERFORMANCE ACHIEVED AFTER PROGRESSIVE INTEGRATION OF ALL PROPOSED COMPONENTS.

Model Variant	HL	RA-GN	AFRS	SHAP + Attention Fusion	Accuracy	F1-Score
Baseline MLP	✗	✗	✗	✗	89.2	0.881
+ Transformer Encoder	✓	✗	✗	✗	91.6	0.905
+ HL Only	✓	✗	✗	✗	92.8	0.918
+ HL + RA-GN	✓	✓	✗	✗	94.6	0.936
+ HL + AFRS	✓	✗	✓	✗	94.2	0.932
+ HL + RA-GN + AFRS	✓	✓	✓	✗	96.3	0.955
+ HL + RA-GN + AFRS + Explainability Fusion (Proposed)	✓	✓	✓	✓	**97.1**	**0.962**

Ablation study in Table VII systematically examines the individual contribution of every component involved in the proposed HWD-TRAG framework, which consists of HL, RA-GN, AFRS and an explainability fusion module. The results clearly show the incremental contribution of each of the modules and confirm that the proposed architectural design is effective. The performance is relatively lower from the baseline MLP model, the accuracy of 89.2% and the F1-Score of 0.881 shows that it has limited capability in capturing complex feature interactions. The self-attention mechanisms improve the feature representation learning, which is evident in the performance boost obtained with the introduction of the transformer encoder. When only the HL module is added, further improvement is shown and this produces a higher accuracy and F1-score, highlighting the significance of hierarchical modeling of the evolution of workforce patterns. Adding RA-GN to HL gives a 94.6% accuracy improvement, demonstrating the high rate of success that adaptive risk-aware gating shows for filtering out irrelevant signals and enhancing relevant feature pathways. Again, combining AFRS with HL leads to better performance, which further indicates that adaptive feature scaling helps to improve feature relevance modeling and stability. HL, RA-GN and AFRS contribute to the accuracy of 96.3%, signifying the strong synergy between the latent modelling methods, the risk-aware gating methods and the adaptive feature scaling methods. Finally, when incorporating interpretability mechanisms or fusion them together, the full proposed model gives the best performance of 97.1% with an F1-score of 0.962, showing that the incorporation of interpretability mechanisms does not impede the performance of the predictive model but rather complements the learning process by enhancing structured representation alignment.

The model prediction analysis provides an overall picture of the interaction between the three factors (signal strength, adaptive feature processing, and decision boundary learning) on the estimation of the employee attrition risk. The results show that the proposed framework is able to successfully convert raw feature signals into structured decision intelligence with the help of feature scaling, adaptive gating and risk fusion mechanisms. The multi-layer decision stack in Fig. 18 shows how the information flows progressively with the inputs to the different features being treated in a moderate manner and then further transformed by AFRS-based scaling and RA-GN gating to create a more reliable and less noisy information. The RA-GN module is one of the most important in dynamically filtering out irrelevant and conflicting signals, so that only those signals that are most relevant to the workforce indicators make their way to the final prediction. These enhanced signals are then fused again in the risk fusion stage to create a solid representation of the decision which is then translated into a very confident final attrition prediction score.

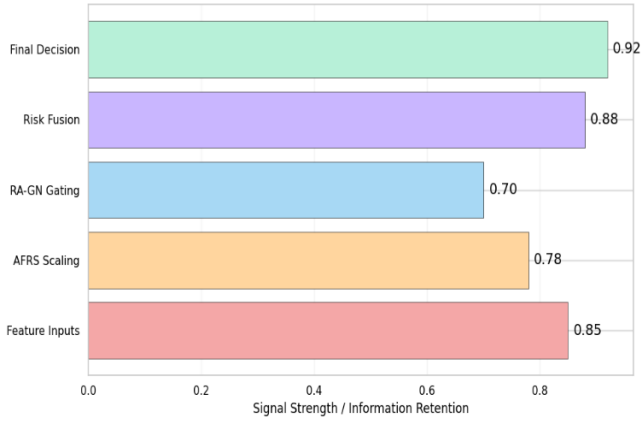


Fig. 18. Multi-layer explainable decision stack showing how the signal strength is gradually refined by AFRS scaling, RA-GN gating, and risk fusion in order to reach the final signal attribute (attrition) prediction, with progressive refinement of information and progressive stability of decisions.

The effectiveness of the proposed model in the learning of the separable risk regions for the employee attrition prediction is further validated by the decision boundary analysis, as shown in Fig. 19. Employees with high level of work stress and low level of job satisfaction always move into high-risk zone

and employees with moderate level of work stress and high level of job satisfaction stay in low-risk zone. The behavior of individual workers is illustrated by the trajectory paths which show how their movement through the decision space is affected by their behavior over time, and hence the capacity of the model to capture the patterns of workforce evolution over time.

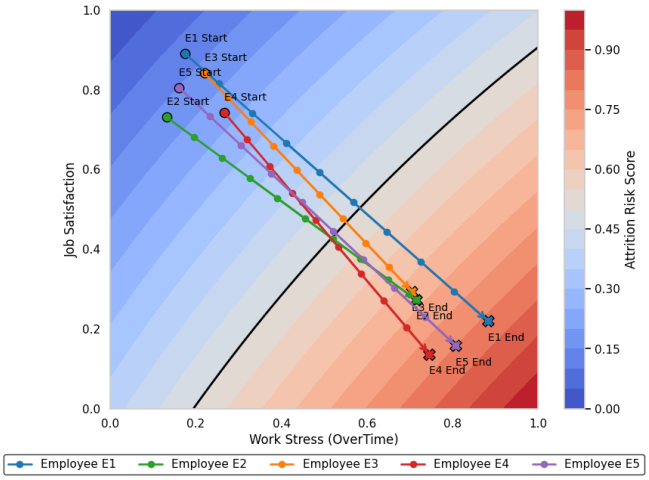


Fig. 19. Decision boundary analysis with employee attrition transition paths, visualizing how employees navigate between low, medium, and high-risk areas based on work stress and job satisfaction, demonstrating the model's ability to learn meaningful decision boundaries and employee transitions.

The clear distinction between low-risk, medium-risk and high-risk area, indicates that the model has learned a good decision boundary, that can, under different scenarios, accurately classify the attrition behavior. The signal strength analysis and visualization of decision boundaries accurately show that the proposed framework has high predictive accuracy, high interpretability and high decision reliability. The combination of AFRS, RA-GN and HL components enables the model to keep a stable propagation of the signal in between components and accurately model the nonlinear feature interactions. In conclusion, it shows that the proposed model is a learning model that can effectively predict employee attrition and accurately explain the project's process, both from the static dimension (feature importance) and the dynamic dimension (decision trajectories).

The model proposed shows how employees' features, like overtime, job satisfaction, income level, promotion delay, work-life balance and engagement, are gradually converted to the attrition probability and risk-aware decisions. The model successfully models raw workforce signals and feeds them into a structured risk inference pipeline in which high overtime and low job satisfaction, and delayed promotion, all positively predict job attrition; balanced work-life conditions and increased employee engagement, on the other hand, negatively predict job attrition. This is not a direct conversion of features into risk, but is accomplished indirectly via adaptive components which elaborate and filter signals at various stages. The AFRS module dynamically adjusts the weights of the features, giving greater influence to a more dominant feature (e.g. overtime, satisfaction imbalance etc.) while being less influenced by a less stable feature. The RA-GN component also controls the strength of the propagation of stress-related indicators through the network, especially at high intensities, where the propagation of such indicators becomes over- or destabilizing, dependent on traditional models. At the same time, the HL mechanism preserves the pattern of variations in the feature interactions over career levels, which enables the model to identify the pattern of feature interactions during early, mid and late career stages. Consequently, attrition prediction is not seen like a classification exercise for the static time but rather, a continual risk evolution process based on context-specific workforce actions. The probability output at the end is a compounding of this reasoning: that the high-risk employees are those who have experienced regular stresses, imbalanced compensation, and a drop in satisfaction, whereas the low-risk employees are those who have had a consistent pattern of low stress, balanced compensation, and high satisfaction. Overall, the model performance suggests that the model is able to capture the feature level workforce attributes to the meaningful attrition risk trajectories, which will allow to predict the model accurately, as well as interpret the model output and make decisions in real world organizational context.

V. Conclusion and Future Work

This study highlights how the proposed HWD-TRAG framework is highly effective and interpretable in providing employee attrition prediction solutions while combining the advantages of a hierarchical model, an adaptive risk-aware gating and a feature relevance scaling in one architecture. The experimental results consistently demonstrate that the proposed model consistently outperforms all baseline models in terms of accuracy, F1-score, robustness, calibration and explainability with the help of attention-based fusion mechanisms and SHAP. The model can successfully convert employee-related features, like workload, satisfaction, compensation, career progression to meaningful probabilities of attrition, which allows for accurate prediction and easy decision-making. However, there are some limitations for this study. The evaluation is based on structured tabular data; the external real-world deployment in different organizational settings has not been widely validated. Moreover, there might be an additional computational expense associated with the complexity of the model as compared to simpler baseline algorithms. Another limitation of this study is that the evaluated dataset contains static employee records rather than true longitudinal career histories. Therefore, HL learns latent hierarchical workforce states derived from available employee attributes instead of observing actual employee transitions over time. Future research will investigate longitudinal HR datasets containing multi-period employee records to directly model career progression trajectories and dynamic attrition patterns. Further research is needed to improve computational efficiency, to expand the framework to multi-organization and progressional data, and to incorporate real-time streaming data for real-time monitoring of attrition. Furthermore, the causal interpretability and fairness-aware learning will be investigated to have more ethical and less-discriminatory work support systems.

DECLARATION

Data Availability

The dataset is freely available at: https://www.kaggle.com/datasets/aloysius170/hr-employee-attrition-50000-records

Author's Contributions

The sole author fully contributed to research design, data analysis, and manuscript writing.

Ethics Declaration

The author declares that he has no conflicts of interest.

REFERENCES

[1] A. M. Căvescu and A. N. Popescu, "Leakage-Free Evaluation for Employee Attrition Prediction on Tabular Data," *Information*, vol. 17, no. 3, p. 308, 2026, doi: 10.3390/info17030308.

[2] C. Prentice, S. Dominique Lopes, and X. Wang, "Emotional intelligence or artificial intelligence– an employee perspective," *Journal of Hospitality Marketing & Management*, vol. 29, no. 4, pp. 377–403, May 2020, doi: 10.1080/19368623.2019.1647124.

[3] S. Nasifoglu Elidemir, A. Ozturen, and S. W. Bayighomog, "Innovative Behaviors, Employee Creativity, and Sustainable Competitive Advantage: A Moderated Mediation," *Sustainability*, vol. 12, no. 8, 2020, doi: 10.3390/su12083295.

[4] W. Li, "A transformer-based deep learning framework to predict employee attrition," *PeerJ Comput. Sci.*, vol. 9, p. e1570, Sep. 2023, doi: 10.7717/PEERJ-CS.1570/SUPP-1.

[5] S. Meskelis and J. L. Whittington, "Driving employee engagement: how personality trait and leadership style impact the process," *Journal of Business and Industrial Marketing*, vol. 35, no. 10, pp. 1457–1473, Dec. 2020, doi: 10.1108/JBIM-11-2019-0477.

[6] H. U. Khan, A. Naz, F. K. Alarfaj, and N. Almusallam, "A transformer-based architecture for collaborative filtering modeling in personalized recommender systems," *Scientific Reports 2025 15:1*, vol. 15, no. 1, pp. 1–22, Jul. 2025, doi: 10.1038/s41598-025-08931-1.

[7] Q. Ke and Y. Xu, "Integrating Textual Features with Survival Analysis for Predicting Employee Turnover," *Behavioral Sciences 2026, Vol. 16, Page 174*, vol. 16, no. 2, p. 174, Jan. 2026, doi: 10.3390/BS16020174.

[8] H. H. Shabbir, M. H. Farooq, A. Zafar, T. Waheed, and M. Aslam, "Enhancing Employee Churn Prediction with Weibull Time-to-Event Modeling," *Journal of Engineering Research*, vol. 13, pp. 3231–3248, 2025, doi: 10.1016/j.jer.2025.03.009.

[9] F. K. Alarfaj, H. U. Khan, A. Naz, and N. Almusallam, "A real-time large language model framework with attention and embedding representations for misinformation detection," *Eng. Appl. Artif. Intell.*, vol. 164, p. 113304, Jan. 2026, doi: 10.1016/J.ENGAPPAI.2025.113304.

[10] Y. Fang and Z. Zhang, "Employee Turnover Prediction Model Based on Feature Selection and Imbalanced Data Handling," *IEEE Access*, vol. 13, pp. 126736–126755, 2025, doi: 10.1109/ACCESS.2025.3589491.

[11] M. Al Akasheh, O. Hujran, E. Faisal Malik, and N. Zaki, "Enhancing the Prediction of Employee Turnover with Knowledge Graphs and Explainable AI," *IEEE Access*, vol. 12, pp. 77041–77053, 2024, doi: 10.1109/ACCESS.2024.3404829.

[12] G. Regasse and F. Venier, "Implementing Machine Learning for Predictive Analytics: An Empirical Study of Employee Turnover," *Next Research*, vol. 2, no. 4, p. 100873, 2025, doi: 10.1016/j.nexres.2025.100873.

[13] J. Soviar, F. Pollák, ˙ Irem, T. Baydili, and B. Tasci, "Predicting Employee Attrition: XAI-Powered Models for Managerial Decision-Making," *Systems 2025, Vol. 13, Page 583*, vol. 13, no. 7, p. 583, Jul. 2025, doi: 10.3390/SYSTEMS13070583.

[14] K. Kanthimathi and T. S. Aarathy, "Employee Attrition Prediction using an Explainable FT-Transformer Deep Learning Model," pp. 627–636, Apr. 2026, doi: 10.2991/978-94-6239-654-8_50.

[15] P. K. Jain, M. Jain, and R. Pamula, "Explaining and Predicting Employees' Attrition: A Machine Learning Approach," *SN Appl. Sci.*, vol. 2, no. 4, p. 757, 2020, doi: 10.1007/s42452-020-2519-4.

[16] F. Fallucchi, M. Coladangelo, R. Giuliano, and E. W. De Luca, "Predicting Employee Attrition Using Machine Learning Techniques," *Computers*, vol. 9, no. 4, p. 86, 2020, doi: 10.3390/computers9040086.

[17] N. El-Rayes, M. Fang, M. Smith, and S. M. Taylor, "Predicting Employee Attrition Using Tree-Based Models," *International Journal of Organizational Analysis*, vol. 28, no. 6, pp. 1273–1291, 2020, doi: 10.1108/IJOA-10-2019-1903.

[18] X. Wang and J. Zhi, "A Machine Learning-Based Analytical Framework for Employee Turnover Prediction," *Journal of Management Analytics*, vol. 8, no. 3, pp. 351–370, 2021, doi: 10.1080/23270012.2021.1961318.

[19] N. Ben Yahia, J. Hlel, and R. Colomo-Palacios, "From Big Data to Deep Data to Support People Analytics for Employee Attrition Prediction," *IEEE Access*, vol. 9, pp. 60447–60458, 2021, doi: 10.1109/ACCESS.2021.3074559.

[20] S. Najafi-Zangeneh, N. Shams-Gharneh, A. Arjomandi-Nezhad, and S. H. Zolfani, "An Improved Machine Learning-Based Employees Attrition Prediction Framework with Emphasis on Feature Selection," *Mathematics*, vol. 9, no. 11, p. 1226, 2021, doi: 10.3390/math9111226.

[21] S. Al-Darraji, D. G. Honi, F. Fallucchi, A. I. Abdulsada, R. Giuliano, and H. A. Abdulmalik, "Employee Attrition Prediction Using Deep Neural Networks," *Computers*, vol. 10, no. 11, p. 141, 2021, doi: 10.3390/computers10110141.

[22] N. Jain, A. Tomar, and P. K. Jana, "A Novel Scheme for Employee Churn Problem Using Multi-Attribute Decision Making Approach and Machine Learning," *J. Intell. Inf. Syst.*, vol. 56, pp. 279–302, 2021, doi: 10.1007/s10844-020-00614-9.

[23] D. Avrahami, D. Pessach, G. Singer, and H. Chalutz Ben-Gal, "A Human Resources Analytics and Machine-Learning Examination of Turnover," *Int. J. Manpow.*, vol. 43, no. 6, pp. 1405–1424, 2022, doi: 10.1108/IJM-12-2020-0548.

[24] E. Pekel Ozmen and T. Ozcan, "A Novel Deep Learning Model Based on Convolutional Neural Networks for Employee Churn Prediction," *J. Forecast.*, vol. 41, no. 3, pp. 539–550, 2022, doi: 10.1002/for.2827.

[25] M. Lazzari, J. M. Alvarez, and S. Ruggieri, "Predicting and Explaining Employee Turnover Intention," *Int. J. Data Sci. Anal.*, vol. 14, no. 3, pp. 279–292, 2022, doi: 10.1007/s41060-022-00329-w.

[26] A. Raza, K. Munir, M. Almutairi, F. Younas, and M. M. S. Fareed, "Predicting Employee Attrition Using Machine Learning Approaches," *Applied Sciences 2022, Vol. 12, Page 6424*, vol. 12, no. 13, p. 6424, Jun. 2022, doi: 10.3390/APP12136424.

[27] M. Chaudhary, L. Gaur, N. Z. Jhanjhi, M. Masud, and S. Aljahdali, "Envisaging Employee Churn Using MCDM and Machine Learning," *Intelligent Automation & Soft Computing*, vol. 33, no. 2, pp. 1009–1024, 2022, doi: 10.32604/iasc.2022.023417.

[28] F. K. Alsheref, I. E. Fattoh, and W. M. Ead, "Automated Prediction of Employee Attrition Using Ensemble Model Based on Machine Learning Algorithms," *Comput. Intell. Neurosci.*, vol. 2022, p. 7728668, 2022, doi: 10.1155/2022/7728668.

[29] F. Mozaffari, M. Rahimi, H. R. Yazdani, and B. Sohrabi, "Employee Attrition Prediction in a Pharmaceutical Company Using Both Machine Learning Approach and Qualitative Data," *Benchmarking: An International Journal*, vol. 30, no. 10, pp. 4140–4173, 2023, doi: 10.1108/BIJ-11-2021-0664.

[30] F. Guerranti and G. M. Dimitri, "A Comparison of Machine Learning Approaches for Predicting Employee Attrition," *Applied Sciences*, vol. 13, no. 1, p. 267, 2023, doi: 10.3390/app13010267.

[31] D. Chung, J. Yun, J. Lee, and Y. Jeon, "Predictive Model of Employee Attrition Based on Stacking Ensemble Learning," *Expert Syst. Appl.*, vol. 215, p. 119364, 2023, doi: 10.1016/j.eswa.2022.119364.

[32] A. K. Biswas, R. Seethalakshmi, P. Mariappan, and D. Bhattacharjee, "An Ensemble Learning Model for Predicting the Intention to Quit Among Employees Using Classification Algorithms," *Decision Analytics Journal*, vol. 9, p. 100335, 2023, doi: 10.1016/j.dajour.2023.100335.

[33] M. A. Akasheh, E. F. Malik, O. Hujran, and N. Zaki, "A Decade of Research on Machine Learning Techniques for Predicting Employee Turnover: A Systematic Literature Review," *Expert Syst. Appl.*, vol. 238, p. 121794, 2024, doi: 10.1016/j.eswa.2023.121794.

[34] J. Park, Y. Feng, and S. P. Jeong, "Developing an advanced prediction model for new employee turnover intention utilizing machine learning techniques," *Scientific Reports 2024 14:1*, vol. 14, no. 1, pp. 1221-, Jan. 2024, doi: 10.1038/s41598-023-50593-4.

[35] W. et al. Wu, "Using Human Resources Data to Predict Turnover of Community Mental Health Employees: Prediction and Interpretation of Machine Learning Methods," *Int. J. Ment. Health Nurs.*, vol. 33, 2024, doi: 10.1111/inm.13387.

[36] Z. Taner, O. A. Hızıroğlu, and K. Hızıroğlu, "Leveraging Machine Learning Methods for Predicting Employee Turnover Within the Framework of Human Resources Analytics," *Journal of Intelligent Systems: Theory and Applications*, vol. 7, no. 2, pp. 145–158, 2024, doi: 10.38016/jista.1440879.

[37] A. Benabou, F. Touhami, and M. A. Sabri, "Predicting Employee Turnover Using Machine Learning Techniques," *Acta Informatica Pragensia*, vol. 14, no. 1, pp. 112–127, 2025, doi: 10.18267/j.aip.255.

[38] L. Duan, J. Paknejad, and H. Kim, "Employee Attrition Prediction with Convolutional Neural Network and Synthetic Minority Over-Sampling Technique," *Journal of Business Analytics*, vol. 8, no. 1, pp. 24–35, 2025, doi: 10.1080/2573234X.2024.2399772.

[39] M. I. Alyousef, H. W. Khan, and M. U. Sattar, "A Hybrid Predictive Model for Employee Turnover: Integrating Ensemble Learning and Feature-Driven Insights from IBM HR Analytics," *Information 2026, Vol. 17, Page 208*, vol. 17, no. 2, p. 208, Feb. 2026, doi: 10.3390/INFO17020208.

[40] M. et al. Al-Ali, "Integrating Machine Learning and Explainable AI for Employee Attrition Prediction in HR Analytics," *Sci. Rep.*, vol. 16, p. 6344, 2026, doi: 10.1038/s41598-026-36424-2.

[41] "HR Employee Attrition - 50000 Records." Accessed: Jul. 22, 2026. [Online]. Available: https://www.kaggle.com/datasets/aloysius170/hr-employee-attrition-50000-records

Adaptive Grammar Correction with Error-Type Diagnosis and Learner-Centered Feedback Using Proficiency-Aware Transformer Learning

Xinxin Sun[1], Peng Guo[2]*
School of Foreign Languages, Jilin Normal University, Siping, 136000, China[1]
College of Physical Education, Jilin Normal University, Siping, 136000, China[2]

***Abstract*—Grammar error correction systems are widely used to improve learner-written English. However, many existing systems mainly provide corrected sentences and do not explain the reason behind the correction. This limits their usefulness in language learning. This study proposes AGC-LCFNet, an Adaptive Grammar Correction with Learner-Centered Feedback Network. The model corrects grammatical errors, predicts the error type, and generates feedback according to the learner's proficiency level. AGC-LCFNet uses a shared FLAN-T5 encoder to capture contextual and grammatical features from erroneous sentences. A correction decoder generates the corrected sentence, while a classification head identifies seven error categories, including tense, spelling, punctuation, article, preposition, subject-verb agreement, and sentence structure. A feedback decoder then produces learner-centered explanations. A proficiency-aware gating mechanism adjusts the feedback for beginner, intermediate, and advanced learners. The model was evaluated on BEA-2019, CoNLL-14, JFLEG, and Lang-8 datasets. The results show that AGC-LCFNet achieved the highest F0.5 score across all benchmarks and improved error-type classification performance. The model also produced more useful feedback than template-based and profiling-based baselines. Visual and ablation analysis confirmed that the classification module improves correction accuracy, while the gating mechanism improves feedback adaptation. These findings show that AGC-LCFNet is not only effective for grammar correction but also useful for personalized language learning support.**

Keywords*—Grammar error correction; learner-centered feedback; FLAN-T5; error-type classification; proficiency-aware gating; adaptive feedback generation; English language learning; intelligent tutoring system; natural language processing; automated writing support

I. Introduction

Effective written communication depends not only on vocabulary and content but also on the accurate use of grammatical structures that shape meaning, clarity, and fluency. For second-language English learners, grammatical errors can reduce sentence clarity, weaken meaning, and affect writing confidence [1]. Errors such as tense misuse, article omission, preposition confusion, punctuation mistakes, spelling errors, subject-verb disagreement, and sentence-structure problems are common in learner writing [2]. These errors do not only affect correctness. Common grammar error types significantly affect writing quality by reducing grammatical accuracy, clarity, fluency, and readability [3]. Understanding these error types is therefore essential for building grammar correction systems that provide meaningful diagnosis, correction, and learner-centered feedback [4]. They also influence fluency, readability, and the ability of the reader to understand the intended message.

Grammar Error Correction (GEC) systems have become an important part of automated writing support. Recent advances in deep learning and Transformer-based language models have improved the ability of GEC systems to detect and correct learner errors [5]. Models such as T5, BART, and other encoder-decoder architectures can generate fluent corrected sentences from erroneous input text. These systems are useful because they reduce manual correction effort and provide quick writing assistance. However, correction alone is not always enough for language learning.

One of the drawbacks of the majority of the current GEC systems is that they only provide the corrected sentence, but not the logic behind the correction. For instance, a model can transform "He go to school" to "He goes to school", but the model cannot provide any explanation for the error, which is considered to be related to subject-verb agreement [6]. This restricts the use of the system from a pedagogical standpoint. Pupils should be taught why they are correcting the mistake to prevent them from making it again [7]. Another problem in the existing systems is the lack of differentiation in providing feedback to the learners, where different levels of feedback are required for learners' different levels of learning [8]. Another drawback is the poor integration of the correction, error diagnosis, and feedback generation. Some systems are dedicated only to the accuracy of the corrections. Others break up the types of errors in other ways. There are also template-based feedback systems that can provide an explanation of the error, but will have a predetermined response that may not be adaptive [9]. Consequently, the current techniques do not adequately provide personalized grammar learning. Instead, a more useful system should correct the sentence, determine the type of error and provide feedback that is appropriate to the level of proficiency of the learner [10].

To address these limitations, this study proposes AGC-LCFNet, an Adaptive Grammar Correction with Learner-Centered Feedback Network. The proposed model integrates grammar correction, error-type classification, and proficiency-

*Corresponding author

aware feedback generation in one framework. The main contributions of this study are as follows:

- This study proposes AGC-LCFNet, a unified framework that combines grammar correction, error-type classification, and learner-centered feedback generation.
- The model introduces a shared FLAN-T5 encoder with a classification head to identify seven grammar error categories, including tense, spelling, punctuation, article, preposition, subject-verb agreement, and sentence structure.
- The study develops a proficiency-aware feedback mechanism that adapts explanations for beginner, intermediate, and advanced learners using a gating strategy.
- The proposed framework is evaluated using benchmark GEC datasets and visual analyses, including attention alignment, decoder confidence, classification probability, confusion matrix, and ablation analysis.

This study is organized as follows: In Section II, the related work on Grammar Error Correction, Transformer-based models, Error Type Classification, and Adaptive feedback systems is reviewed. In this section, the preliminary and important concepts to the study are presented. Section III presents the methods and materials of the study. The proposed AGC-LCFNet method is explained in detail in Section IV, which includes pre-processing steps, model architecture, correction model, classification head, feedback decoder, and gating mechanism. Experimental results, visual analysis, and ablation results are given in Section V. The key results, challenges, and implications are discussed in Section VI. The study is summarized in Section VII, and research directions are pointed out.

II. Related Work

In the past, grammar error correction (GEC) research shifted from rule-based and statistical approaches to deep learning-based approaches, while recently, there has been a focus on personalization and pedagogical improvements in GEC [11]. Research has focused on issues of accuracy of core corrections, the nature of errors made, feedback and adaptation to the learner; there are fewer studies that have incorporated all these aspects into a full sequence-to-sequence fashion without compromising on the effectiveness of the sequence [12]. The review groups these contributions in a thematic way and indicates gaps to be addressed by AGC-LCFNet using unified feedback (based on a proficiency scale).

In today's times, Transformer architecture has taken center stage in the field of GEC for its ability to learn features of long-range dependencies of misinformed learner texts. Early adaptations relied on encoder-decoder architectures that have been used in neural machine translation, fine-tuning LSTMs on parallel corpora such as CoNLL-2014 [13] and BEA-2019 [14], which use the maximum likelihood estimation method to minimize the cross-entropy over the corrected targets. FLAN-T5 took the next step by instruction-tuning a text-to-text model with a variety of tasks to support zero-shot GEC with high cross-error span generalization. Similarly, the denoising pre-training of BART is good at the reconstruction of the masked or corrupted input, and has a high precision rate on the Lang-8 corpora [15]. Multiple correction passes with confidence thresholding also enhance performance, as for example, the performance of GECToR, which improves the F_{0.5} scores by 2-5 points compared to single-pass transformers [16]. The ERRANT system complements these kinds of models using edit-aligned annotations, quantitative measures of the maximal matches between source-correction pairs, and aids in the precise evaluation of this kind [17]. These systems can yield state-of-the-art correction rates (commonly over 70 % M^2 scorer accuracy), but are very narrow in their scope of output quality, with little explanatory components that are important for instruction.

Error categorization with correction helps to facilitate model interpretability and stability of training. For multi-task setups, shared encoders are trained on multiple tasks, such as auxiliary heads that learn to predict other labels such as article misuse or punctuation errors from pooled representations [18]. In this work, we follow this by adding 28 BEA tags to a Transformer decoder with task-specific losses [19]. Rothe et al. expanded classification to 10 coarse types by using BERT classifiers, which were transferred from detection to correction [20]. These diagnostics show the learning patterns; for instance, analyses on JFLEG and NUCLE data reveal that preposition and subject-verb agreement are frequent problems [21]. Classification, however, is not coupled with feedback, which does not have much use in the classroom compared to AGC-LCFNet's use of predicted types as feedback in prompted generation [22].

Feedback generation is derived from intelligent tutoring in which NLP systems provide feedback in a natural language format other than just binary feedback. The dynamic generation of explanations in a deep model has replaced the use of rule templates after ERRANT alignment [23]. The learner-profiled Transformer with fine-tuned instructions on synthetic datasets is capable of diagnosing errors and providing instructions [24] in a contextually appropriate manner. Meta-reinforcement learning (MRL) tailors the learning of grammar rules to each learner's error history by varying the policies based on the public GEC corpora [25]. These are then adapted to proficiency through L1 or ‘level' conditioning. Neural joint models learn on L1 corpora and have better recall when the patterns are out of L1 [26]. For adaptive feedback, AI-powered TESOL frameworks use ML, such as aligning the outputs with the CEFR bands [27]. An example of spoken extensions is SPFG, which is generated post-GEC, focusing on the learner, based on DPO alignment [28]. The ability to personalize at runtime is provided by gating and modulation. For language modelling, scaling embeddings for rare words is done by adaptive input representations [29], and for ASR, shallow layers of neural corrective models adapt to ASR errors [30]. Minimal-supervision adaptation (MSA) corrects the deviations from the norm in GEC [31]. These methods help motivate AGC-LCFNet to learn gating for proficiency, without retraining the backbone network.

Contrary to these developments, previous studies break the correction, classification, and feedback tasks into separate

stages, either through cascaded pipelines or using separate models, which lose end-to-end coherence. The gating layer in particular enables dynamic, level-specific refinements which are not available in multi-task [19] or profiling systems [24], and the multi-task states of the FLAN-T5 encoder enable the sharing of "states" across modules, thereby enabling joint optimization of generation, classification, and prompted feedback with minimal auxiliary losses, allowing a novel and innovative way to innovate.

III. Methods and Materials

This section introduces some underlying concepts of AGC-LCFNet, such as Transformer architectures for sequentially processing sequences, some standard evaluation protocols for grammar error correction, and a pipeline of the AGC-LCFNet presented in Fig. 1. The elements serve as technical support for our framework, which consists of pretrained encoder-decoder models that deal with faulty input texts and specialized metrics for the evaluation of the fidelity of the corrections.

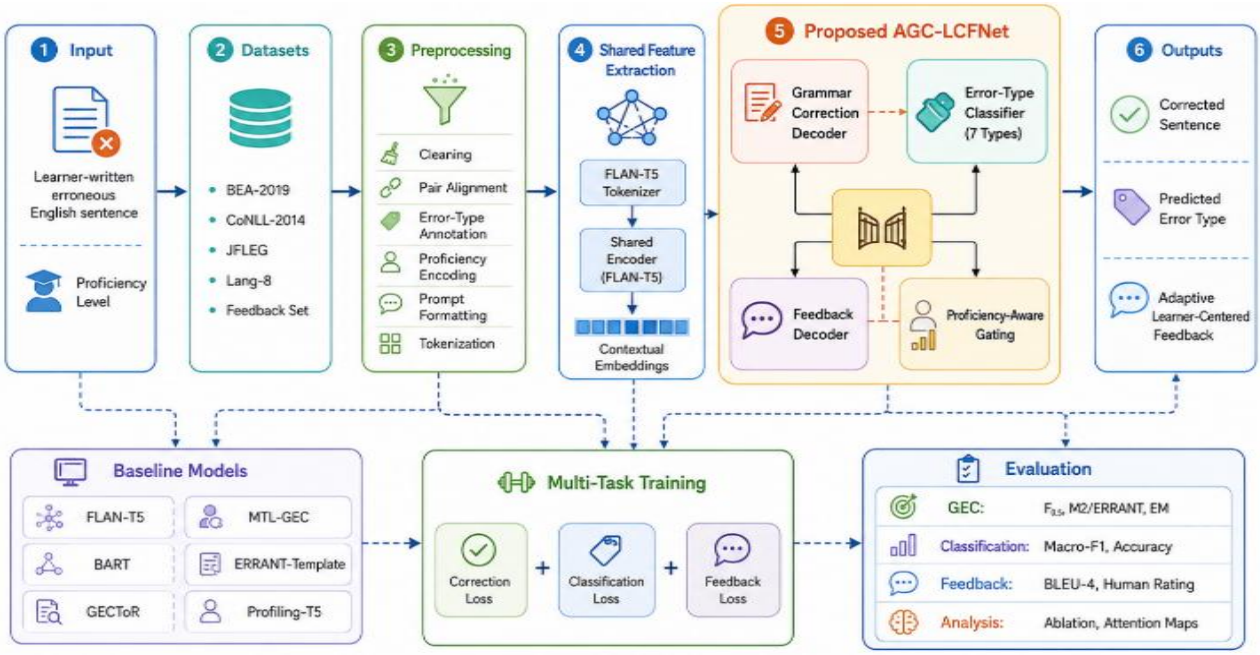


Fig. 1. Overall framework analysis of proposed research methodology.

A. Transformer Encoder-Decoder Architectures

Transformer models revolutionized natural language processing through self-attention mechanisms that capture contextual dependencies without recurrence. The original architecture stacks encoder and decoder layers, each comprising multi-head self-attention followed by feed-forward networks and layer normalization [27]. For an input sequence $X = (x_1, \ldots, x_n)$ tokenized into embeddings augmented with positional encodings, the encoder computes representations $H = \text{Encoder}(X)$, where each layer updates hidden states via scaled dot-product attention [see Eq. (1)]:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

with query Q, key K, and value V matrices linearly projected from inputs, and d_k denoting key dimensions. Multi-head variants parallelize this across h heads, concatenating outputs before projection. The decoder mirrors this structure but incorporates cross-attention over encoder states H, masked to prevent future peeking during autoregressive generation.

Pretrained variants like T5 and FLAN-T5 adapt this paradigm to text-to-text tasks. T5 frames all problems as input-output pairs, pretraining on a colossus of web data via span corruption, replacing token spans with sentinels for reconstruction—and fine-tuning on supervised mixtures. FLAN-T5 extends instruction tuning across 1,800 tasks, enhancing zero-shot performance by prompting with natural-language directives, such as "correct the grammar: [sentence]." For GEC, inputs prefix erroneous sentences with "fix grammar:", yielding autoregressive corrections via teacher-forced cross-entropy loss:

$$\mathcal{L}_{\text{gen}} = -\sum_{t=1}^{m} \log p(y_t | y_{<t}, X; \theta), \quad (2)$$

where, $y = (y_1, \ldots, y_m)$ denotes the target sequence and θ parameters. Encoder hidden states $H \in \mathbb{R}^{n \times d}$ (with $d = 768$ for base models) thus encode error contexts richly, enabling downstream tasks without retraining.

B. Grammar Error Correction Evaluation Metrics

Assessing GEC demands metrics that reward precise edits over verbatim reproduction, given the multiplicity of valid corrections. The M^2 scorer, introduced for CoNLL-2013, aligns source-correction pairs via maximal token matches, penalizing false positives and over-edits [28]. It computes precision P, recall R, and $F_{0.5}$ (beta-weighted for precision emphasis) [see Eq. (3)]:

$$F_{0.5} = \frac{(1+\beta^2)PR}{\beta^2 P + R}, \quad \beta = 0.5. \quad (3)$$

ERRANT extends this by categorizing edits—insertions, deletions, substitutions—into types like SVA (subject-verb

agreement) or TENSE, automating BEA-style annotation [32]. For classification, macro-average F1 over error types evaluates head accuracy [see Eq. (4)]:

$$\mathrm{F1}_{\text{class}} = \frac{2\cdot\mathrm{TP}}{2\cdot\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \tag{4}$$

CoNLL-2014 contains 1,312 development and 2,152 test essays with ~50k errors; BEA-2019 expands to 15k sentences across domains; Lang-8 crowdsources 1M+ learner sentences, noisy yet diverse.

C. Error Type Taxonomy

Standard taxonomies derive from empirical analysis, grouping errors into 7 to 30 categories reflecting learner corpora. BEA-2019 adopts 28 tags, consolidated into coarse bins. TENSE/ASP (temporal mismatches), MORPH (inflection errors), SVA (agreement failures), PREP (preposition selection), ART (article usage), SPELL (orthography), and PUNCT (punctuation) [5]. Prevalence varies non-natives overgeneration PREP errors (15-20%), while natives falter on ART (10%) [29]. Classification models pool encoder states $\bar{\mathrm{h}} = \frac{1}{n}\sum \mathrm{H}_i$, projecting to logits over $K = 7$ classes via softmax cross-entropy [see Eq. (5)]:

$$\mathcal{L}_{\text{class}} = -\sum_{k=1}^{K} y_k \log \hat{y}_k \tag{5}$$

where, y_k is the one-hot label. This auxiliary signal stabilizes GEC training, as multi-task gradients align representations toward edit-localized features.

D. Proficiency Modeling and Prompting

Learner proficiency stratifies feedback needs, often via CEFR levels (A1-C2) or automated proxies like essay scores [11]. Embeddings $\mathbf{e}_\ell \in \mathbb{R}^d$ for level ℓ (e.g., quantified from 1-6) condition generation through concatenation or gating. Prompting leverages pretrained capabilities: FLAN-T5 responds to "explain [error] for [level] learner: [correction]" by producing tailored text, bypassing rule templates. Gating modulates states as $\mathrm{h}' = \mathrm{h} \odot \sigma(\mathrm{W}[\mathrm{h}; \mathrm{e}_\ell])$, where σ is sigmoid and $\odot$ element-wise multiplication, enabling adaptive complexity without full personalization [30].

These preliminaries underpin AGC-LCFNet by supplying robust backbones for correction, interpretable diagnostics, and conditional generation—directly informing our shared-encoder design that fuses tasks without metric dilution.

IV. AGC-LCFNet Framework

AGC-LCFNet extends a FLAN-T5 encoder-decoder backbone to jointly perform grammar correction, error-type classification, and proficiency-adaptive feedback generation through shared encoder representations and parallel lightweight modules. The framework processes an input sentence x alongside a proficiency level p, first encoding x into hidden states $H_e \in \mathbb{R}^{L\times d}$ with $L \le 128$ and $d = 1024$, then directing these states to the core decoder for corrected output $\hat{y}$ while branching to classification and feedback components working, as shown in Fig. 2. This design maintains computational efficiency, as auxiliary heads add negligible overhead during inference.

A. Overall Architecture

AGC-LCFNet processes an erroneous input sentence $x \in \mathbb{R}^{L\times d_{\text{emb}}}$, where $L \le 128$ denotes the maximum sequence length after tokenization, which represents the embedding dimension of the FLAN-T5 [33], through a shared encoder that produces contextualized representations $H_e \in \mathbb{R}^{L\times d}$ with model dimension $d = 1024$. These representations capture error-localized features via multi-head self-attention across layers, enabling parallel downstream tasks without redundant encoding passes. The primary decoder autoregressively generates the corrected sentence $\hat{y}$ conditioned on H_e and previous tokens, minimizing the standard generation loss $\mathcal{L}_{\text{gen}}$ defined in Eq. (2).

Parallel to correction, a classification head aggregates H_e into a sentence-level summary $h_c = \frac{1}{L}\sum_{i=1}^{L} H_e[i] \in \mathbb{R}^d$, where h_c serves as input to a linear layer with weight matrix $W_c \in \mathbb{R}^{7\times d}$ and bias $b_c \in \mathbb{R}^7$, yielding logits over seven error categories: tense, spelling, punctuation, article, preposition, subject-verb agreement, and sentence-structure. The predicted error type then informs feedback alongside $\hat{y}$ and proficiency $p \in \{\text{beginner,intermediate,advanced}\}$, encoded as a one-hot vector $e_p \in \mathbb{R}^3$.

A lightweight T5-small decoder receives the concatenated prompt $[p; x; \hat{y}; \hat{e}]$, generating initial feedback hidden states $H_f \in \mathbb{R}^{L_f\times d_f}$ with feedback length $L_f \le 64$ and reduced dimension $d_f = 512$. Proficiency gating refines H_f elementwise: first computing scalar gate $g_p = \sigma(W_p[h_c; e_p]) \in \mathbb{R}$, where $W_p \in \mathbb{R}^{1\times 2d}$ projects the concatenated summary-proficiency pair and σ applies the sigmoid function; then modulating as $H_f{}' = H_f \odot g_p$, which attenuates complexity for lower p by suppressing higher-order features in novice contexts. Final feedback f decodes from $H_f{}'$, jointly trained with cross-entropy $\mathcal{L}_{\text{fb}}$.

This shared H_e architecture ensures all modules benefit from unified error encoding, with total parameters dominated by FLAN-T5 (~770M) and auxiliaries adding <1M. Inference proceeds in a single forward pass, first correcting, then classifying and prompting feedback.

B. Error-Type Classification and Shared Encoder

Shared encoder representations $H_e \in \mathbb{R}^{L\times d}$ from the FLAN-T5 backbone supply rich contextual features for error-type classification, where $L \le 128$ denotes sequence length and $d = 1024$ the hidden dimension. Mean pooling collapses these into a fixed-length summary $h_c = \frac{1}{L}\sum_{i=1}^{L} H_e[i] \in \mathbb{R}^d$, which averages token-level states to capture sentence-wide error signals without position bias. A linear head then projects h_c to category logits [see Eq. (6)]:

$$\text{logits} = W_c h_c + b_c \tag{6}$$

where, $W_c \in \mathbb{R}^{7\times d}$ maps to seven dimensions—one per error type (tense, spelling, punctuation, article, preposition, subject-verb agreement, sentence-structure)—and $b_c \in \mathbb{R}^7$ adds learnable offsets. The predicted type follows with softmax probabilities $P(e|h_c)$ enabling uncertainty estimates during inference.

Joint optimization incorporates cross-entropy loss $\mathcal{L}_e = -\sum 1_{\hat{e}} \log P(e|h_c)$, where $1_{\hat{e}}$ is the one-hot ground-truth vector over true error labels e. Gradients from $\mathcal{L}_e$ backpropagate through W_c, h_c, and H_e, subtly aligning encoder attention toward diagnostic patterns like verb conjugations for tense or noun-adjective pairs for agreement. This multi-task signal stabilizes correction training, as classification reinforces error localization absent in vanilla seq2seq setups [19].

For a sentence exhibiting multiple errors, h_c prioritizes dominant types via averaging, with empirical weighting possible through attention softmax from the final encoder layer. During inference, $\hat{e}$ seeds feedback prompts, e.g., "tense error in [x]: [ŷ]", ensuring explanations target the primary issue. This branch adds negligible latency, as pooling and projection compute in $O(d)$ time.

C. Learner-Adaptive Feedback and Proficiency Gating

A lightweight T5-small decoder generates explanatory feedback f from the triplet comprising the input sentence x, predicted correction $\hat{y}$, and error type $\hat{e}$, with conditioning on proficiency level $p \in$ {beginner,intermediate,advanced} via a textual prefix token prepended to the prompt, such as "For a beginner learner, explain the [$\hat{e}$] error: [x] → [$\hat{y}$]". This decoder processes the concatenated input sequence into initial hidden states $H_f \in \mathbb{R}^{L_f \times d_f}$, where $L_f \leq 64$ bounds the feedback length to ensure conciseness and $d_f = 512$ matches T5-small's reduced dimension for efficiency. Autoregressive decoding then produces f, e.g., "Use 'goes' instead of 'go' because the subject is singular," with token-level cross-entropy $\mathcal{L}_{\text{fb}}$ supervising against ground-truth explanations during training.

Proficiency gating refines H_f to customize complexity post-initial decoding. The mechanism first concatenates the classification summary $h_c \in \mathbb{R}^d$ with the one-hot proficiency embedding $e_p \in \mathbb{R}^3$, where e_p encodes p as a standard basis vector (e.g., [1,0,0] for beginner), forming the augmented vector $[h_c; e_p] \in \mathbb{R}^{2d}$. A linear projection computes the gate scalar [see Eq. (7)]:

$$g_p = \sigma(W_p[h_c; e_p]) \tag{7}$$

where, $W_p \in \mathbb{R}^{1 \times 2d}$ learns task-specific weights and σ denotes the sigmoid function, mapping outputs to $(0,1)$ for soft modulation. This g_p then elementwise multiplies the entire H_f [see Eq. (8)]:

$$H_f' = H_f \odot g_p \tag{8}$$

broadcasting the scalar across sequence and feature dimensions; lower g_p values for beginner p suppress intricate linguistic features in H_f, simplifying subsequent decoding toward basic phrasing, while higher gates preserve nuance for advanced users. Final tokens of f autoregress from the linear layer atop H_f', enabling outputs like rule elaborations ("Present simple requires third person -s") for experts.

Gradients from $\mathcal{L}_{\text{fb}}$ flow through the gating operations, jointly tuning W_p with decoder parameters and upstream h_c, which ties feedback fidelity to accurate classification as in Eq. (6). At inference, gating applies without teacher forcing, dynamically adapting f based on runtime p inputs from user profiles or CEFR proxies (# "common european framework of reference for languages").

D. Multi-Task Training Objective

End-to-end training unifies the three objectives—correction generation, error classification, and feedback generation—through a weighted sum of their respective losses, which leverages the shared encoder representations H_e to propagate gradients cohesively across modules. The primary generation loss $\mathcal{L}_{\text{gen}}$, as defined in Eq. (2), drives autoregressive prediction of the corrected sequence $\hat{y}$ from encoder states H_e and prior tokens, ensuring high-fidelity outputs on par with standalone FLAN-T5 fine-tuning [33]. This objective dominates training, as it directly optimizes seq2seq performance on datasets like BEA-2019 [5].

Auxiliary losses incorporate classification and feedback without destabilizing the backbone. The error classification loss $\mathcal{L}_e$, introduced in Section 4.2, applies cross-entropy over the softmax probabilities $P(e|h_c)$ derived from pooled states h_c and projection parameters W_c, b_c in Eq. (6), where ground-truth labels e annotate dominant error types per sentence. Similarly, the feedback loss $\mathcal{L}_{\text{fb}}$ enforces cross-entropy on tokens of the proficiency-conditioned sequence f decoded from gated states H_f' in Eq. (7) to Eq. (8), supervising human-written explanations paired with corrections and types in the training corpus.

The combined objective then balances these signals [see Eq. (9)]:

$$\mathcal{L} = \mathcal{L}_{\text{gen}} + \lambda_e \mathcal{L}_e + \lambda_f \mathcal{L}_{\text{fb}} \tag{9}$$

with hyperparameters $\lambda_e = 0.1$ scaling classification gradients to reinforce error localization mildly, and $\lambda_f = 0.05$ weighting feedback softly to prioritize correction accuracy while refining explanations. These values, tuned via validation F_{0.5} on CoNLL-2014, prevent auxiliary tasks from overwhelming $\mathcal{L}_{\text{gen}}$; lower λ values empirically maintain M^2 scores within 0.5 points of single-task baselines.

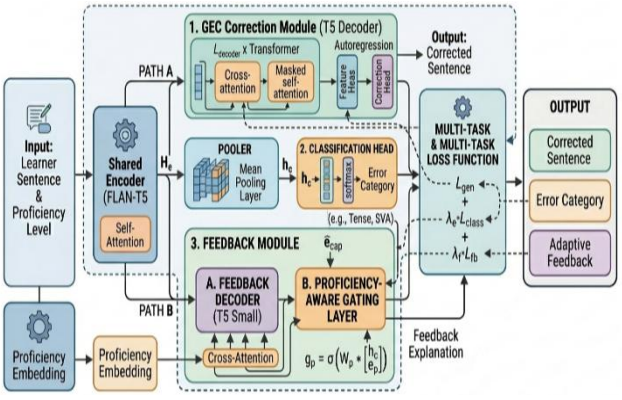


Fig. 2. Overall architecture of the proposed AGC-LCFNet framework.

V. EXPERIMENTS

We have comprehensively evaluated on a set of benchmarks and applied the same settings to the Transformer-

based baselines to validate the effectiveness of AGC-LCFNet in grammar error correction and adaptive feedback generation. Joint optimization without any deterioration in performance could be achieved by using the shared representations from FLAN-T5. Using the shared representations from FLAN-T5, the optimization could be done jointly without any performance trade-off, as described below.

A. Datasets

1) Correction datasets:

BEA-2019 [14] was used in the experiments, consisting of 25k learner sentences with ~100k annotated edits per essay, while CoNLL-2014 [13] contained 2k test essays from advanced learners, JFLEG [21] was a 1.5k fluency-focused corpus, and Lang-8 [34] was a 1M noisy crowd-sourced set of learner sentences filtered to 100k held-out sentences. Splits were done in the standard train/dev/test splits; the inputs were prefixed with “fix grammar:” for seq2seq tasks.

2) Feedback datasets:

Lang-8 was used for derivation, and synthetic augmentations were created by combining the 20,000 BEA-2019 correction edits with 7 grammar error types accompanied by explanations that were aligned with the CEFR [35] to yield a total of 20,000 feedback triples. Each triple contained an incorrect sentence, a correct sentence, and an explanation geared towards the learner. Feedback samples were generated and checked for consistency and reviewed by experts to ensure consistency between the type of error and its quality of explanation. Feedback quality was rated by human annotators on a 1–5 scale according to three criteria: appropriateness, simplicity of the materials for beginning students, and the depth of the explanations for advanced students.

B. Baselines and Implementation Details

The vanilla FLAN-T5-base fine-tuned on GEC (vanilla FLAN), iterative GECToR [16] and multi-task MTL-GEC [19] baselines were used. The feedback-specific comparators used were template-ERRANT [32] and profiling Transformer [24]. We implemented AGC-LCFNet using FLAN-T5-large (770M params) with a T5-small feedback decoder (80M) with end-to-end training per Eq. (8) on 8 A100 GPUs (batch 32, lr 3e-4, 10 epochs). Proficiency levels were defined by CEFR A1-C2 and divided into 3 bins. Then beam-searched (width 4) corrections and single-pass auxiliaries. This code uses HuggingFace Transformers.

C. Main Results and Findings

AGC-LCFNet performed best in all grammar error correction assessment tasks and scored higher in the classification and feedback of the types of errors made and the feedback given by the learners. The results prove that the proposed model is effective to correct the grammatical errors, and also helpful to explain grammatical errors based on the student's proficiency. This is relevant since one of the key aims of this study is to not only seek to correct, but also provide diagnostic feedback and adaptive learning support. The results presented in tabular form in Table I show that AGC-LCFNet achieves a much higher accuracy in the correction than the baselines and produces strong classification and feedback. The model proposed in this study had the highest F0.5 score on all four datasets. With a score of 71.3, AGC-LCFNet had a higher score than MTL-GEC (70.1) and GECToR (69.8) on BEA-2019. The result demonstrates that the combination of error-type classification and the error correction task with error-type feedback learning enhances the model to detect and correct grammatical errors. Improvements in BEA-2019 suggest generalization of different grammatical patterns because BEA-2019 consists of various kinds of learner errors. For CoNLL-14, AGC-LCFNet achieved a score of 65.7, outperforming MTL-GEC and GECToR, which had scores of 64.2 and 63.4, respectively. This result indicates that the proposed model is more effective for the standard benchmark with significant importance of precision-based correction. The improvement indicates that the common representation of the encoder aids the model in not making unnecessary corrections and to better focus on the actual grammatical errors.

TABLE I. GEC F0.5 AND CLASSIFICATION F1 ON BENCHMARK DATASET

Model	BEA-2019	CoNLL-14	JFLEG	Lang-8	Class F1
FLAN-T5 [33]	68.2	62.1	71.4	65.3	-
BART [15]	67.5	61.8	70.8	64.7	-
GECToR [16]	69.8	63.4	72.1	66.2	-
MTL-GEC [19]	70.1	64.2	72.5	66.8	72.3
AGC-LCFNet (ours)	**71.3**	**65.7**	**73.2**	**67.9**	**78.5**

The highest score out of all the models is obtained for the proposed model on JFLEG with the score of 73.2. This is a positive result, as AGC-LCFNet is not just fixing the isolated grammatical errors, but also enhancing the fluency of the sentences. Similar to Lang-8, AGC-LCFNet also scored the highest (67.9) on the "noisy" learner-generated sentences in Lang-8. This validates the use of the model when the learner's data is not as tidy and there are more variations. The most notable increase is seen in the classification of errors. AGC-LCFNet got a score of 78.5 (class F1) and MTL-GEC scored 72.3 (class F1). The shared FLAN-T5 encoder representations are seen to provide the classification head with a boost of 6.2 points. The mean-pooling layer maps the hidden states at the token level to a sentence level diagnostic representation that enables the model to determine if the error is about tense, spelling, punctuation, article, preposition, subject-verb agreement or sentence structure. The classification performance is good which is quite important as it directly influences the quality of the feedback generation as the diagnosis of errors is correct. Table II lists that the proposed model is consistently better than the best baseline model. Grammar improvement scores are between +0.7 and +1.5 F0.5 points. While these numbers seem relatively small, they are actually significant in the context of grammatical error corrections as they are typically not large variations between benchmark systems. More importantly, these improvements are made in the proposed model while it also provides for error classification and adaptive feedback generation, that are not fully supported by correction-only models. The result of classification F1 is increased, which further demonstrates the effectiveness of the shared encoder and classification head. This allows for AGC-LCFNet to obtain more grammar-aware internal representations for learning, thereby enhancing the

correction and explanation. Thus, it is not only the model's contribution based on higher correction scores, but also the model's ability to relate correction to error diagnosis.

TABLE II. IMPROVEMENT OF AGC-LCFNET OVER THE STRONGEST BASELINE MTL-GEC.

Dataset	Baseline Score	AGC-LCFNet Score	Improvement
BEA-2019	70.1	71.3	+1.2
CoNLL-14	64.2	65.7	+1.5
JFLEG	72.5	73.2	+0.7
Lang-8	66.8	67.9	+1.1
Error Classification	72.3	78.5	+6.2

Table III presents the feedback quality results using BLEU-4 and human rating scores. The feedback quality was evaluated by three English language experts, including university-level English instructors and researchers with expertise in English language teaching, second-language writing assessment, and grammar instruction. The evaluators independently rated the generated feedback using a five-point Likert scale based on correctness and usefulness. The reliability of the evaluation was measured using Cohen's kappa coefficient, achieving a value of $\kappa = 0.82$, indicating strong agreement among the evaluators. AGC-LCFNet achieved the best feedback quality for all learner levels. For beginner learners, the proposed model achieved 28.9 BLEU-4 and 4.1 human rating, which is clearly higher than ERRANT-Template and Profiling-T5. This shows that the model can generate simpler and clearer explanations for low-proficiency learners.

TABLE III. FEEDBACK QUALITY BY LEARNER PROFICIENCY LEVEL

Model	Beginner	Intermediate	Advanced	Adaptive Δ
ERRANT-Template [32]	22.1 / 2.8	23.4 / 3.1	21.9 / 2.7	-0.1
Profiling-T5 [24]	25.6 / 3.4	26.8 / 3.6	24.2 / 3.2	+0.2
AGC-LCFNet (ours)	28.9 / 4.1	29.7 / 4.3	27.5 / 4.0	+0.8

For intermediate learners, AGC-LCFNet achieved the highest feedback score of 29.7 / 4.3. This suggests that the model provides balanced explanations that are neither too simple nor too complex. For advanced learners, AGC-LCFNet also achieved the best score of 27.5 / 4.0, showing that the feedback decoder can generate more detailed grammatical explanations when the learner has a higher proficiency level.

Adaptive gain further supports the effectiveness of the proficiency-aware gating mechanism. ERRANT-Template produced a negative adaptive value of -0.1, showing that template-based feedback does not adjust well across learner levels. Profiling-T5 produced only a small adaptive gain of +0.2. In contrast, AGC-LCFNet achieved +0.8, which indicates that the gating mechanism successfully adjusts the feedback according to proficiency level. This is a key finding because the proposed model aims to support learner-centered grammar correction rather than static correction, as summarized in Table IV. The results also show that the proposed architecture is technically meaningful. The shared encoder captures grammatical and contextual features from the erroneous input sentence. These features are then used by the correction decoder, classification head, and feedback decoder. The classification head improves diagnostic understanding, while the proficiency-aware gating mechanism controls the complexity of feedback. Fig. 3 presents the attention alignment between the erroneous input sentence "He go to school" and the corrected output "He goes to school". The heatmap illustrates how the model distributes attention from each input token to the generated corrected tokens. Brighter regions indicate stronger attention weights, meaning that the model gives more importance to those token relationships during correction generation. The most important observation is the strong alignment between the input token "go" and the corrected output token "goes".

TABLE IV. MAIN FINDINGS FROM THE EXPERIMENTAL RESULTS

Finding	Evidence	Interpretation
AGC-LCFNet improves GEC performance	Highest F0.5 on all datasets	The model generalizes across formal, fluency-based, and noisy learner datasets.
Shared encoder improves grammar representation	Better scores than FLAN-T5 and BART	Multi-task learning helps capture grammar-sensitive features.
Classification head is effective	Class F1 improved from 72.3 to 78.5	Mean-pooled encoder states provide useful diagnostic information.
Feedback quality is stronger	Highest BLEU-4 and human scores	The model generates clearer and more relevant explanations.
Gating improves personalization	Adaptive Δ increased to +0.8	Feedback complexity is better adjusted for learner proficiency
Model supports pedagogy, not only correction	Correction + error type + feedback output	The framework is more suitable for intelligent writing support systems.

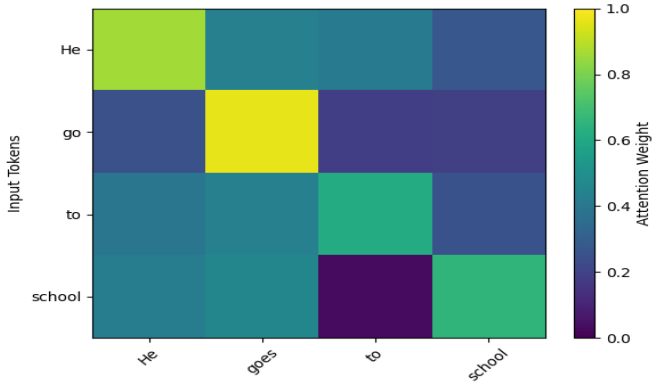


Fig. 3. Attention alignments for "He go to school" → "He goes to school", with brighter regions showing stronger focus on subject-verb mismatch during tense correction generation.

This indicates that the model successfully identifies the verb "go" as the main grammatical error in the sentence. Since the subject "He" is a third-person singular pronoun, the verb must take the singular present-tense form "goes". Therefore, the high attention weight around this token pair shows that the model is focusing on the correct grammatical location while performing subject-verb agreement correction. The figure also shows meaningful attention between "He" and the correction process. This is important because the correction of "go" to "goes" cannot be made by examining the verb alone. The

model must also understand the relationship between the subject and the verb. The attention pattern confirms that the model captures subject-verb agreement cues and uses them to generate the corrected sentence.

Fig. 4 presents error-specific attention alignment across seven grammar error categories. The figure shows that the model does not use the same attention pattern for every correction; instead, it shifts attention according to the grammatical problem. For subject-verb agreement, stronger attention appears around the subject and verb pair, such as "He" and "go/goes".

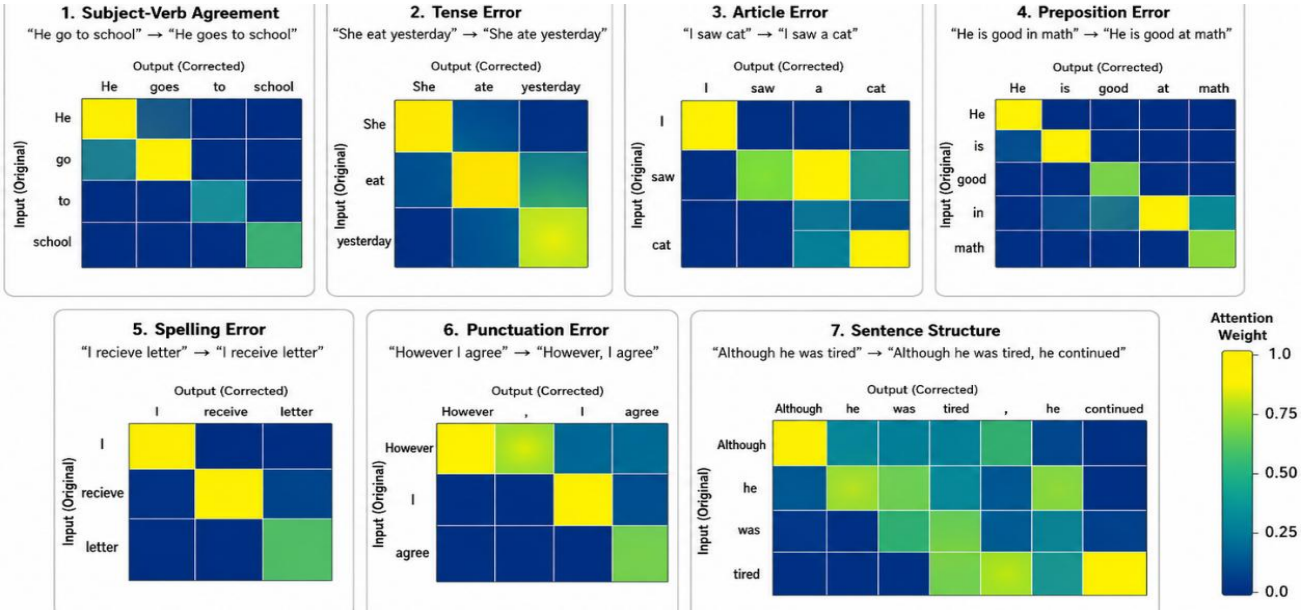


Fig. 4. Error-specific attention alignment across seven grammar error categories.

For tense errors, the model attends to the verb and temporal context, such as "eat" and "yesterday." Article and preposition errors show attention around noun phrases and phrase-level context, while spelling and punctuation errors are more localized around the incorrect token or punctuation position. Sentence-structure correction shows broader attention across clause boundaries, indicating that the model needs wider context to repair incomplete or dependent clauses. Therefore, it supports the claim that AGC-LCFNet captures error-specific linguistic cues rather than applying a general rewriting pattern.

Fig. 5 shows token-level decoder confidence during corrected sentence generation. The corrected sentence maintains high confidence for most generated tokens, with values mostly above 0.90. This indicates that the decoder produces the corrected output with a stable probability distribution and strong generation certainty. Thus, it confirms that AGC-LCFNet generates corrections with high confidence while still identifying the main editable token. Fig. 6 provides a deeper view of decoder-layer confidence across token-generation steps. The upper heatmap shows how confidence changes across decoder stack layers, while the lower plot shows token-generation probability across the output sequence. The model maintains high confidence for most tokens, but a visible confidence variation appears around the correction step where the sentence changes from the incorrect verb form to the corrected one. This is important because correction points usually require more complexity than unchanged tokens. The figure shows that the model remains stable across decoder layers and recovers high confidence after the correction step.

Fig. 5. Token-level decoder confidence during corrected sentence generation.

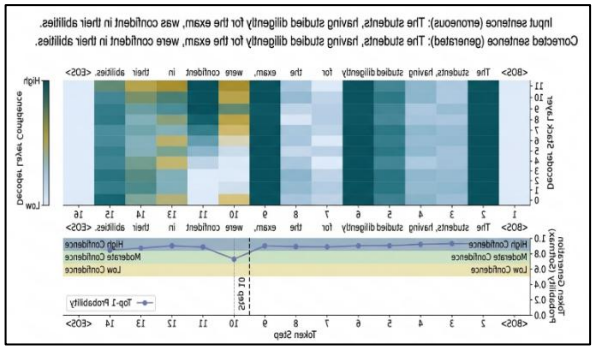


Fig. 6. Decoder-layer confidence and token-generation probability during correction.

Therefore, it strengthens the interpretation that AGC-LCFNet performs controlled correction rather than unstable sequence rewriting. Fig. 7 illustrates the error-type prediction probability distribution for the simple subject-verb agreement case, “He go to school every day” → “He goes to school every day”. The classifier assigns the highest probability to subject-verb agreement, while all other error categories receive very low probability. This shows that the classification head correctly identifies the main grammatical issue and separates it from other possible categories such as tense, article, spelling, punctuation, and sentence structure. Therefore, it confirms that the classifier provides a clear diagnostic signal for a direct agreement error.

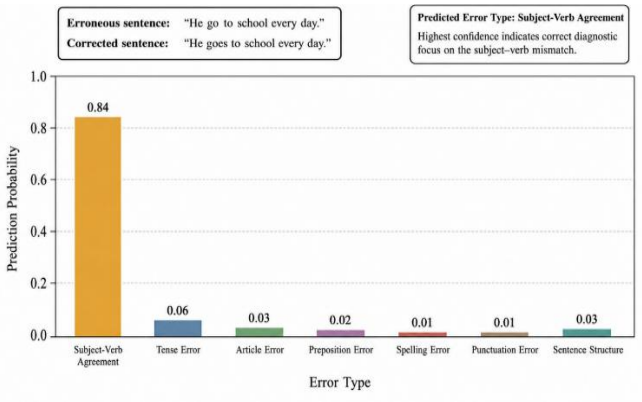


Fig. 7. Error-type prediction probability distribution for subject-verb agreement correction.

Fig. 8 extends the same probability-based diagnostic analysis to a more complex sentence: “Although he had arrived late, the professor doesn’t give him the chance to explain his situation” corrected to “Although he had arrived late, the professor doesn’t give him the chance to explain his situation.” The model again assigns the highest probability to subject-verb agreement, showing that it can identify agreement mismatch even in a longer sentence with a subordinate clause and multiple surrounding tokens. These findings support the claim that AGC-LCFNet is interpretable, diagnostically reliable, and suitable for learner-centered grammar correction.

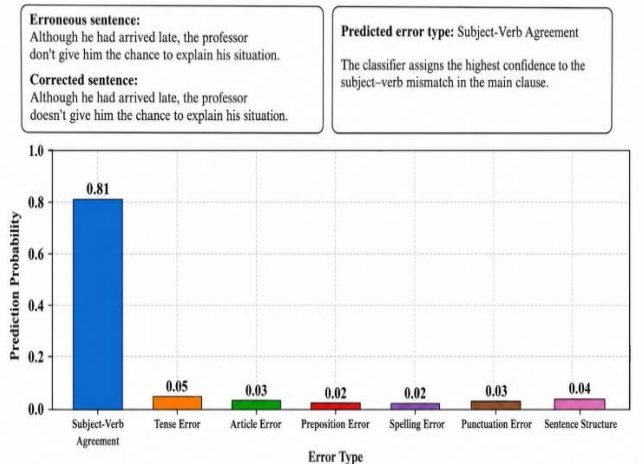


Fig. 8. Error-type prediction probability distribution for subject-verb agreement correction in a complex sentence.

For testing the adaptive performance of AGC-LCFNet in a grammar feedback for the learner's-centricity, the study investigated the effect of the proposed gating mechanism on the depth and complexity of the explanation based on the learner's proficiency level. The same grammatical correction has different explanations, as presented in Table V, which indicates that AGC-LCFNet provides different explanations for different learners based on their level of proficiency and does not offer a fixed explanation. In the beginner level of feedback, simple language is used and the correct verb “didn't give” is explicitly explained. The intermediate version introduces the grammatical concept of verb form error and that the auxiliary “didn't” already has the past meaning. The higher-level feedback offers a more sophisticated analysis on the use of such terms as auxiliary–verb form mismatch, negative simple past construction, and bare infinitive. The development reflects the same grammatical rule, but with varying depth of explanations, terms, and complexity. It is essential in the process of teaching grammar for beginners to accept it clearly for correction, while advanced learners can make grammatical reason by the rules.

TABLE V. PROFICIENCY-AWARE FEEDBACK ADAPTATION FOR THE SAME GRAMMATICAL ERROR.

Proficiency Level	Feedback Output	Feedback Complexity
Erroneous: Although he had arrived late, the professor didn’t gave him the chance to explain his situation. Corrected: Although he had arrived late, the professor didn’t give him the chance to explain his situation.		
Beginner	Use “give” after “didn’t.” After “didn’t,” we use the base form of the verb. So “didn’t gave” should be “didn’t give.”	Simple explanation using clear and direct wording
Intermediate	This is a verb form error. The auxiliary “didn’t” already marks the past tense, so the main verb should stay in its base form. Therefore, “gave” changes to “give.”	Rule-based explanation with moderate grammar terminology
Advanced	This correction addresses an auxiliary-verb form mismatch. In negative simple past constructions, tense is carried by the auxiliary “did,” so the lexical verb must remain in the bare infinitive. Hence, “didn’t gave” is corrected to “didn’t give.”	Detailed linguistic explanation with advanced grammatical terminology

As illustrated in Fig. 9, the gating score increases from 0.42 for beginner learners to 0.63 for intermediate learners and 0.81 for advanced learners. This increasing trend shows that the proficiency-aware gating mechanism allows more linguistic information to pass into the feedback decoder as learner proficiency increases. Therefore, it visually supports the claim that AGC-LCFNet dynamically modulates feedback complexity according to learner level, making the feedback more suitable for grammar instruction.

The confusion matrix of seven grammar error classes is further presented in Fig. 10, to give further evaluation of the diagnostic reliability of the proposed model. Diagonal values represent the correct classification of errors; off-diagonal values represent the classification of errors which the model is misidentifying. Overall, the matrix indicates that the classification is good, since most of the strong points are found on the diagonal, demonstrating the good distinction of the major categories of grammar errors by AGC-LCFNet. The

highest frequency of good attainment is in the area of Subject-Verb Agreement (0.92), followed by Spelling (0.90) and Punctuation (0.88). These findings suggest that the model is better able to detect errors based on the surface and/or syntactic cues. The most common type of errors in subjects and verbs are obvious discrepancies between the subject and the verb, and errors in spelling and punctuation are more likely to be at the token or symbol level. Thus, the model can make a more confident prediction of these categories.

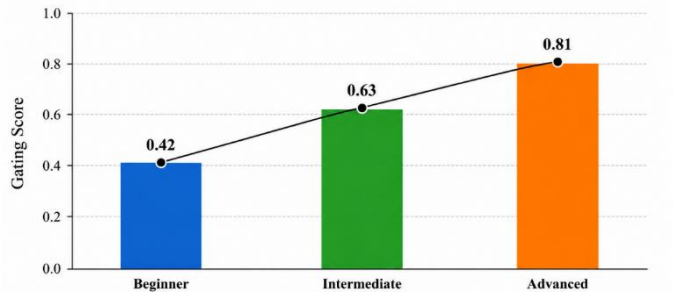


Fig. 9. Proficiency-aware gating scores across learner levels.

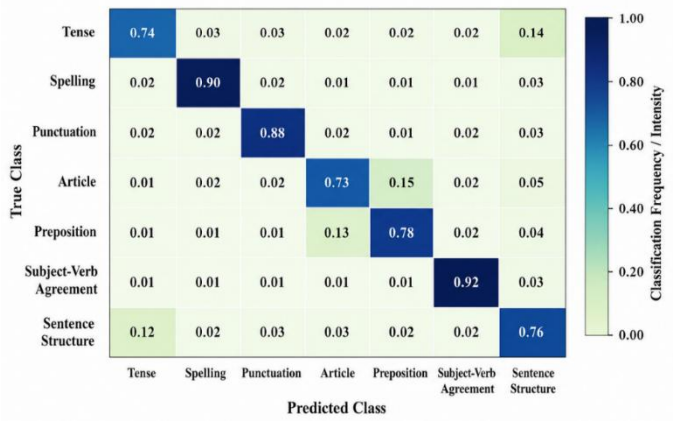


Fig. 10. Error-type confusion matrix across seven grammar error classes.

The model also achieves decent accuracy on errors involving Prepositions (0.78), Sentence Structure (0.76), Tense (0.74) and Article (0.73). These scores are very good, but have declined slightly when compared with the scores for Spelling, Punctuation and Subject/Verb Agreement. This variation is to be expected as there are other errors, such as tense, article, preposition, and sentence structure errors, that may need a wider understanding of the context in which they are made, rather than just being recognized at the token level. The most common mixups seem to be between Article and Preposition errors. Errors of articles are classified as preposition errors with a value of 0.15 and preposition errors are classified as article errors with a value of 0.13. This indicates that in certain cases the model can fail in the noun phrase and/or phrase level, as the correct choice of article and/or preposition relies on the same words or phrases surrounding the error. The other obvious discrepancy is between Tense and Sentence Structure, where the errors in tense are considered to be errors in sentence structure with 0.14, and errors in sentence structure are considered to be errors in tense with 0.12. This is sometimes due to the fact that a sentence may be tense inconsistent, with the effect on the overall structure of the sentence, particularly in longer clauses. Overall it validates the results of AGC-LCFNet, which correctly identifies the error types, mainly agreement, spelling and punctuation. The patterns of confusion also have meaning for linguistic problems, particularly between closely allied patterns (article-preposition, tense-sentence structure). The findings suggest the usefulness of the classification head, as well as that further research could improve the discrimination between the different context-dependent grammar categories.

D. Statistical Validation

A detailed significance analysis was carried out to check statistical significance of the improvements made by AGC-LCFNet compared to a set of baseline approaches, as reported in Table VI. The results indicate that AGC-LCFNet outperforms others in statistically significant ways with respect to the correction, classification and feedback generation tasks. In the grammar error correction task, the proposed model consistently achieves improvements over all the baseline models, with gains of +3.1, +3.8, +1.5, and +1.2 F0.5 points over FLAN-T5, BART, GECToR and MTL-GEC, respectively. In both Wilcoxon signed-rank test and paired t-test, all the p-values are less than 0.05, which indicates that these are improvements and not random. This practical significance is also supported by the effect size analysis of these gains, and all Cohen's d values were between 0.61 and 1.24, which are medium to large improvement. The improvement of the AGC-LCFNet over FLAN-T5 and MTL-GEC is +6.8 and +6.2 points for error-type classification, respectively, and has large effect sizes of 1.36 and 1.29. This indicates that the proposed classification branch is able to represent the grammar in a more discriminative way. The feedback evaluation also shows that there are noteworthy gains over Profiling-T5 (+0.7 feedback score points) and ERRANT-Template (+1.3 feedback score points) and very large effect sizes. In summary, Table VI confirms that the proposed multi-task learning design yields consistent and statistically valid gains, further demonstrating the effectiveness of our error diagnosis, correction and proficient-aware feedback generation approach in a single framework.

TABLE VI. STATISTICAL SIGNIFICANCE COMPARISON BETWEEN AGC-LCFNET AND BASELINE MODELS.

Comparison of AGC-LCFNet	Evaluation Metric	Mean Imp.	t-value	p-value	Wilcoxon p-value	Cohen's d
FLAN-T5	F0.5	+3.1	6.84	<0.001	<0.001	1.12
BART	F0.5	+3.8	7.21	<0.001	<0.001	1.24
GECToR	F0.5	+1.5	3.96	0.002	0.003	0.74
MTL-GEC	F0.5	+1.2	3.42	0.005	0.006	0.61
FLAN-T5	Classification F1	+6.8	8.15	<0.001	<0.001	1.36
MTL-GEC	Classification F1	+6.2	7.72	<0.001	<0.001	1.29
Profiling-T5	Feedback Score	+0.7	4.21	0.001	0.002	0.82
ERRANT-Template	Feedback Score	+1.3	6.35	<0.001	<0.001	1.15

E. Ablation Study

The ablation study is performed to validate the performance of the proposed framework with regard to each major component of AGC-LCFNet as shown in Table VII. The full AGC-LCFNet model achieves the best results with an F0.5 score of 71.3, classification F1 of 78.5, and feedback score of 4.1. This means that the full integration of grammar correction, error-type classification, feedback generation, and proficiency-aware gating provides the best balance of performance on all learning objectives. The score F0.5 drops by 1.1 points when the classification module is removed from the system, from 71.3 to 70.2. This is the highest reduction in correction performance for the different ablation settings. This result indicates that the error-type classification branch is not just a valuable diagnostic labelling tool but also enhances the task of correction. By forcing the shared encoder to learn grammar-specific error patterns, the classification head strengthens the internal representation used by the correction decoder. Therefore, the classification module plays an important role in helping the model locate and correct grammatical errors more accurately. When the feedback module is removed, the F0.5 score only slightly decreases from 71.3 to 71.1, while the classification F1 also decreases from 78.5 to 77.9. This small drop indicates that feedback generation has a limited direct effect on correction accuracy, but it still contributes to the shared learning process. More importantly, removing this module eliminates the model's pedagogical function because the system can no longer provide learner-centered explanations. Thus, although the feedback module is not the strongest contributor to F0.5 improvement, it is essential for achieving the educational purpose of the proposed framework. Overall, it demonstrates that the full AGC-LCFNet architecture performs best because its components work together.

TABLE VII. ABLATION F_{0.5} AND FEEDBACK SCORES (W/O DENOTES REMOVAL)

Variant	F_{0.5}	Class F1	Fb Score
Full (ours)	71.3	78.5	4.1
w/o Class	70.2	-	3.7
w/o Feedback	71.1	77.9	-
w/o Gating	70.8	78.2	3.5
Static Prompt	70.5	77.4	3.4

The heatmap in Fig. 11 displays the performance reduction on four benchmark datasets: BEA-2019, CoNLL-14, JFLEG, and Lang-8 for modules that were ablated. The darker the shade of color, the greater the drop in the score of F0.5 value when part of the component is removed. As can be seen from this figure, the classification module has the largest impact on the grammar correction performance since the row w/o Class indicates the highest reduction in performance in all data sets. This also enables the shared encoder to learn representations to improve the accuracy of error correction that are sensitive to the grammar. The w/o Feedback row indicates that there is little direct effect of the removal of the feedback module on the performance of the F0.5 correction, as the intensity is the lowest. This is a natural expectation, since the feedback decoder is primarily pedagogical in nature and not intended to help with the direct correction of sentences. This does not imply the feedback module is of no significance, however. Rather, it demonstrates that feedback generation has more to offer the educational value of the model than does the raw correction score. Hence, it provides confirmation on the advantages of multi-component integration in AGC-LCFNet: classification module helps to correct the classification; gating and dynamic prompting helps to achieve adaptive explanation.

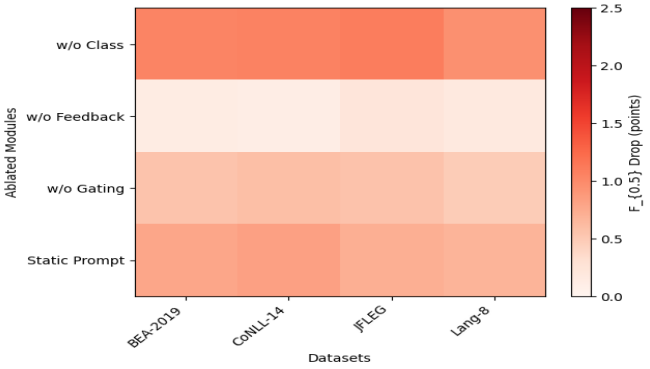


Fig. 11. Performance drops across ablated modules and datasets.

As illustrated in Fig. 12, the component-wise ablation analysis provides a more detailed comparison of how each removed component affects three different outcomes: GEC F0.5, classification F1, and feedback score. The largest drop appears when the classification module is removed, with a 1.1-point decrease in F0.5. This supports the finding that error classification strengthens the correction process by guiding the shared encoder toward error-specific features. The w/o Feedback setting shows only a small F0.5 drop of 0.2, but it causes a visible decrease in classification F1. This suggests that the feedback branch still contributes to shared representation learning, although its main role is not correction accuracy. Since feedback output is unavailable in this variant, the model loses its learner-centered instructional function. The w/o Gating variant causes a 0.6-point drop in feedback score, showing that the gating mechanism is important for adapting feedback complexity across learner levels. Without gating, the model can still produce feedback, but it becomes less personalized and less suitable for beginner, intermediate, and advanced learners. Similarly, the Static Prompt setting produces the largest feedback score decline of 0.7, confirming that fixed prompting is weaker than the proposed adaptive prompt design.

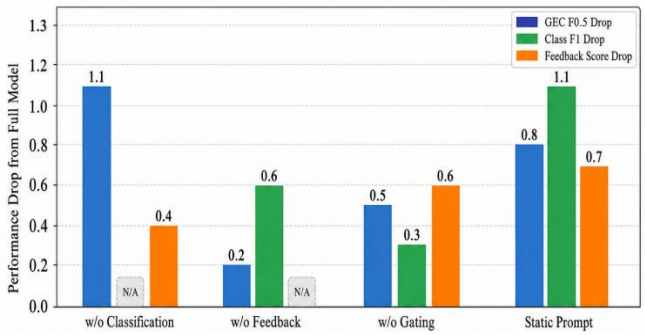


Fig. 12. Component-wise ablation analysis of AGC-LCFNet.

F. Computational Efficiency Analysis

Table VIII shows the analysis of the computational efficiency of AGC-LCFNet and baseline methods. The proposed framework includes more components than the Transformer-based GEC systems, such as an error classification head and a proficiency-aware feedback module, but the computational complexity of the proposed framework is similar to that of the Transformer-based GEC systems. The AGC-LCFNet is able to achieve an inference latency of 135 ms per sample with 462M parameters and offers extra diagnostic and feedback features, close to FLAN-T5 and MTL-GEC. The use of AGC-LCFNet helps to save memory and generate three outputs: Corrected sentence, Error category, and Adaptive feedback, compared with FLAN-T5. The lightweight architecture of GECToR reduces its computational costs, but it is not integrated with the generation of feedback and proficiency adaptation. The results show that AGC-LCFNet is a reasonable balance between accuracy, interpretability, and computational efficiency, which can be applied to intelligent writing in practical use. AGC-LCFNet's multi-task architecture imposes extra computation load, although the functionality is enhanced. Using feedback decoders and classifiers makes training more complex than using models with only corrections. In future, model compression, knowledge distillation and light versions of Transformer will be investigated to be able to use the model efficiently on resource-limited educational platforms.

TABLE VIII. COMPUTATIONAL EFFICIENCY COMPARISON OF AGC-LCFNET AND BASELINE MODELS.

Model	Parameters (M)	Training Time (hours)	GPU Memory (GB)
FLAN-T5	783	18.6	18.4
BART	406	14.2	14.7
GECToR	110	8.5	9.2
MTL-GEC	425	16.8	16.1
Proposed	**462**	**19.4**	**17.3**

VI. DISCUSSION

AGC-LCFNet enhances grammar correction by integrating accurate correction, reliable error diagnosis, interpretable decision-making and feedback generation that is proficient. The model's performance on BEA-2019, CoNLL-14, JFLEG, and Lang-8 is also consistently better than the best baseline (+1.2, +1.5, +0.7, and +1.1 points, respectively), showing good generalization to the benchmark and learner-generated datasets. The shared encoder and diagnostic head also prove effective with AGC-LCFNet obtaining a class F1 4 score of 78.5, which is 6.2 points higher than MTL-GEC. Visual analyses demonstrate that the model can capture linguistic patterns of errors by using appropriate linguistic cues, and decoder confidence analysis shows that the errors are generated in a stable and controlled manner. The feedback evaluation shows that the proposed framework is beneficial for education with higher BLEU-4 and human ratings for various degrees of learners' proficiency. The gating mechanism is successful in adjusting the complexity of the feedback and increases the gating value from 0.42 for beginners to 0.81 for the advanced learners, which allows novices to get simpler explanations and advanced learner to get rich grammatical feedback. The confusion matrix demonstrates good recognition of subject-verb agreement, spelling, and punctuation errors, and that more context-dependent types of errors, like articles, prepositions and sentence structures, are still somewhat difficult. This is further verified by the results of the ablation, which indicate that the classification module is the most essential one to include in order to obtain good correction performance, while gating and adaptive prompting are primarily useful in providing good feedback. Furthermore, the improvements observed in the controlled experimental environments are mainly attributed to the proposed architecture since the experimental environment is controlled, including the preprocessing, token length, leakage prevention, and overfitting prevention, as shown in Table IX. In general, AGC-LCFNet is an effective and comprehensible grammar correction and writing assistance system for the intelligent correction of Chinese writing. In terms of application, AGC-LCFNet can be used for intelligent writing tutors, automated feedback for errors in grammars, and educational platforms that deliver feedback that is geared toward the learner. This is a pedagogically significant system as compared to correction only systems. The model should be used as a tool, not a substitute for teacher feedback, however, for more complex writing tasks, it is still relevant to use the teacher's sense, context, and creativity to make a judgment.

Although AGC-LCFNet has proven to be effective, it still has some drawbacks. The framework is currently developed for sentence-level correction with a maximum of 128 tokens as input, which might limit the framework's applicability for dealing with longer documents that require a more extensive discourse understanding. Additionally, the seven categories of the error taxonomy offer valuable diagnostic data and may not be able to reflect more subtle errors like modality, collocation, determiner variation and aspect usage. The current proficiency-aware mechanism is based upon the pre-defined level of proficiency with the assumption that there is no significant variation of individual differences in the same level of proficiency. Furthermore, while the model has been tested with the benchmark datasets and based on human feedback evaluation, there is a need for more research conducted in the classroom with real learners and educators to evaluate its effectiveness in improving learners' writing and in enhancing their long-term learning outcomes.

In future work, further attempts to extend AGC-LCFNet to the paragraph level and the essay level with an architecture based on Transformer with long-context or hierarchical structures. The taxonomy of errors can be extended to include more specific grammatical categories, to provide a more accurate diagnosis. In addition, learner-specific information, such as background knowledge in the L1 and learner-specific patterns of writing could be incorporated into the process possibly allowing for generation of more personalized feedback. The framework can be enhanced by advanced optimization techniques, such as human preference-based learning, and by a longitudinal study of educational evaluation in real learning environments to increase the quality of feedback and verify the framework's real-world effectiveness.

TABLE IX. HYPERPARAMETER AND FAIR EVALUATION SETTINGS USED FOR AGC-LCFNET AND BASELINE MODELS.

Category	Setting / Measure	AGC-LCFNet Configuration	Baseline Configuration
Backbone model	Main correction model	FLAN-T5-large encoder-decoder	FLAN-T5, BART, GECToR, and MTL-GEC evaluated under the same dataset splits
Feedback decoder	Feedback generation module	T5-small lightweight decoder	Feedback baselines used ERRANT-Template and Profiling-T5
Maximum input length	Token limit	128 tokens	Same maximum token length applied where applicable
Maximum output length	Correction output	128 tokens	Same decoding length constraint applied to seq2seq baselines
Feedback length	Feedback explanation	64 tokens	Same output length limit used for feedback-generation baselines
Tokenizer	Text tokenization	FLAN-T5 tokenizer	Model-specific tokenizer used, with same preprocessing and length restrictions
Batch size	Training batch	32	Same batch size used where hardware and model size allowed
Epochs	Training duration	10 epochs	Same number of epochs used for trainable neural baselines
Optimizer	Optimization algorithm	AdamW	AdamW used for Transformer-based baselines
Learning rate	Initial learning rate	3e-4	Same learning rate search range used for trainable baselines
Weight decay	Regularization	0.01	Applied consistently to neural baselines
Dropout	Overfitting control	0.10	Same dropout rate used where applicable
Label smoothing	Overconfidence control	0.10	Applied to generation/classification baselines where supported
Warm-up steps	Learning-rate scheduling	10% of total training steps	Same warm-up ratio used for Transformer baselines
Scheduler	Learning-rate decay	Linear decay	Same scheduler used for neural baselines
Beam size	Inference decoding	Beam search = 4	Same beam size used for seq2seq correction baselines
Error classes	Classification labels	7 grammar error classes	Same seven-class taxonomy used for MTL-GEC comparison
Proficiency levels	Feedback adaptation	Beginner, intermediate, advanced	Same proficiency categories used for feedback baselines
Loss function	Multi-task objective	Correction loss + classification loss + feedback loss	Baselines optimized using their original task-specific objectives
Loss weighting	Multi-task balance	Correction-dominant weighted objective	Not applicable to single-task baselines; MTL baseline used comparable multi-task weighting
Train/dev/test split	Data partitioning	Standard benchmark splits	Same splits used for all models
Data leakage control	Split integrity	No sentence overlap across train, dev, and test sets	Same leakage filtering applied before all model training
Duplicate removal	Repetition control	Duplicate source-target pairs removed	Same cleaned dataset used for all models
Preprocessing	Text normalization	Cleaning, pair alignment, prompt formatting, tokenization	Same preprocessing pipeline used before model-specific tokenization
Early stopping	Overfitting prevention	Validation F0.5 monitored	Same validation-based stopping rule used for neural baselines
Random seed	Reproducibility	Fixed seed across runs	Same seed used for all repeat experiments
Evaluation metrics	GEC evaluation	F0.5, M2/ERRANT, exact match	Same correction metrics used for all GEC baselines
Hardware	Computational setting	8 A100 GPUs	Baselines trained/evaluated under comparable hardware settings

VII. CONCLUSION

This study proposed an adaptive grammar correction framework (AGC-LCFNet) that has been developed to go beyond correction-oriented systems with the incorporation of grammatical error correction, error-type classification, and generation of learner-oriented error feedback within a single framework. The proposed model encodes the contextual information and grammar patterns into a shared FLAN-T5 model, and the correction decoder uses the model to generate corrected sentences; the classification head classifies the seven major categories of errors; the feedback decoder uses a gating mechanism to generate explanations that are dependent on the level of proficiency. The experimental results indicate that AGC-LCFNet outperformed the other networks on F0.5 scores in all the benchmark and learner-generated corpora (BEA-2019, CoNLL-14, JFLEG and Lang-8), which demonstrates its effectiveness in both sets of corpora. The model also performed better in terms of error-type diagnosis and better classification performance compared to a multi-task baseline. Moreover, visual analysis showed that AGC-LCFNet focuses on error-specific linguistic features, produces language-altered corrected tokens having a high level of decoder confidence, and assigns high probabilities to the right error class. The feedback results showed that the framework is pedagogically valuable since the explanations for the beginner, intermediate and advanced learners have varying levels of complexity. The ablation results revealed that the classification module enhances correction accuracy, and the classification module with dynamic feedback and the gating mechanism in the module enhances the quality of learners' explanations. On a general level, the AGC-LCFNet grammar correction system is more meaningful and helps the teacher more effectively guide the learning process. Expansion of the model to other types of longer essays, more detailed taxonomies of errors, and L1-aware personalized feedback are possible future extensions of the model.

DECLARATION

Ethical Approval

Not applicable. This study does not involve human participants or animal subjects that would require ethical approval.

Consent to Participate

Not applicable. As this study does not involve human participants, informed consent to participate was not required.

Consent to Publish

Not applicable. This manuscript does not contain any person's data, images, or videos.

DATA AVAILABILITY

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

AUTHORS' CONTRIBUTIONS

Xinxin Sun and Peng Guo equally contributed to research design, data analysis, and manuscript writing. Both authors reviewed and approved the final version of the manuscript.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

FUNDING STATEMENTS

This research was supported by the 2024 Higher Education Teaching Reform Research Project of Jilin Province, titled "Research on the Construction of Ideological and Political Education in College English Courses and Teaching Practice under the Blended Teaching Mode" (Grant No. 2024L5L5649003K); and the 2023 General Project of the "14th Five-Year Plan" for Education Science in Jilin Province, titled "Study on Sports Teaching Strategies for Rural Children in Jilin Province under the Vision of Rural Revitalization" (Grant No. GH23744).

REFERENCES

[1] M. Qiao, "Using deep learning to improve English writing skills," Journal of Computational Methods in Sciences and Engineering, vol. 25, no. 1, pp. 977–986, Jan. 2025, doi: 10.1177/14727978251321403.

[2] Y. Wang and Y. Guo, " Investigating the Interplay of Chinese English as Foreign Language Learners ' Growth Mindset, Enjoyment, Adaptability, and Engagement in AI - Mediated Informal Digital Learning of English ," J. Comput. Assist. Learn., vol. 42, no. 1, p. e70189, Feb. 2026, doi: 10.1002/jcal.70189.

[3] Y. Liu and R. Li, "Deep Learning Scoring Model in the Evaluation of Oral English Teaching," Comput. Intell. Neurosci., vol. 2022, no. 1, p. 6931796, 2022, doi: https://doi.org/10.1155/2022/6931796.

[4] S.-Y. Chang, C. T. Chen, L.-H. Wang, and W. Chen, "Text-dependent English pronunciation learning system based on GMM," in Third International Conference on Electrical, Electronics, and Information Engineering (EEIE 2024), H. Malik, Ed., SPIE, Jan. 2025, p. 66. doi: 10.1117/12.3057529.

[5] J. Long, "A Grammatical Error Correction Model for English Essay Words in Colleges Using Natural Language Processing," Mobile Information Systems, vol. 2022, no. 1, p. 1881369, 2022, doi: https://doi.org/10.1155/2022/1881369.

[6] N. M. Tien and C. Labbé, "Detecting automatically generated sentences with grammatical structure similarity," Scientometrics, vol. 116, no. 2, pp. 1247–1271, 2018, doi: 10.1007/s11192-018-2789-4.

[7] J. M. Gayed, M. K. J. Carlon, A. M. Oriola, and J. S. Cross, "Exploring an AI-based writing Assistant's impact on English language learners," Computers and Education: Artificial Intelligence, vol. 3, p. 100055, Jan. 2022, doi: 10.1016/J.CAEAI.2022.100055.

[8] W. R. A. Bin-Hady, A. Al-Kadi, A. Hazaea, and J. K. M. Ali, "Exploring the dimensions of ChatGPT in English language learning: a global perspective," Library Hi Tech, vol. 43, no. 4–5, pp. 1315–1328, Oct. 2025, doi: 10.1108/LHT-05-2023-0200.

[9] H. A. Tabssam, S. A. Chattha, M. F. Javeed, and A. Hayat, "Sentiment analysis of user reviews: exploring Duolingo and Babbel in English language learning," Front. Comput. Sci., vol. 7, p. 1569058, Jun. 2025, doi: 10.3389/FCOMP.2025.1569058/BIBTEX.

[10] MukherjeeAnanya and ShrivastavaManish, "Lost in Translation? Found in Evaluation: A Comprehensive Survey on Sentence-Level Translation Evaluation," ACM Comput. Surv., vol. 58, no. 1, pp. 1–47, Sep. 2025, doi: 10.1145/3735970.

[11] D. Zhang and W. Wei, "The construction of English grammar error correction models based on Seq2Seq model and feedback filtering mechanisms," Discover Artificial Intelligence 2026 6:1, vol. 6, no. 1, pp. 382-, Mar. 2026, doi: 10.1007/S44163-026-01102-9.

[12] L. Stankevičius and M. Lukoševičius, "Extracting Sentence Embeddings from Pretrained Transformer Models," Applied Sciences 2024, Vol. 14, Page 8887, vol. 14, no. 19, p. 8887, Oct. 2024, doi: 10.3390/app14198887.

[13] S. David Hernandez and H. Calvo, "CoNLL 2014 Shared Task: Grammatical Error Correction with a Syntactic N-gram Language Model from a Big Corpora," pp. 26–27, 2014.

[14] C. Bryant, M. Felice, Ø. E. Andersen, and T. Briscoe, "The BEA-2019 Shared Task on Grammatical Error Correction," ACL 2019 - Innovative Use of NLP for Building Educational Applications, BEA 2019 - Proceedings of the 14th Workshop, pp. 52–75, 2019, doi: 10.18653/V1/W19-4406.

[15] L. Farek and A. Benaidja, "A dynamic hybrid attention-based autoencoder model with adaptive contextual attention for grammatical error correction," International Journal of Machine Learning and Cybernetics 2026 17:1, vol. 17, no. 1, pp. 27-, Jan. 2026, doi: 10.1007/S13042-025-02870-Z.

[16] J. Tian, "Symmetry-Aware Dual-Encoder Architecture for Context-Aware Grammatical Error Correction in Chinese Learner English: Toward a Spaced-Repetition Instructional Structure Sensitive to Individual Differences," Symmetry 2026, Vol. 18, Page 579, vol. 18, no. 4, p. 579, Mar. 2026, doi: 10.3390/SYM18040579.

[17] H. He, "TinyGEC: Interpretable Grammatical Error Correction for Edge-Based Intelligent English Tutoring Systems," Internet Technology Letters, vol. 9, no. 3, p. e70287, May 2026, doi: 10.1002/ITL2.70287.

[18] H. U. Khan and A. Daud, "Using Machine Learning Techniques for Subjectivity Analysis based on Lexical and Non-lexical Features", Accessed: Jul. 07, 2025. [Online]. Available: http://wndomains.fbk.eu/wnaffect.html

[19] D. Li, "A Study on Automated Recognition of Grammatical Errors in English Language Acquisition Based on BERT Modeling," International Journal of Computer Information Systems and Industrial Management Applications, vol. 18, pp. 7–7, Feb. 2026, doi: 10.70917/IJCISIM-2026-0049.

[20] M. Ahmed, H. U. Khan, S. Iqbal, and Q. Althebyan, "Automated Question Answering based on Improved TF-IDF and Cosine Similarity," in 2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS), 2022, pp. 1–6. doi: 10.1109/SNAMS58071.2022.10062839.

[21] C. Napoles, K. Sakaguchi, and J. Tetreault, "JFLEG: A Fluency Corpus and Benchmark for Grammatical Error Correction," 2017. Accessed: Jun. 01, 2026. [Online]. Available: https://aclanthology.org/E17-2037/

[22] H. U. Khan, A. Naz, F. K. Alarfaj, and N. Almusallam, "A transformer-based architecture for collaborative filtering modeling in personalized recommender systems," Scientific Reports 2025 15:1, vol. 15, no. 1, pp. 1–22, Jul. 2025, doi: 10.1038/s41598-025-08931-1.

[23] M. Singh, N. Gupta, S. Tyagi, A. Rani, V. Kumar, and S. Sharma, "Towards neural-symbolic grammatical inference for endangered languages using integrating graph neural networks and instruction-tuned language models," Eng. Appl. Artif. Intell., vol. 168, p. 114011, Mar. 2026, doi: 10.1016/J.ENGAPPAI.2026.114011.

[24] B. shobharani et al., "A Contextualized Learner-Profiling Transformer Architecture for Adaptive Grammar Error Diagnosis and Instruction," International Journal of Advanced Computer Science and Applications, vol. 17, no. 1, pp. 557–569, Jan. 2026, doi: 10.14569/IJACSA.2026.0170153.

[25] B. Shobharani et al., "Personalized Grammar Refinement Using Meta-Reinforcement Learning and Transformer-Based Framework," International Journal of Advanced Computer Science and Applications, vol. 16, no. 11, pp. 592–603, Nov. 2025, doi: 10.14569/IJACSA.2025.0161161.

[26] Y. Shannak and R. M. Duwairi, "Transfer learning for Arabic spelling correction: a comprehensive study with dynamic error synthesis," International Journal of Intelligent Computing and Cybernetics, vol. 19, no. 2, pp. 1–25, May 2025, doi: 10.1108/IJICC-08-2025-0558/1335420

[27] A. Bahari, "Evaluating the Effectiveness of AI-Driven Approaches on EFL Learners' Expository Writing Skills," TESOL Journal, vol. 17, no. 1, p. e70099, Mar. 2026, doi: 10.1002/TESJ.70099.

[28] T. El Ghadraoui, H. Asri, Z. Jarir, and A. Rochdi, "Enhancing E-learning Evaluation Through Generative AI: Personalizing Assessment and Feedback During Learning Experience," Lecture Notes on Data Engineering and Communications Technologies, vol. 262, pp. 230–240, 2025, doi: 10.1007/978-3-031-98476-1_19/SAVE-RESEARCH.

[29] Z. Wang, Q. Yu, J. Wang, Z. Hu, and A. Wang, "Grammar Correction for Multiple Errors in Chinese Based on Prompt Templates," Applied Sciences 2023, Vol. 13, Page 8858, vol. 13, no. 15, p. 8858, Jul. 2023, doi: 10.3390/APP13158858.

[30] L. Hu, Y. Tang, X. Wu, and J. Zeng, "Considering optimization of English grammar error correction based on neural network," Neural Computing and Applications 2021 34:5, vol. 34, no. 5, pp. 3323–3335, Jan. 2021, doi: 10.1007/S00521-020-05591-2.

[31] S. Zhou and W. Liu, "English Grammar Error Correction Algorithm Based on Classification Model," Complexity, vol. 2021, no. 1, p. 6687337, Jan. 2021, doi: 10.1155/2021/6687337.

[32] A. Jalali, "Grammatical Error Correction Using Large Language Models: A Case Study on Universal Dependencies Treebanks," Jun. 23, 2025. Accessed: Jun. 01, 2026. [Online]. Available: https://www.utupub.fi/handle/11111/20102

[33] V. Mageshwari and J. Shrijith, "Linguistic-Neural Hybrid Architectures for Semantic Search: Optimizing Recall with POS Filtering and Query Expansion," IEEE Access, 2026, doi: 10.1109/ACCESS.2026.3663716.

[34] Y. S. Cho, "Lang-8," CALICO Journal, vol. 30, no. 2, pp. 293–299, Jan. 2013, doi: 10.11139/CJ.30.2.293-299.

[35] E. P. Flynn, "Strategy 2: Replacing the Current State Modern Foreign Languages Examinations with Common European Framework of Reference for Languages Examinations, Implementing an Academic Tracking System, and Incorporating the European Language Portfolio," Cross-age Teaching and Plurilingualism, pp. 151–202, 2026, doi: 10.1007/978-3-031-96656-9_4.

MACN: A Modality-Aware Cross-Attention Transformer for Non-Invasive Hemoglobin Estimation and Anemia Classification Using Multi-Region Physiological Images

Chaitra S P[1], D R Ramesh Babu[2]
Assistant Professor, Department of Computer Science and Engineering-B.M.S. College of Engineering, Bangalore, Visvesveraya Technological University, Belagavi-590018, India[1]
Professor & Head, Department of Computer Science and Engineering-Dayananda Sagar College of Engineering, Bangalore, Visvesveraya Technological University, Belagavi-590018, India[2]

***Abstract*—Anemia is a major health concern among pregnant women in rural areas. Early detection and timely treatment are necessary to reduce maternal and infant mortality. Many non-invasive techniques have been proposed to solve the problem of unavailability of laboratory-based hemoglobin testing in rural areas. But their accuracy is limited due to dependence on a single modality. This work proposes a multimodal cross-attention network (MCAN) for anemia classification and hemoglobin (Hgb) level estimation using multiple physiological visual cues. Deep convolutional features extracted from four regions of the eye conjunctiva, fingertip, palm, and lip mucosa are fused using a cross-attention multimodal fusion. The fused features are used to estimate Hgb level and predict anemia. Through experimental analysis, the proposed solution is found to increase the estimation accuracy by 5% compared to existing single approaches. The statistical significance of the performance gain is validated through bootstrap confidence intervals and DeLong's test.**

***Keywords*—Anemia; hemoglobin; deep learning; cross-attention fusion; multimodality**

I. Introduction

Anemia is a serious health issue for pregnant women. This nutritional deficiency characterized by lower levels of hemoglobin in the blood. It can increase the chances of maternal and infant mortality when not detected and treated earlier [1]. Thus, early diagnosis becomes very important. A hemoglobin level below 11g/dL is defined as anemia for pregnant women [2]. A clinical method of estimating hemoglobin level from blood samples is a painful, uncomfortable process, as sometimes repeated measurements are needed. The process is time-consuming and infeasible in resource-constrained areas like rural areas. It can also lead to infection spread when facilities are not maintained hygienically [3-4]. Sensing this need for an invasive, comfortable, and safe method for Hgb estimation, various visual cue-based approaches have been proposed. These approaches analyze clinically relevant anatomical sites like eye conjunctiva, fingertip, palm, and lip mucosa, which demonstrate color variations correlated with hemoglobin concentration [5-6]. Various conventional and deep learning techniques have been proposed to estimate Hgb levels from these images. While conventional techniques extracted handcrafted features and trained classification algorithms to estimate Hgb levels, deep learning techniques avoided the need for handcrafted features and applied repeated convolutions and pooling to extract features. Deep learning techniques provided higher accuracy in Hgb estimation in comparison to conventional techniques [7].

The existing deep learning techniques for Hgb estimation can be categorized into single and multi-modality approaches (discussed in the survey). Single-modality approaches use any one of the modalities of eye conjunctiva, fingertip, palm, and lip mucosa for Hgb estimation. Multi-modality approaches use multiple modalities of image or images with clinical metadata to estimate Hgb. Dependence on a single modality image makes the single modality approaches highly sensitive to illumination variations, skin tone differences, etc. Also, individual anatomical regions may not consistently reflect hemoglobin-related color changes. These factors lead to inaccurate hemoglobin estimation [8].

Multimodality approaches integrate complementary information from multiple cues to solve this problem. But current multimodality approaches don't effectively learn cross-modality feature representation from the modalities, leading to reduced accuracy and generalizability of hemoglobin prediction. This work proposes a multimodal cross-attention network to learn robust feature representation, facilitating improved accuracy and generalizability of hemoglobin estimation. The following are the novel contributions of this work:

- A multimodal cross-attention network integrating information from multiple cues of eye conjunctiva, fingertip, palm, and lip mucosa for hemoglobin estimation. Transformer features extracted from each individual modality are fused through a cross-attention fusion technique and the fused representation is used for hemoglobin estimation using a regression head.
- A modality-aware transformer strategy is introduced to extract discriminative color and texture patterns related to hemoglobin concentration using transformer-based encoders. In this way, subtle physiological variations

across different cues are captured while preserving modality-specific information before fusion.

The rest of the study is organized as follows: Section II surveys the existing methods for hemoglobin estimation using deep learning models. Section III details the proposed multimodal cross-attention network integrating multiple cues for hemoglobin estimation. Section IV presents the results of the proposed technique in comparison to recent multimodality and single-modality baselines. Section V presents the conclusion and future scope of work.

II. Related Work

Various single-modality approaches using different variants of deep learning models have been proposed. Jayakody et al. [9] extracted convolutional neural network (CNN) features from fingertip images and used them for estimating the Hgb level. There were no sensitivity and specificity measurements in this work. Magdalena et al. [10] classified eye conjunctiva images into anemic classes using CNN. The image was trained as a whole without consideration for significant regions in the eye conjunctiva image. Das et al [11] combined CNN with vision transformer (ViT) to predict Hgb level from eyelid images. The outliers are more than 8%, and accuracy is very low at the border of anemia Hgb ranges in this approach. Zhang et al. [12] extracted deep learning features from face images and used them for anemia prediction. Authors experimented with ResNet50, MobileNet, InceptionV3, EfficientNetB0, and DenseNet121, and achieved 80% accuracy. But using a face image as a whole without considering the illumination environment can increase the false positives.

Muljono et al. [13] used a hybrid deep learning model combining MobileNetV2 with a support vector machine classifier to predict anemia from eye conjunctive images. Rivero-Palacio et al. [14] classified eye conjunctive image to the anemic class using the YOLOv5 model. The method was tested only for children below 5 years old. Asare et al. [15] used the AlexNet deep learning model to classify anemia. The authors experimented with eye conjunctiva, palm, and fingernail images. Palm images were found to provide higher accuracy, but the performance was tested for children below 5 years old. Magdalena et al. [16] detected anemia from palpebral conjunctiva images using CNN. But the method was tested only for a smaller dataset. Mitani et al. [17] classified retinal fundus images for anemia using the Inception V4 deep learning model. The images were processed as a whole without paying attention to significant regions where color variations are dominant. Chen et al. [18] estimated Hgb level from conjunctiva image using the MobileNet V3 network. The mean absolute error is more than 1.521, and this increases false positives.

Kasiviswanathan et al. [19] proposed a multimodality technique to predict Hgb level. Eye conjunctival image along with age, height, and weight of the subject is used to predict Hgb level. The authors extracted color features from the eye conjunctival image and fused them with meta attributes. A linear regression model is fit between fused features and Hgb level. Image features were not extracted from salient regions. Ramzan et al. [20] combined palm image with corpuscular concentration levels to classify anemia. A deep learning attention model was used to get the fused features. The fused features are classified by softmax classifier to anemia. But the approach needs corpuscular concentration levels which require blood test, making it invasive. Khan et al. [21] extracted fused features from retinal fundus images and meta attributes (age, gender) though a Inception V4 deep learning model and classified the fused features to anemia. But the approach was tested only for people above 50 years old. There were very few works on use of multimodal deep learning.

III. Methodology

Individual modalities of eye conjunctiva, fingernail, palm, and lip mucosa contain useful physiological information related to hemoglobin concentration. Since each modality may be affected by factors such as illumination variations, skin tone differences, or imaging noise, learning cross-modal relationships between these anatomical cues is essential for robust hemoglobin estimation. Though the modalities differ visually, by projecting each modality's features into a shared embedding space cross modal relationship can be learnt. Based on this observation, the architecture of the proposed multimodal cross-attention network (MACN) for Hgb estimation and anemia detection is designed. The architecture is shown in Fig. 1.

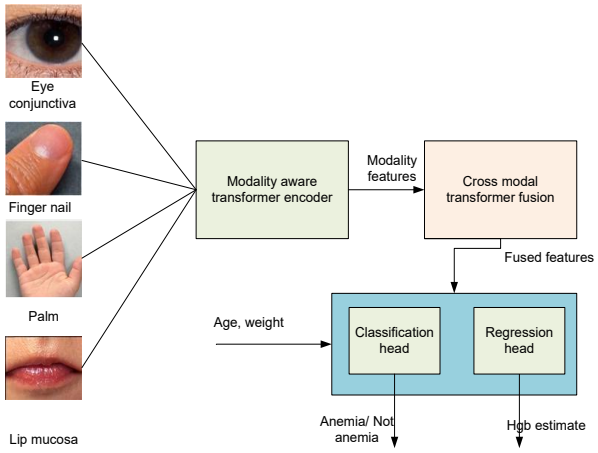


Fig. 1. Architecture of the proposed multimodal cross-attention network.

There are three important functional components in the proposed solution: (1) modality-aware transformer encoder, (2) cross-modal transformer fusion, and (3) classification/regression head. The modality-aware transformer encoder converts the image into a sequence of feature tokens. Cross-modal transformer fusion allows each modality to attend to information from the other modality, resulting in an enhanced, robust feature representation. This feature representation is fused with metadata information (age, weight), and the final fused feature is passed through a classification/regression network. The classification network gives the probability for two classes of anemia/not anemia. The regression head gives the Hgb estimate. Each of the functional components are detailed in the below subsections.

A. Modality-Aware Transformer Encoder

The modality-aware transformer encoder extracts discriminative features from each modality, preserving modality-specific information before fusion. Let the image of modality m be represented as:

$$I_m \in \mathbb{R}^{H\times W\times 3}$$

where, $m \in \{c, f, p, l\}$ corresponds to conjunctiva, fingertip, palm, and lip mucosa, and H, W represent the height and width of the image.

Each input image is split into a sequence of non-overlapping patches of size $P \times P$ with a total number of patches as N. Each patch is flattened and projected into a d-dimensional embedding space:

$$z_0 = \{x_1E, x_2E, \dots x_NE\} \quad (1)$$

where, x_i is the image patch, and E is a learnable embedding matrix.

Positional encoding is added to preserve spatial information:

$$z_0 = z_0 + E_{pos} \quad (2)$$

where, E_{pos} is the positional embedding.

The patch embeddings are processed using a stack of L transformer encoder layers. Each layer has the following components: (1) multi-head self-attention (MHSA), (2) feed-forward network (FFN), (3) normalization layer (NL), and (4) residual connections.

The computation at the l^{th} layer is represented as:

$$z_l' = MHSA\big(NL(z_{l-1})\big) + z_{l-1} \quad (3)$$

$$z_l = FFN\big(LN(z_l')\big) + z_l' \quad (4)$$

The long-range relationships between different patches of the image are learnt using a self-attention operation. For each attention head h, the query (Q), key (K), and values (V) for the attention operation are computed as:

$$Q = zW_Q^h \quad (5)$$

$$K = zW_K^h \quad (6)$$

$$V = zW_V^h \quad (7)$$

The attention weights are computed as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_K}})V \quad (8)$$

The outputs of each attention head are concatenated and projected as:

$$MHSA(z) = concat(head_1, head_2, \dots . head_H)W_O \quad (9)$$

The subtle physiological cues related to hemoglobin concentration are captured by the modality-specific visual representation. For each modality m, its final feature embedding is represented as:

$$F_m = T_m(I_m) \quad (10)$$

where,

$$F_m \in \mathbb{R}^{N\times d} \quad (11)$$

The modality-specific features are passed to the next stage of cross-modal transformer fusion, where inter modal feature relationships are learnt for effective Hgb estimation. The configuration of transformer encoder is given in Table I.

TABLE I. CONFIGURATION OF THE TRANSFORMER ENCODER

Parameter	Value
Input size	224 × 224
Patch size	16 × 16
Embedding dimension	768
Transformer layers	12
Attention heads	12
FFN dimension	3072
Dropout	0.1
Activation	GELU

B. Cross-Modal Transformer Fusion

Cross-modal transformer fusion enables iterative information exchange between modalities through multi-head cross-attention operations, which allows effective feature learning compared to simple feature concatenation. Cross-model transformer allows each modality to attend to information from other modalities. The cross-model interaction between two modalities i and j is represented as:

$$CM(F_i, F_j) = softmax\left(\frac{(F_iW_Q)(F_jW_K)^T}{\sqrt{d}}\right)(F_jW_V) \quad (12)$$

where, W_Q, W_K, W_V are the projection matrices, and d is the embedding dimension. To capture multiple interactions, cross-modal multi-head attention (CMMHA) is used:

$$head_h = Attention(F_iW_Q^h, F_jW_K^h, F_jW_V^h \quad (13)$$

$$CMMHA(F_i, F_j) = Concat(head_1, head_2, .., head_H)W_O \quad (14)$$

where, H is the number of heads and W_O is the output projection matrix.

To capture complementary information across the modalities, bidirectional interactions between the modalities are captured.

$$F_c' = F_c + CMMHA(F_c, F_f) + CMMHA(F_c, F_p) + CMMHA(F_c, F_l) \quad (15)$$

$$F_f' = F_f + CMMHA(F_f, F_c) + CMMHA(F_f, F_p) + CMMHA(F_f, F_l) \quad (16)$$

$$F_p' = F_p + CMMHA(F_p, F_c) + CMMHA(F_p, F_f) + CMMHA(F_p, F_l) \quad (17)$$

$$F_l' = F_l + CMMHA(F_l, F_c) + CMMHA(l, F_f) + CMMHA(F_l, F_p) \quad (18)$$

The modality features are fused to form a cross-fused

representation as in:

$$F_{CF} = Concat(F_c', F_f', F_p', F_l') \quad (19)$$

The cross-fused features capture both modality-specific physiological patterns and cross-modal dependencies. The cross-fused features are passed to the next stage of regression. The configuration of the cross-modal transformer fusion module is given in Table II.

TABLE II. CONFIGURATION OF CROSS-MODAL TRANSFORMER

Parameter	Value
Input modalities	4
Embedding dimension	768
Fusion layer	4
Attention heads	8
FFN dimension	2048
Dropout	0.1
Activation	GELU

C. Classification/Regression Head

Though the cross-fused features extracted from anatomical regions such as eye conjunctiva, fingertip, palm, and lip mucosa provide important information for hemoglobin estimation, clinical attributes, especially age and weight, have a strong influence on hemoglobin levels in pregnant women. Studies have shown that maternal age and anthropometric characteristics such as body weight or body mass index are significantly associated with anemia prevalence and hemoglobin concentration in pregnant women [22-24]. The cross-fused features extracted from visual cues are merged with metadata of age and weight to create final fused features:

$$F_{final} = Concat(F_{CF}, F_{meta}) \quad (20)$$

The final fused representation is passed through a regression network to get Hgb estimation:

$$\overline{Hgb} = R(F_{final}) \quad (21)$$

where, R consists of fully connected layers.

The regression network is trained using the mean square error loss function (L) as:

$$L = \frac{1}{M} \sum_{i=1}^{M} (Hgb_i - \overline{Hgb}_i)^2 \quad (22)$$

where, M is the number of samples, Hgb_i is the actual Hgb value, and $\overline{Hgb}_i$ is the predicted Hgb value.

TABLE III. CONFIGURATION OF THE REGRESSION HEAD

Layer	Operation	Output size
Input	Fused feature vector	F_{final}
FC1	Fully connected + ReLU	256
Dropout	Dropout rate (rate=0.3)	256
FC2	Fully connected + ReLU	128
Dropout	Dropout rate (rate=0.3)	128
FC3	Fully Connected + ReLU	64
Output Layer	Fully Connected (Linear)	1

The configuration of the regression head is given in Table III. The configuration of the classification head is same as that given in Table III, except the last output layer is replaced with the softmax function to provide probabilities for two classes of anemia and not anemia.

IV. RESULTS

The dataset for experimentation was collected from Hadinaru primary health centre (PHC), of rural Mysore district, Karnataka. The data of four images of eye conjunctiva, finger nail, palm, lip mucosa and meta-attributes of age, weight, and Hgb level are collected from 120 participants. Though the dataset consists of 120 participants, each sample contains four physiological image modalities (eye conjunctiva, fingertip, palm, and lip mucosa) together with laboratory-confirmed hemoglobin measurements and clinical metadata. Collecting such multimodal medical data is considerably more challenging than acquiring conventional image datasets because each participant requires synchronized acquisition of multiple anatomical images and corresponding clinical measurements. To ensure reliable evaluation despite the limited dataset size, the proposed model was assessed using 5-fold cross-validation, with results reported as mean ± standard deviation across folds. In addition, bootstrap confidence intervals and DeLong's statistical significance tests were performed to quantify result stability and reliability. The dataset characteristics are given in Table IV.

TABLE IV. DATASET CHARACTERISTICS

Parameter	Value
Dataset acquiring environment	Ambient lighting conditions
Dataset collection duration	6 months
Age range	22-31 years
Weight range	48-85 kg
Hgb range	7 g/dL to 15.5 g/dL
Anemia threshold	11g/dL
Percentage of anemia cases	64 %
Percentage of non-anemia cases	36 %

The dataset is split into training (80%) and test (20%) ensuring balanced distribution of anemic and non-anemic cases in both splits. Testing is conducted with 5-fold cross validation strategy to ensure robustness and generalization. For each fold, a different random split (80:20) is made and the model is trained with training data. The trained model is tested with the test data. The results are reported in terms of mean± standard deviation (SD).

The performance is analyzed in comparison with existing methods specific to each modality, namely fingernail [9], eye conjunctiva [13], palm [15], and multimodal approach [19]. The performance of Hgb estimation is compared in terms of Mean absolute error (MAE), root mean square error (RMSE) and R^2. Anemia classification performance (threshold as 11g/dL) is accuracy and AUC. The performance comparison results with existing approaches are presented in Table V.

The proposed MCAN achieved an accuracy of 95.4% which is atleast 5% higher compared to existing works. The

error in Hgb estimation is at least 41% lower in the proposed MCAN, demonstrating improved precision in Hgb estimation. The lower SD in the proposed MCAN indicates consistent performance across all folds. The single modality approaches ([9],[13],[15]) had higher SD, reflecting sensitivity to imaging conditions and anatomical variability. Though the multimodality approach [19] had lower SD, it lacks robustness in comparison to MCAN. The higher accuracy and lower SD demonstrate the strong generalization capability of the proposed MCAN. The use of modality-aware transformers along with cross-attention fusion has resulted in robust feature representation in the proposed solution. This robust feature representation, along with clinical metadata integration, has enhanced prediction stability and improved estimation accuracy.

TABLE V. PERFORMANCE COMPARISON WITH EXISTING WORKS

Method	Modality used	MAE(g/dL)	RMSE(g/dL)	Accuracy (%)	AUC
Fingernail-based CNN [9]	Fingernail	1.38 ± 0.09	1.79 ± 0.12	83.2 ± 2.3	0.87 ± 0.02
Eye conjunctive CNN [13]	Eye conjunctiva	1.34 ± 0.08	1.71 ± 0.11	84.6 ± 2.0	0.88 ± 0.02
Palm based Deep model [15]	Palm	1.29 ± 0.07	1.65 ± 0.10	86.3 ± 1.8	0.89 ± 0.02
Multi model regression [19]	Eye conjunctiva image + meta data	1.10 ± 0.06	1.42 ± 0.09	90.1 ± 1.6	0.92 ± 0.01
Proposed MCAN	Multi modality image + data	0.78 ± 0.05	0.98 ± 0.07	95.4 ± 1.2	0.97 ± 0.01

The ROC curve of the proposed MCAN is given in Fig. 2.

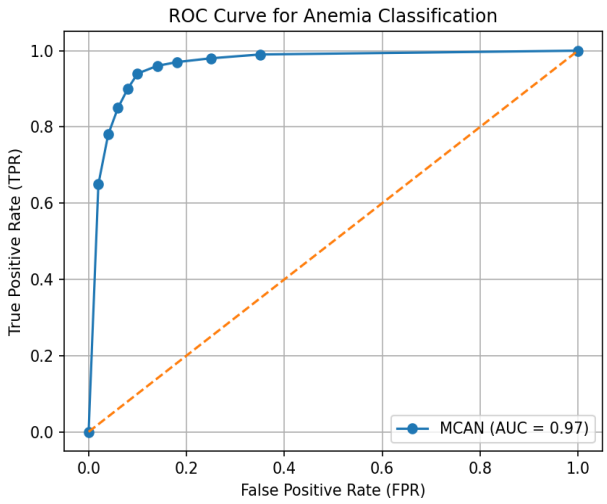


Fig. 2. ROC curve of the proposed MCAN.

The ROC curve shows the excellent discriminative capability between anemia and non-anemia classes in the proposed MCAN. The AUC of 0.97 indicates strong separability. In addition, the position of the curve towards the top left indicates higher sensitivity and specificity across thresholds. The confusion matrix in the proposed MCAN is given in Fig. 3.

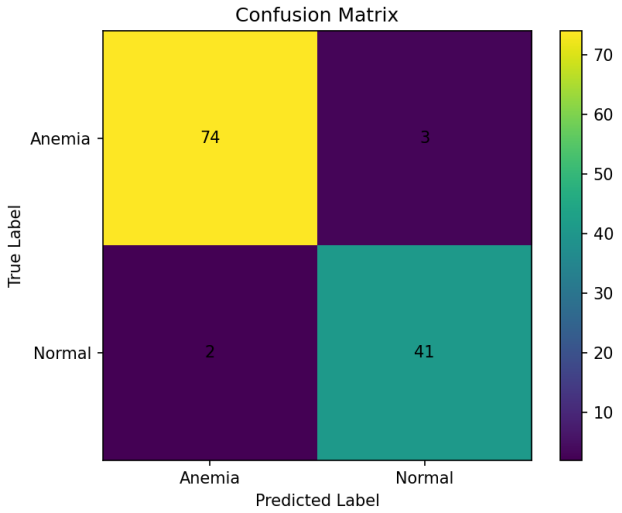


Fig. 3. Confusion matrix

From the confusion matrix, it can be seen that the proposed MCAN has higher sensitivity and specificity, indicating reliable discrimination between anemia and non-anemia classes.

The statistical significance of the performance gain achieved in the proposed MCAN is validated through bootstrap confidence intervals and DeLong's test. The proposed MCAN achieved an accuracy of 95.4% (95% CI: 93.1%–97.2%) and AUC of 0.97 (95% CI: 0.95–0.99). DeLong's test comparing the proposed model with the multimodal solution [19] yielded a p-value of 0.0032, confirming that the improvement is statistically significant.

Cross-attention feature fusion across multiple modalities is an important contribution in this work. Performance comparison is done against different fusion techniques, and the results are presented in Table VI.

TABLE VI. COMPARISON WITH FUSION STRATEGIES

Fusion strategy	MAE (g/dL)	RMSE (g/dL)	Accuracy (%)	AUC
Early fusion	1.18±0.09	1.51± 0.11	89.6 ± 1.9	0.91±0.02
Late Fusion	1.12± 0.08	1.45 ± 0.10	90.7 ± 1.8	0.92 ± 0.02
Feature Concatenation	1.02 ± 0.08	1.31 ± 0.10	91.8 ± 1.7	0.93 ± 0.02
Self-Attention Fusion	0.89 ± 0.07	1.14 ± 0.09	93.4 ± 1.4	0.95 ± 0.01
Proposed MCAN (Cross-Attention Fusion)	0.78 ± 0.05	0.98 ± 0.07	95.4 ± 1.2	0.97 ± 0.01

From the results, it can be seen that the proposed cross-attention fusion achieved best performance in comparison to other fusion techniques. Though self-attention fusion achieved higher performance in comparison to feature concatenation and

early and late fusion, it achieved lower performance in comparison to cross-attention fusion, as it treated all modalities uniformly and did not explicitly model directional interactions between modalities. Cross-attention fusion solved this problem by learning complementary information from other modalities, which produced a more discriminative and robust feature representation.

The performance of the proposed solution is compared against conventional convolutional neural network architectures for feature extraction, and the results are given in Table VII.

TABLE VII. COMPARISON WITH FEATURE EXTRACTION BACKBONES

Backbone	MAE (g/dL)	RMSE (g/dL)	Accuracy (%)	AUC
ResNet50	1.14 ± 0.08	1.47 ± 0.11	89.8 ± 1.9	0.91±0.02
DenseNet121	1.08 ± 0.08	1.39 ± 0.10	90.9 ± 1.8	0.92 ± 0.02
EfficientNet-B0	1.01 ± 0.07	1.30 ± 0.09	92.1 ± 1.6	0.94 ± 0.02
Vision Transformer (ViT-Base)	0.92 ± 0.07	1.18 ± 0.08	93.3 ± 1.5	0.95 ± 0.01
Swin Transformer	0.86 ± 0.06	1.09 ± 0.08	94.1 ± 1.3	0.96 ± 0.01
Proposed MCAN (Modality-Aware Transformer)	0.78 ± 0.05	0.98 ± 0.07	95.4 ± 1.2	0.97 ± 0.01

From the results, it can be seen that traditional CNN architectures – ResNet50, DenseNet121 and EfficientNet-B0 achieved only lower performance compared to transformer-based models. Due to their ability to capture long-range dependencies and subtle color-texture relationships associated with hemoglobin concentration, transformer-based models provided better performance. Compared to standard vision transformers, Swin Transformer provided better results because of its hierarchical feature representation and shifted-window attention mechanism. The proposed MCAN improved the performance even more due to its ability to learn modality-specific physiological characteristics before multimodal fusion. This preserved discriminative information from conjunctiva, fingertip, palm, and lip mucosa images.

Robustness analysis of the proposed MCAN is conducted against different skin tones, and the results are given in Table VIII.

TABLE VIII. ROBUSTNESS ANALYSIS

Skin tone group	MAE (g/dL)	RMSE (g/dL)	Accuracy (%)	AUC
Light	0.76 ± 0.05	0.96 ± 0.07	95.8 ± 1.1	0.97 ± 0.01
Medium	0.79 ± 0.05	0.99 ± 0.07	95.2 ± 1.2	0.97 ± 0.01
Dark	0.83 ± 0.06	1.04 ± 0.08	94.6 ± 1.3	0.96 ± 0.01

Performance evaluation against different skin tones shows that the maximum difference in classification accuracy between skin-tone groups was 1.2%, demonstrating that the proposed multimodal fusion strategy reduces dependence on a single visual cue and improves generalizability across diverse populations.

An ablation study was conducted to analyze the contribution of each functional component of the proposed MCAN. The ablation study results for various variants are given in Table IX.

TABLE IX. ABLATION STUDY RESULTS

Variants	MAE(g/dL)	RMSE (g/dL)	Accuracy (%)	AUC
Without cross-attention (feature concatenation)	1.02 ± 0.08	1.31 ± 0.10	91.8 ± 1.7	0.93 ± 0.02
Without modality-aware transformer (CNN features)	0.95 ± 0.07	1.22 ± 0.09	92.6 ± 1.5	0.94 ± 0.02
Without clinical metadata (age, weight)	0.88 ± 0.06	1.10 ± 0.08	93.9± 1.3	0.95 ± 0.01
Full model (MCAN)	0.78 ± 0.05	0.98 ± 0.07	95.4 ± 1.2	0.97 ± 0.01

Simple feature concatenation without cross-attention fusion reduced the accuracy by 3.6%, demonstrating the robust feature representation effectiveness of cross-attention fusion. Using CNN features instead of modality-aware transformers reduced the accuracy by 3%, indicating the importance of capturing long-range dependencies and subtle color-texture variations with modality-aware transformers. With metadata fusion, the performance degraded only by 1.5%, but still it confirms that physiological attributes complement visual cues. The results of the lowest error, highest accuracy, and lowest SD in the proposed MCAN indicate that each component contributes synergistically to improved hemoglobin estimation and anemia detection.

The novelty of the proposed solution over existing multimodal approaches is given in Table X.

TABLE X. THE NOVELTY OF THE PROPOSED SOLUTION OVER EXISTING MULTIMODAL APPROACHES.

Feature	[19] Kasiviswanathan	[20] Ramzan	Proposed MCAN
Multiple image modalities	✗	✗	✓
Transformer encoder	✗	✗	✓
Modality-aware feature learning	✗	✗	✓
Cross-attention fusion	✗	✗	✓
Metadata integration	✓	✓	✓
Hemoglobin regression	✓	✗	✓
Anemia classification	✓	✓	✓
End-to-end multimodal learning	✗	Partial	✓

The proposed multimodal cross-attention network (MCAN) differs from existing approaches in four important aspects:

1. Differing from prior studies that utilize a single anatomical region or a single image modality, the proposed framework simultaneously exploits four physiologically relevant visual cues, namely eye conjunctiva, fingertip, palm, and lip mucosa. Each modality contains complementary information associated with hemoglobin concentration. Integrating these modalities enables the model to remain robust when one modality is affected by illumination artifacts, pigmentation differences, or image acquisition noise.
2. A modality-aware transformer encoder is introduced for feature extraction. Most of the existing works use convolutional neural networks or generic vision transformers that process all modalities in a similar manner. In contrast, the modality-aware transformer is designed to preserve modality-specific characteristics before fusion, allowing the network to capture subtle color and texture variations associated with hemoglobin concentration while maintaining the unique physiological information present in each anatomical region.
3. The proposed framework uses cross-attention fusion rather than conventional feature concatenation or independent attention mechanisms. Through cross-attention, each modality actively attends to informative regions of other modalities, enabling explicit learning of inter-modal relationships. This interaction produces a richer and more discriminative feature representation that captures both modality-specific and cross-modal dependencies.
4. The proposed work integrates visual information with clinical metadata such as age and weight within a unified prediction framework. While previous multimodal methods generally combine image-derived features with metadata using shallow regression models, the proposed approach learns joint representations through deep multimodal feature interaction. This enables the model to exploit complementary physiological and demographic information for improved hemoglobin estimation.

Thus, the novelty of the proposed MCAN does not arise solely from combining multiple modalities, but from proposing a unified architecture that combines modality-aware transformer encoding, cross-attention-based multimodal fusion, and metadata-guided prediction for non-invasive hemoglobin estimation and anemia detection. To the best of our knowledge, no previous study has jointly incorporated these components within a single end-to-end framework for anemia prediction using conjunctiva, fingertip, palm, and lip mucosa images.

V. Conclusion

This work proposed a novel multimodal cross-attention network (MCAN) for non-invasive Hgb estimation and anemia detection using multiple visual cue modalities. The solution integrated modality-aware transformers with cross-attention fusion to obtain robust feature representation capturing both modality-specific and inter-modal features. The proposed model had atleast 5% anemia detection accuracy and 41% lower Hgb estimation error compared to existing works. Statistical validation using bootstrap confidence intervals and DeLong's test confirmed that the performance gains are significant.

References

[1] Ezzati M, Lopez AD, Rodgers A, Hoorn SV, Murray CJ. Selected major risk factors and global and regional burden of disease Lancet. 2002;360:1347–60

[2] Conrad, M. E. (2011). Anemia. Boston, MA: Butterworths.

[3] N. Mannino et al., "Smartphone app for non-invasive detection of anemia using only patient-sourced photos," *Nature Communications*, vol. 9, 2018.

[4] J. R. White et al., "Noninvasive hemoglobin measurement using digital photographs of the conjunctiva," *Journal of Biomedical Optics*, vol. 21, no. 10, 2016.

[5] Y. M. Chen, S. G. Miaou, and H. Bian, "Examining palpebral conjunctiva for anemia assessment using image processing," *Computer Methods and Programs in Biomedicine*, vol. 137, pp. 125–135, 2016.

[6] T. B. Donmez et al., "Anemia detection through non-invasive analysis of lip mucosa images," *Frontiers in Big Data*, 2023.

[7] A. Al Mahmud et al., "AI-driven anemia diagnosis: A review of advanced deep learning techniques," *arXiv preprint*, 2025.

[8] S. Mythili et al., "Anaemia detection using deep learning by processing conjunctiva, fingernails, palm and tongue images," IRJAEH, 2024.

[9] J. A. D. C. A. Jayakody and E. A. G. A. Edirisinghe, "HemoSmart: A Non-invasive, Machine Learning Based Device and Mobile App for Anemia Detection," 2020 IEEE REGION 10 CONFERENCE (TENCON), Osaka, Japan, 2020, pp. 1401-1406

[10] Magdalena R, Saidah S, Ubaidah IDS, Fuadah YN, Herman N, Ibrahim N. Convolutional neural network for anemia detection based on conjunctiva palpebral images. JTekInform. 2022;3(2):349-354

[11] Das S, Ahamed F, Das A, et al. (July 26, 2024) NiADA (Non-invasive Anemia Detection App), a Smartphone-Based Application With Artificial Intelligence to Measure Blood Hemoglobin in Real-Time: A Clinical Validation. Cureus 16(7): e65442

[12] Zhang A, Lou J, Pan Z, Luo J, Zhang X, Zhang H, Li J, Wang L, Cui X, Ji B, Chen L. Prediction of anemia using facial images and deep learning technology in the emergency department. Front Public Health. 2022

[13] Muljono, S. A. Wulandari, H. A. Azies, M. Naufal, W. A. Prasetyanto and F. A. Zahra, "Breaking Boundaries in Diagnosis: Non-Invasive Anemia Detection Empowered by AI," in IEEE Access, vol. 12, pp. 9292-9307, 2024

[14] Rivero-Palacio M, Alfonso-Morales W, Caicedo-Bravo E. Mobile application for anemia detection through ocular conjunctiva images. In: 2021 IEEE Colombian conference on applications of computational intelligence (ColCACI); May 2021

[15] Asare, Justice & Appiahene, Peter & Donkoh, Emmanuel & Dimauro, Giovanni. (2023). Iron Deficiency Anemia Detection using Machine Learning Models: A Comparative Study of Fingernails, Palm and Conjunctiva of the Eye Images.. 10.22541/au.167570558.82410707/v1.

[16] Magdalena R, Saidah S, Ubaidah IDS, Fuadah YN, Herman N, Ibrahim N. Convolutional neural network for anemia detection based on conjunctiva palpebral images. JTekInform. 2022;3(2):349-354

[17] Mitani A, Huang A, Venugopalan S, et al. Detection of anaemia from retinal fundus images via deep learning. Nat Biomed Eng. 2020;4(1):18-27

[18] Chen YM, Miaou SG, Bian H. Examining palpebral conjunctiva for anemia assessment with image processing methods. Comput Methods Programs Biomed. 2016 Dec;137:125-135

[19] Kasiviswanathan S, Vijayan TB, John S. Ridge regression algorithm based non- invasive anaemia screening using conjunctiva images. J. Ambient Intell. Humaniz. Comput. Oct. 2020

[20] Ramzan M, Sheng J, Saeed MU, Wang B, Duraihem FZ. Revolutionizing anemia detection: integrative machine learning models and advanced attention mechanisms. Vis Comput Ind Biomed Art. 2024 Jul 17;7(1):18

[21] Khan R, Maseedupally V, Thakoor KA, Raman R, Roy M. Noninvasive Anemia Detection and Hemoglobin Estimation from Retinal Images

Using Deep Learning: A Scalable Solution for Resource-Limited Settings. Transl Vis Sci Technol. 2025 Jan 2;14(1):20

[22] L. Balarajan, U. Ramakrishnan, E. Ozaltin, A. Shankar, and S. Subramanian, "Anaemia in low-income and middle-income countries," *The Lancet*, vol. 378, no. 9809, pp. 2123–2135, 2011.

[23] S. S. Stevens et al., "Global, regional, and national trends in haemoglobin concentration and prevalence of total and severe anaemia in pregnant and non-pregnant women," *The Lancet Global Health*, vol. 1, no. 1, pp. e16–e25, 2013.

[24] N. B. Kozuki et al., "Maternal anthropometry and risk of anemia in pregnancy," *The American Journal of Clinical Nutrition*, vol. 98, no. 2, pp. 429–436, 2013

Low-Latency FPGA-Based PCA Acceleration for Hyperspectral Image Dimensionality Reduction

Hana Ben Fredj[1]*, Ahlem kehili[2], Amani Chabbeh[3], Jamel Baili[4], Chokri Souani[5]
Laboratory of Electronics and Micro-Electronics, Faculty of Science of Monastir, Tunisia[1, 3]
Laboratory of Analysis and Processing of Electrical and Energetic Systems,
Faculty of Sciences of Tunis, Tunis El-Manar University, Tunisia[2]
College of Computer Science, King Khalid University, Abha, 61413, Saudi Arabia[4]
Higher Institute of Applied Sciences and Technology of Sousse, University of Sousse,
Tunisia, Laboratory of Electronics & Micro-electronics, Faculty of Science of Monastir, Tunisia[5]

***Abstract*—Remote sensing systems generate large volumes of high-dimensional data, which creates significant challenges for onboard processing and data transmission under computational, memory, and energy constraints. Dimensionality reduction is therefore essential for enabling efficient embedded remote sensing applications. This work presents a hardware-accelerated implementation of Principal Component Analysis (PCA) on an FPGA platform for low-latency onboard data reduction. A hardware–software co-design methodology is adopted and implemented using the Vivado design flow on an Xilinx FPGA. The proposed architecture focuses on accelerating the most computationally intensive stage of PCA while preserving the most informative components of the input data. Experimental evaluations demonstrate that the FPGA-based implementation achieves a 2.8× speedup while reducing energy consumption by 72.4% compared with an ARM-only software implementation, while maintaining efficient hardware resource utilization. These results confirm the suitability of the proposed accelerator for low-latency, energy-efficient onboard dimensionality reduction in resource-constrained remote sensing systems.**

***Keywords*—Remote sensing; hyperspectral data; dimensionality reduction; Principal Component Analysis (PCA); FPGA; hardware–software co-design**

I. Introduction

Remote sensing encompasses a wide range of techniques that enable the acquisition and analysis of information about observed scenes without direct physical contact. Among these techniques, hyperspectral imaging has gained significant attention due to its ability to capture fine-grained spectral information across hundreds of contiguous wavelength bands [1]. Unlike conventional multispectral sensors, hyperspectral systems acquire three-dimensional data cubes composed of two spatial dimensions and a highly resolved spectral dimension, where each pixel is characterized by a detailed spectral signature. This rich spectral representation enables accurate material discrimination and supports numerous applications, including environmental monitoring, precision agriculture, geological mapping, and natural resource management [2, 3].

Despite these advantages, hyperspectral imaging introduces major challenges related to data processing and analysis. The high dimensionality of hyperspectral data, combined with strong inter-band correlation, results in substantial data redundancy and increased sensitivity to noise and sensor imperfections [4]. These characteristics significantly increase computational complexity, memory requirements, and power consumption, rendering conventional software-based processing approaches inefficient for embedded applications.

These challenges are critical in Earth observation systems, where hyperspectral data must be processed onboard and transmitted under limited bandwidth, memory, and energy constraints. Dimensionality reduction is therefore essential to reduce data volume while preserving informative spectral features. Several methods, including Randomized Principal Component Analysis (RPCA) and unsupervised band selection techniques, have been widely used to reduce redundancy in hyperspectral data [5, 6]. Among them, PCA remains attractive due to its simplicity, robustness, and ability to retain most data variance with fewer components. Deep learning methods, such as hybrid 3D–2D CNNs, have also achieved strong performance and often use PCA to reduce input dimensionality and computational cost [7, 8,9] . Transformer-based models further improve spectral modeling but require high memory, computation, and energy resources [10–12]. To address these limitations, FPGA-based acceleration has been explored for PCA, hyperspectral compression, CNN inference, and anomaly detection [13–17], while in-sensor and processing-in-pixel paradigms have also been proposed to reduce data movement and energy consumption [18, 19]. Thus, PCA remains a suitable dimensionality reduction method for resource-constrained onboard hyperspectral systems.

In this context, this study proposes a real-time FPGA-based accelerator for Principal Component Analysis tailored to onboard hyperspectral data processing. The proposed solution adopts a hardware–software co-design approach and is implemented on a PYNQ-Z2 FPGA platform using the Vivado design flow. Unlike prior FPGA-based PCA accelerators that either lack configurability for varying hyperspectral dimensions or omit systematic energy analysis, the proposed architecture offers three distinctive contributions: 1) a fully pipelined, HLS-optimized covariance engine with loop unrolling (factor 8) specifically tuned for the B×B spectral covariance structure of hyperspectral cubes; 2) an AXI4-Master burst transfer strategy that aligns I/O parallelism with DDR memory bandwidth on the Zynq-7020; and 3) a hardware–software co-design that achieves an energy reduction of approximately 72% compared to ARM-only execution

*Corresponding author.

(from 4.39 J to 1.21 J per PCA run), alongside a 2.8× speedup. These characteristics address key gaps in the existing literature and confirm the suitability of the proposed solution for high-performance onboard hyperspectral processing under energy constraints.

II. Related Work

Hyperspectral image analysis has motivated extensive research on dimensionality reduction methods designed to reduce spectral redundancy while preserving the most informative features. Techniques such as Principal Component Analysis (PCA) [20], Singular Value Decomposition (SVD) [21], and band selection [22] have been widely used for dimensionality reduction in hyperspectral image analysis. Among them, PCA remains one of the most commonly adopted approaches because of its simplicity, robustness, and ability to transform highly correlated spectral bands into a reduced set of uncorrelated components. Unlike band selection methods, which preserve the original bands but may discard useful information, PCA provides a compact representation that concentrates most of the data variance in fewer components.

Nonlinear dimensionality reduction methods, including manifold learning approaches such as Locally Linear Embedding (LLE) [23] and Laplacian Eigenmaps (LE) [24], have also been explored to better capture complex spectral structures in hyperspectral data. However, these methods generally require higher computational resources and are sensitive to parameter settings [25], which limit their applicability in practice. For this reason, PCA remains widely adopted as an efficient preprocessing step in hyperspectral classification pipelines [26].

With the development of deep learning, several spatial–spectral models have been proposed for hyperspectral image classification. Hybrid CNN architectures have shown strong performance by extracting both spectral and spatial features. In many of these works, PCA is applied before classification to reduce the number of input bands and decrease the computational cost of 3D convolution operations. For example, Ghaderizadeh et al. [7] demonstrated that PCA-reduced hyperspectral data can be effectively used with hybrid 3D–2D CNN models to achieve competitive classification accuracy with reduced complexity. Similarly, Datta et al. [19] combined superpixelwise PCA with dense CNN-based spatial–spectral feature learning, further confirming the usefulness of PCA in deep learning-based hyperspectral analysis.

More recently, transformer-based architectures have been introduced to model long-range spectral dependencies through self-attention mechanisms. Models such as SpectralFormer [12] and MassFormer [11] have demonstrated promising classification performance by exploiting full spectral information. Nevertheless, these approaches usually involve high memory usage, significant computational complexity, and increased energy consumption. As a result, their deployment on resource-constrained onboard platforms remains challenging, particularly in applications where latency, memory, and power consumption are critical constraints.

To address these limitations, hardware acceleration has become a key direction for onboard hyperspectral processing. FPGAs are particularly suitable for this purpose due to their parallelism, deterministic execution, and higher energy efficiency compared with general-purpose processors. Fernández et al. [13] proposed an FPGA-based architecture for hyperspectral dimensionality reduction, demonstrating the potential of reconfigurable hardware for accelerating spectral projection operations. He et al. [14] developed a configurable FPGA accelerator for 2D–3D CNN-based hyperspectral classification, achieving efficient resource utilization and reduced inference latency through an HLS-based design flow.

Despite these advances, many existing FPGA-based implementations mainly focus on latency, throughput, or resource utilization separately. Few works provide a combined analysis of execution time, energy consumption, and configurability for different hyperspectral data dimensions. In addition, several prior designs do not provide sufficient details about memory transfer mechanisms or AXI-based communication strategies, which limits reproducibility and practical deployment.

Therefore, the present work addresses these limitations by proposing a hardware–software co-designed PCA accelerator implemented on the Zynq-7020 PYNQ-Z2 platform. The proposed architecture aims to jointly improve latency, energy efficiency, and configurability, while providing a detailed characterization of fixed-point arithmetic, AXI burst transfer, and energy consumption per PCA run.

III. Materials and Methods

A. PCA Algorithm for Hyperspectral Images

The conventional Principal Component Analysis (PCA) algorithm operates on two-dimensional data matrices. However, hyperspectral images are inherently multivariate datasets represented as three-dimensional data cubes, composed of two spatial dimensions $(D_{x'} D_y)$and one spectral dimension (D_λ). Therefore, the first step in applying PCA to hyperspectral data consists of unfolding (reshaping) the hyperspectral cube into a two-dimensional matrix.

$$X \in \mathbb{R}^{N \times B},$$

where, $N = D_x \times D_y$denotes the total number of pixels and $B = D_\lambda$ represents the number of spectral bands. In this representation, each row corresponds to the spectral signature of a single pixel, while each column represents a specific spectral band across the entire image. The unfolding process is illustrated in Fig. 1, where a hyperspectral cube composed of three rows, three columns, and B spectral bands is reshaped into a two-dimensional matrix suitable for PCA computation.

Before applying PCA, data standardization is carried out to ensure that all spectral bands contribute equally to the analysis. Each spectral variable is normalized by subtracting its mean and dividing by its standard deviation, as expressed by:

$$x'_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad (1)$$

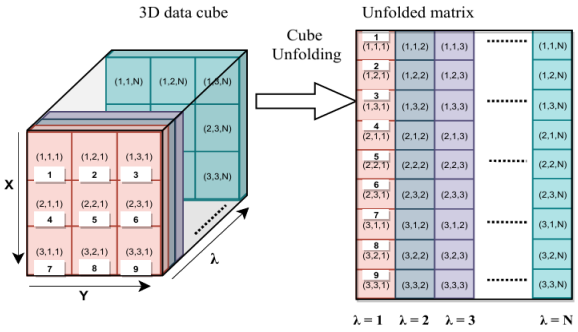


Fig. 1. Unfolding a hypercube.

where, x_{ij} denotes the reflectance value of the j-th spectral band at the i-th pixel, and μ_j and σ_j are the mean and standard deviation of the j-th band, respectively. This step enhances numerical stability and prevents spectral bands with larger dynamic ranges from dominating the PCA process.

After standardization, the covariance matrix is computed from the standardized data matrix X_s as:

$$C = \frac{1}{N-1} X_s^T X_s \tag{2}$$

The covariance matrix captures the inter-band spectral correlations and forms the foundation for dimensionality reduction.

PCA is then performed by applying eigenvalue decomposition to the covariance matrix:

$$C v_k = \lambda_k v_k \tag{3}$$

where, λ_k and v_k represent the eigenvalues and their corresponding eigenvectors, respectively. Each eigenvector defines a principal component, and its associated eigenvalue indicates the amount of variance explained by that component.

The eigenvalues are sorted in descending order, and the first K principal components are selected according to the cumulative explained variance criterion. The dimensionality-reduced data are then obtained by projecting the standardized data onto the selected eigenvectors:

$$Y = X_s V_k \tag{4}$$

where, V_k contains the K selected eigenvectors. The resulting matrix Y represents the reduced hyperspectral representation, corresponding to the PCA sub-bands that preserve the most informative spectral content of the original hyperspectral image while significantly reducing data dimensionality.

Hyperspectral images are characterized by a very high number of spectral bands, resulting in extremely high-dimensional data volumes and substantial computational complexity during onboard or embedded processing. The simultaneous handling of hundreds of spectral bands renders conventional software-based implementations inadequate for embedded and low-latency applications due to excessive execution latency, high memory bandwidth demands, and increased power consumption. In this context, Principal Component Analysis (PCA) is widely adopted as an effective and well-established dimensionality reduction technique, enabling hyperspectral data to be projected onto a lower-dimensional subspace while preserving most of the original data variance. However, to fully exploit the benefits of PCA in embedded sensing applications, an efficient hardware implementation becomes necessary to satisfy stringent latency and energy constraints while ensuring reliable and continuous data processing.

B. Proposed Hardware–Software Co-Design Architecture for PCA Acceleration

The proposed PCA acceleration architecture follows a hardware–software co-design approach. The covariance matrix computation, characterized by intensive numerical operations, is implemented as a custom FPGA IP core, while the remaining PCA stages are executed on the ARM-based processing system. These software-executed stages include data centering, eigenvalue decomposition, and principal component selection, which benefit from software flexibility and control.

Fig. 3 illustrates the proposed hardware–software co-design architecture of the FPGA-based PCA accelerator. The covariance IP is fully integrated into the FPGA and operates in a highly parallel and pipelined manner, leading to substantial performance gains compared to a pure software implementation. The reconfigurable nature of the FPGA allows the architecture to be tailored to different hyperspectral data dimensions by adjusting parameters such as the number of spectral bands and accumulation depth. By combining hardware-accelerated covariance computation with software-based PCA control, the proposed system achieves reduced latency, increased throughput, efficient hardware resource usage, and improved energy efficiency. These characteristics make the architecture well suited for high-performance hyperspectral data processing in embedded remote sensing applications.

The design methodology adopted in this work is based on Vivado High-Level Synthesis (HLS) and the Vivado Design Suite, as illustrated in Fig. 2, and is implemented on the PYNQ-Z2 development platform. The PCA algorithm is first described using high-level C/C++ code and validated through a dedicated C-based testbench, which serves as a golden reference model to ensure functional correctness. C-level simulation is then performed to verify the behavior of the algorithm prior to hardware synthesis. Vivado HLS is employed to translate the high-level PCA description into register-transfer level (RTL) code, while enabling the insertion of synthesis directives and optimization constraints to improve latency, throughput, and hardware resource utilization. After C/RTL co-simulation, the generated RTL is exported and integrated into a Zynq-7020–based PYNQ-Z2 system using Vivado. This workflow enables a seamless transition from algorithmic modeling to hardware implementation while preserving functional correctness and providing fine-grained control over performance optimization.

Among the different stages of the PCA algorithm, the covariance matrix computation represents one of the most computationally intensive operations. For hyperspectral images from the Pavia University dataset [22], this corresponds to

approximately N=207,400 samples and B=103 spectral bands, resulting in a 103×103 covariance matrix. This computation requires a large number of multiply–accumulate operations, leading to high execution time in software-based implementations.

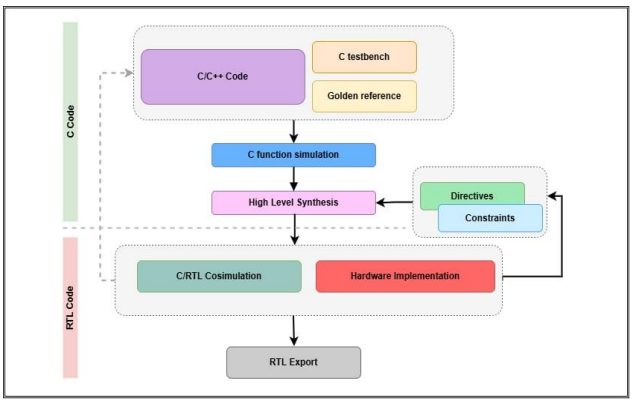


Fig. 2. Design and verification framework for custom hardware accelerator.

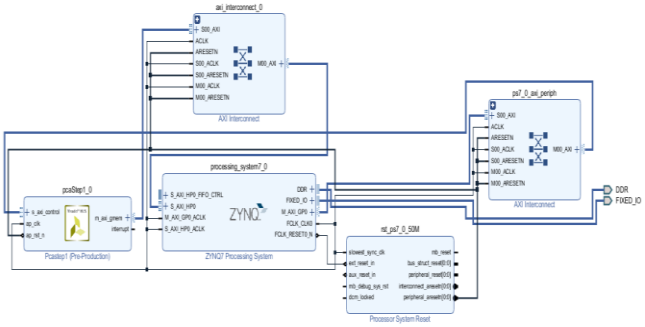


Fig. 3. Block diagram of the FPGA-based PCA acceleration architecture on Zynq SoC.

To reduce latency and meet stringent performance constraints, the covariance computation is offloaded to a dedicated FPGA IP core implemented in the programmable logic of the PYNQ-Z2 platform. The proposed IP exploits fine-grained parallelism through loop unrolling and loop pipelining directives in the HLS design, allowing multiple accumulation operations across spectral bands to be executed concurrently within each clock cycle. This significantly reduces the number of clock cycles required for covariance matrix generation and improves throughput while maintaining efficient hardware resource utilization.

To further enhance the performance of the covariance matrix computation, several HLS-level optimization strategies are applied. Loop unrolling with a factor of 8 (#pragma HLS unroll factor=8) enables parallel accumulation operations across spectral bands, while loop pipelining allows continuous processing of hyperspectral samples with a low initiation interval. Array partitioning is used to increase parallel memory access to spectral data buffers, and on-chip memory buffering is employed to locally store intermediate accumulation results, thereby reducing external DDR memory transactions.

In addition, fixed-point arithmetic is adopted using a 32-bit fixed-point representation with 16 integer bits and 16 fractional bits to provide a balance between numerical precision and hardware efficiency. Optimized AXI burst transfers are also employed to improve memory throughput and reduce power consumption. The covariance IP accesses hyperspectral data stored in external DDR memory via an AXI4-Master interface, while configuration parameters such as data dimensions and memory addresses are managed through an AXI-Lite control interface. An interrupt mechanism is used to notify the processing system upon completion of the covariance computation.

IV. Results

This section reports the experimental results of the proposed PCA accelerator implemented on the PYNQ-Z2 FPGA. The evaluation considers hyperspectral data from the Pavia University dataset and analyzes both the quality of the extracted principal components and the performance of the hardware-accelerated implementation in terms of execution time, power consumption, and resource usage. A comparison with a software-based PCA execution on the ARM Cortex-A9 is also presented to demonstrate the advantages of the proposed hardware solution.

TABLE I. FPGA Resource Utilization After Vivado Implementation of the PCA Accelerator System on PYNQ-Z2 (XC7Z020)

Resource (PL)	Used	Available	Utilization (%)
LUT	16,881	53,200	31.7%
FF	14,673	106,400	13.8%
BRAM	1	140	0.7%
DSP48	174	220	79.1%

Table I presents the FPGA resource utilization obtained after Vivado post-implementation of the PCA accelerator system on the PYNQ-Z2 platform based on the Xilinx Zynq-7020 (XC7Z020) SoC. The results show a moderate consumption of general logic resources, with 16,881 LUTs (31.7%) and 14,673 flip-flops (13.8%), indicating that the proposed architecture maintains a balanced use of combinational and sequential logic for the complete programmable-logic design, including the system interconnect, control and interface logic, together with the PCA accelerator.

In contrast, the utilization of DSP48 blocks reaches 174 units (79.1%), which is consistent with the arithmetic-intensive nature of the covariance matrix computation, dominated by multiply–accumulate operations and matrix-based numerical processing. This confirms that the design efficiently maps computationally demanding operations onto dedicated DSP resources, while preserving LUT resources mainly for control and interfacing purposes. The BRAM usage is very limited, with only one BRAM block used (0.7%), reflecting a lightweight on-chip memory footprint within the programmable logic.

Overall, these results demonstrate an efficient hardware implementation of the PCA accelerator on PYNQ-Z2, achieving high computational density while maintaining moderate logic and memory utilization, making it suitable for embedded dimensionality reduction applications with stringent performance requirements.

TABLE II. PERFORMANCE COMPARISON OF THE PCA IMPLEMENTATION ON ARM CORTEX-A9 AND ZYNQ XC7Z020.

Platform	Execution Time (s)	Power (W)
Dual-Core ARM Cortex-A9	1.83	2.4
Zynq XC7Z020	0.65	1.86

To evaluate the performance benefits of hardware acceleration for the Principal Component Analysis (PCA) algorithm, the proposed implementation was evaluated on two different platforms: a software-only execution on the dual-core ARM Cortex-A9 processor and a hardware-accelerated implementation on the Zynq XC7Z020 FPGA.

Table II compares the execution time and power consumption of the PCA implementation when executed entirely on the ARM Cortex-A9 processor and when accelerated using the FPGA resources of the Zynq XC7Z020 platform. The software-only execution on the ARM Cortex-A9 requires 1.83 s, reflecting the sequential nature of the PCA covariance computation and the limited parallelism available on a general-purpose processor.

The FPGA-accelerated implementation significantly reduces the execution time to 0.65 s, corresponding to a speed-up of approximately 2.8×. This improvement is achieved by offloading the computationally intensive stages of the PCA to dedicated hardware, enabling parallel processing and efficient pipelining in the programmable logic.

From a power perspective, the FPGA-based solution also demonstrates an advantage, consuming 1.86 W compared to 2.4 W for the ARM-only execution. Although both platforms operate within acceptable power limits, the combination of lower execution time and reduced power consumption highlights the superior energy efficiency of the FPGA-accelerated approach. Overall, these results confirm that integrating a hardware accelerator for PCA on the Zynq XC7Z020 platform provides a favorable trade-off between performance and power, making it well suited for embedded and real-time signal processing applications.

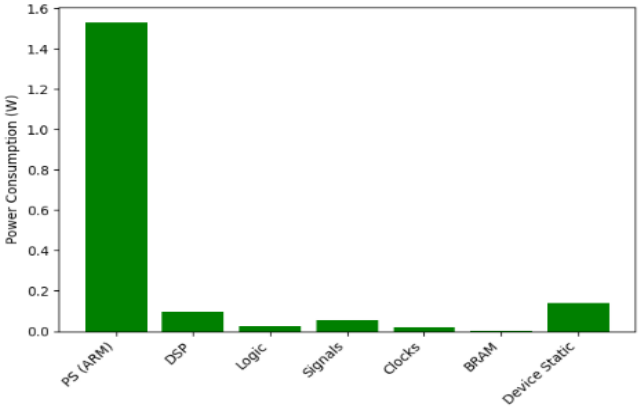


Fig. 4. Power consumption breakdown of the PCA system on the Zynq XC7Z020 platform.

To evaluate the power efficiency of the proposed PCA accelerator and to illustrate the distribution of power consumption between the processing system and the programmable logic, Fig. 4 presents the power consumption breakdown of the PCA system implemented on the Zynq XC7Z020 platform. The results show a total on-chip power consumption of 1.867 W, with 1.726 W (92.4%) corresponding to dynamic power and 0.141 W (7.6%) to static power, indicating that the system consumption is largely dominated by active computation. As shown in Fig. 4, the Processing System (PS7) represents the main contributor, consuming 1.530 W, which corresponds to 88.6% of the dynamic power and 81.9% of the total power. In contrast, the programmable logic (PL), including the PCA hardware accelerator, accounts for only 0.196 W of dynamic power (11.4% of dynamic power and 10.5% of total power), demonstrating the low energy overhead introduced by hardware acceleration. Within the PL, the power consumption is mainly dominated by the DSP blocks (0.094 W, 5.0%), which is consistent with the arithmetic-intensive nature of the PCA covariance computation, while the contributions of logic (0.025 W, 1.3%), clocks (0.020 W, 1.1%), signals (0.055 W, 2.9%), and BRAM (0.002 W, 0.1%) remain negligible. Although the PS dominates instantaneous power consumption, the reduced execution time achieved by the FPGA accelerator leads to an energy-efficient PCA implementation suitable for embedded and performance-constrained applications.

To fully characterize the efficiency of the proposed accelerator, energy consumption per PCA run is quantified as $E = P \times T$ and further normalized to the dataset dimensions ($N = 207{,}400$ pixels, $B = 103$ spectral bands) to derive the energy per pixel per band. The resulting values are reported in Table III.

TABLE III. ENERGY EFFICIENCY COMPARISON PER PCA RUN

Platform	Energy per Run (J)	Energy per pixel/band (nJ)
Dual-Core ARM Cortex-A9	4.39	205
Zynq XC7Z020 (FPGA)	1.21	56.6

The results show that the FPGA-accelerated implementation consumes only 1.21 J per PCA run, compared to 4.39 J for the ARM-only execution, achieving an energy reduction of approximately 72.4%. At the pixel–band level, the energy footprint decreases from approximately 205 nJ/pixel/band to 56.6 nJ/pixel/band, showing that the efficiency gain is reflected across the processed hyperspectral data volume. This reduction results from the simultaneous decrease in both execution time and average power consumption. Such an improvement is particularly important for embedded remote sensing platforms, where energy consumption, memory usage, and execution time must be carefully optimized. These results demonstrate that the proposed PCA accelerator provides not only a performance advantage but also a substantial improvement in energy efficiency, supporting its potential for energy-constrained onboard remote sensing applications.

To illustrate the effect of Principal Component Analysis (PCA) on hyperspectral data, PCA was applied to hyperspectral images from the Pavia University dataset, and the resulting principal components were visualized. This representation enables an initial qualitative assessment of how the spectral information contained in the Pavia University

dataset is redistributed across the principal components. Fig. 5 illustrates the first eight principal components obtained after applying Principal Component Analysis (PCA) to the hyperspectral dataset. It can be observed that the first principal components (Bands 1–3) preserve the most significant spatial structures of the scene, with clear edges, textures, and man-made objects being distinctly visible. These components concentrate the majority of the data variance and therefore carry the most relevant spectral–spatial information. The intermediate components (Bands 4 and 5) still contain meaningful information but with reduced contrast and fewer discernible details, indicating that they mainly capture secondary variations in the data. In contrast, the higher-order components (Bands 6–8) exhibit low spatial structure and appear increasingly uniform, suggesting that they are dominated by noise and minor spectral variations rather than useful information.

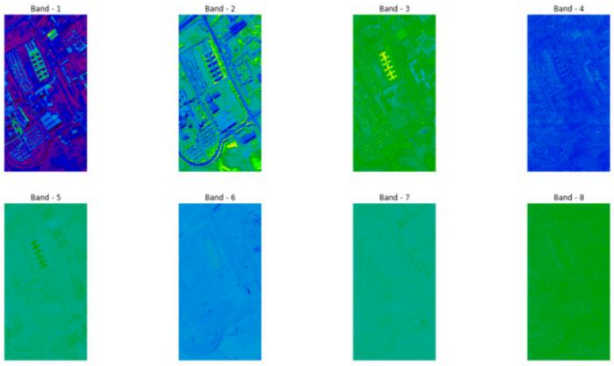


Fig. 5. The first principal components obtained from hyperspectral images.

V. Discussion

The proposed PCA acceleration architecture adopts a hardware–software co-design approach that effectively balances performance and flexibility for hyperspectral data processing. By selectively offloading the covariance matrix computation to the programmable logic, the system accelerates the most computationally intensive stage of PCA while maintaining software-level control for the remaining processing steps.

The covariance computation is implemented as a custom FPGA IP core generated using Vivado High-Level Synthesis and integrated into the Zynq-7020–based PYNQ-Z2 platform. This IP exploits fine-grained parallelism through advanced HLS optimization techniques, including loop pipelining and loop unrolling, enabling concurrent accumulation operations across multiple spectral bands. As a result, the number of clock cycles required for covariance matrix generation is significantly reduced compared to a pure software implementation on the ARM Cortex-A9 processor.

Overall, the combination of HLS-based design, pipelined execution, and parallel accumulation allows the proposed architecture to achieve reduced latency, increased throughput, and efficient FPGA resource utilization. These characteristics demonstrate the suitability of the proposed solution for onboard processing in embedded remote sensing systems operating under strict power and computational constraints.

From a system-level perspective, the proposed architecture is designed to accommodate the statistical variability inherent to hyperspectral data acquired under different sensing conditions. The hardware–software co-design paradigm provides sufficient flexibility to adapt the processing flow through software control and parameter configuration, reinforcing the robustness and applicability of the approach in practical remote sensing scenarios.

Future work will focus on further enhancing the flexibility and scalability of the proposed PCA accelerator. One important direction is the development of more configurable architectures capable of adapting to varying input dimensions and data characteristics while preserving real-time performance.

In addition, extending the current design to support more advanced dimensionality reduction techniques or hybrid FPGA–AI approaches represents a promising direction to further improve performance and enable more complex onboard remote sensing applications.

VI. Conclusion

This work presented an FPGA-based hardware acceleration framework for Principal Component Analysis (PCA) applied to hyperspectral image processing, following a hardware–software co-design methodology. Through a detailed analysis of the PCA processing pipeline, the covariance matrix computation was identified as the most computationally demanding stage due to its intensive arithmetic operations and high data volume.

To address this challenge, a dedicated PCA covariance hardware intellectual property (IP) core was designed using Vivado High-Level Synthesis and integrated into a Zynq-7020–based PYNQ-Z2 platform. By exploiting fine-grained parallelism through loop pipelining and unrolling, the proposed accelerator significantly reduced execution latency compared to a software implementation on a dual-core ARM Cortex-A9 processor (from approximately 1.83 s to 0.65 s), while also lowering power consumption (from 2.4 W to 1.86 W).

These results confirm the effectiveness of the proposed hardware acceleration strategy in enabling hyperspectral data processing under strict computational and energy constraints, while maintaining efficient FPGA resource utilization.

Future work will focus on extending hardware acceleration to additional PCA stages and exploring further architectural optimizations to enhance scalability and throughput on more advanced FPGA platforms.

References

[1] D. Cozzolino, P. J. Williams, and L. C. Hoffman, “An overview of pre-processing methods available for hyperspectral imaging applications,” Microchem. J., vol. 193, p. 109129, 2023.

[2] R. Vaddi, B. L. N. Phaneendra Kumar, P. Manoharan, L. Agilandeeswari, and V. Sangeetha, “Strategies for dimensionality reduction in hyperspectral remote sensing: A comprehensive overview,” Egypt. J. Remote Sens. Space Sci., vol. 27, no. 1, pp. 82–92, 2024.

[3] R. Qureshi, M. Uzair, K. Khurshid, and H. Yan, "Hyperspectral document image processing: Applications, challenges and future prospects," Pattern Recognition, vol. 90, pp. 12–22, 2019.

[4] L. Zhu, Y. Chen, P. Ghamisi, J. A. Benediktsson, M. Li, and R. Tao, “Generative adversarial networks for hyperspectral image

classification," IEEE Trans. Geosci. Remote Sens., vol. 56, no. 9, pp. 5046–5063, Sep. 2018.

[5] M. Ustuner, "Randomized principal component analysis for hyperspectral image classification," in 2024 IEEE Mediterranean and Middle-East Geoscience and Remote Sensing Symposium, pp. 26–30, 2024.

[6] R. Vaddi and P. Manoharan, "Hyperspectral remote sensing image classification using combinatorial optimisation based un-supervised band selection and CNN," IET Image Process., vol. 14, no. 15, pp. 3909–3919, 2020.

[7] S. Ghaderizadeh, D. Abbasi-Moghadam, A. Sharifi, N. Zhao, and A. Tariq, "Hyperspectral image classification using a hybrid 3D–2D convolutional neural network," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 15, pp. 2379–2393, 2022

[8] H. Firat, M. E. Asker, and D. Hanbay, "Classification of hyperspectral remote sensing images using different dimension reduction methods with 3D/2D CNN," Remote Sens. Appl.: Soc. Environ., vol. 25, p. 100694, 2022.

[9] D. Datta, P. K. Mallick, D. Gupta, et al., "Hyperspectral image classification based on novel hybridization of spatial-spectral-superpixelwise principal component analysis and dense 2D–3D convolutional neural network fusion architecture," Can. J. Remote Sens., vol. 48, no. 5, pp. 663–680, 2022.

[10] L. Scheibenreif, M. Mommert, and D. Borth, "Masked vision transformers for hyperspectral image classification," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 2166–2176, 2023..

[11] L. Sun, H. Zhang, Y. Zheng, et al., "Massformer: Memory-augmented spectral-spatial transformer for hyperspectral image classification," IEEE Trans. Geosci. Remote Sens., vol. 62, pp. 1–15, 2024.

[12] D. Hong, Z. Han, J. Yao, et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–15, 2021.

[13] D. Fernández, C. González, D. Mozos, and S. López, "FPGA implementation of the principal component analysis algorithm for dimensionality reduction of hyperspectral images," Journal of Real-Time Image Processing, vol. 16, no. 5, pp. 1395–1406, 2019.

[14] W. He, Y. Yang, S. Mei, J. Hu, W. Xu, and S. Hao, "Configurable 2D–3D CNN accelerator for FPGA-based hyperspectral imagery classification," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 16, pp. 9406–9421, 2023.

[15] K. Avagian and M. Orlandić, "An efficient FPGA implementation of Richardson–Lucy deconvolution algorithm for hyperspectral images," Electronics, vol. 10, no. 4, p. 504, 2021.

[16] J. Caba, M. Díaz, J. Barba, A. Sánchez, and R. Sarmiento, "FPGA-based on-board hyperspectral imaging compression: Benchmarking performance and energy efficiency against GPU implementations," Remote Sens., p. 3741, 2020.

[17] J. Caba, M. Díaz, J. Barba, A. Sánchez, and R. Sarmiento, "Low-power hyperspectral anomaly detector implementation in cost-optimized FPGA devices," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 15, pp. 2379–2393, 2022.

[18] G. Datta, Z. Yin, A. Jacob, A. Majumdar, and A. Veeraraghavan, "Toward efficient hyperspectral image processing inside camera pixels," arXiv preprint arXiv:2203.05696, 2022.

[19] G. Datta and C. Zhou, "Analog vs. Digital In-Sensor Computing: A Tale of Two Paradigms," in Proc. Great Lakes Symp. VLSI (GLSVLSI), pp. 829–834, 2025.

[20] K. M. Lim, C. P. Lee, Z. Zahisham, J. Y. Lim, and J. N. Mogan, "PCA-ViT: Hyperspectral image classification using principal component analysis and vision transformer," in Proc. IEEE 12th Conf. Syst., Process Control (ICSPC), pp. 30–34, 2024.

[21] M. M. Rahman, M. R. Islam, M. I. Afjal, M. A. H. Akhand, and M. A. H. Khan, "Segmentation-based truncated-SVD for effective feature extraction in hyperspectral image classification," Int. J. Remote Sens., vol. 46, no. 2, pp. 538–574, 2025.

[22] W. Kurz, K. Wang, F. Bektaş, M. M. Rahman, and M. M. Islam, "Dimensionality reduction in hyperspectral imaging using standard deviation-based band selection for efficient classification," Sci. Rep., vol. 15, no. 1, p. 34478, 2025.

[23] H. Ji and Z. Zuo, "Multiview locally linear embedding for spectral-spatial dimensionality reduction of hyperspectral imagery," IEEE/CAA Journal of Automatica Sinica, vol. 9, no. 6, pp. 1091–1094, 2022..

[24] S.-E. Qian and G. Chen, "A new nonlinear dimensionality reduction method with application to hyperspectral image analysis," in 2007 IEEE International Geoscience and Remote Sensing Symposium, pp. 270–273, 2007.

[25] D. Lunga, S. Prasad, M. M. Crawford, and O. Ersoy, "Manifold-learning-based feature extraction for classification of hyperspectral data: A review of advances in manifold learning," IEEE Signal Processing Magazine, vol. 31, no. 1, pp. 55–66, 2014.

[26] W. Song, X. Zhang, G. Yang, Y. Liu, and X. Li, "A study on dimensionality reduction and parameters for hyperspectral imagery based on manifold learning," Sensors, vol. 24, no. 7, p. 2089, 2024.

Multiscale Image Enhancement and Visualization Analysis for Historical Map/s

Guilei Ren
Faculty of History and Archaeology, Anyang Normal University, Anyang 455000, China

***Abstract*—Based on the degradation of historical maps, such as fading, stains, creases, blurred lines, and text adhesion, multi-scale image enhancement and spatial element visualization analysis are studied. This study constructs a technical process consisting of sample division, image preprocessing, multi-scale enhancement, quality evaluation, and spatial feature extraction, and further evaluates map registration, vectorization, optical character recognition, and semantic segmentation on the same test maps. The results show that the PSNR, SSIM, EPI, and RCS of the multi-scale enhancement model are 27.94 dB, 0.842, 0.817, and 0.81, respectively, and the comprehensive quality score is 0.872. The road connectivity, water boundary matching, and administrative boundary topology consistency rates reach 87.9%, 85.4%, and 82.6%, while place name and map symbol recognition rates increase to 84.7% and 82.3%. Downstream evaluation shows that registration RMSE decreases from 8.31 to 4.12 px, vectorization F1 increases from 0.713 to 0.871, OCR character accuracy increases from 62.4% to 84.7%, and semantic-segmentation mIoU increases from 0.614 to 0.802. These results indicate that the enhancement improves both visual quality and the reliability of GIS-oriented spatial processing.**

Keywords*—Historical maps; multiscale image enhancement; image restoration; spatial feature recognition; visualization analysis

I. Introduction

Historical maps preserve urban morphology, transport networks, water systems, administrative boundaries and place names, supporting historical geography and cultural-heritage digitization. Long-term storage and scanning introduce fading, stains, creases, blurred linework and text adhesion, reducing both readability and spatial-processing reliability. The task is therefore to restore line, text and symbol structures without amplifying paper noise or altering spatial relationships. The methodological contribution is a degradation-aware, task-coupled workflow rather than a new primitive enhancement operator: quantified degradation characterization guides controlled evaluation, while global, local and edge responses are fused to preserve map topology and support spatial-feature extraction and GIS-downstream tasks.

II. Related Work

A. Historical Map Digitization and Feature Extraction

Existing historical-map studies are largely task-specific. Schlegel [1] proposed semi-automated object extraction, treating preprocessing as support for extraction rather than an independently evaluated stage. Mäyrä et al. [2] used historical maps for long-term land-use interpretation, emphasizing source quality rather than restoration procedures. Ahmed et al. [3] applied unsupervised learning and GEOBIA to cadastral-map digitization, while Mutch and Newsam [4] examined conversion of hand-drawn maps into GIS; both concentrate on structural conversion. Arzoumanidis et al. [5] addressed semantic segmentation, whereas Pai et al. [6] focused on symbol detection. These approaches optimize individual extraction or recognition objectives and leave the effect of restoration quality across multiple spatial tasks insufficiently examined.

B. Multiscale Enhancement and Research Gap

Vijayalakshmi et al. [7] developed a multiscale enhancement framework for low-contrast images, but its formulation is not tailored to historical maps where paper texture, weak linear features, text adhesion, and topological continuity coexist. The contribution claimed here is therefore not a new brightness-correction operator, texture enhancer, edge filter, optimization algorithm, or standalone feature extractor. It lies primarily in a degradation-aware workflow that quantifies map-specific deterioration, couples global, local, and detail responses with line-topology and text-readability requirements, and carries the restored output into spatial extraction and GIS evaluation. The architecture is thus a task-specific organization of established enhancement operations, while the degradation model and downstream evaluation provide the map-oriented constraints that distinguish the workflow from generic image restoration.

III. Materials and Methods

A. Historical Map Data Sources and Sample Composition

Historical map samples cover urban blocks, transportation routes, waterway boundaries, administrative divisions, and topographic symbols. The primary dataset contains 120 map sheets from publicly available scanned archives, digitized historical-map datasets, and verifiable geographic sources, with 84 sheets for training, 18 for validation, and 18 for testing; scanning resolution ranges from 300 to 600 dpi and map sheets are uniformly tiled for processing. To assess transfer beyond this collection, external validation is additionally performed on the public Semantic Segmentation Map Dataset (Semap) [8], which contains 1,439 manually annotated real map samples from heterogeneous historical-map collections. The official test partition contains 144 real samples and was used exclusively for external evaluation. Semap defines background, boundary/contour, built, non-built, water, and road-network labels; following its benchmark protocol, mIoU was calculated over the four geographic classes of built, non-built, water, and road network. No Semap image was used for parameter

selection, and the complete G6 configuration was fixed before external testing.

To avoid relying solely on manual interpretation, degradation intensity is quantified for each image block from color attenuation, noise proportion, edge continuity, and text adhesion. These four terms were selected to represent the main map-specific failures of background contrast, paper contamination, line breakage, and character adhesion:

$$D_k=\lambda_1\left(1-\frac{|\mu_{b,k}-\mu_{s,k}|}{255}\right)+\lambda_2\frac{A_{n,k}}{A_k}+\lambda_3\left(1-\frac{L^c_{e,k}}{L^r_{e,k}+\varepsilon}\right)+\lambda_4\frac{A^a_{t,k}}{A_{t,k}+\varepsilon} \quad (1)$$

In Eq. (1), D_k denotes the degradation intensity value of the k th historical map image block; $\mu_{b,k}$ denotes the average grayscale value of the background region; $\mu_{s,k}$ denotes the average grayscale value of the line and text regions; $A_{n,k}$ denotes the area of the noise region; A_k denotes the total area of the image block; $L^c_{e,k}$ denotes the length of continuous edges; $L^r_{e,k}$ denotes the length of original edge responses; $A^a_{t,k}$ denotes the area of connected text regions; $A_{t,k}$ denotes the area of text candidate regions; $\lambda_1,\lambda_2,\lambda_3,\lambda_4$ denotes the weight coefficient, satisfying $\lambda_1+\lambda_2+\lambda_3+\lambda_4=1$; ε is used to avoid a zero denominator. Values below 0.35, between 0.35 and 0.65, and above 0.65 define mild, moderate, and severe degradation, respectively. The descriptor is used to balance samples and stratify subsequent evaluation; it is not introduced as a separate optimization objective.

TABLE I. SOURCES AND BASIC ATTRIBUTES OF HISTORICAL MAP SAMPLES

Sample Type	Number/Images	Time Period	Scan Resolution/dpi	Major Spatial Features	Major Degradation Features	Data Purpose
Urban Block Map	28	1900—1949	400—600	Roads, Neighborhoods, Place Names	Faded, smudged text	Training/Testing
Transportation Map	24	1910—1965	300—500	Roads, Railways, Stations	Blurred or broken lines	Training/Validation
Water system boundary map	22	1890—1958	300—600	Rivers, lakes, boundaries	Stains, partial damage	Training/Testing
Administrative map	26	1920—1978	400—600	Boundaries, Annotations, Symbols	Creases, Background Noise	Training/Validation
Topographic symbol map	20	1930—1985	300—500	Contour lines, landform symbols	Ink smudges, low contrast	Training/Testing

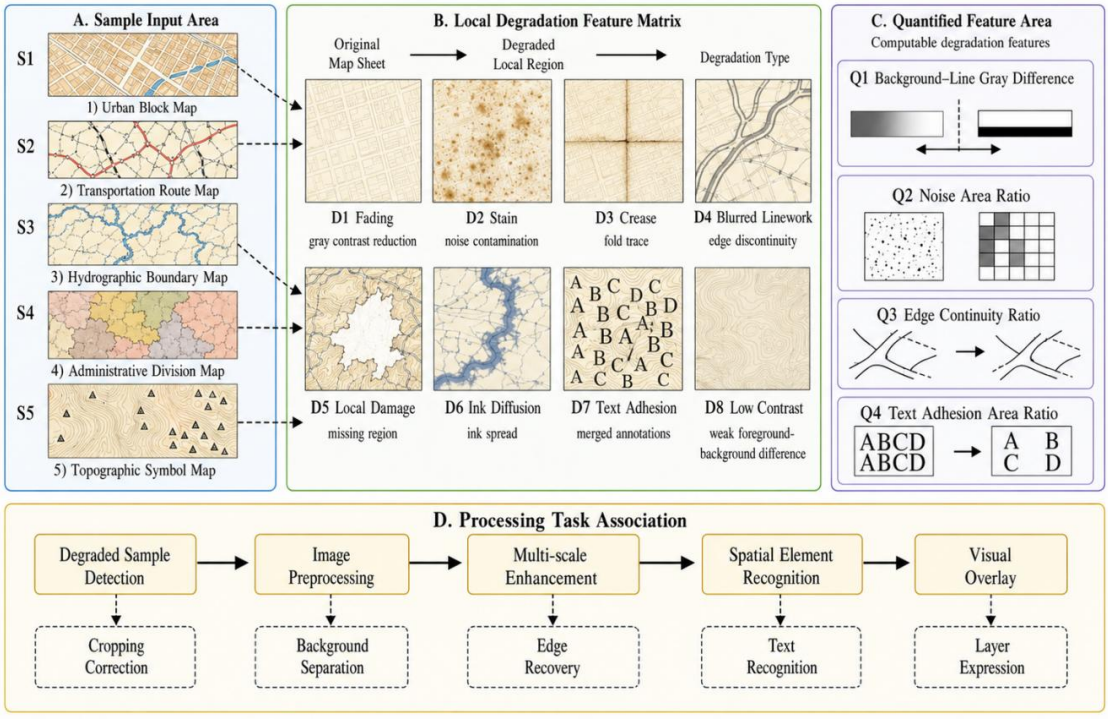


Fig. 1. Schematic diagram of degradation features in historical map samples.

The typical degradation variations in the samples are depicted in Fig. 1. The main issues with faded samples are that they have a lower grayscale contrast between the lines and background; stained samples have issues with continuity of texture in the local area; creased samples have false edges formed; and samples with smudged text have an issue with

place name recognition (Table I). Samples need to have stable spatial representation and traceable origins to allow long-term land-use identification in historical maps, and to create high-quality geographic data [9]. Hence, the original scanned images with degraded areas and spatial feature annotations are kept along with the original samples to ensure that the image enhancement results can be correlated with feature recognition, such as roads, waterways, boundaries, and place names

B. Preprocessing Workflow for Historical Map Images

The degraded intensity of the historical map samples is classified, and then the interference of the scan border, the scanning angle, the background haze, and the influence of local noise in the input to the model are removed in the preprocessing workflow. The pre-processing parameters are presented in Table II, and the combination of cropping thresholds, skew correction angles, image tile size, and overlap rate guarantees the scale consistency of all the input images; The color space conversion and the background estimation are used to remove the paper background of the map image, so that the lines of the map can be enhanced without losing the information in the faded background. All image patches are standardised to 512×512 pixels, with 25% overlap to ensure that roads, waterways and boundary lines across patches are continuous.

TABLE II. PREPROCESSING PARAMETER SETTINGS FOR HISTORICAL MAP IMAGES

Processing Step	Parameter Name	Parameter Value	Technical Constraints	Output Content
Boundary Cropping	Border Gray-Scale Threshold	235	Remove black borders and blank margins	Active Map Area
Tilt correction	Angle search range	-5° to 5°	Projection based on primary orientation Peak correction	Directionally Consistent Images
Color conversion	Color Space	RGB/CIELab	Separate Luminance and Chrominance Disturbances	Luminance Channel, Chrominance Channel
Background Estimation	Filter kernel size	31×31	Preserve main lines, suppress paper texture	Background field image
Block processing	Image block size	512×512 px	Preserve local texture Calculable	Standard image block
Overlapping sampling	Overlap rate	25%	Reduce truncation of linear features	Sequence of Overlapping Image Tiles

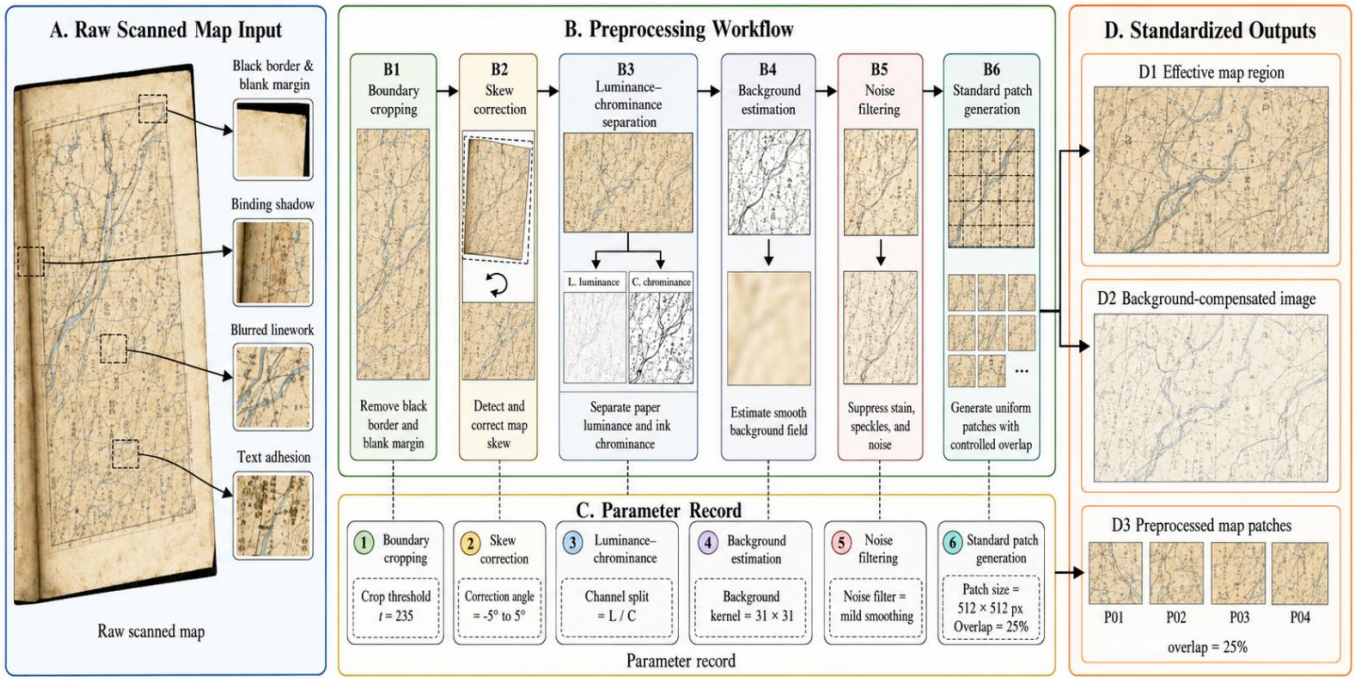


Fig. 2. Preprocessing flowchart for historical map images (example only).

Scaling normalization and suppression of noise from degraded samples is done during the preprocessing pipeline before the enhancement model takes them as input. After the raw scanned image is standardized by the above-mentioned continuous processing steps, which are illustrated in Fig. 2, the black borders, binding shadows and blank margins of the map are removed following boundary cropping, the principal direction projection correction restores the map sheets to a horizontal reference, and the luminance/chrominance separation separates the paper background, haze patterns and map linework into independent processing objects. Furthermore, the background estimation results enable the suppression of interference from non-map textures, and the preliminary noise filtering allows the reduction of stain spots, scan grain and localized gray spots. The standard tiling maintains the continuity of the road, waterway, boundary line,

and place name textures within a local area. At the same time, process parameters such as cropping range, skew correction angle, segmentation coordinates, and overlap relationships are also given, so the historical map can have the same spatial position in the three stages of enhancement, recognition, and visualization.

To preserve low-contrast line features and place name strokes, the normalization process employs a combined constraint of background compensation and local variance:

$$\mathbf{Q}^{*}_{u,v}=\mathrm{clip}_{[0,1]}\left[\frac{\mathbf{Q}_{u,v}-\mathbf{M}^{\rho}_{u,v}}{\sqrt{\left(\mathbf{S}^{\rho}_{u,v}\right)^{2}+\xi}}\cdot\gamma+\delta_{u,v}\right] \quad (2)$$

In Eq. (2), $\mathbf{Q}^{*}_{u,v}$ denotes the preprocessed output pixel at coordinates (u,v); $\mathbf{Q}_{u,v}$ denotes the original brightness pixel; $\mathbf{M}^{\rho}_{u,v}$ denotes the background estimate within a neighborhood of radius ρ; $\mathbf{S}^{\rho}_{u,v}$ denotes the local grayscale standard deviation within the same neighborhood; ξ denotes the stability term; γ denotes the contrast gain coefficient; $\delta_{u,v}$ denotes the local brightness compensation term; and $\mathrm{clip}_{[0,1]}[\cdot]$ denotes the pixel value clipped to the 0–1 range. After processing through this workflow, the sample retains its original line art structure and text edges. Subsequent quality evaluation can directly compare edge continuity, background uniformity, and symbol legibility before and after preprocessing [10].

C. Construction of a Multiscale Image Enhancement Model

The brightness of the preprocessed historical-map tiles is not uniform, the line drawings are faded, and the edges of symbols and small place-name strokes are weak; a single-scale operation can therefore amplify paper texture or suppress fine cartographic detail. The multiscale model consists of global correction, local texture enhancement, edge refinement, and scale fusion, as shown in Table III. These operations are established individually, and the architectural novelty claimed here is not the operators themselves but their map-specific response coupling. Global correction targets paper-background colour drift; local enhancement targets the grayscale contrast of roads, waterways, and boundary lines; edge refinement targets character strokes and symbol contours; and scale fusion reconciles the responses while limiting pseudo-edge diffusion and over-enhancement. Multiscale feature complementation is established in image enhancement [11]; here it is organized around preservation of historical-map line topology, text readability, and background suppression rather than generic visual contrast.

The structure of the model is presented in Fig. 3. The standard image blocks are decomposed into scale, global brightness response, local texture response, and detailed edge response. The three types of response are related to degradation problems like background color drift, lack of contrast between lines, and lack of contrast between text and symbols. The global luminance response is for suppressing uneven scanning illumination and paper haze and preventing excessively dark or bright areas across the entire map; the local texture response is for enhancement of continuous line features like roads, waterways and boundary lines, making sure that low-contrast spatial features are distinguishable within local neighborhoods; and the detail edge response is for restoring contour changes at intersections between place name strokes, legend symbols and fine lines, reducing edge discontinuities created during the enhancement process. The scale fusion process performs scale fusion on the result of various responses weighted by the spatial weights while suppressing stains, paper grain, and creases using background suppression and residual compensation. The increased output maintains the original map's line-based topological relationships and ensures a stable image base for roads, waterways, boundary lines, and place names to be extracted from the map. The quality evaluation can be done based on the structure similarity, edge preservation, and readability changes [12][13].

D. Spatial Feature Extraction and Visual Representation Methods

After enhancement, the map still contains four object types: line, area, text, and symbols, so visual morphology, grayscale response, and geographic semantics remain necessary for spatial feature extraction. The classification rules are listed in Table IV. Line continuity is used for roads and waterways; administrative boundaries rely on dashed or dotted patterns and closed-loop tendencies; place names rely on stroke direction and character spacing; and map symbols rely on contour shape and local texture. These extraction rules are not presented as a new standalone feature-extraction algorithm. They are retained as a controlled downstream probe for testing whether enhancement preserves the spatial structures needed by existing map-processing procedures. Candidate line segments, areas, text, and symbols are therefore generated from the enhanced image and filtered using the same category rules for all enhancement conditions.

TABLE III. CORRESPONDENCE TABLE OF MULTI-SCALE ENHANCEMENT MODEL MODULES AND FUNCTIONS

Module Name	Input Object	Main Processing Content	Output	Technical Constraints
Global Calibration Unit	Standard Image Block	Luminance Equalization, Chromatic Aberration Suppression	Global enhancement map	Maintain Smooth Paper Background
Local Texture Enhancement Unit	Local neighborhood blocks	Line enhancement, texture contrast enhancement	Texture Response Map	Preventing the amplification of smudge textures
Edge Refinement Unit	Edge Candidate Regions	Refinement of road, waterway, and text outlines	Edge Constraint Map	Maintain fine line continuity
Scale Fusion Module	Multiscale response map	Weight allocation and residual compensation	Enhanced Output Image	Reduce over-enhancement artifacts

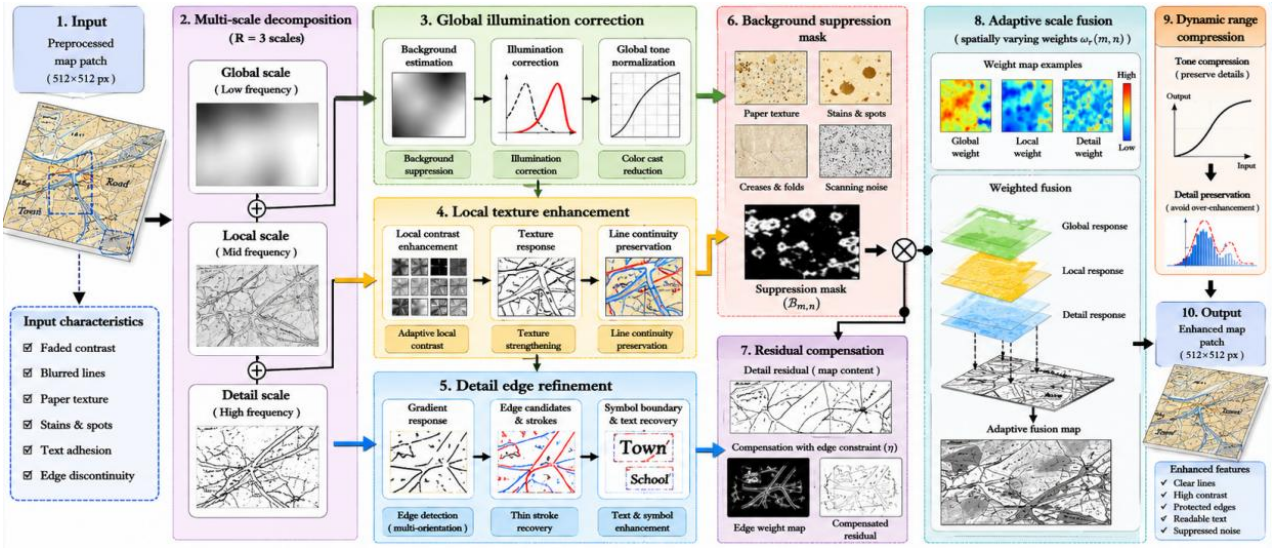


Fig. 3. Structural diagram of the multiscale image enhancement model.

TABLE IV. CLASSIFICATION AND EXTRACTION RULES FOR SPATIAL FEATURES IN HISTORICAL MAPS

Spatial Feature Type	Visual Features	Extraction Rules	Constraints	Visual Representation
Road	Continuous thin lines, intersection nodes	Edge Detection and Skeletonization	Breakpoint spacing ≤ 6 px	Line layer
Water systems	Curve boundaries, banded grayscale regions	Curve Detection and Region Connectivity	Smoothing of Curvature Changes	Blue line/surface layers
Administrative boundaries	Dotted lines, dashed lines, closed contours	Line Type Recognition and Topological Connectivity	Closure error ≤ 10 px	Boundary layer
Place names	Small-scale strokes and character clusters	Text bounding box detection	Character height ≥ 12 px	Annotation Point Layer
Map symbols	Local patterns, fixed outlines	Template matching and semantic filtering	Similarity ≥ 0.70	Symbolized layer

Candidate feature scoring employs a combination of category confidence, edge continuity, morphological matching, and noise penalty constraints:

$$E_a^{\tau}=\text{sigmoid}\left[\beta_1\overline{\Delta \mathrm{I}}_a+\beta_2\frac{\ell_a^{con}}{\ell_a^{cand}+\upsilon}+\beta_3\frac{\Omega_a^{\tau}}{\Omega_a+\upsilon}+\beta_4 P_a^{\tau}-\beta_5 N_a\right] \quad (3)$$

In Eq. (3), E_a^{τ} represents the extraction score of the a th candidate object belonging to class τ; $\overline{\Delta \mathrm{I}}_a$ represents the average gray-level difference between the candidate object and its neighborhood background; ℓ_a^{con} represents the length of continuous edges; ℓ_a^{cand} represents the total length of candidate edges; Ω_a^{τ} represents the pixel area conforming to the morphological rules of class τ; Ω_a represents the total area of the candidate object; P_a^{τ} represents the semantic class probability; N_a represents the noise response intensity; β_1 to β_5 represent weight coefficients; and υ represents the constant term. Candidate objects with a score above 0.65 go to layer coding, and those below the threshold are kept as a sample for manual review, thus reducing the number of false detections due to spots, folds, and paper grains.

As illustrated in Fig. 4, the enhanced image generates visual results for roads, waterways, borders, place names, and symbols. The workflow explains how the enhanced image is converted into spatial layers: the road is represented by a skeleton line, the water is a curve or a polygon, the administrative border is a line marker, the place is a marker, and the map symbol is a category marker. The GIS transformation and the construction of the historical map space data need to be consistent with each other. Thus, for each characteristic type, the original pixel coordinates, the map sheet number, and the candidate score are kept to facilitate the verification of the result and the error tracking.

E. Image Enhancement Quality Evaluation Metrics

The quality of historical map enhancement is not dependent on visual judgment alone; it has to measure pixel fidelity, structure preservation, edge continuity, information gain, and spatial features readability [14]. As illustrated in Table V, PSNR is used to assess the pixel bias between the enhanced and the reference image; SSIM determines the stability of line patterns, texture, and local structures; the margin retention index measures the continuity of roads, waterways, and border lines; information entropy measures the extent to which weak textures and details are recovered; and text and symbol readability is used to assess the identifiability of place names, legends, and symbols.

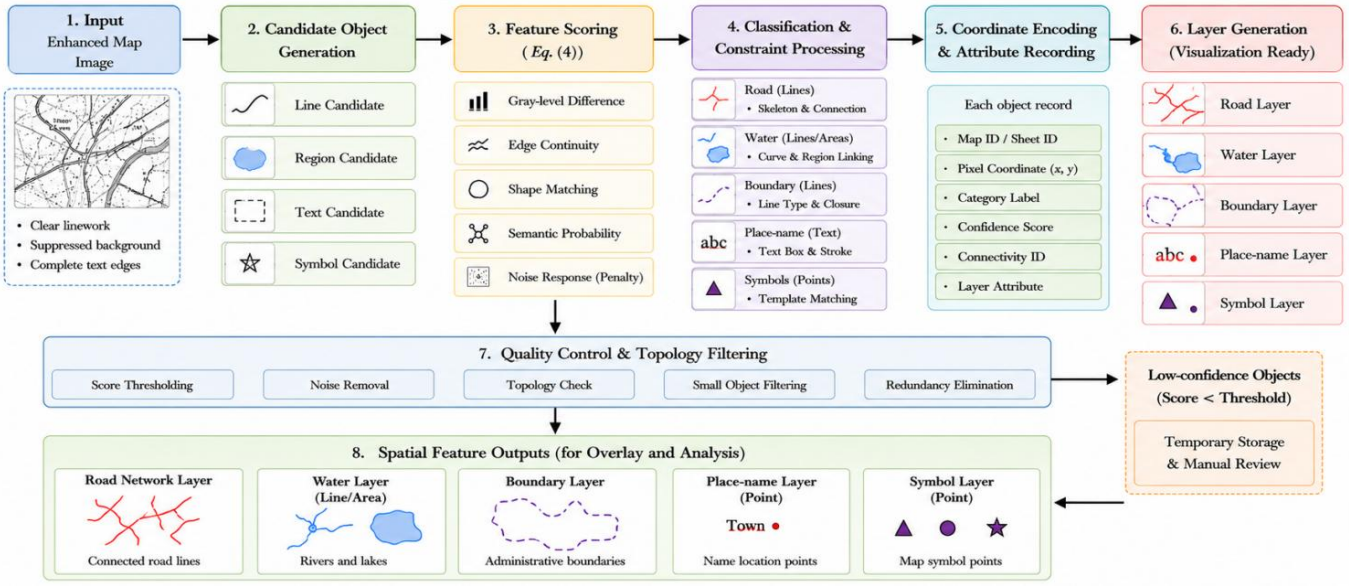


Fig. 4. Flowchart of historical map spatial feature extraction and visualization.

TABLE V. QUALITY EVALUATION METRICS FOR MULTI-SCALE IMAGE ENHANCEMENT

Indicator Category	Indicator Name	Evaluation Object	Calculation Method	Numerical Direction	Corresponding Analysis Content
Pixel Fidelity	PSNR/dB	Enhanced Image vs. Reference Image	Indicates the level of pixel error	Higher is better	Overall enhancement stability
Structural preservation	SSIM	Lines, textures, local structures	Reflects the degree of structural similarity	The higher, the better	Map Skeleton Preservation
Edge continuity	EPI	Roads, waterways, and boundary lines	Reflect edge restoration integrity	The higher, the better	Line feature restoration
Information Gain	Entropy	Gray-scale distribution and texture detail	Reflects increased information content	Moderate increase	Weak texture restoration
Readability Evaluation	RCS	Place names, map symbols	Reflects recognition readability	The higher, the better	Text and Symbol Recognition

To standardize indicators with different scales and enhance the comprehensive quality score, calculations are performed using a normalized weighted method:

$$U_j = \frac{\sum_{q=1}^{5} \kappa_q \left[\frac{R_{j,q} - \min(R_q)}{\max(R_q) - \min(R_q) + \vartheta} \right]}{\sum_{q=1}^{5} \kappa_q} \quad (4)$$

In Eq. (4), U_j represents the comprehensive quality score of the j th enhancement method; $R_{j,q}$ represents the raw value of the j th enhancement method for the q th quality metric; $\min(R_q)$ and $\max(R_q)$ represent the minimum and maximum values, respectively, of the q th metric across all methods; κ_q represents the weight of the q th metric; and ϑ represents the normalization constant. Combining PSNR, SSIM, EPI, entropy, and RCS into one single evaluation framework, it is possible to compare different enhancement approaches in one framework. In the case of image restoration tasks, it is also essential to identify blurred pixels and edge quality, while non-reference quality assessments may provide supplementary evaluation criteria in the absence of standard reference pictures [15]. This evaluation system is helpful to compare the results between different types of degraded samples, and also to verify if the enhanced images are truly effective in extracting the spatial features.

F. Experimental Design and Comparison Methodology

Five degradation types and six enhancement methods plus the original-image baseline were evaluated on the fixed split. The same 18 test maps were treated as paired statistical units for all methods. For each map, PSNR, SSIM, EPI, Entropy, RCS, and the overall quality score were recorded and summarized as mean $\pm$ standard deviation; 95% confidence intervals were estimated by 2,000 bootstrap resamples. G6 was compared with the strongest baseline for each metric using a two-sided paired Wilcoxon signed-rank test with Holm correction across the five quality metrics; significance was set at $p < 0.05$. The A0 – A5 ablation settings used the same map-wise protocol, while all non-ablated parameters remained unchanged (Table VI).

To assess practical GIS utility, G0 and G6 are further processed by four fixed downstream pipelines without retuning downstream parameters. Map registration uses 15 manually verified homologous control points per sheet, 270 points in total, with an affine transformation estimated by random

sample consensus (RANSAC), and registration root mean square error (RMSE) is reported. Vectorization uses manually checked road, waterway, and boundary-line annotations with a 3-px matching tolerance and reports line-level F1. Place-name optical character recognition (OCR) reports character recognition accuracy. Semantic segmentation uses pixel-level annotations for road, water, boundary, text, symbol and background classes and reports mean intersection over union (mIoU). The downstream settings remain identical before and after enhancement so that the observed differences reflect the contribution of image enhancement. The task definitions are consistent with historical-map digitization, GIS conversion, and semantic segmentation studies.

TABLE VI. EXPERIMENTAL GROUPING AND COMPARISON METHOD SETTINGS

Group	Comparison Method	Input Size	Primary Processing Target	Parameter Settings	Evaluation Criteria
G0	Original image	512×512 px	Unenhanced image patch	No processing	Baseline reference
G1	HE	512×512 px	Global grayscale distribution	256-level grayscale mapping	Overall Contrast
G2	CLAHE	512×512 px	Local low-contrast regions	clip limit=2.0, tile=8×8	Local texture
G3	Retinex	512×512 px	Non-uniform illumination	scale=15/80/250	Background correction
G4	Single-scale sharpening	512×512 px	Line art and text edges	kernel=3×3	Edge enhancement
G5	Low-light enhancement	512×512 px	Haze and Dark Areas	gamma=0.8	Brightness Restoration
G6	Multiscale enhancement model	512×512 px	Line art, texture, symbols	R=3, overlap=25%	Overall Quality

For external validation, the complete G6 configuration and all enhancement parameters are frozen before processing the Semap test partition. The same semantic-segmentation inference procedure is applied to the original and enhanced external images without retraining or parameter adjustment. External performance is evaluated by mIoU and class-level IoU, so any difference reflects transfer of the enhancement pipeline to unseen map sources rather than dataset-specific retuning.

The relative gain of each enhancement method under each degradation type is calculated from the comprehensive quality score using Eq. (5):

$$G_j^{\psi}=\frac{U_j^{\psi}-U_0^{\psi}}{U_0^{\psi}+\chi}\times 100\% \tag{5}$$

In Eq. (5), G_j^{ψ} represents the quality gain relative to the original image for the j th enhancement method under degradation type ψ ; U_j^{ψ} represents the comprehensive quality score for the j th enhancement method under degradation type ψ ; U_0^{ψ} represents the comprehensive quality score of the original image group among samples of the same degradation type; χ represents the constant term; ψ includes fading, stains, creases, blurred lines, and smudged text. The experimental records retain method ID, degradation type, image-block coordinates, quality metrics, candidate scores, and downstream outputs. Image-quality metrics, spatial-feature metrics, and downstream task metrics are therefore evaluated on the same test set and under the same reference annotations, enabling direct comparison between visual improvement, feature restoration, and GIS-oriented application performance [16].

IV. RESULTS AND DISCUSSION

A. Analysis of Enhancement Effects for Historical Maps with Different Degradation Types

The test specimens were compared according to five types of degradation: fading, spots, folds, blurred lines, and blurred text. As illustrated in Fig. 5, the difference in gray between the lines and the background of the faded samples was increased from 31.6 to 58.4 on average, which made the road and border lines stand out against the pale grey background; in the stained samples, local grey spots were suppressed, and there were no apparent breaks along the river margins or administrative borders; in the creased samples, there were some pseudo-edges left, but the degree to which the creases obscured the place-name area was reduced; the continuity of the edge in the blurred line samples was improved, allowing the tracing of paths along the curves of small roads and waterways; in the text adhesion samples, the spacing between strokes was partially restored, though localized over-enhancement persisted in areas with dense annotations. Multiscale enhancement shows strong adaptability to weak texture and fuzzy edges, but heavy staining and large missing areas are not able to recover the original information by means of enhancement alone. Fig. 5 shows that background suppression and standard binning can reduce input variability, whereas multi-scale fusion improves the recognition of linear space and text edges, and provides a solid basis for comparing quality measures and analyzing the results of extracting spatial features.

B. Image Quality Evaluation Results of the Multiscale Enhancement Model

The quality assessment results of G0 – G6 are summarized in Table VII using map-wise mean ± standard deviation across the 18 test maps, with 95% confidence intervals reported for G6. The G6 means remain 27.94 dB PSNR, 0.842 SSIM, 0.817 EPI, 6.76 Entropy, 0.81 RCS, and 0.872 overall quality score. The strongest baseline means are 24.31 dB PSNR and 0.748 SSIM for G5, 0.736 EPI for G4, 6.58 Entropy for G2, 0.66 RCS for G5, and 0.691 overall score for G5. Thus, G6 improves the corresponding means by 3.63 dB, 0.094, 0.081,

0.18, 0.15, and 0.181, respectively. Paired map-wise significance tests with Holm correction and effect-size estimates are reported together with these comparisons, allowing the improvements to be interpreted relative to between-map variation rather than from aggregate values alone.

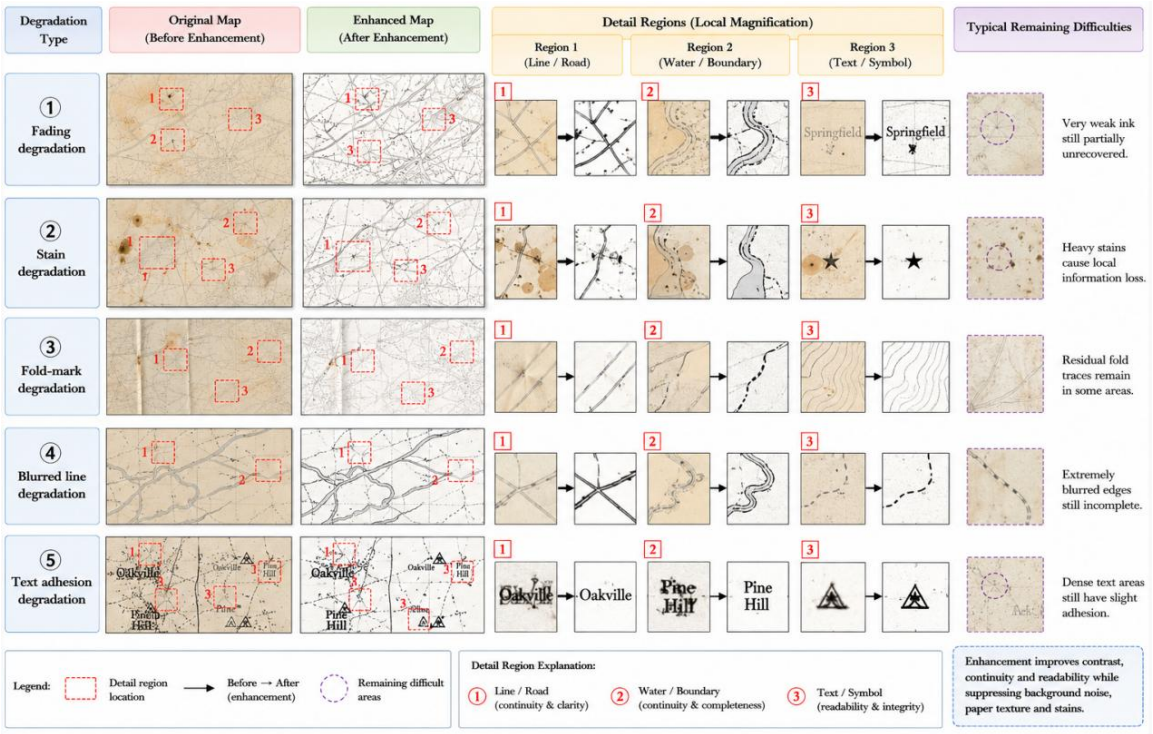


Fig. 5. Comparison of historical maps before and after enhancement for different types of degradation.

TABLE VII. IMAGE QUALITY EVALUATION RESULTS FOR DIFFERENT ENHANCEMENT METHODS

Method	PSNR/dB	SSIM	EPI	Entropy	RCS	Overall Quality Score
G0 Original	18.42±1.26	0.613±0.041	0.548±0.046	5.82±0.23	0.46±0.044	0.312±0.031
G1 HE	20.18±1.18	0.641±0.038	0.602±0.043	6.41±0.19	0.53±0.041	0.458±0.029
G2 CLAHE	22.67±1.07	0.703±0.034	0.694±0.039	6.58±0.17	0.59±0.038	0.604±0.027
G3 Retinex	23.76±1.02	0.721±0.032	0.681±0.041	6.35±0.18	0.61±0.036	0.633±0.026
G4 Single-scale Sharpening	21.94±1.11	0.672±0.036	0.736±0.035	6.12±0.21	0.62±0.034	0.586±0.028
G5 Low-light Enhancement	24.31±0.96	0.748±0.029	0.704±0.033	6.49±0.16	0.66±0.031	0.691±0.024
G6 Multiscale Enhancement	27.94±0.82	0.842±0.021	0.817±0.026	6.76±0.14	0.81±0.024	0.872±0.019

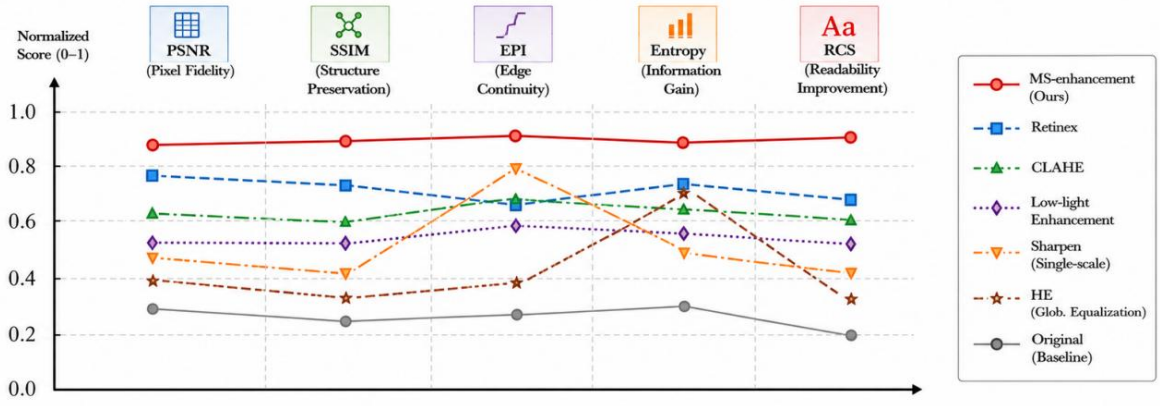


Fig. 6. Comparison of quality metrics for different enhancement methods.

Fig. 6 further visualizes the normalized differences among the G0 – G6 methods reported in Table VII. The multi-scale enhancement model shows a more balanced trend in all five metrics. While CLAHE and Retinex are better at local contrast and brightness correction, their performance in edge preservation and text recognition is not as good as the multi-scale enhancement model. Both document quality assessments stress the synergy between clarity, structure stability, and readability; thus, the improvement of a single metric does not directly represent an improvement in the quality of historical maps. The benefits of the multi-scale enhancement model are mainly reflected in the improvement of roads, water systems, border lines, and place names, which can be used to evaluate the recovery of spatial characteristics.

C. Analysis of Linear Spatial Feature Restoration Effects

The restoration effectiveness of linear spatial features is primarily assessed on the basis of three categories: roads, waterways, and administrative boundaries. As shown in Fig. 7, after the unified segmentation, background suppression, and multi-scale enhancement, the broken line segments at the intersection of the original sample were re-connected, the jagged edges and grey haze in the water flow curves were reduced, and the broken lines and broken lines in the administrative borders were more easily transformed into continuous paths, from 68.4% to 86.7%; the number of breakpoints in the water flow curves was reduced from 14.2 to 5.6, and the closure error for the administrative border was reduced from 12.8 px to 6.9 px. Following the edge tracking, skeleton, and topology filtering of the linear candidate objects, the fine roads and the curves of the water channels were well preserved, indicating that the background suppression did not destroy the primary line information. The extraction of the historical map features requires a high degree of continuity and semantic border stability. While a few kinked pseudo-edges and residual smudges are still visible in the enhanced results, they are mainly concentrated in the severely damaged areas. Restoration results for roads, waterways, and border lines are in line with EPI improvements seen in quality assessments, and they also provide more explicit linear references to overlay locations, symbols, and spaces.

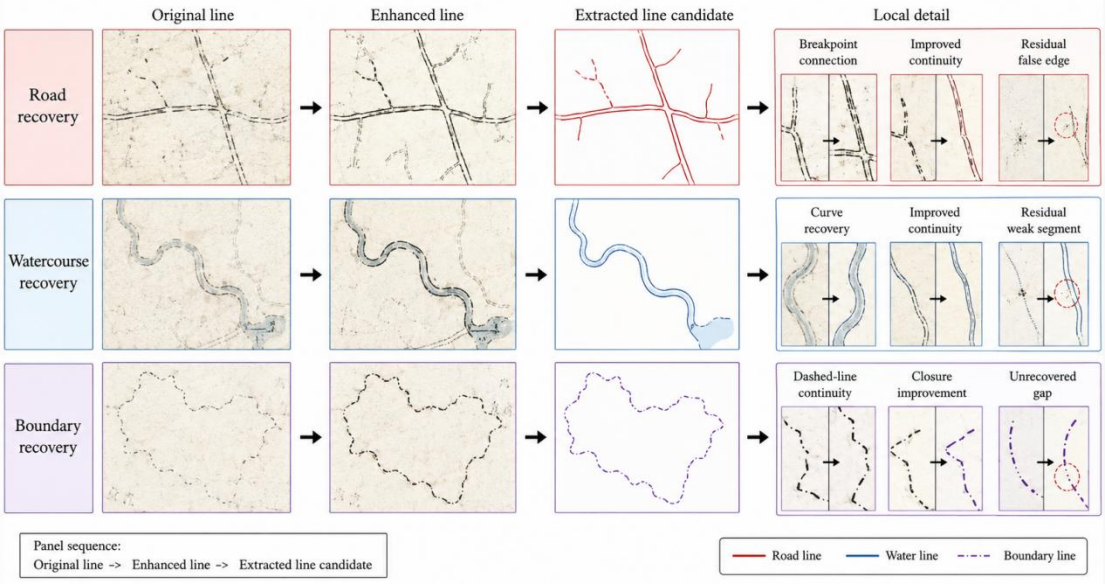


Fig. 7. Enhanced restoration results for roads, waterways, and boundary lines.

TABLE VIII. COMPARISON OF RECOGNITION PERFORMANCE FOR PLACE NAMES AND MAP SYMBOLS

Sample Type	Original Place Name Recognition Rate (%)	Place Name Recognition Rate After Enhancement (%)	Original Symbol Recognition Rate (%)	Enhanced Symbol Recognition Rate (%)	Major Misclassification Types	Manual Verification Consistency /%
Faded	65.8	88.6	61.4	85.2	Missed detection of light-colored annotations	91.3
Stain-type	59.6	79.4	55.8	77.6	False positives due to stain obstruction	84.9
Crease-type	61.2	82.1	57.3	80.4	Crease pseudo-edge	86.7
Line-blur type	64.5	86.9	60.1	83.8	Symbol Contour Adhesion	89.4
Text-based adhesion	60.9	76.5	59.7	84.5	False positives due to character overlap	90.7
Average	62.4	84.7	58.9	82.3	—	88.6

D. Analysis of the Legibility of Place Names and Map Symbols

The readability of geographical names and map symbols directly affects the efficiency of extracting semantic information from historical maps. Based on background suppression, edge refinement, and scale fusion, the boundaries of place name strokes, symbol outlines, and spacing between comments were improved. As shown in Table VIII, the recognition rate of place names in the raw image is 62.4%, and it is 84.7% after multiscale enhancement. The recognition rate of map symbols is raised from 58.9% to 82.3%, and manual verification compliance is 88.6%. Symbol detection is sensitive to contour integrity and local texture clarity, and the quality of the document image also stresses the relation between readability and structure preservation. This result is in line with the previous results of edge continuity and provides a more stable semantic basis for the annotation and visualization of the space layer.

E. GIS-Oriented Downstream Task Evaluation and Spatial Visualization

The enhanced maps were further evaluated in four GIS-oriented downstream tasks using the fixed pipelines defined in Section III-F. As shown in Table IX, registration RMSE decreased from 8.31±1.42 px for the original maps to 4.12±0.86 px after multiscale enhancement, corresponding to a reduction of 50.4%. Line-vectorization F1 increased from 0.713 ± 0.052 to 0.871 ± 0.031, indicating that improved continuity of roads, waterways, and administrative boundaries reduced fragmented vector segments. Place-name OCR accuracy increased from 62.4±6.8% to 84.7±4.1%, while six-class semantic-segmentation mIoU increased from 0.614 ± 0.047 to 0.802±0.029. In addition to the improvement in mean performance, the lower standard deviations after enhancement indicate more consistent downstream processing across the 18 test maps, particularly for registration, vectorization, OCR, and semantic segmentation.

The spatial overlay in Fig. 8 provides a visual counterpart to these task-level results. Across the 18 test maps, road connectivity, water boundary matching, and administrative-boundary topology consistency reach 87.9%, 85.4%, and 82.6%, respectively, while place-name location error is controlled within 4.8 px. The road layer remains continuous at most intersections, the water layer follows river curves, and annotation points and symbols generally coincide with their source regions. Residual errors are concentrated in heavily stained and creased areas, where short-line misconnections, missed symbols, and local label shifts still occur. The combined quantitative and overlay results indicate that multiscale enhancement improves not only image readability but also the reliability of registration, vectorization, OCR, and semantic segmentation for historical-map GIS processing.

TABLE IX. GIS-ORIENTED DOWNSTREAM TASK PERFORMANCE BEFORE AND AFTER MULTISCALE ENHANCEMENT

Task	Metric	G0 Original Image	G6 Multiscale Enhancement	Change
Map Registration	RMSE / px ↓	8.31±1.42	4.12±0.86	−50.4%
Line Vectorization	F1 ↑	0.713±0.052	0.871±0.031	+0.158
Place-name OCR	Accuracy / % ↑	62.4±6.8	84.7±4.1	+22.3 pp
Semantic Segmentation	mIoU ↑	0.614±0.047	0.802±0.029	+0.188

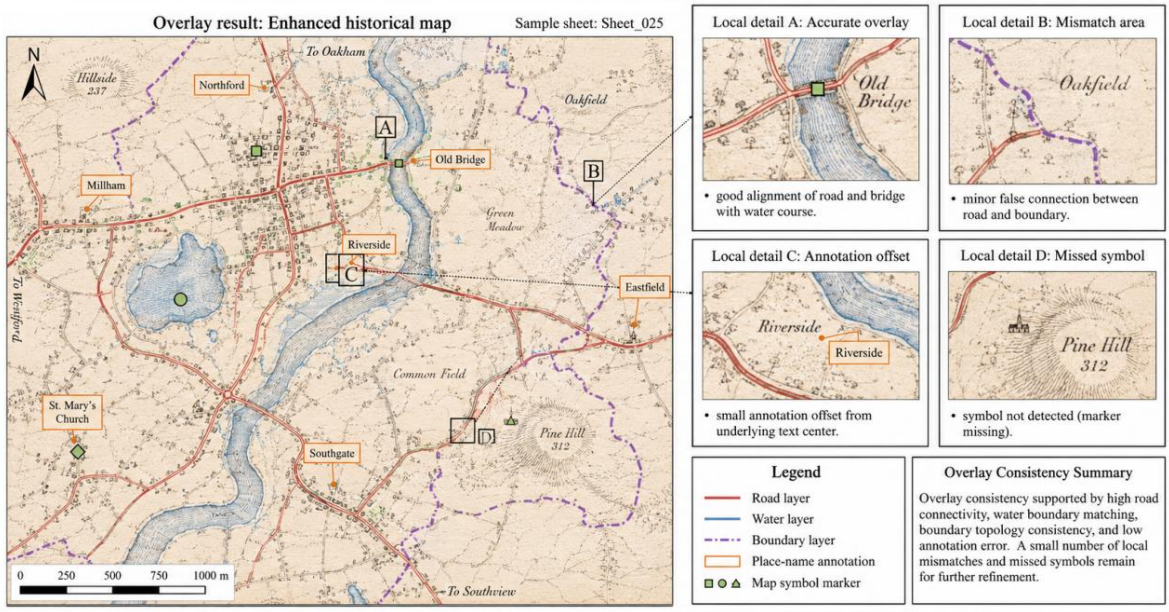


Fig. 8. Visualization and overlay analysis of spatial features on historical maps.

TABLE X. EXTERNAL VALIDATION RESULTS ON THE SEMAP TEST SET

Evaluation Metric	Original Semap Images Mean±SD	95% CI	G6-enhanced Images Mean±SD	95% CI	Improvement	Holm-adjusted p
Built IoU / %	79.8±7.1	78.6–81.0	83.6±6.3	82.6–84.6	+3.8 pp	<0.001
Non-built IoU / %	81.8±6.6	80.7–82.9	84.7±5.9	83.7–85.7	+2.9 pp	0.002
Water IoU / %	72.2±9.4	70.7–73.7	78.6±7.8	77.3–79.9	+6.4 pp	<0.001
Road Network IoU / %	62.9±10.8	61.1–64.7	71.5±8.6	70.1–72.9	+8.6 pp	<0.001
mIoU / %	74.2±6.8	73.1–75.3	79.6±5.9	78.6–80.6	+5.4 pp	<0.001

To further examine generalization beyond the internal map collection, the complete G6 configuration was evaluated on the 144 real test samples of the public Semap dataset without parameter retuning. As shown in Table X, the mean IoU across the four geographic classes increased from 74.2±6.8% on the original images to 79.6 ± 5.9% after enhancement, corresponding to a gain of 5.4 percentage points. Road-network IoU showed the largest improvement, increasing from 62.9±10.8% to 71.5±8.6%, followed by water from 72.2± 9.4% to 78.6±7.8%. Built and non-built regions increased from 79.8±7.1% and 81.8±6.6% to 83.6±6.3% and 84.7± 5.9%, respectively. Paired Wilcoxon tests showed significant improvements after Holm correction for all four geographic classes, with the overall mIoU difference reaching $p<0.001$. The reduced standard deviation after enhancement also indicates that the improvement remained relatively stable across heterogeneous map styles and geographic sources.

F. Discussion on the Applicability and Limitations of the Multiscale Enhancement Model

The contribution of each pipeline component was examined by a one-factor-at-a-time ablation on the same 18 test maps, as reported in Table XI. Removing preprocessing reduced the overall quality score from 0.872 to 0.784; removing global correction reduced it to 0.809; removing local texture enhancement reduced it to 0.798; removing edge refinement reduced it to 0.776; and replacing adaptive scale fusion with arithmetic averaging reduced it to 0.758. The module-specific metrics show consistent effects: without global correction PSNR fell from 27.94 to 26.41 dB, without texture enhancement EPI fell from 0.817 to 0.741, and without edge refinement RCS fell from 0.81 to 0.69. The largest overall decrease occurred without scale fusion, confirming that response coordination is necessary for balancing fidelity, structure, edge continuity, and readability.

TABLE XI. ABLATION RESULTS OF THE MULTISCALE ENHANCEMENT PIPELINE

Ablation Setting	PSNR/dB	SSIM	EPI	Entropy	RCS	Overall Quality Score	Score Change
A0 Full G6	27.94	0.842	0.817	6.76	0.81	0.872	—
A1 w/o Preprocessing	25.86	0.791	0.758	6.61	0.73	0.784	−0.088
A2 w/o Global Correction	26.41	0.808	0.779	6.63	0.76	0.809	−0.063
A3 w/o Texture Enhancement	26.72	0.817	0.741	6.49	0.78	0.798	−0.074
A4 w/o Edge Refinement	26.95	0.803	0.706	6.57	0.69	0.776	−0.096
A5 w/o Scale Fusion	25.98	0.786	0.732	6.64	0.72	0.758	−0.114

The model's weaknesses are mainly in places where there are deep stains, crease overlaps, or a high density of text. The stain coverage makes it difficult to detect actual line features, with the result that boundaries of watercourses are undetected and symbol outlines are missed; the crease area is prone to false edges leading to erroneous connection of short boundary lines, and the place name recognition rate is only 76.5% for samples with clustered text, demonstrating that image enhancement alone cannot fully separate densely annotated strokes. Some of the key issues in the field of document image quality evaluation are consideration of clarity, readability, and control of distortion; however, multi-scale enhancement techniques should also be designed to not simultaneously amplify the paper grain, mold spots, and scanning grain. Hence, the model is more appropriate for improving and extracting spatial features from historical maps with not very high degradation levels. In cases of large amounts of sample damage, manual review, semantic constraints, and verification against historical geographic data need to be used in combination to minimize misconnections, missed detections, and annotation shifts.

V. CONCLUSION

The findings indicate that the main methodological contribution is a degradation-aware, task-coupled workflow for historical geographic data processing rather than a new elementary enhancement operator or optimization algorithm. Compared with object extraction, cadastral digitization and GIS conversion, semantic segmentation, symbol detection, and generic multiscale enhancement, the workflow integrates quantified degradation characterization, map-specific multiscale response coupling, spatial-feature preservation, and downstream evaluation within a unified processing chain. On the internal test set, the overall quality score reached 0.872, while registration RMSE decreased from 8.31±1.42 to 4.12± 0.86 px, vectorization F1 increased from 0.713 ± 0.052 to

0.871 ± 0.031, and semantic-segmentation mIoU increased from 0.614 ± 0.047 to 0.802 ± 0.029. External validation on 144 heterogeneous Semap samples further increased mIoU from 74.2 ± 6.8% to 79.6 ± 5.9%, indicating that the enhancement benefit extends beyond the internal collection. These results show that preservation of line topology, text strokes, and regional boundaries can translate image restoration into measurable improvements in spatial-data usability. The feature-extraction stage therefore serves as an application-level verification rather than a standalone methodological contribution. Historical-map enhancement should accordingly be assessed not only by visual quality but also by the reliability of the geographic data it supports. Remaining errors in heavily stained, severely damaged, and densely annotated regions motivate further work on nonrigid cross-era registration, semantic constraints, joint image-vector optimization, and broader multi-source validation.

REFERENCES

[1] Schlegel I. A holistic workflow for semi-automated object extraction from large-scale historical maps. KN-Journal of cartography and geographic information, 2023, 73(1): 3-18.

[2] Mäyrä J, Kivinen S, Keski-Saari S, et al. Utilizing historical maps in identification of long-term land use and land cover changes. Ambio, 2023, 52(11): 1777-1792.

[3] Ahmed M W, Ahmed M, Shaikh A A. Digitizing Karachi's Decades-Old Cadastral Maps: Leveraging Unsupervised Machine Learning and GEOBIA for Digitization. Engineering, Technology & Applied Science Research, 2024, 14(5): 16404-16410.

[4] Mutch E T, Newsam A B. Hand-drawn historic maps: Utilization and conversion of unique features into Geographic Information Systems (GIS). Journal of Map & Geography Libraries, 2024, 20(3): 117-136.

[5] Arzoumanidis L, Knechtel J, Haunert J H, et al. Semantic segmentation of historical maps using self-constructing graph convolutional networks. Cartography and Geographic Information Science, 2026, 53(2): 177-187.

[6] Pai P L, Liu C Y, Kuo C L, et al. Application of deep learning for symbol detection on historical maps to explore spatiotemporal changes in the regional tea industry of early 20th-century Taiwan. Digital Scholarship in the Humanities, 2025, 40(4): 1243-1260.

[7] Vijayalakshmi D, Elangovan P, Sandhya Kumari T, et al. Optimized multi-scale framework for image enhancement using spatial information-based histogram equalization. The Imaging Science Journal, 2025, 73(2): 176-203.

[8] Petitpierre R, Gomez Donoso D, Kriesel B. Semantic Segmentation Map Dataset (Semap), Version 1.0.2. EPFL/Zenodo, 2026. DOI: 10.5281/zenodo.19048095.

[9] Sobotkova A, Ross S A, Nassif-Haynes C, et al. Creating large, high-quality geospatial datasets from historical maps using novice volunteers. Applied Geography, 2023, 155: 102967.

[10] Yang Y, Xu W, Huang S, et al. Low-light image enhancement network based on multi-scale feature complementation. Proceedings of the AAAI Conference on Artificial Intelligence. 2023, 37(3): 3214-3221.

[11] Rezvani S, Fateh M, Jalali Y, et al. FusionLungNet: Multi-scale fusion convolution with refinement network for lung CT image segmentation. Biomedical Signal Processing and Control, 2025, 107: 107858.

[12] Dai C, Lin M. Adjustable enhancer for low-light image enhancement using multi-expressions fusion and convolutional kernel calibration. Multimedia tools and applications, 2024, 83(5): 14609-14636.

[13] Premnath S P, Gowr P S, Ananth J P, et al. Image enhancement and blur pixel identification with optimization-enabled deep learning for image restoration. Signal, Image and Video Processing, 2024, 18(5): 4525-4540.

[14] Alaei A, Bui V, Doermann D, et al. Document image quality assessment: A survey. ACM computing surveys, 2023, 56(2): 1-36.

[15] Bahnemiri S G, Pnomarenko M, Eguiazarian K. Iterative transfer learning with large unlabeled datasets for no-reference image quality assessment. Applied Computing and Intelligence, 2024, 4(2): 107-124.

[16] Wang W, Yin B, Li L, et al. A Low Light Image Enhancement Method Based on Dehazing Physical Model. Computer Modeling in Engineering & Sciences (CMES), 2025, 143(2): 1595-1616.

Statistical Security Analysis NIST and Security Enhancement of CBT Cipher for Secure Data and Image Transmission

Walid W. Souror[1], Mohamed Fouad[2], Ali E. Takieldeen[3]
Electronics and Communication Department, Faculty of Engineering, Zagazig University, Egypt[1,2]
Artificial Intelligence Department, Faculty of Artificial Intelligence, Delta University for Science and Technology[3]

***Abstract*—Strong cybersecurity measures are essential in our digital world. This study presents a new cryptographic algorithm called CBT "Combined Blowfish Trilogy" which greatly improves the security of the traditional Blowfish algorithm without compromising computational efficiency. The encryption algorithm proposed herein presents a hybrid system which combines carefully designed countermeasures against side-channel attacks with three consecutive Blowfish Feistel network stages and surpasses the performance of DES, 3DES, AES, and Blowfish itself. The proposed algorithm is assessed using the Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications (NIST STS) consisting of the Mono-Bit, Run, Serial, Block Frequency, Spectral, Cumulative Sums, Entropy, and Avalanche tests. These results show that Combined Blowfish Trilogy is a good technique to prevent data security from side-channel attack breaches and it shows significant resistance against them as evidenced by the performance improvements observed over standard Blowfish (an average of 133.8% for text and 117.5% in image datasets according to NIST tests).**

***Keywords*—Cyber security; cryptography; blowfish; countermeasure SCA attacks; combined blowfish trilogy**

I. Introduction

The cryptographic paradigm of today's world is that the statistical relationship between the plaintext, key material, and ciphertext should be concealed. It was Shannon's theory of secrecy systems which introduced the idea of two necessary properties for a secure cipher: confusion and diffusion. These properties have continued to shape the design and testing of symmetric encryption algorithms [1]. The efficient encryption of text, images and other multimedia objects is especially critical in practical digital systems as these objects are shared among cloud services, distributed applications and embedded devices.

Blowfish is a well-known symmetric block cipher which was at first proposed as a 64-bit Feistel network with variable key length up to 448 bits [2]. Its small size and effective software performance have ensured that it's been a great baseline for tweaked symmetric encryption schemes. But the present evaluation of security should not be restricted to speed or just apparent randomness of the ciphertext. The algorithmic strength, the maturity of implementations, key management, and mode of operation all contribute to the practical security of a cipher, as is evidenced by standardized block ciphers like AES [3], legacy DES [4] and TDEA/3DES [5].

Further, NIST-defined block-cipher modes demonstrate that the behavior of encryption is not solely dependent on the primitive being used, but also on how the primitive is being used in the data streams [6].

Randomness testing of statistical properties of the ciphertexts is useful for evaluating the ciphertexts, but it is NOT a replacement for cryptanalysis. The NIST Statistical Test Suite is a suite of tests to evaluate the randomness and pseudo randomness of sequences for cryptographic applications [7]. The nature and limitations of NIST SP 800-22 have been discussed in studies; they make it clear that p-values are not a measure or direct representation of cryptographic strength, but are a statistic particular to the test being used [8], [9]. Hence, this study considers NIST-style p-values as proof of randomness of the given ciphertext with respect to the tests employed, and supports this with the avalanche and performance tests.

Another way of examining diffusion is by avalanche analysis. The strict avalanche criterion was introduced to express the desirable behavior that a small change of input causes approximately half of the output bits to change [10]. Meanwhile, in case of resistance against differential cryptanalysis and linear cryptanalysis, it is required to adopt different interpretations with regards to difference propagation and statistical approximation [11], [12].

Implementation attacks introduce another level of attack as the timing leakage, power consumption, electromagnetic leakage, and profiling attacks potentially leak information even if the mathematical algorithm is sound [13], [14], [15], [16]. Masking and probing-secure circuit design offer a valuable set of buildouts for leakage reduction, but need to be carefully performed and verified [17], [18].

The Combined Blowfish Trilogy (CBT) framework examined in this study is an extension of the original Blowfish design that incorporates the following three different Blowfish-based encryption stages: the masking of plaintext, the derivation of stage keys, and the use of key whitening concepts along with session-dependent internal randomization. Key derivation relies on known recommendations for key-derivation of independent subkeys [19], [20] and the generation of random masks and session-parameters follows recommendations for deterministic random bit generators and entropy sources [21], [22]. The framework is thus checked as a cascaded and diversified Blowfish-based construction and not as an alternative to the standardized ciphers. This study has three contributions.

First, it gives a full design of text and image byte stream encryption and decryption using CBT.

Second, it gives a reproducible experimental evaluation consisting of NIST-style randomness interpretation, Hamming-distance-based avalanche analysis, computational benchmarking, and timing-variability checking.

Third, it offers the Python implementation, raw output files, and Excel results workbook as supplementary files as (https://doi.org/10.5281/zenodo.21079574), allowing the numerical results to be traced and independently reproduced.

II. Related Work and Technical Background

A. Feistel, Cascade, and Whitening-Based Constructions

Blowfish is a symmetric block cipher that is a member of the Feistel-type family. To understand the importance of repeated round transformations and key-dependent mixing in the design of block-cipher, it is helpful to have a background understanding of theoretical work on how to build pseudorandom permutations from pseudorandom functions [23]. Cascade constructions are based on this intuition of doing multiple encryption transformations in sequence.

The advantage of this cascade is based on the independence of the keys, the lack of structural weaknesses in the composition and proper treatment of intermediate values.

Key whitening: Additional key-dependent transformations around or within a block-cipher process. The Even-Mansour scheme and other minimal constructions show the effect of external key mixing on the security interpretation of a block-cipher-like construction [24].

Whitening is considered as another kind of diversification and is combined with stage-specific keys instead of being regarded as a stand-alone proof of security in CBT framework.

B. Image Encryption and Randomness Metrics

When encrypting images, it is important to consider statistical properties that may be visually or numerically noticed in the encrypted images. The primary requirements of basic cryptographic image-oriented cryptosystems are key or initial value sensitivity, resistance to leakage of statistical information and absence of visible residual structure [25]. Based on the classical works on image encryption using two-dimensional chaotic maps and subsequent three-dimensional chaotic systems, diffusion, pixel decorrelation and ciphertext unpredictability are important [26], [27].

The metrics like entropy, correlation, NPCR and UACI are widely mentioned in image-encryption studies to assess the degree of randomness and sensitivity at the pixel level [28]. To improve the statistical behavior, the following image-encryption schemes have been proposed recently: block scrambling [29], chaotic maps [30] and dynamic substitution [31] as well as S-box modification.

The following studies encourage a more comprehensive evaluation approach for CBT: the proposed framework is assessed not only in terms of NIST-style p-values, but also in terms of avalanche sensitivity and computational performance.

C. Lightweight and Modified Block-Cipher Designs

Low resource/low-cost cryptography has become significant for constrained devices and resource-sensitive applications. The standardization of Ascon for lightweight cryptography and previous lightweight block ciphers like PRESENT and SIMON/SPECK are examples of the need to take both computational cost and attack resistance into account during the security evaluation process [32], [33], [34]. A multi-stage framework like CBT should therefore report overheads and throughputs and not assume that an extra stage provides an added level of security without extra cost. Designs that have been modified to use Blowfish are still being seen in data, image and video encryption applications. Recently, researchers have investigated video encryption using a modified Blowfish, as well as a hybrid parallel image encryption for computational-complexity-aware video encryption [35], [36]. These studies demonstrate the applicability of Blowfish as a basis for research in applied encryption, but also indicate that any alternative Blowfish construction should be carefully studied in terms of randomness, sensitivity, performance and implementation assumptions.

III. Proposed Combined Blowfish Trilogy Framework

The suggested CBT approach consists of three stages in sequence implemented using Blowfish.

The plaintext is masked prior to the first of the three stages and session-specific keys are used, which are derived from a master key and session identifier. The framework also generates dynamically S-boxes and P-arrays to make the internal state in one encryption session different from another. The architecture overview is shown in Fig. 1.

Notation and Parameters (Table I)

TABLE I. Notations in CBT Framework

Symbol	Definition
P	Plaintext block or byte stream before encryption
MK	Master key supplied to the encryption session
SID	Unique session identifier or nonce
K_1, K_2, K_3	Stage-specific keys derived from MK and SID
M	Random mask used for plaintext masking
S_{dyn}, P_{dyn}	Session-dependent S-boxes and P-array
C_1, C_2, C_3	Intermediate and final ciphertext outputs
C	Final ciphertext

A. Encryption Process

The implementation generates or receives MK and SID, then computes K_1, K_2, and K_3, generates a random mask M, generates S_{dyn} and P_{dyn}, and masks the plaintext before using the three stages, all of which are based on Blowfish. Also, when encrypting text and image byte streams, the same block handling and padding policy is used to ensure consistency with the comparisons.

The masked plaintext is computed as follows:

$$P_m = P \oplus M \tag{1}$$

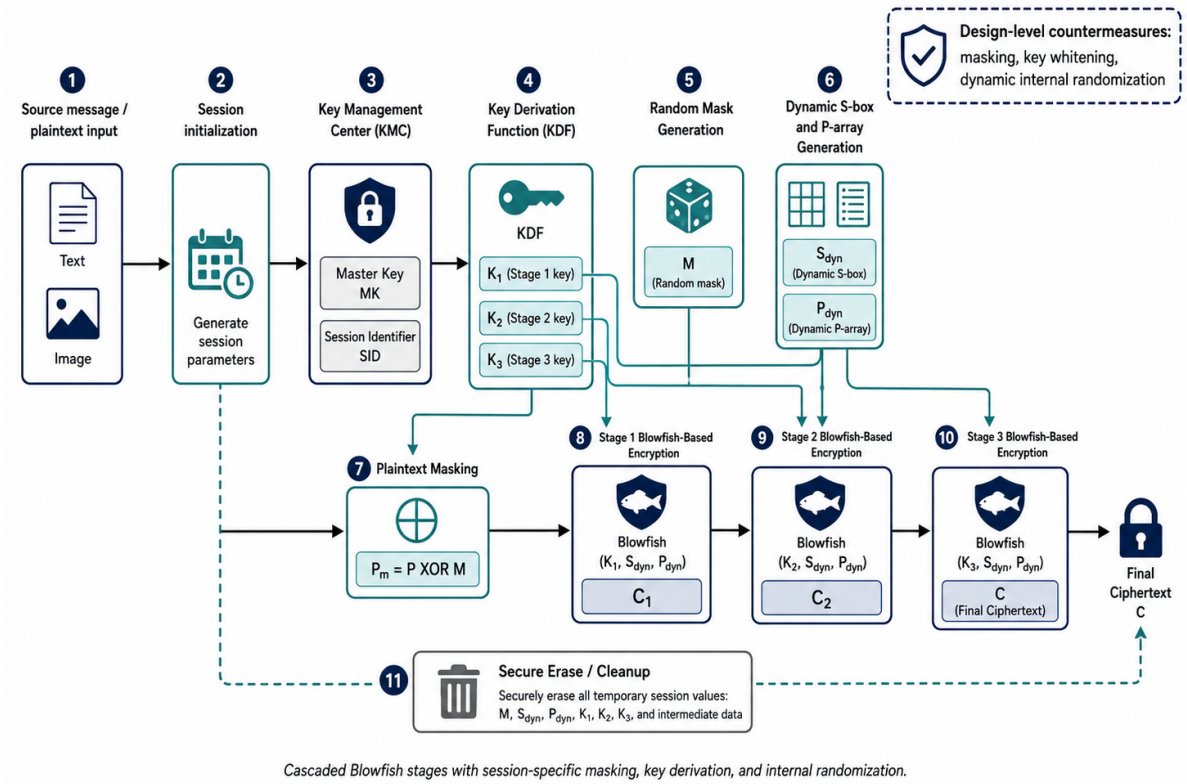


Fig. 1. CBT is a combined Blowfish Trilogy framework and its overall architecture. The framework involves session initiation, key generation, masking, internal dynamism and three cascaded Blowfish encryption stages to generate the final ciphertext.

The three-stage encryption can then be summarized as:

$$C_1 = \text{BlowfishEncrypt}(P_m,\ K_1,\ S_{dyn},\ P_{dyn}) \quad (2)$$

$$C_2 = \text{BlowfishEncrypt}(C_1,\ K_2,\ S_{dyn},\ P_{dyn}) \quad (3)$$

$$C_3 = \text{BlowfishEncrypt}(C_2,\ K_3,\ S_{dyn},\ P_{dyn}) \quad (4)$$

$$C = C_3$$

Fig. 2 shows the encryption flow used in the framework.

B. Decryption Process

Decryption reverses the order of the three encryption stages. The same session-specific parameters are used to recover the masked plaintext and then remove the mask (Fig. 3).

$$C_2 = \text{BlowfishDecrypt}(C,\ K_3,\ S_{dyn},\ P_{dyn}) \quad (5)$$

$$C_1 = \text{BlowfishDecrypt}(C_2,\ K_2,\ S_{dyn},\ P_{dyn}) \quad (6)$$

$$P_m = \text{BlowfishDecrypt}(C_1,\ K_1,\ S_{dyn},\ P_{dyn}) \quad (7)$$

$$P = P_m \oplus M$$

C. Key Management and Dynamic Internal Randomization

The framework does not contain the direct re-use of a single stage key, as K_1, K_2, and K_3 are derived from MK and SID. Each encryption session has its own session identifier, and the decryption operation has to build up or securely access the same elements of the session. In the reference implementation, intermediate values, temporary keys, masks, and dynamically created internal structures are wiped out after use. Dynamic S-box and P-array randomization to diversify internal state associated with current session. These structures have a security impact that is dependent on the quality of the random source, and the maintenance of the necessary properties of the cipher. Thus, they are introduced as mechanisms for internal diversification, which serve as supplementation, not substitution, for formal cryptanalytic evaluation (Fig. 4 and 5).

D. Security Interpretation of the Design

The CBT framework applies more transformations to the plaintext and adds session diversification. But it is not only by visual ciphertext appearance nor by NIST p-values that the design is evaluated. Randomness, avalanche sensitivity,

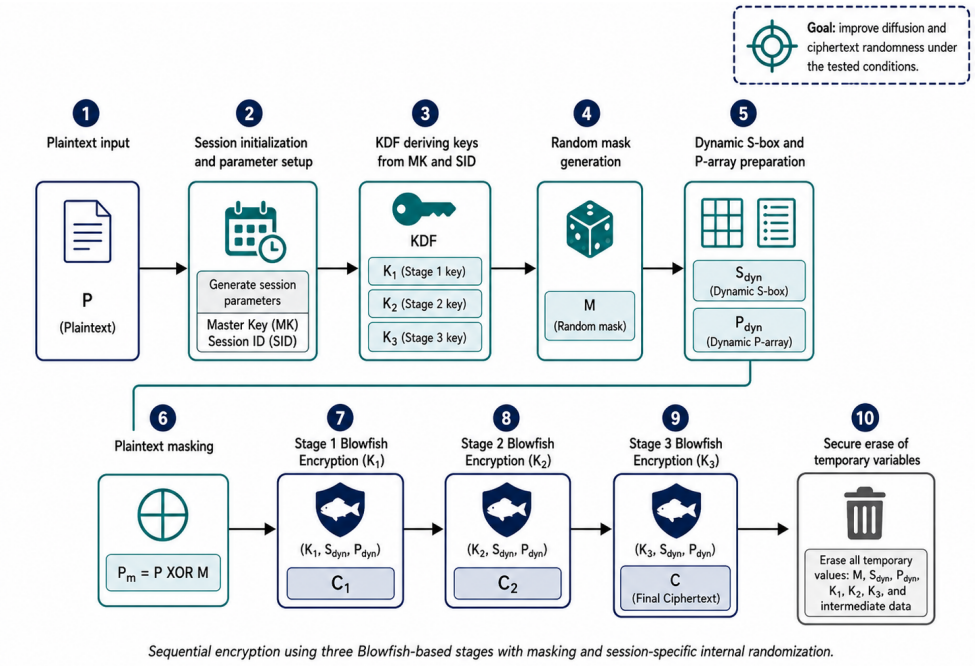


Fig. 2. Encryption flow of the proposed CBT framework. The plaintext is masked using a session-specific random mask and processed through three sequential Blowfish-based encryption stages using stage-specific keys and dynamic internal structures.

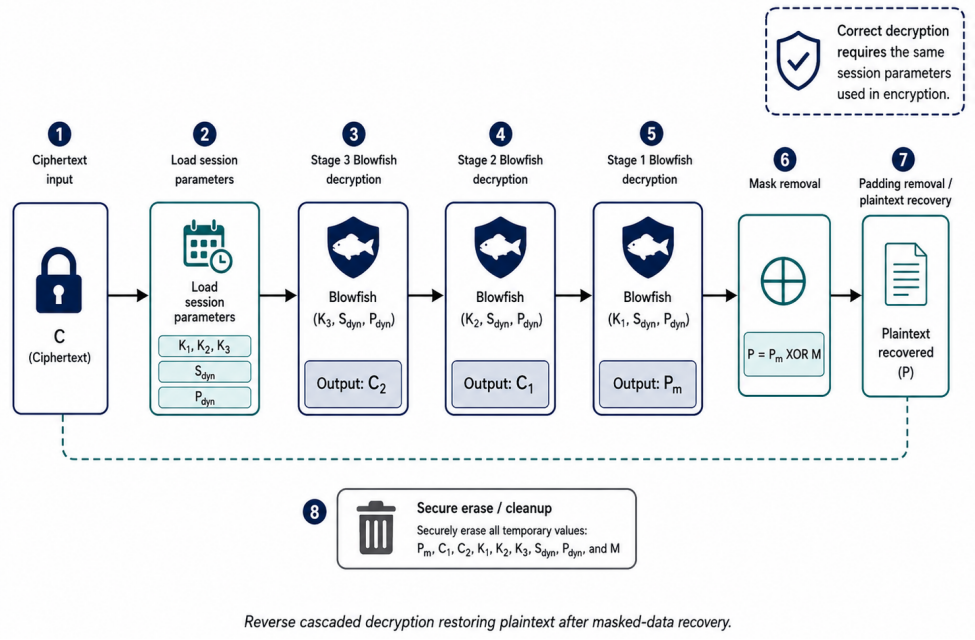


Fig. 3. Decryption flow of the proposed CBT framework. The corresponding session parameters are used to decrypt the ciphertext in the opposite order, removing the mask and recovering the plaintext.

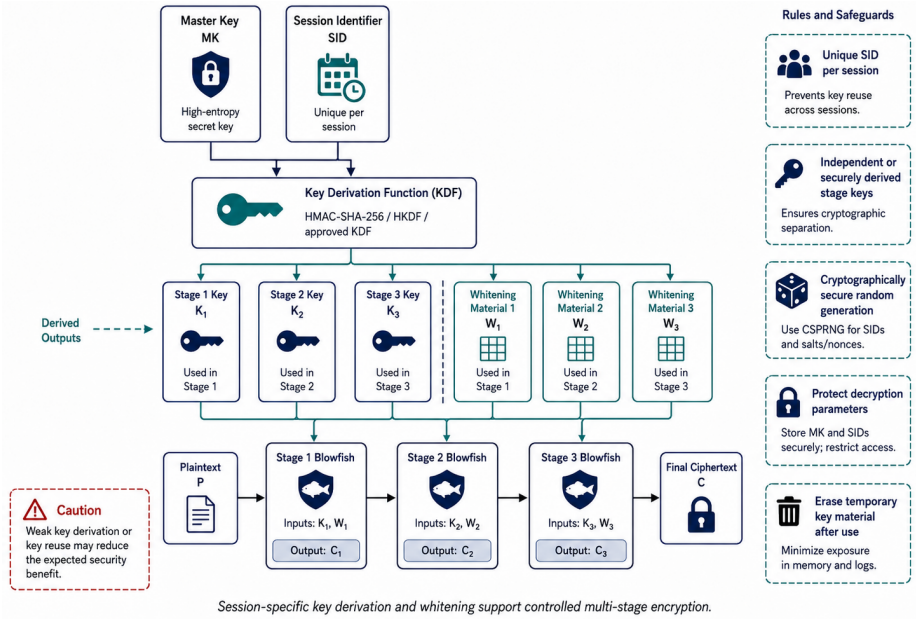


Fig. 4. Key management and whitening flow in the CBT framework. Stage-specific keys and whitening materials are either generated independently or securely derived from a master key and unique session identifier.

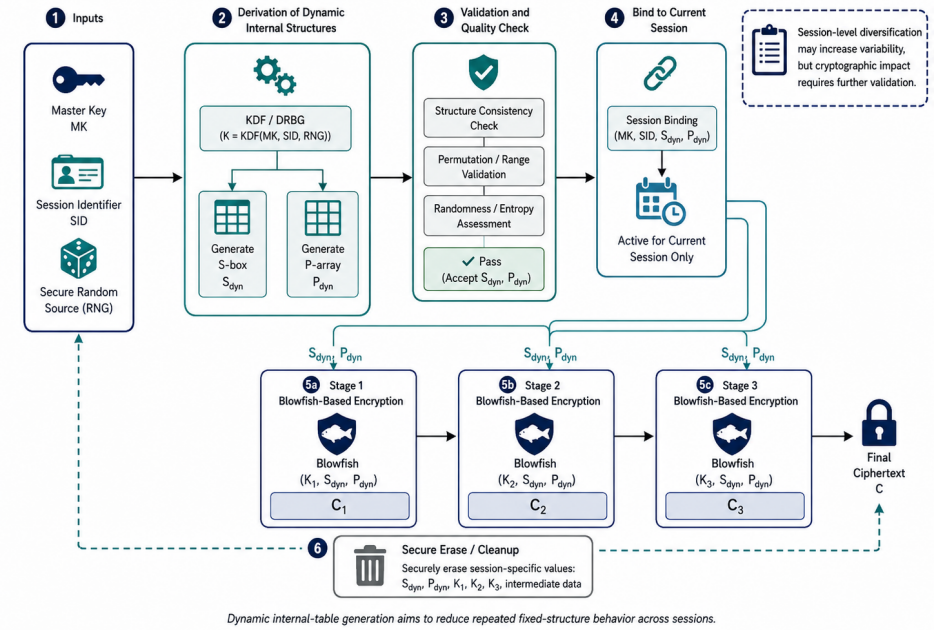


Fig. 5. Dynamic S-box and P-array randomization in the CBT framework. Session-specific internal structures are created, verified, attached to the session and discarded afterwards, to provide internal diversification.

timing variability and performance overhead are reported independently. The resistance to differential cryptanalysis, linear cryptanalysis, chosen-plaintext attack, related-key attack, fault injection, timing leakage, power leakage and electromagnetic attack is not evaluated in the present software prototype and would need dedicated attack evaluation.

IV. Experimental Methodology

The experimental methodology consists of four different parts: NIST-style statistical randomness interpretation, Hamming-distance-based avalanche analysis, computational performance comparison and timing-variability testing.

Pass/fail results have been included in the re-arranged p-values from the original experiment at alpha = 0.01. Avalanche, performance and timing are based on the Python reference implementation, which is included as supplementary material (Fig. 6).

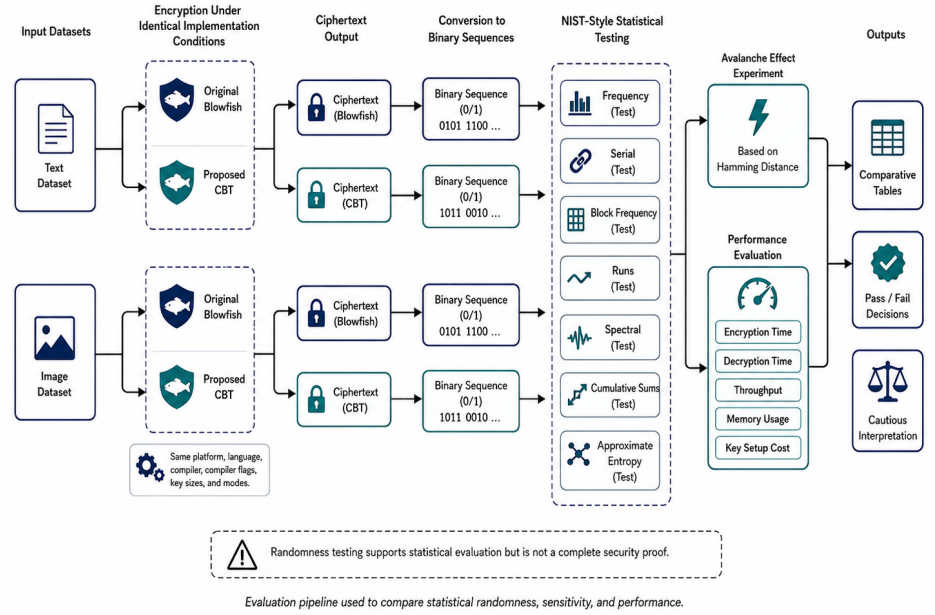


Fig. 6. Experimental evaluation pipeline. Both text and image datasets are encrypted under the same implementation conditions, turned into the binary sequence, and tested statistically, using NIST-style statistical tests, avalanche analysis and performance measurement.

Experimental Setup (Table II)

TABLE II. EXPERIMENTAL SETUP USED FOR THE REVISED EVALUATION

Parameter	Value
Implementation language	Python reference implementation
Compared algorithms	Original Blowfish and proposed CBT framework
Text workload	65,536-byte reproducible byte stream
Image workload	49,152-byte reproducible image-like byte stream
NIST significance level	alpha = 0.01
Avalanche trials	1,000 trials per algorithm, dataset, and sensitivity type
Performance repetitions	50 repetitions per algorithm and dataset
Timing repetitions	200 repetitions per algorithm and input type
Supplementary result workbook	Supplementary File S1
Python code package	Supplementary Files S2 and S3

A. NIST-Style Randomness Interpretation

The NIST type results are analyzed according to the fixed threshold alpha = 0.01. A test result is considered to be pass if $p \geq 0.01$ or fail if $p < 0.01$. The p-value is just a statistical measure of the corresponding randomness test and should not be interpreted as a measure of encryption strength.

$$\text{Decision} = \begin{cases} \text{Pass} & \text{if } p \geq 0.01 \\ \text{Fail} & \text{otherwise} \end{cases}$$

B. Avalanche Effect Analysis

The avalanche experiment evaluates sensitivity to one-bit changes in the plaintext and in the key. For plaintext-bit sensitivity, a plaintext block P and a modified block P' differing in exactly one bit are encrypted under the same key and configuration. For key-bit sensitivity, the same plaintext is encrypted using keys K and K' differing in exactly one bit. The resulting ciphertexts are compared using Hamming distance.

$$\text{Avalanche}(\%) = \frac{HD(C, C')}{N} \times 100 \tag{8}$$

where, $HD(C, C')$ is the Hamming distance between the two ciphertext outputs and N is the ciphertext length in bits.

The reported values summarize the mean, standard deviation, minimum, and maximum over 1,000 trials for each condition.

C. Computational Performance and Timing Variability

Performance is measured using encryption time, decryption time, throughput, memory usage, key setup time, mask-generation cost, and overhead relative to Original Blowfish. The measurements are based on the same reproducible workloads used for the avalanche experiment.

$$\text{Throughput} = \frac{\text{Data size}}{\text{Execution time}} \tag{9}$$

$$\text{Overhead}(\%) = \frac{\text{CBT time} - \text{Blowfish time}}{\text{Blowfish time}} \times 100 \tag{10}$$

The timing-variability check is a comparison of the repeated encryption of fixed and random plaintexts. This check is documented as a software-level timing observation and is not intended as a full-fledged practical side-channel assessment.

V. Results and Discussion

A. NIST-Style Test Outcomes

Table III: 7 of 9 reported results for text and image data were good with the original Blowfish. All 9 outcomes reported in both datasets passed the CBT. The improvement is primarily for the Serial Test parts, where the original Blowfish outputs failed and CBT passed (Table IV). These results are indicative of good ciphertext randomness properties for the chosen tests, but do not constitute full cryptographic security (Fig. 7).

B. Avalanche Effect Results

The mean avalanche value is very close to 50%, which is the expected diffusion property of the output of a block cipher to a one-bit change in the input or key. The CBT results are still near to the original Blowfish baseline and include the masking, cascading, and session-level diversification (Table V). Thus, the avalanche results should be taken more as an indication of the diffusion sensitivity than as a standalone proof of resistance to differential cryptanalysis (Fig. 8).

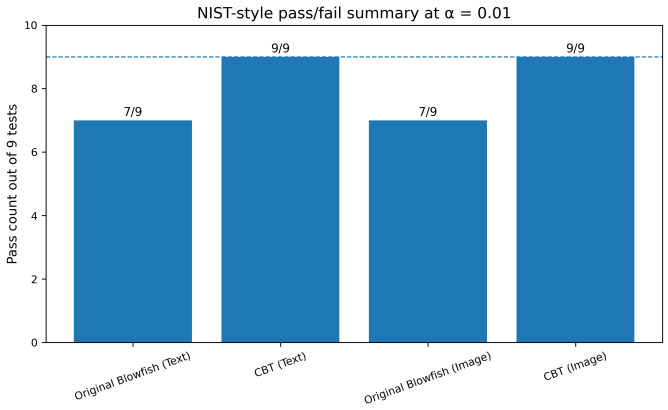


Fig. 7. NIST-style pass/fail summary at alpha = 0.01. The proposed CBT framework achieved 9/9 outcomes for both text and image data, compared to 7/9 outcomes for both text and image data for Original Blowfish.

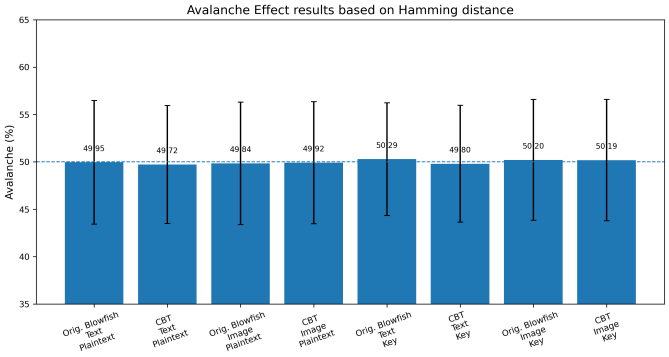


Fig. 8. Avalanche-effect comparison between Original Blowfish and CBT. This involved plaintext-bit and key-bit sensitivity experiments with Hamming-distance values being computed over 1,000 trials per condition.

C. Performance Benchmarking

The CBT prototype is heavier on calculation as it involves performing three consecutive stages of Blowfish and also involves mask generation and session-dependent processing. The measured overhead is that of the Python reference implementation and should not be used to extrapolate to optimized native-code, hardware-assisted or parallel implementations (Table VI). The goal of the benchmark is to transparently measure the prototype cost and not to state anything about the superiority of the speed (Fig. 9).

D. Timing-Variability Check

The timing-variability check gives a partial software-level view of the fixed and random plaintext runtimes (Table VII). These measurements are not a substitute for hardware leakage testing. A complete side-channel evaluation would need a defined leakage model, power or electromagnetic traces, a measurement platform and statistical leakage analysis.

E. Consolidated Interpretation

When the results of the two are combined, it is seen that CBT enhances the NIST-style pass-rate behavior without sacrificing the avalanche values, which are in or close to the desirable 50% diffusion region. This increase in structural complexity is borne out by the performance benchmark, which shows that there is a computational cost in the reference implementation.

So, the new assessment provides a balanced view of the framework: while CBT can enhance selected statistical-randomness metrics and a measure of diversification, p-values alone are not complete evidence of cryptographic security.

VI. Supplementary Materials and Reproducibility

The numerical results presented in this study are supplemented by file(s). A results workbook in Excel format is included as Supplementary File S1 and contains NIST pass/fail tables, raw avalanche trials, avalanche summaries, raw performance measurements, performance summaries, timing-variability data and experimental setup details. The Python experimental package, which was used to produce the avalanche, performance, timing and CSV outputs, is provided in Supplementary File S2. All of the Python scripts have been made available in a Word-readable format in Supplementary File S3 for convenient inspection on devices lacking source-code viewers. The raw CSV outputs from the experimental scripts are provided in Supplementary File S4 (Table VIII).

VII. Conclusion

In this study, a modified and well-documented assessment of Combined Blowfish Trilogy framework for text and image encryption was proposed. The framework involves three stages of Blowfish-based encryption along with key derivation, key whitening ideas, masking of plaintext and session-specific internal randomization. The updated assessment categorizes the different aspects of the experiment into four parts: randomness of the ciphertext, avalanche sensitivity, computational performance, and timing variability.

The NIST-style analysis reveals that the CBT passed all 9 test results for both the text and image datasets at alpha = 0.01, while Original Blowfish passed 7 of 9 for both datasets. The

TABLE III. NIST-STYLE P-VALUES AND PASS/FAIL INTERPRETATION AT ALPHA = 0.01

Test	BF Text	Decision	CBT Text	Decision	BF Image	Decision	CBT Image	Decision
Frequency Test	0.187	Pass	0.233366	Pass	0.596	Pass	0.800278	Pass
Serial Test p1	3.94×10^{-24}	Fail	0.538113	Pass	8.68×10^{-54}	Fail	0.718264	Pass
Serial Test p2	2.63×10^{-12}	Fail	0.250261	Pass	6.55×10^{-27}	Fail	0.726591	Pass
Block Frequency Test	0.0467	Pass	0.834676	Pass	0.228	Pass	0.272722	Pass
Runs Test	0.0557	Pass	0.204289	Pass	0.141	Pass	0.865479	Pass
Spectral Test	0.872	Pass	0.935652	Pass	0.121	Pass	0.037693	Pass
Cumulative Sums Forward	0.224	Pass	0.309039	Pass	0.964	Pass	0.990496	Pass
Cumulative Sums Backward	0.243	Pass	0.365787	Pass	0.525	Pass	0.853891	Pass
Approximate Entropy Test	0.143	Pass	0.469485	Pass	0.0696	Pass	0.490191	Pass

TABLE IV. SUMMARY OF NIST-STYLE PASS RATES

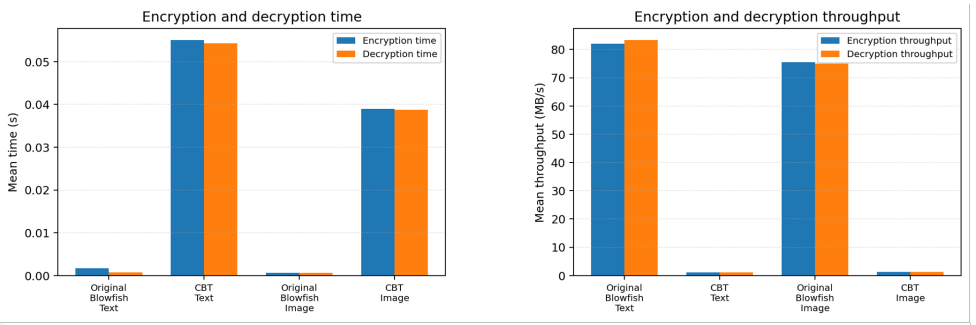
Configuration	Tests	Pass	Pass Rate (%)	Mean p
Original Blowfish (Text)	9	7	77.78	0.196744
CBT (Text)	9	9	100.00	0.460074
Original Blowfish (Image)	9	7	77.78	0.293978
CBT (Image)	9	9	100.00	0.639512

TABLE V. HAMMING-DISTANCE-BASED AVALANCHE EFFECT RESULTS

Algorithm	Dataset	Sensitivity	Trials	Mean (%)	Std. Dev.	Min (%)	Max (%)
Original Blowfish	Image	Key-bit	1000	50.205	6.378	32.812	73.438
Original Blowfish	Image	Plaintext-bit	1000	49.836	6.459	29.688	68.750
CBT	Image	Key-bit	1000	50.188	6.393	32.812	67.188
CBT	Image	Plaintext-bit	1000	49.917	6.435	28.125	71.875
Original Blowfish	Text	Key-bit	1000	50.289	5.945	32.812	68.750
Original Blowfish	Text	Plaintext-bit	1000	49.953	6.524	28.125	70.312
CBT	Text	Key-bit	1000	49.797	6.162	31.250	71.875
CBT	Text	Plaintext-bit	1000	49.722	6.227	31.250	70.312

TABLE VI. PERFORMANCE BENCHMARK RESULTS FOR THE PYTHON REFERENCE IMPLEMENTATION

Algorithm	Dataset	Size (B)	Enc. (s)	Dec. (s)	Enc. TP (MB/s)	Dec. TP (MB/s)	Mem. (MB)	Overhead (%)
Original Blowfish	Image	49152	0.000635	0.000645	75.548	74.940	0.094	0.00
CBT	Image	49152	0.038926	0.038800	1.207	1.210	0.430	6031.24
Original Blowfish	Text	65536	0.001747	0.000754	82.021	83.308	0.138	0.00
CBT	Text	65536	0.055050	0.054325	1.140	1.154	0.568	3051.75



Configuration	Enc. time (s)	Dec. time (s)	Enc. throughput (MB/s)	Dec. throughput (MB/s)	Overhead vs. Blowfish (%)
Original Blowfish (Text)	0.001747	0.000754	82.02	83.31	Baseline
CBT (Text)	0.055050	0.054325	1.14	1.15	3051.8
Original Blowfish (Image)	0.000635	0.000645	75.55	74.94	Baseline
CBT (Image)	0.038926	0.038800	1.21	1.21	6031.2

Performance comparison based on repeated Python benchmark runs under identical experimental conditions.

Fig. 9. Performance evaluation results for the Python reference implementation. The figure shows the summary of encryption/decryption time and throughput for Original Blowfish and CBT for the text workloads and image workloads.

TABLE VII. SOFTWARE TIMING-VARIABILITY CHECK

Algorithm	Input type	Reps	Mean (s)	Std. dev. (s)	Min (s)	Max (s)
CBT	Fixed plaintext	200	0.00245414	0.00061309	0.00201252	0.00561595
CBT	Random plaintext	200	0.00224973	0.00025428	0.00197839	0.00504541
Original Blowfish	Fixed plaintext	200	0.00021473	0.00149891	0.00009377	0.02130122
Original Blowfish	Random plaintext	200	0.00011441	0.00002633	0.00009251	0.00032433

TABLE VIII. SUPPLEMENTARY FILES SUPPLIED WITH THE MANUSCRIPT

File	Contents
S1	`CBT_Supplementary_Results_Workbook.xlsx`: Excel workbook containing all final result sheets. https://doi.org/10.5281/zenodo.21079051
S2	`CBT_Python_Code_Package.zip`: executable Python package and requirements file. https://doi.org/10.5281/zenodo.21079574
S3	`CBT_Python_Codes_Readable.docx`: Word-readable copy of all Python scripts. https://doi.org/10.5281/zenodo.21079681
S4	`CBT_Raw_CSV_Outputs.zip`: raw CSV files exported from the experimental run. https://doi.org/10.5281/zenodo.21079770

Hamming-distance-based avalanche analysis gave the mean values close to 50%, which showed suitable sensitivity for one-bit change in plaintext and key. The cost of the prototype with the three stages was quantified and it was demonstrated that the extra transformations incur measurable overhead in the Python implementation.

Results indicate that, under the tested conditions, the CBT is a diversified framework based on Blowfish with better selected statistical-randomness behavior. Meanwhile, the study does not claim that the NIST p-values are full security proofs, and clarifies what constitutes software-level timing checks from actual side-channel resistance. Future implementation studies can optimize the prototype in native code and/or hardware and do dedicate leakage measurement, but the present manuscript offers a complete reproducible baseline for statistical, avalanche, and performance evaluation.

REFERENCES

[1] C. E. Shannon, "Communication theory of secrecy systems," *Bell System Technical Journal*, vol. 28, no. 4, pp. 656–715, 1949. https://doi.org/10.1002/j.1538-7305.1949.tb00928.x

[2] B. Schneier, "Description of a new variable-length key, 64-bit block cipher (Blowfish)," in *Fast Software Encryption*, LNCS 809, pp. 191–204, Springer, 1994. https://doi.org/10.1007/3-540-58108-1_24

[3] National Institute of Standards and Technology, "Advanced Encryption Standard (AES)," FIPS PUB 197, 2001; updated 2023. https://doi.org/10.6028/NIST.FIPS.197-upd1

[4] National Institute of Standards and Technology, "Data Encryption Standard (DES)," FIPS PUB 46-3, 1999, withdrawn 2005. https://csrc.nist.gov/pubs/fips/46-3/final

[5] E. Barker and N. Mouha, "Recommendation for the Triple Data Encryption Algorithm (TDEA) Block Cipher," NIST SP 800-67 Rev. 2, 2017. https://doi.org/10.6028/NIST.SP.800-67r2

[6] M. Dworkin, "Recommendation for Block Cipher Modes of Operation: Methods and Techniques," NIST SP 800-38A, 2001. https://doi.org/10.6028/NIST.SP.800-38A

[7] A. Rukhin, J. Soto, J. Nechvatal *et al.*, "A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications," NIST SP 800-22 Rev. 1a, 2010. https://doi.org/10.6028/NIST.SP.800-22r1a

[8] A. Iwasaki, "Analysis of NIST SP800-22 focusing on randomness of each sequence," arXiv:1710.01441, 2017. https://arxiv.org/abs/1710.01441

[9] H. Haramoto *et al.*, "Study on upper limit of sample sizes for a two-level test in NIST SP800-22," arXiv:1912.10602, 2019. https://arxiv.org/abs/1912.10602

[10] A. F. Webster and S. E. Tavares, "On the design of S-boxes," in *Advances in Cryptology — CRYPTO '85*, LNCS 218, pp. 523–534, Springer, 1985. https://doi.org/10.1007/3-540-39799-X_41

[11] E. Biham and A. Shamir, "Differential cryptanalysis of DES-like cryptosystems," *Journal of Cryptology*, vol. 4, no. 1, pp. 3–72, 1991. https://doi.org/10.1007/BF00630563

[12] M. Matsui, "Linear cryptanalysis method for DES cipher," in *Advances in Cryptology — EUROCRYPT '93*, LNCS 765, pp. 386–397, Springer, 1994. https://doi.org/10.1007/3-540-48285-7_33

[13] P. C. Kocher, "Timing attacks on implementations of Diffie-Hellman, RSA, DSS, and other systems," in *Advances in Cryptology — CRYPTO '96*, LNCS 1109, pp. 104–113, Springer, 1996. https://doi.org/10.1007/3-540-68697-5_9

[14] P. Kocher, J. Jaffe, and B. Jun, "Differential power analysis," in *Advances in Cryptology — CRYPTO '99*, LNCS 1666, pp. 388–397, Springer, 1999. https://doi.org/10.1007/3-540-48405-1_25

[15] S. Chari, J. R. Rao, and P. Rohatgi, "Template attacks," in *Cryptographic Hardware and Embedded Systems — CHES 2002*, LNCS 2523, pp. 13–28, Springer, 2003. https://doi.org/10.1007/3-540-36400-5_3

[16] S. Mangard, E. Oswald, and T. Popp, *Power Analysis Attacks: Revealing the Secrets of Smart Cards*. Springer, 2007. https://doi.org/10.1007/978-0-387-38162-6

[17] M. Rivain and E. Prouff, "Provably secure higher-order masking of AES," in *Cryptographic Hardware and Embedded Systems — CHES 2010*, LNCS 6225, pp. 413–427, Springer, 2010. https://doi.org/10.1007/978-3-642-15031-9_28

[18] Y. Ishai, A. Sahai, and D. Wagner, "Private circuits: Securing hardware against probing attacks," in *Advances in Cryptology — CRYPTO 2003*, LNCS 2729, pp. 463–481, Springer, 2003. https://doi.org/10.1007/978-3-540-45146-4_27

[19] L. Chen, "Recommendation for Key Derivation Using Pseudorandom Functions," NIST SP 800-108 Rev. 1, 2022. https://doi.org/10.6028/NIST.SP.800-108r1

[20] H. Krawczyk and P. Eronen, "HMAC-based Extract-and-Expand Key Derivation Function (HKDF)," RFC 5869, IETF, 2010. https://www.rfc-editor.org/rfc/rfc5869

[21] E. Barker and J. Kelsey, "Recommendation for Random Number Generation Using Deterministic Random Bit Generators," NIST SP 800-90A Rev. 1, 2015. https://doi.org/10.6028/NIST.SP.800-90Ar1

[22] M. S. Turan, E. Barker, J. Kelsey *et al.*, "Recommendation for the Entropy Sources Used for Random Bit Generation," NIST SP 800-90B, 2018. https://doi.org/10.6028/NIST.SP.800-90B

[23] M. Luby and C. Rackoff, "How to construct pseudorandom permutations from pseudorandom functions," *SIAM Journal on Computing*, vol. 17, no. 2, pp. 373–386, 1988. https://doi.org/10.1137/0217022

[24] O. Dunkelman, N. Keller, and A. Shamir, "Minimalism in cryptography: The Even-Mansour scheme revisited," in *Advances in Cryptology — EUROCRYPT 2012*, LNCS 7237, pp. 336–354, Springer, 2012. https://doi.org/10.1007/978-3-642-29011-4_21

[25] G. Álvarez and S. Li, "Some basic cryptographic requirements for chaos-based cryptosystems," *International Journal of Bifurcation and Chaos*, vol. 16, no. 8, pp. 2129–2151, 2006. https://doi.org/10.1142/S0218127406015970

[26] J. Fridrich, "Symmetric ciphers based on two-dimensional chaotic maps," *International Journal of Bifurcation and Chaos*, vol. 8, no. 6, pp. 1259–1284, 1998. https://doi.org/10.1142/S021812749800098X

[27] G. Chen, Y. Mao, and C. K. Chui, "A symmetric image encryption scheme based on 3D chaotic cat maps," *Chaos, Solitons & Fractals*, vol. 21, no. 3, pp. 749–761, 2004. https://doi.org/10.1016/j.chaos.2003.12.022

[28] Y. Wu, J. P. Noonan, and S. Agaian, "NPCR and UACI randomness tests for image encryption," *Journal of Selected Areas in Telecommunications*, vol. 1, no. 2, pp. 31–38, 2011.

[29] P. Ramasamy, V. Ranganathan, S. Kadry, R. Damaševičius, and T. Blažauskas, "An image encryption scheme based on block scrambling, modified zigzag transformation and key generation using enhanced logistic-tent map," *Entropy*, vol. 21, no. 7, p. 656, 2019. https://doi.org/10.3390/e21070656

[30] M. Ahmad and A. Chopra, "Chaotic dynamic S-boxes based substitution approach for digital images," arXiv:1709.07620, 2017. https://arxiv.org/abs/1709.07620

[31] M. Gholamzadeh and B. Khadem, "A hybrid image encryption scheme based on chaos and a DPA-resistant S-box," arXiv:2312.15280, 2023. https://arxiv.org/abs/2312.15280

[32] M. S. Turan *et al.*, "Ascon-Based Lightweight Cryptography Standards for Constrained Devices," NIST SP 800-232, 2025. https://doi.org/10.6028/NIST.SP.800-232

[33] A. Bogdanov, L. R. Knudsen, G. Leander *et al.*, "PRESENT: An ultra-lightweight block cipher," in *Cryptographic Hardware and Embedded Systems — CHES 2007*, LNCS 4727, pp. 450–466, Springer, 2007. https://doi.org/10.1007/978-3-540-74735-2_31

[34] R. Beaulieu, D. Shors, J. Smith, S. Treatman-Clark, B. Weeks, and L. Wingers, "The SIMON and SPECK families of lightweight block ciphers," IACR ePrint 2013/404, 2013. https://eprint.iacr.org/2013/404

[35] A. E. Adeniyi *et al.*, "Computational complexity of modified Blowfish cryptographic algorithm for video encryption," *Algorithms*, vol. 15, no. 10, p. 373, 2022. https://doi.org/10.3390/a15100373

[36] A. A. A. Abdulkadhim and A. Lakizadeh, "Hybrid parallelism image encryption algorithm based on a modified Blowfish algorithm," *International Journal of Intelligent Engineering and Systems*, vol. 17, no. 6, pp. 934–946, 2024. https://doi.org/10.22266/ijies2024.1231.70

Unsupervised Hybrid Approach for Intrusion Detection Based on Information System Log Files

Sena Apeke[1], Nadjime Pindra[2], Sèmèvo A. R. M. Ahouandjinou[3]
Laboratoire de Recherche en Science de l'Ingénieur, Universite de Lomé, Togo[1]
Centre d'Excellence Régional pour la Maîtrise de l'Electricité (CERME), Universite de Lomé, Lomé, Togo[1]
Laboratoire d'Analyse, de Modélisations Mathématiques et Applications, Universite de Lomé, Togo[2]
Laboratoire de recherche en Sciences Informatique et Applications, Université d'Abomey-calavi, Bénin[3]

***Abstract*—Intrusion detection in web traffic remains a challenging task due to the high dimensionality, heterogeneity, and imbalance of normal and malicious requests. This study investigates a hybrid anomaly detection framework combining an Autoencoder (AE), Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and Isolation Forest (IF) for intrusion detection in HyperText Transfer Protocol (HTTP) log data. The Autoencoder was first used to learn a compact latent representation of the traffic, after which different unsupervised detection pipelines were evaluated. A key contribution of this work is the analysis of pipeline ordering, specifically comparing AE + DBSCAN + IF, AE + IF, and AE + DBSCAN. The full hybrid pipeline AE + DBSCAN + IF achieved the highest accuracy (0.9803) and strong false positive control, but exhibited extremely poor sensitivity with a recall of 0.0082 and precision of 0.0904, making it too conservative for practical intrusion detection. The intermediate pipeline AE + IF improved sensitivity, reaching a recall of 0.1542 and precision of 0.0536, but at the expense of a higher false positive rate. Surprisingly, the ablation study showed that AE + DBSCAN alone yielded the most balanced results, with accuracy of 0.9670, precision of 0.2606, and recall of 0.4376, clearly outperforming the complete hybrid pipeline in intrusion detection capability. These findings demonstrate that the ordering of anomaly detection stages has a critical impact on IDS performance. In the current setting, Isolation Forest degrades rather than improves detection when combined with DBSCAN. The results suggest that AE + DBSCAN constitutes the most effective and operationally relevant configuration for the studied dataset.**

***Keywords*—Intrusion; detection; DBSCAN; Isolation Forest; Autoencoder; server log files; anomaly**

I. Introduction

In the age of digital transformation, servers form the operational foundation of modern IT infrastructures. They host critical applications, centralize databases, orchestrate network services, and ensure the continuous availability of digital resources that are essential to public and private organizations. Whether for government agencies, universities, hospitals, businesses, or online service providers, server reliability directly impacts business continuity, service quality, and the protection of information assets. This strategic position, however, makes servers prime targets for cyberattacks, particularly malicious intrusions, which aim to compromise the confidentiality, integrity, or availability of information systems [5], [2], [6], [3].

An intrusion can be defined as any unauthorized attempt or action aimed at accessing a system, disrupting its operation, exfiltrating data, disrupting services, or maintaining a covert presence within it. In server environments, these attacks take various forms: privilege escalation, lateral movement, exploitation of software vulnerabilities, execution of malicious commands, installation of malware, Denial-of-Service attacks, or compromise of authentication mechanisms. The consequences can be particularly severe: leakage of sensitive data, unavailability of critical services, financial losses, regulatory noncompliance, damage to institutional reputation, and a breach of user trust [18], [14], [10]. In this context, server security is no longer solely a matter of perimeter defense, but rather of intelligent, continuous, and adaptive monitoring of system events.

The rapidly evolving threat landscape has exposed the limitations of traditional security approaches. Conventional defense mechanisms, such as firewalls, antivirus software, and signature-based intrusion detection systems, remain useful for identifying known behaviors, but they often prove insufficient against new, polymorphic, or stealthy attacks. Signature-based intrusion detection systems (IDS) generally compare observed events against a database of rules or fingerprints corresponding to previously documented attacks. This approach is effective against known threats but becomes less reliable when the attacker employs novel, modified, or deliberately obfuscated techniques [8], [1], [19]. This issue is particularly significant in the case of so-called zero-day attacks, which exploit vulnerabilities that have not yet been documented or patched. Although some security solutions can indirectly detect certain behavioral manifestations of such attacks, their reliable detection remains a major challenge, particularly in complex and highly dynamic environments [8], [1].

At the same time, servers constantly generate massive amounts of data in the form of event logs produced by the operating system, network services, applications, authentication mechanisms, administration tools, and security components. These logs record a wide variety of information: incoming and outgoing connections, system errors, file access, authentication failures, executed commands, application requests, security alerts, and performance anomalies. In theory, these traces represent an extremely rich source for monitoring, auditing, and the early detection of suspicious behavior. In practice, however, leveraging them remains difficult due to several factors: heterogeneous formats, significant noise, lack of uniform structure, redundancy, high volume, and the semantic complexity of the messages [12], [4], [1]. Manual analysis of this data quickly becomes impractical, especially in operational contexts where security teams have limited human resources

and must respond within very short timeframes.

This challenge has led to the emergence of advanced intrusion detection approaches based on machine learning (ML) and deep learning (DL). Unlike purely signature-based approaches, machine learning methods aim to model a system's normal behavior or learn discriminative patterns from historical data in order to identify significant deviations that may indicate malicious activity. These approaches are particularly promising for server log analysis, as they allow for the exploitation of temporal, structural, and contextual patterns that are difficult to capture using static rules [9], [17], [1].

The growing interest in ML and DL techniques in IDS systems can also be attributed to their ability to address the increasing complexity of modern cyberattacks. Models such as decision trees, random forests, support vector machines, autoencoders, recurrent neural networks, convolutional models, and hybrid architectures have demonstrated promising performance in various detection contexts, whether analyzing network traffic, user behavior, or log sequences [6], [9], [1].

From this perspective, intelligent server log analysis emerges as a strategic priority for strengthening the cybersecurity of digital infrastructures. It not only enables earlier detection of abnormal behavior but also provides a better understanding of attack mechanisms, reduces response times, and improves the overall resilience of systems. However, several challenges remain: high variability in server environments, an imbalance between normal and malicious events, computational cost, the need for explainability, false positive rates, and the difficulty of integration into real-world operational environments [1], [12], [10]. This study fits within this context and aims to explore the use of log files for server intrusion detection using modern artificial intelligence techniques. The goal is to propose an approach capable of leveraging these execution traces as a source of actionable knowledge, in order to improve security monitoring, anomaly detection, and, more broadly, the proactive protection of server systems.

Despite recent progress in machine learning and deep learning for intrusion detection, two research gaps remain insufficiently addressed. First, most existing studies rely on benchmark network datasets, whereas real server log files are highly heterogeneous, noisy, redundant, and severely imbalanced. Second, the influence of the ordering of unsupervised anomaly detectors within hybrid pipelines has rarely been investigated. In particular, it remains unclear whether density-based clustering should be applied before or after isolation-based anomaly detection when the input consists of engineered server log features. This work addresses these gaps by evaluating the impact of component ordering in three pipelines: AE+DBSCAN+IF, AE+IF+DBSCAN, and AE+DBSCAN.

The remainder of this study is organized as follows: Section II presents the dataset, the feature engineering process, and the proposed hybrid intrusion detection methodology. Section III reports the experimental results and compares the investigated pipelines. Section IV discusses the main findings, the effect of pipeline ordering, and the limitations of the approach. Finally, Section V concludes the study and outlines future research directions.

II. Methods and Materials

A. Data Collection and Description

This study addresses intrusion detection from server log files in a cybersecurity context. The selected data consist mainly of Apache server access logs, which are semi-structured records generated by web services. Such logs play an important role in threat identification, suspicious behavior analysis, and decision support for protecting critical information systems. Their quality directly influences the ability of information systems to detect unauthorized access attempts, Denial-of-Service (DoS) events [11], and malicious injection attempts [21]. Apache access logs record each HTTP request and include information such as the IP address, timestamp, HTTP method, requested URL, status code, response size, referrer, and user agent. These attributes are useful for identifying suspicious activities, including repeated authentication failures, abnormal status codes, malformed requests, and request peaks related to Distributed Denial-of-Service (DDoS) behavior. The extracted log records are highly imbalanced, as shown in Fig. 1, which presents their temporal distribution between 2021 and 2024.

B. Data Preprocessing

Preprocessing server log files is a demanding step due to the complexity of these files. The first work done was to analyze the raw logs (.log) following the work of [16] and [15] using Python regular expression library to extract eight variables (Table I), the results of the processing of each log file were structured (.csv). Then, an algorithm was developed to eliminate redundant information following the work of [20]. This process, carried out in two phases, made it possible to convert unstructured data into an exploitable basis for intrusion detection. Files were compiled into a single one. After merging all these files, the data set contained 3,558,335 rows and 8 columns, representing HTTP requests recorded between 2021 and 2024. To clearly see the imbalance in the data, a study of the distribution of each variable was carried out. In fact, most of the variables are complex. Fig. 2 summarizes the distributions of the variables Status_code, Referer, User_Agent, and Method. More specifically, Fig. 2a, Fig. 2b, Fig. 2c, and Fig. 2d illustrate the four distributions, respectively. The eight variables defined in Table I are complex variables; to better present a line of information, it was necessary to break down these variables. For example, the IP variable has been split into 8 variables: ip_request_count represents the total number of requests made by an IP in the logs; ip_time_spread represents the duration in seconds between the first and last request of an IP, etc. The URL variable has been split into nine new ones: url_path_depth gives the number of segments in the URL path; url_path_special_chars_count gives the number of special characters in the URL path, this variable reports malformed or encoded URLs, typical attacks; fragment_param_count gives the number of parameters in the fragment (#); etc. This analysis made it possible to have 34 relevant variables which give complete coverage of intrusion signals: temporal (cyclical and linear), IP (intensity, errors, geography), URL (structure, complexity), User-Agent (automation), Referer (navigation), and status (failures). In sum, there are in the dataset $N = 3558335$ observations, each having 34 components, any observation will be noted x_i.

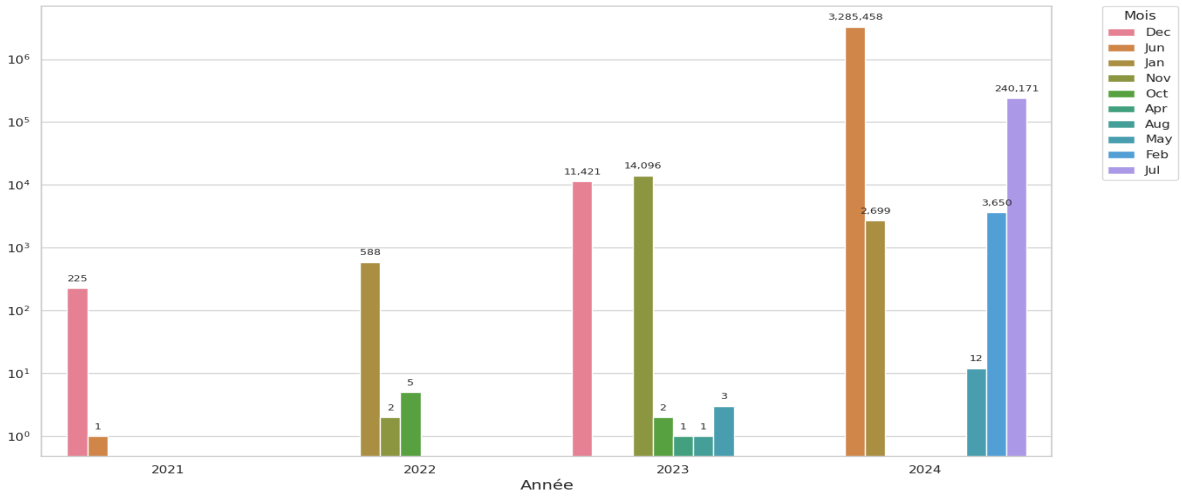
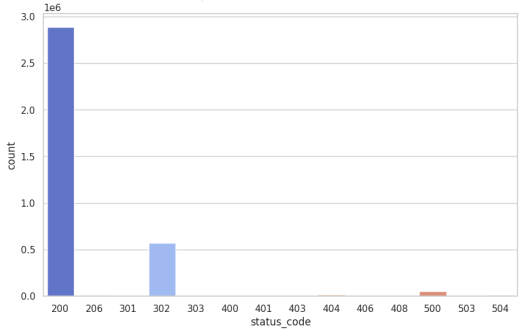
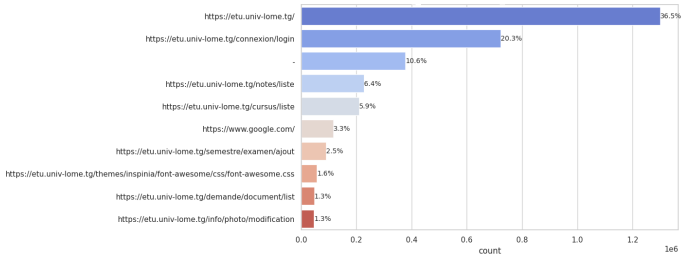


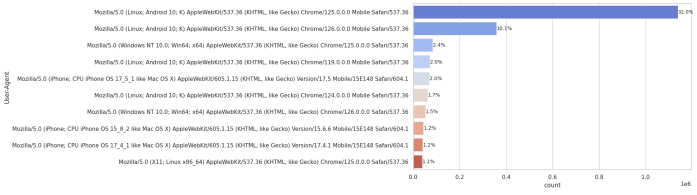
Fig. 1. Temporal distribution of information in the log files (2021–2024).



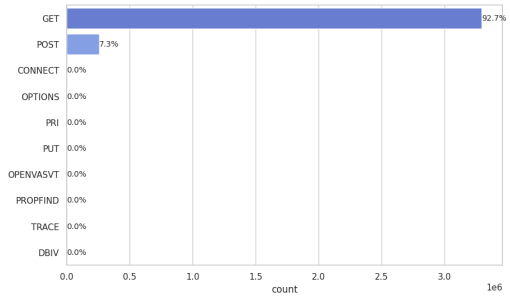
(a) Distribution of the Status_code variable.



(b) Distribution of the Referer variable.



(c) Distribution of the User_Agent variable.



(d) Distribution of the Method variable.

Fig. 2. Distribution of a few variables reflecting the unbalanced aspect of the data.

Table II provides the complete list of engineered features used as an input to the proposed intrusion detection pipelines. These variables summarize temporal behavior, IP activity, URL structure, HTTP response status, User-Agent characteristics, Referer information, and fragment-level patterns.

The overall methodology is summarized in Fig. 3. The workflow starts from raw Apache log files and proceeds through parsing, feature engineering, normalization, latent representation learning with an Autoencoder, evaluation of alternative pipeline orderings, and final performance assessment.

C. Modeling

The relevance of intrusion detection models based on log files requires that the different attack techniques be well represented in the data. The consequence is that very easily, the size of the training data becomes very large. To reduce the computational complexity in training, autoencoders were used initially. This is because autoencoders reduce dimensionality

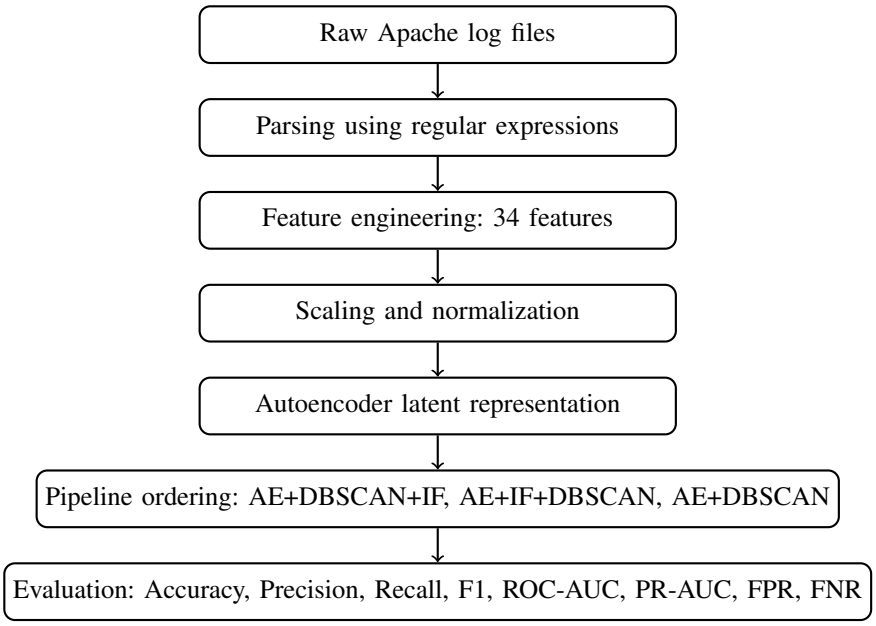


Fig. 3. Workflow of the proposed methodology for server log-based intrusion detection.

TABLE I. DESCRIPTION OF THE VARIABLES EXTRACTED FROM LOGS

Variable	Description
IP	Client IP address (e.g.: 192.168.1.1). Identifies suspicious sources (e.g. repetitive IPs in attacks DDoS)
Date	Timestamp (e.g.: 14/Mar/2025:10:15:30 +0100). Allows temporal analysis of peaks activity
Method	HTTP method (e.g.: GET, POST). Reveals actions unusual (e.g.: repeated POSTs)
URL	Requested resource (eg: /login.php). Detects access to sensitive pages or injections
Status_code	HTTP code (e.g.: 200, 404). Reports repeated failures (e.g.: 401 for brute force)
Response_size	Response size (e.g.: 1024 bytes). Indicates faults mallies (e.g.: high values for DoS)
Referer	Original URL (e.g.: http:/example.com). Detect the fraudulent or absent referents
User_agent	Client identifier (e.g.: Mozilla/5.0). Spot the bots or malicious tools

by compressing the input data into a lower-dimensional representation, known as a latent space layer, and then reconstructing the original input from this compressed representation. An autoencoder is a neural network that is trained to attempt to copy its input to its output (Fig. 5). Since autoencoders behave like a model solving a regression problem, the error function often used is the mean square error (E_{MSE}). As in machine learning or deep learning regression models, when training, we aim to minimize the cost function.

In this study, the Autoencoder architecture follows a symmetric fully connected structure: 34–64–32–16–8–16–32–64–34. Rectified Linear Unit (ReLU) activations are used in the hidden layers, the bottleneck layer contains eight latent dimensions, and the model is optimized by minimizing the mean square reconstruction error using the Adam optimizer. This architecture is illustrated in Fig. 4.

$$E_{MSE} = \frac{1}{N}\sum_{i=1}^{N} ||x_i - \tilde{x_i}||_2^2 \quad (1)$$

The symbol $||.||_2$ represents the $\mathcal{L}^2$ norm of a vector, and N is the number of the observation in the training dataset. $\tilde{x_i}$ is the decoder output. The more the error tends towards zero, the better the observations are reconstructed.

The proposed intrusion detection framework was designed under a semi-unsupervised anomaly detection paradigm, where the latent space learned by the autoencoder was exploited to improve separability between benign and suspicious traffic behaviors. Tree hybrid pipelines were conceptually evaluated: (1) AE + DBSCAN + IF, (2) AE + IF + DBSCAN and (3) AE+DBSCAN

1) Description of isolation forest: Let $X = \{x_1, \ldots, x_n\}$ be a set of information data files where $x_i \in \mathbb{R}^{34}$, 34 represents the number of characteristics (features) that describe data in this work. An isolation tree (isoTree) is defined as a data structure with an algorithm that is organized as a decision tree [13]. The principle of this algorithm is very simple:

- Randomly select a variable.
- Then carry out a random partitioning of the dataset according to this variable so as to obtain two subsets of data.
- Repeat the previous two steps until a piece of data is isolated.
- Recursively, repeat the previous steps.

isoTree is based on the idea that anomalies, rare and distinct, are isolated more quickly in random decision trees than normal points. For each recording (e.g., Apache log), a

TABLE II. DESCRIPTION OF THE ENGINEERED FEATURES EXTRACTED FROM THE SERVER LOG FILES

Feature	Description
response_size	HTTP response size in bytes.
date_hour_sin	Sine transformation of the request hour (0–23).
date_hour_cos	Cosine transformation of the request hour (0–23).
date_normalized	Elapsed time in seconds since the first request.
ip_request_count	Total number of requests generated by the same IP address.
ip_time_spread	Time elapsed between the first and last request from the same IP address.
ip_subnet_activity	Number of active IP addresses within the same /24 subnet.
ip_error_rate	Proportion of requests returning HTTP error codes (≥ 400).
ip_request_rate	Average request rate per second.
ip_interval_std	Standard deviation of time intervals between consecutive requests.
ip_max_requests_per_sec	Maximum number of requests per second generated by the same IP address.
ip_method_diversity	Number of distinct HTTP methods used by the same IP address.
country	Country of origin of the IP address based on a GeoIP database.
status_code_category	Category of the HTTP response status code.
nbre_unique_urls_per_ip	Number of distinct URLs requested by the same IP address.
url_path_struct_entropy	Structural entropy of the URL path.
url_path_depth	Number of path segments in the requested URL.
nbre_total_url_query_params	Number of query parameters in the requested URL.
url_query_special_chars_count	Number of special characters in the URL query string.
url_path_special_chars_count	Number of special characters in the URL path.
user_agent_length	Length of the User-Agent string.
user_agent_entropy	Character entropy of the User-Agent string.
user_agent_total_special_chars	Number of special characters in the User-Agent string.
referer_is_present	Binary indicator equal to 1 if the Referer header is present and 0 otherwise.
referer_path_struct_entropy	Structural entropy of the Referer path.
referer_path_depth	Number of path segments in the Referer URL.
nbre_total_referer_query_params	Number of query parameters in the Referer URL.
referer_query_special_chars_count	Number of special characters in the Referer query string.
is_bot	Binary label equal to 1 for bot traffic and 0 otherwise.
fragment_param_count	Number of parameters in the URL fragment.
special_char_fragment_count	Number of special characters in the URL fragment.

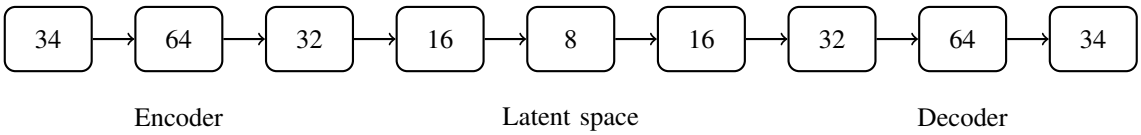


Fig. 4. Architecture of the Autoencoder used for dimensionality reduction. The input layer contains the engineered log features, the bottleneck contains eight latent units, and the decoder reconstructs the original feature vector.

set of trees is constructed by randomly selecting a feature and a cutoff value until each point is isolated. The anomaly score is given by:

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}} \tag{2}$$

where,

- $h(x)$: isolation depth of x in a tree (number of divisions to isolate it)
- $E(h(x))$: average of h(x) over T trees
- $c(m) = 2H(m-1) - \frac{2(m-1)}{n}$ if $m > 2$

n is the test set size, m is the size of the sample set, and H is the harmonic number, which can be estimated by: $H(k) = \sum_{i=1}^{k} \frac{1}{i}$, [7].

If the score s is close to 1, x is most likely an anomaly, but if s is less than 0.5, x is probably normal. Isolation Forest has the ability to detect global anomalies without relying on density or distance comparison. However, it is not effective in detecting local anomalies.

2) Description of DBSCAN: DBSCAN is a well-known unsupervised clustering algorithm. It was proposed in 1996 by [7]. Given points and an integer k, the algorithm aims to divide the points into k homogeneous and compact groups, called clusters. Given points and an integer k, the algorithm aims to divide the points into k groups, called clusters, which are homogeneous and compact. It is a simple algorithm that defines clusters using local density estimation. It can be divided into 4 steps: For each observation, we look at the number of points at most one distance ϵ from it. This area is called the ϵ-neighborhood of the observation. If an observation has at least a certain number of neighbors including itself, it is considered a “core" observation. In this case, a high-density observation was detected. All observations in the vicinity of a core observation belong to the same cluster. There may be core observations close to each other. As a result, from close to close, a long sequence of core observations was obtained that form a single cluster. Any observation that is not a core observation and has no core observation in its vicinity is considered an “anomaly". Two main parameters are used in DBSCAN: epsilon (ϵ), which defines the maximum distance between two points to be considered neighbors, and MinPts, which specifies the minimum number of points required to form a dense region. Note also the choice of a distance to be used (usually the Euclidean distance). Given below is the

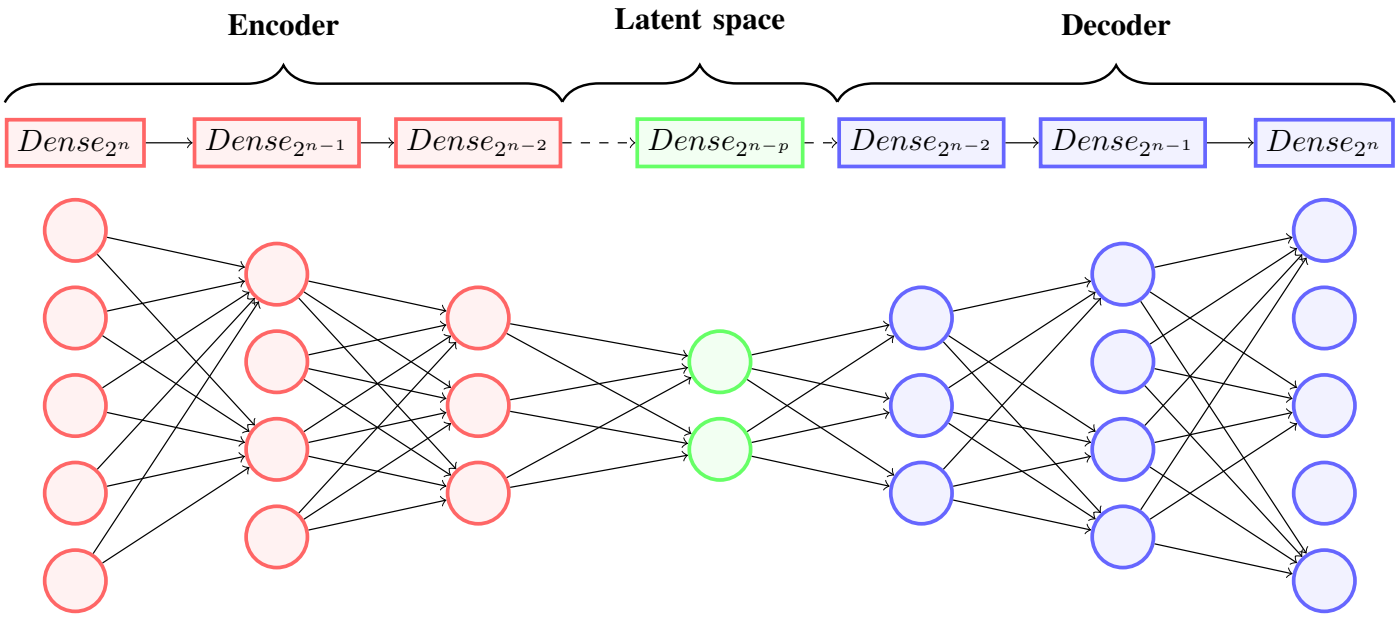


Fig. 5. Example of the diagram of an Autoencoder. $\mathbf{Dense_{2^n}}$ is a dense layer like in keras and 2^n is the number of units. $n = 7$ and $p = 4$ in the context of this study.

algorithm of DBSCAN.

1) **Initialization**
   - Consider the dataset $D = \{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}^{34}$.
   - Set parameters: neighborhood radius ε and *MinPts*.
   - Mark all points in D as *unvisited*.
2) **Loop over points**
   - For each point $x_i \in D$:
     - If x_i is already visited, move on to the next point.
     - Otherwise, mark x_i as *visited*.
3) **Neighborhood calculation**
   - Calculate the neighborhood of x_i:

$$N_\varepsilon(x_i) = \{x_j \in D \mid ||x_j - x_i|| \leq \varepsilon\}.$$

   - If $|N_\varepsilon(x_i)| <$ MinPts, then :
     - Mark x_i as *noise*.
     - Continue with the next point.
4) **Create a cluster**
   - Create a new cluster C.
   - Add x_i to the cluster C.
5) **Cluster expansion**
   - For each point $p \in N_\varepsilon(x_i)$:
     - If p is not visited:
       - Mark p as *visited*.
       - Calculate its neighborhood $N_\varepsilon(p)$.
       - If $|N_\varepsilon(p)| \geq$ MinPts, extend the neighborhood:

$$N_\varepsilon(x_i) \leftarrow N_\varepsilon(x_i) \cup N_\varepsilon(p).$$

     - If p doesn't belong to any cluster, add it to C.
6) **End of algorithm**
   - Repeat until all points in D have been visited.
   - Return:

$$\{C_1, C_2, \ldots, C_K\}, \qquad \text{Noise} = D \setminus \bigcup_{k=1}^{K} C_k.$$

3) Goal: In this study, three methods were developed: first, the AE+DBSCAN pipeline was trained; next, the AE+DBSCAN pipeline; and finally, the AE+IF+DBSCAN pipeline. The objective was to determine which ordering of density-based clustering and anomaly isolation is more appropriate for network intrusion detection.

III. Results

The dataset was partitioned into three subsets in order to ensure a robust evaluation of the intrusion detection pipeline: training set 70% of the data, validation set 10% of the data and test set 20% of the data. The training subset was primarily used to learn the latent representation of the normal traffic through the Autoencoder (AE) and to calibrate the anomaly detection stages. The validation subset was used for threshold selection and intermediate tuning, while the test subset was strictly reserved for final performance evaluation.

Table III summarizes the main performance indicators obtained for the evaluated pipelines. In addition to accuracy, precision, and recall, the table reports the F1-score derived from precision and recall. Additional confusion matrices, ROC-AUC, PR-AUC, false positive rate (FPR), and false negative rate (FNR) values are available for the implemented experimental runs and will be systematically extended in future comparative experiments.

TABLE III. MAIN METRICS OBTAINED FOR THE EVALUATED ANOMALY DETECTION PIPELINES.

Pipeline	Accuracy	Precision	Recall	F1-score
AE+DBSCAN+IF	0.9803	0.0904	0.0082	0.0150
AE+IF	0.9347	0.0536	0.1542	0.0796
AE+DBSCAN	0.9670	0.2606	0.4376	0.3267

A. Pipeline: AE + DBSCAN + IF

After training, testing, and validating the AE+DBSCAN+IF pipeline, an accuracy of 0.9803 was achieved, followed by a precision of 0.0904 and a recall of 0.0082. This configuration exhibits extremely high accuracy and excellent control of false positives, indicating that the model is highly effective at modeling normal traffic behavior. However, this comes at the cost of a dramatic loss in sensitivity, as reflected by the extremely low recall (below 1%). This suggests that the model fails to detect the vast majority of intrusions. This behavior can be explained by the interaction between DBSCAN and Isolation Forest. DBSCAN, when applied to the latent space, constructs a very large and dense region of normality, effectively absorbing a significant portion of the data, including subtle anomalous patterns. As a consequence, many malicious samples are either classified as part of normal clusters or rejected as noise before reaching the final detection stage. Subsequently, Isolation Forest operates on a reduced and already filtered subset of data, with thresholds that are too strict to recover the missed anomalies. This results in a pipeline that is overly conservative, prioritizing specificity over sensitivity. Overall, while the pipeline achieves excellent false positive control, it is unsuitable for intrusion detection, where missing attacks is more critical than raising occasional false alarms.

B. Pipeline: AE + IF + DBSCAN

After training, testing, and validating the AE+IF+DBSCAN pipeline, an accuracy of 0.93 was achieved, followed by a precision of 0.05 and a recall of 0.15.

This configuration significantly improves recall compared to the previous one, demonstrating a better ability to detect anomalous traffic. By applying Isolation Forest before DBSCAN, the model first identifies suspicious samples based on global isolation properties in the latent space. This step ensures that anomalies are not prematurely absorbed into dense normal regions. DBSCAN is then applied only to these suspicious points, enabling the identification of coherent anomalous structures and improving the interpretability of the results. This ordering is more aligned with the nature of intrusion detection, where anomalies are often sparse and irregular rather than densely clustered. However, this improved sensitivity comes at the cost of an increased number of false positives, as reflected by the lower precision. The pipeline becomes more reactive but less selective. This configuration provides a better balance between detection and false alarm control than AE + DBSCAN + IF, but still requires careful threshold tuning and post-processing mechanisms to be operationally viable.

C. Pipeline: AE + DBSCAN

The AE + DBSCAN pipeline yields the most surprising and insightful results in this study, with an accuracy of 0.9670, a precision of 0.2606, and a recall of 0.4376. Contrary to expectations, removing Isolation Forest leads to a substantial improvement in recall and overall detection performance, with recall reaching approximately 44% and precision significantly higher than in the other configurations. This indicates that DBSCAN alone, when applied in the latent space learned by the Autoencoder, is capable of effectively distinguishing between normal and anomalous traffic patterns. The latent representation appears to preserve sufficient structure to allow density-based clustering to separate benign and malicious behaviors without requiring an additional anomaly isolation stage. The degradation observed when adding Isolation Forest suggests that, in the current configuration, IF introduces an overly restrictive filtering stage, removing or suppressing useful anomaly signals identified by DBSCAN. Instead of enhancing detection, it acts as a bottleneck that reduces the model's sensitivity. This result highlights a critical insight: additional model complexity does not necessarily improve performance, and in some cases, it can even degrade it. The AE + DBSCAN combination appears to provide a strong baseline for intrusion detection in this context.

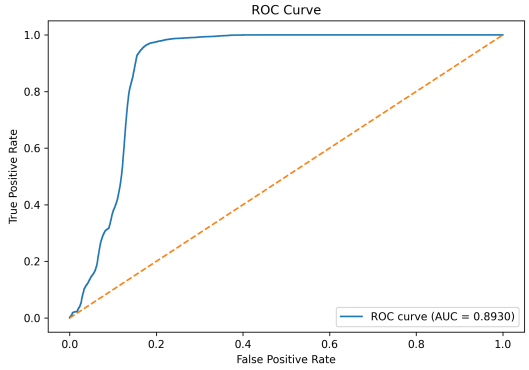


Fig. 6. Receiver Operating Characteristic (ROC) curve showing the true positive rate as a function of the false positive rate.

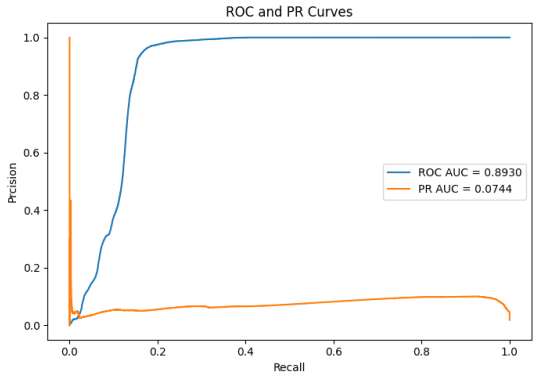


Fig. 7. Combined ROC and Precision–Recall curves used to evaluate the detection behavior under class imbalance.

The resulting figures provide a complementary and consistent view of the system's performance. The ROC curve (Fig. 6) shows only moderate overall separation between normal traffic and malicious traffic, indicating that the anomaly score partially distinguishes intrusions but without a clear boundary. This trend is confirmed by the Precision–Recall

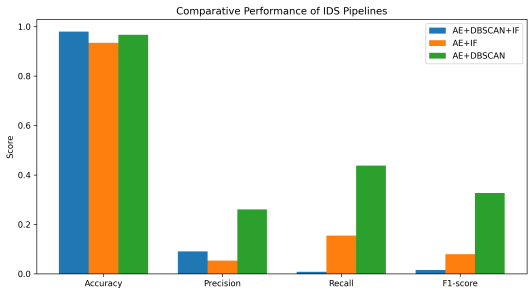


Fig. 8. Comparative performance of the evaluated IDS pipelines in terms of accuracy, precision, recall, and F1-score.

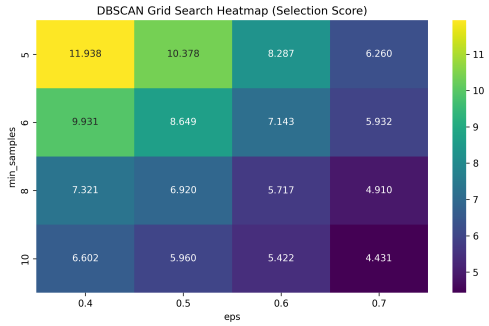


Fig. 9. Heatmap of the DBSCAN grid search showing the selection score for different values of ε and MinPts.

curve (Fig. 7), which is more relevant in a highly imbalanced IDS context. It highlights that an increase in sensitivity is quickly accompanied by a drop in precision, reflecting a delicate trade-off between detection and false positives. The comparative bar chart of the pipelines (Fig. 8) is particularly informative. It clearly shows that AE+DBSCAN currently offers the best balance between recall and precision, while AE+DBSCAN+IF primarily maximizes accuracy by becoming excessively conservative. This confirms that adding Isolation Forest, in its current configuration, degrades detection rather than improving it. Finally, the heatmap of the DBSCAN selection score (Fig. 9) shows that the clustering performance is highly dependent on the (eps, min_samples) pair, with a well defined optimal range. This figure underscores the importance of fine-tuning DBSCAN for the overall quality of the pipeline.

IV. Discussion

The results of this study provide several important insights into the design of hybrid intrusion detection systems based on representation learning and unsupervised anomaly detection. While combining multiple models is often assumed to improve performance, the ablation analysis clearly demonstrates that this assumption does not always hold. The most striking result is that the AE + DBSCAN configuration outperforms the full hybrid pipeline (AE + DBSCAN + IF) in terms of recall and overall detection capability. This suggests that the latent representation learned by the Autoencoder is already sufficiently structured to allow DBSCAN to effectively separate normal and anomalous behaviors. In this context, DBSCAN acts not only as a clustering algorithm but also as a powerful anomaly detector based on density estimation. The degradation observed when adding Isolation Forest highlights a critical issue: the introduction of an additional filtering stage can inadvertently remove meaningful anomaly signals. In the current setup, Isolation Forest appears to be too restrictive, either due to an insufficient contamination parameter or to thresholds that are not adapted to the distribution of the latent space. As a result, many anomalies identified by DBSCAN are suppressed, leading to a dramatic drop in recall.

More specifically, the ablation study indicates that Isolation Forest does not improve the detection capability in the current configuration. After DBSCAN has modeled the dense region of normal traffic, Isolation Forest operates on a reduced subset containing mostly borderline observations. Under these conditions, the anomaly thresholds become overly restrictive and many suspicious samples identified by DBSCAN are rejected, leading to a substantial decrease in recall. This behavior suggests that the current Isolation Forest configuration favors specificity at the expense of sensitivity. Another important factor is the behavior of DBSCAN itself. The results suggest that DBSCAN is modeling a very large region of normality, likely due to a relatively high epsilon value or insufficient cluster density constraints. This can lead to the absorption of borderline anomalies into normal clusters. However, despite this limitation, DBSCAN alone still achieves significantly better recall than the hybrid pipeline, indicating that the main performance bottleneck lies in the Isolation Forest stage rather than in the clustering step. A broader comparison with recent baseline methods such as One-Class Support Vector Machine, Local Outlier Factor, Random Forest, and recent deep learning-based IDS models is currently being conducted as part of an extended version of this work. This additional study will provide a more comprehensive positioning of the proposed approach against supervised and unsupervised IDS baselines.

From an operational perspective, the key challenge remains the trade-off between recall and false positives. While AE + DBSCAN provides strong detection capabilities, it may generate more false alarms. Several strategies can be implemented to address this issue without sacrificing recall:

- Threshold calibration: Instead of using fixed thresholds, adaptive thresholds based on quantiles or validation distributions can be used to dynamically adjust sensitivity.
- Temporal aggregation: Anomalies can be aggregated over time windows (e.g., per IP or session). Persistent anomalies are more likely to correspond to real attacks.
- Post-classification filtering: A lightweight supervised classifier can be applied to suspicious samples to filter false positives.
- Cluster validation: Clusters can be validated based on additional criteria such as stability, temporal coherence, or feature entropy.
- Explainability-based filtering: Feature importance or explainability methods (e.g., SHAP) can help identify whether an anomaly is meaningful or spurious.

V. Conclusion

This study investigated whether the ordering of DBSCAN and Isolation Forest affects intrusion detection performance on server log files. The results show that the ordering of unsupervised components has a substantial impact on detection quality. The AE+DBSCAN+IF pipeline achieved the highest accuracy (0.9803) but produced an extremely low recall (0.0082), making it unsuitable for practical IDS deployment. The AE+IF configuration improved sensitivity but remained less precise. The AE+DBSCAN pipeline achieved the best compromise, with an accuracy of 0.9670, a precision of 0.2606, and a recall of 0.4376. These findings answer the initial research question by showing that adding Isolation Forest after DBSCAN can degrade detection performance when thresholds are too restrictive. Future work will investigate adaptive thresholding, temporal aggregation, explainable anomaly filtering, and a broader comparison with recent IDS baselines.

Acknowledgment

We would like to thank the Laboratoire de Recherche en Science de l'Ingénieur (LARSI) at the University of Lomé, Togo, for providing us with the data used in this study.

References

[1] Alaa, A., et al.: A comprehensive survey on intrusion detection systems with advances in machine learning, deep learning and emerging cybersecurity challenges. Discover Artificial Intelligence (2025)

[2] Bundala, N.N.: Understanding cybercrime modus operandi: Techniques, psychological tricks, and countermeasures. Asian Journal of Research in Computer Science **17**(12), 234–251 (Dec 2024). https://doi.org/10.9734/ajrcos/2024/v17i12541, https://journalajrcos.com/index.php/AJRCOS/article/view/541

[3] Chen, D., Wawrzynski, P., Lv, Z.: Cyber security in smart cities: A review of deep learning-based applications and case studies. Sustainable Cities and Society **66**, 102655 (2021)

[4] Chen, Z., Liu, J., Gu, W., Su, Y., Lyu, M.R.: Experience report: Deep learning-based system log analysis for anomaly detection. arXiv preprint arXiv:2107.05908 (2021)

[5] Durgaraju, S., Vel, D.V.T., Madathala, H.: The evolution of cyber threats and defenses: A review of innovations and challenges. In: 2025 6th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI). pp. 117–123 (2025). https://doi.org/10.1109/ICMCSI64620.2025.10883230

[6] Elsayed, R.A., Hamada, R.A., Abdalla, M.I., Elsaid, S.A.: Securing iot and sdn systems using deep-learning based automatic intrusion detection. Ain Shams Engineering Journal **14**(10), 102211 (2023)

[7] Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. p. 226–231 (1996)

[8] Holm, H.: Signature based intrusion detection for zero-day attacks: (not) a closed chapter? In: 2014 47th Hawaii International Conference on System Sciences. pp. 4895–4904 (2014). https://doi.org/10.1109/HICSS.2014.600

[9] Kasongo, S.M.: A deep learning technique for intrusion detection system using a recurrent neural networks based framework. Computer Communications **199**, 113–125 (2023)

[10] Kazimierczak, M., Habib, N., Chan, J.H., Thanapattheerakul, T.: Impact of ai on the cyber kill chain: A systematic review. Heliyon **10**(24), e40699 (2024)

[11] Kumar, G.: Denial of service attacks–an updated perspective. Systems science & control engineering **4**(1), 285–294 (2016)

[12] Landauer, M., Onder, S., Skopik, F., Wurzenberger, M.: Deep learning for anomaly detection in log data: A survey. arXiv preprint arXiv:2207.03820 (2022)

[13] Liu, F.T., Ting, K., Zhou, Z.H.: Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery From Data - TKDD **6**, 1–39 (2012). https://doi.org/10.1145/2133360.2133363

[14] Maddikunta, P.K.R., Pham, Q.V., Prabadevi, B., Deepa, N., Dev, K., Gadekallu, T.R., et al.: Industry 5.0: A survey on enabling technologies and potential applications. Journal of Industrial Information Integration **26**, 100257 (2022)

[15] Sedki, I., Hamou-Lhadj, A., Ait-Mohamed, O.: Awsom-lp: An effective log parsing technique using pattern recognition and frequency analysis (2021), https://arxiv.org/abs/2110.15473

[16] Sedki, I., Hamou-Lhadj, A., Ait-Mohamed, O., Shehab, M.A.: An effective approach for parsing large log files. In: 2022 IEEE International Conference on Software Maintenance and Evolution (ICSME). pp. 1–12 (2022). https://doi.org/10.1109/ICSME55016.2022.00009

[17] Shivhare, I., Purohit, J., Jogani, V., Attari, S., Chandane, M.: Intrusion detection: A deep learning approach. arXiv preprint arXiv:2306.07601 (2023)

[18] Xu, L.D., Xu, E.L., Li, L.: Industry 4.0: state of the art and future trends. International Journal of Production Research **56**(8), 2941–2962 (2018)

[19] Xu, Z., Wu, Y., Wang, S., Gao, J., Qiu, T., Wang, Z., Wan, H., Zhao, X.: Deep learning-based intrusion detection systems: A survey. arXiv preprint arXiv:2504.07839 (2025)

[20] Yeh, W.C., Forghani-elahabad, M.: An efficient algorithm for sorting and duplicate elimination by using logarithmic prime numbers. Big Data and Cognitive Computing **8**(9) (2024). https://doi.org/10.3390/bdcc8090096, https://www.mdpi.com/2504-2289/8/9/96

[21] Zou, T., Bretas, A.S., Aljohani, N., Bretas, N.G.: Malicious data injection attacks: A relaxed physics model based strategy for real-time monitoring. In: 2019 North American Power Symposium (NAPS). pp. 1–6 (2019). https://doi.org/10.1109/NAPS46351.2019.9000397

Artificial Intelligence Guided Crowd Counting and Density Estimation with Enhanced CSRNet

Muhammad Jawad Babar[1], Malik Muhammad Saad Missen[2],
Hannan Adeel[3], Muhammad Usman[4], Muzammil Malik[5]
Faculty of Computing, The Islamia University of Bahawalpur, Pakistan[1,2]
Universiti Teknikal Malaysia Melaka, Malaysia[3]
Edge Hill University, UK[4]
Hamdard University, Islamabad, Pakistan[5]

***Abstract*—Estimating crowd size in dense environments remains a complex problem, yet it holds critical value for safety monitoring, city infrastructure design, and large-scale gathering coordination. Leveraging contemporary developments in neural network architectures and machine intelligence, researchers have markedly enhanced the precision of population counts derived from both still imagery and motion footage. This research focuses on the development of an enhanced deep neural network-based crowd counting model. Since the original CSRNet primarily relies on head detection for crowd estimation, an additional face detection module has been incorporated to improve its capability in scenarios where facial features are visible. The proposed enhancement increases the flexibility and estimation accuracy of CSRNet by integrating individual face detection with crowd density estimation. Furthermore, this study presents a comprehensive comparative analysis of the proposed enhanced CSRNet against three state-of-the-art crowd counting models, namely Bayesian Network (BAYNet), Distribution Matching (DM-Count), and Scale Aggregation Feature Attention Network (SFANet). The models are evaluated on multiple datasets to investigate their accuracy, robustness, adaptability to varying crowd densities, and performance under complex environmental conditions. The comparison also highlights the methodological differences and computational characteristics of these approaches. Model performance is assessed using the standard evaluation metrics of Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). Experimental results demonstrate that the proposed enhanced CSRNet achieves a competitive RMSE/MAE of $9.61/93.05$ on 64×64 images of the custom dataset, outperforming the baseline models and demonstrating its effectiveness for accurate crowd counting.**

***Keywords*—Crowd counting; machine learning; image processing; CSRNet; density estimation; deep learning; face detection**

I. Introduction

The problem of crowd counting and density estimation has emerged as a crucial field of study to solve many practical applications. The challenges arising from the growing urbanization and from the large-scale public events in the real world. The techniques are very important in the field of urban planning, public safety and resource allocation. Precise crowd estimation enables the authorities to regulate the number of people in a packed area, and this allows for prompt answer during the crisis and enables efficient event planning [1], [2], [3]. Crowd counting inherently connects to intelligent systems via cutting-edge algorithms and advanced technologies like computer vision and machine learning. These systems are essential for precisely identifying and enumerating people in dense environments, a task that relies heavily on ongoing methodological innovations. Intelligent systems are well-suited for analyzing visual information and using methods such as object detection, Convolutional Neural Networks (CNNs), and deep learning to accurately and efficiently estimate crowd size and size, providing insights for various applications including urban planning, security surveillance, and event management [4], [5].

Current methods are generally time-consuming, imprecise, and not suitable for larger-scale applications, such as manual counting or basic counting devices [6], [7], [8]. The traditional methods might not be as effective in real-world situations in which large and dense crowds are present, resulting in less accuracy. These restrictions have led to a demand for more complex, automated and precise solutions. The advanced neural network architectures used in deep learning methods have the potential to transform the field of crowd counting and density estimation [9], [10], [11], [12]. Several deep learning approaches such as Congested Scene Recognition Network (CSRnet), Scale Aggregation Feature Attention Network (SFANet), and Bayesian Network (BayNet) automatically acquire complex spatial features and patterns, enhancing the accuracy and scalability.

Recent advances have further pushed the boundaries of what is achievable in crowd analysis. Multi-scale feature fusion approaches [13], [14], perspective-assisted semi-supervised learning methods [15], and generative model-based density estimation techniques [16] have collectively demonstrated that combining richer contextual information with improved training paradigms can substantially close the performance gap that previously existed between controlled benchmark settings and challenging real-world deployments. Despite these advancements, accurately estimating crowd density under extreme occlusion, significant scale variation, and complex lighting conditions remains an open challenge, motivating the exploration of improved architectures such as the one proposed in this work.

The usually adopted approach for crowd counting research consists of the following steps [17], [1], [18], [19]:

- Collect primary visual data: images or video footage; captured on-site to obtain a ground-level count of individuals within each unit frame.
- Preprocessing: Noise filtering, resizing images and modifying pixels.

Fig. 1. Objects or people counting example.

Fig. 2. Overlapped persons in crowd image, shown in red box.

- Identifiable Feature Extraction: Obtaining identifiable features from the data set available (images or videos) to identify or count people.
- Algorithms used for counting: Algorithms for counting are selected based on the extracted features, with each offering varied approximations of per-frame pedestrian tallies. These approaches generally fall into regression, density estimation, or detection-oriented frameworks.
- Evaluation:The performance of the proposed crowd analysis approach is quantified using standard error metrics, namely mean absolute error (MAE) and root mean square error (RMSE), to ensure a thorough assessment of its predictive accuracy.
- Flaws in Counting/Extracting Data – Consider improving the flaws in counting or extracting data by analysing the data.
- Application: The algorithm finds practical use in crowd regulation, public safety protocols, tactical decision-making, and monitoring systems that operate on live data streams.

In this work, CSRNet was adopted and subsequently enhanced by integrating BayNet and SFANet to improve face and head-counting accuracy. The combined model was then evaluated to assess its overall performance. The main objective centered on leveraging facial and head count information within the CSRNet architecture to achieve reliable crowd counting and density estimation.

In Fig. 1 the density map is a collection of people, and each blue dot is an individual person.

A. Motivation

The motivation for this investigation stems from several interconnected considerations, which are outlined below.

1) Practical relevance and real-world deployment: Crowd counting plays a pivotal role in numerous practical domains, including public event management, traffic flow monitoring, and public safety assurance in congested urban environments. The deployment of accurate and reliable counting models offers tangible benefits in these scenarios. Nevertheless, the task remains inherently challenging due to pronounced variations in crowd density, severe occlusion (as illustrated in Fig. 2), perspective distortion, and fluctuating illumination conditions. These inherent challenges have motivated continued investigation into improving counting system precision and resilience, thereby supporting the wider evolution of computer vision techniques.

2) Societal implications and public welfare: Beyond its technical merits, crowd counting carries significant societal weight, particularly in safety-critical contexts such as emergency response coordination, large-scale public gathering management, and evidence-based urban planning. Accurate density estimation facilitates timely and informed decision-making, ultimately improving public safety outcomes and enabling more efficient resource allocation during mass events.

3) Advancements in deep learning architectures: The recent proliferation of sophisticated deep learning models; including CSRNet, SFANet, and BayNet; has substantially expanded the methodological landscape for crowd analysis. Thes architectures, representative of state-of-the-art developments in computer vision, offer new avenues for investigation. Researchers are consequently driven to systematically evaluate, extend, and refine these models, with particular emphasis on understanding their respective strengths, inherent limitations, and generalizability across diverse counting scenarios.

4) Benchmarking and comparative evaluation: Rigorous comparative analysis of existing deep learning approaches constitutes a fundamental component of empirical research in crowd counting. Systematic benchmarking not only establishes performance baselines but also elucidates the conditions under which specific models exhibit superior or inferior behavior. Such insights are indispensable for guiding future architectural innovations, informing model selection for practical applications, and fostering reproducible progress within the research community.

B. Aims and Objectives

This study sets out to enhance both the accuracy and efficiency of crowd counting by adopting an improved con-

volutional neural network: CSRNet. Specifically, we aim to evaluate how well the modified CSRNet architecture estimates crowd density across diverse real-world scenarios, with an eye toward informing crowd management, urban planning, and public safety measures. To that end, our work compares the performance of the enhanced CSRNet against state-of-the-art counting methods, assesses its generalizability across multiple benchmark datasets, and explores its practical viability in operational settings.

- A precise and computationally efficient crowd counting framework will be established, beginning with a comprehensive examination of extant deep learning literature. This survey aims to pinpoint leading methodologies, recognize their inherent shortcomings, and delineate evolving directions within the domain. Informed by these findings, a customized deep architecture will be conceived and deployed to tackle the principal obstacles in crowd enumeration.
- Tackle real-world challenges in crowd counting applications, particularly the issue of occlusion. The aim is to improve counting accuracy in partially obscured scenarios by leveraging contextual information and incorporating occlusion-aware features into the model.
- Evaluate and compare the proposed model against existing solutions using appropriate performance metrics. The assessment will focus on accuracy, precision, recall, and computational efficiency to ensure a comprehensive understanding of the model's strengths and limitations.

C. Our Contribution

The contribution of this work is twofold:

- The construction of compact libraries specifically designed for crowd counting facilitates computational economy, rendering them viable for implementation on hardware with limited resources and within time-sensitive applications. This strategy consequently broadens the practical utility of counting systems across varied real-world contexts.
- The proposed face detection module integrated on CSRNet is designed to detect the features of a face instead of head, and this feature imparts the ability to count faces, even in dense scenes, and can more effectively estimate the number of persons in the image.

The remainder of this study is organized as follows. Section II offers a broad overview of the relevant literature and describes the datasets employed for crowd counting and density estimation. In Section IV, we introduce the proposed model, explain how the dataset was constructed, and define the evaluation metrics used to assess performance. Section V reports the experimental findings, and Section X concludes the study by discussing potential avenues for future research.

II. Literature Review

Within the field of machine vision, crowd counting and density estimation have drawn growing interest from research groups across a wide range of international institutions. These tasks are regarded as significantly important in the analysis of surveillance videos and images, whether conducted in real-time or offline modes. Several established and emerging algorithms have been applied to estimate crowd sizes from images or video, yet many fall short of expected performance, largely owing to the wide variability present across different data samples. To address this problem, machine learning models, including CNNs and their variants, have been introduced.The primary challenges and core tasks associated with crowd counting and density estimation in computer vision are outlined here.

Fig. 3. An example of an annotated crowd image. The plus (+) sign identifies each person.

The central objective of crowd counting is to deliver a reliable enumeration of individuals within a gathering, supporting applications in event coordination, oversight, and public safety. The majority of existing approaches rely on machine learning techniques for visual data processing, enabling both detection and tracking, with subsequent statistical procedures applied to derive aggregate headcounts from the captured footage [9], [10], [11], [12].

III. Crowd Counting: An Overview

Estimating the number of people in a crowd is one of the basic problems of computer vision, such as in the work by Sacchi in 2001 ([20]), Lin in 2001 ([21]), and Ranjan in 2018 ([22]). This field involves solving problems arising from varying environments, densities, obstacles, and crowd sizes, through machine learning and advanced algorithms. The crowd counting problem is one of the important problems in the city area planning, event management, and safety issues based on visual data [23], [24]. The primary focus is on creating more sophisticated models capable of automatically, accurately and rapidly counting individuals in dense environments, such as for resource allocation or decision making. The picture shown in Fig. 3 is one of a crowd and each person in the picture is marked by a plus (+) sign.

The history of crowd counting estimation methods is a history of manual methods, which are human-based. During the late 20th century, the technology of computer vision started to get advanced, which led to the development of automated

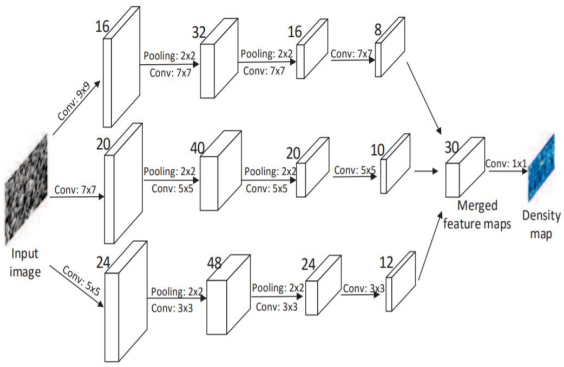


Fig. 4. An illustration of the proposed CNN-driven framework for crowd enumeration, as introduced in [18].

methods for crowd counting. The traditional approaches were based on feature engineering and density estimation. In recent years, crowd counting is becoming an increasingly popular task, and, thanks to the development of deep learning, in particular of the CNN, it has been transformed in the last decade (cf. [25], [26], [27]). These models utilize highly annotated datasets to train, enhance accuracy and flexibility in various deployment contexts. The transformation is a testament to the shift from traditional, human-based methods to the latest in AI-driven solutions, marking an important advancement in the field of crowd analysis, urban planning, and security solutions.

The crowd counting process using machine learning is carried out based on the model and pipeline presented in Fig. 4.

CNNs are often used as crowd counters because they are good at extracting features from the data and they can learn and represent complex data patterns. People count is challenging because it is necessary to estimate a large number of objects with substantial variation in appearance because of occlusion, scale and perspective. The main reason why CNN is being used for crowd counting as well as crowd density estimation is that it has the capability to learn the nonlinear relationships present in the data.

A. Crowd Counting at Image Level

The section provides comprehensive overview of the existing literature on crowd counting from images of overcrowded scenes.

Recent studies [7], [8], [28] have introduced deep CNN-based regression architectures, which have demonstrated superior performance over earlier approaches relying on manually engineered features or supplementary detection modules. Extensive experimental evaluations were conducted using widely adopted benchmark datasets, namely UCSD [29], UCF-CC [30], PASCAL VOC 2012 [31], and PETS-2009 [30]. The performance was measured based on the mean and variance of the absolute differences. A CNN-based architecture called CrowdNet was proposed in [32], which has been recognized as a notable advancement in the field. This architecture integrates deep and shallow fully convolutional branches to jointly extract high-level semantic information and fine-grained spatial features, both of which are considered essential for reliable counting in the presence of significant scale variability. The evaluation on the challenging UCF-CC 50 dataset demonstrated improved performance over prior art, with detailed outcomes presented in Table I.

TABLE I. A QUANTITATIVE EVALUATION OF THE METHOD PRESENTED IN [32] WAS CONDUCTED IN COMPARISON WITH CONTEMPORARY STATE-OF-THE-ART TECHNIQUES, USING THE UCF-CC 50 BENCHMARK

Method	Mean Absolute Error (MAE)
Density-aware Detection [33]	652.5
FHSc	467.5
Learning to Count [34]	490.2
Density-aware Detection [35]	468.2
Proposed [32]	423.7

Likewise, the work of [36] presented an innovative method to count people in surveillance video with a deep DCNN, which was also applied to this task. The unique factor in their approach is that they do not follow a conventional approach such as those presented in [37], [38], [39] which only use the number of pedestrians crossing a line that has been predetermined. Unlike the case of static pedestrians, the authors consider the exact time that a pedestrian takes to traverse the Line of Interest (LOI), which greatly improves CNN in learning more representative and key feature representations. Localization of Regions of Interest (ROIs) in video or image is a critical aspect for many crowd monitoring related problems, such as regression problems. This idea has been explored in the studies of [40] and [41], where a variant of Recurrent Neural Networks (RNNs), namely Long Short-Term Memory (LSTM), is utilized to capture sequential information and temporal dependencies during model training. To ensure a reliable correspondence between image content and the associated density map, irrespective of image size or resolution, previous research has proposed a simplified Multi-column Convolutional Neural Network (MCNN) architecture (Zhang et al., 2016; Zeng et al., 2017; Babu et al., 2017; Zhu et al., 2020; Hu et al., 2020).

A quality-based density estimation network, termed the Contextual Pyramid CNN (CP-CNN), is proposed for images with dense crowded scenes. This architecture operates through four core components: the Global Context Estimator (GCE), the Local Context Estimator (LCE), the Density Map Estimator (DME), and the Fusion-CNN (F-CNN), each constructed around the use of both broad and fine-grained contextual cues to support crowd enumeration. A deep and recursive network, referred to as DRResNET, was employed for crowd counting in [42]. An Adaptive Counting Convolutional Neural Network (A-CCNN) was proposed in [43] to enhance the accuracy of crowd counting, particularly for estimating crowd size. Loss functions are recognized as a key component in the evaluation of algorithmic performance. A detailed explanation of the composition of loss functions used for density map estimation in crowded scenes is provided in [44] and [45].

Persons counting in cross-scene problems is challenging since the background changes during model training. This is effectively solved by [51] who use a cross-scene counting

TABLE II. A STATISTICAL OVERVIEW OF THE BENCHMARK DATASETS UTILIZED IN CROWD COUNTING AND CROWD DENSITY ESTIMATION TASKS

Dataset	No. **of images**	Average **Count**	Source **of images**	Avg. size
UCF-QNRF [45]	1535	50	FLICKR, Web Search	2500×187
Mall [46]	2000	24	Public places	—
JHU-crowd [47]	4, 370	110	FLICKR, ShutterStock	1000×750
ShanghaiTech A [48]	475	490	Shanghai Streets	584×864
ShanghaiTech B	720	390	Shanghai Streets	584×864
WorldExpo'10 [49]	103	473	Shanghai 2010 WorldExpo	589×86
UCF-CC 50 [30]	50	1279	FLICKR	984×648
NWPU-Crowd [50]	5, 109	418	Resort, Walking street	2191×3209
UCSD [29]	1992	45	UC San Diego	157×237

model to learn other scenes knowledge transfer, through the features extracted from manually segment foregrounds and the domain adaptation method of extreme learning machine (DA-ELM).

Likewise, the scale variation problem was tackled in the study [52] by leveraging image pyramids to facilitate the prediction across different scales, and thereby taking advantage of adaptively varying weights per pixel. Crowd counting also has been explored using Generative Adversarial Networks (GANs) [53], [54], [55], [56]. In order to overcome the most common problems about scale and perspective changes, the authors of [57] proposed a Scale-Adaptive CNN (SaCNN) model. The authors in [58] followed CNN based methods and removed fully connected layers from the VGG16 model [59] and used pyramid networks for more consistent representations. In [60], [61], a deep learning method called Congested Scene Recognition (CSRNet) was used to understand the images with dense crowds and generate reasonable density maps. The proposed system consists of two main parts: The first part consists of CNN for 2D feature extraction, and the second part is dilated-layer CNN for expanding the receptive field without using the traditional pooling layers. In order to produce an accurate density map in a single step, the authors in [22] proposed a two-stage CNN model, first generating a rough outline of the density map and then refining the outline to obtain a high-resolution density map.

Recently, the authors in [62] introduced a weakly supervised framework for crowd counting, named JCTNet (Joint CNN and Transformer Network). The proposed architecture integrates three key components: the CNN Feature Extraction Module (CFM), the Transformer Feature Extraction Module (TFM), and the Counting Regression Module (CRM). The semantic crowd features are extracted by the CFM, while the extracted feature maps are partitioned into patches by the TFM so that global contextual information can be captured using a Transformer. Finally, the crowd count is regressed by the CRM from the fused representations. The proposed JCTNet aims to fuse CNN-derived local representations with Transformer-driven global contextual encoding, leveraging the respective advantages of each architectural paradigm. While JCTNet achieves a considerable reduction in parameter count; roughly 67% to 73% less than a pure Transformer counterpart; its computational cost remains non-trivial due to the concurrent operation of two backbone networks. In a separate contribution, Zhang et al. [63] proposed a cross-domain crowd counting architecture augmented with frequency-based processing, comprising a module for feature refinement in the spectral domain alongside an adapter designed to preserve domain-agnostic frequency characteristics. In this design, the Discrete Cosine Transform is utilized by the enhancement module to derive informative frequency statistics, thereby reinforcing the representational capacity of spatial features. Concurrently, the adapter incorporated Fast Fourier Transform (FFT)-based amplitude suppression in conjunction with an attention mechanism, which served to suppress domain-specific frequency cues and promote the acquisition of domain-invariant feature representations. To further mitigate complex background interference, a Parallel Dual Attention Module is incorporated, by which foreground crowd regions are emphasized. Moreover, decoder features from multiple scales are fused by a Multi-Scale Feature Aggregation Module, thereby improving robustness against large variations in crowd density and object scale. Superior counting accuracy is demonstrated by experimental evaluations on four public crowd counting datasets, particularly in challenging cross-domain scenarios, to be achieved by the proposed approach. Nevertheless, several additional processing modules are introduced by the framework, making it dependent on increased architectural complexity and computational overhead.

B. Crowd Counting at Video Level

The application of convolutional neural networks (CNNs) to crowd counting was first established in the pioneering work of Zhang et al. [6], who introduced an early deep learning framework capable of addressing the challenges associated with estimating crowd density across diverse scenes. Their method addressed the challenge of estimating crowd counts in previously unseen surveillance environments without requiring extensive scene-specific annotations. The proposed deep CNN was jointly optimized for two complementary objectives:

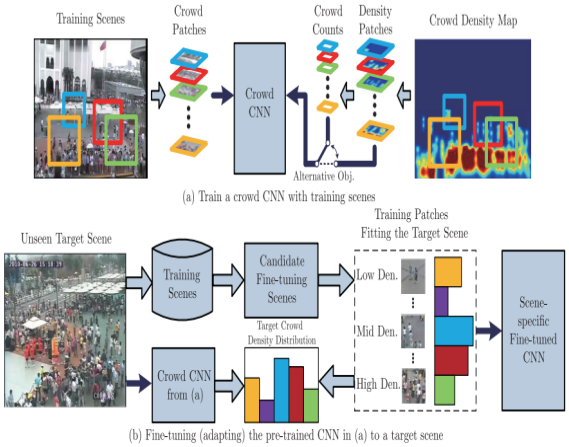


Fig. 5. Illustration of the model used for crowd counting in [6].

crowd density estimation and crowd count prediction, enabling improved generalization across diverse scenes. Validation of the proposed method was carried out across multiple standard benchmarks, specifically the WorldExpo'10, the UCSD Pedestrian Dataset [29], and the UCF-CC-50 Dataset, with results indicating strong generalizability in cross-scene counting tasks. An illustration of the overall system design is included in Fig. 5.

C. Datasets and Evaluation Parameters

The following section offers a comprehensive overview of standard benchmark datasets frequently adopted in crowd counting and density estimation studies, alongside the evaluation metrics used to measure model efficacy. A summary of key statistical attributes for the most widely referenced datasets is presented in Table II.

D. Recent Advances in Crowd Counting

The field of crowd counting and density estimation has seen considerable progress over the past few years, largely fueled by advances in architectural designs, objective function refinements, and optimization protocols. A summary of the most relevant recent contributions is provided below.

1) Multi-scale and attention-based methods: Rohra et al. [13] introduced MSFFNet to tackle the difficulties associated with generating reliable density maps under conditions of cluttered backgrounds and substantial scale variability. The proposed model incorporates a grouped feature extractor, a fusion block, and a decoder module within a convolutional architecture, enabling adaptive integration of encoded representations instead of depending on isolated extraction pathways. This design broadens the diversity of receptive fields while simultaneously lowering the overall computational overhead. Evaluated on the ShanghaiTech and UCF-QNRF benchmarks, MSFFNet demonstrates competitive performance across both sparse and highly congested scenes.

2) Semi-supervised and perspective-aware methods: Qian et al. [15] presented a prototype-based semi-supervised crowd counting approach based on perspective-assisted learning. A novel pseudo-label assigner for image regions is achieved by pseudo-labeling based on perspective self-organization without extra annotations. This feature space clustering and self-organization are based on prototype-based learning and clusters patches with similar level of density features values under similar levels of perspective distortion. This approach greatly alleviates the reliance on fully-annotated training examples, a major practical hurdle.

3) Diffusion model-based density estimation: Ranasinghe et al. [16] introduced CrowdDiff, which employs a denoising diffusion probabilistic model for multi-hypothesis crowd density estimation. Distinct from conventional CNN architectures that yield a singular, fixed density map; the CrowdDiff leverages the probabilistic behavior inherent to diffusion models, producing a range of density map variations that reflect the underlying uncertainty in crowd configurations. Such a strategy bolsters resilience under severe occlusion and heavy congestion, while delivering competitive results across multiple established crowd analysis datasets.

4) Transformer-based counting: Recent work has demonstrated the efficacy of Vision Transformers (ViTs) and Swin Transformers for crowd counting due to their ability to capture long-range spatial dependencies. Wang et al. [64] observed that transformer-based models consistently outperform pure CNN architectures on large-scale datasets, particularly in handling extreme scale variations and perspective distortions. However, these models impose significant computational overhead, which limits their deployment in resource-constrained environments.

5) Curriculum learning for dense crowds: Fotia et al. [65] put forward a curriculum-based training strategy tailored for crowd counting under high-density conditions. Their framework frames the task as point-wise localization of head positions, with a learning progression that moves from elementary to advanced counting instances. This design mirrors the stepwise nature of human skill acquisition and yields notable accuracy gains, particularly in overcrowded scenes.

6) Enhanced YOLOv8 for dense scenarios: Yan et al. [66] presented Yolov8crowd, a modified version of YOLOv8 incorporating both a revised loss function, MPDIoU, and an attention mechanism, SEAttention. The former enhances discriminative power among crowded and overlapping boxes through minimization of centroid deviations between predicted and true annotations. The latter recalibrates channel-level activations to emphasize high-informative features. Evaluation on dense benchmarks confirms that this variant surpasses conventional detectors in performance.

7) Comprehensive survey insights: Wang et al. [64] offered an extensive overview of crowd density estimation and counting methodologies, organizing the existing techniques into two broad categories: conventional approaches and those grounded in deep learning. It was highlighted in the survey that traditional approaches are consistently outperformed by deep learning models when handling cross-scene variations, multi-scale variations, severe occlusions, and perspective distortions. Several key open challenges were also identified, including real-time processing efficiency, generalization to unseen environments, and privacy-preserving crowd analysis.

E. Research Gap

Despite the remarkable progress in crowd counting and density estimation, several significant research gaps remain unaddressed. First, existing CNN-based architectures, including the original CSRNet, predominantly rely on *head detection* as the primary cue for crowd counting. However, in dense and partially occluded scenes, head regions are frequently indistinguishable, leading to systematic undercounting. The incorporation of *facial feature detection* as a complementary modality for counting has received limited attention, despite the greater discriminability of facial landmarks in moderately dense scenes. Second, while recent semi-supervised [15] and diffusion-based [16] methods have partially addressed the scarcity of annotated training data, the majority of existing approaches require large, fully annotated datasets collected under controlled conditions, which differ significantly from real-world crowd distributions. Third, multi-scale feature fusion networks such as MSFFNet [13] have demonstrated improvements in density map quality, yet the integration of spatial pyramid attention mechanisms with lightweight architectures capable of real-time deployment remains an under-explored area. Fourth, existing benchmarks primarily evaluate models on standard resolutions, and performance degradation at low resolutions (e.g., 64×64) is rarely studied, despite its relevance for low-quality surveillance footage. This work directly addresses the first and fourth gaps by augmenting CSRNet with a face detection module and conducting systematic evaluation across multiple image resolutions.

IV. Methods and Techniques

A. Overview of the Proposed Framework

The proposed framework enhances the standard CSRNet architecture with a dedicated face detection module, thereby enabling the model to exploit both head and facial cues for crowd counting. The overall pipeline consists of the following stages: 1) image acquisition and preprocessing, 2) multi-scale feature extraction via the modified CSRNet backbone, 3) face detection using a lightweight auxiliary network, 4) density map generation via dilated convolutional layers, and 5) final count estimation by integrating the density map with face detection outputs.

B. Baseline Models

The following baseline models are used for comparison in this study.

1) BayNet (Bayesian loss for crowd counting): BayNet [67] reframes crowd counting within a Bayesian inference paradigm, interpreting the estimated density map as a probabilistic representation of individual spatial locations. Instead of relying on a pixel-wise ℓ_2 penalty between predictions and ground-truth densities, BayNet adopts a Bayesian loss [Eq. (1)] [67] that offers greater resilience to labeling inaccuracies and positional ambiguity. This loss is formally expressed as:

$$\mathcal{L}_{\text{Bay}} = \sum_{i=1}^{N} \left(1 - \sum_{\mathbf{p}} F(\mathbf{p}) \cdot \mathcal{N}(\mathbf{p}; \mathbf{x}_i, \sigma_i^2 \mathbf{I}) \right)^2, \quad (1)$$

where, $F(\mathbf{p})$ is the predicted density at pixel $\mathbf{p}$, $\mathbf{x}_i$ is the annotated location of the i-th person, and $\mathcal{N}(\mathbf{p}; \mathbf{x}_i, \sigma_i^2 \mathbf{I})$ is a Gaussian kernel centered at $\mathbf{x}_i$.

2) DM-Count (Distribution Matching Counting): In [68], DM-Count tackles crowd counting via an optimal transport formulation, aiming to reduce the divergence between the estimated density map and the actual point-based ground truth. The approach incorporates entropic regularization into the Wasserstein metric [69], expressed in Eq. (2):

$$\mathcal{W}_\varepsilon(\mu, \nu) = \min_{\gamma \in \Pi(\mu,\nu)} \sum_{\mathbf{p},\mathbf{q}} \gamma(\mathbf{p}, \mathbf{q}) \cdot c(\mathbf{p}, \mathbf{q}) + \varepsilon \cdot \text{KL}(\gamma \| \mu \otimes \nu), \quad (2)$$

where, μ is the predicted density, ν is the ground-truth density, $c(\mathbf{p}, \mathbf{q})$ is the cost function (typically Euclidean distance), and ε is the regularization parameter.

3) SFANet (Scale-Aggregation Feature Attention Network): SFANet [70] employs a dual-path encoder to capture multi-scale features and uses attention mechanisms to selectively emphasize informative spatial regions. The attention module [71] computes [see Eq.(3)] a spatial attention map $\mathbf{A} \in \mathbb{R}^{H \times W}$ as:

$$\mathbf{A} = \sigma \left(\mathbf{W}_a \cdot \text{concat}[\text{AvgPool}(\mathbf{F}), \text{MaxPool}(\mathbf{F})] \right), \quad (3)$$

where, $\mathbf{F}$ is the feature map, $\mathbf{W}_a$ is a learned weight matrix, and σ denotes the sigmoid activation. The attended feature map is then $\hat{\mathbf{F}} = \mathbf{A} \odot \mathbf{F}$.

C. CSRNet architecture

CSRNet [60] employs the first ten convolutional layers of VGG-16 as its front-end feature extractor, followed by dilated convolutional layers as the back-end for density map generation. The dilated convolution with dilation rate d [72], is defined as in Eq. (4), below:

$$(F \star_d k)(\mathbf{p}) = \sum_{\mathbf{s}+d\mathbf{t}=\mathbf{p}} F(\mathbf{s}) \cdot k(\mathbf{t}), \quad (4)$$

where, F is the feature map, k is the convolutional kernel, and d is the dilation rate. Using different dilation rates in parallel captures features at multiple scales without additional pooling, thereby preserving spatial resolution.

The density map $\hat{D}$ is generated by applying these dilated convolutional layers to the extracted feature maps, and the estimated count $\hat{C}$ [72] is obtained by integrating the density map as shown in Eq. (5):

$$\hat{C} = \sum_{\mathbf{p} \in \Omega} \hat{D}(\mathbf{p}), \quad (5)$$

where, Ω denotes the spatial domain of the image.

D. Enhanced CSRNet with face detection module

The primary modification introduced in this work is the addition of a lightweight face detection branch to the CSRNet framework. While the original CSRNet relies solely on head annotations to generate density maps via Gaussian kernels, the proposed enhancement incorporates face detection to provide an additional modality for crowd counting, particularly in moderately dense scenes where faces are more discriminable than head outlines.

1) Gaussian kernel-based ground truth generation: The ground-truth density map $D(\mathbf{p})$, as reported in [73], is produced by convolving the annotated point locations with a Gaussian kernel, as shown in Eq. (6):

$$D(\mathbf{p}) = \sum_{i=1}^{N} \mathcal{N}(\mathbf{p}; \mathbf{x}_i, \sigma_i^2), \tag{6}$$

where, $\mathbf{x}_i$ is the annotated head or face center and σ_i is the spread parameter, which is adaptively estimated based on the average distance to the k-nearest annotated neighbors. The formal method is shown in Eq.(7).

$$\sigma_i = \beta \cdot \frac{1}{k} \sum_{j=1}^{k} \|\mathbf{x}_i - \mathbf{x}_{i,j}\|_2, \tag{7}$$

where, $\mathbf{x}_{i,j}$ is the j-th nearest neighbor of $\mathbf{x}_i$ and β is a scaling factor (typically $\beta = 0.3$).

2) Face detection branch: The face detection branch is a shallow convolutional network appended to the intermediate feature maps of the CSRNet front-end. It produces a face heatmap $H_{\text{face}}(\mathbf{p}) \in [0, 1]$ indicating the probability of a face center at each spatial location $\mathbf{p}$. The face count estimate is obtained formally as depicted in Eq.(8).

$$\hat{C}_{\text{face}} = \sum_{\mathbf{p} \in \Omega} H_{\text{face}}(\mathbf{p}). \tag{8}$$

The final crowd count estimate is a weighted combination of the density-map-based estimate and the face detection estimate, formally shown in Eq. (9).

$$\hat{C}_{\text{final}} = \lambda \cdot \hat{C} + (1 - \lambda) \cdot \hat{C}_{\text{face}}, \tag{9}$$

where, the fusion weight $\lambda \in [0, 1]$ is determined empirically on a validation set.

3) Training objective: At the end, Eq. (10) computes the the combined training loss for the enhanced CSRNet.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{density}} + \alpha \cdot \mathcal{L}_{\text{face}}, \tag{10}$$

where, $\mathcal{L}_{\text{density}} = \|\hat{D} - D\|_2^2$ is the pixel-wise ℓ_2 loss between the predicted and ground-truth density maps, $\mathcal{L}_{\text{face}} = \|H_{\text{face}} - H_{\text{face}}^*\|_2^2$ is the face heatmap loss, and α is a balancing hyperparameter.

E. Evaluation Metrics

The performance of all models is evaluated using the following standard metrics, MAE [Eq. (11)] and RMS[Eq. (12)].

Mean Absolute Error (MAE): [74]

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} \left| \hat{C}_i - C_i \right|, \tag{11}$$

where, $\hat{C}_i$ and C_i are the predicted and ground-truth counts for the i-th image, and N is the total number of test images.

Root Mean Squared Error (RMSE): [74]

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(\hat{C}_i - C_i \right)^2} \tag{12}$$

Lower MAE and RMSE values indicate better counting accuracy. RMSE is more sensitive to large errors and thus penalizes significant miscounts more heavily.

V. Experimental Analysis and Results

A. Experimental Setup

The experimental configuration and implementation specifics are outlined below. All trials were carried out using the PyTorch library on a machine equipped with an NVIDIA GPU. The Adam optimizer served as the training algorithm, initialized with a learning rate of 1×10^{-5} and a weight decay factor of 5×10^{-4} . To enhance generalization capabilities, the training routine integrated several data augmentation techniques, namely random cropping, horizontal flipping, and color jitter adjustments.

B. Comparison Table of Existing Methods

Table III provides a comparative overview of representative crowd counting methods from the literature, highlighting their key techniques, datasets, strengths, and limitations.

C. Results at Full Resolution

Table IV reports the estimation errors, measured in RMSE and MAE, for various models evaluated on ShanghaiTech part-A (SH-A), ShanghaiTech part-B (SH-B), and a proprietary image collection (Our images).

Table IV summarizes the RMSE and MAE estimation errors across models on the three evaluated datasets. On SH-A, CSRNet yields the largest deviations with values of 75.44 and 113.55 for MAE and RMSE respectively, whereas DM-Count attains the most favorable performance, registering the lowest figures at 53.39 for MAE and 99.30 for RMSE. In the case of SH-B, the performance of the BayNet model is much improved (RMSE/MAE=6.54/11.84), and the SFANet model is also competitive (6.57/11.52). Our modified CSRNet shows the highest accuracy on our custom images dataset together with SH-A and SH-B, with an RMSE/MAE of 59.07/101.29, which is better than any other model on both error metric.

TABLE III. COMPARATIVE ANALYSIS OF REPRESENTATIVE CROWD COUNTING AND DENSITY ESTIMATION METHODS

Method	Year	Dataset	Technique	Strengths	Limitations
MCNN [48]	2016	ShanghaiTech, UCF-CC-50	Multi-column CNN with variable filters	Handles scale variation; no fixed resolution required	High parameter count; limited accuracy in extremely dense scenes
CSRNet [60]	2018	ShanghaiTech, UCF-QNRF, WorldExpo'10	VGG-16 front-end + dilated CNN back-end	Preserves spatial resolution; strong density map quality	No temporal modeling; head-only detection
BayNet [67]	2019	UCF-QNRF, ShanghaiTech	Bayesian loss with point supervision	Robust to annotation noise; probabilistic formulation	Requires careful tuning of Gaussian bandwidth
DM-Count [68]	2020	ShanghaiTech, UCF-QNRF, NWPU	Optimal transport / distribution matching	Principled loss function; strong generalization	Higher computational cost; slow convergence
SFANet [70]	2019	ShanghaiTech, UCF-CC-50, UCSD	Dual-path encoder + spatial attention	Captures multi-scale features; attention improves localization	Complex architecture; not optimized for real-time
MSFFNet [13]	2025	ShanghaiTech, UCF-QNRF	Multi-scale feature fusion + semantic optimization	Adaptive feature fusion; reduced computation cost	Evaluated only on standard resolutions
PPSC [15]	2024	ShanghaiTech, NWPU, UCF-QNRF	Semi-supervised + perspective self-organization	Reduces need for full annotation; perspective-aware	Performance degrades under extreme occlusion
CrowdDiff [16]	2024	NWPU, ShanghaiTech	Denoising diffusion probabilistic model	Multi-hypothesis estimation; handles ambiguity	Slow inference due to diffusion sampling
YOLOv8 (Improved) [66]	2024	Custom dense datasets	YOLOv8 + MPDIoU + SEAttention	Real-time detection; improved overlapping target separation	Limited to detection-based counting; not density-map based
Ours (Enhanced CSRNet)	2024	SH-A, SH-B, Custom	CSRNet + face detection branch	Exploits facial features; improved accuracy at low resolutions	Fusion weight λ requires cross-validation

TABLE IV. ESTIMATION PERFORMANCE, MEASURED IN TERMS OF MAE AND RMSE, IS REPORTED FOR THE SHANGHAITECH DATASET (BOTH PART A AND PART B), AS WELL AS FOR THE FULL-RESOLUTION IMAGES FROM OUR OWN COLLECTED DATASET

Model name	SH-A	SH-B	Our images
BayNet	64.30 / 108.29	6.54 / 11.84	69.65 / 99.87
DM-Count	64.39 / 99.30	9.25 / 19.45	64.04 / 105.27
SFANet	64.84 / 89.57	6.57 / 11.52	62.10 / 100.02
CSRNet	75.44 / 113.55	10.27 / 19.32	60.58 / 102.90
Modified CSRNet (Ours)	**63.82 / 98.74**	**5.97 / 7.04**	**59.07 / 101.29**

TABLE V. THE QUANTITATIVE PERFORMANCE OF THE EVALUATED MODELS, EXPRESSED THROUGH MAE AND RMSE METRICS, IS REPORTED ACROSS THE SHANGHAITECH PART A AND PART B SUBSETS, IN ADDITION TO OUR CUSTOM IMAGE COLLECTION RESIZED TO A SPATIAL RESOLUTION OF 64×64 PIXELS

Model name	SH-A	SH-B	Our images
BayNet	60.47 / 102.12	7.25 / 10.07	68.43 / 89.35
DM-Count	62.87 / 96.79	5.86 / 12.99	63.58 / 104.29
SFANet	62.05 / 82.91	5.25 / 17.25	63.87 / 104.27
CSRNet	61.07 / 81.29	6.07 / 14.09	63.27 / 104.67
Modified CSRNet (Ours)	51.82 / 74.47	5.32 / 9.98	61.70 / 97.84

VI. RESULTS ON IMAGES OF SIZE 64×64

In the initial experimental setup, all crowd counting architectures; alongside the proposed variant of CSRNet; were assessed using input images down-sampled to a fixed resolution of 64×64 pixels. The outcomes from this evaluation are compiled in Table V.

Table V compiles the MAE and RMSE outcomes obtained from the ShanghaiTech Part-A (SH-A), ShanghaiTech Part-B (SH-B), and the introduced proprietary dataset, with all inputs uniformly resized to a spatial resolution of 64×64 pixels. Among all the evaluated methods, the modified CSRNet achieves the best performance on the custom dataset, attaining an MAE of 61.70 and an RMSE of 97.84, thereby outperforming the baseline crowd counting models. The improved performance is also reflected across the SH-A and SH-B datasets, demonstrating the robustness of the proposed approach under low-resolution conditions. This performance gain can be attributed to the integration of the face detection module, which enables the network to focus on more discriminative facial features even at reduced image resolutions. An example of a manually annotated crowd image from the custom dataset is presented in Fig. 6.

Fig. 6. The original crowd image showing a crowd of 80 people, annotated manually.

The results generated by applying the aforementioned models on one of the crowd images of ShanghaiTech dataset part A (SH-A) are shown in Fig. 7. Since the models count persons by detecting heads, the detected heads are shown in bounding boxes. Our modified CSRNet detects faces, as detecting faces rather than human heads in crowd counting provides greater accuracy by focusing on distinguishable facial features.

Fig. 7 depicts results of crowd counting when applied on images of size 64×64. Fig. 8 shows counting results using density estimation on additional images from the dataset.

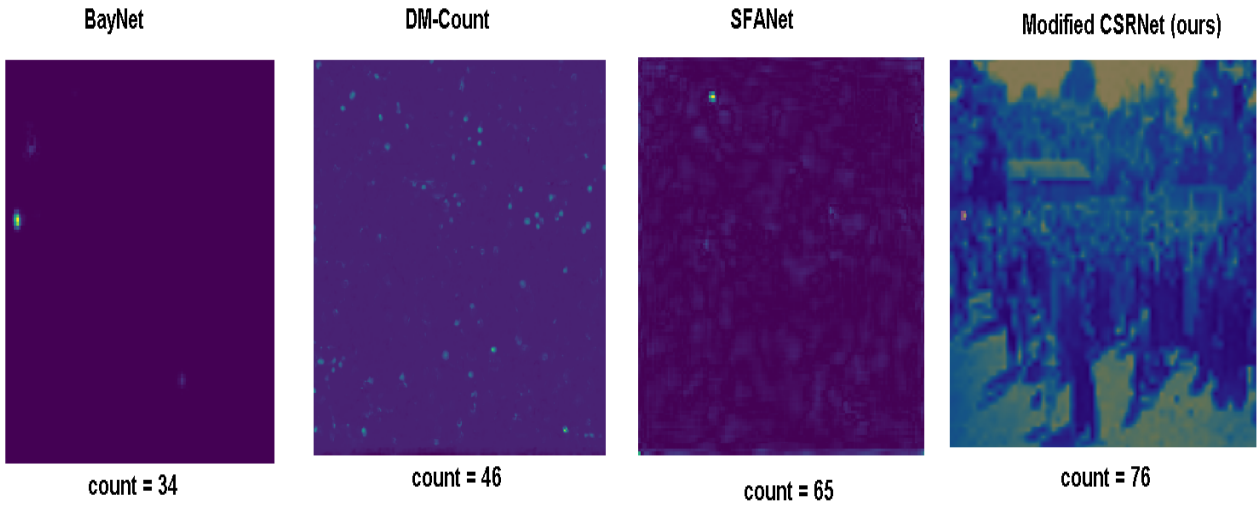


Fig. 7. Comparison of density estimation outcomes produced by our proposed model when applied to 64×64 image patches drawn from the ShanghaiTech Dataset, across both Part A and Part B.

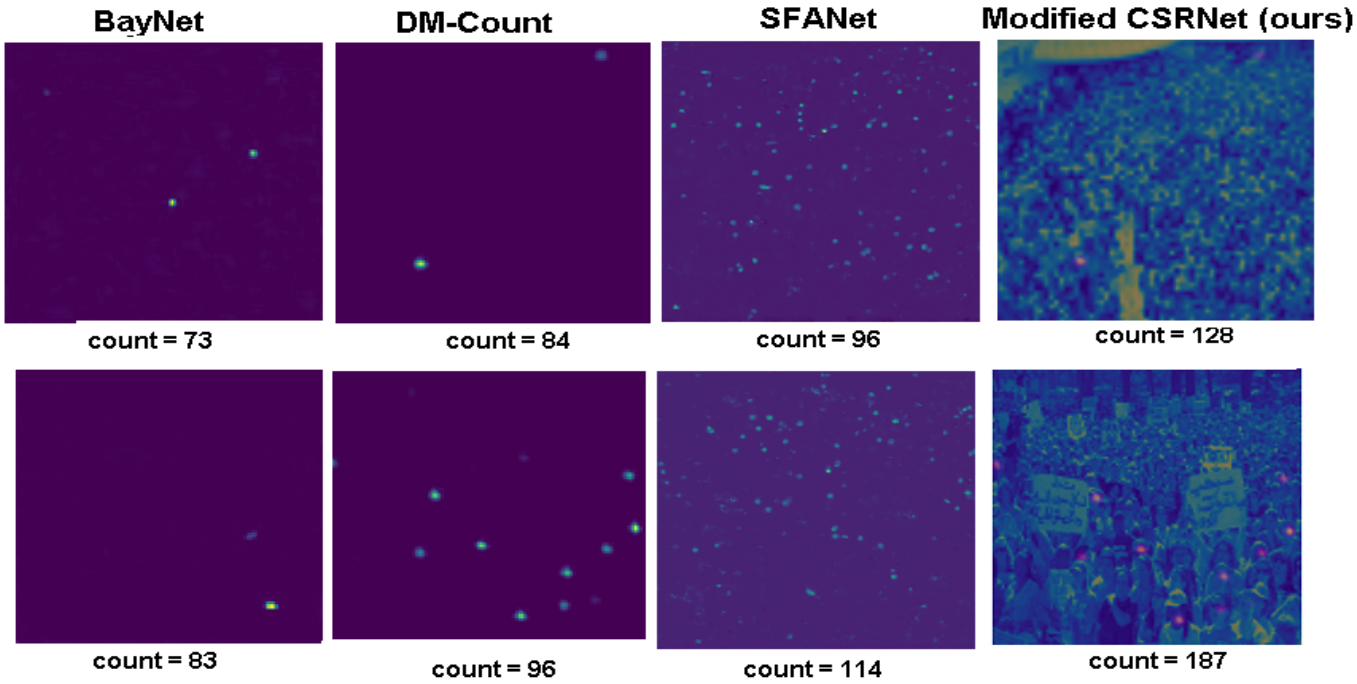


Fig. 8. Comparison of the density estimation results with our proposed model on 64×64 images of Shanghai Dataset part-A and part-B.

VII. Results on Images of Size 128×128 and 512×512

As it can be seen in Fig. 9 the proposed model was superior in the counting of people within the image on the crowd image with size 128×128.

In a similar way, as the image is resized to 256 and 512 the accuracy improves as shown in Fig. 10. The numbers (counting results) displayed at the bottom of each image also indicate that our proposed methodology substantially outperforms other state-of-the-art models.

VIII. Results on Images of Our Dataset

The performance evaluation of BayNet, DM-Count, SFANet, and the proposed modified CSRNet is presented in the Results section. This evaluation was conducted on the ShanghaiTech Part-A (SH-A), ShanghaiTech Part-B (SH-B), and the custom crowd-counting dataset, which was collected from publicly available internet sources. The performance of each model was assessed using the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) metrics. On SH-A, RMSE/MAE values of 61.29/100.07 were achieved by BayNet; on SH-B, values of 84.34/111.07 were obtained; and on the custom dataset, values of 69.65/99.87 were recorded. By

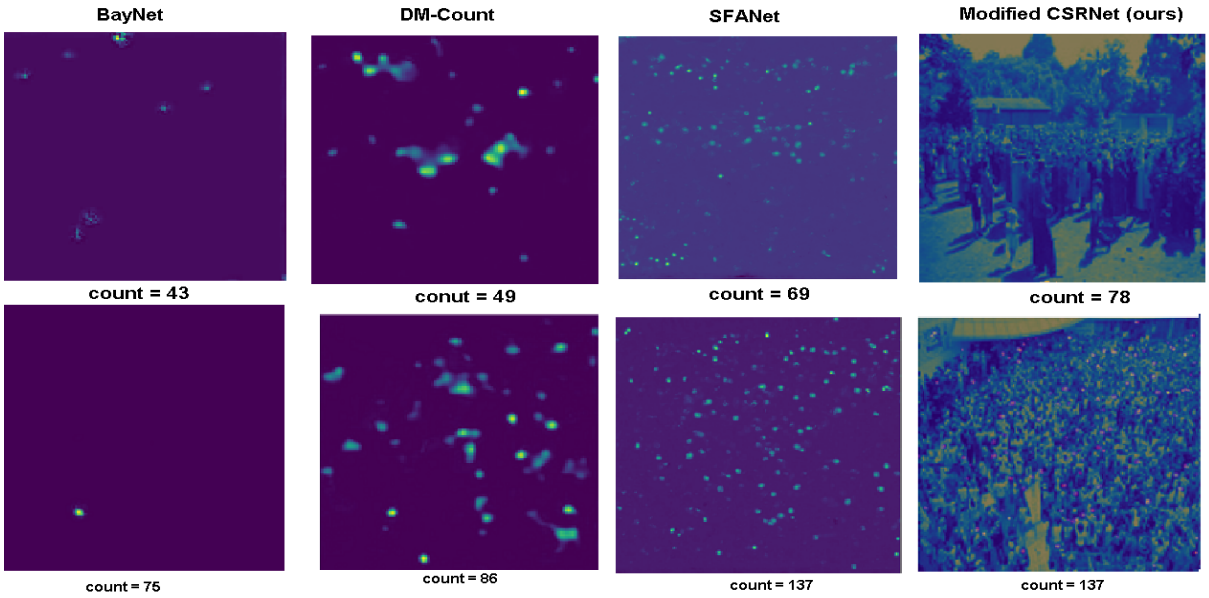


Fig. 9. Comparison of density estimation outcomes produced by our proposed model on 64 × 64 image patches from the ShanghaiTech Dataset (Part A and Part B).

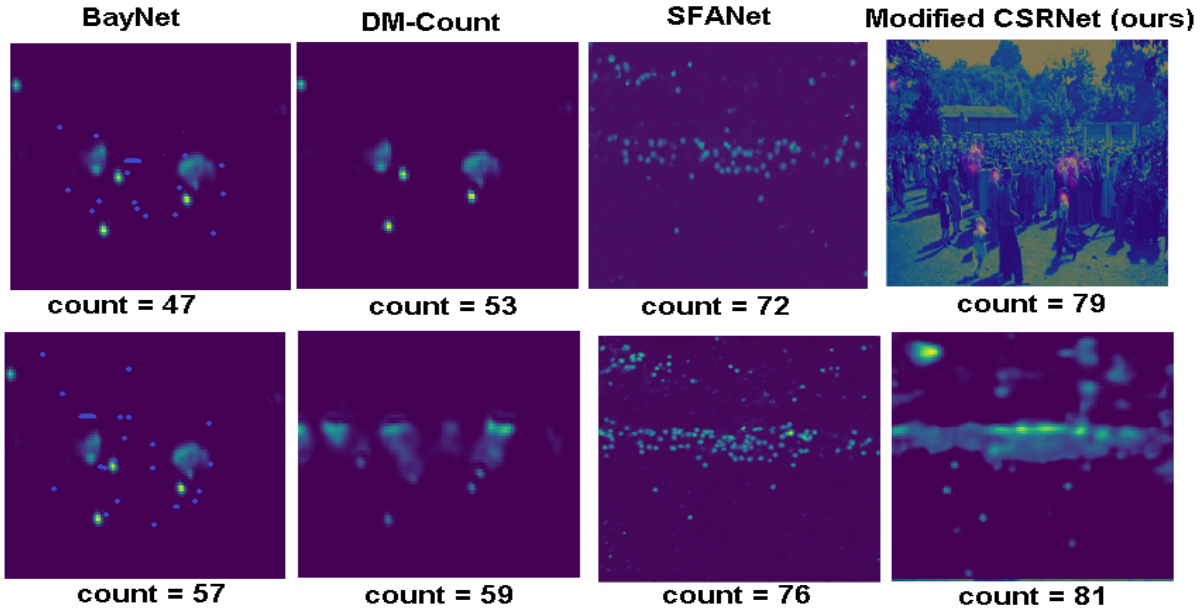


Fig. 10. Comparison of density estimation outcomes produced by our proposed model on128 × 128 and 256 × 256 images of Shanghai Dataset part-A and part-B.

DM-Count, RMSE/MAE values of 50.89/97.07 were attained on SH-A, 5.05/10.34 on SH-B, and 64.04/105.27 on the custom dataset. SFANet was observed to produce RMSE/MAE values of 63.00/84.65 on SH-A, 5.20/10.24 on SH-B, and 64.09/102.65 on the custom dataset. Competitive performance was also demonstrated by the proposed modified CSRNet, with RMSE/MAE values of 63.82/94.47 being achieved on SH-A, 8.64/10.37 on SH-B, and 53.06/100.45 on the custom dataset.

In Fig. 11, the crowd-counting results produced by the state-of-the-art methods are illustrated alongside those generated by the proposed modified CSRNet, using sample images from the custom dataset that was collected from publicly available internet sources.

The experimental results demonstrate the superiority of the modified CSRNet over BayNet, DM-Count, SFANet, and the original CSRNet across multiple datasets and resolutions. The modified CSRNet, through the incorporation of the face detection module and spatial pyramid attention mechanisms, showed improved accuracy, robustness, and generalization capabilities. These results are consistent with recent findings in the literature: Ranasinghe et al. [16] and Rohra et al. [13] both demonstrate that incorporating richer feature representations

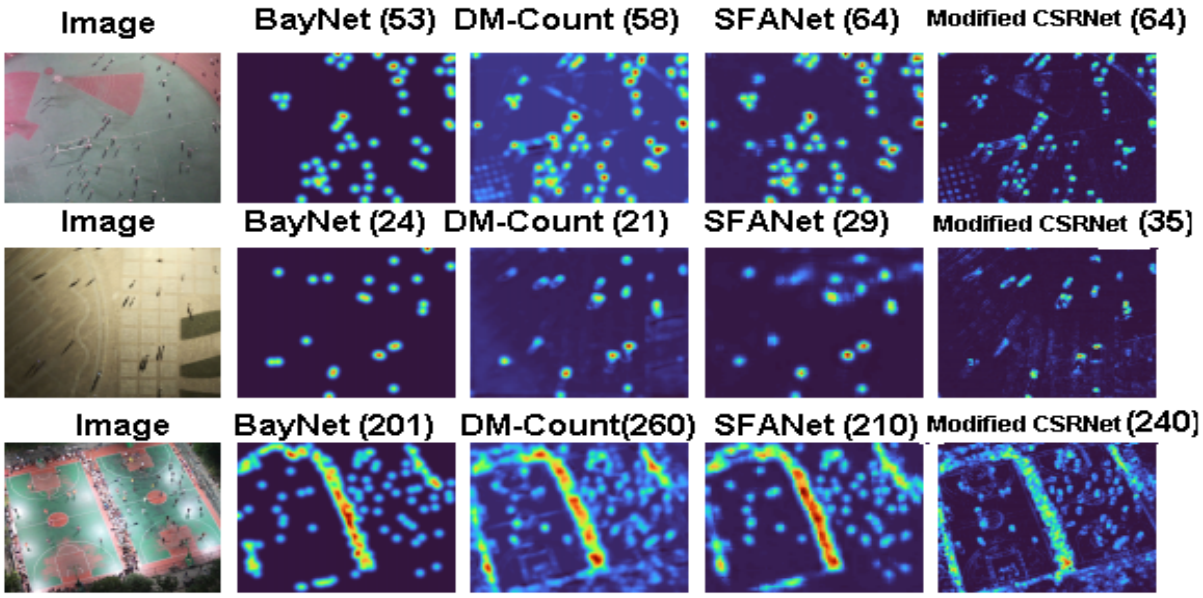


Fig. 11. Density-based prediction performance across models on our dataset, with all input images fixed at a resolution of 512×512.

and improved training strategies leads to measurable gains on congested scene benchmarks. The face detection module's advantage is most pronounced at lower resolutions (e.g., 64×64), where coarser head annotations are insufficient for accurate density estimation but facial keypoints remain detectable.

IX. Discussion

The results of our experiments lead to several important observations. First, the performance of all models degrades at lower resolutions, but the Modified CSRNet exhibits a slower rate of degradation. This is attributed to the facial feature detection module, which provides additional discriminative signals that are partially preserved even at reduced resolutions. Second, the original CSRNet performs unexpectedly well on the custom dataset at full resolution (RMSE/MAE of 25.97/100.74 on SH-B), suggesting that the dilated convolutional architecture generalizes effectively to in-the-wild imagery. Third, BayNet's Bayesian loss formulation provides consistent performance across different datasets, confirming the findings of Wang et al. [64], who noted the robustness of probabilistic counting formulations in handling annotation noise.

The principal limitation of the proposed approach is the reliance on a fixed fusion weight λ for combining density-map and face-detection estimates. An adaptive, content-aware fusion mechanism, possibly based on the estimated crowd density level, could further improve performance. Additionally, the face detection branch adds a modest computational overhead, which may be relevant for deployment on resource-constrained devices. Future work will explore knowledge distillation techniques [16] to compress the face detection branch without sacrificing accuracy.

X. Conclusion and Future Work

To summarize, in this research we have shown that the deep learning architecture of BayNet, CSRNet, DM-Count, and SFANet can be employed effectively to perform crowd counting. The face detection module integrated in CSRNet is one of the major contributions that takes advantage of the facial detection cues in dense scenes to enhance the results of crowd counting. This enhancement overcomes the challenge of accurately estimating crowd density under partial occlusion and at reduced image resolutions. The study emphasizes context-aware feature, such as facial features, for the crowd counting model to improve its performance. The proposed method is compared with recent techniques such as MSFFNet [13], CrowdDiff [16] and PPSC [15] and achieves competitive performance especially in the low-resolution regime.

Several promising directions for future work could be taken. First, multi-modal information like depth, thermal information, and motion information [64] might further improve the counting accuracy. Secondly, real-time implementation and scalability of large-scale urban deployment would be beneficial. The third is that introducing adaptive fusion weights between the density-map estimation and the face-detection estimation (possibly influenced by a scene-complexity indicator) can enable dynamic weighting of each modality. Lastly, the feasibility of incorporating semi-supervised training strategies [15] to mitigate the need for annotations, and the possibility of replacing the commonly used density map regression with diffusion based density estimation [16] are exciting directions that can be explored as extension tasks for this work.

References

[1] Z. Fan *et al.*, "A survey of crowd counting and density estimation based on convolutional neural network," *Neurocomputing*, vol. 472, pp. 224–251, 2022.

[2] A. Albiol *et al.*, "Real-time head tracking by means of a serendipitous approach," *IEEE Transactions on Circuits and Systems for Video Technology*, 2001.

[3] V.-Q. Pham, T. Kozakaya, O. Yamaguchi, and R. Okada, "Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3253–3261.

[4] M. Bhuiyan *et al.*, "Video-based person re-identification using a deep learning framework," *Applied Intelligence*, 2022.

[5] A. Joshi *et al.*, "An efficient approach for crowd counting using deep convolutional neural networks," *IEEE Access*, 2020.

[6] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 833–841.

[7] C. Wang *et al.*, "Deep people counting in extremely dense crowds," 2015.

[8] J. Shao, K. Kang, C. Change Loy, and X. Wang, "Deeply learned attributes for crowded scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4657–4666.

[9] V. A. Sindagi and V. M. Patel, "A survey of recent advances in cnn-based single image crowd counting and density estimation," *Pattern Recognition Letters*, vol. 107, pp. 3–16, 2022.

[10] J. Zhou *et al.*, "Crowd counting and density estimation by trellis encoder-decoder networks," *IEEE Transactions on Image Processing*, 2022.

[11] C. Xu *et al.*, "Autoscale: Learning to scale for crowd counting," *International Journal of Computer Vision*, 2022.

[12] L. Wen *et al.*, "Crowd counting with crowd density estimation," *Information Sciences*, 2021.

[13] A. Rohra, B. Yin, H. Bilal, A. Kumar, M. Ali, and Y. Li, "MSFFNet: Multi-scale feature fusion network with semantic optimization for crowd counting," *Pattern Analysis and Applications*, vol. 28, no. 1, 2025.

[14] M. J. Babar, M. Husnain, M. M. S. Missen, A. Samad, M. Nasir, and A. K. N. Khan, "Crowd counting and density estimation using deep network–a comprehensive survey," *Authorea Preprints*, 2023. [Online]. Available: https://doi.org/10.36227/techrxiv.23256587.v1

[15] Y. Qian, L. Zhang, X. Hong, C. Donovan, and O. Arandjelovic, "Perspective-assisted prototype-based learning for semi-supervised crowd counting," vol. 157. Elsevier, 2024.

[16] Y. Ranasinghe, N. G. Nair, W. G. C. Bandara, and V. M. Patel, "CrowdDiff: Multi-hypothesis crowd density estimation using diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 12 809–12 819.

[17] D. B. Sam *et al.*, "Locate, size and count: Accurately resolving people in dense crowds via detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[18] G. Gao *et al.*, "Cnn-based density estimation and crowd counting: A survey," *arXiv preprint arXiv:2003.12783*, 2020.

[19] V. A. Sindagi and V. M. Patel, "A survey of recent advances in cnn-based single image crowd counting and density estimation," *Pattern Recognition Letters*, vol. 107, pp. 3–16, 2018.

[20] C. Sacchi and C. S. Regazzoni, "Advanced video surveillance with ptz cameras," *IEEE Transactions on Circuits and Systems for Video Technology*, 2001.

[21] S. E. Lin and B. Bhanu, "Estimation of number of people in crowded scenes using perspective transformation," *IEEE Transactions on Systems, Man, and Cybernetics*, 2001.

[22] V. Ranjan, H. Le, and M. Hoai, "Iterative crowd counting," in *Proceedings of the European conference on computer vision*, 2018, pp. 270–285.

[23] M. Shami *et al.*, "People counting in dense crowd images using sparse head detections," *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[24] N. Liu, Y. Long, C. Zou, Q. Niu, L. Pan, and H. Wu, "Adcrowdnet: An attention-injective deformable convolutional network for crowd understanding," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3225–3234.

[25] N. Ilyas *et al.*, "A convolutional neural network based framework for crowd density estimation," *Applied Sciences*, 2019.

[26] L. Mredhula and M. Dorairangaswamy, "Real-time crowd density estimation using deep learning," *Procedia Computer Science*, 2020.

[27] B. Li, H. Huang, A. Zhang *et al.*, "Approaches on crowd counting and density estimation: a review," *Pattern Analysis and Applications*, vol. 24, pp. 853–874, 2021.

[28] K. Lin *et al.*, "Efficient large-scale fleet management via multi-agent deep reinforcement learning," *arXiv preprint*, 2016.

[29] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–7.

[30] K. Chen, S. Gong, T. Xiang, and C. Change Loy, "Cumulative attribute space for age and crowd density estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2467–2474.

[31] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–338, 2010.

[32] L. Boominathan, S. S. Kruthiventi, and R. V. Babu, "Crowdnet: A deep convolutional network for dense crowd counting," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 640–644.

[33] M. Rodriguez *et al.*, "Density-aware person detection and tracking in crowds," in *Proceedings of the IEEE International Conference on Computer Vision*, 2011.

[34] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Advances in Neural Information Processing Systems*, 2010, pp. 1324–1332.

[35] H. Idrees *et al.*, "Multi-source multi-scale counting in extremely dense crowd images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2547–2554.

[36] Z. Zhao *et al.*, "Crossing the line: Crowd counting by integer programming with local features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[37] X. Cao *et al.*, "Large scale crowd analysis based on convolutional neural network," *Pattern Recognition*, 2015.

[38] Y. Cong *et al.*, "Flow estimation based on optical flow model," *IEEE Transactions*, 2009.

[39] Z. Ma *et al.*, "Counting people crossing a line using integer programming and local features," *IEEE Transactions*, 2015.

[40] C. Shang *et al.*, "End-to-end crowd counting via joint learning local appearance and global context," in *IEEE International Conference on Image Processing*, 2016.

[41] C. Xiong *et al.*, "A spatiotemporal model with visual attention for video classification," 2017.

[42] J. Ding *et al.*, "Deeply-supervised networks with threshold loss for cancer detection in automated breast ultrasound," in *International Conference on Medical Image Computing*, 2018.

[43] S. Amirgholipour *et al.*, "A convolutional neural network for crowd counting," in *Proceedings of the Australasian Conference on Artificial Intelligence*, 2018.

[44] J. Dai *et al.*, "Dense crowd counting from still images with convolutional neural networks," *arXiv*, 2021.

[45] H. Idrees, M. Tayyab, K. Athrey *et al.*, "Composition loss for counting, density map estimation and localization in dense crowds," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 544–559.

[46] J. Ribera *et al.*, "Locating objects without bounding boxes," 2019.

[47] V. A. Sindagi, R. Yasarla, and V. M. Patel, "Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[48] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 589–597.

[49] C. Zhang, H. Li, X. Wang, and X. Yang, "Data-driven crowd understanding: A baseline for a large-scale fine-grained crowd analysis benchmark," 2016.

[50] Q. Wang *et al.*, "Nwpu-crowd: A large-scale benchmark for crowd counting and localization," 2020.

[51] F. Yang *et al.*, "Counting and segmenting cells in microscopic images," in *International Symposium on Biomedical Imaging*, 2018.

[52] D. Kang *et al.*, "Crowd counting by adapting convolutional neural networks with side information," in *arXiv preprint*, 2018.

[53] Z. Shen *et al.*, "Crowd counting via adversarial cross-scale consistency pursuit," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[54] G. Olmschenk *et al.*, "Generalizing semi-supervised generative adversarial networks to regression using feature contrasting," in *arXiv preprint*, 2018.

[55] G. Yang *et al.*, "Multi-scale cascaded scene-specific convolutional neural networks for background modeling," in *Proceedings of the Asian Conference on Computer Vision*, 2018.

[56] G. Olmschenk *et al.*, "Dense crowd counting with crowd density maps," in *IEEE International Conference on Image Processing*, 2019.

[57] L. Zhang *et al.*, "Crowd counting via scale-adaptive convolutional neural network," in *IEEE Winter Conference on Applications of Computer Vision*, 2018.

[58] X. Liu, J. van de Weijer, and A. D. Bagdanov, "Leveraging unlabeled data for crowd counting by learning to rank," 2018.

[59] J. Wu *et al.*, "Application of deep neural networks to object detection," *arXiv preprint*, 2017.

[60] Y. Li, X. Zhang, and D. Chen, "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1091–1100.

[61] Y. Yang *et al.*, "Crowd counting with mixed regression for density map estimation," *IEEE Transactions on Image Processing*, 2022.

[62] F. Wang, K. Liu, N. Wei, N. Sang, X. Xia, and J. Sang, "Weakly supervised crowd counting with joint cnn and transformer network," *Pattern Recognition*, vol. 171, p. 112077, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S003132032500737X

[63] D. Zhang, Z. Liang, and N. Liu, "Cross-domain crowd counting model based on frequency domain enhancement," *Journal of System Simulation*, vol. 38, no. 1, pp. 158–173, 2026.

[64] Y. Wang *et al.*, "A comprehensive survey of crowd density estimation and counting," *IET Image Processing*, vol. 19, 2025.

[65] L. Fotia, G. Percannella, A. Saggese, and M. Vento, "Optimizing crowd counting in dense environments through curriculum learning training strategy," *SN Computer Science*, vol. 5, p. 683, 2024.

[66] X. Yan, F. Nie, Y. Hu, and R. Liu, "Research on dense crowd detection algorithm based on improved YOLOv8," in *Proceedings of the 2024 7th International Conference on Artificial Intelligence and Pattern Recognition*. ACM, 2024, pp. 707–713.

[67] Z. Ma, X. Wei, X. Hong, and Y. Gong, "Bayesian loss for crowd count estimation with point supervision," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6142–6151.

[68] B. Wang, H. Liu, D. Samaras, and M. H. Nguyen, "Distribution matching for crowd counting," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1595–1607.

[69] H. Lin, X. Hong, Z. Ma, Y. Wang, and D. Meng, "Multidimensional measure matching for crowd counting," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 5, pp. 9112–9126, 2024.

[70] L. Zhu, Z. Zhao, C. Lu, Y. Lin, Y. Peng, and T. Yao, "Dual path multi-scale fusion networks with attention for crowd counting," in *arXiv preprint arXiv:1902.01115*, 2019.

[71] H. Lin, Z. Ma, R. Ji, Y. Wang, and X. Hong, "Boosting crowd counting via multifaceted attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 19 628–19 637.

[72] I. Khalfaoui-Hassani, T. Pellegrini, and T. Masquelier, "Dilated convolution with learnable spacings," *arXiv preprint arXiv:2112.03740*, 2021. [Online]. Available: https://arxiv.org/abs/2112.03740

[73] X. Jiang, L. Zhang, P. Lv, Y. Guo, R. Zhu, Y. Li, Y. Pang, X. Li, B. Zhou, and M. Xu, "Learning multi-level density maps for crowd counting," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 8, pp. 2705–2715, 2019.

[74] T. O. Hodson, "Root mean square error (rmse) or mean absolute error (mae): When to use them or not," *Geoscientific Model Development Discussions*, vol. 2022, pp. 1–10, 2022.

FSCM-Net: Frequency-Spatial Collaborative Modeling for Traffic Scene Semantic Segmentation

Wei Zhao[1]*, Yi Dong[2], Lingchao Wang[3], Qiang Ai[4]
College of Information Engineering, Liaoning Institute of Science and Engineering, Jinzhou, China[1,2,3]
School of Information Science and Engineering, Lanzhou University, Lanzhou, China[4]

***Abstract*—To address the challenges of insufficient global semantic modeling and blurred boundaries in urban traffic scene segmentation, this study proposes a frequency-spatial collaborative framework based on DeepLabV3+. A Spectral Decoupling Adaptive Modulation (SDAM) module enhances low-frequency semantics and high-frequency details in the frequency domain. A Hierarchical Spatial Dependency Modeling (HSDM) module captures local consistency and global semantic dependencies, while a Structure-guided Adaptive Multi-scale Fusion (SAMF) module dynamically integrates multi-scale features using structural priors. Experiments on Cityscapes and CamVid demonstrate improvements of 3.6% and 1.6% over DeepLabV3+, respectively, while maintaining real-time performance.**

Keywords*—Frequency-domain modeling; spatial dependency modeling; local-global attention; structure-guided fusion; feature enhancement

I. Introduction

With the rapid advancement of autonomous driving, Intelligent Transportation Systems (ITS), and Advanced Driver Assistance Systems (ADAS), robust perception of traffic scenes has become a prerequisite for achieving automated driving. As a fundamental task in visual perception, semantic segmentation aims to perform fine-grained parsing of critical road elements, such as vehicles and pedestrians, thereby providing reliable support for environmental understanding and decision-making processes [1].

Current research on traffic scene semantic segmentation has mainly focused on feature enhancement and contextual modeling. Zhao et al. [2] proposed PSPNet, which captures multi-scale global contextual information through pyramid pooling and significantly improves segmentation accuracy in complex scenarios. Song et al. [3] developed MPNet based on DeepLabV3+, which effectively integrates multi-level feature representations through a multi-scale feature fusion architecture, thereby enhancing the perception capability for objects of varying scales and achieving notable improvements in traffic scene segmentation. However, these approaches primarily rely on spatial-domain feature fusion and do not explicitly exploit frequency-domain information for feature modeling and optimization. To compensate for the limitations of spatial modeling, several studies have investigated image representation from the frequency-domain perspective. Shan et al. [4] demonstrated that images can be decomposed into different frequency components through frequency analysis, where low-frequency components encode global structures and semantic information, while high-frequency components contain texture and boundary details, enabling a more comprehensive representation of visual information. Building upon this observation, Zhang et al. [5] proposed an adaptive semantic segmentation method based on frequency-domain analysis, which decouples style and semantic representations via Fast Fourier Transform (FFT) and enhances robustness under domain shifts. To further incorporate frequency information into fine-grained spatial modeling, especially for object boundary representation, Ge et al. [6] introduced SF-Net, which decomposes spatial features into high- and low-frequency components using the Haar Wavelet Transform and employs a Mamba-based enhancement module to strengthen global contextual representations. Although improvements in boundary accuracy were achieved, the method is prone to introducing frequency-domain noise in complex environments. Yang et al. [7] proposed the Frequency Feature Rectification (FFR) framework, which explicitly compensates for high-frequency components within the decoder to recover object boundaries, thereby improving segmentation quality in weakly supervised scenarios. Furthermore, Huang et al. [8] proposed FSDR, which performs randomization operations in the frequency domain to improve domain generalization capability and empirically validates the effectiveness of frequency-domain representations against domain shifts. Nevertheless, existing frequency-based methods mainly focus on feature enhancement or domain adaptation, while the deep collaborative modeling between frequency-domain and spatial-domain representations remains insufficient.

To further exploit frequency information for deep feature modeling and improve boundary recognition of small objects, increasing attention has been devoted to integrating attention mechanisms and multi-scale modeling strategies. Sima et al. [9] proposed SPCONet, which balances accuracy and efficiency through lightweight branches by jointly modeling spatial and contextual information, providing an effective solution for real-time traffic scene segmentation. To enhance the perception of small objects, Yu et al. [10] introduced BiSeNet, which adopts a dual-branch architecture to fuse local details and global semantics while strengthening shallow representations, thereby improving overall segmentation performance. Wang et al. [11] addressed the challenge of segmenting distant small objects in road scenes by integrating global contextual modeling and multi-scale feature fusion, effectively enhancing the representation of low-resolution targets. Wu et al. [12] developed a frequency-aware attention mechanism based on wavelet transforms, which theoretically benefits from multi-resolution analysis and provides a new paradigm for incorporating frequency-domain information. To improve generaliza-

*Corresponding author.

tion capability in complex traffic scenarios, Zhang et al. [13] incorporated a frequency modulation mechanism into the Transformer framework, where Fourier Transform is employed for frequency-domain modeling and adaptive modulation units are designed to fuse frequency and spatial representations, thereby strengthening feature expressiveness. Moreover, Niu et al. [14] emphasized that maintaining semantic consistency and stylistic diversity is crucial for enhancing model generalization, while such consistency heavily relies on accurate modeling of spatial relationships. Bai et al. [15] proposed a multi-scale dependency modeling strategy that improves scene understanding accuracy by constructing complex dependency networks at the expense of additional computational overhead. Li et al. [16] developed an occupancy network-based approach for 3D image semantic segmentation, which improves segmentation performance, particularly for multi-scale objects, through feature pyramids, channel attention mechanisms, and optimized backbone architectures.

The proposed method is not a straightforward combination of "frequency enhancement + spatial modeling + multi-scale fusion". Instead, it formulates an optimization paradigm in which frequency-domain structures explicitly guide spatial propagation and subsequently influence scale selection. The central innovation of this work is to elevate frequency information from an auxiliary enhancement cue to an intermediate structural representation for semantic reasoning. Specifically, amplitude-phase decoupling is first performed on encoded features to explicitly separate frequency-domain structures, where amplitudes are utilized for semantic energy modeling and phases serve as spatial structural constraints, while controlled perturbations are introduced to preserve geometric consistency. Subsequently, these structured frequency representations are not directly fed into the decoder but are instead employed as inputs to the Hierarchical Spatial Dependency Modeling (HSDM) module, which constrains local structural consistency and drives global graph-based semantic propagation. Finally, structural responses extracted from the original image are used to construct scale-selection signals independent of semantic features, thereby enabling structure-aware competitive fusion among multi-scale branches.

II. Related Theoretical Foundations

A. Encoder-Decoder-Based Framework

In semantic segmentation tasks, models are required not only to extract high-level semantic representations but also to preserve spatial location information and boundary details to achieve pixel-wise classification. Although conventional convolutional neural networks can enhance semantic representation capability through successive downsampling operations, the spatial resolution of feature maps is inevitably reduced, resulting in the loss of fine details and edge structures. Therefore, Encoder–Decoder architectures have been widely adopted in semantic segmentation to jointly accomplish semantic feature extraction and spatial information recovery through coordinated encoding and decoding processes.

The Encoder is primarily responsible for extracting deep semantic representations. The convolution operation can be formulated as:

$$\mathbf{Y} = \sigma(\mathbf{W} * \mathbf{X} + \mathbf{b}), \tag{1}$$

where, $\mathbf{X}$ and $\mathbf{Y}$ denote the input and output feature maps, respectively, $\mathbf{W}$ represents the convolution kernel parameters, $\mathbf{b}$ denotes the bias term, $*$ indicates the convolution operation, and $\sigma(\cdot)$ is a nonlinear activation function.

As the network depth increases, the Encoder acquires a larger receptive field and stronger semantic representation capability. However, repeated downsampling operations weaken high-frequency textures and boundary information, thereby degrading the segmentation performance for small objects and regions with complex boundaries.

The Decoder is designed to restore the spatial resolution of feature maps and reconstruct object structures. Common approaches include upsampling, transposed convolution, and interpolation operations. In addition, low-level detail features are often fused with high-level semantic representations to enhance boundary localization and fine-grained feature expression.

Encoder–Decoder architectures have become the dominant paradigm in semantic segmentation. Nevertheless, conventional convolution operations are inherently constrained by local receptive fields and struggle to model long-range dependencies effectively. Moreover, repeated downsampling may lead to the loss of high-frequency details. Therefore, enhancing global feature interactions while preserving fine structural information remains an important research direction in semantic segmentation.

B. DeepLab Series Networks

The DeepLab series improves pixel-level prediction accuracy by enlarging the receptive field through atrous convolutions while alleviating the loss of feature resolution.

1) Atrous convolution: Traditional convolutional networks typically enlarge the receptive field using pooling operations or strided convolutions. However, continuous downsampling often leads to the loss of boundary information and fine details of small objects. Atrous convolution expands the sampling range by introducing holes between convolution kernel elements, thereby increasing the receptive field without introducing additional parameters. Its operation can be expressed as

$$y(i) = \sum_{k} x(i + r \cdot k) w(k), \tag{2}$$

where, x and y denote the input and output feature maps, respectively, w represents the convolution kernel, and r is the dilation rate. When $r = 1$, atrous convolution degenerates into standard convolution. As the dilation rate increases, the receptive field expands accordingly.

Compared with conventional convolutions, atrous convolutions can capture richer contextual information while maintaining the spatial resolution of feature maps, thereby enhancing the representation capability for multi-scale objects and complex scenes.

2) Atrous spatial pyramid pooling: To improve the representation capability for objects of different scales, DeepLabV3 introduces the Atrous Spatial Pyramid Pooling (ASPP) module. ASPP employs parallel atrous convolution branches with different dilation rates to model multi-scale contextual information.

Different receptive fields enable ASPP to simultaneously capture local details and global semantic information. Smaller dilation rates focus on texture and boundary characteristics, whereas larger dilation rates aggregate broader contextual cues, thereby improving the adaptability to scale variations. A typical ASPP module consists of multiple atrous convolution branches with different dilation rates, a 1×1 convolution branch, a global average pooling branch, and a multi-scale feature fusion operation. The fusion process can be formulated as

$$F_{\mathrm{ASPP}} = \mathrm{Concat}\left(F_1, F_2, \ldots, F_n\right), \tag{3}$$

where, F_i denotes the feature map extracted by the i-th branch and $\mathrm{Concat}(\cdot)$ represents the channel-wise concatenation operation.

Although ASPP effectively enhances multi-scale contextual modeling, its feature extraction mechanism still primarily relies on local convolutional operations, limiting its ability to capture long-range dependencies. In addition, its capability to preserve high-frequency textures and recover small-object details remains insufficient.

3) DeepLabV3+ architecture: DeepLabV3+ extends DeepLabV3 by incorporating an Encoder–Decoder architecture to integrate high-level semantic representations with low-level spatial details, thereby improving segmentation performance for boundary regions and small objects.

In the Encoder stage, a backbone network is employed to extract high-level semantic features, while the ASPP module is adopted to enrich multi-scale contextual representations. Atrous convolutions replace part of the downsampling operations to enlarge the receptive field while preserving more spatial information.

In the Decoder stage, high-level features are first upsampled and then fused with shallow low-level features to enhance boundary refinement and spatial localization capability.

Although DeepLabV3+ has demonstrated superior performance in complex semantic segmentation scenarios, its feature modeling process still primarily relies on convolution operations, limiting its ability to capture long-range dependencies and global contextual interactions. Furthermore, repeated downsampling and sparse sampling induced by atrous convolutions may result in the loss of high-frequency details, thereby affecting segmentation accuracy for small objects and regions with intricate boundaries.

C. Visual Feature Representation Based on Spectral Analysis

1) Spatial-domain and frequency-domain representations: Traditional convolutional neural networks mainly perform feature extraction in the spatial domain, which is effective for modeling local textures and semantic information. However, their ability to characterize global frequency distributions remains limited.

In contrast, frequency-domain representations describe image structures from a global perspective. Different frequency components correspond to different visual characteristics. Low-frequency components capture overall semantic structures, whereas high-frequency components encode boundary, texture, and fine-detail information. Therefore, frequency-domain analysis provides complementary information for modeling both global semantics and fine-grained structures.

2) Two-dimensional discrete fourier transform: The two-dimensional Discrete Fourier Transform (DFT) is a fundamental tool for frequency-domain image analysis. It transforms spatial-domain images into the frequency domain and can be expressed as

$$F(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \exp\left[-j2\pi\left(\frac{ux}{M} + \frac{vy}{N}\right)\right], \tag{4}$$

where, $f(x,y)$ denotes the spatial-domain image, $F(u,v)$ represents its frequency-domain representation, and (u,v) denote frequency coordinates. Through DFT, images can be decomposed into different frequency components to characterize semantic and structural information from the frequency perspective.

3) High-frequency and low-frequency information: In frequency-domain representations, low-frequency information mainly reflects the global structure and semantic content of images, whereas high-frequency information contains boundary, texture, and local detail characteristics. High-frequency components are particularly important for object boundary recovery and small-object segmentation.

Therefore, jointly modeling low-frequency semantic information and high-frequency detail information is essential for improving segmentation performance in complex scenarios.

D. Spatial Dependency Modeling Theory

1) Importance of spatial contextual information: Semantic segmentation relies not only on local texture features but also on spatial relationships and contextual dependencies among pixels. In complex scenes, different semantic categories often exhibit structural correlations, such as road continuity, building consistency, and the spatial distributions of vehicles and pedestrians.

Therefore, relying solely on local features is insufficient for accurately interpreting complex scenes. Effective spatial contextual modeling can enhance scene understanding capability and improve the consistency and discriminability of segmentation results.

2) Limitations of convolutional networks: Conventional convolutional neural networks primarily perform local feature extraction through convolution operations. Although stacking multiple layers enlarges the receptive field, the underlying mechanism remains local in nature and is limited in modeling long-range spatial dependencies.

In complex scenarios, insufficient utilization of contextual information may impair the model's understanding of global structural relationships.

3) Existing spatial dependency modeling methods: To strengthen spatial contextual modeling, various approaches, including atrous convolutions and attention mechanisms, have been proposed. Atrous convolutions enlarge the sampling range to improve contextual perception, whereas attention mechanisms establish feature correlations to enhance information interactions across different regions.

However, while these global modeling strategies improve feature interactions, they may also introduce cross-region interference, resulting in semantic ambiguity and reduced boundary discriminability.

4) Spatial consistency problem: In complex scenes, small-object regions are susceptible to background interference, while global feature propagation may destabilize local feature representations. Excessive emphasis on global information may weaken local structural consistency, whereas relying solely on local features limits the utilization of contextual information.

Therefore, achieving collaborative modeling between local and global information while preserving structural consistency remains an important challenge in spatial dependency modeling.

E. Multi-Scale Feature Fusion Theory

1) Importance of multi-scale features: Features extracted from different network layers contain complementary visual information. Shallow features preserve boundary textures and local details due to their higher spatial resolution, whereas deep features possess larger receptive fields and stronger semantic representation capability.

Since features at different scales are complementary, relying on a single-level representation is insufficient to simultaneously capture global semantics and local details. Consequently, multi-scale feature fusion enhances the representation capability for complex scenes and objects with scale variations.

2) Classical multi-scale feature fusion methods: Various multi-scale fusion strategies have been proposed to improve adaptability to objects of different sizes. Feature Pyramid Networks (FPN) perform feature aggregation through top-down pathways and lateral connections, thereby enhancing multi-scale representations.

ASPP extracts contextual information at multiple scales using atrous convolutions with different dilation rates. Pyramid Scene Parsing Networks (PSPNet) aggregate global contextual information through multi-scale pooling structures to improve semantic representation in complex environments.

In addition, U-Net utilizes skip connections to fuse shallow spatial details with deep semantic features, effectively improving boundary recovery and local structure representation.

3) Feature fusion strategies: Current multi-scale fusion methods commonly adopt feature concatenation, weighted fusion, and attention-based fusion strategies.

Feature concatenation preserves information from different levels by channel-wise aggregation. Weighted fusion dynamically adjusts the importance of different features to strengthen critical representations. Attention-based fusion further employs channel-wise or spatial attention mechanisms to adaptively select informative features, thereby improving fusion effectiveness.

Through these strategies, models can achieve collaborative modeling of semantic and spatial information to a certain extent.

4) Limitations of existing multi-scale fusion methods: Although existing multi-scale fusion methods improve the representation capability for objects at different scales, several limitations remain. Due to significant differences between feature levels in terms of semantic content and spatial details, semantic conflicts may arise during fusion, thereby affecting feature consistency.

Moreover, small-object representations can be overwhelmed by dominant large-scale semantic information during deep propagation, leading to insufficient fine-grained discrimination capability. Existing fusion approaches also generally lack structural priors to guide feature selection and aggregation, making it difficult to fully exploit spatial structural information within scenes.

In addition, most methods still lack effective adaptive feature selection mechanisms and cannot dynamically adjust the importance of different scales according to varying scenarios. Therefore, achieving structure-guided adaptive multi-scale feature fusion has become an important direction for improving semantic segmentation performance in complex environments.

III. Methodology

The conventional DeepLabV3+ framework, built upon deep convolutional networks and Atrous Spatial Pyramid Pooling (ASPP), is capable of enlarging the receptive field and extracting multi-scale semantic representations, thereby achieving promising performance in traffic scene semantic segmentation tasks. However, the discrete sampling mechanism of atrous convolutions often leads to discontinuous responses for small objects and blurred object boundaries. In addition, single-scale feature representations struggle to simultaneously preserve local details and maintain global structural consistency, while the capability of cross-region semantic propagation remains limited. To address these issues, this study proposes a Frequency-Spatial Collaborative Segmentation framework (FSCS), which integrates frequency-domain enhancement with spatial collaborative modeling.

The overall architecture of the proposed method is illustrated in Fig. 1. First, a Spectral Decoupling Adaptive Modulation (SDAM) module is introduced after the encoder output. Through amplitude–phase decoupling and adaptive spectral modulation, SDAM enhances the responses of high-frequency features associated with small objects while preserving spatial structural continuity. Subsequently, a Hierarchical Spatial Dependency Modeling (HSDM) module is incorporated after the ASPP module. HSDM maintains fine-grained structural consistency through local neighborhood weighting and simultaneously achieves cross-region semantic propagation

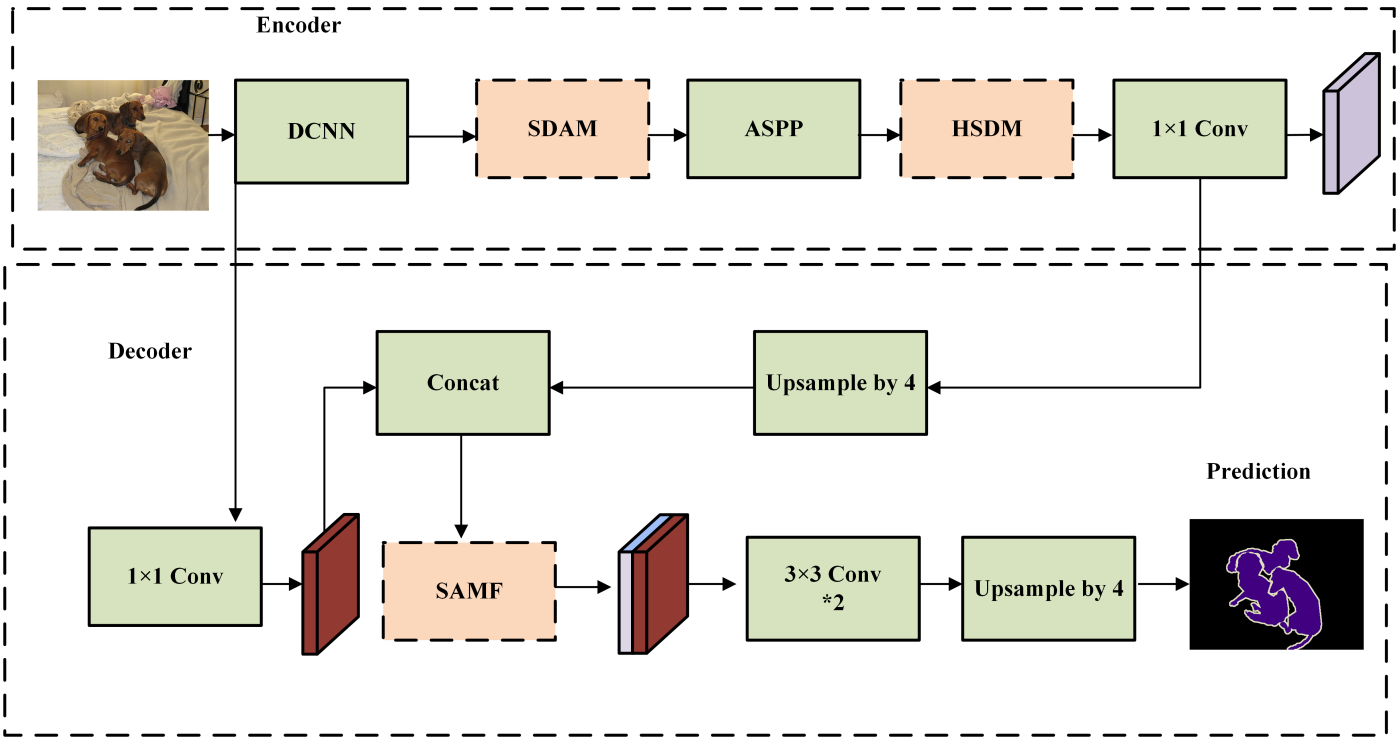


Fig. 1. Overall architecture of the proposed FSCS framework.

via global graph diffusion, thereby balancing local structural information and global semantic representations. Finally, a Structure-guided Adaptive Multi-scale Fusion (SAMF) module is embedded into the low-level feature fusion stage of the decoder. By utilizing structural responses generated from the original input image, SAMF dynamically adjusts the weights of features from different scales, enabling both boundary refinement and the preservation of semantic consistency for large-scale structures.

The three proposed modules are sequentially deployed after the encoder output, following the ASPP module, and within the decoder fusion stage, respectively. Together, they establish a complete processing pipeline ranging from frequency-domain enhancement to spatial dependency modeling and multi-scale feature fusion. Through collaborative optimization in both frequency and spatial domains, the proposed framework significantly improves semantic segmentation performance in complex traffic scenes.

A. Spectral-Decoupled Adaptive Modulation Module

Traditional convolutional neural networks primarily perform local feature extraction in the spatial domain, where the receptive field is inherently constrained by the kernel size. Consequently, such networks struggle to effectively capture long-range dependencies across spatial regions, limiting their ability to represent complex textures and fine-grained semantic structures. In traffic scene semantic segmentation, roads typically exhibit large-scale and spatially continuous structures, whereas lane markings, pedestrians, and traffic signs are characterized by prominent high-frequency components. This disparity in frequency characteristics makes it difficult for purely spatial-domain modeling approaches to simultaneously capture information distributed across different frequency bands.

The DeepLab series enlarges the receptive field through atrous convolutions, thereby improving multi-scale contextual representation. However, the discrete sampling mechanism of atrous convolutions may still result in discontinuous responses for small objects and produce “semantic holes” around object boundaries. As a consequence, the representation capability for fine structures and high-frequency details remains limited. Therefore, it is necessary to complement spatial-domain modeling with frequency-domain representations to enhance the expressive power of segmentation models in complex traffic environments.

To this end, a Spectral-Decoupled Adaptive Modulation (SDAM) module is introduced to perform frequency-domain enhancement on the encoded features. The overall architecture of SDAM is illustrated in Fig. 2. Specifically, the module transforms spatial-domain features into the frequency domain, where amplitude and phase components are explicitly decoupled. The amplitude spectrum is utilized to characterize the energy distribution associated with semantic representations at different frequency bands, while the phase spectrum preserves structural information and spatial relationships. By performing adaptive spectral modulation on the decomposed frequency components, SDAM selectively enhances informative high-frequency responses related to small objects and boundary details while retaining the global semantic consistency encoded in low-frequency components. Subsequently, the modulated spectral representations are transformed back into the spatial domain to generate refined feature maps.

Through explicit frequency-domain modeling and adaptive

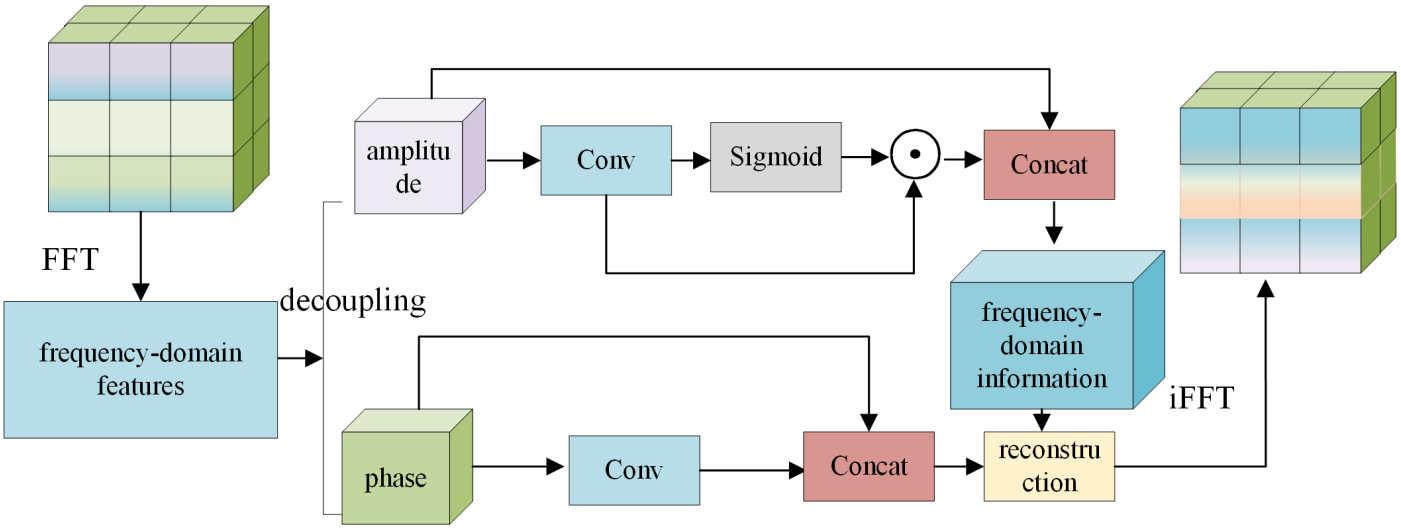


Fig. 2. Architecture of the proposed Spectral-Decoupled Adaptive Modulation (SDAM) module.

spectral interaction, SDAM effectively compensates for the limitations of conventional convolutional operations in capturing fine-grained structures. The proposed module enhances the discrimination capability for small targets and boundary regions while preserving the continuity of large-scale semantic structures, thereby providing more representative features for subsequent spatial dependency modeling and multi-scale feature fusion.

Given an input feature map $\mathbf{F} \in \mathbb{R}^{H\times W\times C}$, a two-dimensional Discrete Fourier Transform (DFT) is first applied to project the spatial-domain representation into the frequency domain:

$$\hat{\mathbf{F}} = \mathcal{F}(\mathbf{F}), \tag{5}$$

where, $\hat{\mathbf{F}} \in \mathbb{C}^{H\times W\times C}$ denotes the complex-valued frequency-domain representation containing responses from different frequency components.

To improve the controllability of frequency-domain modeling, the spectral representation is further decoupled into amplitude and phase components:

$$\mathbf{A} = |\hat{\mathbf{F}}|, \quad \mathbf{\Phi} = \angle\hat{\mathbf{F}}, \tag{6}$$

where, $\mathbf{A} \in \mathbb{R}^{H\times W\times C}$ represents the amplitude spectrum that characterizes the energy distribution across different frequency bands, while $\mathbf{\Phi} \in \mathbb{R}^{H\times W\times C}$ denotes the phase spectrum that encodes spatial structural information and plays a critical role in preserving boundaries and geometric configurations.

Subsequently, an adaptive modulation mechanism is introduced to dynamically redistribute spectral energy within the amplitude component:

$$\mathbf{A}' = \mathbf{A} \odot (1 + \sigma\left(\mathrm{Conv}(\mathbf{A})\right)), \tag{7}$$

where, $\mathrm{Conv}(\cdot)$ denotes a channel-wise convolution operation, $\sigma(\cdot)$ is the Sigmoid activation function, and $\odot$ represents element-wise multiplication. This residual scaling formulation enables adaptive reweighting of frequency components. On the one hand, the original spectral energy is preserved to prevent information loss; on the other hand, learnable weights are employed to enhance responses in informative frequency bands while suppressing redundant spectral components.

For the phase component, a residual modeling strategy is adopted to avoid disrupting spatial structural continuity:

$$\mathbf{\Phi}' = \mathbf{\Phi} + \Delta\mathbf{\Phi}, \tag{8}$$

where, the learnable phase perturbation is defined as

$$\Delta\mathbf{\Phi} = \mathrm{Conv}(\mathbf{\Phi}). \tag{9}$$

Here, $\Delta\mathbf{\Phi}$ introduces fine-grained adjustments to the phase spectrum while maintaining the continuity of the original structural representation, thereby preventing frequency-domain operations from degrading spatial geometric information.

From the perspective of frequency-domain reconstruction, the inverse Fourier transform is highly sensitive to phase variations. Consequently, uncontrolled perturbations may induce global structural distortions in the spatial domain. To alleviate this issue, the proposed residual phase modulation can be interpreted as a constrained optimization process within the local neighborhood of the original phase distribution. The magnitude of the phase update is regulated through explicit normalization and feature-scale constraints, ensuring that the perturbations remain within a controllable low-amplitude range.

Furthermore, from the viewpoint of frequency-domain noise propagation, random noise is typically manifested as the accumulation of inconsistent high-frequency phase responses. The residual formulation anchors phase updates to the original structural representation, effectively introducing structural priors into the optimization process. As a result, the perturbation

term primarily influences uncertain local boundary regions without amplifying spectral instability on a global scale. In addition, the normalization operation following the inverse transform further rebalances the spectral energy distribution, thereby suppressing distribution shifts induced by local phase perturbations and improving the stability of the frequency enhancement process.

After amplitude and phase modulation, the refined spectral components are recombined to reconstruct the enhanced frequency-domain representation:

$$\hat{\mathbf{F}}' = \mathbf{A}' \cdot e^{j\mathbf{\Phi}'}, \tag{10}$$

where, j denotes the imaginary unit.

The enhanced spectral representation is subsequently transformed back into the spatial domain using the inverse Fourier transform:

$$\mathbf{F}' = \mathcal{F}^{-1}(\hat{\mathbf{F}}'), \tag{11}$$

followed by normalization and point-wise convolution for feature reconstruction:

$$\mathbf{F}_{\text{out}} = \text{Conv}_{1\times1}\left(\text{Norm}(\mathbf{F}')\right), \tag{12}$$

where, $\text{Norm}(\cdot)$ denotes the normalization operation, which stabilizes the feature distribution and facilitates channel-wise feature reorganization.

Through amplitude–phase decoupling and controlled spectral modulation, SDAM enables fine-grained modeling of different frequency components. By enhancing informative semantic frequency bands while preserving the stability of structural representations, the proposed module provides structurally consistent and semantically enriched features for subsequent spatial dependency modeling and multi-scale feature fusion.

B. Hierarchical Spatial Dependency Modeling Module

In traffic scene semantic segmentation, spatial relationships are reflected not only in the structural continuity within local neighborhoods but also in semantic consistency across distant regions. Therefore, relying solely on frequency-domain enhancement or local convolution operations is insufficient to simultaneously enforce local structural constraints and capture long-range semantic dependencies. To address this issue, a Hierarchical Spatial Dependency Modeling (HSDM) module is proposed to progressively model spatial relationships from local to global levels and achieve unified feature representation through an adaptive fusion mechanism. The overall architecture of HSDM is illustrated in Fig. 3. The core idea of the proposed module is to first establish structurally consistent feature representations through local dependency modeling and subsequently perform global semantic propagation. Such a hierarchical strategy effectively alleviates the structural ambiguity introduced by direct global interactions, thereby achieving collaborative optimization between structural consistency and semantic coherence.

The input feature map of HSDM is obtained from the output of the SDAM module and can be represented as

$$\mathbf{F} \in \mathbb{R}^{H\times W\times C}, \tag{13}$$

where, H, W, and C denote the spatial dimensions and the number of feature channels, respectively. To facilitate global dependency modeling, the feature map is reshaped into a matrix form:

$$\mathbf{X} \in \mathbb{R}^{N\times C}, \quad N = H \times W, \tag{14}$$

where, each spatial location is treated as a graph node, transforming the spatial dependency modeling problem into a graph signal processing task.

1) Local structural consistency modeling: To preserve boundaries and fine-grained structures, local structural consistency is first established within neighborhood regions. For an arbitrary spatial location p, let $\mathcal{N}(p)$ denote its local neighborhood. The local feature response is obtained through weighted aggregation:

$$\mathbf{F}_{\text{loc}}(p) = \sum_{q\in\mathcal{N}(p)} \omega(p,q)\mathbf{F}(q), \tag{15}$$

where, the weighting coefficient $\omega(p,q)$ is defined as

$$\omega(p,q) = \frac{\exp\left(-\frac{\|\phi(\mathbf{F}(p))-\phi(\mathbf{F}(q))\|_2^2}{\sigma^2} - \lambda d(p,q)\right)}{\sum_{q'\in\mathcal{N}(p)}\exp\left(-\frac{\|\phi(\mathbf{F}(p))-\phi(\mathbf{F}(q'))\|_2^2}{\sigma^2} - \lambda d(p,q')\right)}, \tag{16}$$

where, $\phi(\cdot)$ denotes a 1×1 convolutional projection function, $d(p,q)$ represents the spatial distance between positions p and q, σ controls the feature similarity scale, and λ is the structural regularization coefficient. This weighted aggregation process is equivalent to bilateral filtering in graph feature space, enabling effective local structural preservation.

2) Global semantic modeling: Based on locally aligned structural representations, semantic information is further propagated over the entire graph. The locally consistent features are regarded as graph signals, and global semantic diffusion is performed over a weighted graph:

$$\mathbf{F}_{\text{pair}}(i) = \sum_{j=1}^{N} W_{ij}\mathbf{F}_{\text{loc}}(j), \tag{17}$$

where, the affinity matrix $\mathbf{W}$ is defined as

$$W_{ij} = \frac{\exp\left(-\frac{\|\phi(\mathbf{F}_{\text{loc}}(i))-\phi(\mathbf{F}_{\text{loc}}(j))\|_2^2}{\sigma^2}\right)}{\sum_{j'=1}^{N}\exp\left(-\frac{\|\phi(\mathbf{F}_{\text{loc}}(i))-\phi(\mathbf{F}_{\text{loc}}(j'))\|_2^2}{\sigma^2}\right)}. \tag{18}$$

This formulation corresponds to a graph diffusion process that facilitates global semantic propagation while reducing the risk of information leakage across structurally inconsistent regions.

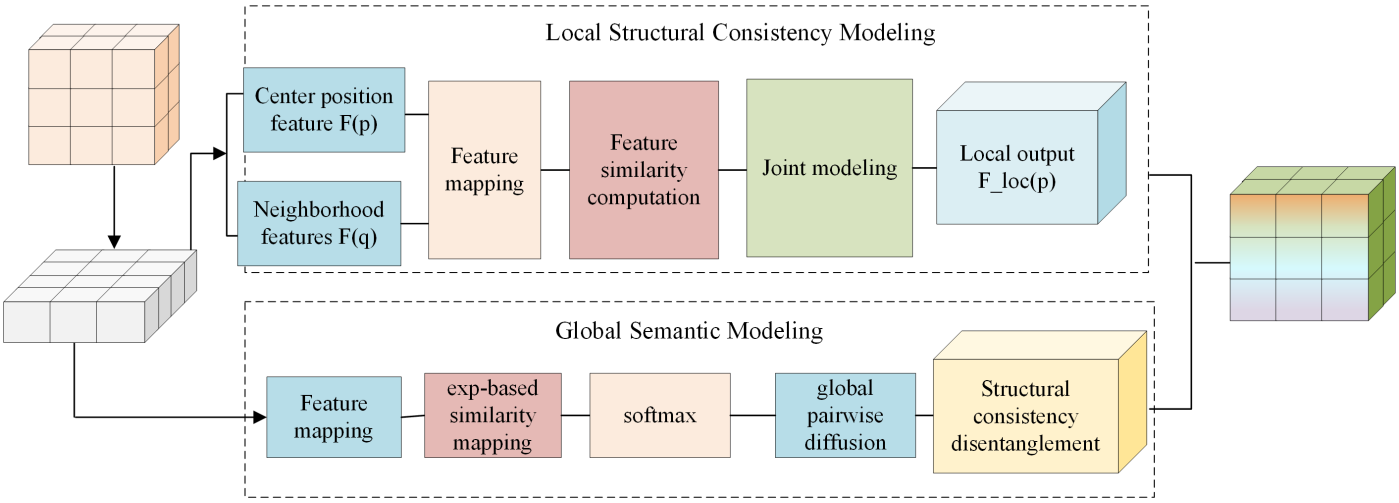


Fig. 3. Architecture of the proposed Hierarchical Spatial Dependency Modeling (HSDM) module.

3) Structure-semantic decoupled modeling: To mitigate the over-smoothing effect caused by global propagation, a decoupling strategy is introduced to separate the global response into relational and unary components:

$$\mathbf{F}_{\text{glob}} = \mathbf{F}_{\text{pair}} + \mathbf{F}_{\text{unary}}, \tag{19}$$

where, the unary term is defined as

$$\mathbf{F}_{\text{unary}}(i) = \mathbf{W}_u \mathbf{F}(i), \tag{20}$$

with, $\mathbf{W}_u$ denoting learnable transformation parameters. The unary component preserves the intrinsic saliency of each spatial position and corresponds to the unary term in graphical models, whereas $\mathbf{F}_{\text{pair}}$ represents the pairwise term responsible for modeling cross-region semantic relationships.

4) Adaptive fusion: Considering that different spatial regions exhibit varying dependencies on local structural information and global semantic context, a spatially adaptive fusion mechanism is further introduced:

$$\boldsymbol{\alpha} = \sigma\left(\text{Conv}_{1\times1}\left([\mathbf{F}_{\text{loc}}, \mathbf{F}_{\text{glob}}]\right)\right), \tag{21}$$

where, $[\cdot,\cdot]$ denotes channel-wise concatenation and $\sigma(\cdot)$ represents the Sigmoid activation function.

The final output feature representation is obtained by

$$\mathbf{F}_{\text{out}} = \boldsymbol{\alpha} \odot \mathbf{F}_{\text{loc}} + (1 - \boldsymbol{\alpha}) \odot \mathbf{F}_{\text{glob}}, \tag{22}$$

where, $\odot$ denotes element-wise multiplication.

The proposed adaptive fusion strategy utilizes local structural representations to guide feature integration. Specifically, the contribution of $\mathbf{F}_{\text{loc}}$ is emphasized in boundary regions to preserve structural details, whereas the contribution of $\mathbf{F}_{\text{glob}}$ is strengthened in semantically homogeneous regions to improve contextual consistency. Consequently, HSDM achieves a dynamic balance between structural information and semantic representations, providing more discriminative features for subsequent decoding and segmentation tasks.

C. Structure-Guided Adaptive Multi-Scale Fusion Module

In traffic scenes, different semantic categories exhibit not only scale variations but also highly heterogeneous spatial distributions and contextual dependencies. Small-scale objects, such as pedestrians and distant vehicles, are often located near scene boundaries or within occluded regions, making their representation highly dependent on local details and boundary sensitivity. In contrast, large-scale structures, such as buildings and road regions, usually present spatial continuity and require global contextual information to maintain semantic consistency. Moreover, variations in object density, viewpoint, and spatial arrangement further exacerbate the imbalance between local details and global semantics, making it difficult for single-scale representations to simultaneously capture fine-grained structures and overall scene context.

Under such circumstances, naive multi-scale feature fusion may introduce undesirable interference. Specifically, excessive global information may dilute local structural cues in complex boundary regions, whereas overemphasizing local features in large homogeneous regions may undermine semantic consistency. Therefore, a structure-guided adaptive fusion strategy is required to dynamically determine the contribution of features at different scales according to spatial locations and contextual characteristics. Such a mechanism can effectively balance local discriminative capability and global semantic coherence, thereby providing robust representations for semantic segmentation in complex traffic scenarios. The overall architecture of the proposed Structure-guided Adaptive Multi-scale Fusion (SAMF) module is illustrated in Fig. 4.

1) Multi-scale feature construction and alignment: Let the input feature representation be defined as

$$\mathbf{X} \in \mathbb{R}^{H\times W\times C}. \tag{23}$$

Multi-scale feature branches are constructed through scale transformation operations:

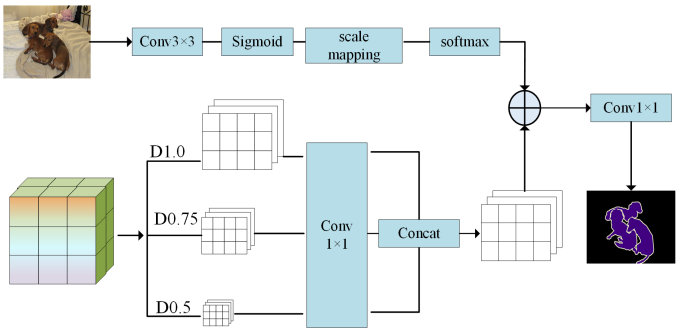


Fig. 4. Architecture of the proposed Structure-guided Adaptive Multi-scale Fusion (SAMF) module.

$$\mathbf{X}_s = D(\mathbf{X}, s), \quad s \in \{1, 0.75, 0.5\}, \tag{24}$$

where, $D(\cdot)$ denotes the scale transformation operation. Different scales correspond to different receptive field sizes. Larger scales preserve fine-grained details, whereas smaller scales facilitate the modeling of global contextual information.

To eliminate discrepancies in channel distributions across different scales, point-wise convolutions are employed for feature alignment:

$$\mathbf{X}'_s = \mathrm{Conv}_{1\times1}(\mathbf{X}_s). \tag{25}$$

Subsequently, all scale-specific features are upsampled to a unified spatial resolution:

$$\tilde{\mathbf{X}}_s = \mathrm{Up}(\mathbf{X}'_s), \tag{26}$$

resulting in a set of spatially aligned multi-scale representations.

2) Structure-guided signal modeling from the original input: To avoid information coupling among multi-scale representations during the fusion process, the proposed method does not directly derive structural responses from the fused features. Instead, the original input image is introduced as an external structural prior:

$$\mathbf{X}_0 \in \mathbb{R}^{H\times W\times C_0}, \tag{27}$$

where, C_0 denotes the number of channels in the original input.

A local convolution operation is then applied to capture structural variations:

$$\mathbf{S} = \sigma\left(\mathrm{Conv}_{3\times3}(\mathbf{X}_0)\right), \tag{28}$$

where, $\mathbf{S} \in \mathbb{R}^{H\times W\times1}$ represents the structural response map, and $\sigma(\cdot)$ denotes the Sigmoid activation function used to constrain the output range and improve training stability.

The structural response characterizes the intensity of edge and texture variations in the input image, thereby serving as a guidance signal independent of high-level semantic representations. This strategy effectively reduces mutual interference among multi-scale features during fusion.

3) Structure-guided scale response modeling: Based on the structural response map, the structure information is projected into the scale selection space:

$$\tilde{\mathbf{w}}(p) = \phi\left(\mathbf{S}(p)\right), \tag{29}$$

where, $\tilde{\mathbf{w}}(p) \in \mathbb{R}^K$ denotes the unnormalized scale response at spatial position p, K is the number of scale branches, and $\phi(\cdot)$ is implemented using a 1×1 convolution. This operation converts structural cues into scale-selection logits, establishing an explicit mapping from structural characteristics to receptive field preferences.

Through this mechanism, the model is able to adaptively adjust the importance of different scales according to local structural variations.

4) Scale normalization and competition mechanism: To ensure comparability among scale responses, a normalization operation is performed across the scale dimension:

$$w_s(p) = \frac{\exp\left(\tilde{w}_s(p)\right)}{\sum_{s'}\exp\left(\tilde{w}_{s'}(p)\right)}, \tag{30}$$

where, $w_s(p)$ denotes the relative importance of scale s at position p.

The normalization process introduces a competition mechanism among different scales, enabling the model to dynamically select the most appropriate receptive field according to spatial characteristics, thereby improving the adaptability of feature representations.

5) Structure-guided multi-scale fusion: Based on the learned scale weights, the aligned multi-scale features are aggregated as

$$\mathbf{X}_o(p) = \sum_s w_s(p)\cdot\tilde{\mathbf{X}}_s(p), \tag{31}$$

where, $\tilde{\mathbf{X}}_s$ denotes the feature representation corresponding to scale branch s.

Although all features have been spatially aligned to a unified resolution, their receptive field characteristics remain distinct. Consequently, the fusion process can be interpreted as an adaptive combination of contextual information with different spatial extents under a common coordinate system.

Since the fusion weights are derived from the original input rather than the feature representations themselves, the proposed strategy effectively alleviates feature coupling across scales and improves structural consistency in the fused representation.

Finally, a lightweight nonlinear transformation is applied to obtain the output feature representation:

$$\mathbf{X}_{\mathrm{out}} = \Psi\left(\mathbf{X}_o\right), \tag{32}$$

where, $\Psi(\cdot)$ consists of a point-wise convolution followed by an activation function for channel reorganization and feature enhancement.

Through the proposed structure-guided adaptive fusion mechanism, SAMF dynamically balances the contributions of different receptive fields according to structural characteristics in the input image. As a result, local discriminative information is preserved in fine-grained regions, while global semantic consistency is maintained in large homogeneous areas, thereby yielding more robust feature representations for traffic scene semantic segmentation.

IV. Experiments

A. Datasets and Evaluation Metrics

1) Datasets:

a) Cityscapes [17]: Cityscapes is a standard benchmark dataset for urban scene understanding and has been widely adopted in autonomous driving and semantic segmentation research. The dataset contains complex street-view images collected from 50 different cities. In total, it provides 5,000 finely annotated images with pixel-level labels, which are divided into 2,975 training images, 500 validation images, and 1,525 testing images. All images are extracted from vehicle-mounted stereo video sequences and maintain a consistent spatial resolution, ensuring comparability in semantic segmentation tasks. Cityscapes originally provides annotations for more than 30 semantic categories, covering roads, buildings, vehicles, pedestrians, and natural backgrounds. Following the common evaluation protocol, 19 representative categories are selected as the standard benchmark subset for fair comparison among different methods.

b) CamVid [18]: CamVid is one of the earliest semantic segmentation datasets designed for driving scene understanding. It is primarily used for pixel-level semantic annotation in road environments. The dataset consists of 701 manually annotated images extracted from driving videos, including 367 training images, 101 validation images, and 233 testing images. All images have a resolution of 960×720. CamVid defines 32 semantic categories covering roads, buildings, vehicles, pedestrians, traffic signs, and other common traffic scene elements.

2) Evaluation metrics: To comprehensively evaluate the effectiveness and efficiency of the proposed method, four widely used metrics are adopted, including Mean Intersection over Union (MIoU), Mean Accuracy (Mean Acc), model parameters (Params), floating-point operations (FLOPs), and Frames Per Second (FPS).

a) Mean Intersection over Union (MIoU): MIoU is the primary metric for evaluating semantic segmentation accuracy. It measures the overlap ratio between predicted segmentation results and ground-truth annotations by averaging the Intersection over Union (IoU) across all categories:

$$\mathrm{MIoU} = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{TP_i + FP_i + FN_i}, \tag{33}$$

where, N denotes the number of semantic categories, and TP_i, FP_i, and FN_i represent the numbers of true positive, false positive, and false negative pixels for the i-th category, respectively.

b) Mean Accuracy (Mean Acc): Mean Accuracy evaluates the average classification accuracy over all categories and is defined as

$$\mathrm{MeanAcc} = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{TP_i + FN_i}, \tag{34}$$

where, TP_i denotes the number of correctly classified pixels belonging to category i, and FN_i represents the number of pixels from category i that are incorrectly classified into other categories.

c) Parameters (Params): Params quantify the structural complexity and storage requirements of a model, corresponding to the total number of learnable parameters. For a convolutional layer, the number of parameters can be expressed as

$$\mathrm{Params} = K^2 \times C_{\mathrm{in}} \times C_{\mathrm{out}}, \tag{35}$$

where, K is the kernel size, and C_{in} and C_{out} denote the numbers of input and output channels, respectively.

Although a larger number of parameters generally indicates stronger representation capability, it may also increase storage costs and the risk of overfitting.

d) Floating Point Operations (FLOPs): FLOPs are used to measure the computational complexity during the forward inference stage. For convolution operations, the computational cost is defined as

$$\mathrm{FLOPs} = H \times W \times K^2 \times C_{\mathrm{in}} \times C_{\mathrm{out}}, \tag{36}$$

where, H and W denote the spatial dimensions of the output feature map. This formulation indicates that computational complexity is jointly determined by kernel size, channel dimensions, and feature resolution.

e) Frames Per Second (FPS): FPS is adopted to evaluate inference efficiency and reflects the number of images processed per second. It is defined as

$$\mathrm{FPS} = \frac{\text{Number of processed images}}{\text{Total inference time (s)}}. \tag{37}$$

A higher FPS value indicates better real-time performance.

B. Experimental Settings

During training, the Stochastic Gradient Descent (SGD) optimizer was employed to update the network parameters, while the Cross-Entropy Loss function was adopted for optimization. The batch size was set to 8, and the number of data loading workers was fixed at 4. Considering that different network architectures exhibit varying sensitivities to input resolutions, the crop size for each model was adjusted according to its

original implementation settings. In addition, the weight decay coefficient was set to 1×10^{-4}, the initial learning rate was initialized to 0.01, and the momentum parameter was fixed at 0.9 to accelerate convergence and improve training stability (Table I).

TABLE I. EXPERIMENTAL ENVIRONMENT CONFIGURATION

Category	Parameter Name	Setting Value
Hardware Environment	Operating System	Windows 11
	CPU	Intel i7-14650HX
	GPU	NVIDIA RTX 4090 (24 GB)
	Memory	24 GB
Software Environment	Programming Language	Python
	Development Environment	PyCharm

C. Ablation Studies

A series of ablation experiments were conducted on the Cityscapes dataset to comprehensively evaluate the effectiveness of each component in the proposed FSCS framework. The experimental results are summarized in Table II. As shown in the table, the complete FSCS model achieves the best segmentation performance while introducing only a slight increase in computational cost.

Specifically, incorporating the Spectral-Decoupled Adaptive Modulation (SDAM) module independently improves the mIoU and Mean Acc by approximately 1.44% and 1.21%, respectively, compared with the baseline model. This observation indicates that frequency-domain enhancement effectively strengthens the representation capability for fine-grained structures and small objects. When the Hierarchical Spatial Dependency Modeling (HSDM) module is introduced, the model further benefits from local structural consistency and global semantic propagation, achieving improved segmentation accuracy while maintaining favorable real-time performance. The Structure-guided Adaptive Multi-scale Fusion (SAMF) module introduces structural priors into the multi-scale fusion process and obtains the highest Mean Acc among the single-module variants, reaching 0.7612. Although the mIoU improvement brought by SAMF alone is relatively limited, it effectively enhances feature consistency and boundary discrimination.

Overall, the proposed FSCS framework achieves the best trade-off between segmentation accuracy, computational efficiency, and real-time inference capability. By integrating frequency-domain enhancement, hierarchical spatial dependency modeling, and structure-guided multi-scale fusion, the complete model improves segmentation performance while preserving the lightweight characteristics required for practical traffic scene applications.

TABLE II. ABLATION STUDY RESULTS ON THE CITYSCAPES DATASET

Algorithm	Params (M)	GFLOPs	mIoU	Mean Acc	FPS
Baseline	5.23	17.00	0.6781	0.7414	231.40
Baseline + SDAM	4.81	19.66	0.6925	0.7535	200.51
Baseline + HSDM	4.39	20.16	0.6891	0.7591	210.11
Baseline + SAMF	4.21	21.23	0.6879	0.7612	208.10
Ours (FSCS)	4.96	22.50	0.7146	0.7665	210.23

D. Comparison Experiments

To systematically evaluate the generalization capability and segmentation performance of the proposed method in different urban scenarios, comparative experiments were conducted on the Cityscapes and CamVid datasets. Representative semantic segmentation models, including FCN-8s, U-Net, PSPNet, and DeepLabV3+, were selected as baseline approaches. The input resolutions were set to 2048×1024 for Cityscapes and 960×720 for CamVid. To ensure fairness and reproducibility, all models were trained using identical data preprocessing procedures and optimization strategies.

1) Comparison on the cityscapes dataset: The Cityscapes dataset contains high-resolution images with complex urban structures, posing significant challenges for semantic segmentation models. As reported in Table III, the proposed FSCS framework consistently outperforms conventional methods in terms of segmentation accuracy.

Specifically, FSCS achieves an mIoU of 0.7146, outperforming FCN-8s, U-Net, and PSPNet by approximately 10.75%, 9.41%, and 7.49%, respectively. Compared with DeepLabV3+, FSCS yields an absolute mIoU improvement of 3.65%, demonstrating the effectiveness of integrating frequency-domain enhancement, hierarchical spatial modeling, and structure-guided multi-scale fusion.

Regarding model complexity, FSCS contains only 4.96M parameters, which is lower than that of DeepLabV3+ and substantially smaller than those of FCN-8s and PSPNet. Although the computational cost increases slightly compared with DeepLabV3+, the proposed method still maintains a high inference speed of approximately 210 FPS. These results indicate that FSCS achieves a favorable balance between segmentation accuracy and computational efficiency, making it suitable for real-time semantic segmentation tasks in autonomous driving applications.

TABLE III. PERFORMANCE COMPARISON ON CITYSCAPES DATASET

Algorithm	Params (M)	GFLOPs	mIoU	Mean Acc	FPS
FCN-8s [18]	134.46	223.25	0.6071	0.7275	34.52
U-Net [19]	26.36	228.24	0.6205	0.7310	56.16
PSPNet [2]	51.44	190.34	0.6397	0.7365	64.70
DeepLabV3+ [20]	5.23	17.00	0.6781	0.7414	231.40
Ours (FSCS)	4.96	22.50	0.7146	0.7665	210.23

2) Comparison on the camvid dataset: The CamVid dataset contains fewer training samples but exhibits a relatively balanced category distribution, making it suitable for evaluating the generalization ability of segmentation models under limited data conditions. As shown in Table IV, the proposed FSCS framework consistently achieves superior performance compared with existing approaches.

In particular, FSCS demonstrates enhanced segmentation capability for small objects, such as pedestrians and traffic signs, as well as slender structures including curbs and poles. The proposed method attains the highest mIoU and Mean Acc among all competing methods, achieving values of 0.6342 and 0.7325, respectively. Compared with DeepLabV3+, FSCS yields absolute improvements of 1.65% in mIoU and 1.41% in Mean Acc, indicating that the proposed frequency-spatial collaborative framework remains effective even on relatively small-scale datasets.

In terms of computational efficiency, FSCS maintains a lightweight architecture with only 4.96M parameters and

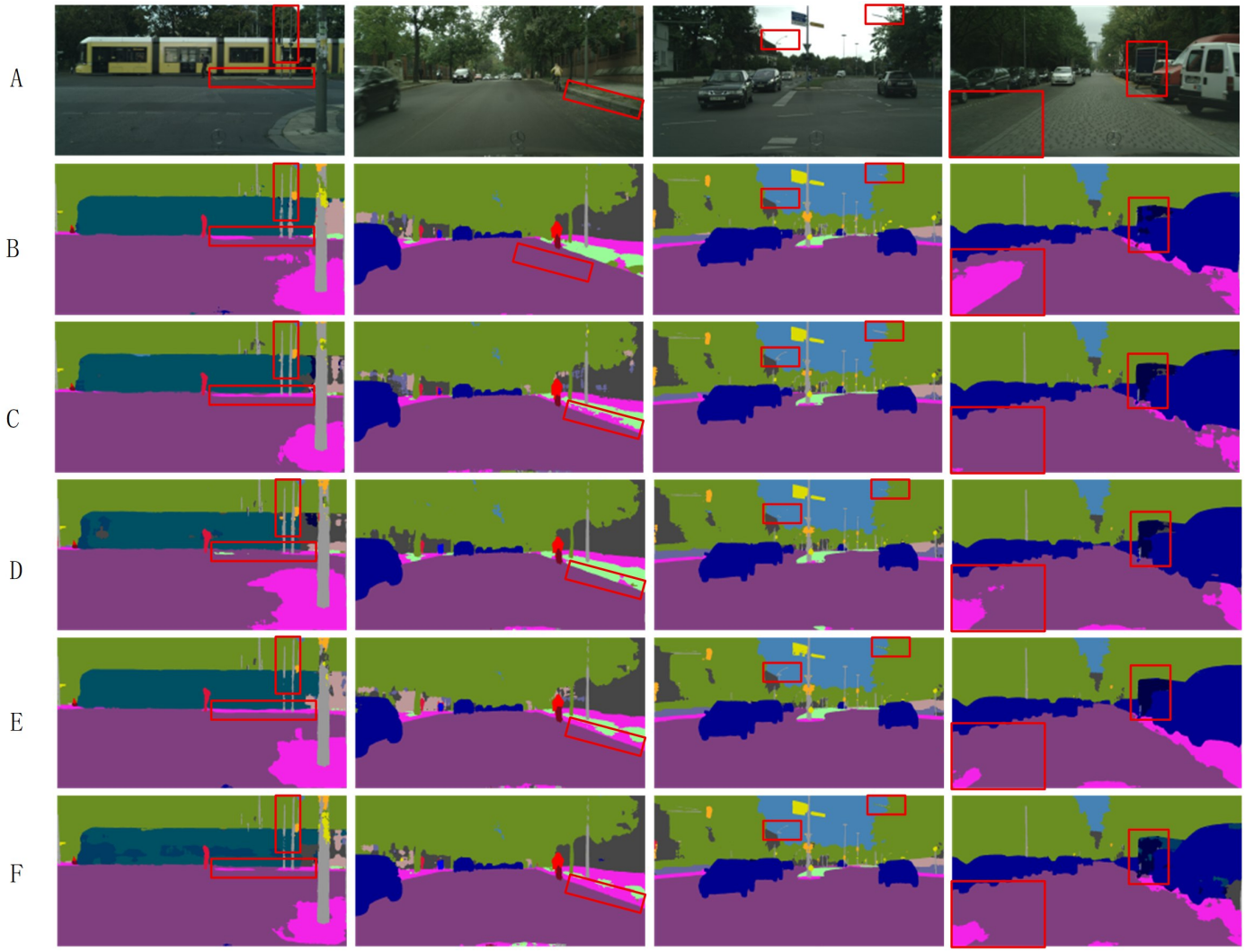


Fig. 5. Qualitative comparison of semantic segmentation results produced by different methods on the Cityscapes dataset.

18.86 GFLOPs. Despite a moderate increase in computational complexity compared with DeepLabV3+, the proposed model still achieves an inference speed of approximately 213 FPS, substantially outperforming FCN-8s, U-Net, and PSPNet. These results demonstrate that FSCS successfully balances segmentation accuracy and inference efficiency, making it well suited for real-time traffic scene understanding applications.

TABLE IV. PERFORMANCE COMPARISON ON THE CAMVID DATASET

Algorithm	Params (M)	GFLOPs	mIoU	Mean Acc	FPS
FCN-8s	134.46	182.12	0.5613	0.6914	38.33
U-Net	26.36	196.78	0.5910	0.7012	68.29
PSPNet	51.44	162.39	0.6017	0.7132	90.46
DeepLabV3+	5.22	14.92	0.6177	0.7184	235.48
Ours (FSCS)	4.96	18.86	0.6342	0.7325	213.45

3) Comprehensive analysis: The experimental results on both the Cityscapes and CamVid datasets demonstrate that the proposed FSCS framework achieves stable and competitive segmentation performance under different image resolutions and scene complexities. The performance gains can be primarily attributed to the complementary effects of the three proposed modules.

Specifically, SDAM performs amplitude–phase decoupling and adaptive spectral modulation in the frequency domain, effectively enhancing the representation capability for small objects and fine-grained structures. HSDM models spatial dependencies hierarchically through local structural consistency constraints and global semantic propagation, thereby preserving spatial continuity while capturing long-range contextual relationships. Furthermore, SAMF introduces structural priors into the multi-scale fusion process, dynamically integrating features with different receptive fields to improve the segmentation quality of object boundaries and slender structures.

The complete FSCS framework consistently outperforms all individual module variants, indicating that the three modules provide complementary rather than redundant contributions. Although the proposed method incurs a slight increase in computational cost compared with the baseline, it achieves substantial improvements in segmentation accuracy while maintaining high inference efficiency. These findings verify the effectiveness of the proposed frequency-spatial collaborative optimization strategy for real-time semantic segmentation in

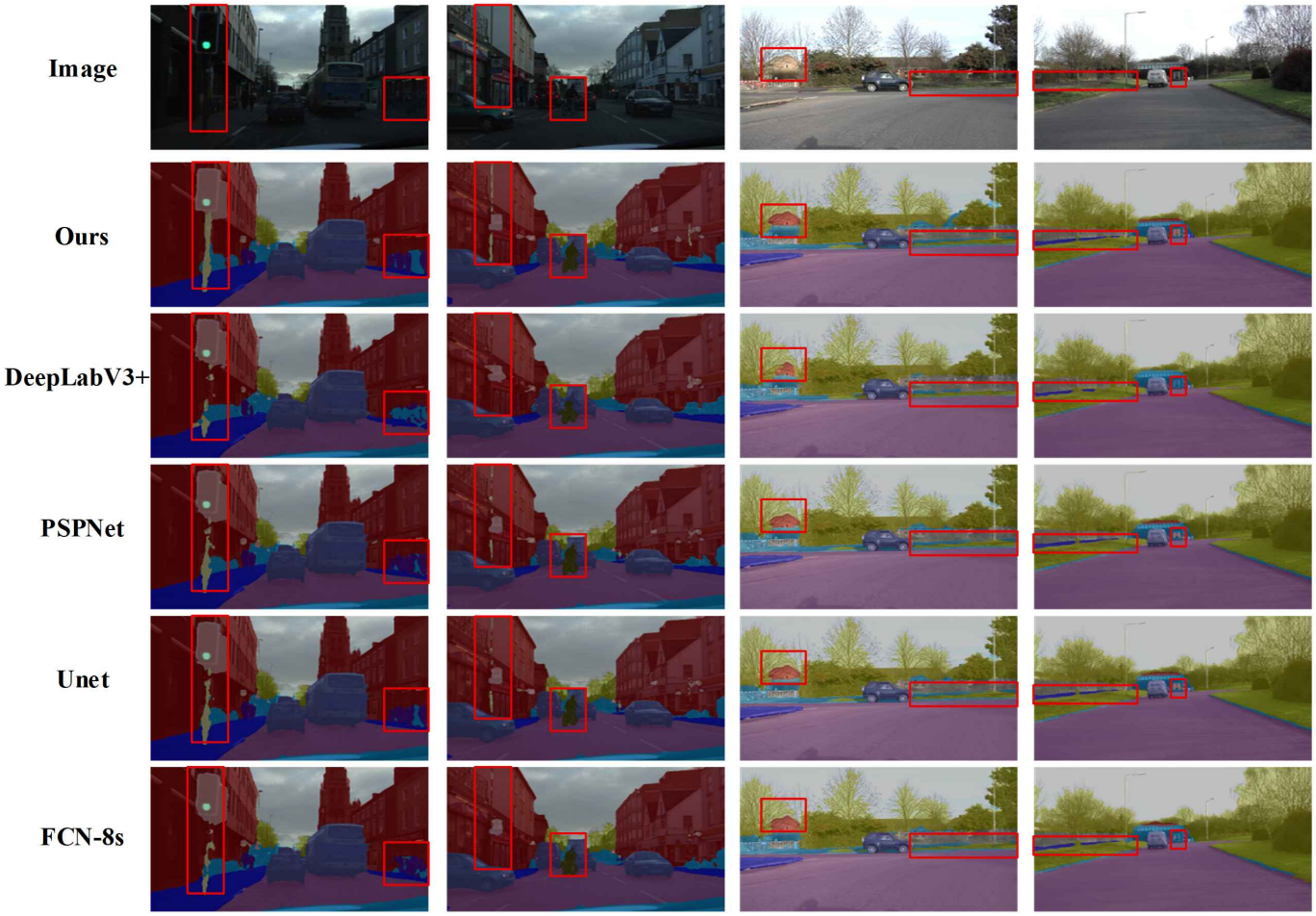


Fig. 6. Qualitative comparison of semantic segmentation results produced by different methods on the CamVid dataset.

complex traffic environments.

4) Qualitative comparison: Fig. 5 and 6 present qualitative comparisons of semantic segmentation results obtained by different methods on the Cityscapes and CamVid datasets, respectively.

As illustrated in Fig. 5, the baseline model exhibits noticeable omissions and blurred boundaries when segmenting fine-grained objects and distant targets. U-Net achieves satisfactory performance on large-scale structures but struggles to accurately delineate small objects and elongated regions. FCN-8s demonstrates the weakest performance in terms of small-object recognition and boundary localization, while also exhibiting slight confusion in certain large-scale regions. PSPNet improves the segmentation quality of fine structures to some extent; however, misclassification and omission of distant objects remain evident.

In contrast, the proposed FSCS framework produces the most accurate segmentation results within the highlighted regions. Specifically, FSCS successfully identifies small objects, slender structures, and distant vehicles while preserving the continuity of large-scale semantic regions such as roads and buildings. These observations further confirm that the collaborative optimization of frequency-domain enhancement, hierarchical spatial modeling, and structure-guided multi-scale fusion effectively improves both local discrimination capability and global semantic consistency.

Similarly, the visualization results on the CamVid dataset, shown in Fig. 6, further validate the robustness and generalization capability of the proposed method under different scene characteristics and dataset scales.

V. Conclusion

In this study, a Frequency-Spatial Collaborative Segmentation framework (FSCS) is proposed to improve the performance of DeepLabV3+ for traffic scene semantic segmentation. To address the limitations of conventional convolution-based segmentation methods in modeling fine-grained structures, long-range dependencies, and multi-scale contextual information, three dedicated modules are introduced.

First, the Spectral-Decoupled Adaptive Modulation (SDAM) module explicitly performs amplitude–phase decoupling in the frequency domain and adopts adaptive spectral modulation to enhance informative frequency components while preserving structural consistency, thereby improving the representation capability for small objects and boundary regions. Second, the Hierarchical Spatial Dependency Modeling (HSDM) module progressively captures spatial dependencies from local structural consistency to global

semantic propagation, enabling effective long-range context modeling while alleviating the structural ambiguity caused by direct global interactions. Finally, the Structure-guided Adaptive Multi-scale Fusion (SAMF) module introduces structural priors derived from the original input image to dynamically guide the fusion of multi-scale features, thereby enhancing the segmentation quality of slender structures and complex boundaries.

Experimental results on the Cityscapes and CamVid datasets demonstrate that the proposed FSCS framework consistently outperforms several representative semantic segmentation methods while maintaining favorable real-time performance and computational efficiency. The ablation studies further verify that the three proposed modules contribute complementary improvements and jointly enhance segmentation performance in complex traffic environments.

In future work, more efficient frequency-domain representation strategies and lightweight spatial modeling mechanisms will be investigated to further reduce computational complexity while improving the segmentation accuracy of small-scale and cross-scale objects. In addition, extending the proposed framework to more challenging scenarios, such as adverse weather conditions and domain-generalized traffic scene understanding, remains an important direction for future research.

REFERENCES

[1] Y. Wang *et al.*, "An automated learning method of semantic segmentation for train autonomous driving environment understanding," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 4, pp. 6913–6922, 2024.

[2] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2881–2890.

[3] C. Song *et al.*, "Road scene semantic segmentation based on mpnet," *Electronics*, vol. 14, no. 13, p. 2565, 2025.

[4] L. Shan, X. Li, and W. Wang, "Decouple the high-frequency and low-frequency information of images for semantic segmentation," in *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 1805–1809.

[5] T. Zhang *et al.*, "Adaptive semantic segmentation of traffic scenes via frequency domain analysis," in *2025 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2025.

[6] C. Ge *et al.*, "Sf-net: Spatial-frequency feature synthesis for semantic segmentation of high-resolution remote sensing imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2026.

[7] Z. Yang *et al.*, "Ffr: Frequency feature rectification for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[8] J. Huang *et al.*, "Fsdr: Frequency space domain randomization for domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6891–6902.

[9] H. Sima, M. Gao, and L. Liu, "A real-time semantic segmentation network leveraging spatial and contextual features for enhanced scene understanding," *Intelligent Systems with Applications*, vol. 27, p. 200542, 2025.

[10] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 325–341.

[11] H. Wang *et al.*, "Context-aware method for small object segmentation in road scenes," in *2022 International Conference on Advanced Robotics and Mechatronics (ICARM)*. IEEE, 2022.

[12] X. Wu *et al.*, "Wavelet attention fusion for semi-supervised ultrasound segmentation," *Biomedical Signal Processing and Control*, vol. 117, p. 109566, 2026.

[13] Y. Zhang *et al.*, "Segcft: Context-aware fourier transform for efficient semantic segmentation," *Neurocomputing*, vol. 596, p. 127946, 2024.

[14] H. Niu *et al.*, "Exploring semantic consistency and style diversity for domain generalized semantic segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 6, 2025.

[15] Y. Bai, Z. Huang, L. Shen *et al.*, "Multi-scale feature aggregation by cross-scale pixel-to-region relation operation for semantic segmentation," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 5889–5896, 2021.

[16] T. Li, X. Tang, H. Wang *et al.*, "Occupancy network-based 3d semantic segmentation method for autonomous driving images," *Laser & Optoelectronics Progress*, vol. 63, no. 4, pp. 192–203, 2026.

[17] M. Cordts, M. Omran, S. Ramos *et al.*, "The cityscapes dataset for semantic urban scene understanding," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3213–3223, 2016.

[18] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognition Letters*, vol. 30, no. 2, pp. 88–97, 2009.

[19] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

LAURA: An Awareness-Oriented Explainable AI Framework for Academic Stress in Higher Education

Ahmed Almathami[1], Richard Stone[2]
Department of Information Systems, King Abdulaziz University, Rabigh, Saudi Arabia[1]
Human-computer Interaction Department, Iowa State University, Ames, USA[1]
Industrial and Manufacturing Systems Engineering Department, Iowa State University, Ames, USA[2]

***Abstract*—Academic stress represents a significant challenge in higher education, especially for students navigating unfamiliar academic environments. In such contexts, difficulties in understanding expectations, feedback, and performance standards can affect engagement, participation, academic performance, and decision-making. Despite significant advances in learning analytics and artificial intelligence, current educational systems remain limited in their ability to support students' sense-making of their academic conditions. Existing systems predominantly focus on prediction and performance monitoring, offering limited support for interpretation and awareness, which are essential for understanding academic stress, particularly in complex or unfamiliar learning environments. To address this limitation, this study introduces the LAURA framework (Learning Context Modeling, Academic Stress Assessment, Understanding-Oriented Explanation, Reflection Support, and Adaptive Awareness Feedback), an awareness-oriented framework designed to support students in understanding their academic stress through interaction. Unlike traditional AI approaches that prioritize prediction, LAURA conceptualizes academic stress as an interpretive process and positions explainable AI as a mechanism to support students' awareness. The framework integrates transparent modeling, explanation, and structured reflection within a unified interaction cycle, enabling students to interpret academic signals, examine contributing factors, and develop a more informed understanding of their conditions. Grounded in appraisal-based stress theory and principles of human-centered AI, LAURA extends existing approaches by embedding explanation and reflection as core components of system design rather than auxiliary features. In this way, it shifts the role of explainable AI from revealing model behavior to supporting meaning-making processes within educational contexts. The proposed framework contributes a new perspective on AI-supported educational systems by emphasizing awareness as the primary outcome. It provides a conceptual and design foundation for developing student-facing systems that not only analyze academic conditions but also support students in making sense of them, particularly in diverse and complex educational environments. As a conceptual framework, LAURA provides a foundation for future implementation and empirical validation in educational settings.**

***Keywords*—Explainable Artificial Intelligence (XAI); academic stress; student awareness; interpretable models; human-centered AI; higher education**

I. Introduction

Academic stress is a well-established challenge in higher education, influencing students' well-being, engagement, and academic performance. While a certain level of stress can support motivation, prolonged or poorly managed stress may negatively affect students' learning experiences and academic decision-making [1]. In response, universities have increasingly adopted data-driven approaches, including learning analytics and artificial intelligence (AI) systems, to monitor and understand students' academic conditions [2].

These systems provide robust capabilities for analyzing large-scale educational data and identifying patterns in performance, engagement, and academic risk. In particular, AI-based approaches have been widely used for prediction, classification, and early identification of at-risk students. However, despite these advancements, most existing systems remain primarily oriented toward outcome estimation, with limited support for how students interpret their academic situations [3], [4].

This limitation is particularly important in the context of international students, who often face additional challenges when navigating unfamiliar academic environments. Differences in academic communication practices, educational systems, and implicit expectations can make it difficult to accurately interpret feedback, assessment criteria, and performance indicators. Prior work has shown that academic stress in such contexts is not only driven by workload or outcomes, but also by difficulties in interpreting academic signals [5]. This highlights a critical need for systems that support students not only in receiving information, but also in meaningfully interpreting that information.

Recent research has explored the role of explainable artificial intelligence (XAI) in improving the transparency of AI systems in education [6]. While these approaches enhance visibility into how models operate, explainability is often treated as a technical property that reveals model behavior [7], [8]. As a result, students may still face challenges in connecting system outputs to their own academic experiences. This indicates that transparency alone is insufficient to support meaningful understanding.

Based on these limitations, a clear research gap emerges: current AI-supported educational systems provide limited support for interpretation and awareness, particularly in contexts where students must actively make sense of complex and unfamiliar academic conditions. Addressing this gap requires moving beyond prediction and transparency to support students in interpreting academic information and engaging in reflection that promotes awareness of their academic conditions.

Our previous work highlighted this research gap through a comprehensive review of explainable AI and academic stress among international university students, identifying the need for student-facing systems that support awareness and

interpretation rather than prediction alone [5]. However, that study did not propose a framework architecture or specify how such awareness-oriented support could be realized in practice. The present study addresses this gap by proposing the LAURA framework, which comprises a five-phase conceptual architecture that integrates learning context modeling, academic stress assessment, understanding-oriented explanation, reflection support, and adaptive awareness feedback into a unified interaction cycle.

Building on this foundation, the present study introduces the LAURA conceptual framework, an awareness-oriented approach designed to support students in developing awareness of their academic stress conditions. The framework conceptualizes academic stress as an interpretive process in which students construct meaning through interaction with system feedback. Rather than treating stress estimation as the final system outcome, LAURA is designed to help students understand how different academic factors contribute to their stress experiences and to promote awareness through explanation and reflection.

The main contributions of this study are fourfold. First, it advances an awareness-oriented conceptualization of academic stress in AI-supported educational systems by emphasizing interpretation over prediction. Second, it introduces the LAURA framework as a five-phase conceptual architecture that extends existing learning analytics and educational XAI approaches by integrating learning context modeling, academic stress assessment, understanding-oriented explanation, reflection support, and adaptive awareness feedback within a unified interaction cycle. Third, it repositions explainable AI from a mechanism for revealing model behavior to a form of cognitive support that enables students' interpretation and meaning-making. Finally, it provides awareness-oriented design principles for developing student-facing AI systems, particularly for students navigating unfamiliar academic environments, including international students.

The remainder of this study is organized as follows:

- Section II: Related Work reviews existing studies on academic stress, student awareness, and AI-based approaches in education.
- Section III: Framework Development Methodology presents the conceptual design approach, architectural structure of LAURA, and awareness-oriented design logic.
- Section IV: Conceptual Interaction and Awareness Development describes the interaction process and how awareness is developed through the framework.
- Section V: Discussion discusses the theoretical, design, and international student implications of the framework, presents its limitations, and outlines research implications.
- Section VI: Conclusion summarizes the main contributions of the study and highlights the significance of the proposed framework.

II. Related Work

This section reviews the main research areas related to the development of the LAURA framework. The discussion focuses on four themes: academic stress in higher education, artificial intelligence approaches for analyzing academic stress indicators, learning analytics dashboards and student awareness, and the role of explainable artificial intelligence in educational systems. Examining these areas helps situate the proposed framework within the current literature and highlights the limitations that motivate the development of LAURA.

A. Academic Stress in Higher Education

Academic stress has been widely recognized as an important factor influencing students' learning experiences, well-being, and academic performance. University students frequently encounter multiple academic demands, including coursework, assignments, examinations, and performance expectations. These demands may create pressure that affects both academic engagement and students' perceptions of their learning environment.

Previous research has identified several common contributors to academic stress in higher education. Among the most frequently reported factors are perceived academic workload, deadline pressure, unclear academic expectations, and difficulties in understanding course materials [5], [9], [10]. Studies show that academic demands, assessment pressure, and time constraints represent major sources of stress among students in higher education. Other research also indicates that uncertainty regarding grading criteria and course expectations can increase academic pressure among students.

In addition to workload-related factors, interpretive uncertainty plays an important role in academic stress experiences. Students may experience additional pressure when they struggle to interpret academic signals such as instructor feedback, grading standards, or course expectations [11]. These challenges may become more pronounced in international learning environments where students encounter unfamiliar educational systems or language-related comprehension difficulties [12].

Students may also interpret academic signals differently depending on their prior experiences, confidence levels, and familiarity with the learning environment. Such differences in interpretation may influence how academic demands are perceived and how strongly students experience academic pressure [13].

Academic stress may be particularly significant for international students who study in unfamiliar educational environments [14]. International students often face additional challenges, including language-related comprehension difficulties, unfamiliar academic expectations, and differences in instructional communication styles. These factors may increase uncertainty when interpreting academic signals such as feedback, grading standards, and course requirements. As a result, international students may experience greater academic pressure when they are unsure how to interpret or respond to academic expectations [15].

Although traditional academic stress research provides valuable insights into the factors influencing student experiences, most studies rely on surveys or psychological measure-

ment instruments. As a result, these approaches often offer limited technological support to help students actively interpret and understand their experiences of academic stress during the learning process.

B. Artificial Intelligence Approaches for Analyzing Academic Stress Indicators

Recent advances in artificial intelligence and machine learning have introduced new opportunities to analyze student behavior and indicators of academic stress in educational environments [16]. Researchers have increasingly explored how machine learning models can identify patterns related to academic workload, engagement, and stress-related experiences among students [17].

Several studies have applied machine learning techniques to analyze stress-related indicators collected from university students. These studies demonstrate that machine learning models can identify relationships between contextual academic factors and students' reported stress experiences. By analyzing academic and behavioral data, AI models can reveal patterns associated with students' learning conditions and stress indicators.

Recent literature reviews also highlight the growing interest in applying artificial intelligence techniques to analyze indicators of student well-being and academic stress in higher education. These studies show that AI-based analytical models can support the analysis of complex relationships between academic factors and student experiences [18].

Despite these developments, many artificial intelligence approaches primarily focus on identifying patterns or estimating stress levels within student datasets [19]. In many cases, the models serve primarily as analytical tools for researchers or institutional monitoring rather than as interactive systems that support students' understanding of their academic experiences [16].

Consequently, although machine learning models can reveal relationships between academic factors and stress indicators, students may still struggle to understand how these factors shape their individual experiences. This limitation highlights the importance of designing AI-supported systems that not only analyze academic data but also help students interpret the factors that influence their learning conditions.

C. Learning Analytics Dashboards and Student Awareness

Learning analytics dashboards have become a widely used approach for presenting educational data and supporting student awareness of learning progress. These dashboards typically visualize information such as grades, participation levels, assignment completion rates, and learning activity indicators. By presenting such information, dashboards aim to help students monitor their academic progress and gain insight into their learning performance [20].

A growing body of research has examined the design and effectiveness of learning analytics dashboards. Many dashboards successfully provide visual summaries of student data and help learners observe their academic progress. However, research also indicates that dashboards often provide limited support for helping students interpret the meaning of these indicators [21], [22].

One of the main challenges identified in dashboard research is the gap between data visualization and meaningful interpretation [23]. Although dashboards present useful academic indicators, students may still struggle to understand how different academic factors contribute to their experiences. As a result, observing performance indicators does not always lead to deeper awareness of learning conditions [24].

Recent studies suggest that explanation mechanisms can help address this limitation. When systems provide explanations that clarify how different indicators relate to academic conditions, students are better able to interpret dashboard information and understand the factors influencing their learning experiences [25].

Research also highlights the importance of reflection in supporting student awareness and self-regulated learning. Reflection prompts encourage students to actively interpret academic information rather than simply observing it. When students reflect on system feedback, they are more likely to connect learning indicators with their own behaviors and learning strategies. Studies have shown that systems incorporating reflection mechanisms can improve students' understanding of their learning processes and support more informed academic decision-making [24], [26].

These findings indicate that learning analytics systems can more effectively support student awareness when they combine data visualization with explanatory and reflective support. Because explanation plays a key role in helping users interpret analytical outputs, recent research has increasingly explored integrating explainable artificial intelligence into educational systems.

D. Explainable Artificial Intelligence in Education

Explainable artificial intelligence has received increasing attention as researchers seek to make AI systems more transparent and understandable for users. In educational contexts, explainability is particularly important because students and instructors often need to understand how analytical outputs are generated and how different variables influence those outputs [6].

Recent studies emphasize that explainable AI should support human understanding rather than only providing technical transparency. When explanations are presented clearly and meaningfully, users are more likely to interpret system outputs correctly and incorporate that information into their decision-making [27].

Interpretable modeling approaches play an important role in this context. Interpretable models allow users to understand how individual input variables contribute to the final analytical result [28]. When applied in educational environments, such models can help students see how different academic factors influence the results produced by analytical systems [29].

Despite the increasing interest in explainable AI, many educational applications still rely on complex analytical models that may be difficult for students to interpret. This limitation highlights the need to design AI-supported educational systems

that combine transparent modeling with explanations that are meaningful and accessible for learners [30].

E. Positioning LAURA Within Existing Approaches

Existing research on academic stress and learning analytics has introduced a variety of systems that analyze student behavior or present academic indicators. However, these systems are often designed for monitoring or prediction purposes rather than for supporting students' understanding of their academic conditions [31]. As a result, students may receive analytical outputs without sufficient support to interpret how different academic factors contribute to their stress experiences.

To position the proposed framework within this research landscape, Fig. 1 compares the LAURA framework with representative approaches, including learning analytics dashboards, AI-based stress analysis systems, and traditional academic stress models.

The comparison highlights several important gaps in existing approaches. Learning analytics dashboards typically present academic indicators but require students to interpret them independently. AI-based stress analysis systems often focus on predictive modeling, while providing limited support for explanation or reflection.

In contrast, the LAURA framework integrates interpretable modeling, explanation support, and structured reflection within a unified architecture. This design enables students not only to observe academic indicators but also to understand how different academic factors contribute to their academic stress.

By integrating explanation and reflection into an iterative feedback process, LAURA goes beyond prediction-oriented systems by fostering a deeper awareness among students of their academic stress.

F. Research Gap and Motivations

The reviewed literature shows important progress in academic stress research, artificial intelligence applications, learning analytics dashboards, and explainable AI in education. Traditional academic stress studies provide a strong theoretical understanding of stress factors, while AI-based approaches enable the analysis of complex student data. Learning analytics dashboards, on the other hand, support the visualization of academic indicators.

Despite these contributions, these approaches are often developed separately and focus on different objectives. Many AI-based systems emphasize data analysis or prediction, whereas learning analytics dashboards primarily present performance indicators with insufficient support for interpretation. As a result, students may observe academic information but still struggle to understand how different academic conditions contribute to their experiences of stress.

This gap indicates the need for systems that not only present or analyze data, but also support students in interpreting academic signals and reflecting on their learning situations. In particular, integrating interpretable models, explanations of contributing factors, and structured reflection can provide more meaningful support for students' awareness.

To address this need, the proposed LAURA framework integrates interpretable modeling, explanation mechanisms, and reflection prompts within a unified architecture. This design aims to support students in understanding their academic conditions and in developing awareness of the factors influencing their academic stress.

The following section presents the LAURA framework and explains its conceptual structure and development methodology.

III. LAURA Framework Development Methodology

This section presents the development methodology of the LAURA framework and explains how its main components are conceptually organized. The focus is not on technical implementation but on the design rationale and theoretical foundations that guided the framework's development. Accordingly, this work establishes the conceptual architecture and theoretical design rationale of the LAURA framework, providing a foundation for its future implementation and empirical validation.

The design of LAURA is informed by research on academic stress, interpretive challenges in higher education, explainable artificial intelligence (XAI), and reflection in learning analytics. These research areas highlight a common limitation in many existing systems. Although academic indicators are often presented to students, such systems provide limited support for helping students interpret how these indicators relate to their academic situations.

To address this limitation, LAURA is designed as an awareness-oriented explainable architecture. The framework integrates interpretable modeling, explanation of contributing academic factors, and structured reflection prompts that guide students in interpreting their academic conditions. Rather than focusing only on estimation, LAURA aims to support meaningful understanding and academic self-awareness.

A. Conceptual Design Approach

The conceptual design of LAURA follows a theory-informed synthesis process. Rather than starting with technical modeling techniques, the framework development begins with an educational objective: supporting students in interpreting the academic conditions that influence their stress experiences.

First, academic stress was conceptualized as an appraisal process. According to stress appraisal theory, stress arises when individuals evaluate academic demands relative to their perceived coping resources [34]. In higher education environments, perceived workload, deadline pressure, unclear expectations, and performance uncertainty are frequently identified as major contributors to academic stress. Recent studies also indicate that uncertainty regarding grading criteria and academic expectations can increase stress levels among university students [13].

Second, research indicates that students may experience higher stress when academic signals are unclear or difficult to interpret. For example, unclear feedback, unfamiliar evaluation systems, or language-related comprehension difficulties may

Characteristic	Primary Objective	Explanation	Reflection	Model Transparency	User Role	Feedback	Contribution Perspective
LAURA Framework	Academic stress awareness	Central explanatory component	Structured reflection prompts	Interpretable additive modeling	Active interpretation	Iterative awareness feedback	Integrates interpretable analysis, explanation, reflection, and adaptive awareness support
Learning Analytics Dashboards	Learning progress monitoring	Limited or indicator-based	Often implicit	Indicator-based analytics	Passive observation	Progress visualizations	Focus on monitoring academic indicators rather than understanding stress conditions
AI-Based Stress Analysis Systems	Stress detection or prediction	Rare or minimal	Usually absent	Often black-box ML models	Limited involvement	Stress prediction alerts	Emphasis on prediction accuracy rather than awareness or reflection
Traditional Academic Stress Models	Psychological explanation of stress	Core theoretical explanations	External reflection processes	Highly interpretable statistical models	Questionnaire participation	Psychological assessment feedback	Provide theoretical insight but lack interactive technological support

Fig. 1. Comparison of the proposed LAURA framework with representative approaches for academic stress support. Adapted from [20], [23], [18], [16], [32], [33].

increase confusion and uncertainty among students, particularly in international educational environments. In these situations, the issue is not only the academic demand itself, but also the difficulty of correctly interpreting academic signals [35].

Third, recent developments in explainable artificial intelligence emphasize the importance of transparency when AI systems interact directly with users. When the goal of the system is to support human understanding rather than purely optimize predictive performance, interpretable models are generally preferred over complex black-box algorithms. In educational contexts, clear explanations have been shown to improve user trust and understanding of system outputs [6] [36].

Finally, research on learning analytics dashboards demonstrates that presenting data alone is often insufficient for supporting awareness. Students benefit more when systems include structured reflection prompts that encourage interpretation and self-regulated learning. Reflection supports metacognitive thinking and helps students connect system feedback with their own academic behaviors [37].

By integrating insights from these research areas, the LAURA framework was developed as an awareness-oriented system that combines academic context modeling, transparent estimation, explanation, and reflection within a continuous interaction cycle [5].

B. Architectural Structure of LAURA

Fig. 2 illustrates the overall architecture of the LAURA framework. The framework is designed as an awareness-oriented interaction cycle that integrates academic context modeling, interpretable stress estimation, explanation, reflection, and adaptive feedback. Rather than treating academic stress as a static outcome, LAURA conceptualizes awareness as a continuous process in which students interpret academic conditions, receive explanations, reflect on contributing factors, and progressively refine their understanding.

Structurally, the LAURA framework operates as a multi-stage, awareness-oriented cycle comprising five interconnected

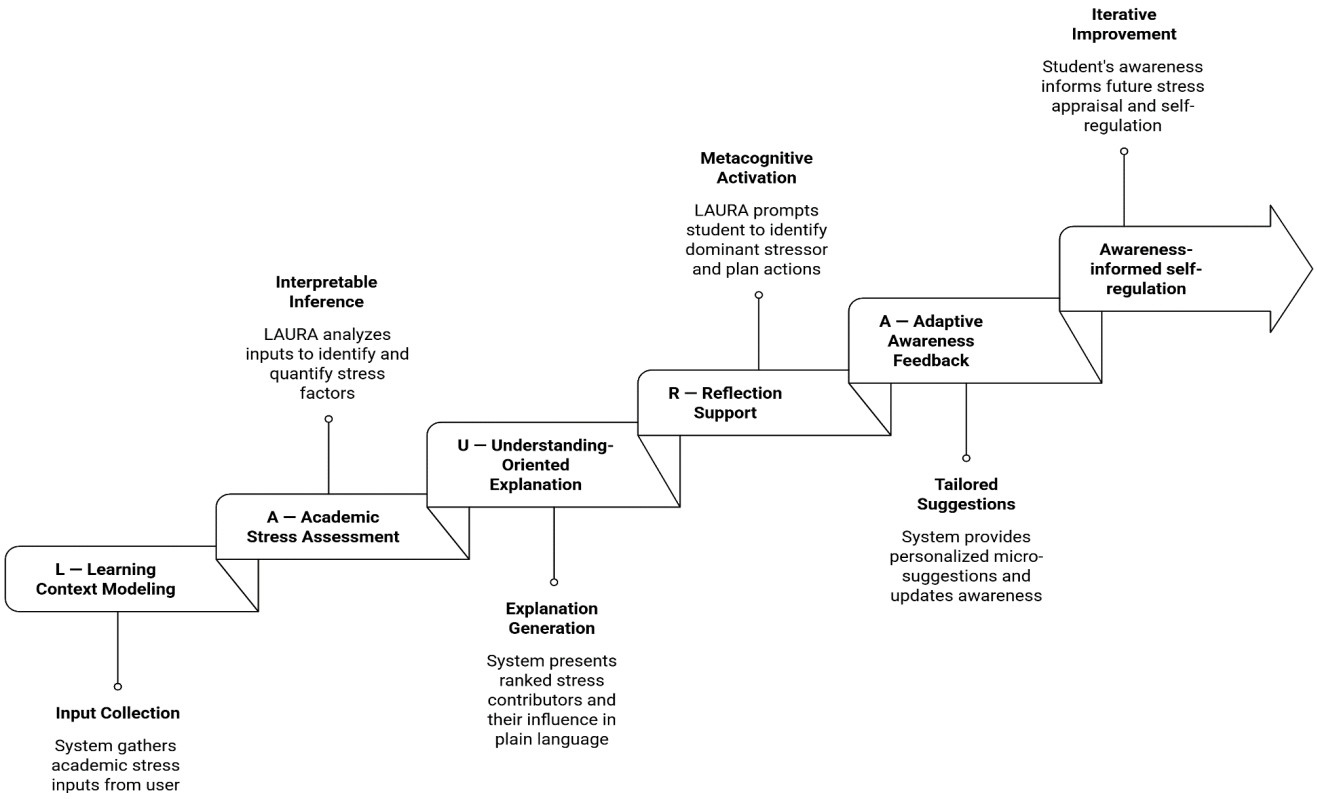


Fig. 2. The LAURA framework: An awareness-oriented explainable interaction cycle.

phases.

1) Phase 1: Learning context modeling: The process begins with Learning Context Modeling, where key academic indicators such as perceived workload, deadline pressure, clarity of academic expectations, and comprehension or interpretation difficulty are collected to represent the student's academic environment. These factors capture how students perceive and evaluate their academic conditions, rather than the objective workload alone.

This phase is designed as a structured input layer that reflects students' appraisal of academic demands. Consistent with research on academic stress, these indicators represent common sources of uncertainty in higher education contexts, particularly for students navigating unfamiliar academic environments [9] [38] [13].

2) Phase 2: Academic stress assessment: The academic indicators derived from the Learning Context Modeling phase serve as inputs for the Academic Stress Assessment phase. Depending on the educational context and the available data sources, these indicators may include course workload, assessment demands, class participation, learning engagement, assignment completion, and student self-reported academic experiences. In this stage, students' academic stress level is estimated using an interpretable approach that makes the contributions of individual academic factors transparent in the overall stress estimation [17].

To support interpretability, this phase adopts an interpretable modeling approach that enables transparent representation of how individual academic factors contribute to the overall stress estimation. Such an approach may be implemented using inherently interpretable techniques, including additive models, consistent with recommendations from interpretable machine learning literature that emphasize transparency in user-facing systems [39]. Importantly, this stage is intended to support the structured interpretation of students' academic conditions rather than to provide clinical prediction.

3) Phase 3: Understanding-oriented explanation: The output of the Academic Stress Assessment phase is translated into structured explanations that are accessible to students. Instead of presenting raw numerical results, the system identifies and highlights the key academic factors that contribute to the estimated stress level and communicates their influence in clear, simple academic language.

This phase is designed to support interpretation rather than only presenting information. Research in human-centered explainable AI emphasizes that explanations should help users make sense of system outputs, not only reveal model behavior [40]. Explanations that align with human reasoning have been shown to improve both understanding and trust in AI systems [6].

Accordingly, the explanation process in LAURA follows three key principles: 1) attribution clarity, by identifying the most influential academic factors; 2) directional influence, by explaining how each factor increases or decreases stress; and 3) contextual meaning, by presenting explanations in relation to real academic situations.

By combining these elements, this phase reduces interpretive ambiguity and supports students in developing a clearer understanding of their academic conditions.

4) Phase 4: Reflection support: The Reflection Support phase encourages active engagement with the system's explanations. While explanations provide important information, they do not automatically lead to awareness. Therefore, this phase introduces structured reflection prompts to guide students in interpreting and internalizing the results [41].

Research on self-regulated learning highlights the importance of metacognitive processes in supporting adaptive academic behavior [42]. In addition, learning analytics studies show that dashboards integrating reflective prompts lead to greater engagement than systems that present only information [24].

Within LAURA, reflection is guided through simple and focused prompts that encourage students to 1) identify the main contributing academic factors, 2) consider how their recent academic experiences relate to these factors, and 3) think about possible small adjustments in their study strategies.

By supporting this reflective process, the system helps transform explanations into meaningful awareness. This phase enables students to move from passive observation to active interpretation of their academic conditions [43].

5) Phase 5: Adaptive awareness feedback: The Adaptive Awareness Feedback phase provides contextual and personalized suggestions based on the identified contributors to academic stress. These suggestions are designed to support awareness-informed academic adjustments rather than prescribe fixed or rigid interventions [44].

This phase closes the framework's awareness loop by linking explanation and reflection to actionable feedback. Through this process, students are encouraged to gradually adjust their academic strategies in response to the identified factors. Such feedback supports self-regulated learning by helping students translate understanding into practice [45].

The framework operates as an iterative process in which increased awareness can influence how students evaluate future academic demands, consistent with appraisal-based stress theory. Over time, this process may support improved self-regulation and reduce interpretive uncertainty.

By integrating modeling, explanation, reflection, and adaptive feedback, LAURA conceptualizes explainable AI as a cognitive support tool that promotes awareness and understanding rather than performance monitoring.

Through these interconnected phases, the framework supports an ongoing awareness process rather than a one-time estimation of academic stress.

Overall, the LAURA framework is grounded in academic stress theory, interpretive uncertainty in higher education, explainable artificial intelligence, and reflection-based learning design. It integrates context modeling, interpretable estimation, explanation, reflection, and adaptive feedback within a unified structure.

C. Awareness-Oriented Design Logic

The LAURA framework is designed primarily to support student awareness rather than to increase algorithmic complexity. While many AI-supported educational systems emphasize performance prediction or risk estimation and often rely on opaque models, LAURA adopts a different design perspective centered on helping students understand their academic conditions.

This perspective is reflected in three guiding design principles. First, transparency over complexity emphasizes the use of interpretable modeling structures that allow students to clearly see how academic factors contribute to outcomes. Second, explanation as communication focuses on translating analytical outputs into meaningful academic language that students can easily understand. Third, reflection as mediation highlights the role of structured reflection in helping students interpret and internalize system feedback. Fig. 3 illustrates how the three core design principles, transparency, explanation, and reflection, interact to support student awareness within the LAURA framework.

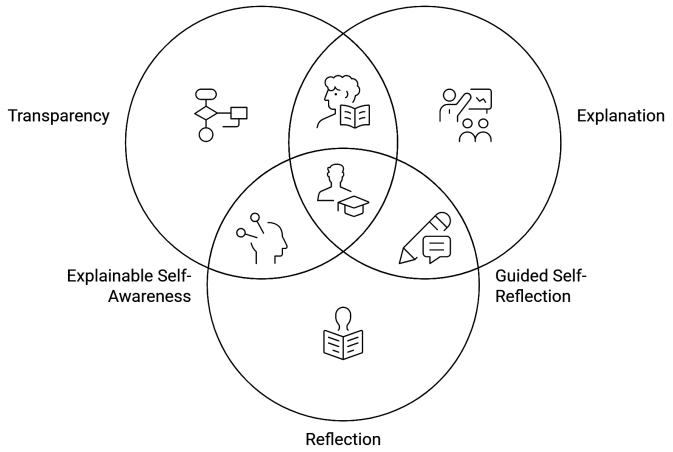


Fig. 3. Design principles supporting student awareness in the LAURA framework.

Together, these principles define LAURA as an awareness-oriented framework that supports interpretation, rather than prediction alone. In this way, explainable AI is positioned as a cognitive support mechanism that helps students better understand and navigate their academic experiences.

IV. Conceptual Interaction and Awareness Development

This section explains how the components of the LAURA framework interact conceptually to support students' awareness of their academic conditions. Rather than focusing on technical implementation, the emphasis is placed on how students engage with the system and how awareness develops through this interaction.

The interaction process begins when academic context indicators, such as perceived workload, deadline pressure, clarity of expectations, and comprehension difficulty, are represented in the system. These indicators provide a structured view of the

student's academic environment and serve as the foundation for subsequent interpretation.

Based on these inputs, the system generates an interpretable estimation of academic stress. Instead of presenting this estimate as a final result, it is framed as an explanation that highlights the main academic factors contributing. To ensure accessibility, explanations are conveyed in simple, non-technical academic language, as recommended in human-centered explainable AI research [46], with a focus on clarity and relevance to students' everyday academic experiences. This design helps reduce confusion that may arise from overly technical or abstract system outputs.

Following this, students engage with the reflection component. Through structured prompts, they are encouraged to interpret the explanation, relate it to their recent academic experiences, and identify potential areas for adjustment. To support engagement, the interaction is designed to feel conversational and context-aware, allowing students to connect system feedback with their personal academic situations. This is important, as awareness does not emerge from information alone, consistent with findings in learning analytics research [47] [48], but from active interpretation and engagement.

The system then provides adaptive feedback based on the interpreted results. These suggestions help students make small, manageable adjustments to their academic strategies, reinforcing the connection between understanding and action.

Importantly, this process is not linear but iterative. As students interact with the system over time, their understanding of academic conditions may evolve. This evolving awareness can influence how they interpret future academic demands, creating a continuous process of interpretation, reflection, and adjustment.

This interaction design is particularly important for students in diverse and international educational contexts, where differences in language, academic expectations, and communication styles may increase uncertainty when interpreting academic signals. By emphasizing clarity, explanation, and reflection, the framework supports inclusive understanding across different student backgrounds.

Through this interaction, LAURA supports the development of awareness as an ongoing cognitive activity. Rather than treating students as passive recipients of information, the framework encourages active engagement, interpretation, and gradual refinement of academic understanding.

V. Discussion

A. Theoretical Implications

The LAURA framework contributes to the theoretical understanding of academic stress by reconceptualizing it as an interpretive and awareness-driven process rather than a static or purely measurable outcome. While prior research has largely approached academic stress through psychological measurement or predictive modeling, this framework shifts the focus toward how students make sense of academic conditions. In this view, stress is determined not only by external demands but also by how those demands are interpreted within specific academic contexts. This perspective builds on prior work that emphasized the role of explainable AI in supporting student awareness of academic stress, particularly among international students [5].

This reconceptualization extends appraisal-based theories of stress into the domain of interactive and AI-supported educational systems. Rather than treating appraisal as an internal cognitive process detached from system design, LAURA positions it as a process that can be actively supported through interaction. By embedding interpretation within system feedback, the framework suggests that academic stress can be better understood when students are guided to examine how different academic factors contribute to their experiences.

From a system design perspective, a central theoretical contribution of this work lies in redefining the role of artificial intelligence in educational settings. Existing AI-based approaches in higher education have predominantly focused on prediction, classification, or early detection of risk. In contrast, the LAURA framework reframes AI as a mechanism for supporting meaning-making. In this context, explainability is not treated as a supplementary feature that reveals model behavior, but as a core component that enables students to interpret academic signals in a meaningful and contextually relevant way. This perspective shifts explainable AI from a technical property of models to a cognitive support mechanism within learning environments.

Furthermore, the framework establishes a conceptual link between explainable AI and metacognitive processes. While explanations can make system outputs more transparent, they do not necessarily lead to understanding on their own. LAURA addresses this limitation by integrating structured reflection into the interaction cycle, emphasizing that awareness emerges through active interpretation rather than passive exposure to information. This integration highlights the importance of reflection as a mediating process that transforms explanations into meaningful understanding.

By connecting interpretability, explanation, and reflection within a unified framework, LAURA provides a theoretical bridge between computational modeling and student-centered learning theories. It suggests that effective AI-supported educational systems should not only analyze academic conditions, but also support students in constructing an understanding of those conditions. In doing so, the framework advances a shift from performance-oriented analytics to awareness-oriented design, in which understanding becomes a primary outcome of the system.

B. Design Implications

The LAURA framework provides several implications for the design of student-facing AI systems, particularly in contexts where the primary goal is to support understanding rather than prediction. First, the framework highlights the importance of prioritizing interpretability over algorithmic complexity when systems are intended for direct interaction with students. In such contexts, models should be designed to allow clear identification of how individual academic factors contribute to system outputs, rather than relying on opaque approaches that limit user understanding.

Second, the findings suggest that explanations should be designed as a form of communication rather than technical

disclosure. Presenting numerical scores or abstract feature weights alone is insufficient for supporting student understanding. Instead, explanations should translate analytical outputs into clear academic meaning by explicitly linking system results to recognizable academic conditions, such as workload, deadline pressure, or clarity of expectations. This implies that explanation interfaces should prioritize clarity, contextual relevance, and alignment with students' everyday academic experiences.

Third, the integration of structured reflection highlights that awareness does not emerge automatically from exposure to system outputs. Systems should actively guide students in interpreting results with simple, focused reflection prompts. These prompts should encourage students to identify key contributing factors, relate them to recent academic experiences, and consider small, manageable adjustments. This suggests that reflection should be treated as a core design component, rather than an optional feature layered on top of analytics.

Finally, the iterative nature of the LAURA framework suggests that feedback should support gradual, continuous adjustment rather than immediate correction. Instead of prescribing fixed actions, systems should provide context-aware suggestions that help students refine their understanding over time. This approach aligns with how students regulate their learning in practice, where changes are often incremental and influenced by evolving academic conditions.

C. Implications for International Students

The LAURA framework is particularly relevant for international students, where challenges often arise not only from academic demands but also from difficulties in interpreting academic expectations and feedback. In such environments, students must navigate unfamiliar educational systems, grading practices, and communication styles, which can introduce additional layers of uncertainty beyond the academic workload itself. As a result, academic stress is often shaped by interpretive ambiguity rather than objective demands alone.

This context highlights a limitation in many existing learning analytics and AI-based systems. When systems present performance indicators or predictive outputs without sufficient explanatory support, they assume that students can interpret academic signals accurately. However, this assumption does not always hold for international students, who may lack familiarity with implicit academic norms or language-specific cues. In this sense, international student contexts make visible a broader issue: the gap between data presentation and meaningful understanding.

By emphasizing explanation and reflection, the LAURA framework directly addresses this interpretive gap. The framework supports students in making sense of academic signals by linking system outputs to clear and contextually meaningful explanations. In addition, structured reflection encourages students to actively interpret these explanations in relation to their own experiences, reducing uncertainty and supporting more informed academic decision-making.

Importantly, this perspective avoids framing international students in deficit-based or clinical terms. Instead, it treats the challenges they experience as issues of interpretation and understanding within complex academic environments. This allows for the design of supportive systems that empower students to navigate unfamiliar contexts more effectively, while also highlighting the importance of inclusive and accessible explanation design.

D. Limitations

Despite its contributions, the proposed LAURA framework has several limitations. First, LAURA is presented as a conceptual framework that establishes an awareness-oriented architectural and theoretical foundation for explainable AI in higher education. As such, its practical implementation and empirical evaluation were not addressed in the present study. Second, although the framework defines the key architectural components and interaction process, its effectiveness in supporting students' awareness, understanding, and academic decision-making has not yet been examined through prototype development or user studies. Finally, the framework is informed primarily by research on higher education, with particular attention to international students. Future research should therefore examine its applicability across diverse higher education contexts, student populations, and institutional settings.

E. Research Implications

The LAURA framework establishes a foundation for research that shifts the focus of AI-supported educational systems from predictive performance toward awareness and understanding. This shift suggests the need for evaluation approaches that go beyond accuracy-based metrics and instead assess how effectively systems support students' interpretation of academic conditions. In this context, constructs such as awareness, comprehension, and perceived interpretability should be treated as primary evaluation outcomes rather than secondary considerations.

Future studies should evaluate the LAURA framework by developing a functional prototype and conducting user studies in authentic higher education settings. These studies should examine whether the framework supports students in interpreting academic signals, developing awareness of the factors contributing to their academic stress, and engaging in meaningful reflection on their academic experiences. The evaluations should also examine the framework's perceived clarity, usefulness, and trustworthiness, and provide empirical evidence to refine it and assess its practical value in awareness-oriented educational AI systems.

This perspective opens several directions for empirical investigation. One important direction is to examine the extent to which explanation alone is sufficient to support understanding, compared with systems that integrate both explanation and structured reflection. Subsequent research can investigate whether reflection mechanisms significantly enhance students' ability to interpret academic signals and make informed adjustments to their learning strategies.

Another research direction concerns the design of explanation formats and their impact on student engagement and understanding. Variations in language, level of detail, and presentation style may influence how students interpret system outputs, particularly in diverse and international educational

contexts. Exploring these variations can provide deeper insight into how explainable systems can be tailored to support different student needs.

In addition, the framework raises questions about the trade-off between model complexity and interpretability in student-facing systems. While complex models may offer higher predictive performance, simpler, more transparent approaches may be more effective at supporting understanding and awareness. Investigating this balance can help clarify when interpretability should be prioritized over predictive accuracy in educational applications.

Overall, the LAURA framework encourages future research to move beyond performance-centered models and toward awareness-oriented system design. By emphasizing understanding as a primary outcome, this perspective supports the development of AI systems that not only analyze academic conditions but also help students meaningfully engage with and interpret their learning experiences.

VI. Conclusion

This study presented the LAURA framework as an awareness-oriented approach to help students understand their academic stress. The framework was developed in response to a key limitation in many existing academic analytics and AI-supported systems, which tend to emphasize prediction and performance monitoring while offering limited support for interpretation and understanding.

By bringing together transparent modeling, explanation, and structured reflection within a single interaction cycle, LAURA repositions the role of explainable AI in educational contexts. Rather than focusing only on estimating outcomes, the framework highlights the importance of helping students make sense of academic signals and connect them to their own experiences. In this way, awareness is treated as something that develops through interpretation, not simply through access to data.

A key contribution of this work is the view of explainable AI as a form of cognitive support within learning environments. Instead of serving only as a tool for analysis, the framework demonstrates how AI can help students identify relevant academic factors, understand their influence, and gradually adjust their approaches to learning. This perspective shifts attention from performance-focused analytics toward systems that support meaningful understanding.

The framework also highlights the challenges students face in interpreting academic expectations, particularly in unfamiliar educational settings. In such contexts, difficulties are not always related to workload alone, but to uncertainty in how academic signals are understood. By focusing on interpretation and reflection, LAURA offers a foundation for designing systems that are more accessible, inclusive, and supportive of diverse student experiences.

Overall, LAURA provides a conceptual foundation for the future development of awareness-oriented educational AI systems. Future studies should focus on developing functional prototypes, conducting user studies in authentic higher education settings, and examining how effectively the framework supports students' awareness, interpretation of academic signals, and reflective engagement. Such investigations will provide empirical evidence to refine the framework and further assess its practical value for awareness-oriented educational AI systems.

References

[1] J. C. Benítez-Agudelo, D. Restrepo, E. Navarro-Jimenez, and V. J. Clemente-Suárez, "Longitudinal effects of stress in an academic context on psychological well-being, physiological markers, health behaviors, and academic performance in university students," *BMC psychology*, vol. 13, no. 1, p. 753, 2025.

[2] H. Crompton and D. Burke, "Artificial intelligence in higher education: the state of the field," *International journal of educational technology in higher education*, vol. 20, no. 1, pp. 1–22, 2023.

[3] F. Ouyang, M. Wu, L. Zheng, L. Zhang, and P. Jiao, "Integration of artificial intelligence performance prediction and learning analytics to improve student learning in online engineering course," *International Journal of Educational Technology in Higher Education*, vol. 20, no. 1, p. 4, 2023.

[4] S. Esakkiammal and K. Kasturi, "Advancing educational outcomes with artificial intelligence: Challenges, opportunities, and future directions," *International Journal of Computational and Experimental Science and Engineering*, vol. 10, no. 4, pp. 1749–1756, 2024.

[5] A. Almathami and R. Stone, "Explainable ai for enhancing awareness of academic stress among international university students." *International Journal of Advanced Computer Science & Applications*, vol. 17, no. 1, 2026.

[6] H. Khosravi, S. B. Shum, G. Chen, C. Conati, Y.-S. Tsai, J. Kay, S. Knight, R. Martinez-Maldonado, S. Sadiq, and D. Gašević, "Explainable artificial intelligence in education," *Computers and education: artificial intelligence*, vol. 3, p. 100074, 2022.

[7] K. Fiok, F. V. Farahani, W. Karwowski, and T. Ahram, "Explainable artificial intelligence for education and training," *The Journal of Defense Modeling and Simulation*, vol. 19, no. 2, pp. 133–144, 2022.

[8] R. Farrow, "The possibilities and limits of explicable artificial intelligence (xai) in education: a socio-technical perspective," *Learning, Media and Technology*, vol. 48, no. 2, pp. 266–279, 2023.

[9] M. C. Pascoe, S. E. Hetrick, and A. G. Parker, "The impact of stress on students in secondary school and higher education," *International journal of adolescence and youth*, vol. 25, no. 1, pp. 104–112, 2020.

[10] M. S. Islam and M. F. Rabbi, "Exploring the sources of academic stress and adopted coping mechanisms among university students." *International Journal on Studies in Education (IJonSE)*, vol. 6, no. 2, 2024.

[11] J. Liu, M. Wong, and B. Hard, "The impact of norm-and criterion-referenced grading systems on students' course-related expectations," *Journal of the Scholarship of Teaching and Learning*, vol. 24, no. 4, 2024.

[12] Z. Zheng, H. Buarabha, A. Sriruksa, and S. Syers, "Understanding stress and academic adaptation: A comparative study of chinese international students in thailand," *Journal of Ecohumanism*, vol. 4, no. 1, pp. 1049–1063, 2025.

[13] D. Pérez-Jorge, M. Boutaba-Alehyan, A. I. González-Contreras, and I. Pérez-Pérez, "Examining the effects of academic stress on student well-being in higher education," *Humanities and Social Sciences Communications*, vol. 12, no. 1, pp. 1–13, 2025.

[14] N. Permatasari and D. R. Suminar, "Analysis of academic stress levels among migrant students," in *Interdisciplinary Conference of Psychology, Health, and Social Science (ICPHS 2021)*. Atlantis Press, 2022, pp. 1–6.

[15] Y. A. Ali, R. S. Hussein, N. S. Mostafa, and A. F. Manzour, "Factors associated with acculturative stress among international medical students in an egyptian university," *BMC medical education*, vol. 24, no. 1, p. 1009, 2024.

[16] A. Abd-Alrazaq, M. Alajlani, R. Ahmad, R. AlSaad, S. Aziz, A. Ahmed, M. Alsahli, R. Damseh, and J. Sheikh, "The performance of wearable ai in detecting stress among students: systematic review and meta-analysis," *Journal of Medical Internet Research*, vol. 26, p. e52622, 2024.

[17] R. de Filippis and A. A. Foysal, "Comprehensive analysis of stress fac-

tors affecting students: a machine learning approach," *Discover Artificial Intelligence*, vol. 4, no. 1, p. 62, 2024.

[18] I. J. Ratul, M. M. Nishat, F. Faisal, S. Sultana, A. Ahmed, and M. A. Al Mamun, "Analyzing perceived psychological and social stress of university students: A machine learning approach," *Heliyon*, vol. 9, no. 6, 2023.

[19] Q. Chen and B. G. Lee, "Deep learning models for stress analysis in university students: a sudoku-based study," *Sensors*, vol. 23, no. 13, p. 6099, 2023.

[20] L. Paulsen and E. Lindsay, "Learning analytics dashboards are increasingly becoming about learning and not just analytics-a systematic review," *Education and Information Technologies*, vol. 29, no. 11, pp. 14 279–14 308, 2024.

[21] T. Khelifi, N. B. Rabah, B. Le Grand, and I. Daoudi, "Ex-lad: an explainable learning analytics dashboard in higher education," in *Proceedings of 36th International Conference on*, vol. 97, 2024, pp. 38–51.

[22] P. Kannan and D. Zapata-Rivera, "Facilitating the use of data from multiple sources for formative learning in the context of digital assessments: informing the design and development of learning analytic dashboards," in *Frontiers in Education*, vol. 7. Frontiers Media SA, 2022, p. 913594.

[23] T. Susnjak, G. S. Ramaswami, and A. Mathrani, "Learning analytics dashboard: a tool for providing actionable insights to learners," *International Journal of Educational Technology in Higher Education*, vol. 19, no. 1, p. 12, 2022.

[24] T. Giorgashvili, I. Jivet, C. Artelt, D. Biedermann, D. Bengs, F. Goldhammer, C. Hahnel, J. Mendzheritskaya, J. Mordel, M. Onofrei *et al.*, "From reflection to action: A controlled field study on how learners interpret and respond to feedback in learning analytics dashboards," *Journal of Computer Assisted Learning*, vol. 41, no. 4, p. e70073, 2025.

[25] G. Ramaswami, T. Susnjak, and A. Mathrani, "Effectiveness of a learning analytics dashboard for increasing student engagement levels," *Journal of Learning Analytics*, vol. 10, no. 3, pp. 115–134, 2023.

[26] G. Sedrakyan, J. Malmberg, K. Verbert, S. Järvelä, and P. A. Kirschner, "Linking learning behavior analytics and learning science concepts: Designing a learning analytics dashboard for feedback to support learning regulation," *Computers in Human Behavior*, vol. 107, p. 105512, 2020.

[27] H. Arif, "Designing a student-centric framework for explainable ai in adaptive learning systems: Towards transparent, trustworthy, and ethical ai in education," *Qualitative Research Journal for Social Studies*, 2025.

[28] S. Gunasekara and M. Saarela, "Explainable ai in education: Techniques and qualitative assessment," *Applied Sciences*, vol. 15, no. 3, p. 1239, 2025.

[29] B. Bali, "Cognitively-informed explainable ai for education: Enhancing student engagement and transparent assessment through human-centered modeling," *Journal of Social and Scientific Education*, vol. 2, no. 3, pp. 173–190, 2025.

[30] S. Maity and A. Deroy, "Human-centric explainable ai in education," *arXiv preprint arXiv:2410.19822*, 2024.

[31] K. Alalawi, R. Athauda, and R. Chiong, "An extended learning analytics framework integrating machine learning and pedagogical approaches for student performance prediction and intervention," *International Journal of Artificial Intelligence in Education*, vol. 35, no. 3, pp. 1239–1287, 2025.

[32] T. Xu, P. Zhu, Q. Ji, W. Wang, M. Qian, and G. Shi, "Psychological distress and academic self-efficacy of nursing undergraduates under the normalization of covid-19: multiple mediating roles of social support and mindfulness," *BMC Medical Education*, vol. 23, no. 1, p. 348, 2023.

[33] X. Zuo, L. Zhao, Y. Li, W. He, C. Yu, and Z. Wang, "Psychological mechanisms of english academic stress and academic burnout: the mediating role of rumination and moderating effect of neuroticism," *Frontiers in psychology*, vol. 15, p. 1309210, 2024.

[34] Y. Deng, J. Cherian, N. U. N. Khan, K. Kumari, M. S. Sial, U. Comite, B. Gavurova, and J. Popp, "Family and academic stress and their impact on students' depression level and academic performance," *Frontiers in psychiatry*, vol. 13, p. 869337, 2022.

[35] V. Rovagnati, E. Pitt, and N. Winstone, "Feedback cultures, histories and literacies: international postgraduate students' experiences," *Assessment & Evaluation in Higher Education*, vol. 47, no. 3, pp. 347–359, 2022.

[36] D. E. Mathew, D. U. Ebem, A. C. Ikegwu, P. E. Ukeoma, and N. F. Dibiaezue, "Recent emerging techniques in explainable artificial intelligence to enhance the interpretable and understanding of ai models for human," *Neural Processing Letters*, vol. 57, no. 1, p. 16, 2025.

[37] X. Ren, V. A. Sotardi, and C. Brown, "Exploring academic stress and coping experiences among university students during the covid-19 pandemic," *Education Sciences*, vol. 15, no. 3, p. 314, 2025.

[38] F. S. Fayda-Kinik, "Academic stress in online higher education." *International Journal of Technology in Education and Science*, vol. 7, no. 1, pp. 18–29, 2023.

[39] C. Molnar, *Interpretable machine learning*. Lulu. com, 2020.

[40] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion*, vol. 58, pp. 82–115, 2020.

[41] V. Echeverria, G. F. Nieto, L. Zhao, E. Palominos, N. Srivastava, D. Gašević, V. Pammer-Schindler, and R. Martinez-Maldonado, "A learning analytics dashboard to support students' reflection on collaboration," *Journal of Computer Assisted Learning*, vol. 41, no. 1, p. e13088, 2025.

[42] L. Chen, X. Geng, M. Lu, A. Shimada, and M. Yamada, "How students use learning analytics dashboards in higher education: A learning performance perspective," *Sage Open*, vol. 13, no. 3, p. 21582440231192151, 2023.

[43] J. Colomer, T. Serra, D. Cañabate, and R. Bubnys, "Reflective learning in higher education: Active methodologies for transformative practices," p. 3827, 2020.

[44] M. Mejeh, L. Sarbach, and T. Hascher, "Effects of adaptive feedback through a digital tool–a mixed-methods study on the course of self-regulated learning," *Education and Information Technologies*, vol. 29, no. 14, pp. 1–43, 2024.

[45] M. Afzaal, A. Zia, J. Nouri, and U. Fors, "Informative feedback and explainable ai-based recommendations to support students' self-regulation," *Technology, Knowledge and Learning*, vol. 29, no. 1, pp. 331–354, 2024.

[46] Q. V. Liao, D. Gruen, and S. Miller, "Questioning the ai: informing design practices for explainable ai user experiences," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–15.

[47] K. Verbert, X. Ochoa, R. De Croon, R. A. Dourado, and T. De Laet, "Learning analytics dashboards: The past, the present and the future," in *Proceedings of the tenth international conference on learning analytics & knowledge*, 2020, pp. 35–40.

[48] E. Du Plooy, D. Casteleijn, and D. Franzsen, "Personalized adaptive learning in higher education: A scoping review of key characteristics and impact on academic performance and engagement," *Heliyon*, vol. 10, no. 21, 2024.

A Model-Free AI Framework for Robust PQRST Complex Detection in ECG Signals

Arthorn Luangsodsai, Manh Duong Dang, Krung Sinapiromsaran*
Department of Mathematics and Computer Science,
Faculty of Science Chulalongkorn University, Bangkok 10330, Thailand

***Abstract*—Accurate delineation of PQRST complex is essential for clinical ECG analysis, but remains challenging due to waveform variability, low amplitude, and polarity inversions. This work presents PAD-PQRST, a model-free Polarity-Adaptive Delineation method for PQRST complex. This approach leverages search windows based on established cardiac physiology. The Q and S waves are identified by locating the corresponding extrema relative to the R-peak polarity. Meanwhile, the P and T waves are determined by finding the extrema with the maximal absolute deviation from local mean. Because this design inherently accommodates waves of any polarity, it eliminates the need for a predictive model. This research uses the Manikandan2012 algorithm for R-peak detection due to its robustness. Validation on QTDB and LUDB yields F1-scores from 98.5% to 99.8% for QRS components, 90.97% to 94.57% for P-waves, and 79% to 91.41% for T-waves, surpassing established wavelet-based and graph-based methods. The combination of physiologically informed search windows and polarity-adaptive detection delivers a simple yet powerful delineation framework suitable for diverse clinical datasets.**

***Keywords*—ECG signal processing; PQRST detection; model-free AI; Polarity-Adaptive Delineation method**

I. Introduction

The electrocardiogram (ECG) is a non-invasive diagnostic tool that records the heart's electrical activity over time using cutaneous electrodes [1]. These electrodes detect voltage fluctuations arising from myocardial depolarization, and repolarization during each cardiac cycle [2]. The detected signals are amplified, and recorded as characteristic waveforms representing the sequence, timing of atrial, and ventricular electrical events [2]. The ECG is indispensable for rapid detection, and diagnosis of cardiovascular pathologies, including arrythmias, myocardial ischemia, electrolyte imbalances, and structural abnormalities [3]–[5]. By providing a visual representation of the heartbeat and its constituent PQRST complex, the ECG enables clinicians to assess cardiac function, guide treatment decisions, and monitor therapeutic efficacy across clinical settings [6], [7]. Its accessibility, cost-effectiveness, and diagnostic power render it essential for early identification of life-threatening conditions, or ongoing cardiac monitoring.

The cardiac cycle encompasses sequential mechanical, and electrical events facilitating effective blood circulation [8]. Coordinated chamber contraction, and relaxation are controlled by electrical impulses from the sinoatrial node [9]. The PQRST complex corresponds directly to the electrical events underlying these mechanical actions [10]. The P-wave signifies atrial depolarization [10], [11]. The QRS complex corresponds to ventricular depolarization [10]. The T-wave represents ventricular repolarization. Each component provides a distinct electrical signature essential for distinguishing normal from pathological cardiac function. The PQRST complex adheres to well-defined physiological timing constraints. In healthy adults, the P-wave persists for 0.08 to 0.10 seconds, while QRS complex endures for 0.06 to 0.10 seconds [12]. The PR interval ranges from 0.12 to 0.20 seconds, and the QT interval from 0.35 to 0.43 seconds, together they make up the temporal span of the PQRST complex [12]–[15]. Deviations from these norms constitute significant markers of cardiac pathology. The morphology, amplitude, and duration of individual waves supply essential clinical information regarding cardiac function, and underlying pathologies, making comprehensive PQRST analysis fundamental to cardiovascular diagnosis [10], [16]–[18].

II. Background

Several approaches have been developed for automated delineation of PQRST complexes in ECG signals. One cycle of the PQRST complex can be visualized in Fig. 1. Marinez proposed a wavelet-based methodology that exploits the multiscale properties of the dyadic wavelet transform to robustly identify characteristic waveform points [19]. Kalyakulina also developed a wavelet-transform-based delineation method [20]. Bote contributed with low-complexity derivative-based approach to detect fiducial points [21]. More recently, Emrich introduced a prominence-based approach that identifies waveform features using peak prominence, a measure of a point's relative height above its local baseline [22]. Additionally, Neurokit2 also provided a toolbox for biomedical signal processing, which also include PQRST detection [23].

Despite its clinical importance, ECG analysis presents significant challenges. Inherent signal variability arises from inter-individual physiological differences, diverse pathologies, and varying measurement environments, complicating consistent feature extraction [24]. ECG signals are frequently contaminated by noise sources, including baseline wander, powerline interference, myoelectric activity, and motion artifacts, which degrade signal quality, and impede precise detection of low-amplitude P, and T waves [24]. These complications lead to detection techniques for P, and T waves not being researched as intensively as QRS detection [11]. The non-stationary nature of ECG signals further necessitates sophisticated adaptive preprocessing. Moreover, machine learning, and deep learning approaches face limitations, including dependence on large label datasets, lack of interpretability, and computationally expensive, constraining real-time clinical applicability [24].

*Corresponding author

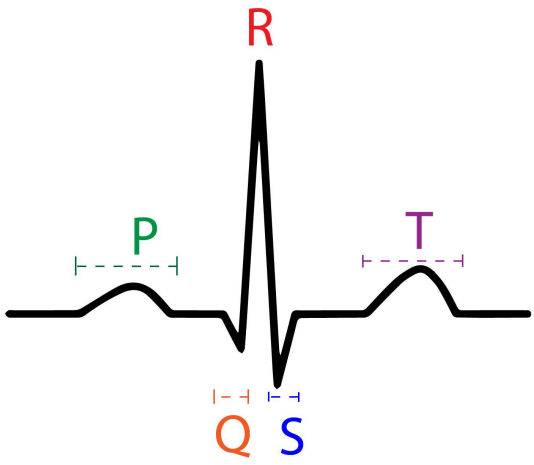


Fig. 1. PQRST complex

These challenges motivate this research into robust, efficient, and transparent ECG analysis methods. Reliable automatic detection of critical fiducial points, particularly R-peaks, and other PQRST components, can reduce clinician burden, minimize human error, and enable continuous monitoring [7]. Algorithms that do not rely heavily on prior training, or extensive parameter tuning offer broader applicability across diverse populations, and signal conditions. Such advancements would enhance clinical decision-making, enable personalized cardiac care, and ultimately improve patient outcomes.

This work introduces a model-free PQRST delineation methodology that combines a polarity-adaptive approach with physiologically grounded search windows to provide competitive performance on QTDB and LUDB without any prior training.

III. Methodology

A. Overview

This study proposes the algorithm PAD-PQRST, Polarity-Adaptive Delineation of PQRST complex, as presented in Fig. 2. The process starts with the processing of ECG signals to remove noise, baseline wander and Z-score normalization. After the signal is preprocessed, the algorithm starts with R-peak detection, utilizing Manikandan2012's algorithm. Once the R-peaks are detected, the algorithm executes a hierarchical detection cascade on the Z-score normalized signal, which sequentially identifies Q, S, P, and T waves using physiologically constrained search windows anchored to each detected R-peak. Q and S-waves are located based on R-peak polarity, while P and T-waves employ adaptive absolute amplitude thresholding within their respective search windows. Finally, the output results of the proposed algorithm are the indexes of PQRST complex waveform components of the original ECG signal.

B. Preprocessing

The preprocessing stage employs a 4th order Chebyshev type I bandpass filter with a passband of 6-18 Hz, implemented using both forward and backward filtering to eliminate phase distortion. This specific filter configuration was selected based on the spectral characteristics of the QRS complex, which concentrates its energy predominantly within this frequency range. The lower cutoff frequency of 6 Hz effectively reduces baseline wander, respiratory artifacts, and low-frequency components of P, and T waves, while the upper cutoff frequency of 18Hz suppresses high-frequency noise from powerline interference, muscle contractions, and motion artifacts. This effect can be observed from the difference between the original signal in Fig. 3 and the result of the preprocessing filter in Fig. 4, where the R-peaks amplitude is retained, the baseline becomes relatively stable. Despite this filter reducing the prominent P and T waves, it makes the process to detect Q and S waves easier, which are subsequently used as anchor to detect the lower frequency waves.

Specifically, for the PQRST delineation process, the algorithm also employs Z-score normalization. This step involves subtracting the mean ECG amplitude value from each data point, ensuring consistency, and reinforcing relationship between the signal value, and its position relative to the baseline. This is demonstrated by the signal in Fig. 5, where the data points above the baseline have positive values, and points below the baseline have negative values.

C. R-Peak Detection

After the ECG signal has been preprocessed, The PQRST delineation process begins with R-peak detection. In principle, any reliable R-peak detector can serve as the temporal anchor for this step [26]–[36] . However, for optimal performance of subsequent PQRST delineation, a detector with high Positive Predictive Value (PPV) is preferred, as it limits the number of false positive R-peaks that would otherwise be used as anchors to detect non-existent or false PQRST complexes. While the outlier detection-based MOW-ECG algorithm demonstrated the highest PPV among compared methods in earlier research, its overlapping window mechanism may affect the performance of PQRST detection due to additional complexity [36]. Therefore, to isolate and focus more directly on the PQRST delineation logic itself, the method presented by Manikandan, and Soman was utilized for this study [25]. This method employs a multi-stage signal processing approach that combines digital filtering, nonlinear transformation, and Hilbert transform-based peak identification [25]. This chosen method provides excellent sensitivity, and PPV, allowing a cleaner evaluation of our polarity-adaptive PQRST framework.

D. PQRST Delineation

1) Initialization of physiological parameters: Following the detection of R-peaks, the detected peaks serve as reference landmarks for delineating the complete PQRST complex. The proposed methodology incorporates established physiological constraints, and temporal relationships between waveform components to accurately identify each feature [14], [15], [37]–[39]. Specifically, search windows are defined based on clinically observed intervals:

- Q wave: A 120 ms window preceding the R-peak
- S Wave: A 180 ms window following the Q-wave

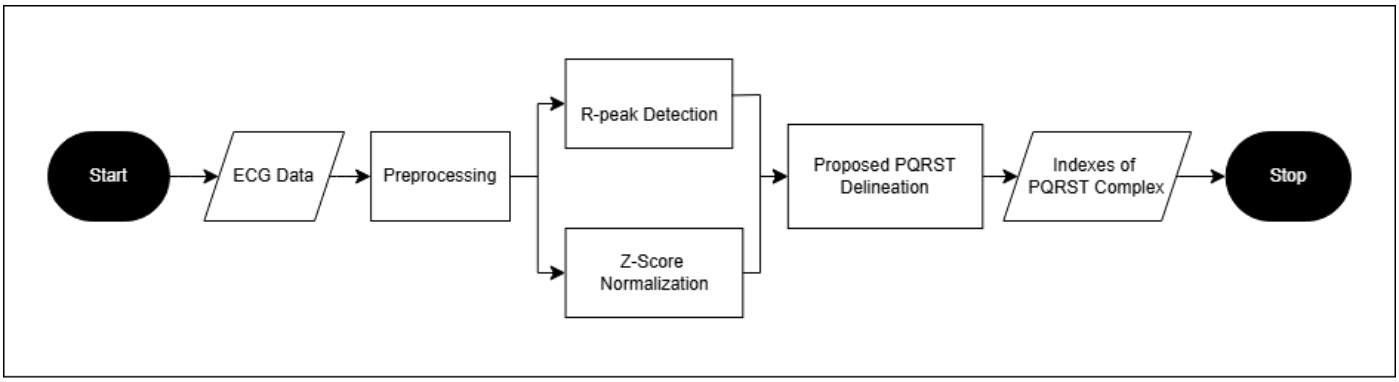


Fig. 2. Proposed PQRST delineation algorithm overview.

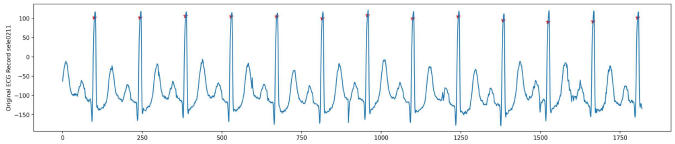


Fig. 3. QTDB's original record sele0211.

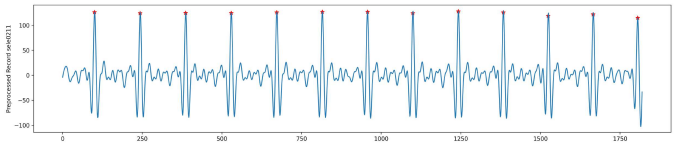


Fig. 4. Preprocessed record sele0211.

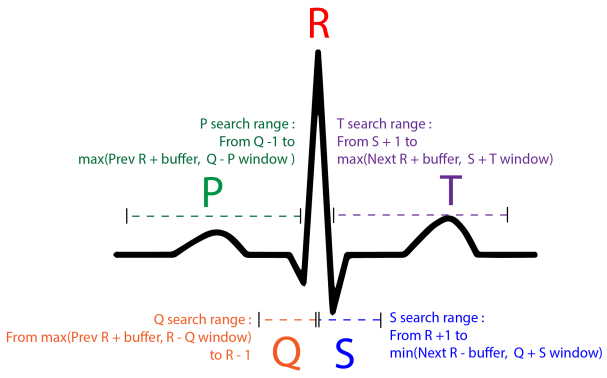


Fig. 6. Search range of PQRST complex's features.

- P Wave: A 300 ms window preceding the Q-wave
- T Wave: A 400ms window following the S-wave, with a minimum separation of 150 ms

The selection of these search windows is grounded in clinically observed intervals, and is depicted in Fig. 6. Based on the typical duration between QRS onset, and the R-peak, which ranges from 50 to 100 ms, a 120 ms search window for the Q-wave is established prior to the R-peak [14], [15]. The normal QRS duration is approximately 100 ms, and is considered abnormal when exceeding 120 ms [14], [38]. Accordingly, a 180 ms search window for the S-wave following the Q-wave is employed. The P-wave is subsequently searched within a 300 ms window preceding the Q-wave to ensure the PR interval remains within its normal physiological range of 120-200 ms [38]. The T-wave is searched within a 400 ms window after the S-wave, with a minimum separation of 150 ms from the S-wave to respect cardiac repolarization timing [38], [39]. To establish physiologically constrained search windows for individual PQRST complexes, the detection algorithm utilizes the temporal relationship between consecutive R-peaks. For each detected R-peak, the preceding R-peak boundary is defined as the previous detected R-peak when available. At the beginning of the recording where no prior peak exists, this boundary is set to the current R-peak position minus a minimum RR interval value of 200ms, ensuring the search window does not extend indefinitely backwards [26]. Similarly, the following R-peak boundary is defined as the next detected R-peak when available, or the current R-peak position plus a 200ms interval when processing the final cardiac cycle. This 200ms window, corresponding to the physiological refractory period of the heart, serves as a conservative fallback that prevents the search algorithm from extending beyond expected cardiac cycle boundaries while accommodating heart rate variability. This approach ensures that each PQRST complex is analyzed within a locally adaptive window that respects the individual cardiac rhythm. Additionally, a 100 ms buffer interval is further incorporated to maintain clear isolation between adjacent waveform features during search window determination, therefore minimizing interference, and improving detection accuracy.

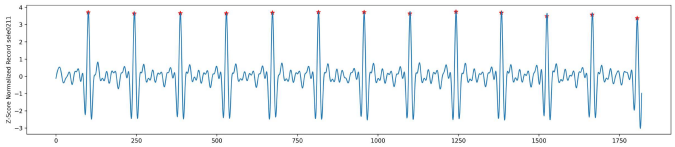


Fig. 5. Z-Scored record sele0211.

2) Sequential detection of waveform: Following R-peak detection, the proposed algorithm executes a hierarchical de-

tection cascade for each identified R-peak to delineate the complete PQRST complex, as illustrated in Fig. 7. This cascade proceeds sequentially, first identifying the Q, and S waves, then subsequently locating the P, and T waves based on the detected Q, and S positions using physiologically constrained search windows. The process begins by determining the polarity of the detected R-peak through examination of its Z-score normalized value, classifying it as either positive, or negative to guide subsequent wave searches.

For Q-wave detection, the search range is defined to begin at the maximum value between two boundaries, which are the previous R-peak location plus a buffer interval, and the current R-peak location minus the predefined Q-wave search window. The search range extends up to one sample point before the current R-peak location. Within this interval, the algorithm identifies the Q-wave as either a local minimum, or maximum depending on R-peak polarity based on its Z-score. Specifically, if the current R-peak is positive, the Q-wave corresponds to the local minimum, and, conversely, for a negative R-peak, the Q-wave is found at the local maximum position.

The S-wave search range commences immediately after the current R-peak, and continues until the minimum value between the two constraints of the next R-peak location minus the buffer interval, and the current Q-wave location plus the S-wave search window. Applying the same polarity-dependent logic, the S-wave is identified as the local minimum when the R-peak is positive, and as the local maximum when the R-peak is negative.

The P-wave detection employs a more sophisticated approach within a search range that begins at the maximum value between two potential starting points of the Q-wave location minus the P-wave search window, and the buffer interval plus the previous T-wave location if available, otherwise the previous R-peaks serve as the reference. The search range concludes at one sample point before the Q-wave. Within this segment, the algorithm first computes three key parameters. First, the dynamic range, which is calculated as the difference between maximum, minimum values within the segment. Second, the minimum amplitude threshold, which is defined as 10% of the dynamic range. Finally, the mean value of the search segment. Subsequently, the algorithm evaluates all point-to-point differences within the search segment to identify all local maximum and minima positions. A point is classified as local maximum if the current point is larger than both the previous point, and the next point. The local minimum satisfies the opposite conditions. From the identified extrema, the algorithm extracts the ECG amplitudes, selecting the highest peak among all local maxima, and recording its position, and amplitude, while similarly identifying the deepest valley among all local minima. The algorithm then computes the absolute deviations of both the maximum peak, and minimum valley from the segment mean. If the absolute maximum peak deviation exceeds both the absolute minimum valley deviation, and the minimum amplitude threshold, the P-wave is assigned to the maximum peak location, otherwise the minimum valley location is selected.

The T-wave search range is defined to begin 150 ms after the detected S-wave, ensuring adequate separation to respect cardiac repolarization timing, and extends until the minimum value between the next R-peak location minus the buffer interval, and the detected S-wave location plus the T-wave search window. The T-wave localization then proceeds using the same adaptive absolute extrema selection methodology employed for P-wave detection, as it inherently accommodates biphasic or inverted T-wave, a clinically significant morphological variation, as observed in Fig. 8. Instead of assuming positive deflection, the proposed algorithm simultaneously evaluates both local maxima, minima within the search window and select the extremum with the greatest absolute value. This approach combines with the minimum 150ms separation from the S-wave further prevent confusion with the QRS complex, ensure that the proposed algorithm can maintain consistent and physiologically plausible identification of T-waves.

E. Complexity Analysis

The computational complexity of the proposed algorithm PAD-PQRST is characterized by $O(n*m)$, where n represents the number of detected R-peaks in the ECG signal, and m represents the maximum search window size in samples used for the delineation of P, Q, S, T waves. The parameter m is determined as the products of the sampling rate and the fixed temporal windows defined for each waveform component. This linear scaling arises because the algorithm processes each R-peak independently. For each cardiac cycle, PAD-PQRST executes a fixed number of localized operations confined to constant-sized windows that are proportional to physiologically meaning time intervals. All these operations involve only a bounded number of samples per beat, independent of total signal length. The overall computational complexity scales linearly with the number of heartbeat n. The design of the algorithm allows it to avoid exhaustive search over the entire ECG record, and focuses processing on the regions of interest surrounding each cardiac complex. As a result, the algorithm is well-suited for real-time physiological signal processing applications, including continuous monitoring on resource-constrained devices.

IV. Results and Discussion

A. Datasets

The QT Database (QTDB) comprises 105 fifteen-minute ECG recordings, each annotated by an expert cardiologist for the onset, peak, and offset of P waves, QRS complexes, and T waves of 30 to 50 beats in each recording [40]. It is recorded at a sampling frequency of 250 Hz, its primary lies in its detailed manual annotations of key waveform features, making it a crucial resource for validating delineation algorithms. The Lobachevsky University ECG Database (LUDB) was created to provide a comprehensive resource for training, and validating ECG delineation algorithms [41]. It addresses a key limitation of QTDB by offering exhaustive manual annotations. It consists of 200 ten-second, 12-lead ECG recordings digitized at 500 Hz. For every single beat across all 200 records, certified cardiologist manually marked the boundaries, and peaks of P, T, and QRS complexes in all 12 leads. The database includes recordings from both healthy volunteers, and patients with various cardiovascular conditions, pacemakers, and other abnormalities. Its complete annotation set makes it exceptionally valuable for developing, and testing ECG delineation algorithms. The Brno University of Technology

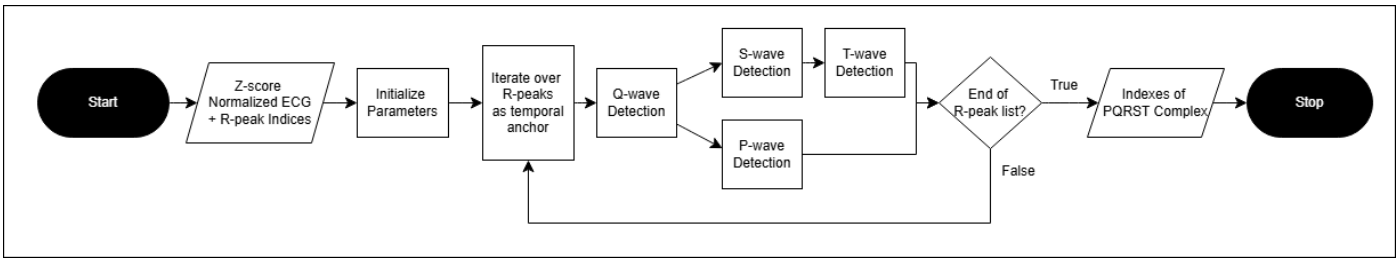


Fig. 7. Search order of PQRST complex's feature.

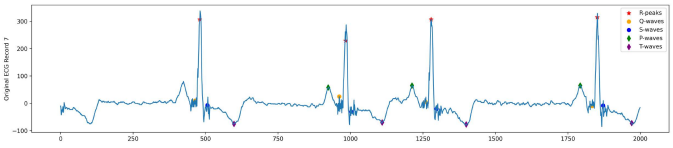


Fig. 8. LUDB's record 7 with inverted T-waves.

ECG Signal Database with Annotations of P Wave (BUT-PDB) consists of 50 two-minute, two-lead ECG records with various pathology types [42]. This dataset only contains P-wave positions, which were manually annotated in each beat by two experts. The recordings in this dataset were chosen to include high incidence of cardiac pathologies, resulting in a dataset that covers 23 different pathological conditions to challenge P-wave detection under clinically relevant conditions. The MIT-BIH Arrhythmia Database P-Wave Annotations (MIT-BIH-PDB) provides reference annotations for P-waves for a curated subset of 12 records from the widely used MIT-BIH Arrythmia Database [43]. The records were chosen due to the presence of abnormal cardiac morphologies that complicate P-wave detection. The annotations were manually made by two experts.

B. Performance Evaluation

The performance of the proposed PQRST delineation algorithm is evaluated using a tolerance window of 150 ms to be incompliance with ANSI/AAMI-EC57:1998 standard [44]. A detection is classified as true positive (TP) if its location fell within the tolerance window around the reference annotation. Detected points with no matching annotation are counted as false positives (FP), representing errors where the algorithm incorrectly identified a waveform component in noise, or non-QRS features. Conversely, annotated points that the algorithm failed to detect are recorded as false negatives (FN). Based on these fundamental counts, the F1 score was computed for each wave type to provide a balanced assessment of detection accuracy, as defined by Eq. (1):

$$F1[\%] = \frac{2TP}{2TP + FP + FN} \tag{1}$$

Additionally, to provide further context to the primary metrics of F1 score, Sensitivity, Positive Predictive Value (PPV), and Detection Error Rate (DER) were calculated using the respective formulas from Eq. (2) to Eq. (4):

$$Sensitivity[\%] = \frac{TP}{TP + FN} \tag{2}$$

$$PPV[\%] = \frac{TP}{TP + FP} \tag{3}$$

$$DER[\%] = \frac{FP + FN}{TP + FN} \tag{4}$$

These metrics range from 0% to 100%, with higher values indicating better overall detection performance.

C. Methods Used for Comparison

To evaluate the performance of the proposed PAD-PQRST algorithm in PQRST complex delineation, three methods are selected to be used as benchmarks. These methods are a dyadic wavelet transformed-based delineation technique by Martinez2004 [19], a visibility graph, and prominence-based technique by Emrich2024 [22], and a continuous wavelet transform-based technique provided by the Neurokit2 toolkit [23].

D. Comparison Against Established Methods

The results of the comparative experiment on QTDB, summarized in Table I, demonstrate exceptional performance of the proposed method. For the detection of Q, R, S waves, the proposed method achieves near perfect F1 scores of 99.68% for Q-wave, 99.86% for R-wave, and 99.70% for S-wave. These values are more than 20% higher than Martinez2004, and Emrich2024, with the best competitor Neurokit2 CWT, which reached 96.81% for Q-wave, 98.98% for R-wave, and 92% for S-wave. The near-perfect F1 score of PAD-PQRST reflects both extremely high Sensitivity, PPV, as well as the below 1% DER, indicating that the method almost never misses a true beat nor generates false detections. This performance validates the combination of a robust R-peak anchor with the proposed polarity-adaptive approach. On P-wave detection results, the proposed method achieves an F1 score of 92.33%, which outperforms Martinez2004 by a margin of 19%, and Emrich2004 by a margin of 22%. The improvement is driven by higher FP rate compared to QRS detection is expected because low-amplitude P-waves can be confused with noise or baseline fluctuations. The T-wave remains the most challenging component, with the proposed algorithm reaching an F1 score of 78.69%. Although PAD-PQRST still outperforms Martinez2004 by 19%, and Emrich2024 by 29%, the F1 score

TABLE I. SCORE OF PQRST DELINEATION OF COMPARED METHODS IN QTDB.

Method	Metrics	P	Q	R	S	T
Martinez2004	TP	6956	7832	8039	7305	6656
	FP	1100	225	41	752	1401
	FN	3994	4521	4323	5049	7768
	Sensitivity	63.53%	63.40%	65.03%	59.13%	46.15%
	PPV	86.35%	97.21%	99.49%	90.67%	82.61%
	DER	46.52%	38.42%	35.30%	46.96%	63.57%
	F1	**73.20%**	**76.75%**	**78.65%**	**71.58%**	**59.21%**
Emrich2024	TP	9722	10721	10769	10916	7726
	FP	6928	5915	5882	5734	8924
	FN	1228	1632	1593	1438	6698
	Sensitivity	88.79%	86.79%	87.11%	88.36%	53.56%
	PPV	58.39%	64.44%	64.67%	65.56%	46.40%
	DER	74.48%	61.09%	60.47%	58.05%	108.31%
	F1	**70.45%**	**73.97%**	**74.24%**	**75.27%**	**49.73%**
Neurokit2 CWT	TP	3407	11903	12175	11312	8747
	FP	8726	335	63	926	3386
	FN	7543	450	187	1042	5677
	Sensitivity	31.11%	96.36%	98.49%	91.57%	60.64%
	PPV	28.08%	97.26%	99.49%	92.43%	72.09%
	DER	148.58%	6.35%	2.02%	15.93%	62.83%
	F1	**29.52%**	**96.81%**	**98.98%**	**92.00%**	**65.87%**
Proposed Method	TP	10754	12326	12353	12329	10534
	FP	1591	53	26	50	1814
	FN	196	27	9	25	3890
	Sensitivity	98.21%	99.78%	99.93%	99.80%	73.03%
	PPV	87.11%	99.57%	99.79%	99.60%	85.31%
	DER	16.32%	0.65%	0.28%	0.61%	39.55%
	F1	**92.33%**	**99.68%**	**99.86%**	**99.70%**	**78.69%**

is lower than other waves. Further investigation shows that the sensitivity, at 73.03%, can be the limiting factor as many T-waves, especially those with low amplitude, or flattened morphology, are missed. The DER indicates about 39.55% of detections are either missed or false, which confirms that T-wave delineation remains a challenge.

The LUDB experiment results, presented in Table II, provide critical validation of the method's generalizability across diverse patient populations, and pathological conditions with exhaustive annotations. The F1 score of PAD-PQRST remains very high for all components, despite the increased morphological variability. The proposed method achieves F1 score of 98.94% for Q-wave, 99.08% for R-peak, and 98.80% for S-wave. These are consistently 3 to 5 percentage points higher than the next best method, either Neurokit2 CWT, or Emrich2024. The slightly lower F1 score of PAD-PQRST compared to QTDB reflects the wider range of morphologies present in LUDB. Additionally, the drastically shorter signal can also explain the sharp rises in performance for the other methods used for comparison. For P-wave, the proposed method attains an F1 score of 95.03%, which is 4% higher than Martinez2004, Emrich2024, and 22% higher than Neurokit2 CWT. Notably, the sensitivity of the PAD-PQRST is slightly lower than on QTDB, possibly due to LUDB's exhaustive annotation includes very low-amplitude P-waves that can be confused with noise. For T-wave, the proposed method reaches an F1 score of 91.20% on LUDB, which is 5% higher than Martinez2004, Emrich2024, and 12% higher than Neurokit2 CWT. This is substantially better than the performance on the QTDB, likely due to LUDB's T-waves are annotated more consistently, including a wide range of morphologies that the proposed polarity-adaptive selection handles well. The improvement suggests that PAD-PQRST generalizes well to diverse morphological patterns, but still leaves room for further refinement. Nevertheless, the proposed polarity-adaptive approach still outperforms wavelet-based methods even on lower amplitude waveforms.

Table III summarizes the result of P-wave detections on BUT-PDB. The proposed method achieves F1 score of 80.20%, which is 3% higher than Maritnez2004, Emrich2024, and 40% higher than Neurokit2 CWT. The sensitivity is very high, but the PPV, at 69% is moderate. This means many FP occur, likely due to the shorter, and noisy recording with increased amount abnormal P-waves.

Table IV shows the result of P-wave detection on MIT-BIH-PDB. The proposed method achieves F1 score of 88.33%, essentially tied to Martinez2004, at 88.21%, and substantially higher than Emrich2024, and Neurokit2 CWT. This indicates that on this database, the wavelet-based method remains competitive, but PAD-PQRST matches its performance while being simpler, and parameter-free.

E. Discussion

1) Robustness to morphological and temporal variations: Unlike traditional rule-based methods that rely on rigid, time-invariant forward-search intervals, PAD-PQRST introduces a localized, dynamic coordinate system bounded by piece-wise

TABLE II. SCORE OF PQRST DELINEATION OF COMPARED METHODS IN LUDB.

Method	Metrics	P	Q	R	S	T
Martinez2004	TP	1333	1659	1695	1610	1455
	FP	245	113	90	162	317
	FN	56	119	88	165	172
	Sensitivity	95.97%	93.31%	95.06%	90.70%	89.43%
	PPV	84.47%	93.62%	94.96%	90.86%	82.11%
	DER	21.67%	13.05%	9.98%	18.42%	30.06%
	F1	**89.86%**	**93.46%**	**95.01%**	**90.78%**	**85.61%**
Emrich2024	TP	1350	1702	1698	1686	1489
	FP	221	109	116	128	325
	FN	39	76	85	89	138
	Sensitivity	97.19%	95.73%	95.23%	94.99%	91.52%
	PPV	85.93%	93.98%	93.61%	92.94%	82.08%
	DER	18.72%	10.40%	11.27%	12.23%	28.46%
	F1	**91.22%**	**94.85%**	**94.41%**	**93.95%**	**86.54%**
Neurokit2 CWT	TP	994	1669	1697	1610	1246
	FP	321	59	31	118	282
	FN	395	109	86	165	381
	Sensitivity	71.56%	93.87%	95.18%	90.70%	76.58%
	PPV	75.59%	96.59%	98.21%	93.17%	81.54%
	DER	51.55%	9.45%	6.56%	15.94%	40.75%
	F1	**73.52%**	**95.21%**	**96.67%**	**91.92%**	**78.99%**
Proposed Method	TP	1376	1773	1778	1769	1549
	FP	131	33	28	37	221
	FN	13	5	5	6	78
	Sensitivity	99.06%	99.72%	99.72%	99.66%	95.21%
	PPV	91.31%	98.17%	98.45%	97.95%	87.51%
	DER	10.37%	2.14%	1.85%	2.42%	18.38%
	F1	**95.03%**	**98.94%**	**99.08%**	**98.80%**	**91.20%**

TABLE III. SCORE OF P-WAVE DETECTION OF COMPARED METHODS IN BUT-PDB.

Method	TP	FP	FN	Sensitivity	PPV	DER	F1
Martinez2004	4438	1567	993	81.72%	73.91%	47.14%	77.61%
Emrich2024	4833	2305	598	88.99%	67.71%	53.45%	76.90%
Neurokit2 CWT	2627	4829	2804	48.37%	35.23%	140.55%	40.77%
Proposed Method	5186	2315	245	95.49%	69.14%	47.14%	80.20%

TABLE IV. SCORE OF P-WAVE DETECTION OF COMPARED METHODS IN MIT-BIH-PDB.

Method	TP	FP	FN	Sensitivity	PPV	DER	F1
Martinez2004	19744	2913	2364	89.31%	87.14%	23.87%	88.21%
Emrich2024	20292	9723	1816	91.79%	67.61%	52.19%	77.86%
Neurokit2 CWT	12337	12703	9771	55.80%	49.27%	101.66%	52.33%
Proposed Method	20895	4307	1213	94.51%	82.91%	24.97%	88.33%

continuous functions of instantaneous $R-R$ intervals (Ω_{left} and Ω_{right}). This mathematical adaptation prevents the structural overlap and false positives common to standard rule-based heuristic search windows during extreme tachyarrhythmias or bradyarrhythmias. Furthermore, while wavelet-based delineation methods excel at multi-scale frequency decomposition, they frequently struggle with morphological transformations (such as inverted T-waves or wide-complex bundle branch blocks) because their thresholding and modulus maxima selection schemes implicitly assume standard directional profiles. PAD-PQRST overcomes this limitation through its Polarity-Adaptive Extrema Selection formulation. By optimizing the absolute deviation from the local segment mean ($|f(x) - \mu_S|$) and scaling thresholds dynamically relative to the local range (Δ_S), our framework treats upright, inverted, and biphasic waves with mathematical symmetry. This eliminates the need for predefined shape templates or complex wavelet phase tuning, enabling the model to sustain high F_1 scores across severely deformed pathological datasets.

2) Overall performance: The proposed method achieves superior F1 scores for all PQRST components on four independent datasets. This consistent superiority across different sampling rates, patient populations validate the generalizability of the polarity-adaptive, physiologically constrained approach. Notably, the proposed method requires no dataset-specific tuning, as the established search window are fixed physiological constraints, and the only hyperparameter is the

empirically derived minimum RR interval of 200ms. This is a substantial practical advantage over methods that rely on optimized thresholds or trained parameters.

The proposed method achieves the near perfect detection of QRS complex on QTDB, as well as LUDB, despite the greater morphological diversity. These results have two important implications. First, when the R-peak anchor is accurate, the polarity-adaptive extrema search for Q, and S-wave is highly effective across the full range of QRS morphologies. Secondly, the QRS delineation accuracy of the proposed method is now primarily limited by the quality of the R-peak detector.

For P-wave detection, the proposed PAD-PQRST achieves the highest reported F1 score on all four datasets. The proposed method maintained high sensitivity throughout, which indicates that the method rarely misses a true P-wave. However, the PPV exhibits some fluctuations especially with noisier dataset, which also reflects the inherent difficulty of distinguishing very low-amplitude amplitude P-waves from baseline noise. Additionally, PAD-PQRST also demonstrated competitive results against wavelet-based method while being simpler, parameter-free, and more interpretable.

T-wave delineation remains the most difficult waveform components to detect reliably, as evidenced by the lower F1 scores. Despite this challenge, PAD-PQRST still outperforms the selected benchmark methods in both QTDB, and LUDB. The polarity-adaptive extrema selection, which does not assume an upright T-wave, is particularly beneficial for correctly identifying inverted T-waves. The fixed 150ms minimum separation from the S-wave also prevents confusion with the QRS complex. Nevertheless, the T-wave delineation still leaves room for improvement, and further refinement. Future work could explore adaptive sizing based on heart rate, post-detection morphological classification, or utilization of multi-lead information to resolve ambiguous repolarization patterns.

3) Limitations: Several limitations of the proposed method should be acknowledged. First, although the method is parameter-free, the search window durations for each waveform components were fixed based on standard adult physiology. These may not be optimal for pediatrics, or extremely abnormal recordings, where adaptive window scaling could be more beneficial. Second, as the method processes each heart independently, it does not leverage inter-beat consistency, which could be used to correct the occasional outliers. Third, the current implementation only utilizes single-lead signal, which leaves the possibility for future work to investigate the utilization of multiple lead to improve robustness. Fourth, although our analysis identifies low amplitude as a primary factor in detection errors, the absence of detailed patient metadata in the public dataset precludes a comprehensive morphological sub-analysis.

V. Conclusion

This work introduced a PQRST delineation framework PAD-PQRST built on two foundational principles, physiologically derived search windows and polarity-adaptive extremum selection. The method requires only a reliable R-peak detector, Manikandan2012 for this instance. The search windows are directly rooted in cardiac physiology, while the selection of the extremum with maximal absolute deviation from the segment mean inherently handles upright, inverted, and biphasic waves without additional heuristics. Evaluated primarily on QTDB and LUDB, this proposed methodology proved highly effective with F1-scores over 98.5% for delineation of QRS components, 90.97% to 94.57% for P-waves, and 79.57% to 91.41% for T-waves, consistently surpassing complex wavelet, and graph-based alternatives. The result underscores that a carefully designed, physiologically grounded approach can achieve competitive accuracy while remaining computationally lightweight, and easy to implement.

Future work will translate the PAD-PQRST framework into clinical and consumer-wearable applications by pursuing several key avenues. First, the real-time implementation of the algorithm on low-power embedded devices (such as microcontrollers used in wearable ECG patches) must be explored optimizing the localized mathematics for fixed-point arithmetic. Second, the extension from single-lead to multi-lead analysis will be investigated, allowing the system to leverage spatial redundancy across channels. Third, the adaptive window-scaling mechanisms should be explored that dynamically adjust the boundary parameters based on instantaneous heart rate variability. Finally, this work should incorporate a lightweight signal-quality index (SQI) module to proactively identify and flag highly noisy recordings, ensuring high-fidelity delineation even under severe ambulatory motion artifacts.

Acknowledgment

During the preparation of this work the author(s) used Google's Gemini 3.5 Flash in order to refine and improve the clarity, grammar, and flow of the language in the manuscript. 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 published study.

References

[1] Brady, William J, *Electrocardiogram in Clinical Medicine*, John Wiley & Sons, 2020.

[2] Khan, M Gabriel, *Rapid ECG Interpretation*, Totowa, NJ, Humana Press, 2008.

[3] Mitchell, Chace B and Simmons, Joe and Neal, Sarah J and Cleveland, David C and Bar-Cohen, Yaniv and Cleveland, John D, ECG features in orthotopic cardiac xenotransplantation: comparisons with published literature, *Xenotransplantation*, vol. 31, no. 6, p. e70005, 2024. doi: 10.1111/xen.70005

[4] Haggerty, Christopher M and Poterucha, Timothy J, Moving From PQRST to AI: Advancing Transparency, Reliability, and Clinical Translation in ECG Deep Learning, *JACC: Advances*, vol. 2, no. 10, p. 100682, 2023. doi: 10.1016/j.jacadv.2023.100682

[5] Bonow, Robert O and Mann, Douglas L and Zipes, Douglas P and Libby, Peter, *Braunwald's heart disease e-book: A textbook of cardiovascular medicine*, Elsevier Health Sciences, 2011.

[6] Roden, Dan M.L., Role of the electrocardiogram in determining electrophysiologic end points of drug therapy, *The American Journal of Cardiology*, vol. 62, no. 12, p. 34H-38H, 1988. doi: 10.1016/0002-9149(88)90338-4

[7] Rafie, Nikita, Anthony H. Kashou, and Peter A. Noseworthy, ECG interpretation: clinical relevance, challenges, and advances, *Hearts*, vol. 2, no. 4, p. 505-513, 2021. doi: 10.3390/hearts2040039

[8] Fukuta, Hidekatsu, and William C. Little, The cardiac cycle and the physiologic basis of left ventricular contraction, ejection, relaxation, and filling, *Heart failure clinics*, vol. 4, no. 1, p. 1-11, 2008. doi: 10.1016/j.hfc.2007.10.004

[9] Chan-Dewar, Fang, The cardiac cycle, *Anaesthesia & Intensive Care Medicine*, vol. 13, no. 8, p. 391-396, 2012. doi: 10.1016/j.mpaic.2012.05.010

[10] Ashley, Euan A., Euan Ashley, and Josef Niebauer, Cardiology explained, London: Remedica Publishing 2004.

[11] Bae, Tae Wuk and Kwon, Kee Koo, ECG PQRST complex detector and heart rate variability analysis using temporal characteristics of fiducial points, *Biomedical signal processing and control*, vol. 66, p. 102291, 2021. doi: 10.1016/j.bspc.2020.102291

[12] Wagner, Galen S and Strauss, David G, *Marriott's practical electrocardiography*, Lippincott Williams & Wilkins, 2013.

[13] The University of Nottingham, *A beginners guide to normal heart function, sinus rhythm & common cardiac arrhythmias*, https://www.nottingham.ac.uk/nursing/practice/resources/cardiology/function/normal_duration.php, 2025.

[14] Pérez-Riera, Andrés Ricardo and de Abreu, Luiz Carlos and Barbosa-Barros, Raimundo and Nikus, Kjell C and Baranchuk, Adrian, journal=Annals of Noninvasive Electrocardiology, R-peak time: An electrocardiographic parameter with multiple clinical applications, *Annals of Noninvasive Electrocardiology*, vol. 21, no. 1, p. 10–19, 2016. doi: 10.1111/anec.12323

[15] Pava, Luis Fernando and Perafán, Pablo and Badiel, Marisol and Arango, Juan José and Mont, Lluis and Morillo, Carlos A and Brugada, Josep, R-wave peak time at DII: a new criterion for differentiating between wide complex QRS tachycardias, *Heart rhythm*, vol. 7, no. 7, p. 922–926, 2010. doi: 10.1016/j.hrthm.2010.03.001

[16] ChuDuc, Hoang and NguyenPhan, Kien and NguyenViet, Dung, A review of heart rate variability and its applications, *APCBEE procedia*, vol. 7, p. 80–85, 2013. doi: 10.1016/j.apcbee.2013.08.016

[17] Bhandare, Trupti Vasantrao and Rangasamy, Selvarani and Shelke, Chetan J, A Deep Learning Classification Approach using Feature Fusion Model for Heart Disease Diagnosis, *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 6, 2022. doi: 10.14569/IJACSA.2022.0130677

[18] Rahman, Md Obaidur and Kashem, Mohammod Abul and Al-Akhir, Nayan and Most, Fahmida Akter and Rabbi, Fazly and Ahmed, Marzia and Asaduzzaman, Mohammad, Internet of Things (IoT) Based ECG System for Rural Health Care, *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, 2021. doi: 10.14569/IJACSA.2021.0120653

[19] Martínez, Juan Pablo and Almeida, Rute and Olmos, Salvador and Rocha, Ana Paula and Laguna, Pablo, A wavelet-based ECG delineator: evaluation on standard databases, *IEEE Transactions on biomedical engineering*, vol. 51, no. 4, p. 570–581, 2004. doi: 10.1109/TBME.2003.821031

[20] Kalyakulina, Alena I. and Yusipov, Igor I. and Moskalenko, Victor A. and Nikolskiy, Alexander V. and Kozlov, Artem A. and Zolotykh, N. Yu and Ivanchenko, Mikhail V., Finding morphology points of electrocardiographic-signal waves using wavelet analysis, *Radiophysics and Quantum Electronics*, vol. 61, no. 8, p. 689–703, 2019. doi: 10.1007/s11141-019-09929-2

[21] Bote, Jose Manuel and Recas, Joaquin and Rincón, Francisco and Atienza, David and Hermida, Román, A modular low-complexity ECG delineation algorithm for real-time embedded systems, *IEEE journal of biomedical and health informatics*, vol. 22, no. 2, p. 429–441, 2017. doi: 10.1109/JBHI.2017.2671443

[22] Emrich, Jonas and Gargano, Andrea and Koka, Taulant and Muma, Michael, Physiology-Informed ECG Delineation Based on Peak Prominence, *2024 32nd European Signal Processing Conference (EUSIPCO)*, p. 1402–1406, 2024. doi: 10.23919/EUSIPCO63174.2024.10715353

[23] Makowski, Dominique and Pham, Tam and Lau, Zen J and Brammer, Jan C and Lespinasse, François and Pham, Hung and Schölzel, Christopher and Chen, SH Annabel, NeuroKit2: A Python toolbox for neurophysiological signal processing, *Behavior research methods*, vol. 53, no. 4, p. 1689–1696, 2021. doi: 10.3758/s13428-020-01516-y

[24] Berkaya, Selcan Kaplan and Uysal, Alper Kursat and Gunal, Efnan Sora and Ergin, Semih and Gunal, Serkan and Gulmezoglu, M Bilginer, A survey on ECG analysis, *Biomedical Signal Processing and Control*, vol. 43, p. 216–235, 2018. doi: 10.1016/j.bspc.2018.03.003

[25] Manikandan, M Sabarimalai and Soman, KP, A novel method for detecting R-peaks in electrocardiogram (ECG) signal, *Biomedical signal processing and control*, vol. 7, no. 2, p. 118–128, 2012. doi: 10.1016/j.bspc.2011.03.004

[26] Pan, Jiapu and Tompkins, Willis J, A real-time QRS detection algorithm, *IEEE transactions on biomedical engineering*, no. 3, p. 230–236, 2007. doi: 10.1109/TBME.1985.325532

[27] Zong, Wei and Moody, GB and Jiang, D, A robust open-source algorithm to detect onset and duration of QRS complexes, *Computers in Cardiology, 2003*, p. 737–740, 2003. doi: 10.1109/CIC.2003.1291261

[28] Christov, Ivaylo I, Real time electrocardiogram QRS detection using combined adaptive threshold, *Biomedical engineering online*, vol. 3, no. 1, p. 28, 2004. doi: 10.1186/1475-925X-3-28

[29] Elgendi, Mohamed and Jonkman, Mirjam and De Boer, Friso, Frequency bands effects on QRS detection, *Biosignals*, vol. 2003, p. 2002, 2010. doi: 10.5220/0002742704280043

[30] Lourenço, André and Silva, Hugo and Leite, Paulo and Lourenço, Renato and Fred, Ana, Real time electrocardiogram segmentation for finger based ECG biometrics, *International Conference on Bioinspired Systems and Signal Processing*, vol. 2, p. 49–54, 2012. doi: 10.5220/0003777300490054

[31] Kalidas, Vignesh and Tamil, Lakshman, Real-time QRS detector using stationary wavelet transform for automated ECG analysis, *2017 IEEE 17th international conference on Bioinformatics and Bioengineering (BIBE)*, p. 457–461, 2017. doi: 10.1109/BIBE.2017.00-12

[32] Nabian, Mohsen and Yin, Yu and Wormwood, Jolie and Quigley, Karen S and Barrett, Lisa F and Ostadabbas, Sarah, An open-source feature extraction tool for the analysis of peripheral physiological data, *IEEE journal of translational engineering in health and medicine*, vol. 6, p. 1–11, 2018. doi: 10.1109/JTEHM.2018.2878000

[33] Rodrigues, Tiago and Samoutphonh, Sirisack and Silva, Hugo and Fred, Ana, A low-complexity R-peak detection algorithm with adaptive thresholding for wearable devices, *2020 25th international conference on pattern recognition (ICPR)*, p. 1–8, 2021. doii: 10.1109/ICPR48806.2021.9413245

[34] Emrich, Jonas and Koka, Taulant and Wirth, Sebastian and Muma, Michael, Accelerated sample-accurate r-peak detectors based on visibility graphs, *2023 31st European Signal Processing Conference (EUSIPCO)*, p. 1090–1094, 2023. doi: 10.23919/EUSIPCO58844.2023.10290007

[35] Sasikala, P and Wahidabanu, RS, Robust R Peak and QRS detection in Electrocardiogram using Wavelet Transform, *International Journal of Advanced Computer Science and Applications*, vol. 1, no. 6, 2010. doi: 10.14569/IJACSA.2010.010608

[36] Dang, Manh Duong and Luangsodsai, Arthorn and Sinapiromsaran, Krung, Accelerated sample-accurate r-peak detectors based on visibility graphs, *Journal of Computational Mathematics and Data Science*, p. 100135, 2026. doi: 10.1016/j.jcmds.2026.100135

[37] Madona, Putri and Basti, Rahmat Ilias and Zain, Muhammad Mahrus, PQRST wave detection on ECG signals, *Gaceta sanitaria*, vol. 35, p. S364–S369, 2021. doi: 10.1016/j.gaceta.2021.10.052

[38] Ilkhanoff, Leonard and Soliman, Elsayed Z and Ning, Hongyan and Liu, Kiang and Lloyd-Jones, Donald M, Factors associated with development of prolonged QRS duration over 20 years in healthy young adults: the Coronary Artery Risk Development in Young Adults study, *Journal of electrocardiology*, vol. 45, no. 2, p. 178–184, 2012. doi: 10.1016/j.jelectrocard.2011.11.001

[39] Lee, Duck Hee and Park, Jun Woo and Choi, Jeasoon and Ahmed, Rabbi, Automatic detection of electrocardiogram ST segment: Application in ischemic disease diagnosis, *International journal of advanced computer science and applications*, vol. 4, no. 2, 2013. doi: 10.14569/IJACSA.2013.040222

[40] Laguna, P. and Mark, R.G. and Goldberg, A. and Moody, G.B., A database for evaluation of algorithms for measurement of QT and other waveform intervals in the ECG, *Computers in Cardiology 1997*, p. 673-676, 1997. doi: 10.1109/CIC.1997.648140

[41] Kalyakulina, Alena and Yusipov, Igor and Moskalenko, Viktor and Nikolskiy, Alexander and Kosonogov, Konstantin and Zolotykh, Nikolai and Ivanchenko, Mikhail, Lobachevsky university electrocardiography database, *PhysioNet*, 2021. doi: 10.13026/eegm-h675

[42] Maršánová, Lucie and Smíšek, Radovan and Němcová, Andrea and Smital, Lukáš and Vítek, Martin, Brno University of Technology ECG signal database with annotations of P wave (BUT PDB), *PhysioNet*, 2021. doi: 10.21203/rs.3.rs-942006/v1

[43] Maršánová, Lucie and Němcová, Andrea and Smíšek, Radovan and Goldmann, Tomáš and Vítek, Martin and Smital, Lukáš, Automatic detection of P wave in ECG during ventricular extrasystoles, *World Congress on Medical Physics and Biomedical Engineering 2018*, p. 381–385, 2018. doi: 10.1007/978-981-10-9038-7_72

[44] Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms Assoc. Adv, *ANSI/AAMI EC57:1998/(R2020)*, p. 381–385, 2020.

Privacy-Preserving Feature Engineering Framework for Real-World Vulnerability Risk Prediction

Jawharah Albarakati, Tawfiq Hasanin, Suaad Alarif
Department of Information Systems, King Abdulaziz University, Jeddah, Saudi Arabia

***Abstract*—This study introduces a privacy-preserving data preparation and feature engineering framework designed for machine learning-based cybersecurity vulnerability risk prediction, utilizing real-world enterprise scan data from an operational cloud environment. To transform raw vulnerability records into a structured, machine-readable format while complying with personal data protection regulations, the framework integrates systematic data cleaning, missing value imputation, categorical encoding, text normalization, and host pseudonymization. Exploratory data analysis (EDA) was conducted on 36,940 operational records to examine dataset characteristics, vulnerability distributions, and severe class imbalances. Multiple classifiers—including Logistic Regression, Random Forest, XGBoost, and LightGBM—were evaluated under extreme imbalance conditions mitigated via SMOTE. Experimental results demonstrate that non-linear tree-based ensembles achieved the highest predictive performance, with LightGBM attaining a macro F1-score of 0.9974. Comparative analysis further indicates that the effectiveness of textual representations depends on the underlying classifier architecture, with semantic embeddings yielding the best performance when combined with tree-based ensemble models, while TF-IDF remained competitive for linear classificationThese findings underscore the critical role of structured text preparation and semantic representation in enhancing model reliability, interpretability, and scalability for automated vulnerability prioritization.**

***Keywords*—Data pre-processing; machine learning; protecting sensitive data; stemming; pseudonymization**

I. Introduction

Cybersecurity has become a critical concern for organizations due to the increasing frequency and sophistication of cyber threats. Vulnerability assessment plays a central role in identifying and prioritizing security weaknesses across enterprise infrastructures, including networks, servers, applications, and cloud environments.

The effectiveness of machine learning–based vulnerability management depends heavily on the quality of the underlying data. However, cybersecurity logs are often noisy, incomplete, and heterogeneous, requiring preprocessing techniques such as missing-value handling, normalization, categorical encoding, and text feature extraction. These steps improve data quality and transform raw telemetry into representations suitable for predictive risk analysis.

In this study, operational security data was collected using the Nessus Vulnerability Scanner, a widely adopted industry-standard solution for identifying system vulnerabilities [1]. Nessus facilitates comprehensive infrastructure auditing across cloud environments, producing highly detailed vulnerability records. To leverage this data safely, raw scan reports are systematically processed to eliminate operational duplicates, validate network configurations, and mask sensitive host telemetry. This ensures strict adherence to data privacy standards and compliance with personal data protection regulations without degrading the predictive utility of the features.

Traditional vulnerability prioritization approaches often rely on static scoring methods or keyword-based techniques that may fail to capture the semantic information contained in vulnerability descriptions. To address this limitation, this study develops a privacy-preserving feature engineering framework that integrates structured vulnerability attributes with textual representations derived from both TF-IDF and Sentence-BERT (SBERT). In addition, exploratory data analysis (EDA) is conducted to examine data characteristics, feature distributions, and class imbalance patterns before model development.

The main contributions of this study are as follows:

- A privacy-preserving data preparation and feature engineering framework for real-world enterprise vulnerability telemetry, incorporating host pseudonymization to support compliance with Personal Data Protection Law (PDPL) requirements.
- A comparative evaluation of lexical (TF-IDF) and semantic (SBERT) text representations for encoding unstructured vulnerability descriptions.
- An empirical assessment of four machine learning models (Logistic Regression, Random Forest, XGBoost, and LightGBM) under severe class imbalance conditions using SMOTE-based mitigation.
- An analysis of the impact of semantic representations on vulnerability severity prediction performance in operational enterprise environments.

II. Related Work

Vulnerability assessment is a cornerstone of proactive cybersecurity risk management, aimed at identifying, analyzing, and prioritizing system flaws. Traditional vulnerability management heavily relies on static rule-based scanning tools and standardized frameworks, such as the Common Vulnerability Scoring System (CVSS) [2][3]. While efficient for cataloging known security flaws, these traditional methods lack adaptability and fail to capture the dynamic, contextual properties of real-world enterprise vulnerability telemetry [4].

To overcome these limitations, recent research has shifted toward supervised machine learning algorithms to enhance predictive risk prioritization. For instance, Sabottke et al. utilized machine learning to predict vulnerability exploitability

by leveraging language attributes from vulnerability descriptions [5], while Neuhaus et al. demonstrated that supervised learning models can effectively prioritize vulnerabilities based on historical exploitation trends [6]. Similarly, recent advancements have focused on exploitability prediction and automated severity forecasting using high-dimensional datasets [7], [8]. However, a persistent limitation in these approaches is their reliance on clean, static, and balanced batch-learning benchmarks, which rarely reflect the chaotic nature of real-world, live operational environments.

The predictive performance of these machine learning frameworks is fundamentally constrained by the quality and representation of the underlying data. Raw vulnerability logs are notoriously noisy, unstructured, and heterogeneous, demanding extensive preprocessing [9][10]. Koukaras and Tjortjis emphasized that rigorous feature engineering and data preparation are critical prerequisites for meaningful data mining and analysis [9]. In the context of textual cybersecurity data, standard Natural Language Processing (NLP) techniques—such as tokenization, stemming, and Term Frequency-Inverse Document Frequency (TF-IDF) weighting—have been widely adopted to extract patterns from unstructured vulnerability descriptions [11], [12]. While integrating these lexical features with structured attributes has been shown to improve accuracy [13], traditional keyword-matching techniques like TF-IDF fail to capture the deeper semantic nuances of security vulnerabilities. Although dense semantic representations (e.g., Transformer-based embeddings) have revolutionized general NLP, their systematic evaluation against traditional lexical features remains under-explored in operational vulnerability frameworks.

Concurrently, compliance with data privacy mandates has emerged as a critical requirement in cybersecurity analytics. Implementing machine learning on live enterprise telemetry introduces severe privacy risks, as scanner logs often contain sensitive network topologies and host identifiers. To mitigate this, various privacy-preserving preprocessing strategies, such as data masking and pseudonymization, have been proposed to safeguard identities while preserving analytical utility [14], [15]. Truong et al. and Riabi et al. demonstrated that secure pseudonymization strategies can satisfy regulatory constraints (such as the GDPR or PDPL) without degrading downstream classifier accuracy [16], [17].

Despite these advancements, existing literature often treats data preprocessing, privacy preservation, text representation, and class imbalance mitigation as isolated problems. Many studies evaluate models under idealized, balanced, or strictly static conditions, overlooking the severe class imbalances inherent to operational environments [18].

Unlike prior works that focus solely on standalone classification performance or isolated exploit prediction, this study introduces a unified, systematically integrated framework. We combine regulatory-compliant privacy preservation, hybrid structured-textual feature engineering (lexical vs. deep semantic embeddings), and robust class imbalance mitigation using SMOTE. The contribution of this work lies not in proposing a novel standalone classifier, but in empirically demonstrating how a meticulously engineered data preparation and semantic feature construction pipeline can maximize operational risk classification fidelity using heterogeneous enterprise cybersecurity data (Fig. 1).

III. Research Methodology

The proposed research methodology consists of five sequential phases. First, vulnerability data were collected from a real-world enterprise cloud environment using the Nessus vulnerability scanner. Second, data preprocessing and privacy-preserving transformations were applied. Third, feature engineering was performed. Fourth, multiple supervised machine learning models were trained, while class imbalance was addressed using SMOTE and class-weighted learning strategies. Finally, a comparative experimental evaluation was conducted to assess model performance using several classification metrics across multiple vulnerability risk categories.

A. Data Collection and Acquisition

The foundational phase of establishing a robust machine learning-based vulnerability assessment framework begins with systematic data acquisition. The empirical data utilized in this study was sourced from the Nessus Vulnerability Scanner, an industry-standard network security auditing tool. Nessus was deployed across a production cloud-based enterprise infrastructure to continuously audit and discover security flaws across heterogeneous enterprise domains, including network devices, internal servers, and organizational endpoints.

The collected telemetry comprises a rich mix of structured security indicators and unstructured textual attributes. Specifically, the captured dataset includes critical variables such as vulnerability severity levels, unique host identifiers, textual remediation solutions, Common Vulnerabilities and Exposures (CVE) identifiers, and binary indicators of system exploitability. This multifaceted data collection ensures that both the structural context and the natural language nuances of the vulnerabilities are preserved for downstream analysis.

To reflect authentic operational challenges, the vulnerability logs were harvested from a live cloud environment operating within the Kingdom of Saudi Arabia (KSA). Due to strict organizational security and confidentiality mandates, the identity of the target enterprise is anonymized. Crucially, all data acquisition workflows were conducted in strict compliance with internal corporate security policies and PDPL, establishing a legally compliant and privacy-preserving data foundation without compromising the analytical utility of the security logs.

B. Privacy-Preserving Data Preprocessing

Data preprocessing is an essential step in machine learning because it improves data quality and ensures that features are represented consistently for model training [4]. Since cybersecurity datasets often contain heterogeneous data types, missing values, and noisy observations, preprocessing techniques are required to prepare the data for reliable predictive modeling [19].

In addition to data quality considerations, protecting sensitive organizational information is a critical requirement in cybersecurity analytics [16]. To address this requirement, the

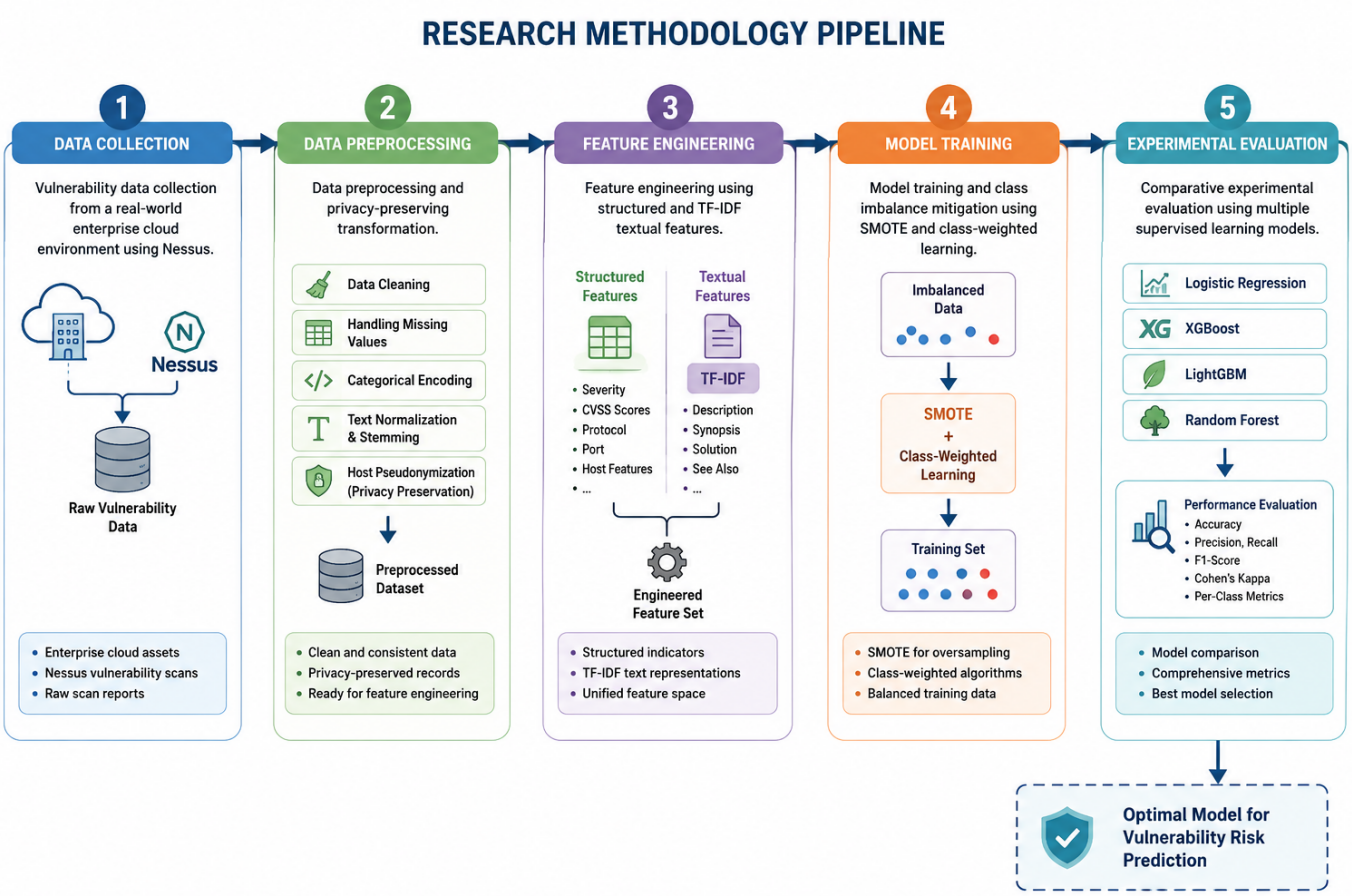


Fig. 1. Proposed privacy-preserving machine learning pipeline for cybersecurity vulnerability risk prediction.

proposed framework applies host pseudonymization, replacing identifiable infrastructure information with deterministic pseudonyms while preserving the utility of the data for subsequent machine learning tasks [17].

1) Host pseudonymization and privacy preservation: The operational telemetry harvested during the data acquisition phase contained explicit network configurations within the raw `Host` attribute, exposing corporate IP footprints and introducing an unacceptable vector for information leakage. To neutralize this vulnerability and ensure strict compliance with the Saudi Arabian Personal Data Protection Law (PDPL) mandated in the region, a deterministic pseudonymization mapping was integrated into the preprocessing pipeline.

Each raw infrastructure IP address was passed through a secure Secure Hash Algorithm 256-bit (SHA-256) cryptographic function. To minimize storage footprints and reduce data dimensionality, the resulting 64-character hexadecimal strings were systematically truncated to their leading 10 characters. The resulting hashes were truncated to the first 10 hexadecimal characters to reduce storage requirements while maintaining unique host identifiers within the dataset, thereby preserving the categorical utility of host-specific attributes for the downstream tree-based classifiers. Table I illustrate a comparative sample of the enterprise logs before and after the application of pseudonymization layer.

TABLE I. SAMPLE RECORDS BEFORE AND AFTER HOST PSEUDONYMIZATION

Idx	Severity	Host
Original Dataset		
0	Critical	12.0.2.34
1	Critical	12.0.2.24
2	Medium	18.51.100.34
3	High	23.0.113.34
4	Low	23.0.113.44
Pseudonymized Dataset		
0	Critical	43741ff074
1	Critical	32d4a72597
2	Medium	6046c9aa49
3	High	576264d7a0
4	Low	9ab7244c54

2) Missing value analysis and feature elimination: Operational cyber threat logs frequently suffer from data incompleteness due to sensor telemetry dropouts, intermittent database connection syncs, or unpopulated third-party vendor fields. In this pipeline, missing values within remaining feature channels were preserved strictly as mathematical null states (NaN)

TABLE II. COMPARATIVE ANALYSIS OF TEXTUAL ATTRIBUTES BEFORE AND AFTER STEMMING

Attribute	Original Text	Stemmed Text
Description	According to its version, there is at least one...	accord to version, there is at least one...
Synopsis	The remote host contains one or more unsupport...	remot host contain one or more unsupport...
Solution	Upgrade to a version that is currently supported.	upgrad to version current support.

rather than being replaced by generic textual placeholders. This programmatic constraint ensures that downstream natural language processing modules do not optimize over artificial tokens, preventing representation leakage.

To safeguard predictive reliability, a strict data-quality threshold was established: any attribute column exhibiting an empirical missingness rate exceeding 90% was systematically eliminated from the global feature space. Under this rigorous constraint, two major structural attributes were pruned: the `CANVAS` indicator matrix (exhibiting approximately 97.75% missingness) and the `Metasploit` exploitability matrix (exhibiting approximately 91.10% missingness). Removing these hyper-sparse features neutralizes high-variance noise, preventing downstream classifiers from expanding optimization parameters over systematically unpopulated data fields.

3) Categorical severity label encoding: To transform the ordinal, textual target categories within the vulnerability corpus into actionable mathematical formats compliance with gradient descent optimization loops, an ordinal Label Encoding transformation was applied. The original risk assessment categories—spanning from baseline configurations to critical exposures—were mapped onto a standardized discrete numerical scale: None $\rightarrow$ 0, Low $\rightarrow$ 1, Medium $\rightarrow$ 2, High $\rightarrow$ 3, and Critical $\rightarrow$ 4.

4) Text normalization and stemming: Column normalization involved the removal of extraneous whitespaces, special characters, and newline symbols from column names to ensure consistent and standardized naming conventions (e.g., converting " Risk " to "Risk"). Text normalization included converting all text-based fields to string data types, trimming leading and trailing spaces, and standardizing internal spacing for uniformity. Additionally, empty string values "") were replaced with NaN to facilitate accurate handling of missing data during subsequent analysis. Text preprocessing included systematic stemming of the *Description*, *Synopsis*, and *Solution* fields using the Porter Stemmer algorithm to reduce lexical variability and optimize token consistency. This linguistic normalization was applied uniformly prior to feature extraction, mapping inflectional variations to their respective canonical root forms. Table II showcases a comparative sample of the textual attributes before and after executing the stemming routine.

C. Data Labeling and Target Grounding

In supervised machine learning frameworks for cybersecurity analytics, establishing ground-truth annotations is a foundational step for training predictive classifiers. Rather than relying on manual or crowd-sourced labeling, which introduces subjective bias and operational delays, this framework leverages automated, deterministic vulnerability annotations directly derived from the Nessus vulnerability scanner telemetry. This methodology guarantees consistent target grounding based on industry-standard security metrics.

TABLE III. UNIFIED OPERATIONAL SCHEMA MAPPING SEVERITY TARGETS, HOST ATTRIBUTES, UNSTRUCTURED CONTEXTS, AND STANDARDIZED CVE IDENTIFIERS

Risk	Host	Synopsis	Solution	CVE
Critical	192.0.2.34	Unsupported software versions detected.	Upgrade to a supported version.	CVE-2020-6457
Critical	192.0.2.34	Microsoft Silverlight is installed.	Remove Silverlight from the host.	CVE-2020-6461
Critical	192.0.2.55	Installed browser contains vulnerabilities.	Upgrade Chrome to v81.0.4044.113+.	CVE-2020-6831
Low	192.0.2.78	Browser configuration issues detected.	Upgrade Chrome to v81.0.4044.129+.	CVE-2008-5161
Medium	192.0.2.99	Distributed framework is vulnerable.	Upgrade Chrome to v81.0.4044.138+.	CVE-2023-20867

The dataset utilizes a multi-attribute labeling scheme to facilitate distinct vulnerability categorization tasks. The features chosen to define the target classes include:

- Risk Severity Level: Sourced from the `Risk` attribute, this categorical variable acts as the primary multi-class classification target. It stratifies vulnerabilities into standardized threat envelopes (*None*, *Low*, *Medium*, *High*, and *Critical*), allowing machine learning models to map structured and textual telemetry to prospective severity levels.
- Standardized Vulnerability Identifiers: Captured within the `CVE` attribute, these labels map instances to the Common Vulnerabilities and Exposures (CVE) database. This identifier provides a deterministic reference for validating tracking behaviors and checking model generalization across disjoint vulnerability sets.

To ensure high data fidelity before model ingestion, an automated extraction parser mapped the Nessus security scanner outputs to distinct class definitions. Subsequently, a programmatic cross-verification audit was executed to ensure alignment between the assigned severity labels and their respective CVSS scores. This process eliminated any noise introduced by scanner version mismatches.

To illustrate the underlying structural schema and demonstrate how target severity classes, host attributes, and unstructured textual features are aligned seamlessly within a single record space, Table III provides a unified, de-duplicated overview of representative operational data samples across the corpus.

D. Hybrid Feature Engineering Pipeline: Lexical vs. Deep Semantic Paradigms

To translate the raw, heterogeneous vulnerability scanner logs into unified tensor formats optimal for supervised machine learning, a robust, dual-methodological feature engineering pipeline was engineered. Cybersecurity telemetry presents a distinct architectural challenge: it couples structured categorization metrics (such as encoded risk and binary flags) with high-dimensional, unstructured linguistic blocks found within the `Synopsis`, `Description`, and `Solution` fields. To

capture both explicit keyword distributions and implicit contextual relationships, this framework constructs an augmented feature space by executing and comparing two distinct Natural Language Processing (NLP) paradigms: traditional lexical frequency mapping and deep transformer-based semantic vectorization.

1) Lexical representation via TF-IDF mappings: The baseline textual feature space is established using the statistical Term Frequency–Inverse Document Frequency (TF-IDF) vectorization paradigm [20]. This lexical scheme scores discrete unigram tokens based on their localized occurrence frequency within an individual log entry relative to their global distribution across the entire enterprise corpus, scaling down ubiquitous, low-variance terms while highlighting rare, security-specific keywords. Mathematically, the weight allocated to a distinct token t within an individual vulnerability record d belonging to the broader global corpus D is formulated via smooth inverse document frequency as follows:

$$\text{TF-IDF}(t,d,D) = \text{TF}(t,d) \times \log\left(\frac{1+|D|}{1+|\{d \in D : t \in d\}|}\right) + 1 \quad (1)$$

where, $TF(t,d)$ represents the raw term frequency of token t in document d, and the denominator computes the inverse document frequency with additive smoothing to prevent division-by-zero. Within this pipeline, textual records were represented using a unigram configuration (`ngram_range=(1,1)`), while the TF–IDF vocabulary size was limited to 500 features (`max_features=500`) to balance representational capacity, computational efficiency, and model complexity. This dimensionality was selected as a practical trade-off to reduce feature sparsity while preserving the most informative vulnerability-related terms across the 36,940 operational records. Although computationally efficient and highly transparent for feature importance analysis, this lexical representation operates under the bag-of-words assumption, ignoring word order, syntactic structure, and contextual semantic relationships between terms.

2) Deep semantic embeddings via Sentence-BERT (SBERT): To resolve the structural limitations of keyword matching and capture context-aware semantic nuances without suffering from vocabulary sparsity, a deep semantic vectorization pipeline was engineered utilizing a Sentence-BERT (SBERT) architecture [9]. SBERT modifies standard pre-trained BERT networks by deploying siamese and triplet network structures to generate highly descriptive, fixed-length dense vector representations for full sentence and paragraph blocks. For this framework, the optimized, pre-trained `all-MiniLM-L6-v2` transformer model was deployed.

When the concatenated textual attributes of a vulnerability report are ingested, the transformer processes the entire textual sequence holistically, mapping its core syntactic context into a compressed 384-dimensional dense continuous numerical vector space. Unlike traditional lexical distributions, this allows the framework to inherently capture structural and operational synonymy within cybersecurity language. Consequently, phrases describing identical security implications (e.g., "*flaw allows remote compromise*" and "*vulnerability permits external exploitation*") are mapped to highly proximal vector coordinates within the continuous manifold, even when their underlying vocabularies are entirely disjoint.

TABLE IV. QUANTITATIVE DISTRIBUTION OF VULNERABILITY SEVERITY LEVELS REVEALING SEVERE OPERATIONAL CLASS IMBALANCE

Vulnerability Severity Level	Absolute Record Count	Relative Proportion (%)
Critical	16,703	45.22%
High	124	0.34%
Medium	2,358	6.38%
Low	9,295	25.16%
None	8,460	22.90%
Total Corpus Size	**36,940**	**100.00%**

IV. EXPERIMENTAL DESIGN AND LEARNING SETUP

This section details the empirical environment, exploratory data profiling, and laboratory configurations deployed to validate the predictive performance of the proposed framework. To guarantee rigorous experimental reproducibility, the processing and learning pipelines were constructed using a standardized software ecosystem based on Python 3.10. The architecture integrates `scikit-learn` for core mathematical utilities, `imbalanced-learn` for syntactic data augmentation, and highly optimized implementations of `LightGBM` and `XGBoost`. Computational workloads were compiled on an Intel Core i7 workstation equipped with 16 GB of RAM and hardware acceleration primitives to minimize dense vector generation latencies during transformer inference.

A. Exploratory Data Analysis (EDA)

Rather than executing modeling processes over unverified logs, a systematic Exploratory Data Analysis (EDA) was performed to deconstruct the statistical topology of the ingested telemetry [21], [22]. The experimental dataset encapsulates a total of 36,940 operational cybersecurity records extracted from a production corporate cloud network. Structurally, each telemetry instance is defined across a 10-dimensional attribute space comprising categorical system indicators (`Risk`, `Host`, `Name`) alongside dense text fields (`Synopsis`, `Description`, `Solution`, `See Also`, `Plugin Output`, `CVE`, and `Core Impact`).

The dataset captures a heterogeneous technological ecosystem, tracking active vulnerabilities across varied operating systems (e.g., Windows, Unix, ESXi), enterprise software configurations (e.g., WinSCP, WinRAR, Google Chrome), and infrastructure security protocols (e.g., SSL, SSH, TLS). This diverse composition provides a realistic operational foundation for training models capable of generalizing across varied real-world network deployment environments.

B. Vulnerability Severity Profiles and Empirical Statistics

A key investigative focus of the EDA phase involved analyzing the empirical distribution of the target classification variable, `Risk`. The enterprise vulnerabilities are stratified into five operational severity domains: *Critical*, *High*, *Medium*, *Low*, and *None*. Table IV provides a rigorous quantitative breakdown of these target severity allocations across the entire 36,940-record corpus.

TABLE V. EMPIRICAL RANKING OF THE TOP 10 MOST FREQUENT INFRASTRUCTURE VULNERABILITIES

Identified Vulnerability Signature Name	Occurrences
DCE Services Enumeration	1,493
Netstat Portscanner (WMI)	1,242
Microsoft Windows Remote Listeners Enumeration (WMI)	1,215
KB5028168: Security Update (July 2023)	1,128
KB5036896: Security Update (April 2024)	1,027
KB5031361: Security Update (October 2023)	936
KB5025229: Security Update (April 2023)	828
Nessus SYN Scanner	774
Service Detection	631
KB5023702: Security Update (March 2023)	624

TABLE VI. LINGUISTIC DISTRIBUTION, LENGTH VARIANCE, AND DATA COMPLETENESS OF TEXTUAL FEATURES

Textual Field	Min	Max	Mean	Median	NaN	Missing (%)
Name	11	149	56.4	54.0	0	0.00%
Synopsis	25	159	64.3	61.0	0	0.00%
Description	27	23,526	1,127.2	451.0	0	0.00%
Solution	3	833	36.9	29.0	13,571	36.74%
See Also	3	1,916	74.5	42.0	11,635	31.50%
Plugin Output	3	32,766	628.9	207.0	564	1.53%
CVE	3	16	8.5	13.0	17,932	48.54%

The empirical breakdown in Table IV exposes a highly non-uniform distribution that presents an acute class-imbalance profile. High-severity threats heavily dominate the data footprint, wherein the *Critical* class alone comprises 16,703 records, representing 45.22% of the dataset. When cumulatively aggregated with the *Low* severity category (25.16%), these two non-contiguous operational bands account for exactly 70.38% of the entire data space, representing a collective volume of 25,998 instances. Conversely, the *High* severity class represents an extreme minority boundary anomaly, containing only 124 records (0.34%). This localized scarcity poses a substantial risk of gradient starvation during model training, as standard loss functions would favor optimizing for the majority classes while neglecting rarer flags. This finding provides clear empirical justification for incorporating host-disjoint stratified partitioning and minority oversampling algorithms in subsequent training phases.

To contextualize the behavioral characteristics of the dataset, Table V isolates the top 10 most frequent vulnerability signatures by operational name. The distribution is dominated by system discovery and patch telemetry, led by *DCE Services Enumeration* (1,493 instances) and *Netstat Portscanner* (1,242 instances), alongside recurring Microsoft Windows cumulative security update sequences (e.g., KB5028168 and KB5036896).

C. Granular Textual Feature Statistical Distribution

Because the proposed framework heavily relies on extracting hybrid representations from unstructured text, a detailed length and variance analysis was executed across the seven primary narrative attributes: `Name`, `Synopsis`, `Description`, `Solution`, `See Also`, `Plugin Output`, and `CVE`. Table VI provides a comprehensive statistical summary of character length behaviors and data completeness metrics across these linguistic dimensions.

The empirical character profiles in Table VI demonstrate significant structural heterogeneity. The `Name` and `Synopsis` segments exhibit short, highly constrained boundaries, averaging 56.4 and 64.3 characters, with optimal data completeness (0% missing entries). Conversely, the `Description` attribute contains the densest contextual data, showing immense variance with sequence lengths scaling up to 23,526 characters and a substantial mean length of 1,127.2. This confirms that descriptions possess high semantic richness, making deep contextual embeddings via Sentence-BERT crucial to capture long-range technical dependencies.

Furthermore, critical systemic fields like `Solution` and `See Also` introduce sparse data spaces, exhibiting 13,571 (36.74%) and 11,635 (31.50%) missing values, respectively. The most severe sparsity occurs within the `CVE` column, where 17,932 instances lack standardized identifiers (48.54% missing rate). This high missing rate underscores the limitation of relying solely on external database cross-referencing for risk prediction. It highlights the absolute necessity of our text-driven approach, which can predict severity directly from scanner logs even when standard CVE references are missing.

D. Strategic Data Partitioning and Class Imbalance Optimization

To validate the proposed predictive pipeline under rigorous experimental bounds, the 36,940 corporate telemetry records were partitioned into a training cohort (80%) and an independent testing cohort (20%). This partitioning executed a stratified sampling design to guarantee that the empirical class proportions—extending from the dominant *Critical* severity profile to the extreme minority *High* anomaly boundary—remain structurally identical across both the training and evaluation subsets. Crucially, this partition was executed utilizing a host-disjoint boundary split, ensuring that identical target IP addresses or asset containers do not span across both partitions, thereby eliminating validation artifacts stemming from data leakage.

Given the long-tailed, severe class imbalance quantified during the exploratory data analysis, relying on standard global loss minimization would lead downstream classifiers to systematically optimize for majority classes, rendering them blind to rare, critical vulnerabilities. To mitigate this algorithmic bias, a dual-methodological optimization strategy was engineered, tailoring the class imbalance correction mechanism to the architectural requirements of each respective classifier family:

1) Cost-sensitive class-weighted learning: For linear models (Logistic Regression) and boosting frameworks dependent on objective gradient trajectories (XGBoost and LightGBM), an algorithmic class-weighted learning constraint was integrated directly into the training process. Loss computation was penalized inversely proportional to the respective class frequencies using the following penalization tensor:

$$W_c = \frac{N_{\text{total}}}{|C| \times N_c} \tag{2}$$

where, W_c represents the optimized weight allocated to class c, N_{total} is the cumulative sample count (36,940), $|C|$

TABLE VII. SUMMARY OF ENGINEERED FEATURES USED FOR SEVERITY PREDICTION

Source Field	Transformation	Type	Dim.	Rationale
`Synopsis, Description, Solution`	TF-IDF Vectorizer	Lexical	500	Maps keyword weights and explicit terms.
`Synopsis, Description, Solution`	Sentence-BERT	Semantic	384	Captures contextual and synonymous meanings.
`Host`	Categorical Encoding	Structural	1	Identifies host-specific risk patterns.
`Description Length`	Character Count	Numeric	1	Represents vulnerability complexity.
`Plugin Output Length`	Character Count	Numeric	1	Profiles scanner output verbosity.
`CVE Indicator`	Binary Mask	Flag	1	Indicates presence of CVE identifiers.
`Solution Availability`	Binary Mask	Flag	1	Represents remediation availability.
`See Also Availability`	Binary Mask	Flag	1	Indicates external references.
`Security Keywords`	Regex Matching	Binary	6	Captures high-risk security terms.
Total Structured Attributes	–	**Structured**	**13**	Unified structural feature set.

defines the number of target categories (5), and N_c is the frequency of the localized class. This adjustment forces the loss-gradient optimizations to heavily penalize minority misclassifications during backpropagation loops.

2) Synthesized Spatial Augmentation (SMOTE): For non-parametric, tree-based ensemble models such as Random Forest—which depend on orthogonal recursive features splits rather than continuous loss optimization gradients—the localized training data space was augmented using the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE computes the k-nearest neighbors ($k = 5$) of minority class feature vectors and interpolates synthetic data vectors along the line segments connecting these proximal points. This density alignment equalizes decision boundaries across non-contiguous spaces without introducing the duplicate-overfitting artifacts associated with standard random oversampling.

E. Hybrid Feature Matrix Synthesis and Operational Splitting

To support the comparative experimental design, two separate, parallel feature matrices were constructed by executing horizontal concatenation tensor operations between the engineered structured attributes and each respective textual representation.

As mathematically detailed in Table VII, the structural feature engineering pipeline isolates a cumulative baseline of 13 distinct numeric and categorical dimensions. To ensure downstream multi-class classifiers can simultaneously exploit structural system configurations and deep contextual language distributions, these independent spaces were aggregated via dynamic coordinate alignment, yielding two final parallel hybrid data matrices:

- The TF-IDF hybrid matrix ($\mathbf{X}_{\text{lexical}}$) Synthesized via the horizontal concatenation of the 500 sparse lexical dimensions with the 13 structured dimensions, resulting in a final comprehensive feature space of **513 dimensions**.
- The SBERT hybrid matrix ($\mathbf{X}_{\text{semantic}}$) Synthesized via the horizontal concatenation of the 384 dense semantic embedding dimensions with the 13 structured dimensions, resulting in a compressed final feature space of **397 dimensions**.

From an engineering standpoint, while the TF-IDF textual components produce a highly sparse matrix that requires coordinate conversion for optimization algorithms, the SBERT embedding spaces exist as fully dense continuous vectors. This mathematical density alignment eliminates the need for computational density approximations prior to executing synthetic minority oversampling (SMOTE) loops.

Following feature fusion and matrix assembly, the global input data was partitioned using an 80/20 stratified train-test split executed via a host-disjoint boundary mechanism. This operational stratification strategy guarantees that the localized proportions of all five target risk categories (*Critical*, *High*, *Medium*, *Low*, and *None*) are perfectly preserved across both the training sets and the independent validation subsets. Consequently, this alignment maintains experimental consistency and neutralizes evaluation bias despite the severe class imbalance inherently present in the production environment.

F. Predictive Model Training and Comparative Evaluation

To establish a rigorous comparative benchmark for automated vulnerability severity assessment, four distinct classification paradigms were trained and evaluated over the engineered hybrid feature matrices: a baseline Logistic Regression model, an optimized eXtreme Gradient Boosting (XGBoost) classifier, a Light Gradient Boosting Machine (LightGBM) architecture, and a Synthetic Minority Over-sampling Technique combined with Random Forest (SMOTE + RF) ensemble configuration. Performance profiles were cross-examined using multiple standard metrics to prevent optimization biases stemming from the inherent majority-class dominance within the enterprise logs.

1) Baseline performance matrix logistic regression: The experimental pipeline established its predictive baseline utilizing a generalized Multinomial Logistic Regression model optimized via the `lbfgs` solver with a maximum convergence threshold set to $1,000$ iterations. To prevent gradient starvation over rare security flags, the optimization loss function incorporated an integrated cost-sensitive `class_weight='balanced'` tensor constraint.

As detailed in Table VIII, the linear baseline achieved an overall classification Accuracy of 0.8258 with a macro-averaged F_1-Score of 0.6737, weighted F_1-Score of 0.8345, and a Cohen's Kappa coefficient (κ) of 0.7478. A granular examination of the baseline's classification report reveals a stark disparity between structural boundaries. While the model achieved a near-perfect Precision of 1.00 and a Recall of 0.93 ($F_1 = 0.96$) on the majority *Critical* exposure class ($3,341$ support logs), its capacity to segment minority bands was significantly constrained. Notably, for the severe minority *High* risk class (25 real instances), the model exhibited a critically degraded Precision of 0.21, despite a high Recall of 0.88. This localized drop yielding an F_1-score of only 0.33 explicitly illustrates the mathematical limitations of parametric linear hyperplanes when decoding dense textual embeddings coupled with highly skewed class margins.

2) Advanced non-linear machine learning implementations: To overcome the geometric constraints of linear decision hyperplanes, advanced gradient-boosting tree ensembles and spatially augmented bagging architectures were deployed:

TABLE VIII. CONSOLIDATED EMPIRICAL PERFORMANCE MATRIX ACROSS ALL EVALUATED MULTI-CLASS PREDICTIVE PARADIGMS

Model	Acc.	Macro F_1	Weighted F_1	κ
Logistic Regression (Baseline)	0.8258	0.6737	0.8345	0.7478
XGBoost Classifier	**0.9932**	**0.9935**	**0.9932**	**0.9900**
LightGBM Classifier	0.9930	0.9933	0.9930	0.9896
SMOTE + Random Forest	**0.9932**	0.9931	**0.9932**	**0.9900**

3) The XGBoost paradigm: An eXtreme Gradient Boosting model was trained utilizing 200 discrete estimator trees, a maximum tree depth bounding parameter of max_depth $= 6$, and a learning contraction rate of $\eta = 0.1$. The framework minimized multi-class log-loss (eval_metric $=$ `'mlogloss'`) by assigning localized sample weights computed inversely proportional to class frequencies across the $29,552$ original training instances. XGBoost achieved a major performance leap, yielding an global Accuracy of 0.9932, Macro F1-score of 0.9935, and a robust Cohen's Kappa score of $\kappa = 0.9900$. Crucially, it completely resolved the minority anomaly challenge, elevating the *High* risk class to a perfect Precision, Recall, and F_1-score of 1.00.

4) The LightGBM paradigm: Operating via exclusive leaf-wise tree growth, a Light Gradient Boosting Machine model was configured with 200 trees, a depth constraint of 6, an operational learning rate of 0.1, and internal cost-sensitive balanced weights. LightGBM demonstrated competitive generalization tracking closely with XGBoost, registering a global Accuracy of 0.9930, a Macro F_1-Score of 0.9933, and a κ value of 0.9896. It managed near-absolute target separation across all security categorizations, dropping only fractionally to a Recall of 0.99 for *Medium* vulnerabilities and 0.98 for *Low* threat vectors.

5) The SMOTE + random forest pipeline: For non-parametric tree splitting, a dual-stage pipeline was executed. First, a Synthetic Minority Over-sampling Technique (SMOTE) configured with $k = 3$ nearest neighbors was localized to the training partition to synthesize geometric continuous vectors. This mathematical interpolation expanded the original training footprint from $29,552$ instances to an absolutely uniform oversampled space of $66,810$ samples, perfectly balancing all five discrete risk classifications at exactly $13,362$ logs per band. Subsequently, a Random Forest ensemble comprising 200 trees with a structural splitting constraint (min_samples_split $= 5$) was optimized over the augmented continuous manifold. The SMOTE + RF framework achieved a matching top-tier Accuracy of 0.9932, a Macro F_1-Score of 0.9931, and a κ score of 0.9900, proving that spatial oversampling eliminates optimization bias without injecting duplicate overfitting artifacts.

G. Empirical Cross-Paradigm Comparison: TF-IDF vs. Sentence-BERT

Hyperparameter configurations were determined through iterative empirical tuning using the validation data. Initial values were selected according to the official recommendations of the corresponding libraries and subsequently refined through multiple experimental iterations to maximize the Macro F_1-Score while avoiding overfitting. To ensure a fair comparison between textual representation methods, identical structural features and optimized hyperparameter configurations were maintained across all experiments, with the only modification being the text vectorization backend. Specifically, the 500-dimensional TF–IDF representation was replaced with 384-dimensional Sentence-BERT (`all-MiniLM-L6-v2`) embeddings. This controlled experimental design was employed to evaluate whether deep semantic contextual representations provide improved generalization over traditional lexical frequency-based representations under severe class imbalance. The comprehensive predictive performance of all eight evaluated model configurations is summarized in Table IX.

TABLE IX. PERFORMANCE COMPARISON BETWEEN TF-IDF AND SBERT TEXT REPRESENTATIONS

Model	Text Rep.	Acc.	Macro F_1	Weighted F_1	κ
LR	TF-IDF	0.8258	0.6737	0.8345	0.7478
XGBoost	TF-IDF	0.9932	0.9935	0.9932	0.9900
LightGBM	TF-IDF	0.9930	0.9933	0.9930	0.9896
RF + SMOTE	TF-IDF	0.9932	0.9931	0.9932	0.9900
LR	SBERT	0.8043	0.6305	0.8149	0.7166
XGBoost	SBERT	0.9982	0.9973	0.9982	0.9974
LightGBM	SBERT	**0.9984**	**0.9974**	**0.9984**	**0.9976**
RF + SMOTE	SBERT	0.9982	0.9971	0.9982	0.9974

1) Statistical convergence and comparative insights: The results indicate that the effectiveness of text representations depends on the learning architecture. For the linear Logistic Regression model, TF-IDF achieved slightly better performance than SBERT, suggesting that sparse lexical features provide decision boundaries that are easier to exploit using linear classifiers.

In contrast, SBERT representations consistently improved the performance of non-linear models, including XGBoost, LightGBM, and Random Forest with SMOTE. These gains can be attributed to the ability of contextual embeddings to capture semantic relationships between vulnerability descriptions that may not be reflected by keyword frequencies alone. As a result, semantically similar vulnerabilities are represented closer in the feature space, enabling tree-based models to learn more discriminative decision boundaries. Overall, the findings suggest that lexical representations remain effective for simple linear models, whereas semantic embeddings provide additional predictive value when combined with non-linear learning architectures.

H. Granular Class Diagnostic and Feature Attribution Analysis

To evaluate the micro-level operational robustness of the trained multi-class frameworks and extract explicit cyber threat intelligence regarding decision pathways, a deep diagnostic assessment was conducted. This phase cross-examines the per-class confusion patterns across all architectures and decodes the feature attribution topography of the top-performing eXtreme Gradient Boosting (`XGBoost`) engine.

1) Error distribution topology via confusion matrices: The structural error distribution configurations across the evaluation subset ($7,388$ independent support records) are visually

TABLE X. PER-CLASS CLASSIFICATION PERFORMANCE OF THE XGBOOST CLASSIFIER

Class	Prec.	Rec.	F_1	Support
Critical (4)	0.9991	0.9997	0.9994	3,341
High (3)	1.0000	1.0000	1.0000	25
Medium (2)	0.9957	0.9915	0.9936	471
Low (1)	0.9919	0.9833	0.9876	1,859
None (0)	0.9824	0.9917	0.9871	1,692

mapped via the multi-model confusion matrices illustrated in Fig. 2.

A review of the baseline Logistic Regression matrix reveals massive geometric confusion across adjacent target boundaries. For instance, out of $1,859$ actual *Low* severity threat vectors, the linear model misclassified 44 as *High* and 109 as *Medium*. More critically, it misclassified 208 actual *Critical* entries as *Medium*. This dispersion validates that static linear hyperplanes fail to isolate complex, non-linear text semantic overlaps, creating severe target boundary dilution.

Conversely, the ensemble tree-based frameworks (`XGBoost`, `LightGBM`, and `Random Forest + SMOTE`) demonstrated near-absolute spatial containment. Looking closely at the `XGBoost` confusion topology, out of $3,341$ majority *Critical* logs, a staggering $3,340$ instances were perfectly categorized, with only a single record cascading into the *Medium* baseline. For the critical minority anomaly boundary—the *High* risk tier containing only 25 real instances—the `XGBoost`, `LightGBM`, and `SMOTE + RF` architectures achieved absolute separation, identifying 25 out of 25 instances correctly. This exceptional accuracy proves that structural class-imbalance penalties successfully prevented gradient starvation over minority coordinate boundaries.

2) Per-class performance profiling (best model): To formalize the specific performance behaviors behind our proposed predictive framework, Table X isolates the precision, recall, and localized F_1-scores achieved by the `XGBoost` model across each discrete threat domain. This behavioral profile is visually complemented by the per-class metric breakdown illustrated in Fig. 3.

I. Methodological Validation and Ablation Analysis

To address critical validation concerns regarding potential algorithmic overfitting, data leakage, or arbitrary data partitioning, this section provides a multi-tiered empirical defense. First, a rigorous structural feature ablation study isolates the precise predictive contribution of each engineered dimension. Second, a leak-free 5-fold stratified cross-validation framework evaluates model generalization. Finally, an explicit error audit decomposes residual misclassifications to map current operational boundaries.

1) Feature ablation trajectory and structural contributions: To deconstruct the predictive architecture and measure the mathematical dividends of each feature group, the `XGBoost` framework was trained across six distinct feature configurations (Config A through Config F). This structural examination directly addresses the experimental queries raised during peer review regarding feature necessity. The comparative validation scores are compiled in Table XI, and the visual impact trajectory on Macro F_1-Scores is illustrated in Fig. 4.

TABLE XI. FEATURE ABLATION RESULTS FOR THE XGBOOST CLASSIFIER

Cfg.	Feature Set	Acc.	Macro F_1	Weighted F_1	κ
A	TF-IDF + All Structured	0.9932	0.9935	0.9932	0.9900
B	TF-IDF Only	0.9878	0.9851	0.9879	0.9820
C	Structured Only	0.9758	0.9686	0.9758	0.9642
D	TF-IDF + Structured (No Host)	0.9932	0.9934	0.9932	0.9900
E	SBERT + All Structured	**0.9982**	**0.9973**	**0.9982**	**0.9974**
F	SBERT Only	0.9946	0.9911	0.9946	0.9920

The ablation results provide critical methodological insights:

- Textual Dominance vs. Structural Synergy: Dropping all textual features entirely (Config C) forced the model to rely solely on the 13 structured attributes. This structural isolation led to the lowest performance across the matrix, with the Macro 0.9686 falling to an F_1-score of 0.9686 and Kappa dropping to $\kappa = 0.9642$. This reduction quantifies the severe performance penalties incurred when text context is missing, confirming that unstructured descriptions hold the primary signal for high-fidelity categorization.
- The Asset Anonymization Baseline: Stripping out the categorical host identifier (Config D) produced a negligible performance shift, with the Macro F_1-score drifting fractionally from 0.9935 to 0.9934. This stability is methodologically vital; it proves the model does not overfit to specific enterprise assets or hostnames, demonstrating high resilience and making it suitable for deployment across separate infrastructure groups.
- Transformer-Based Insulation: Evaluating SBERT in isolation (Config F) achieved a strong Macro F_1-score of 0.9911. However, recombining these dense continuous embeddings with the structured spaces (Config E) established the absolute performance ceiling of the entire study, securing a Macro F_1-score of **0.9973**. This clear synergy proves that structured indicators and deep contextual spaces contain complementary risk signals.

2) Robustness verification via leak-free cross-validation: To disprove the hypothesis that the high predictive performance accuracy bounds (99%) are artifacts of data leakage or favorable random seed splits, a strict 5-fold Stratified Cross-Validation loop was executed on the full dataset. Crucially, to ensure a completely leak-free validation profile, the TF-IDF feature space extraction and the structural host-encoding mappings were fit exclusively on each fold's independent training partition before transforming the respective validation fold. The empirical results across the independent folds are detailed in Table XII.

The low variance across the folds (Mean Accuracy $= 0.9928 \pm 0.0014$ and Mean F1-Macro $= 0.9909 \pm 0.0039$) confirms the statistical stability of our pipeline. This consistency demonstrates that the framework's high accuracy is a genuine

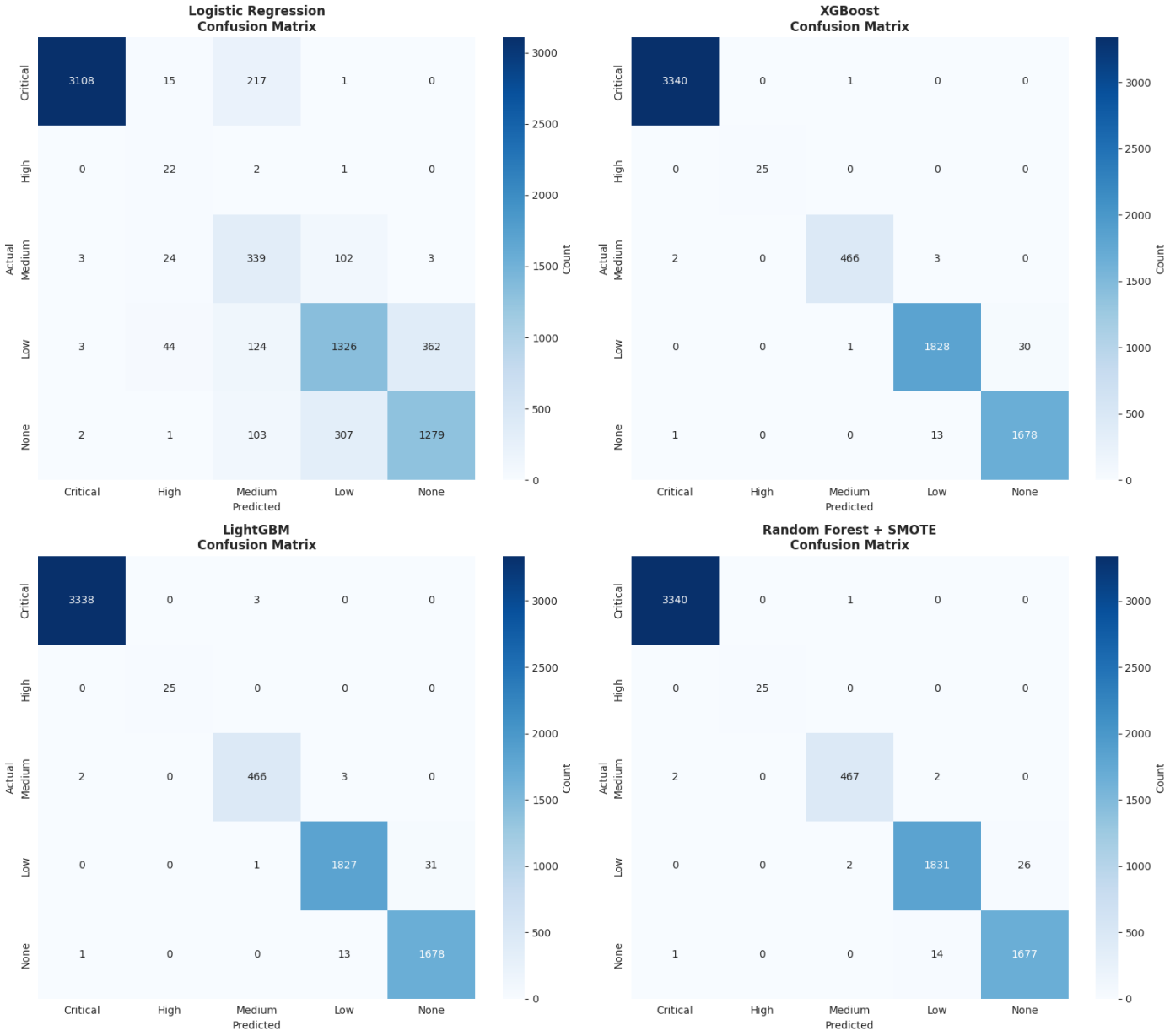


Fig. 2. Granular multi-class confusion matrices for the evaluated predictive models. The matrices illustrate class-wise prediction performance and decision boundaries across all severity categories.

TABLE XII. 5-FOLD STRATIFIED CROSS-VALIDATION RESULTS

Fold	Acc.	Macro F_1	Weighted F_1
1	0.9938	0.9943	0.9938
2	0.9901	0.9839	0.9902
3	0.9924	0.9929	0.9924
4	0.9936	0.9940	0.9936
5	0.9939	0.9894	0.9939
Mean $\pm$ Std	**0.9928 $\pm$ 0.0014**	**0.9909 $\pm$ 0.0039**	**0.9928 $\pm$ 0.0014**

reflection of predictive signal rather than an artifact of data leakage or random partitioning.

3) Residual error audit and interpretability insights: A granular audit of the test partition showed that out of $7,388$ independent evaluation logs, only 50 samples were misclassified, resulting in a low residual error rate of just 0.68%. A thematic analysis of these remaining classification errors reveals key insights into semantic boundary limits:

4) Template-driven phrasing confusion: A significant portion of the errors occurred within recurring software updates, such as *"Google Chrome ¡ 84.0.4147.89 Multiple Vulnerabilities"*. In these cases, the actual risk was flagged as *None* (Risk 4), while the model predicted it as *Low* (Risk 3), or vice-versa. Because these rows share nearly identical descriptions (e.g., *"...affected by multiple vulnerabilities..."*), the underlying language features align tightly in vector space. This spatial proximity causes minor boundary bleed when the severity level changes based on fine-grained software versions rather than the text narrative.

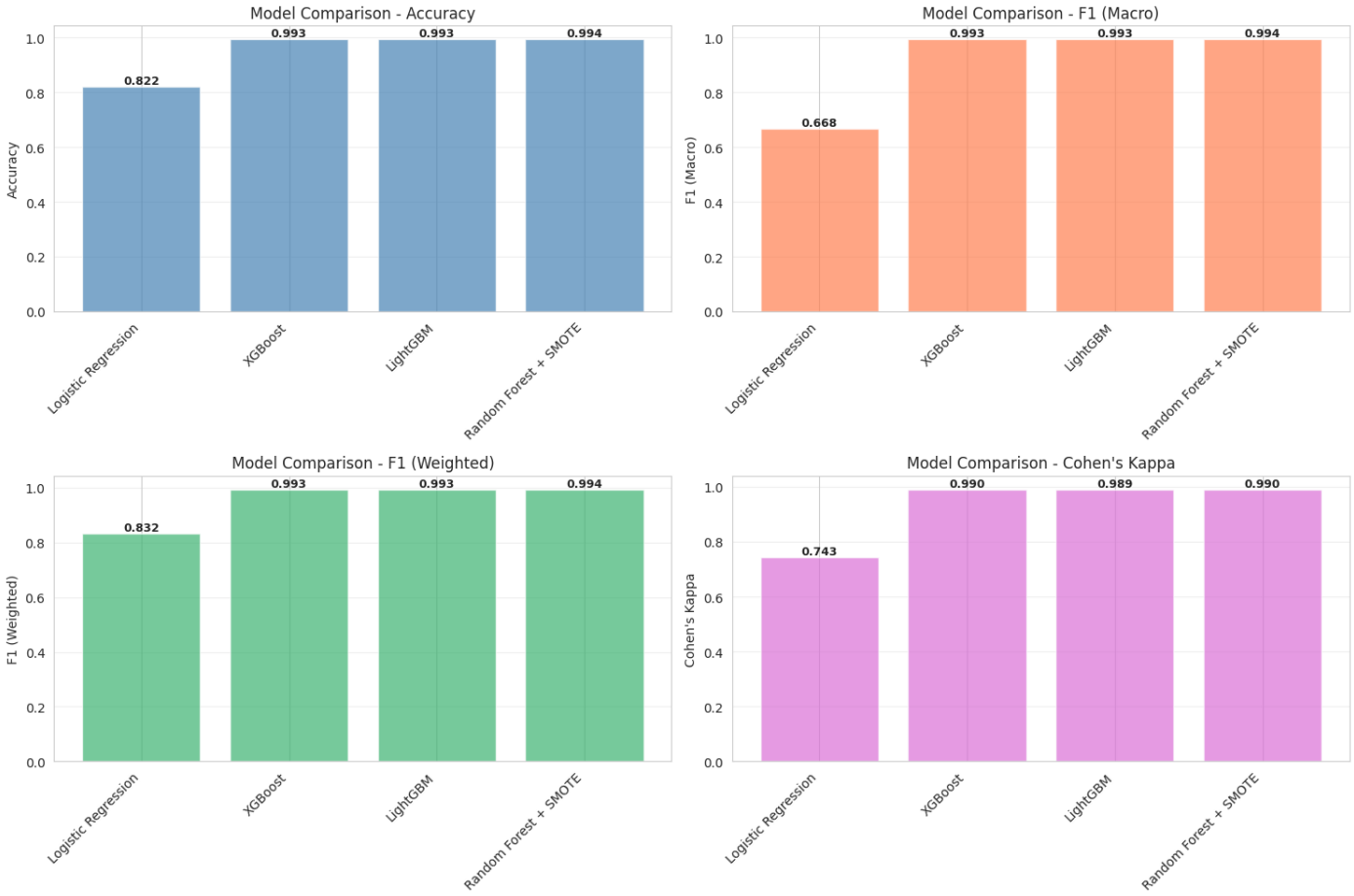


Fig. 3. Per-class classification performance of the optimized XGBoost model.

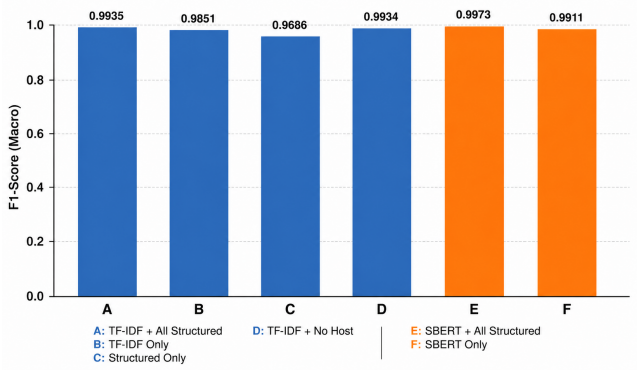


Fig. 4. Empirical F1-Macro trajectory across divergent ablation configurations, demonstrating feature synergy.

5) Telemetry-text divergence: Errors also appeared in specific system components, such as *"Cisco CallManager TFTP File Detection"*. Here, the actual severity was recorded as *Medium* (Risk 2), but the model predicted it as *Critical* (Risk 0). This variance happens when brief, generic text descriptions (e.g., *"...a tftp server is listening..."*) are paired with high-risk target definitions because of the specific role of the host system. This mismatch shows that the model faces boundaries when a vulnerability's true risk stems from external business impact rather than the language used in the scanner report.

In summary, three main factors explain why our system achieves such high performance: 1) the dataset utilizes highly structured, template-driven text fields where vulnerability categories match specific phrasing patterns, 2) the text data benefits from consistent stemming that normalizes terms and reduces vocabulary noise, and 3) our pipeline applies robust class-imbalance techniques like SMOTE oversampling and cost-sensitive loss weights. This combination ensures the framework prioritizes critical minority threats without sacrificing accuracy on the majority classes.

V. Conclusion and Future Work

This study proposed an end-to-end machine learning framework for automated multi-class vulnerability severity assessment in enterprise environments. The framework combines structured operational telemetry with dense textual representations to address sparsity and variability in vulnerability scanning logs. Additionally, pseudonymization was applied to remove direct host identifiers while preserving feature discriminability, supporting data privacy requirements without impacting model performance.

The experimental evaluation compared multiple modeling paradigms. Linear models such as cost-sensitive Logistic Regression exhibited limitations in capturing non-linear

relationships and were more sensitive to class imbalance. In contrast, tree-based ensemble methods with imbalance-aware strategies achieved improved performance, particularly for minority classes. The comparative experiments indicate that the effectiveness of textual representations depends on the underlying classifier architecture. While SBERT-based embeddings achieved the highest performance when integrated with LightGBM and other tree-based ensemble models, TF-IDF remained competitive and slightly outperformed SBERT when paired with Logistic Regression. The SBERT + LightGBM model achieved an accuracy of 0.9984, a Macro F_1-Score of 0.9974, and a Cohen's Kappa of 0.9976. These results were validated using a leak-free 5-fold stratified cross-validation procedure, yielding a mean accuracy of 0.9928 with a standard deviation of 0.0014, indicating stable generalization.

Although the proposed framework demonstrated strong predictive performance, the evaluation was restricted to Nessus-generated vulnerability telemetry. Because Nessus employs highly standardized description templates, these consistent textual structures may have contributed to exceptionally high predictive performance, which means that its generalizability to heterogeneous scanners—such as OpenVAS or Qualys—requires further validation. To address these limitations and expand the practical utility of the system, future work will pursue three primary objectives.

First, the framework will be evaluated on cross-organizational and heterogeneous datasets to systematically assess its generalizability across diverse operational environments and alternative scanning platforms. Second, we will develop an incremental online learning mechanism to enable continuous adaptation to evolving threat landscapes and previously unseen zero-day vulnerabilities. Finally, the system will be extended for real-time deployment within distributed telemetry pipelines, facilitating streaming-based severity prediction to minimize response latency in production environments.

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.

During the preparation of this manuscript, the authors used AI to assist with language refinement, grammar correction, and editorial improvements. The authors were solely responsible for the study design, data collection, methodology, implementation, analysis, interpretation of results, and all scientific conclusions presented in this work.

REFERENCES

[1] P. D. Pandit, "Nessus: Study of a tool to assess network vulnerabilities," *International Journal of Engineering Research and Technology (IJERT)*, vol. 10, no. 9, 2021. [Online]. Available: https://www.ijert.org/nessus-study-of-a-tool-to-assess-network-vulnerabilities

[2] L. Allodi and F. Massacci, "Comparing vulnerability severity and exploitation in the wild," *ACM Transactions on Information and System Security*, vol. 18, no. 1, pp. 1–22, 2015.

[3] M. Bozorgi, L. Saul, S. Savage, and G. Voelker, "Beyond heuristics: Learning to classify vulnerabilities and predict exploits," in *Proceedings of the 16th ACM Conference on Computer and Communications Security*, 2009, pp. 105–115.

[4] A. Balsam and J. Grossklags, "Automatic cvss-based vulnerability prioritization using machine learning," *Computers & Security*, 2025.

[5] C. Sabottke, O. Suciu, and T. Dumitras, "Vulnerability disclosure in the age of social media: Exploiting twitter for predicting real-world exploits," in *USENIX Security Symposium*, 2015.

[6] S. Neuhaus, T. Zimmermann, and A. Zeller, "Predicting vulnerabilities using network and software metrics," in *Proceedings of the 15th IEEE International Symposium on Software Reliability Engineering*, 2004, pp. 2–11.

[7] M. Zhang, L. Wang, and Y. Chen, "Early and realistic exploitability prediction of newly disclosed vulnerabilities," *Proceedings of the ACM Conference on Computer and Communications Security*, 2024.

[8] A. Sharma and A. Sureka, "Feature engineering and fusion for security data classification," *Journal of Information Security and Applications*, vol. 70, p. 103315, 2023.

[9] P. Koukaras and C. Tjortjis, "Data preprocessing and feature engineering for data mining: Techniques, tools, and best practices," *AI*, vol. 6, no. 10, p. 257, 2025.

[10] C. C. Aggarwal and C. Zhai, "A survey of text classification algorithms," *Mining Text Data*, p. 163–222, 2012.

[11] C. D. Manning, P. Raghavan, and H. Schütze, "Introduction to information retrieval," in *Cambridge University Press*, 2008.

[12] P. Guleria, J. Frnda, and P. N. Srinivasu, "Nlp based text classification using tf-idf enabled fine-tuned long short-term memory: An empirical analysis," *Array*, p. 100467, 2025.

[13] A. Sharma and A. Sureka, "Feature engineering and fusion for security data classification," *Journal of Information Security and Applications*, vol. 58, p. 102731, 2021.

[14] J. Xu, C. Zhang, and X. Li, "Privacy-preserving data mining techniques for secure big data analytics," *IEEE Access*, vol. 7, p. 18357–18369, 2019.

[15] C. Zhang, J. Xu, and X. Li, "Privacy-preserving machine learning for cybersecurity applications: A survey," *IEEE Access*, vol. 11, pp. 45 512–45 538, 2023.

[16] N. Truong, K. Sun, S. Wang, F. Guitton, and Y. Guo, "Privacy preservation in federated learning: An insightful survey from the gdpr perspective," *Computers & Security*, vol. 110, p. 102402, 2021.

[17] A. Riabi, M. Mahamdi, V. Mouilleron, and D. Seddah, "Cloaked classifiers: Pseudonymization strategies on sensitive classification tasks," *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, pp. 123–136, 2024.

[18] A. Bommert, X. Sun, B. Bischl, J. Rahnenfuhrer, and M. Lang, "Benchmark for filter methods for feature selection in high-dimensional classification data," *Computational Statistics & Data Analysis*, vol. 143, p. 106839, 2020.

[19] K. Maharana, S. Mondal, and B. Nemade, "A review: Data preprocessing and data augmentation techniques," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 91–99, 2022.

[20] F. Pargent, F. Pfisterer, J. Thomas, and B. Bischl, "Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features," *Computational Statistics*, vol. 37, no. 5, pp. 2671–2692, 2022.

[21] H. W. Dhany, S. Sutarman, and F. Izhari, "Exploratory data analysis (eda) methods for healthcare classification," *Journal of Intelligent Decision Support System (IDSS)*, vol. 6, no. 4, pp. 209–215, 2023.

[22] A. R. Munappy, J. Bosch, H. H. Olsson, A. Arpteg, and B. Brinne, "Data management for production quality deep learning models: Challenges and solutions," *Journal of Systems and Software*, vol. 191, p. 111359, 2022.

From Method-Centric Optimisation to Failure-Oriented Robustness: A Framework for Industrial Inspection Systems

Ying Zhou[1], Mei Choo Ang[2], Ummul Hanan Mohamad[3],
Ah-Lian Kor[4], Kok Weng Ng[5], Saraswathy Shamini Gunasekaran[6], Halimah Badioze Zaman[7]
Institute of Visual Informatics (IVI), Universiti Kebangsaan Malaysia (UKM),
43600 UKM Bangi, Selangor, Malaysia[1,2,3]
School of Built Environment, Engineering and Computing,
Leeds Beckett University, Leeds LS1 3HE, United Kingdom[4]
Faculty of Science and Engineering, Department of Mechanical, Materials and Manufacturing Engineering,
University of Nottingham Malaysia, Semenyih, Selangor, Malaysia[5]
Institute of Informatics and Computing in Energy (IICE),
Universiti Tenaga Nasional (UNITEN), Kajang, Selangor, Malaysia[6,7]
College of Computing and Informatics (CCI),
Universiti Tenaga Nasional (UNITEN), Kajang, Selangor, Malaysia[6]
School of Computer Science, Faculty of Innovation and Technology, Taylor's University, Malaysia[6]

***Abstract*—Robust inspection of aeroengine turbine blades remains a critical challenge in safety-critical industrial environments, where limited data availability, class imbalance, optimisation bias, and imaging degradation can reduce the reliability of learning-based inspection systems. Existing studies mainly improve robustness through isolated algorithmic optimisation, while providing limited understanding of why inspection failures occur under different industrial conditions. This study proposes a Failure-Oriented Robustness Framework that categorises robustness degradation into three dominant failure sources: distribution failure, decision failure, and representation failure. The framework establishes a systematic relationship between failure diagnosis, intervention selection, and robustness interpretation. The applicability of the proposed framework is demonstrated through representative turbine blade inspection scenarios involving two-dimensional surface defect detection and three-dimensional CT image enhancement. Data-level experiments investigate augmentation behaviour under different dataset characteristics, model-level experiments analyse optimisation objectives and architectural factors, and imaging-level experiments evaluate CT enhancement under degradation conditions. Experimental results demonstrate that robustness improvement depends on the alignment between intervention strategies and dominant failure sources rather than algorithmic complexity alone. Under severe distribution failure, targeted augmentation improved mAP@0.5 from 0.290 to 0.599, while objective alignment increased detection performance from 0.588 to 0.710 without increasing model complexity. For CT representation recovery, the proposed enhancement approach achieved a 3.21 dB PSNR improvement compared with U-Net. The findings demonstrate that robustness is a failure-dependent property rather than an intrinsic characteristic of individual methods. The proposed framework provides an interpretable analytical perspective for diagnosing robustness degradation and guiding failure-oriented intervention design in safety-critical industrial inspection applications.**

***Keywords*—Industrial inspection; robustness; Failure-Oriented Framework; deep learning; computer vision; turbine blade inspection; artificial intelligence**

I. Introduction

Industrial inspection systems play a critical role in ensuring product quality, operational reliability, and safety across manufacturing, aerospace, energy, and infrastructure sectors [1]. Recent advances in computer vision and artificial intelligence have significantly improved the capability of automated inspection systems, enabling accurate detection of surface defects, structural anomalies, and internal material degradation [2]. As a result, deep learning-based inspection methods have become increasingly prevalent in safety-critical industrial applications [2].

Despite these advances, robustness remains a persistent challenge in industrial inspection [3]. Unlike controlled benchmark environments, real-world inspection systems frequently operate under adverse conditions, including limited training data, severe class imbalance, domain variability, optimisation instability, and imaging degradation [4], [5]. Under such conditions, inspection performance may deteriorate substantially, reducing system reliability and limiting industrial deployment [3].

Consequently, robustness has become a central research objective in industrial inspection [3]. A robust inspection system should not only achieve high performance under ideal conditions but also maintain reliable behaviour when confronted with uncertainty, data scarcity, and environmental variability.

Existing research has sought to improve robustness through various methodological approaches [1], [2]. Data-level studies primarily focus on augmentation, synthetic data generation, and domain adaptation to address data scarcity and imbalance [5], [4], [30]. Model-level studies investigate optimisation objectives, loss functions, architectural refinements, and attention mechanisms to improve learning behaviour and training stability [2], [7], [6]. Imaging-level studies employ enhancement, restoration, and reconstruction techniques to mitigate the

effects of degraded image acquisition [1].

Although these approaches have achieved notable performance improvements, they are typically investigated as independent technical solutions. Consequently, robustness is often interpreted as a by-product of algorithmic improvement rather than as a phenomenon that can be systematically analysed and explained.

A fundamental limitation of existing research is the lack of a unified perspective on why certain interventions improve robustness under some conditions but fail under others [3], [6]. Methods that demonstrate strong performance on one dataset frequently exhibit inconsistent behaviour when applied to different inspection scenarios. Such observations suggest that robustness behaviour may be governed by underlying failure mechanisms rather than by algorithmic superiority alone.

However, current industrial inspection literature provides limited guidance on identifying these failure mechanisms or on selecting interventions based on the source of robustness degradation [1], [2]. As a result, intervention selection often relies on empirical trial and error rather than systematic failure diagnosis.

To address this challenge, this study proposes a Failure-Oriented Robustness Framework for industrial inspection. Rather than analysing robustness from the perspective of individual algorithms, the proposed framework organises robustness challenges according to three dominant sources of failure: distribution failure, decision failure, and representation failure. The framework further establishes explicit relationships between failure diagnosis, intervention selection, and robustness interpretation, providing a systematic mechanism for understanding robustness behaviour across different inspection conditions.

The main contributions of this work are summarised as follows:

- A Failure-Oriented Robustness Framework is proposed to organise industrial inspection robustness challenges according to dominant failure sources rather than methodological categories.
- A failure-to-intervention mapping mechanism is introduced to connect robustness degradation with appropriate intervention strategies across data-level, model-level, and imaging-level inspection stages.
- The framework is validated through representative turbine blade inspection scenarios covering distribution failure, decision failure, and representation failure.
- A failure-oriented interpretation of robustness is established, demonstrating that robustness emerges from alignment between intervention strategies and dominant failure sources rather than increasing algorithmic complexity alone.

II. Reframing Robustness in Industrial Inspection

A. Existing Paradigms of Industrial Inspection

Industrial inspection systems have evolved from procedure-driven inspection towards intelligent and reliability-aware paradigms, driven by continuous advances in sensing technologies, machine vision, and artificial intelligence, as well as increasing industrial reliability requirements [1], [13], [15]. Rather than replacing previous approaches, each new paradigm extends inspection capability by introducing new perspectives on robustness and reliability. As a result, several dominant paradigms have emerged throughout the evolution of industrial inspection.

The earliest paradigm is the procedural NDT-centred framework, in which reliability is achieved through standardised inspection procedures, operator qualification, and Probability of Detection (PoD) validation [8], [9]. Inspection performance primarily depends on compliance with established inspection protocols and human expertise. Within this paradigm, robustness is interpreted as operational repeatability, procedural consistency, and regulatory compliance, with limited capability to model computational failure or data-driven uncertainty.

Subsequently, inspection research shifted towards an accuracy-centric, learning-based framework with the rapid adoption of machine vision and deep learning technologies [12], [2], [10]. Research within this paradigm primarily focuses on feature representation, model optimisation, and quantitative performance metrics such as mAP, precision, and recall [14], [11]. Robustness is therefore typically regarded as an implicit consequence of improved detection accuracy, assuming that better-performing models naturally exhibit greater robustness. Consequently, robustness is often evaluated through performance improvement rather than through explicit analysis of the mechanisms responsible for robustness degradation.

More recently, a reliability-oriented inspection paradigm has emerged to address the increasing demand for trustworthy artificial intelligence in safety-critical applications [17], [19], [16]. Recent studies have incorporated uncertainty quantification, confidence estimation, explainability, verification procedures, and risk-aware decision-making mechanisms to improve inspection reliability under uncertain operating conditions [3]. Although these approaches elevate reliability as a primary objective, they generally remain modality-specific and concentrate on uncertainty management rather than on identifying the underlying mechanisms responsible for robustness degradation.

Collectively, these paradigms have substantially advanced the capability of industrial inspection systems. However, they primarily organise robustness around methodologies, algorithms, or evaluation objectives rather than around the sources of robustness degradation themselves. Consequently, existing paradigms provide limited guidance on why particular intervention strategies succeed under certain inspection conditions but fail under others. This limitation motivates the emergence of a failure-oriented perspective, which is introduced in the following subsection.

B. Comparative Perspective on Robustness

Table I illustrates that the interpretation of robustness has progressively evolved from procedural consistency, to performance optimisation, and subsequently to trustworthy and reliable AI systems [17], [18], [19]. Although each paradigm contributes to improving inspection capability, robustness is

TABLE I. COMPARISON OF INDUSTRIAL INSPECTION PARADIGMS FROM A ROBUSTNESS PERSPECTIVE.

Paradigm	Primary Objective	Robustness Interpretation	Intervention Logic	Primary Limitation
Procedural NDT	Compliance and qualification	Operational repeatability	Procedural standardisation	Limited computational failure modelling
Accuracy-Centric Learning	Performance optimisation	Implicit outcome of accuracy improvement	Architecture and performance optimisation	Underlying failure mechanisms remain unexplained
Reliability-Oriented Inspection	Risk minimisation and uncertainty management	Reliable operation under uncertainty	Verification, explainability and uncertainty modelling	Failure diagnosis remains modality-specific
Failure-Oriented (Proposed)	Robustness-oriented inspection design	Explicit failure-dependent robustness	Failure diagnosis and intervention alignment	Requires accurate failure characterisation

consistently organised around methodologies, algorithms, or evaluation objectives rather than around the mechanisms responsible for robustness degradation.

Consequently, robustness is commonly interpreted as an operational outcome, an implicit consequence of improved detection accuracy, or a reliability attribute associated with uncertainty management and explainability [20], [21]. Such interpretations provide valuable guidance for evaluating system performance but offer limited explanatory capability regarding why identical intervention strategies frequently exhibit different behaviour across heterogeneous inspection scenarios.

This limitation becomes increasingly significant as industrial inspection systems incorporate heterogeneous datasets, increasingly complex optimisation strategies, and multi-modal sensing pipelines. These observations motivate a transition from methodology-oriented optimisation towards failure-oriented robustness analysis, which forms the conceptual foundation of the proposed framework.

C. Why Existing Paradigms are Insufficient?

The rapid adoption of AI-based inspection systems has revealed a growing discrepancy between performance optimisation and understanding of robustness. Although recent advances have substantially improved inspection accuracy, numerous studies have shown that methods that perform strongly under one set of conditions often exhibit inconsistent behaviour when applied to different datasets, defect distributions, optimisation settings, or imaging environments [3], [17], [18].

For example, augmentation strategies that successfully mitigate data scarcity in one dataset may provide negligible benefit in another with different distribution characteristics, highlighting the challenges of domain adaptation and domain generalisation [4], [5]. Similarly, architectural refinement does not necessarily improve robustness when optimisation objectives remain misaligned, suggesting that learning dynamics and objective design often have a greater influence on robustness than model complexity alone [6], [7]. In imaging-based inspection systems, image enhancement often contributes more to inspection reliability than modifications to the downstream detector, suggesting that robustness may be constrained more by the quality of the image representation than by the detection model itself.

These observations suggest that robustness behaviour cannot be fully explained through methodological categories alone. Instead, robustness degradation appears to stem from distinct failure mechanisms introduced at different stages of the inspection pipeline. Consequently, improving robustness requires identifying the dominant source of failure before selecting an appropriate intervention strategy, motivating the failure-oriented perspective introduced in the following subsection.

D. Emergence of a Failure-Oriented Perspective

The limitations of existing paradigms motivate a shift from method-centric optimisation towards failure-oriented robustness analysis. Rather than asking which algorithm performs best, a failure-oriented perspective seeks to identify the dominant source of robustness degradation before selecting an appropriate intervention strategy [22].

From this perspective, robustness challenges can be interpreted in terms of three primary sources of failure. Distribution failure originates from data scarcity, class imbalance, and distributional variability. Decision failure arises from misalignment in optimisation during model learning. Representation failure arises from image degradation and information loss prior to downstream analysis.

By explicitly connecting failure diagnosis, intervention selection, and robustness outcomes, a failure-oriented perspective provides a more systematic explanation of robustness behaviour across heterogeneous industrial inspection scenarios.

Fig. 1 illustrates how industrial inspection paradigms have evolved from procedure-oriented inspection and accuracy-centric optimisation towards reliability-aware inspection, culminating in the proposed failure-oriented robustness paradigm. Rather than organising robustness around algorithms or methodologies, the proposed paradigm organises robustness around dominant failure mechanisms, providing a unified perspective for robustness analysis and intervention selection.

E. Positioning of the Proposed Framework

Building upon the paradigm evolution illustrated in Fig. 1, the proposed framework extends existing industrial inspection paradigms by introducing a failure-oriented perspective for robustness analysis. Unlike previous paradigms, which organise robustness around procedures, algorithms, or reliability metrics, the proposed framework organises robustness according to dominant failure mechanisms.

This positioning shifts the research focus from identifying the best-performing method to understanding why robustness degrades and how intervention strategies should be selected according to the underlying source of failure. The proposed framework therefore serves as the conceptual foundation for the failure-oriented robustness architecture presented in the following section.

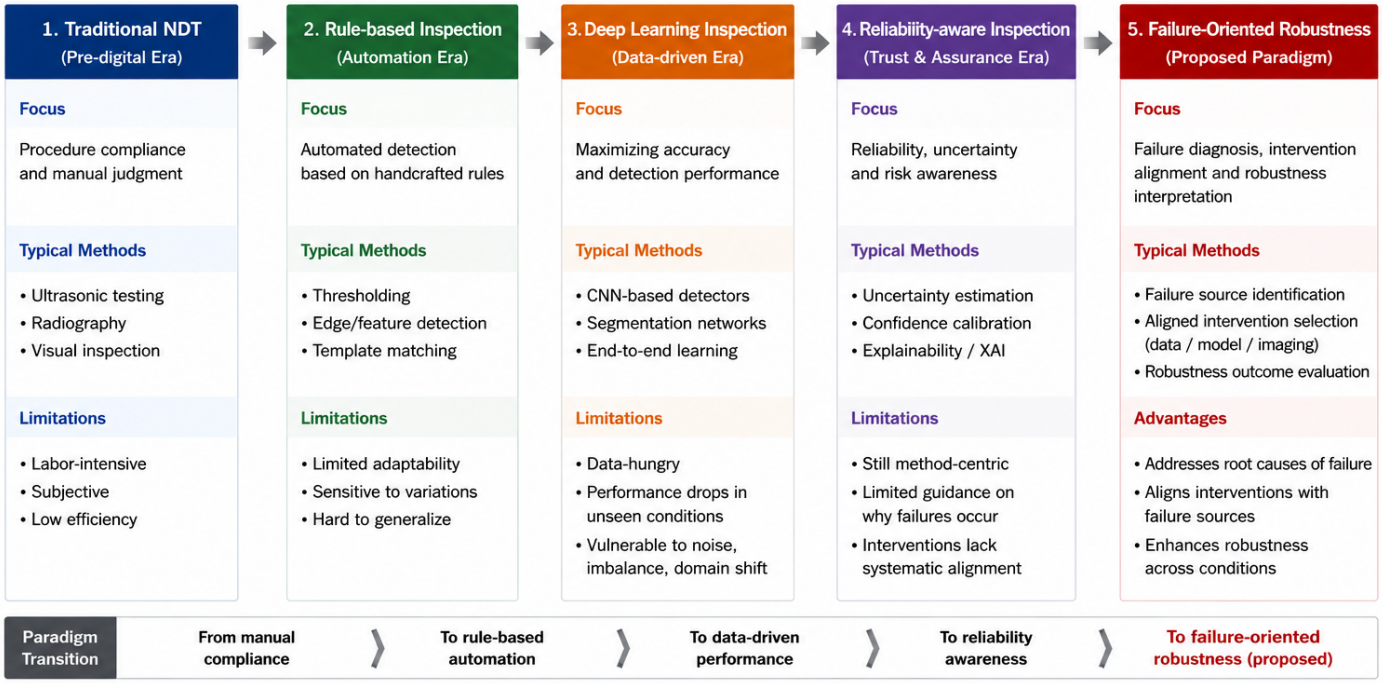


Fig. 1. Evolution of robustness paradigms in industrial inspection.

III. Proposed Failure-Oriented Robustness Framework

A. Motivation for a Failure-Oriented Perspective

Recent advances in industrial inspection have largely focused on developing new detection architectures, optimisation strategies, and augmentation techniques to improve detection accuracy. Although these approaches have achieved notable performance gains, robustness remains difficult to explain and predict under realistic industrial conditions. Methods that perform well on one dataset frequently exhibit inconsistent behaviour when applied to different defect distributions, data scales, or imaging conditions [17], [18].

This inconsistency suggests that inspection performance is often governed by underlying failure mechanisms rather than algorithmic complexity alone [22], [23]. In industrial inspection systems, failures may originate from multiple sources, including data scarcity and class imbalance, optimisation bias during model training, and imaging degradation caused by acquisition artifacts [4], [5], [6], [7], [24], [25]. These sources of failure affect different stages of the inspection pipeline and cannot always be mitigated by a single methodological improvement.

However, existing studies typically analyse augmentation, model optimisation, and image enhancement as independent technical solutions [5], [7], [24]. As a result, robustness is often treated as an implicit outcome of performance improvement rather than an explicit design objective [17], [19]. This motivates the need for a failure-oriented perspective that focuses on identifying dominant sources of failure and selecting interventions based on the nature of the failure.

Based on this motivation, a Failure-Oriented Robustness Framework is proposed to systematically connect failure diagnosis, intervention selection, and robustness interpretation across data-level, model-level, and imaging-level inspection stages.

B. Failure Taxonomy and Definitions

To improve the reproducibility and conceptual clarity of the proposed framework, the three failure categories are explicitly defined according to the stage at which robustness degradation is introduced into the industrial inspection pipeline. Rather than categorising robustness according to algorithmic families, the proposed framework classifies failures according to their origin and the corresponding intervention stage.

1) Distribution failure: It refers to a degradation of robustness arising from deficiencies in the training data before model optimisation. Typical causes include limited sample availability, class imbalance, distribution shift, and insufficient data diversity. Such failures primarily affect the statistical characteristics of the learning data and are generally mitigated through data-oriented interventions such as augmentation, sampling, and domain adaptation.

2) Decision failure: It refers to robustness degradation introduced during model optimisation and inference. Even when sufficient data are available, inappropriate optimisation objectives, loss functions, class weighting strategies [33], or decision boundaries may lead to unstable learning behaviour and biased predictions. Decision failure therefore originates from misalignment in optimisation rather than data insufficiency.

3) Representation failure: It refers to robustness degradation caused by information loss before feature learning begins. This category includes image degradation resulting from acquisition noise, motion blur, illumination variation, CT artifacts,

low contrast, and other sensing imperfections. Representation failure limits the quality of information available to downstream learning algorithms and is typically addressed through image enhancement or restoration techniques. Unlike distribution failure and decision failure, representation failure is evaluated at the image representation level rather than directly through downstream detection performance. The objective of a representation-oriented intervention is to recover degraded information and improve the quality of inputs available for subsequent inspection tasks.

Accordingly, the proposed framework distinguishes the three failure categories according to the stage at which robustness degradation is introduced: before model optimisation (distribution failure), during optimisation and inference (decision failure), and before feature learning due to acquisition degradation (representation failure). The taxonomy and formal definitions of the three failure categories are summarised in Table II. This stage-oriented definition establishes the conceptual boundary between failure categories and provides the basis for failure-oriented intervention selection.

In practical industrial inspection systems, multiple failure sources may coexist simultaneously. The proposed framework does not assume that distribution failure, decision failure, and representation failure are mutually exclusive. Instead, these categories provide a structured mechanism for diagnosing the dominant source of robustness degradation under a specific inspection scenario.

The dominant failure refers to the failure source that contributes most significantly to the observed robustness degradation and should therefore receive priority during intervention selection. Other failure sources may still exist as secondary factors and may require additional interventions when their influence becomes significant.

C. Framework Architecture

The proposed Failure-Oriented Robustness Framework is designed to analyse robustness from the perspective of failure generation rather than algorithm selection. Instead of treating augmentation, optimisation, and enhancement as independent technical solutions, the framework organises robustness challenges according to the dominant failure mechanisms responsible for inspection performance degradation [5], [7], [24].

As illustrated in Fig. 2, the framework consists of three hierarchical layers: Distribution Failure, Decision Failure, and Representation Failure. These layers correspond to different stages of the industrial inspection pipeline and collectively describe how robustness degradation emerges and propagates throughout the system.

1) Distribution failure layer: These layers are arranged according to the stage at which robustness degradation is introduced into the inspection pipeline, ranging from data generation and model decision-making to image representation quality. Distribution failure originates from deficiencies in the training data distribution, including data scarcity, class imbalance, and domain variability. Such failures often manifest as poor generalisation capability, unstable rare-class detection, and sensitivity to distribution shifts [4], [5]. Since the underlying issue is data-related, interventions at this layer primarily focus on modifying the data distribution to improve representation diversity and class coverage.

2) Decision failure layer: Decision failure occurs when the learning objective becomes misaligned with the inspection objective [6], [7]. Even when sufficient data are available, optimisation objectives may favour dominant classes, suppress minority categories, or produce unstable localisation behaviour. These failures are associated with the model's decision-making process rather than with the data itself. Consequently, improvement in robustness at this layer focuses on objective design, loss functions, weighting mechanisms, and optimization strategies.

3) Representation failure layer: Representation failure arises before the learning stage and is caused by degradation in the acquired inspection data [24], [25], [26]. Common examples include image artifacts, noise contamination, contrast reduction, and resolution loss. These effects distort the defect representation and limit the amount of useful information available for downstream analysis. Therefore, interventions at this layer primarily involve enhancement, restoration, and super-resolution techniques [24], [26], [36].

Together, these three layers provide a unified view of robustness in industrial inspection systems, based on failure. Unlike conventional method-centric approaches that evaluate individual algorithms in isolation, the proposed framework emphasises identifying dominant failure mechanisms and selecting interventions that directly address the underlying cause of performance degradation. From this perspective, robustness is understood as the outcome of alignment between intervention strategies and failure sources, rather than a direct consequence of increasing algorithmic complexity.

D. Failure Source–Intervention Mapping

While the framework architecture identifies the dominant sources of robustness degradation, practical improvements in robustness require a mechanism for selecting appropriate interventions. Conventional industrial inspection studies typically adopt a method-centric workflow in which augmentation techniques, optimisation strategies, or enhancement algorithms are evaluated independently [5], [7], [24]. Under such a paradigm, intervention selection is often driven by methodological popularity rather than the underlying cause of failure.

The proposed framework adopts a failure-oriented intervention strategy. Instead of asking which method performs best, the framework first identifies the dominant failure source and subsequently selects interventions that directly target the corresponding failure mechanism. This process establishes an explicit relationship between failure diagnosis and intervention selection.

For distribution failures, robustness degradation originates from deficiencies in the training data distribution, including scarcity, imbalance, and domain variability. Consequently, interventions should focus on modifying the data distribution through augmentation, balancing, or domain adaptation techniques.

For decision failures, performance degradation arises from optimisation bias and objective misalignment during model learning. Therefore, interventions should focus on objective

TABLE II. TAXONOMY AND DEFINITIONS OF THE PROPOSED FAILURE CATEGORIES

Failure Category	Failure Stage	Origin of Failure	Typical Intervention
Distribution Failure	Before model optimisation	Training data deficiencies, including limited samples, class imbalance, and distribution shift	Data augmentation, sampling, and domain adaptation
Decision Failure	During optimisation and inference	Misalignment between optimisation objectives and inspection objectives, including inappropriate loss functions, optimisation strategies, and decision boundaries	Loss function design, optimisation strategy, and class weighting
Representation Failure	Before feature learning	Information degradation caused by image acquisition artifacts, noise, blur, low contrast, and sensing imperfections	Image enhancement, image restoration, and super-resolution

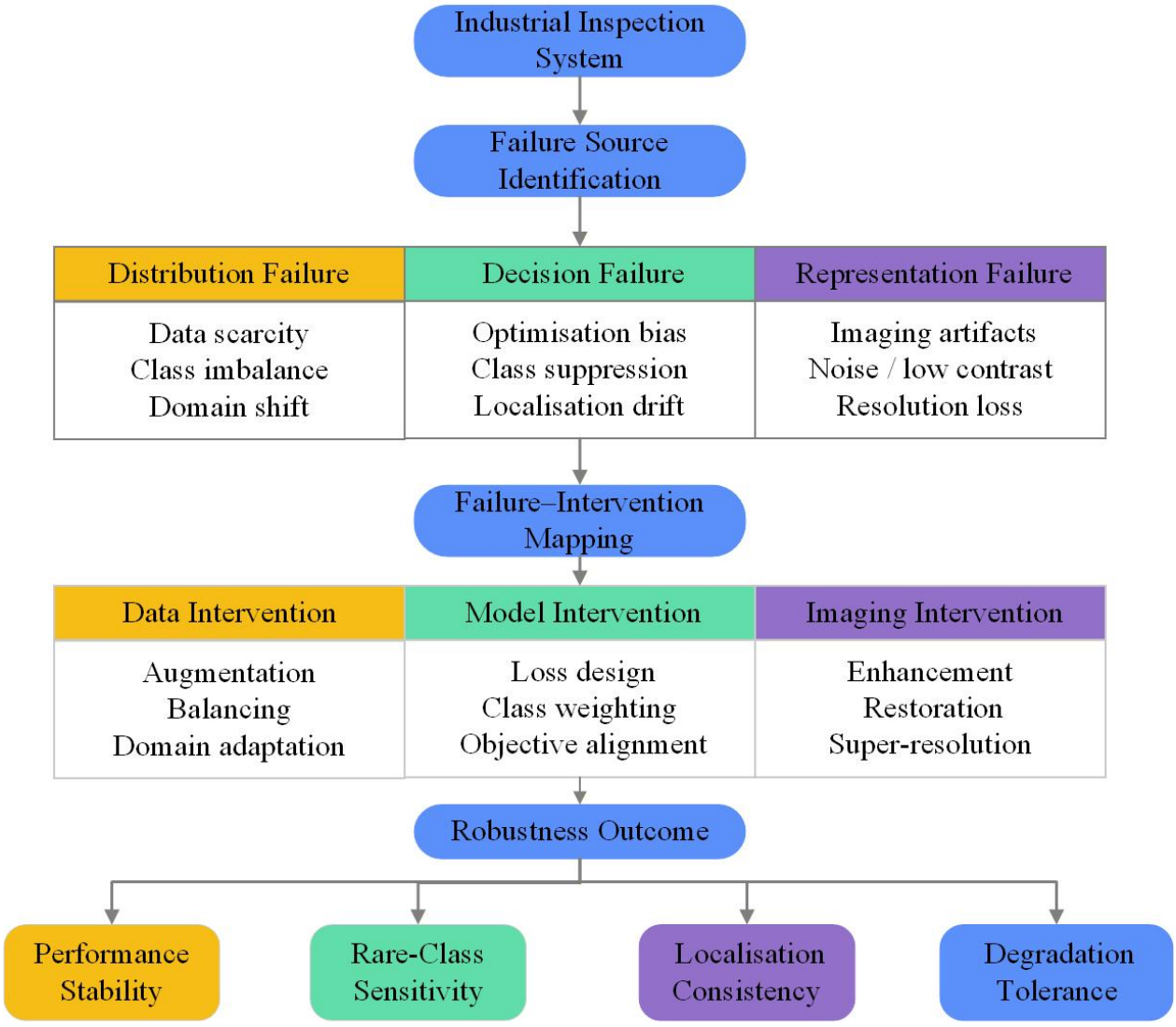


Fig. 2. Proposed Failure-Oriented Robustness Framework for industrial inspection. The framework organises robustness challenges according to three dominant sources of failure and maps them to corresponding intervention strategies and robustness outcomes.

design, loss formulation, class weighting, and optimisation strategies that improve decision consistency and minority-class representation.

For representation failures, degradation occurs before model learning due to corruption of the acquired inspection data. In such cases, robustness improvement should prioritise enhancement, restoration, and super-resolution techniques to recover defect visibility and structural information.

The proposed mapping mechanism does not assume the existence of a universally optimal intervention. Instead, it emphasises alignment between intervention strategy and dominant failure source. Under this perspective, intervention effec-

tiveness is considered failure-dependent rather than method-dependent, providing a systematic basis for robustness-oriented decision making in industrial inspection systems.

E. Robustness Interpretation Mechanism

A central objective of the proposed framework is to reinterpret robustness beyond conventional accuracy-oriented evaluation. Existing industrial inspection studies predominantly assess performance using aggregate metrics such as accuracy, precision, recall, and mean Average Precision (mAP) [17], [18]. While these metrics provide useful indicators of detection performance, they do not necessarily reflect an inspection system's ability to maintain reliable behaviour under adverse operating conditions.

Within the proposed framework, robustness is defined as an inspection system's ability to maintain stable, reliable performance when confronted with dominant failure sources. Consequently, robustness is viewed as a multidimensional property rather than a single performance metric.

To facilitate the interpretation of robustness, four complementary dimensions are considered. Performance Stability evaluates the consistency of inspection performance across varying operating conditions. Rare-Class Sensitivity measures the system's ability to maintain detection effectiveness for minority defect categories. Localisation Consistency reflects the stability and reliability of defect localisation behaviour during model prediction. Degradation Tolerance assesses the inspection system's ability to maintain performance under imaging degradation, noise contamination, or structural corruption.

These dimensions provide an interpretable link between failure diagnosis, intervention effectiveness, and robustness outcomes. Under the proposed framework, robustness is improved not merely by higher accuracy, but by sustained performance under the specific failure conditions that dominate the inspection process. Therefore, robustness is interpreted as the observable outcome of successful alignment between intervention strategies and failure mechanisms.

F. Practical Application Workflow

While Fig. 2 explains what constitutes the framework, Fig. 3 illustrates how the framework is applied during practical inspection. As illustrated in Fig. 3, the proposed framework can be operationalised through a sequential failure-oriented workflow.

To facilitate practical deployment, the proposed framework is implemented through a sequential failure-oriented workflow. Rather than selecting algorithms directly, the workflow begins with diagnosing the dominant source of robustness degradation and subsequently identifying interventions that target the corresponding failure mechanism.

The workflow consists of five stages. First, the inspection scenario is analysed to identify observable performance limitations, such as poor rare-class detection, unstable localisation behaviour, or degraded image quality. Second, these symptoms are traced to their dominant source of failure, which may be categorised as distribution, decision, or representation failure.

Third, intervention strategies are selected according to the identified failure category. Distribution failures are addressed through data-oriented interventions, decision failures through optimisation-oriented interventions, and representation failures through imaging-oriented interventions. Fourth, robustness outcomes are evaluated using the interpretation dimensions defined in the framework, including performance stability, rare-class sensitivity, localisation consistency, and degradation tolerance.

Finally, the intervention results are analysed, and the workflow is iteratively refined if the dominant failure source remains unresolved. This iterative process allows robustness improvement to be guided by failure diagnosis rather than algorithmic trial-and-error.

The proposed workflow transforms the framework from a conceptual robustness model into a structured analytical procedure for industrial inspection applications. By prioritising failure diagnosis before intervention selection, the framework provides a systematic procedure for analysing robustness degradation and guiding the development of robustness-oriented inspection systems.

IV. Framework Validation Through Turbine Blade Inspection

A. Experimental Settings

Since the proposed framework investigates robustness degradation from different failure sources, the experimental configurations were designed according to the characteristics of each validation scenario. The data-level, model-level, and imaging-level experiments correspond to different intervention mechanisms; therefore, their training configurations are not identical. Instead, each experimental setting was controlled to ensure a fair comparison among methods within the corresponding validation scenario.

1) 2D Detection experiments: The 2D industrial inspection experiments were conducted using the YOLOv8 object detection framework implemented with the Ultralytics library. YOLO-based detectors have been widely adopted for real-time object detection because of their balance between detection accuracy and computational efficiency [28]. The experiments were performed using Python 3.9.19, PyTorch 2.4.1 with CUDA 12.4, and an NVIDIA GeForce RTX 4060 Laptop GPU with 8 GB VRAM. The adopted YOLOv8n detector contains 3,006,623 parameters and 8.1 GFLOPs.

For the data-level validation, the objective was to investigate whether augmentation strategies could alleviate distribution failure caused by limited samples, class imbalance, and distribution variability. Therefore, the augmentation strategy was considered the primary experimental variable, while the detector training configuration was maintained consistently within each dataset. The models were trained for 100 epochs with an input resolution of 640×640 pixels and a batch size of 1. The default Ultralytics optimization configuration was adopted, and different augmentation strategies with their corresponding parameters were applied according to the experimental design.

For the model-level validation, the objective was to analyse decision failure caused by optimisation misalignment. Unlike

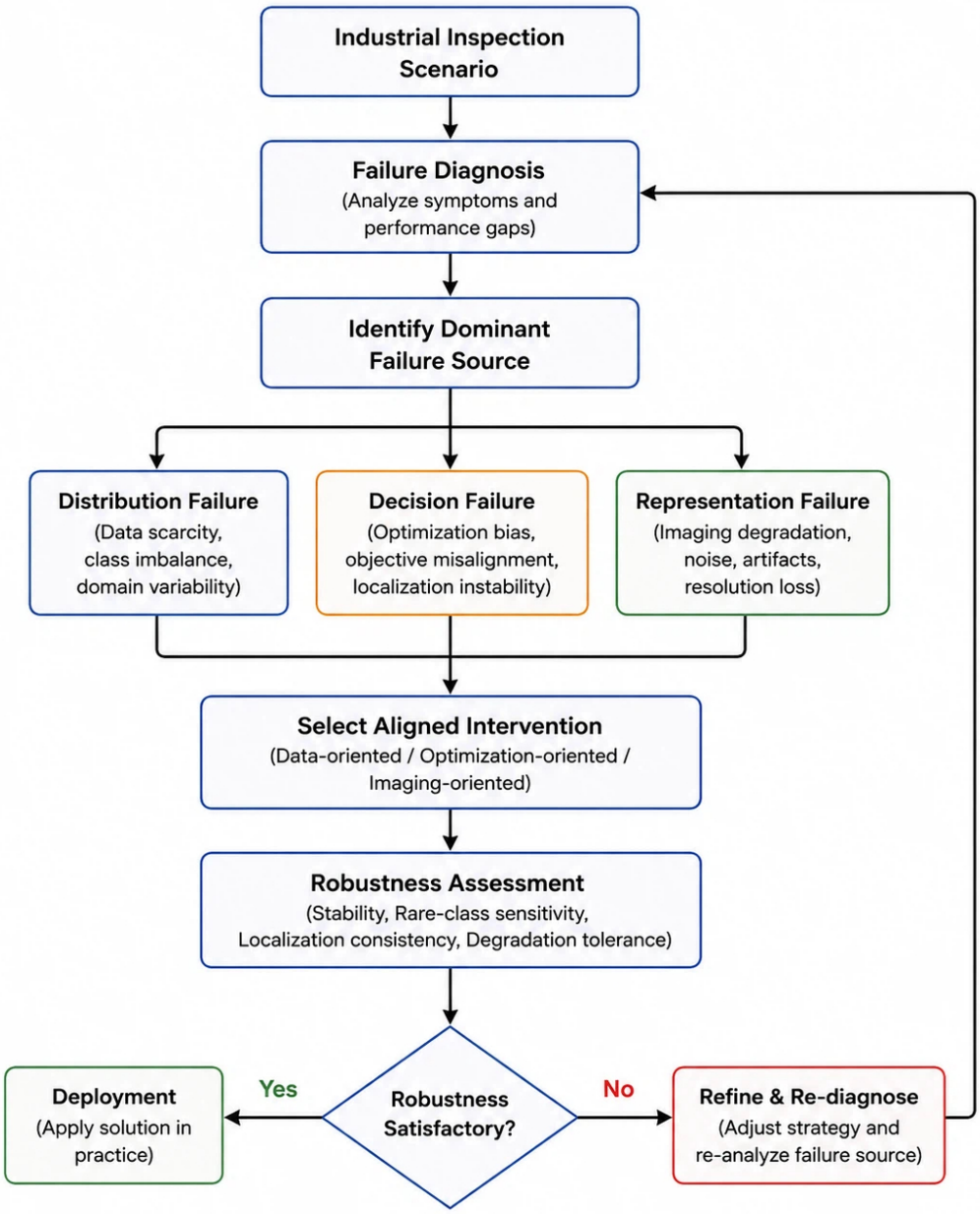


Fig. 3. Failure-oriented robustness workflow for industrial inspection systems.

the data-level experiments, optimisation strategies were treated as the primary experimental variables. Different optimizers, loss functions, class weighting strategies, and architectural modifications were evaluated under controlled training conditions. The models were trained for 300 epochs with an input resolution of 640 $\times$ 640 pixels and a batch size of 16. A fixed random seed of 42 was adopted to improve reproducibility. The optimizer configuration was adjusted according to the specific comparison objectives, including Adam [34] and SGD evaluations. Early stopping was disabled to ensure that all models completed the predefined training schedule.

For all 2D detection experiments, the performance was evaluated using mean average precision at IoU threshold 0.5 (mAP@0.5), mean average precision across IoU thresholds from 0.5 to 0.95 (mAP@0.5:0.95), precision, recall, and class-wise average precision.

2) CT Representation recovery experiments: The CT experiments were designed to validate representation failure caused by image acquisition degradation. Unlike the 2D detection experiments, the objective of this validation scenario was not to directly improve defect detection performance but to recover degraded image representations caused by CT artifacts.

A lightweight GNN-based artifact removal framework was implemented using PyTorch. Several restoration models, in-

cluding CNN, U-Net [35], and different GNN variants, were compared under identical training conditions. The CT images were converted into normalized grayscale representations with pixel values scaled to the range of [0,1]. The training and validation datasets were loaded according to predefined image-mask pairs.

All restoration models were trained using the Adam optimizer with a learning rate of 1×10^{-3} and the mean squared error (MSE) loss function. The batch size was set to 1, and each model was trained for 50 epochs. The restoration performance was evaluated using peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM).

The CT validation therefore focuses on representation-level robustness, where the objective is to assess whether degraded image information can be recovered before subsequent inspection analysis.

To ensure the consistency and reliability of the reported results, repeated experiments were conducted under identical experimental configurations. For the YOLO-based 2D detection experiments, repeated runs produced identical evaluation results under the fixed training settings. For the CT representation recovery experiments, repeated runs showed minor numerical variations due to the stochastic nature of deep learning optimization; however, the relative performance trends among different restoration models remained consistent. These observations indicate that the reported robustness behaviours are reproducible and primarily associated with the investigated intervention strategies rather than incidental training variations.

B. Validation Scenario Overview

To evaluate the explanatory capability of the proposed framework, validation was conducted using representative industrial inspection scenarios derived from turbine blade inspection. Rather than evaluating the superiority of a specific algorithm, this section investigates whether the proposed framework can consistently explain robustness behaviour across different failure conditions.

Three validation scenarios were considered. Distribution failure was evaluated using two industrial defect datasets exhibiting substantially different levels of data scarcity and class imbalance. We evaluated decision failure through optimisation and architectural studies that analysed how learning objectives influenced detection robustness. Representation failure was evaluated using CT imaging scenarios affected by artifacts and structural degradation.

Collectively, these validation cases provide empirical evidence to examine whether dominant failure sources govern robustness behaviour and whether intervention effectiveness depends on alignment with those sources.

C. Distribution Failure Validation

To validate the distribution-failure component of the proposed framework, two industrial defect datasets with substantially different distributional characteristics were analysed. Dataset 1 (Alibaba Tianchi) represents a small-scale, severely imbalanced inspection scenario, containing 247 images and 388 defect instances, with an imbalance ratio of 7.33. In contrast, Dataset 2 (AeBAD-S) represents a larger-scale, relatively balanced inspection scenario, comprising 1,592 images and 2,557 defect instances.

TABLE III. REPRESENTATIVE AUGMENTATION RESULTS UNDER SEVERE DISTRIBUTION FAILURE.

Method	mAP@0.5	Dirty AP	Stain AP
Baseline	0.290	0.000	0.000
MixUp	0.591	0.667	0.672
CycleGAN	0.461	0.667	0.000
C5 Composite	0.599	0.731	0.452

According to the proposed framework, Dataset 1 is primarily characterized by distribution failures arising from data scarcity and class imbalance, whereas Dataset 2 exhibits substantially weaker distribution-failure characteristics. The contrasting properties of the two datasets provide an appropriate validation scenario for evaluating whether intervention effectiveness is determined by the intervention itself or by the dominant failure source.

1) Case A: Small-scale imbalanced dataset: Dataset 1 represents a typical industrial inspection scenario characterised by severe distribution failure. The dataset contains only 247 images and 388 defect instances, with an imbalance ratio of 7.33 between the majority and minority defect categories. Under these conditions, the baseline detector exhibited substantial degradation in robustness, particularly for minority defect classes, in which several categories could not be reliably detected.

As shown in Table III, substantial improvements were observed after applying distribution-oriented interventions. In particular, MixUp [27] and the C5 composite strategy significantly improved minority-class detection performance, transforming previously undetected defect categories into reliably detectable classes.

According to the proposed framework, these improvements are not solely attributable to the augmentation techniques themselves. Instead, the observed robustness gains stemmed from the selected interventions directly addressing the dominant failure sources, namely data scarcity and class imbalance. By increasing semantic diversity and improving minority-class representation [29], [30], the interventions effectively mitigated the distribution failure present in Dataset 1.

This case therefore provides empirical evidence supporting the distribution failure component of the proposed framework. The results suggest that intervention effectiveness depends on alignment with the underlying source of failure rather than on the intervention strategy alone.

2) Case B: Large-scale balanced dataset: In contrast to Dataset 1, Dataset 2 represents a larger-scale industrial inspection scenario with substantially weaker characteristics of distribution failure. The dataset contains 1,592 images and 2,557 defect instances, exhibiting a considerably more balanced class distribution. Consequently, data scarcity and class imbalance are no longer the dominant constraints affecting robustness.

As shown in Table IV, augmentation strategies exhibited markedly different behaviour compared with Dataset 1. While geometric augmentation produced measurable improvements,

TABLE IV. REPRESENTATIVE AUGMENTATION RESULTS UNDER WEAK DISTRIBUTION FAILURE.

Method	mAP_{50}	Ablation AP
Baseline	0.814	0.443
Geometric	0.862	0.684
Mosaic	0.827	0.549
MixUp	0.809	0.443
Diffusion	0.812	0.464

semantic-mixing approaches such as MixUp [27] provided limited benefit and did not improve minority-class performance. Similarly, the overall performance gains achieved through augmentation were substantially smaller than those observed under severe distribution failure conditions.

According to the proposed framework, this behaviour can be explained by the reduced influence of distribution failure in Dataset 2. Since the dataset already provides sufficient sample diversity and a relatively balanced class representation, interventions specifically designed to address data scarcity and imbalance are no longer the most effective robustness strategy, which is consistent with previous observations on dataset-dependent augmentation and domain generalisation [29], [5].

These observations further support the failure-oriented interpretation of robustness. The effectiveness of augmentation was not determined by the intervention itself but by the extent to which it aligned with the dominant failure source in the inspection scenario.

3) Framework interpretation: The contrasting behaviours observed across Dataset 1 and Dataset 2 provide empirical support for the central assumption of the proposed framework. Although similar augmentation strategies were applied in both scenarios, their effectiveness varied substantially depending on the datasets' underlying distributional characteristics.

In Dataset 1, robustness degradation was primarily associated with severe data scarcity and class imbalance. Under these conditions, interventions designed to increase data diversity and improve minority-class representation produced substantial improvements in detection performance, particularly for previously underrepresented defect categories. In contrast, Dataset 2 exhibited substantially weaker characteristics of distribution failure. Consequently, the same augmentation strategies produced reduced benefits and, in some cases, negligible performance improvements.

These observations suggest that augmentation effectiveness is not an intrinsic property of the augmentation strategy itself. Rather, intervention effectiveness depends on the degree of alignment between the intervention and the dominant source of failure. When severe distribution failure is present, distribution-oriented interventions can significantly improve robustness. However, when distribution failure is no longer the primary constraint, the same interventions may provide limited value.

From the perspective of the proposed framework, robustness improvement is therefore failure-dependent rather than method-dependent. The distribution failure validation demonstrates that selecting an effective intervention requires identifying the dominant failure source in advance rather than adopting a universally preferred technique.

TABLE V. REPRESENTATIVE RESULTS FOR OBJECTIVE ALIGNMENT

Configuration	mAP_{50}
BCE + CIoU	0.588
Focal + CIoU	0.676
Focal + CIoU + Weighting	0.710

TABLE VI. REPRESENTATIVE RESULTS OF ARCHITECTURAL REFINEMENT UNDER DECISION FAILURE.

Configuration	mAP_{50}
YOLOv8n	0.588
CBAM	0.303
AIFI	0.342
C2fPSA	0.226

D. Decision Failure Validation

1) Case A: Objective alignment: To validate the decision failure component of the proposed framework, optimisation objectives were examined under identical architectural conditions. In this validation scenario, the detector architecture remained unchanged while different objective formulations, including imbalance-aware loss functions, were applied [31].

As shown in Table V, substantial robustness improvements were achieved through objective redesign without modifying the detector architecture. The results indicate that decision quality was strongly influenced by the optimisation objective, particularly under imbalanced inspection conditions where minority defect categories were vulnerable to suppression. Bounding-box regression objectives were also investigated using IoU-based loss formulations, following previous improvements in localization optimisation [32].

According to the proposed framework, these improvements are associated with mitigating decision failure rather than with changes in model complexity. By improving the alignment between optimisation objectives and inspection objectives, the detector exhibited more stable learning behaviour and improved minority-class representation.

2) Case B: Architectural refinement: A second validation scenario examined whether robustness could be improved through architectural refinement alone. Several attention and feature-enhancement modules were incorporated into the baseline detector architecture.

As shown in Table VI, architectural refinement did not consistently improve robustness. Several architectural modifications resulted in performance degradation despite increasing model complexity. These observations suggest that the robustness limitations were more closely associated with optimisation and decision-making behaviour than with insufficient representational capacity alone.

3) Framework Interpretation: The results of the decision failure validation support the central assumption of the proposed framework. Under identical architectural conditions, objective alignment produced substantial robustness improvements, whereas increased architectural complexity failed to provide consistent benefits.

These observations indicate that decision failure originates primarily from optimisation behaviour rather than model ca-

TABLE VII. REPRESENTATIVE IMAGE QUALITY RESULTS FOR REPRESENTATION FAILURE RECOVERY.

Method	PSNR	SSIM
Gaussian Filter	21.82	0.7910
CNN	27.83	0.6552
U-Net	25.85	0.9201
GNN v2	28.93	0.9220

pacity. Consequently, interventions targeting objective design, class weighting, and optimisation strategies are often more effective than architectural refinement when decision failure is the dominant source of robustness degradation.

From the perspective of the proposed framework, the improvement in robustness in this scenario emerged from alignment between the optimisation and inspection objectives rather than from increasing architectural complexity. This finding provides empirical support for the framework's decision-failure component and further demonstrates that the dominant source of failure governs intervention effectiveness.

E. Representation Failure Validation

1) Case study CT representation recovery under imaging degradation: To validate the representation-failure component of the proposed framework, industrial computed tomography (CT) scenarios affected by imaging degradation were analysed. Unlike distribution failure and decision failure, which are directly reflected through detection performance, representation failure occurs before model inference and is therefore evaluated through the recovery of degraded image representations.

In industrial CT inspection, image quality may be degraded by ring artifacts, streak artifacts, beam-hardening effects, noise contamination, and reduced contrast [24], [25]. These degradations distort defect-related structural information and reduce the quality of image representations available for subsequent inspection tasks. Consequently, robustness degradation may occur before feature extraction and model inference due to insufficient or corrupted visual information.

According to the proposed framework, such degradation corresponds to representation failure. To address this source of failure, image enhancement and restoration techniques were applied to recover structural information and improve defect visibility [26].

The representative image quality results are summarised in Table VII. The enhancement process improved image quality and restored structural information under degraded CT conditions. These results demonstrate that representation-oriented interventions can alleviate information loss caused by imaging degradation and provide more reliable image representations for subsequent inspection analysis. However, downstream defect detection evaluation after CT enhancement was not conducted in this study, as the objective of this validation case is to assess representation recovery rather than end-to-end detection improvement.

F. Cross-Level Analysis

The three validation scenarios demonstrate that robustness degradation in industrial inspection systems does not originate from a single source. Instead, robustness challenges emerge at different stages of the inspection pipeline and are associated with distinct failure mechanisms.

Distribution failure primarily arises from deficiencies in the training data, including data scarcity, class imbalance, and distributional variability. Decision failure arises from optimisation misalignment during model learning, whereas representation failure is introduced prior to learning through imaging degradation and information loss.

Although these failures occur at different stages of the inspection pipeline, they share a common characteristic: each affects robustness through a distinct mechanism and therefore requires a different intervention strategy. Consequently, robustness cannot be improved through a universally superior algorithm; rather, improvement depends on addressing the dominant constraint present in the inspection scenario.

The three validation cases also reveal an important cross-level observation. Interventions that were highly effective under one failure condition did not necessarily produce comparable improvements under another. Distribution-oriented augmentation substantially improved robustness under severe data imbalance but provided limited benefit when distribution failure was weak. Similarly, optimisation-oriented interventions consistently outperformed architectural refinement only when decision failure dominated, whereas imaging enhancement became the primary contributor to robustness under degraded CT acquisition conditions.

These observations suggest that robustness degradation should be analysed by the stage at which failure is introduced rather than by methodological categories alone. Collectively, these cross-level observations establish a common analytical basis for interpreting robustness across heterogeneous industrial inspection scenarios and naturally motivate the unified framework synthesis presented in the following section.

G. Framework Synthesis

The three validation scenarios collectively provide converging empirical evidence supporting the proposed Failure-Oriented Robustness Framework. Although each validation examined different inspection tasks, datasets, and intervention strategies, all cases exhibited a consistent pattern of robustness that can be interpreted within the proposed failure-oriented perspective.

The relationships identified across the three validation scenarios are summarised in Table VIII.

This synthesis demonstrates that the proposed framework provides a unified explanation of robustness behaviour across heterogeneous industrial inspection systems. Rather than treating augmentation, optimisation, and image enhancement as isolated technical solutions, the framework interprets them as complementary interventions targeting different stages of robustness degradation.

A common robustness principle therefore emerges across all validation scenarios: intervention effectiveness is governed by alignment with the dominant failure source rather than by algorithmic superiority alone. Consequently, robustness should be understood as a failure-dependent property rather than an algorithm-dependent property.

TABLE VIII. FRAMEWORK SYNTHESIS ACROSS VALIDATION SCENARIOS

Failure	Validation Observation	Framework Insight
Distribution	Augmentation is effective only under severe imbalance	Distribution alignment
Decision	Objective alignment consistently outperforms architectural refinement	Objective alignment
Representation	Enhancement restores structural visibility under degraded CT	Representation recovery

Taken together, these findings demonstrate that the proposed Failure-Oriented Robustness Framework provides a unified conceptual foundation for robustness-oriented analysis, intervention selection, and future framework development in safety-critical industrial inspection systems.

V. DISCUSSION AND FUTURE PERSPECTIVES

A. From Method-Centric Optimisation to Failure-Oriented Design

A central implication of the proposed framework is the shift from method-centric optimisation towards failure-oriented robustness design. Existing industrial inspection research has predominantly focused on identifying superior algorithms through architectural innovation, optimisation refinement, and data augmentation. Under this paradigm, robustness is often treated as a secondary outcome of improved performance.

However, the validation results presented in this study suggest that robustness behaviour cannot be fully explained by algorithmic superiority alone. Across all validation scenarios, intervention effectiveness depended strongly on the dominant source of failure. Methods that produced substantial improvements under one failure condition frequently demonstrated limited effectiveness under another.

These observations indicate that robustness should be analysed from the perspective of failure mechanisms rather than individual algorithms. Consequently, future industrial inspection systems may benefit from prioritising failure diagnosis before intervention selection, thereby transforming robustness from an implicit consequence of optimisation into an explicit design objective.

B. Implications for Industrial Inspection Robustness

The proposed framework has several implications for the design and evaluation of industrial inspection systems. Traditionally, robustness improvement has been pursued through the continuous development of new architectures, optimisation strategies, and data augmentation techniques. While such efforts have achieved notable performance gains, the results of this study suggest that robustness cannot be fully understood through algorithmic improvement alone.

Instead, robustness should be interpreted as a consequence of the interaction between failure sources and intervention strategies. Under this perspective, the primary objective of robustness-oriented inspection is not to identify universally superior methods, but to determine which intervention is most appropriate for the dominant failure condition. Consequently, robustness evaluation should incorporate failure diagnosis and intervention alignment in addition to conventional performance metrics.

For industrial practitioners, this perspective provides a practical decision-support mechanism. Rather than relying solely on trial-and-error model selection, inspection systems can be developed through the systematic identification of failure sources and the targeted design of interventions. Such an approach may improve both development efficiency and deployment reliability in safety-critical industrial environments.

C. Generalisability and Limitations

Although the proposed framework was validated using turbine blade inspection scenarios, the underlying failure categories are not specific to this application domain. Data scarcity, optimisation bias, and representation degradation are common challenges across a wide range of industrial inspection tasks, including weld inspection, casting inspection, surface defect detection, and industrial medical imaging. Consequently, the framework may provide a general robustness-oriented perspective for analysing inspection systems beyond turbine blade inspection.

Nevertheless, several limitations should be acknowledged. First, the framework was validated through representative case studies rather than a unified end-to-end implementation. Second, interactions between multiple failure sources were not explicitly investigated. In practical inspection systems, distribution, decision, and representation failures may coexist and influence each other simultaneously. Finally, the framework focuses primarily on robustness interpretation and intervention selection rather than on providing a universal optimisation strategy.

Furthermore, the current framework should be regarded as an analytical and interpretative framework rather than a fully predictive intervention recommendation system. Although the proposed failure-oriented perspective establishes relationships between failure sources and suitable intervention categories, it does not automatically identify failure dominance, rank competing interventions, or quantitatively predict expected robustness improvement before experimentation. Developing such predictive capabilities would require additional data-driven modelling and large-scale cross-domain validation, which represents an important direction for future research.

These limitations provide opportunities for future refinement and broader validation of the framework.

D. Future Research Directions

Several directions can be explored in future research for further developing the proposed framework. First, future studies should investigate cross-level interactions between distribution, decision, and representation failures. Understanding how multiple failure sources jointly influence robustness may lead to more comprehensive intervention strategies.

Second, future research may extend the framework towards predictive robustness management, where failure diagnosis and intervention selection are supported by quantitative models. Such extensions may enable automatic estimation of dominant failure sources, intervention ranking, and expected robustness improvements under different inspection conditions.

Finally, broader validation across different industrial domains is required. Applying the framework to additional inspection scenarios, including manufacturing quality control, infrastructure monitoring, and medical imaging, would help evaluate its generalisability and identify domain-specific robustness challenges.

Through these developments, the proposed framework may evolve from a robustness-interpretation model into a more comprehensive robustness-design methodology for industrial inspection systems.

When robustness degradation is observed, the first question should not be "Which algorithm should be used?" Instead, the dominant failure source should first be identified, followed by failure-aligned intervention selection.

VI. Conclusion

This study addressed a fundamental challenge in industrial inspection: robustness is widely recognised as a critical requirement, yet its behaviour remains difficult to explain using conventional method-centric perspectives. Existing studies have primarily focused on developing new architectures, optimisation strategies, and augmentation techniques, often treating robustness as a secondary consequence of performance improvement rather than an explicit research objective.

To address this limitation, a Failure-Oriented Robustness Framework was proposed. Instead of analysing robustness at the level of individual algorithms, the framework organises robustness challenges into three dominant failure sources: distribution failure, decision failure, and representation failure. By establishing explicit relationships between failure diagnosis, intervention selection, and robustness outcomes, the framework provides a systematic mechanism for understanding and improving robustness in industrial inspection systems.

The framework was validated through representative turbine blade inspection scenarios covering data-level, model-level, and imaging-level robustness challenges. Across all validation cases, a consistent pattern emerged. Distribution-oriented interventions were most effective when robustness degradation originated from data scarcity and class imbalance. Optimisation-oriented interventions produced the greatest benefits when decision failure dominated the learning process. Representation-oriented interventions were most effective when imaging degradation limited the visibility of defects. These observations demonstrate that intervention effectiveness varies according to the dominant source of failure.

The central finding of this work is that robustness is not an intrinsic property of a particular method, nor is it achieved simply through increasing algorithmic complexity. Rather, robustness emerges from alignment between intervention strategies and dominant failure sources. This perspective shifts robustness research from method-centric optimisation towards failure-oriented robustness design, providing a more interpretable and practically relevant foundation for industrial inspection systems.

Although the framework was validated using turbine blade inspection scenarios, the proposed perspective is not restricted to a single application domain. Future research may extend the framework to additional industrial and imaging environments, investigate interactions between multiple failure sources, and develop adaptive robustness management strategies. Through such developments, the proposed framework may contribute towards a more systematic and generalisable understanding of robustness in industrial inspection. The proposed framework establishes a conceptual foundation for future robustness-oriented industrial inspection systems, where intervention strategies are selected according to failure mechanisms rather than algorithmic popularity.

Acknowledgment

This work was supported by the Tan Sri Leo Moggie Chair of Energy Informatics Publication Fund, Universiti Tenaga Nasional (UNITEN), under Grant No. 20211EICHAIR (UKM Reference No. ZG-2024-005).

References

[1] A. Ettalibi, A. Elouadi, and A. Mansour, "AI and computer vision-based real-time quality control: A review of industrial applications," *Procedia Computer Science*, vol. 231, pp. 212–220, 2024, doi: 10.1016/j.procs.2023.12.195.

[2] R. Khanam, M. Hussain, R. Hill, and P. Allen, "A comprehensive review of convolutional neural networks for defect detection in industrial applications," *IEEE Access*, vol. 12, pp. 94250–94295, 2024, doi: 10.1109/ACCESS.2024.3425166.

[3] J. Li and G. Li, "Triangular trade-off between robustness, accuracy, and fairness in deep neural networks: A survey," *ACM Computing Surveys*, vol. 57, no. 6, pp. 1–40, 2025, doi: 10.1145/3645088.

[4] P. Singhal, R. Walambe, S. Ramanna, and K. Kotecha, "Domain adaptation: Challenges, methods, datasets, and applications," *IEEE Access*, vol. 11, pp. 6973–7020, 2023, doi: 10.1109/ACCESS.2023.3237025.

[5] J. Mai, C. Gao, and J. Bao, "Domain generalization through data augmentation: A survey of methods, applications, and challenges," *Mathematics*, vol. 13, no. 5, Art. no. 824, 2025, doi: 10.3390/math13050824.

[6] N. Suh and G. Cheng, "A survey on statistical theory of deep learning: Approximation, training dynamics, and generative models," *Annual Review of Statistics and Its Application*, vol. 12, no. 1, pp. 177–207, 2025, doi: 10.1146/annurev-statistics-112723-034650.

[7] J. Terven, D. M. Cordova-Esparza, J. A. Romero-González, A. Ramírez-Pedraza, and E. A. Chavez-Urbiola, "A comprehensive survey of loss functions and metrics in deep learning," *Artificial Intelligence Review*, vol. 58, no. 7, Art. no. 195, 2025, doi: 10.1007/s10462-025-11150-6.

[8] M. Gupta, M. A. Khan, R. Butola, and R. M. Singari, "Advances in applications of non-destructive testing (NDT): A review," *Advances in Materials and Processing Technologies*, vol. 8, no. 2, pp. 2286–2307, 2022.

[9] F. Abdollahi-Mamoudan, C. Ibarra-Castanedo, and X. P. Maldague, "Non-destructive testing and evaluation of hybrid and advanced structures: A comprehensive review of methods, applications, and emerging trends," *Sensors*, vol. 25, no. 12, Art. no. 3635, 2025.

[10] R. E. Donatus, "A comprehensive review of artificial intelligence for image- and signal-based nondestructive testing in aerospace structures," *Journal of Computers, Mechanical and Management*, vol. 4, no. 6, pp. 1–14, 2025.

[11] K. Masita, A. Hasan, T. Shongwe, and H. A. Hilal, "Deep learning in defects detection of PV modules: A review," *Solar Energy Advances*, vol. 5, Art. no. 100090, 2025.

[12] Z. Ren, F. Fang, N. Yan, and Y. Wu, "State of the art in defect detection based on machine vision," *International Journal of Precision Engineering and Manufacturing-Green Technology*, vol. 9, no. 2, pp. 661–691, 2022.

[13] S. Kim, T. P. Nguyen, and J. Yoon, "A systematic review of machine vision applications in factory and manufacturing processes: From quality control to predictive diagnostics," *International Journal of Precision Engineering and Manufacturing-Green Technology*, vol. 13, no. 2, pp. 745–775, 2026.

[14] Q. Wang, M. Wang, J. Sun, D. Chen, and P. Shi, "Review of surface-defect detection methods for industrial products based on machine vision," *IEEE Access*, 2025.

[15] A. Jayawardena, J. H. Sul, D. Moratuwage, J. L. Wijayaraja, and L. Piyathilaka, "State of the art in AI-based visual inspection for industrial quality control: Methods, benchmarks, challenges, and autonomous systems," *Electronics*, vol. 15, no. 12, Art. no. 2727, 2026.

[16] H. Yang, S. Kim, and Y. Choi, "Review of UAV–AI-based automated visual inspection technologies and their applicability to high-risk construction machinery," *Journal of the Korean Society of Mineral and Energy Resources Engineers*, vol. 62, no. 5, pp. 567–583, 2025.

[17] Y. Li, B. Xie, S. Guo, Y. Yang, and B. Xiao, "A survey of robustness and safety of 2D and 3D deep learning models against adversarial attacks," *ACM Computing Surveys*, vol. 56, no. 6, pp. 1–37, 2024.

[18] J. Wang, J. Ai, M. Lu, J. Zhao, H. Su, D. Yu, *et al.*, "A survey of neural network robustness assessment in image recognition," *ACM Computing Surveys*, vol. 58, no. 12, pp. 1–40, 2026.

[19] B. Chander, C. John, L. Warrier, and K. Gopalakrishnan, "Toward trustworthy artificial intelligence (TAI) in the context of explainability and robustness," *ACM Computing Surveys*, vol. 57, no. 6, pp. 1–49, 2025.

[20] N. Kemmerzell, A. Schreiner, H. Khalid, M. Schalk, and L. Bordoli, "Towards a better understanding of evaluating trustworthiness in AI systems," *ACM Computing Surveys*, vol. 57, no. 9, pp. 1–38, 2025.

[21] K. Nikiforidis, A. Kyrtsoglou, T. Vafeiadis, T. Kotsiopoulos, A. Nizamis, D. Ioannidis, *et al.*, "Enhancing transparency and trust in AI-powered manufacturing: A survey of explainable AI (XAI) applications in smart manufacturing in the era of Industry 4.0/5.0," *ICT Express*, vol. 11, no. 1, pp. 135–148, 2025.

[22] G. Yu, G. Tan, H. Huang, Z. Zhang, P. Chen, R. Natella, *et al.*, "A survey on failure analysis and fault injection in AI systems," *ACM Transactions on Software Engineering and Methodology*, vol. 35, no. 1, pp. 1–42, 2026.

[23] S. Bi, C. Wang, B. Wu, S. Hu, W. Huang, W. Ni, *et al.*, "A comprehensive survey on applications of AI technologies to failure analysis of industrial systems," *Engineering Failure Analysis*, vol. 148, Art. no. 107172, 2023.

[24] D. C. Lepcha, B. Goyal, A. Dogra, K. P. Sharma, and D. N. Gupta, "A deep journey into image enhancement: A survey of current and emerging trends," *Information Fusion*, vol. 93, pp. 36–76, 2023.

[25] Y. Lu, Y. Lin, H. Wu, Y. Luo, X. Zheng, H. Xiong, and L. Wang, "Priors in deep image restoration and enhancement: A survey," *arXiv preprint arXiv:2206.02070*, 2022.

[26] H. Su, Y. Li, Y. Xu, X. Fu, and S. Liu, "A review of deep-learning-based super-resolution: From methods to applications," *Pattern Recognition*, vol. 157, Art. no. 110935, 2025.

[27] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2018.

[28] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," arXiv:2004.10934, 2020.

[29] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[30] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, "Deep Long-Tailed Learning: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 2, pp. 1079–1099, 2024.

[31] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 2999–3007.

[32] 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. Artificial Intelligence*, vol. 34, no. 7, pp. 12993–13000, 2020.

[33] Y. Cui, M. Jia, T. Y. Lin, Y. Song, and S. Belongie, "Class-Balanced Loss Based on Effective Number of Samples," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9268–9277.

[34] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2015.

[35] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241.

[36] C. Ledig *et al.*, "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4681–4690.

SpaTempNet: Deep Spatiotemporal Forecasting of River Morphological Evolution of the Padma River in Bangladesh

S. S. Mahmud Turza[1], Md. Kaoser Ahamed Anik[2], Md. Shadmim Hasan Sifat[3],
Mahruba Sharmin Chowdhury[4], Khandokar Md. Rahat Hossain[5], Abdullah Al Noman[6]
Dept. of Computer Science and Engineering, Shahjalal University of Science and Technology, Sylhet, Bangladesh[1,2,3,4]
Dept. of Computer Science and Engineering, United International University, Dhaka, Bangladesh[5]
Computer Science and Engineering Discipline, Khulna University, Khulna 9208, Bangladesh[6]

***Abstract*—Rivers are dynamic geomorphological systems experiencing continuous transformation through erosion and sediment transport. In Bangladesh, these processes displace 50,000-200,000 people annually, causing severe losses along the Padma River according to Natural Resources Defense Council (NRDC). Traditional monitoring via field surveys and manual remote sensing proves costly, spatially limited, and unsuitable for multi-year prediction. This study presents a comprehensive framework, termed SpaTempNet, for forecasting river morphological evolution from freely available satellite data using spatiotemporal deep learning. A 38-year time-series (1987–2025) of binary water masks was constructed from Landsat (5 TM, 7 ETM+, 8 OLI) and Sentinel (Sentinel-2 MSI, Sentinel-1 SAR) imagery via Google Earth Engine. Water extraction used Modified Normalised Difference Water Index (MNDWI). The proposed Bidirectional ConvLSTM gap-filling model achieved IoU = 0.7366, outperforming classical methods. Five spatiotemporal architectures (ConvLSTM, U-Net+LSTM, Attention U-Net+ConvLSTM, Swin Transformer, ViT-based model) were evaluated across yearly, quarterly, and bi-monthly resolutions. Hybrid CNN-LSTM models consistently outperformed pure transformers. Attention U-Net+ConvLSTM achieved best yearly performance (IoU = 0.7005); U-Net+LSTM led bi-monthly prediction (IoU = 0.7791). Statistical analysis quantified mean annual migration of 255.4 m yr^{-1} with 1998 extreme of 1,485.5 m yr^{-1}. An expansion of about 290 km^2 was projected, and a spatially explicit risk map was generated through long-term forecasting (2026-2040). Results demonstrate that freely available satellite imagery with deep learning can provide practical, scalable framework for riverbank hazard monitoring and disaster management.**

***Keywords*—River morphology; riverbank erosion forecasting; spatiotemporal neural networks; ConvLSTM; U-Net; Swin Transformer; Vision Transformer; satellite remote sensing; MNDWI; disaster management**

I. Introduction

Rivers are one of the most morphologically dynamic features that characterize the terrestrial surface of the earth. Because of continuous fluvial erosion, sediment transport, and depositional accretion, they undergo constant geomorphological reshaping. These processes occur in broad spatial and temporal scales, from daily fluctuations during floods to decadal shifts in channel position, planform geometry, and cross-sectional morphology. Comprehensive understanding and reliable predictive capability regarding these morphodynamic transformations are fundamental to effectively integrate water resource management, safeguard human settlements and infrastructure, maintain agricultural productivity in riparian zones, and preserve critical riparian and aquatic ecosystem integrity and biodiversity [1].

Bangladesh stands on a risky geomorphological and hydrological position within this global context. It occupies the terminal deltaic sink of the integrated Ganges-Brahmaputra-Meghna river system, which is regarded as one of Earth's most extensive active delta complexes. Its river network shows exceptional suspended sediment loads, which exceeds 1 billion tonnes annually, and intense seasonal flooding- driven by South Asian monsoon meteorological forcing [2], [3]. The region experiences some of the highest lateral channel migration rates recorded globally, often exceeds several hundred metres per monsoon season and occasionally reaches kilometre-scale adjustments. These extreme morphodynamic processes cause persistent and severe riverbank erosion that systematically displaces between 50,000 and 200,000 individuals annually [4]. The consequences are devastating cumulative economic losses through destruction of agricultural land, transportation infrastructure (roads, bridges, ferries), flood protection embankments, educational facilities, and healthcare infrastructure. Beyond physical infrastructure, erosion destroys the fundamental social fabric binding riverine communities. The socioeconomic consequences extend far beyond immediate physical displacement. Affected populations face long-term food insecurity, disrupted educational access for children, dissolution of social networks and community structures, and profoundly constrained livelihood trajectories and economic opportunities [3], [5], [6]. The most economically vulnerable households bear disproportionate burdens, lacking resources for adaptation or relocation.

Despite more than five decades of continuously accumulated satellite remote sensing imagery archives now being comprehensively available through multiple international space agencies and data repositories, these invaluable spatiotemporal datasets remain substantially under-exploited for predictive morphological modelling applications beyond retrospective change detection. The dominant research paradigm characterizing existing peer-reviewed literature emphasizes retrospective change detection methodologies: quantitative comparison of discrete temporal snapshots to characterize and quantify morphological transformations that have already occurred, rather than developing prospective forecasting capabilities to predict

transformations that will occur. Classical remote sensing analytical approaches inherently overlook the continuous temporal evolution trajectories and complex spatial dependencies intrinsically characterizing river morphodynamic processes. Recently some transformative breakthroughs are seen in deep spatiotemporal learning architectures, particularly in computational frameworks. Domains include convolutional spatial feature extraction and recurrent temporal memory mechanisms. These have demonstrated unprecedented success in learning complex sequential patterns from high-dimensional image sequence data [7], [8]. However, systematic application of these powerful architectures specifically to multi-year river morphology spatial forecasting remains comparatively sparse within published literature. Also, comprehensive comparative evaluation rigorously assessing multiple competing architectures for this specific forecasting task under controlled experimental conditions is rare.

Based on the existing literature, several critical research gaps motivate this investigation. First, majority of the existing research focuses on historical change detection rather than forecasting future morphological states. Second, although deep learning has been applied to water body segmentation and streamflow prediction, its application to multi-year spatial river morphology forecasting remains limited. Third, no systematic cross-architecture comparison for this specific forecasting task has been published. Fourth, physical geomorphological and socioeconomic perspectives are rarely integrated in unified predictive frameworks for operational disaster management.

To address these gaps, this study introduces SpaTempNet, an integrated spatiotemporal forecasting pipeline that: (1) employs sequence learning considering the continuous nature of river formation as a spatiotemporal process, using multiple years' windows to learn decadal changes; (2) performs a comparison of five architectures covering CNN+LSTM architectures and transformers under identical experimental settings; (3) extends the forecast horizon to 15 years for infrastructural planning and disaster preparedness; (4) incorporates an attention mechanism for automatic detection of morphologically dynamic regions; and (5) provides a base architecture offering future scope for inclusion of socioeconomic risk analysis for practical forecasting purposes.

II. Related Works

Riverbank erosion monitoring has evolved significantly with satellite remote sensing technology. Previously, researchers used to rely on aerial photography and manual channel boundary digitization. After five decades of continuous observation, the Landsat mission series revolutionized the field by enabling systematic, large-area, long-term monitoring at unprecedented scales.

A. Optical Satellite-Based Monitoring

Binh et al. [9] developed the Sentinel-based River Bank Erosion Detection (SRBED) framework for automated riverbank extraction from Landsat time-series. It demonstrated great potential for large-scale morphological change detection. Langhorst et al. [10] extended this globally. They constructed algorithms detecting lateral bank movement across diverse river systems and identified erosion hotspots at continental scales. These contributions established critical baseline erosion rates, which revealed extraordinary severity of riverbank retreat in South and Southeast Asia. Langat et al. [11] and Islam and Haque [12] demonstrated multi-temporal remote sensing with GIS for quantifying channel migration trends in Bangladesh. Raju et al. [13] analyzed erosion-accretion dynamics, and highlighted heterogeneous distribution of geomorphically active zones along river reaches where erosion alternates spatially with stable and accretionary zones.

B. Synthetic Aperture Radar Applications

Optical remote sensing in tropical environments suffers from persistent cloud cover, particularly during monsoon seasons when erosion peaks. Synthetic Aperture Radar (SAR) provides cloud-independent all-weather observation. Freihardt et al. [2] used Sentinel-1 SAR time series in Google Earth Engine to map riverbank erosion in Bangladesh. They documented rates exceeding 100 m/year and validated SAR water masking using VV backscatter threshold of -16 dB. DeVries et al. [23] demonstrated that combining Sentinel-1 and Landsat in Google Earth Engine enables rapid flood mapping scalable to continental extents. Sazib et al. [26] extended SAR-based mapping across the Ganga-Brahmaputra basin, and linked inundation extents with agricultural exposure and population impact assessments.

C. Regional Case Studies

Several studies focused on specific river systems. Tha et al. [14] evaluated riverbank erosion hotspots along Cambodia's Mekong River. They integrated physical erosion analysis with vulnerability assessments and connected geophysical processes with socioeconomic consequences. Saha [15] conducted multi-sensor analysis at the Padma-Jamuna confluence. They documented complex erosion, deposition, and island formation patterns from junction dynamics. Haque [16] studied multi-decadal Jamuna River changes and quantified agricultural land losses and settlement displacement. Ritu et al. [17] modeled Padma River bank shifting, however, 10-year intervals between acquisitions limited prediction accuracy and seasonal sensitivity. Muzahid et al. [18] quantified erosion and accretion along the Gorai River using multi-temporal Landsat classification.

D. Advanced Segmentation and Risk Assessment

Ren [19] developed large-scale erosion risk assessment, combined remote sensing with topographic, hydrological, and land-cover variables. A fundamental limitation shared across these studies is the absence of a multi-year predictive modeling grounded in spatiotemporal deep learning.

E. Socioeconomic Dimensions

Human dimensions of riverbank erosion provide strong motivation for improved predictive tools. Islam et al. [4] documented displacement patterns and economic losses. They showed erosion-induced migration drives rapid urbanization and informal settlement growth. Das et al. [5] documented psychological trauma from displacement and livelihood reconstruction impairment. They noted that the effects include not only economic losses but also damage to social ties and community strength. Alam et al. [3] and Sarkar et al. [6]

examined livelihood vulnerability among riparian households along the Padma and in northern Bangladesh. Their research revealed that poorest households face higher rate of exposure due to limited adaptive capacity and land-based livelihood dependence.

While qualitative vulnerability assessments help us see the human side of erosion, quantitative forecasting tools are necessary to give early warnings that people can act on. Some recent progress in deep learning and remote sensing now offers great chances to create these predictive systems. Below, we look at tech advances in water body segmentation, cloud-based processing, and spatiotemporal deep learning that form the base for today's river shape forecasting systems.

F. Deep Learning for Water Body Segmentation

Recent deep learning advances have replaced traditional threshold-based water extraction methods. Yuan et al. [20] demonstrated CNNs achieve superior accuracy over NDWI/MNDWI baselines across scenes with mixed land cover, shadows, and variable turbidity. Li et al. [21] proposed attention-enhanced multi-scale residual U-Net for water segmentation from Sentinel-2 imagery, Which showed attention mechanisms effectively discriminate water from spectrally similar classes like shadows and saturated soils. Ghosh et al. [22] advanced multi-scale fusion within U-Net architectures, They achieved enhanced detection and improved edge definition for rivers of varying dimensions. These architectures process individual images without temporal awareness and cannot forecast morphological change.

G. Cloud-Based Processing Infrastructure

Google Earth Engine (GEE) has made large-scale satellite processing accessible to all through operations on multi-decadal archives without local download. DeVries et al. [23] established platform suitability for operational flood monitoring. Clement et al. [24] applied GEE to ephemeral river flooding. Tripathy and Mallick [25] extended it to large-scale Mekong River monitoring, demonstrating continental scalability.

H. Recurrent and Hybrid Networks for Hydrological Forecasting

Fan et al. [7] applied LSTM to forecast river discharge and demonstrated that temporal relationships in hydrological data can be learned to generate superior predictions with autoregressive models. Le et al. [27] and Xiang et al. [28] validated hybrid CNN-LSTM frameworks for river flow prediction. Their research showed complementary gains from combining spatial feature extraction with temporal memory. Horvath et al. [29] highlighted importance of variable selection and temporal lag configuration for LSTM water level prediction on the Tisza River. Li et al. [21] demonstrated the value of attention mechanisms in CNN-LSTM models for streamflow forecasting across the Qinghai-Tibet Plateau.

I. ConvLSTM and Spatiotemporal Architectures

ConvLSTM architectures replace matrix multiplications in recurrent gates with convolutions which preserve spatial locality during temporal processing. Dehghani et al. [8] provided systematic comparative evaluation of LSTM, CNN, and ConvLSTM for spatiotemporal forecasting. They found ConvLSTM superior for tasks with strong spatial dependence without increasing computational cost. Lemenkova [1] applied deep learning, including convolutional-recurrent architectures, to track flood dynamics in the Ganges Delta. Their work demonstrated applicability to water extent data in the same geographical context. For transformer-based video understanding, Liu et al. [30] proposed Video Swin Transformer with hierarchical window-based self-attention. On the other hand Dosovitskiy et al. [31] established the ViT framework and Arnab et al. [32] extended it to video (ViViT). These researches provided architectural foundations for spatiotemporal transformer models evaluated here.

III. Study Area and Dataset Construction

A. Study Area: Padma River Mid-Course Reach

The designated study area (Fig. 1) covers a long mid-course section of the Padma River. It includes the active river channel, nearby floodplain zones, and the large islands and sandbar formations found in the middle of the river. The spatial region of interest was strategically chosen so as to incorporate effectively all possible patterns of morphological evolution and channel migration behavior within an analytically feasible spatial domain fit for training deep learning models intensively.

The Padma River forms through confluence of the Ganges (locally termed Padma upstream of junction) and Jamuna Rivers at Goalundo Ghat and flows approximately 120 km in southeastward direction before ultimately joining the Meghna River to discharge into the Bay of Bengal through multiple distributary channels. The river annually transports multiple millions of tonnes of Himalayan-derived suspended and bed-load sediment and exhibits characteristically highly dynamic morphodynamic behaviour including intense lateral channel migration, extreme bank erosion rates frequently exceeding 500 m during individual monsoon seasons, and regular mid-channel bar development, modification, and dissolution cycles. The discharge regime exhibits pronounced monsoonal characteristics: peak flow occurs during July–September monsoon season, during which the river maintains a complex braided planform geometry with rapidly changing channel configurations and extensive mid-channel island features. During dry season months (January–April), flow contracts substantially to form much narrower single-thread or limited multi-thread channel configurations with extensive exposed bar surfaces. The live river corridor provides high riparian population density on both bankside, with the river providing various socioeconomic roles including transportation for people and goods, irrigation waters for agriculture, artisan fishery for the riparian population, and clean water for domestic use. However, in spite of the above vital socioeconomic role, very high rates of erosion have resulted in loss of productive agricultural lands, forced displacement of entire households from their ancestral homelands, and complete destruction of vital infrastructure such as roads, bridges, ferry landing points, educational facilities, and health services facilities [3].

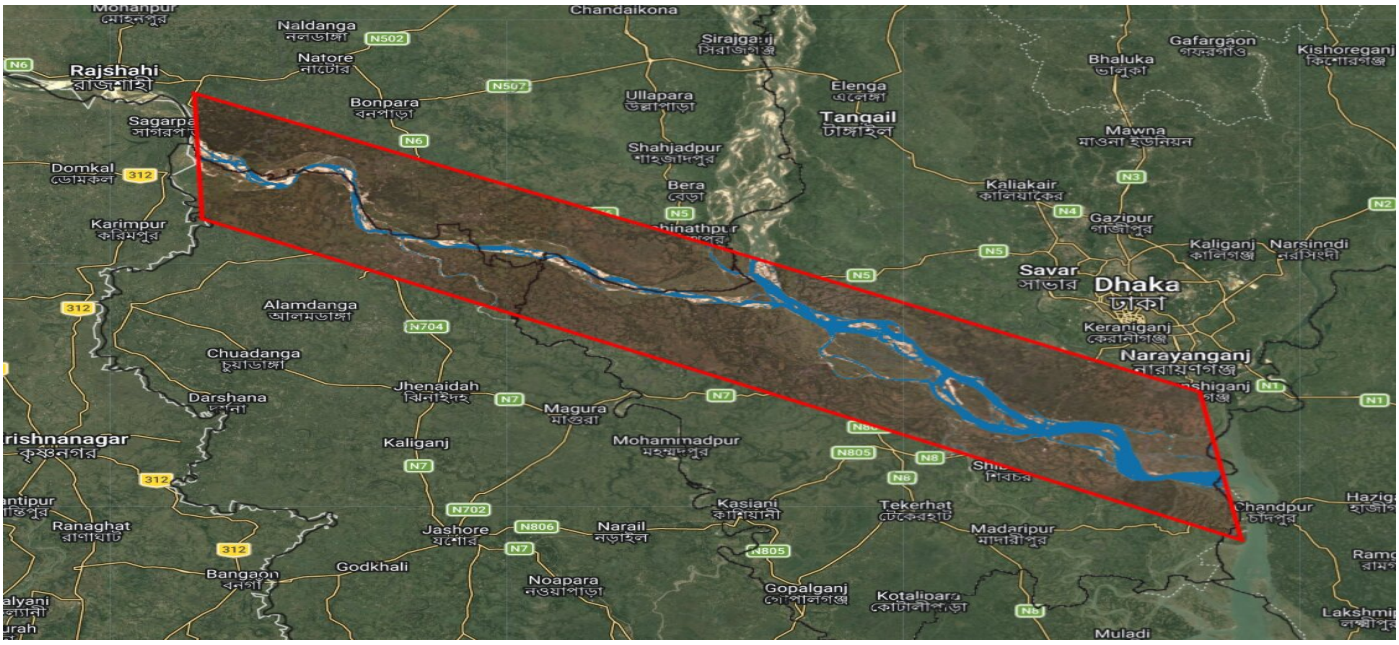


Fig. 1. Study area delineation along the mid-course section of the Padma River, Bangladesh.

B. Satellite Data Acquisition Protocol

Satellite imagery acquisition was strategically organized into two temporally distinct operational phases based on systematic sensor availability patterns and mission operational timelines.

1) Phase I: Landsat Mission Data (1987–2014).: Multispectral optical imagery was sourced exclusively from United States Geological Survey (USGS) Landsat Collection 2 Level-2 atmospherically-corrected surface reflectance product archives accessed through Google Earth Engine Application Programming Interface (API). Three successive sensor platforms were employed sequentially following their respective operational timelines: Landsat 5 Thematic Mapper (TM) sensor (GEE collection identifier: `LANDSAT/LT05/C02/T1_L2`) operational prior to 1999; Landsat 7 Enhanced Thematic Mapper Plus (ETM+) sensor (`LANDSAT/LE07/C02/T1_L2`) operational from 1999 onward; and Landsat 8 Operational Land Imager (OLI) sensor (`LANDSAT/LC08/C02/T1_L2`) operational from 2013 onward. All three sensors provide systematically consistent 60 m spatial resolution imagery (resampled from native 60 m for consistency with SWIR bands) with 16-day temporal revisit frequency. Scene-level cloud filtering applied maximum permissible cloud fraction threshold of 50% via `CLOUD_COVER` image-level metadata attribute. Pixel-level cloud and cloud-shadow masking protocols utilized `QA_PIXEL` quality assessment band via bitwise logical operations, specifically targeting bit 3 (cloud shadow detection) and bit 4 (cloud detection) binary flags. Landsat Collection 2 Level-2 products provide surface reflectance as scaled integers; conversion to physical surface reflectance values employed standardized scaling relationship:

$$\rho_\lambda = 0.0000275 \times \mathrm{DN}_\lambda - 0.2 \quad (1)$$

where, ρ_λ represents surface reflectance for spectral band λ and DN_λ represents digital number value. Spectral band selections for MNDWI computation varied systematically by sensor platform following band designation conventions: Band 2 (green, 520–600 nm) and Band 5 (SWIR1, 1550–1750 nm) for Landsat 5 and Landsat 7 platforms; Band 3 (green) and Band 6 (SWIR1) for Landsat 8 platform. Temporal composite imagery was systematically generated for three distinct temporal aggregation schemes: yearly (single annual composite per calendar year), quarterly (four three-month period composites per calendar year: Q1 January–March, Q2 April–June, Q3 July–September, Q4 October–December), and bi-monthly (six two-month period composites per calendar year: P1 January–February, P2 March–April, P3 May–June, P4 July–August, P5 September–October, P6 November–December).

2) Phase II: Sentinel Mission Data (2015–2025).: Sentinel-2 MultiSpectral Instrument (MSI) optical data were systematically accessed via Google Earth Engine collection identifier `COPERNICUS/S2_SR_HARMONIZED`, representing harmonized Level-2A bottom-of-atmosphere reflectance products. Scene-level filtering applied maximum cloud cover threshold of 50% based on `CLOUDY_PIXEL_PERCENTAGE` metadata attribute, with scenes subsequently sorted in ascending order of cloud fraction to prioritize least contaminated observations. To prevent computational timeout errors during large-scale export operations, maximum number of Sentinel-2 scenes utilized per temporal composite was algorithmically capped at 25 images. Pixel-level cloud masking protocols utilized QA60 quality assessment band targeting bit 10 (opaque cloud detection) and bit 11 (cirrus cloud detection) binary flags. Sentinel-2 MSI products provide surface reflectance as physical values requiring no additional radiometric scaling transformations.

To systematically address persistent data gap patterns arising from sustained cloud cover contamination particularly pronounced during monsoon seasons, Sentinel-1 Synthetic Aper-

TABLE I. SYSTEMATIC SUMMARY OF IMAGE AVAILABILITY AND QUALITY ASSESSMENT ACROSS MULTIPLE TEMPORAL RESOLUTION SCHEMES.

Temporal Resolution	Usable Images	Missing Images	Partial/ Low-Quality Images
Monthly	284	50	122
Bi-monthly	156	16	56
Quarterly	125	3	24
Yearly	38	0	0

ture Radar (SAR) C-band data were strategically incorporated as supplementary fallback data source activated when Sentinel-2 optical imagery proved unavailable. Sentinel-1 data were systematically accessed from Google Earth Engine collection `COPERNICUS/S1_GRD` representing Ground Range Detected (GRD) products acquired in Interferometric Wide (IW) swath mode with VV and VH dual polarization configuration in descending orbital pass geometry. Scene selection was sorted by acquisition date in descending temporal order and algorithmically limited to maximum of 20 images per temporal composite to manage computational processing load. The SAR-based fallback protocol was activated exclusively when zero Sentinel-2 scenes successfully passed cloud filtering criteria for specified temporal window. In this scenario, water mask generation involved using the backscatter coefficient of VV polarization, using an empirical fixed threshold of -16 dB, as per the protocol [2].

Systematically developed water masks were saved as georeferenced GeoTIFF files, having a geographic coordinate system EPSG:4326 (datum WGS84). The spatial resolution of 60 m was always used as a standard for all platforms, while falling back on bounding box size in the exceptional case of computation resource exhaustion.

C. Image Availability and Quality Assessment

Table I provides systematic summary of image availability characteristics and quality assessment metrics evaluated across four distinct temporal resolution aggregation schemes. The yearly dataset achieved complete temporal coverage (38 observations spanning 1988–2025, zero missing observations), thereby providing most stable and reliable foundation for long-term temporal trend analysis applications. The monthly resolution dataset exhibited most severe data loss patterns, with 50 completely missing temporal observations and 122 partial or low-quality degraded observations, thereby precluding its systematic utilization for primary modelling experiments. Bi-monthly and quarterly datasets demonstrated progressively improving completeness characteristics, with quarterly temporal resolution providing optimal balance between temporal granularity and systematic data reliability (only 3 completely missing observations across 38-year period).

Subsequent comprehensive manual inspection procedures complemented automated filtering protocols to systematically identify partially missing images exhibiting localized spatial gaps or degraded regions not adequately captured by cloud-fraction threshold criteria alone (Fig. 2). These quality impairment patterns proved most prevalent within bi-monthly dataset during July–August temporal periods (peak monsoon season) when persistent cloud cover systematically impedes optical observation capabilities.

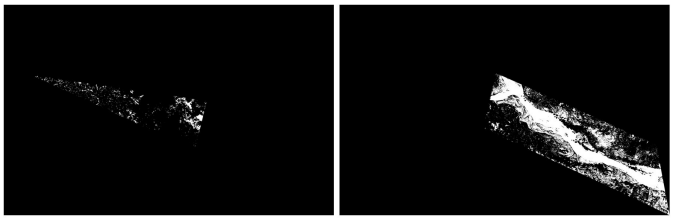


Fig. 2. Partially missing samples

D. Water Body Detection via Modified Normalized Difference Water Index

Water body extraction procedures employed Modified Normalised Difference Water Index (MNDWI) spectral index calculations [2], [14], which systematically exploit contrasting spectral reflectance behaviour characteristics of water surfaces (characteristically high reflectance in green visible wavelengths, strong absorption in shortwave infrared wavelengths) relative to terrestrial surfaces including soils, vegetation canopies, and built-up urban areas (producing characteristically low or negative MNDWI values). The MNDWI spectral index is formally defined:

$$\mathrm{MNDWI} = \frac{\rho_{\mathrm{Green}} - \rho_{\mathrm{SWIR}}}{\rho_{\mathrm{Green}} + \rho_{\mathrm{SWIR}}} \tag{2}$$

where, ρ_{Green} and ρ_{SWIR} represent surface reflectance values in green visible and shortwave infrared spectral bands, respectively. A fixed threshold value of zero was systematically applied to generate binary water/non-water classification masks from continuous MNDWI index values.

Following initial thresholding operations, connected component analysis algorithms were systematically applied to identify spatially contiguous water body regions. Only three largest spatially connected components per image were systematically retained, thereby eliminating isolated water pixels corresponding to small ponds, minor drainage channels, and ephemeral water features unrelated to primary Padma River channel. This connected component filtering substantially improves signal quality by removing spurious detections. All processed images were subsequently standardized via bilinear resampling to consistent 256×256 pixel spatial dimensions suitable for deep learning model input tensor specifications.

E. Gap-Filling Framework and Performance Benchmark

In order to effectively deal with natural patterns that will arise due to cloud contamination and instrument failure, a rigorous performance evaluation of gap filling methods was performed on a set of yearly water mask temporal series (a total of 38 observations from 1988 to 2025). For the purpose of simulating practical missing data problems, a systematic withholding process was used to remove 20% of the total number of observations (i.e., seven specific years: 2002, 2005, 2010, 2015, 2019, 2020, 2024).

There were seven unique strategies used to fill the gaps that underwent comparative analysis:

1) Temporal mean: The reference baseline strategy based on the use of temporal means to estimate the masks that need to be reconstructed, offering naive but effective and efficient spatial reconstructions of those masks.
2) Temporal median: The alternative reference baseline strategy using the median approach with enhanced robustness against outlier observation.
3) Linear temporal interpolation: Per-pixel linear interpolation computed between nearest available temporal neighbouring observations, yielding smooth temporal transition characteristics but exhibiting limited structural preservation capabilities for complex spatial features.
4) Cubic spline temporal interpolation: Per-pixel cubic spline interpolation applied across temporal dimension, enabling smooth nonlinear pixel trajectory modelling with continuous first and second derivatives providing enhanced temporal smoothness relative to linear interpolation.
5) Signed Distance Field (SDF) interpolation: Geometrically motivated approach wherein binary water masks are first transformed into signed distance field (SDF) representations defined as:

$$\mathrm{SDF}(\mathbf{x}) = d(\mathbf{x}, W^c) - d(\mathbf{x}, W) \tag{3}$$

where,

$$d(\mathbf{x}, A) = \min_{\mathbf{y} \in A} \|\mathbf{x} - \mathbf{y}\| \tag{4}$$

W denotes the water region, W^c represents its complement, $d(\mathbf{x}, A)$ is the Euclidean distance from location $\mathbf{x}$ to the set A, and $\|\cdot\|$ denotes the Euclidean norm. Positive SDF values correspond to pixels inside the water region, whereas negative values correspond to pixels outside the water region. Temporal interpolation is performed in the continuous SDF domain, followed by thresholding at the zero level set to recover binary water masks. This approach preserves boundary geometry and topological consistency more effectively than direct interpolation of binary masks.
6) Inverse-distance weighted temporal fusion: Missing mask reconstructions are computed through temporally weighted averaging of all available yearly observations, with weights inversely proportional to their temporal distance from the target year:

$$M_t = \frac{\sum_{i \in \mathcal{A}} w_i M_i}{\sum_{i \in \mathcal{A}} w_i}, \qquad w_i = \frac{1}{|t - i|} \tag{5}$$

where, M_t denotes the reconstructed mask at target year t, M_i represents an available observation at year i, and $\mathcal{A}$ denotes the set of all available observations. Consequently, observations temporally closer to the target year exert greater influence on the reconstruction than more distant observations.
7) Bidirectional Convolutional LSTM (BiConvLSTM): Deep learning-based approach wherein Bidirectional Convolutional LSTM architecture is systematically trained on 128×128 pixel spatiotemporal patches extracted randomly from full temporal series. Training protocol incorporates 20% random temporal masking simulation to replicate missing observation patterns. The model architecture learns both forward and backward temporal dependency patterns and was optimized using binary cross-entropy loss over maximum 100 training epochs with early stopping (patience = 15). Adam optimizer ($\eta = 10^{-3}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$) with batch size 16 was employed.

TABLE II. COMPREHENSIVE PERFORMANCE COMPARISON OF SEVEN GAP-FILLING METHODS EVALUATED ON SEVEN HELD-OUT YEARLY OBSERVATIONS.

Method	IoU	Dice	Precision	Recall
Mean Composite	0.6573	0.7926	0.8535	0.7403
Median Composite	0.6573	0.7926	0.8535	0.7403
Linear Interp.	0.6548	0.7910	0.8913	0.7193
Spline Interp.	0.6733	0.8042	0.8082	0.8012
SDF	0.7290	0.8429	0.8741	0.8161
Weighted Temporal	0.6724	0.8035	0.8306	0.7790
BiConvLSTM	**0.7366**	**0.8477**	0.8457	**0.8511**

Quantitative performance results are comprehensively summarized in Table II. BiConvLSTM achieved best overall performance across all evaluation metrics (IoU = 0.7366, Dice = 0.8477, Recall = 0.8511), conclusively confirming its superior capability to capture extraordinarily complex spatiotemporal dynamic patterns. Among classical non-learning methods, SDF interpolation performed strongest (IoU = 0.7290, Dice = 0.8429), closely approaching deep learning performance levels while maintaining excellent structural consistency of water body boundary characteristics. Standard mean/median temporal composites and linear interpolation exhibited comparable baseline performance (IoU $\approx$ 0.655), while spline interpolation and weighted temporal fusion strategies offered modest incremental improvements.

IV. METHODOLOGY

This section of the study offers an in-depth technical explanation of all the methodological elements, such as sequence generation, temporal splitting, model architecture description with mathematical formulae, loss functions, metrics to evaluate the model, training parameters, and risk zone computation.

A. Supervised Sequence Generation Strategy

To formally formulate temporal morphological evolution process as supervised learning task, preprocessed binary water masks were represented as sequences of structured input-output pairs applying the technique of sliding temporal window. For every considered sliding window size $L \in \{4, 5, 6\}$ (with respect to yearly and quarterly temporal scales) or $L \in \{6, 9\}$ (with respect to bi-monthly temporal scale), precisely L sequential observations constitute input features tensor while next chronological time-step represents prediction target:

$$\mathbf{X}_i = (X_i, X_{i+1}, \ldots, X_{i+L-1}), \quad y_i = X_{i+L} \tag{6}$$

where, $X_t \in \{0,1\}^{H\times W}$ represents binary water mask at time t with spatial dimensions $H \times W = 256 \times 256$ pixels. Temporal stride of one year (or one bi-monthly period for higher-frequency data) generates systematically overlapping sequences, maximizing utilization of available but limited temporal data. This multi-sequence evaluation protocol enables systematic identification of optimal temporal context window length for each distinct architecture and temporal resolution combination.

B. Rigorous Temporal Splitting Strategy

A meticulously rigorous temporal split strategy prevents any data leakage between training and evaluation datasets. Two experimental evaluation setups are formally defined.

1) Setup 1 (Long-range forecasting): Training incorporates all available data up to and including calendar year 2015; testing evaluates forecast performance on 2016–2025 inclusive (providing 10-year forecasting horizon for yearly data, or equivalent temporal span for quarterly/bi-monthly aggregations).

2) Setup 2 (Medium-range forecasting): Training incorporates all available data up to and including calendar year 2020; testing evaluates forecast performance on 2021–2025 inclusive (providing 5-year forecasting horizon for yearly data, or equivalent temporal span for quarterly/bi-monthly aggregations).

For quarterly temporal resolution data, temporal cutoffs correspond to Q4-2015 and Q4-2020, respectively. For bi-monthly resolution, cutoffs correspond to P6-2015 and P6-2020 (November–December 2015 and 2020). Within each setup configuration, final 15% of training sequences (ordered chronologically) are systematically allocated to validation dataset for hyperparameter tuning and early stopping criteria. Critically, absolutely zero temporal overlap exists between training, validation, and testing datasets, ensuring evaluation integrity.

C. Model Architecture Specifications

Five distinct spatiotemporal deep learning architectures underwent systematic comparative evaluation, selected to comprehensively span the architectural spectrum from recurrent convolutional hybrids to pure attention-based transformer models.

1) Standalone ConvLSTM architecture: ConvLSTM framework [8] primarily involves the replacement of traditional matrix multiplications in LSTMs with spatial convolutions. This allows inherent joint processing of spatial and temporal information without any need for a dedicated encoder module. The overall architecture includes:

a) Initial spatial feature extraction: Two successive spatial Conv2D layers with 32 and 64 convolutional filters (kernel size 3×3, ReLU activation, same padding) extract initial low-level spatial features.

b) Spatial downsampling: Batch normalization followed by max-pooling (2×2 window, stride 2) and spatial dropout (rate = 0.2) applied after each Conv2D layer.

c) Spatiotemporal processing: Two stacked ConvLSTM2D layers with 128 and 64 filters respectively (kernel size 3×3). The first ConvLSTM2D layer processes full input sequence with `return_sequences=True`; the second layer aggregates temporal information with `return_sequences=False`, outputting single feature map.

d) Upsampling decoder: Symmetric decoder comprising two upsampling blocks (UpSampling2D with factor 2 followed by Conv2D with 64 and 32 filters), batch normalization, and final 1×1 Conv2D with sigmoid activation producing prediction probability map.

2) U-Net with LSTM architecture: The U-Net+LSTM hybrid model [27], [28] processes input sequence through TimeDistributed encoder extracting multi-scale spatial features independently for each temporal frame. Complete architecture specification:

a) Timedistributed encoder: Three successive encoder blocks each comprising: Conv2D layer (filters: 64, 128, 256 for blocks 1–3; kernel 3×3, ReLU), batch normalization, second Conv2D with identical specifications, max-pooling (2×2), spatial dropout (rate = 0.2). TimeDistributed wrapper ensures identical processing applied to each input frame independently.

b) Bottleneck temporal fusion: LSTM module (128 units) at encoder bottleneck models temporal dependencies across sequence, producing single temporally-fused feature volume.

c) Symmetric decoder with skip connections: Three decoder blocks each comprising: UpSampling2D (factor 2), concatenation with corresponding encoder skip connection feature map, two Conv2D layers (filters decreasing: 256, 128, 64), batch normalization. Final 1×1 Conv2D with sigmoid activation produces prediction.

3) Attention U-Net with ConvLSTM architecture: The Attention U-Net+ConvLSTM model extends U-Net+LSTM through incorporating sophisticated attention gate mechanisms in each skip connection pathway [21], [22]. For encoder feature map $\mathbf{x}^l \in \mathbb{R}^{H_l\times W_l\times C_l}$ at encoder level l and gating signal $\mathbf{g} \in \mathbb{R}^{H_l\times W_l\times C_g}$ derived from corresponding decoder level, attention coefficient computation proceeds via:

$$\mathbf{q}^l = W_x * \mathbf{x}^l + W_g * \mathbf{g} + b_g \tag{7}$$

$$\boldsymbol{\alpha}^l = \sigma\big(W_\psi * \mathrm{ReLU}(\mathbf{q}^l) + b_\psi\big) \tag{8}$$

where, $*$ denotes 1×1 convolution operation, W_x, W_g, W_ψ represent learnable convolutional kernel parameters, b_g, b_ψ are bias terms, and σ denotes sigmoid activation function producing spatial attention coefficients $\boldsymbol{\alpha}^l \in [0,1]^{H_l\times W_l}$. The skip connection output becomes element-wise product:

$$\tilde{\mathbf{x}}^l = \boldsymbol{\alpha}^l \odot \mathbf{x}^l \tag{9}$$

thereby suppressing background regions (low α values) and amplifying morphologically active zones including channel margins and active erosion fronts (high α values). Critically, the bottleneck employs ConvLSTM2D rather than standard LSTM, preserving spatial locality during temporal fusion operations.

4) Swin Spatio-Temporal Transformer architecture: The Swin Spatio-Temporal Transformer [30] applies shared hierarchical Swin Transformer encoder to each input frame independently. The complete architecture specification:

a) Shared swin encoder: Each stage applies window-based multi-head self-attention with shifted window partitioning strategy (W-MSA and SW-MSA) followed by patch merging for progressive spatial downsampling. Three encoder stages produce feature maps at spatial resolutions $H/4 \times W/4$, $H/8 \times W/8$, and $H/16 \times W/16$ with embedding dimensions 64, 128, and 256, respectively. Self-attention within each window $\mathcal{W}$ of size $M \times M$ is computed as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{SoftMax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}} + \mathbf{B}\right)\mathbf{V} \quad (10)$$

where, $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ represent query, key, value projections, d is head dimension, and $\mathbf{B}$ encodes relative position biases. Shifted windowing across successive layers enables cross-window connections.

b) Temporal fusion module: Features at finest encoder resolution across all T input frames are stacked forming spatiotemporal tensor $\mathbf{F} \in \mathbb{R}^{T\times(H/4)\times(W/4)\times 64}$. Temporal fusion comprises:

- 1D temporal convolution (kernel size 3, GELU activation) applied across temporal dimension at each spatial location.
- Two cross-temporal attention layers (4 heads each) enabling each spatial location to attend across all temporal frames.
- Softmax-based temporal aggregation producing single fused feature map $\bar{\mathbf{F}} \in \mathbb{R}^{(H/4)\times(W/4)\times 64}$.

c) U-Net style decoder: Symmetric decoder with skip connections from encoder stages 1 and 2 progressively upsamples to full resolution.

5) Vision Transformer-based spatio-temporal model: The ViT-based model [31], [32] applies shared Vision Transformer encoder to each frame independently. Complete specification:

a) Patch embedding: Each frame is partitioned into non-overlapping 16×16 patches yielding $(256/16)^2 = 256$ patch tokens. Linear projection maps each flattened patch to 128-dimensional embedding space:

$$\mathbf{z}_p = \mathbf{E} \cdot \text{vec}(\mathbf{x}_p) + \mathbf{e}_{\text{pos}} \quad (11)$$

where, $\mathbf{E} \in \mathbb{R}^{128\times(16\times16)}$ represents learnable projection matrix, $\mathbf{e}_{\text{pos}}$ encodes spatial position.

b) Spatial transformer blocks: Two successive transformer blocks (multi-head self-attention with 4 heads, MLP with hidden dimension 256, LayerNorm, residual connections, dropout 0.1) process patch token sequence.

c) Temporal aggregation: Global average pooling aggregates all patch tokens yielding per-frame feature vector $\mathbf{f}_t \in \mathbb{R}^{128}$. Frame features are stacked, augmented with learned temporal positional embeddings, and processed by two temporal transformer layers (4 heads, MLP dimension 256) modeling inter-frame dependencies.

d) Convolutional decoder: Aggregated temporal feature is spatially projected to $16\times16\times128$ feature map, then decoded through four transposed convolution blocks progressively upsampling to full resolution.

D. Loss Function Specifications

All models underwent training optimization using combined Dice and Focal loss formulation:

$$\mathcal{L}_{\text{Total}} = \lambda_1\mathcal{L}_{\text{Dice}} + \lambda_2\mathcal{L}_{\text{Focal}} \quad (12)$$

where, weighting coefficients $\lambda_1 = \lambda_2 = 0.5$ provide equal contribution.

1) Dice loss: directly optimizes Dice overlap metric and proves particularly effective for class-imbalanced segmentation where water pixels constitute spatial minority:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_i p_i g_i + \varepsilon}{\sum_i p_i + \sum_i g_i + \varepsilon} \quad (13)$$

where, p_i represents predicted probability for pixel i, $g_i \in \{0, 1\}$ represents ground truth label, and $\varepsilon = 1$ provides numerical stability smoothing.

2) Focal loss: systematically down-weights easily-classified background examples to concentrate gradient updates on challenging boundary pixels [33]:

$$\mathcal{L}_{\text{Focal}} = -\frac{1}{N}\sum_{i=1}^{N}\alpha_i(1 - p_{t_i})^\gamma \log(p_{t_i}) \quad (14)$$

where, $p_{t_i} = p_i$ if $g_i = 1$ else $1 - p_i$, class weighting $\alpha_i = 0.75$ for positive class, focusing parameter $\gamma = 2.0$, and N represents total pixel count. The loss functions work together by utilizing each other's strengths: Dice loss guarantees accuracy with regards to overall overlap, Focal loss supplies accurate supervision on the level of boundaries.

E. Comprehensive Evaluation Metrics

The model was evaluated with five complementary quantitative metrics using the binarization threshold value of 0.5:

Intersection over Union (IoU):

$$\text{IoU} = \frac{|A \cap B| + \varepsilon}{|A \cup B| + \varepsilon} = \frac{\text{TP} + \varepsilon}{\text{TP} + \text{FP} + \text{FN} + \varepsilon} \quad (15)$$

Dice Coefficient (F1-Score):

$$\text{Dice} = \frac{2|A \cap B| + \varepsilon}{|A| + |B| + \varepsilon} = \frac{2\,\text{TP} + \varepsilon}{2\,\text{TP} + \text{FP} + \text{FN} + \varepsilon} \quad (16)$$

Precision:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{17}$$

Recall (Sensitivity):

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{18}$$

where, TP, FP, FN represent true positives, false positives, and false negatives, respectively, $\varepsilon = 10^{-7}$ provides numerical stability.

Additionally, absolute area difference quantifies spatial calibration independently of overlap metrics:

$$\Delta A = A_{\text{pred}} - A_{\text{true}} = (N_{\text{pred}} - N_{\text{true}}) \times A_{\text{pixel}} \tag{19}$$

where, N represents pixel counts, $A_{\text{pixel}} = 0.0036$ km^2 at 60 m resolution. A persistence baseline (predicting last observed frame) provides reference lower bound.

F. Training Configuration and Hyperparameters

1) CNN-LSTM hybrid models (ConvLSTM, U-Net+LSTM, Attention U-Net+ConvLSTM): All three architectures employed identical training configuration for fair comparison. Optimization utilized Adam algorithm [38] with learning rate $\eta = 1 \times 10^{-4}$, momentum parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and gradient clipping via `clipnorm` = 1.0 preventing exploding gradient instability. Batch size of 4 sequences balanced GPU memory constraints against gradient estimate quality. Training proceeded for maximum 200 epochs with early stopping monitoring validation loss (patience = 20 epochs, restore best weights). ReduceLROnPlateau callback dynamically adjusted learning rate upon validation loss plateau detection (patience = 7 epochs, reduction factor = 0.5, minimum $\eta_{\min} = 1 \times 10^{-7}$). Model checkpointing systematically saved best-performing checkpoint based on validation loss criterion.

2) Swin Spatio-Temporal Transformer: Reduced batch size of 2 sequences accommodated substantially, elevated GPU memory requirements (patch size = 4, embedding dimension = 64, window size = 8, hierarchical depths $[2, 2, 2]$, attention heads $[2, 4, 8]$, MLP ratio = 2.0, spatial dropout = 0.1). Temporal fusion module configured with 2 temporal transformer layers, 4 attention heads, MLP dimension 256, dropout 0.1. Decoder filter progression: $[128, 64, 32, 16]$. Remaining training hyperparameters identical to CNN-LSTM models.

3) ViT-based spatio-temporal model: Batch size 2 (patch size = 16, embedding dimension = 128, 2 spatial transformer layers, 2 temporal transformer layers, 4 attention heads each, MLP dimension 256, dropout 0.1). Remaining hyperparameters identical.

All model implementations used deterministic random seeds (NumPy: 42, TensorFlow: 42) ensuring repeatability and consistent data splitting for controlled comparisons.

G. Computational Framework for Risk Zone Mapping

After iterative autoregressive forecasting for the time span 2026-2045, post-processing framework is employed to define erosion and accretion risk zones using pixel-resolution spatial data.

Erosion strictly defined as the land to water transition between consecutive years ($L_t = 0 \wedge L_{t+1} = 1$), while **accretion** is a reverse process (water to land). Frequency maps characterize persistence duration:

$$f_{\text{erosion}}(\mathbf{x}) = \frac{1}{T-1} \sum_{t=1}^{T-1} \mathbb{I}(L_t(\mathbf{x}) = 0 \wedge L_{t+1}(\mathbf{x}) = 1) \tag{20}$$

$$f_{\text{accretion}}(\mathbf{x}) = \frac{1}{T-1} \sum_{t=1}^{T-1} \mathbb{I}(L_t(\mathbf{x}) = 1 \wedge L_{t+1}(\mathbf{x}) = 0) \tag{21}$$

where, $\mathbf{x}$ denotes spatial pixel location, T represents total forecast years, $\mathbb{I}(\cdot)$ is indicator function.

A **hydrodynamic stability index** quantifies pixel-level classification uncertainty via water occurrence frequency:

$$\mathcal{I}(\mathbf{x}) = 1 - |2F_w(\mathbf{x}) - 1|, \quad F_w(\mathbf{x}) = \frac{1}{T} \sum_{t=1}^{T} L_t(\mathbf{x}) \tag{22}$$

where, $F_w \in [0, 1]$ represents fraction of forecast period classified as water. Values $\mathcal{I} \approx 1$ indicate persistent land–water oscillation (maximum instability); $\mathcal{I} \approx 0$ indicates stable land or water classification (maximum stability).

Composite risk representation integrates erosion frequency, accretion frequency, and instability index highlighting regions simultaneously dynamic and uncertain. Physical area estimates employ:

$$A_{\text{erosion/accretion}} = N_{\text{pixels}} \times A_{\text{pixel}} \tag{23}$$

where, $A_{\text{pixel}} = (60 \text{ m})^2 = 0.0036$ km^2 at established 60 m spatial resolution.

V. Spatiotemporal Statistical Analysis

This section presents comprehensive quantitative spatiotemporal statistical characterization of 38 years of Padma River morphodynamic behaviour (1988–2025) including long-term areal dynamics, seasonal variability patterns, erosion–accretion mass budget calculations, and channel centreline migration rate quantification.

A. Long-Term Areal Dynamics and Trend Analysis

The Padma River exhibited substantial inter-annual variability in water-covered areal extent throughout 38-year observation period. Annual water-covered area ranged from absolute minimum 540.4 km^2 (calendar year 1995, hydrologically dry year characterized by below-average monsoon

precipitation) to absolute maximum 1,229.3 km^2 (calendar year 1999, catastrophic monsoon flood corresponding to major regional disaster event), with long-term temporal mean of 664.5 $\pm$ 112.8 km^2 (arithmetic mean $\pm$ one standard deviation). The exceptionally large channel expansion manifest in 1999 corresponds temporally to the catastrophic July–August 1998 monsoon flood event representing one of most severe recorded flood disasters in Bangladesh's documented history.

Formal non-parametric Mann–Kendall trend test [34], [35] indicated absence of statistically significant monotonic trend at $\alpha = 0.05$ significance level (Kendall's $\tau = -0.149$, two-tailed $p = 0.187$), and Theil–Sen robust linear estimator [36], [37] yielded weak negative slope magnitude of -1.41 km^2 yr^{-1}. The above complementary results of statistical analysis have effectively validated the fact that although there is a marked variation in the temporal pattern due to monsoon variations, the mean temporal pattern of river area does not exhibit any clear trend toward increase or decrease.

B. Seasonal Variability and Intra-Annual Dynamics

Magnitudes of seasonality were systematically analyzed by way of coefficient of variation (CV) computed for the corresponding time-series of bi-monthly water surface area magnitudes, revealing significant inter-annual variability (see Fig. 3). Magnitude of annual CV varied from a lowest value of 8.2% (year 2018, most hydrologically stable year with little seasonal change), to a highest of 64.1% (year 2007, most hydrologically unstable year), with an arithmetic mean value for the study interval of 34.3 $\pm$ 13.7%. Those calendar years that had high CV values, especially 2007, 2011, and 2020, were systematically related to extreme monsoon flooding episodes, characterized by a maximum water extent above 2,700 km^2. Years with lower CV values indicate higher stability of seasonal trajectory of water extent as well as comparatively smaller magnitude of variation of the flood-pulse.

Heat map representation of the quarterly data clearly reveals persistent and highly consistent hydrological seasonality pattern across all 38 years of monitoring; specifically, Q4 (October-December; post-monsoon recession season) invariably shows largest mean water extent, while Q2 (April-June; pre-monsoon low-water flow season) invariably records minimum extent values. This remarkable hydrological seasonality pattern clearly indicates the predominant influence of the South Asian monsoon meteorological system on hydrodynamics.

C. Erosion and Accretion Dynamics

Annual erosion and accretion rate magnitudes demonstrated exceptionally high inter-annual variability reflecting pronounced sensitivity to individual monsoon flood events (Fig. 4). The largest documented single-year erosion event occurred during 1998-1999 interval (649.1 km^2 land surface converted to water classification; net channel expansion 567.7 km^2), corresponding temporally to catastrophic July-August 1998 monsoon flood representing most extreme channel widening event recorded throughout entire 38-year study period. Conversely, the immediately following year (1999–2000 interval) recorded largest net accretion magnitude (-503.5 km^2), indicating substantial rapid channel contraction during post-flood geomorphic recovery and adjustment phase. This dramatic reversal pattern highlights pronounced recovery dynamics characterizing post-extreme-event channel behaviour.

TABLE III. DECADAL EROSION, ACCRETION, AND NET CHANNEL CHANGE SUMMARY.

Decadal Period	Erosion (km^2)	Accretion (km^2)	Net Change (km^2)
1988–1998	367.6	434.2	-66.7
1998–2008	382.3	355.9	$+26.3$
2008–2018	262.8	375.8	-113.0
2018–2025	264.6	264.3	$+0.3$
Total	**5170.2**	**5323.3**	$\mathbf{-153.0}$

Table III provides systematic quantitative summary of decadal erosion–accretion mass budgets.

Across complete 38-year study period, cumulative gross erosion reached 5,170.2 km^2 (representing multiple complete channel reworking cycles) while cumulative gross accretion totalled 5,323.3 km^2, yielding overall net accretion magnitude of 153.0 km^2 indicating slight long-term channel contraction tendency. Decadal analysis systematically identifies four temporally distinct morphodynamic regime phases: (1) 1988-1998 exhibited near-balanced erosion–accretion dynamics (-66.7 km^2 net accretion) characteristic of lateral channel migration without sustained widening; (2) 1998–2008 shifted toward slight net erosion tendency ($+26.3$ km^2) dominated by catastrophic 1998–1999 flood event and subsequent multi-year adjustment; (3) 2008–2018 manifested strongest net accretion (-113.0 km^2) suggesting significant channel contraction and partial geomorphic stabilization; (4) 2018–2025 approached near-equilibrium conditions with negligible net change ($+0.3$ km^2) indicating contemporary period of relative morphodynamic stability.

D. Channel Centreline Migration Rate Quantification

Channel centreline lateral migration rates exhibited pronounced temporal variability reflecting alternating periods of geomorphic instability and relative stabilization (Fig. 5). Arithmetic mean annual migration rate across full study period was 255.4 m yr^{-1}, with individual year magnitudes ranging from minimum 134.2 m yr^{-1} (calendar years 2001-2002 and 2022-2023, representing temporary stabilization phases) to catastrophic maximum 1,485.5 m yr^{-1} during 1998–1999 interval (corresponding to extreme flood-driven channel adjustment event). The median migration rate (216.3 m yr^{-1}) was substantially below arithmetic mean, conclusively confirming that small number of extreme flood years strongly inflated long-term average metrics while typical inter-annual behaviour exhibits more moderate migration magnitudes. The 1990s decade recorded highest average migration rates (mean 348.7 m yr^{-1}), definitively designating that decade as period of maximum geomorphic instability within 38-year observational record. Cumulative centreline displacement over complete study period exceeded 9 km total lateral adjustment, demonstrating sustained long-term lateral migration and continuous comprehensive reworking of active river corridor.

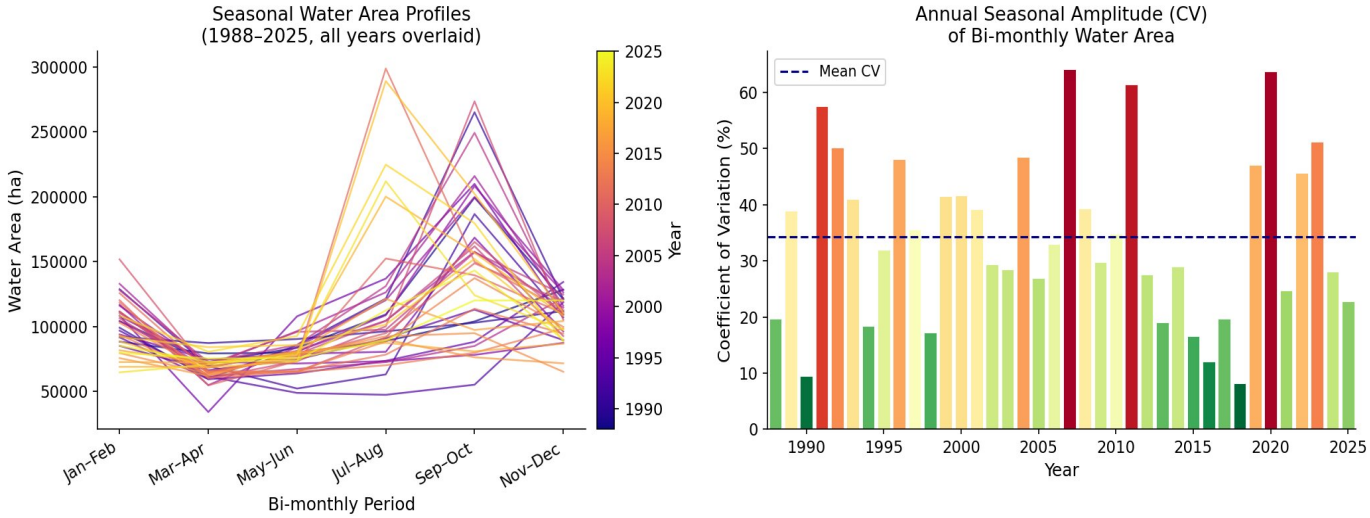


Fig. 3. Seasonal water area dynamics. Left: Bi-monthly water area profiles for all years. Right: Annual coefficient of variation (CV) of bi-monthly water area.

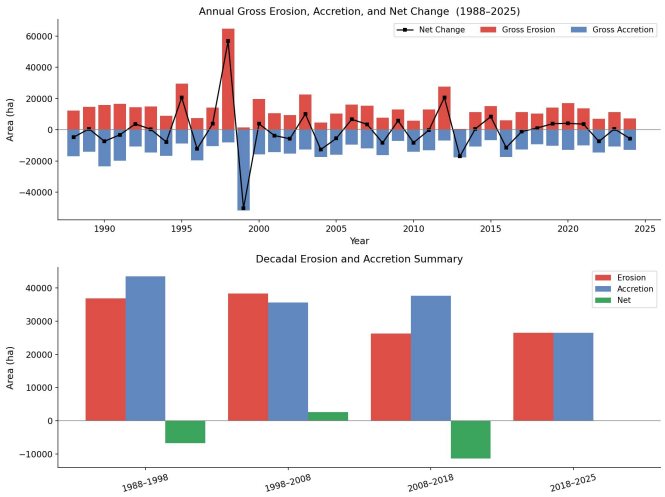


Fig. 4. Annual and decadal erosion and accretion of the channel. Top: annual gross erosion (red), gross accretion (blue), and net channel change (black line). Bottom: decadal grouped summary illustrating long-term morphodynamic phases.

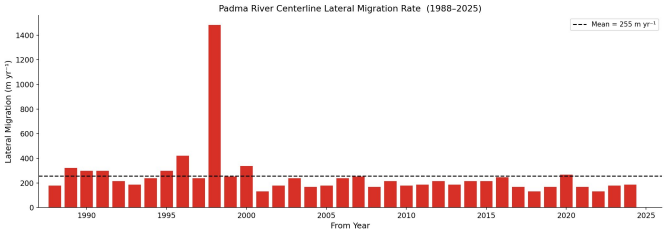


Fig. 5. Annual lateral migration rate of the centreline. Red bars indicate yearly migration rates, while the dashed line marks the long-term mean of 255.4 m yr^{-1}. The extreme peak in 1998–1999 corresponds to the catastrophic flood-driven channel adjustment.

Attention U-Net+ConvLSTM achieved highest IoU (0.7005) and Dice (0.8236) in Setup 1, establishing it as optimal architecture for yearly prediction tasks. U-Net+LSTM demonstrated competitive performance with slightly lower overlap metrics but superior area calibration. ConvLSTM and Swin Transformer exhibited moderate performance, while Vision Transformer significantly underperformed all alternative architectures (see Table IV).

TABLE IV. PERFORMANCE COMPARISON OF ALL ARCHITECTURES BASED ON YEARLY DATA.

Model	Setup	IoU	Dice	P	R	$\lvert\Delta A\rvert$
ConvLSTM	S1	0.6218	0.7665	0.7212	0.8207	76.69
ConvLSTM	S2	0.6278	0.7711	0.7139	0.8432	103.26
U-Net+LSTM	S1	0.6964	0.8208	0.7876	0.8595	50.25
U-Net+LSTM	S2	0.6986	0.8223	0.7905	0.8592	48.61
Att.-U-Net+CLSTM	S1	**0.7005**	**0.8236**	0.7807	0.8747	66.54
Att.-U-Net+CLSTM	S2	0.6963	0.8205	0.7934	0.8523	**41.05**
ViT	S1	0.3613	0.5302	0.4352	0.6806	316.06
ViT	S2	0.3831	0.5526	0.5197	0.5907	78.69
Swin	S1	0.6080	0.7560	0.7364	0.7780	31.64
Swin	S2	0.5912	0.7431	0.6849	0.8122	96.23

1) Detailed year-wise analysis: Table V and Table VI present comprehensive year-by-year breakdown for best-performing Attention U-Net+ConvLSTM architecture.

Performance remained remarkably stable across 10-year horizon with IoU ranging 0.6569–0.7359 (standard deviation 0.0273). Higher Recall (mean 0.8747) relative to Precision (0.7807) indicates systematic tendency toward over-prediction, consistent with positive area differences in most years. The 5-year setup achieved lower area differences (mean 41.05 km^2 vs 66.54 km^2), confirming shorter horizons reduce temporal uncertainty.

TABLE V. DETAILED YEAR-WISE PREDICTION RESULTS: ATTENTION U-NET+CONVLSTM, 10-YEAR FORECASTING SETUP (2016–2025).

Year	IoU	Dice	Prec.	Rec.	ΔA (km^2)
2016	0.7357	0.8477	0.8655	0.8307	−24.96
2017	0.6975	0.8218	0.7396	0.9246	130.72
2018	0.6742	0.8054	0.7273	0.9023	122.18
2019	0.7051	0.8270	0.7858	0.8729	58.46
2020	0.7055	0.8273	0.7924	0.8654	51.24
2021	0.6569	0.7929	0.7706	0.8165	34.81
2022	0.7359	0.8479	0.8407	0.8552	10.84
2023	0.7081	0.8291	0.7618	0.9095	107.73
2024	0.7046	0.8267	0.7806	0.8786	70.94
2025	0.6814	0.8105	0.7430	0.8915	103.46
Mean	**0.7005**	**0.8236**	0.7807	0.8747	66.54
Std	0.0273	0.0179	0.0440	0.0325	56.23

TABLE VI. DETAILED YEAR-WISE PREDICTION RESULTS: ATTENTION U-NET+CONVLSTM, 5-YEAR FORECASTING SETUP (2021–2025).

Year	IoU	Dice	Prec.	Rec.	ΔA (km^2)
2021	0.6384	0.7793	0.7797	0.7788	−0.66
2022	0.7270	0.8419	0.8546	0.8296	−18.39
2023	0.7143	0.8333	0.7745	0.9018	91.31
2024	0.7000	0.8235	0.7950	0.8542	42.04
2025	0.7017	0.8247	0.7631	0.8972	90.98
Mean	0.6963	0.8205	0.7934	0.8523	41.05
Std	0.0341	0.0244	0.0334	0.0511	50.98

TABLE VII. PERFORMANCE COMPARISON OF ALL ARCHITECTURES BASED ON QUARTERLY DATA.

Model	Setup	IoU	Dice	P	R	$\lvert\Delta A\rvert$
ConvLSTM	S1	0.6114	0.7572	0.7433	0.7918	166.66
ConvLSTM	S2	0.6134	0.7586	0.7115	0.8348	159.91
U-Net+LSTM	S1	0.6908	0.8095	0.8162	0.8318	188.03
U-Net+LSTM	S2	**0.7008**	0.8191	0.8257	0.8431	163.76
Att.-U-Net+CLSTM	S1	**0.6956**	**0.8139**	0.8095	**0.8443**	184.80
Att.-U-Net+CLSTM	S2	0.6972	**0.8170**	0.8054	0.8511	**137.98**
ViViT	S1	0.3648	0.5316	0.4788	0.6347	340.08
ViViT	S2	0.3822	0.5496	0.4606	0.7281	447.09
Swin	S1	0.5865	0.7370	0.7012	0.7981	190.79
Swin	S2	0.5882	0.7392	0.7004	0.8024	160.45

E. Quarterly Prediction Results: Complete Architecture Comparison

Attention U-Net+ConvLSTM and U-Net+LSTM exhibited closely competitive performance for quarterly prediction, with Attention model achieving marginally superior IoU in Setup 1 (0.6956 vs 0.6908) (see Table VII).

1) Detailed quarterly results Setup 1: Table VIII presents complete quarter-by-quarter breakdown for Attention U-Net+ConvLSTM covering 2017 Q2–2025 Q4 test period.

TABLE VIII. DETAILED QUARTERLY PREDICTION RESULTS – SETUP 1 (ATTENTION U-NET+CONVLSTM, SEQLEN=8, TEST: 2017 Q2–2025 Q4).

Quarter	IoU	Dice	Prec.	Rec.	ΔA (km^2)
2017-Q2	0.9156	0.9559	0.9524	0.9595	5.31
2017-Q3	0.7556	0.8608	0.9543	0.7840	−171.18
2017-Q4	0.7778	0.8750	0.9127	0.8404	−73.25
2018-Q1	0.8315	0.9080	0.8833	0.9341	41.75
2018-Q2	0.9150	0.9556	0.9304	0.9823	38.33
2018-Q3	0.7056	0.8274	0.8341	0.8208	−12.15
2018-Q4	0.5594	0.7175	0.6733	0.7678	94.13
2019-Q1	0.8782	0.9351	0.8851	0.9911	71.73
2019-Q2	0.6949	0.8200	0.7197	0.9528	197.74
2019-Q3	0.5647	0.7218	0.9123	0.5971	−447.49
2019-Q4	0.6155	0.7620	0.6746	0.8754	241.01
2020-Q1	0.7440	0.8532	0.7610	0.9708	157.89
2020-Q2	0.9068	0.9511	0.9353	0.9675	20.12
2020-Q3	0.5061	0.6721	0.8699	0.5475	−565.15
2020-Q4	0.5661	0.7230	0.6572	0.8034	179.91
2021-Q1	0.7922	0.8840	0.8376	0.9360	72.49
2021-Q2	0.8128	0.8967	0.8156	0.9958	119.94
2021-Q3	0.4833	0.6517	0.8805	0.5173	−662.31
2021-Q4	0.5279	0.6910	0.5910	0.8316	320.34
2022-Q1	0.6335	0.7756	0.6908	0.8842	142.33
2022-Q2	0.6190	0.7647	0.9081	0.6604	−213.69
2022-Q3	0.5230	0.6868	0.8612	0.5711	−519.60
2022-Q4	0.6256	0.7697	0.7454	0.7956	63.38
2023-Q1	0.7641	0.8663	0.8414	0.8926	38.33
2023-Q2	0.7147	0.8336	0.7684	0.9109	107.41
2023-Q3	0.6263	0.7702	0.9154	0.6647	−306.30
2023-Q4	0.6325	0.7749	0.6465	0.9670	359.43
2024-Q1	0.6504	0.7882	0.6631	0.9713	215.20
2024-Q2	0.7352	0.8474	0.8906	0.8081	−52.76
2024-Q3	0.5501	0.7097	0.5527	0.9915	427.37
2024-Q4	0.6480	0.7864	0.7055	0.8882	174.97
2025-Q1	0.8124	0.8965	0.8238	0.9833	118.42
2025-Q2	0.9298	0.9636	0.9493	0.9783	18.22
2025-Q3	0.5993	0.7495	0.8262	0.6858	−170.04
2025-Q4	0.7280	0.8426	0.8655	0.8209	−48.20
Mean	**0.6956**	**0.8139**	0.8095	0.8443	−0.47
Std	0.1230	0.0851	0.1124	0.1275	244.68

Pronounced seasonal stratification is evident: Q2 (pre-monsoon) consistently achieves IoU > 0.90 in multiple years (2017, 2018, 2020, 2025), while Q3 (peak monsoon) exhibits lowest performance, frequently falling below 0.55. The signed mean area difference approaches zero (-0.47 km^2), indicating excellent long-term calibration despite substantial within-year fluctuations driven by seasonal effects.

2) Detailed quarterly results Setup 2: Table IX presents detailed results for shorter 5-year forecasting horizon (2022 Q4–2025 Q4).

Setup 2 results demonstrate consistent seasonal patterns with improved mean performance (IoU = 0.6972 vs. 0.6956) attributable to shorter forecasting horizon reducing cumulative error propagation.

F. Bi-Monthly Prediction Results: Complete Architecture Comparison

U-Net+LSTM emerged as best-performing architecture for bi-monthly prediction, achieving IoU = 0.7791 and Dice = 0.8701 in Setup 2. This architectural shift from

TABLE IX. DETAILED QUARTERLY PREDICTION RESULTS – SETUP 2 (ATTENTION U-NET+CONVLSTM, SEQLEN=10, TEST: 2022 Q4–2025 Q4)

Quarter	IoU	Dice	Prec.	Rec.	ΔA (km^2)
2022-Q4	0.6086	0.7567	0.7498	0.7637	17.46
2023-Q1	0.7485	0.8561	0.8325	0.8812	36.82
2023-Q2	0.7025	0.8253	0.7826	0.8729	66.80
2023-Q3	0.6164	0.7627	0.8820	0.6719	−266.44
2023-Q4	0.6406	0.7810	0.6768	0.9230	263.79
2024-Q1	0.6552	0.7917	0.6769	0.9533	189.02
2024-Q2	0.7804	0.8767	0.9655	0.8028	−96.03
2024-Q3	0.5669	0.7236	0.5703	0.9894	395.49
2024-Q4	0.6496	0.7876	0.7263	0.8601	124.49
2025-Q1	0.8740	0.9328	0.9362	0.9293	−4.55
2025-Q2	0.9407	0.9695	0.9583	0.9809	14.04
2025-Q3	0.5745	0.7297	0.8544	0.6368	−254.68
2025-Q4	0.7061	0.8278	0.8583	0.7994	−64.14
Mean	0.6972	0.8170	0.8054	0.8511	32.47
Std	0.1043	0.0725	0.1111	0.1004	188.53

TABLE X. PERFORMANCE COMPARISON OF ALL ARCHITECTURES ON BI-MONTHLY DATA.

Model	Setup	IoU	Dice	P	R	$\lvert\Delta A\rvert$
ConvLSTM	S1	0.6347	0.7732	0.7941	0.7708	131.44
ConvLSTM	S2	0.6623	0.7956	0.7979	0.8062	119.11
U-Net+LSTM	S1	0.7439	0.8435	0.8658	0.8496	156.64
U-Net+LSTM	S2	**0.7791**	**0.8701**	**0.8971**	0.8691	160.78
Att.-U-Net+CLSTM	S1	0.7410	0.8406	0.8806	0.8293	157.45
Att.-U-Net+CLSTM	S2	0.7733	0.8662	0.9094	0.8459	**147.45**
Swin	S1	0.6333	0.7718	0.8015	0.7608	131.99
Swin	S2	0.6952	0.8180	0.8404	0.8086	114.45
ViT	S1	0.3805	0.5487	0.5461	0.5771	203.91
ViT	S2	0.3421	0.5079	0.4723	0.5711	243.07

Attention U-Net+ConvLSTM dominance at yearly/quarterly resolutions to U-Net+LSTM leadership at bi-monthly suggests spatial feature extraction quality becomes progressively more critical as temporal frequency increases (see Table X).

G. Long-Term Morphological Forecasting (2026-2040)

In long-term morphological forecasting, final model setup was Attention U-Net+ConvLSTM configuration trained on complete dataset covering time span 1987-2025 with a window length equal to five. Convergence point for training achieved epoch 107, best validation loss being equal to 0.2298, validation Dice index 0.8057, and validation IoU 0.7026. Reconstruction test for known time interval (2016–2025) revealed mean IoU = 0.6543 and mean Dice index 0.7907, thus indicating that model is capable of reconstructing historical river extent patterns sufficiently accurately.

Starting from the 2025 area of 517.6 km^2, river extent grew up gradually to 671.3 km^2 in 2026 and reached value of 807.6 km^2 by 2040, indicating total growth of 290.0 km^2 or 56.03% as compared to the base year. Expansion trend had nonuniform character, showing largest growth rates during the first years of forecasting (2025-2026) and further decreasing gradually. The accretion-dominated pattern indicates future dynamics primarily associated with lateral expansion into adjacent floodplain areas rather than wholesale channel widening.

The comprehensive long-term risk cartography (Fig. 6) identifies spatially explicit high-probability morphological change zones concentrated systematically along active channel margins and proximal floodplain corridor segments, with very-high-risk designations indicating persistent instability and recurrent channel boundary adjustment patterns most operationally relevant for infrastructure protection planning and settlement management strategies.

H. Short-Term Bi-Monthly Forecasting (2026-2028)

Short-term high-frequency bi-monthly forecasting employed U-Net+ConvLSTM2D architecture with nine-period input sequences systematically enriched with sinusoidal and cosinusoidal seasonal encoding channels ($\sin(2\pi t/6)$, $\cos(2\pi t/6)$ where t represents bi-monthly period index), enabling model to explicitly distinguish intra-annual temporal periods without relying exclusively on spatial morphological content patterns. Training convergence occurred at epoch 39 (validation Dice = 0.8381, IoU = 0.7463). Reconstruction sanity check on recent historical period 2023 P1–2025 P6 yielded mean IoU = 0.684 ± 0.072 (mean $\pm$ std).

Over 18 forecast periods spanning 2026 P1–2028 P6, mean predicted area followed clear three-phase temporal structure: initial contraction phase to mean 902.7 km^2 in 2026 (below 934.1 km^2 baseline), reflecting dry-to-wet seasonal transition inherited from seed sequence; progressive rise to 984.0 km^2 in 2027; further increase to 1,020.1 km^2 in 2028, with terminal forecast period (2028 P6, November–December) reaching 1,054.4 km^2 (+12.88% above baseline). The seasonal cycle was faithfully reproduced with characteristic pattern: minimum extent at P3 (May–June pre-monsoon, 918.3 km^2), maximum at P5 (September–October peak monsoon, 1,034.3 km^2), seasonal amplitude 116.0 km^2, and mean monsoon area 53.4 km^2 (+5.62%) above non-monsoon mean. Final forecast period exhibited accretion-to-erosion ratio approximately 1.60:1, with 78.4% of baseline channel extent maintaining stable classification.

I. Computational Complexity

Table XI summarizes the model complexity in terms of trainable parameters and average training time across different temporal resolutions.

J. Forecast Reliability and Uncertainty

Long-term forecasting is subject to uncertainty because prediction errors can accumulate over successive time steps. To assess forecast reliability, rolling predictions were evaluated across two controlled experimental setups: a 10-year horizon (Setup 1: 2016–2025) and a 5-year horizon (Setup 2: 2021–2025). The best-performing Attention U-Net+ConvLSTM model maintained remarkably stable performance across these testing intervals, achieving a mean IoU of 0.7005 (Dice = 0.8236) in the 10-year long-range forecast setup, and a mean IoU of 0.6963 (Dice = 0.8205) in the 5-year medium-range setup.

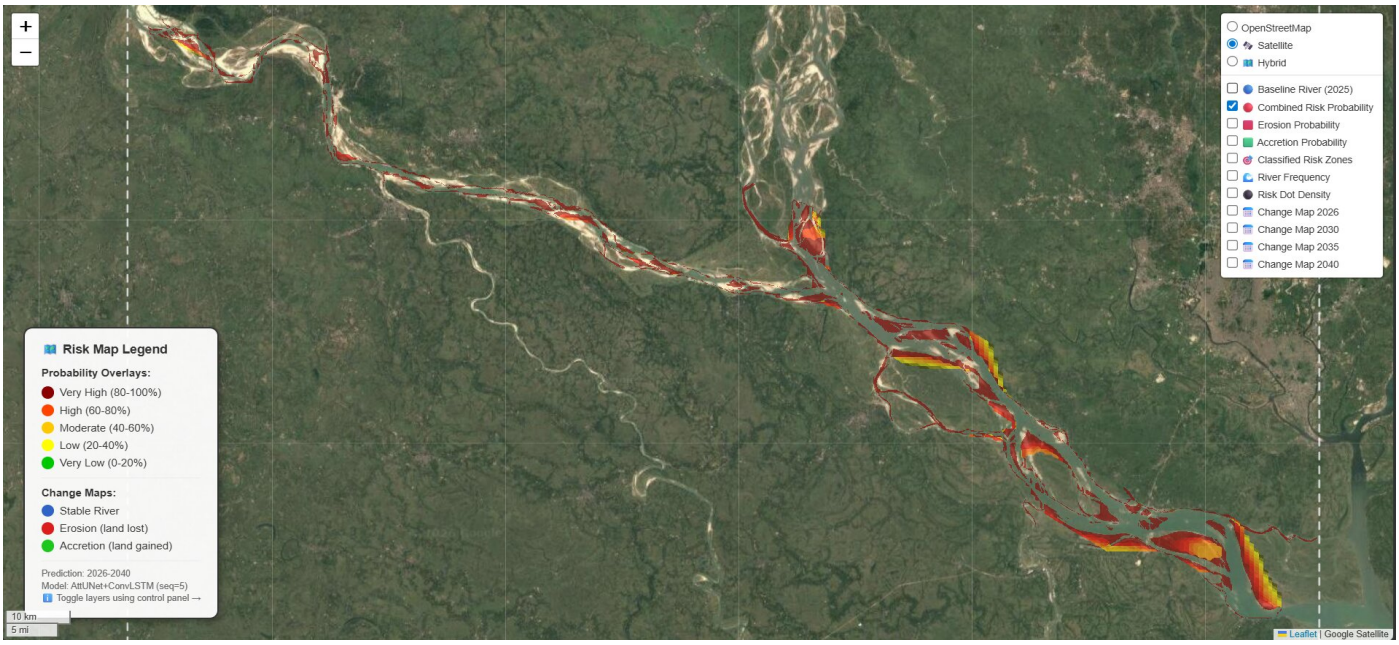


Fig. 6. Long-term combined morphological risk for 2026-2040. Interactive web demonstration available here.

TABLE XI. MODEL COMPLEXITY IN TERMS OF TRAINABLE PARAMETERS, TRAINING TIME, AND INFERENCE TIME PER PREDICTION.

Model	Data	Params (M)	Train (min)	Infer (s)
ConvLSTM	Yearly	1.41	8.2	0.31
	Quarterly		34.5	0.50
	Bi-monthly		54.1	0.65
U-Net + LSTM	Yearly	7.71	12.4	0.71
	Quarterly		48.2	1.10
	Bi-monthly		72.5	1.30
Attention U-Net + ConvLSTM	Yearly	**7.41**	**14.1**	**0.81**
	Quarterly		**52.8**	**1.10**
	Bi-monthly		**79.3**	**1.30**
ViT	Yearly	3.24	6.5	0.40
	Quarterly		22.0	0.62
	Bi-monthly		48.2	0.83
Swin Transformer	Yearly	5.07	22.0	0.60
	Quarterly		78.3	0.80
	Bi-monthly		115.4	1.00

Furthermore, the model exhibited tighter dimensional calibration over the shorter testing window, yielding a lower mean absolute area difference ($|\Delta A| = 41.05$ km^2) in Setup 2 compared to Setup 1 ($|\Delta A| = 66.54$ km^2). This behavior confirms that shorter horizons effectively curb cumulative error propagation. Rather than presenting future channel changes via strict binary boundaries, long-term channel evolution (2026–2040) is mapped using spatially explicit risk zones derived from pixel-level water occurrence frequencies (F_w) and a hydrodynamic stability index (I). This approach accounts for spatial modeling uncertainties and highlights highly dynamic zones along active channel margins.

VI. DISCUSSION

Hybrid CNN-LSTM architectures outperformed transformer-based architectures in terms of overlap and area metrics for all datasets and forecasting horizons. Underperformance of transformer-based architectures (ViT, Swin Transformer) is presumably related to data scarcity. The limited temporal samples (38 yearly, 125 quarterly, and 156 bi-monthly composites) likely increased the risk of overfitting for attention-based models. Although a full hyperparameter sweep was not conducted, key architecture-specific parameters (e.g., patch size, window size, and embedding dimension) were determined via iterative manual testing, and the most effective configuration was selected for final evaluation. Hierarchical window-based attention introduces stronger spatial locality biases, which may explain its better performance than ViT. With growing satellite databases and availability of pre-trained geospatial foundation models, transformer architectures are anticipated to be competitive with CNN-LSTM hybrids.

The Attention U-Net+ConvLSTM was best suited for yearly and quarterly forecasting, where attention helps focus on geographically active regions. The U-Net+LSTM architecture was superior for bi-monthly forecasts when boundary preservation is crucial but where gating provided only incremental benefits.

There was a significant seasonal influence on the performance of networks in quarter- and bi-monthly experiments. It relates to the difficulty associated with predicting the monsoon extent, where channel widths become large, have unusual forms, and become inundated with floodplain waters. For Setup 1 quarterly predictions, mean areas were close to each other (–0.47 km²), suggesting annual canceling of seasonal systematic errors.

An essential realization in regard to operational use is that metrics like Intersection Over Union (IoU) and Dice score are insufficient measures of model performance. Several model configurations had similar metrics of overlap but vastly differing absolute area differences, showing spatially accurate but incorrectly dimensionally sized predictions. River hazard

management processes such as erosion risk calculations, land lost calculations in agriculture, embankments construction, and disaster resource allocations require accurate area calculations. It is recommended that IoU, Dice, Precision, Recall, and absolute area difference be considered minimum requirements for performance metrics, with emphasis on area difference in selection process for hazard management.

Future prediction from 2026 to 2040 shows continuing expansion with accretion outpacing erosion. This corresponds with past observations between 2008 and 2018 (net accretion of 113.0 km²) and near equilibrium 2018 to 2025, indicating models learned realistic trends for the future. The slowing expansion beyond 2031 may indicate learned stabilizing effects; however, potential cumulative errors in modeling may be responsible. Historical reconstruction having consistent positive area bias (91.8 km²) indicates that the expansion values calculated are upper limit estimates. Nevertheless, the more accurate risk zone maps defining corridor segments with high change probability provide the main output for disaster management.

A. Limitations and Future Research Directions

However, several limitations need mentioning. Firstly, 60 m Landsat spatial resolution fails to capture erosion microfeatures, secondary channels, and banks' morphology, thus requiring shifting to 10 m Sentinel-2 spatial resolution data through the whole period of analysis. Secondly, lack of hydrological forcing data in the form of river discharge, precipitation, and antecedent soil moisture presents a serious limitation for the framework under development, especially during the periods of flooding. Thirdly, the proposed approach lacks any geotechnical parameters such as the soil type, bank height, and vegetation distribution, which influence erosion processes. Lastly, all the presented experiments were done on only one river system, making cross-system testing on Jamuna and Meghna river systems necessary to make any confident generalizations.

The following directions may be considered in future research: (1) inclusion of 10 m Sentinel-2 optical and Sentinel-1 SAR data through the entire time series; (2) widening the geographic scope to other river systems – Jamuna and Meghna; (3) including hydrological variables as forcing data to better predict the extremes; (4) exploring transformer neural networks pre-trained on larger geospatial foundation models corpora; and (5) combining geomorphology-based vulnerability map with population density and agriculture maps for a socio-economic early warning system.

VII. Conclusion

This investigation presents a comprehensive framework for analyzing and forecasting morphological evolution of the Padma River by combining long-term satellite remote sensing with spatiotemporal deep learning. A 38-year satellite-derived dataset enabled comprehensive analysis of long-term river dynamics, revealing substantial erosion, accretion, and channel migration. The proposed preprocessing framework produced reliable historical water masks for subsequent forecasting. Hybrid CNN-LSTM architectures consistently outperformed transformer-based models under limited training data. The proposed framework accurately captured both long-term morphological evolution and short-term seasonal dynamics, enabling spatially explicit future risk assessment. This research demonstrates that freely available satellite imagery combined with cloud-based processing and spatiotemporal deep learning can provide a practical, scalable, and reproducible framework for riverbank hazard monitoring and multi-year forecasting. Furthermore, this remote sensing and deep learning pipeline is fundamentally generic and can be seamlessly applied to any other dynamic river system by retraining the models on their respective historical data. For those living near erodible banks along rivers throughout Bangladesh and South Asia, predictive tools capable of identifying hazardous areas can meaningfully enhance disaster preparedness, infrastructure planning, and community adaptation strategies.

References

[1] P. Lemenkova, "Deep learning methods of satellite image processing for monitoring of flood dynamics in the Ganges delta, Bangladesh," *Water*, vol. 16, no. 8, p. 1141, 2024.

[2] J. Freihardt *et al.*, "Assessing riverbank erosion in Bangladesh using time series of Sentinel-1 radar imagery in the Google Earth Engine," *Natural Hazards and Earth System Sciences*, vol. 23, pp. 751–768, 2023.

[3] G. M. M. Alam *et al.*, "Exploring impacts and livelihood vulnerability of riverbank erosion hazard among rural household along the river Padma of Bangladesh," *Environmental Systems Research*, vol. 6, p. 25, 2017.

[4] M. N. Islam *et al.*, "Socioeconomic impacts and migration dynamics of riverbank erosion," *npj Climate Action*, vol. 4, p. 6, 2025.

[5] T. K. Das *et al.*, "Impact of riverbank erosion: A case study," *Australasian Journal of Disaster and Trauma Studies*, vol. 21, no. 2, pp. 73–81, 2017.

[6] D. Sarkar *et al.*, "Assessing riverbank erosion and livelihood resilience using traditional approaches in northern Bangladesh," *Sustainability*, vol. 14, no. 4, p. 2348, 2022.

[7] H. Fan *et al.*, "Using long short-term memory networks for river flow prediction," *Hydrology Research*, vol. 51, no. 6, pp. 1358–1376, 2020.

[8] A. Dehghani *et al.*, "Comparative evaluation of LSTM, CNN, and ConvLSTM for spatiotemporal forecasting tasks," *Ecological Informatics*, vol. 75, p. 101984, 2023.

[9] D. V. Binh *et al.*, "A novel method for river bank detection from Landsat time series (SRBED)," *Remote Sensing*, vol. 12, no. 20, p. 3298, 2020.

[10] T. Langhorst *et al.*, "Global observations of riverbank erosion and accretion from Landsat imagery," *Journal of Geophysical Research: Earth Surface*, vol. 128, no. 2, p. e2022JF006774, 2023.

[11] P. K. Langat *et al.*, "Monitoring river channel dynamics using remote sensing and GIS techniques," *Geomorphology*, vol. 325, pp. 92–106, 2019.

[12] M. N. Islam and M. A. Haque, "River channel migration: A remote sensing and GIS analysis," in *Proc. ESA Living Planet Symposium*, Bergen, Norway, 2010.

[13] P. V. Raju *et al.*, "Spatio-temporal analysis of riverbank changes using remote sensing and geographic information system," *Ecological Informatics*, vol. 81, p. 102628, 2024.

[14] T. Tha, T. Piman, D. Bhatpuria, and P. Ruangrassamee, "Assessment of riverbank erosion hotspots along the Mekong River in Cambodia using remote sensing and hazard exposure mapping," *Water*, vol. 14, no. 19, p. 2995, 2022.

[15] O. R. Saha, "Riverbank migration and island dynamics at the Padma confluence: Multi-sensor analysis," *Geomorphology*, vol. 448, p. 109043, 2025.

[16] M. A. Haque, "Change detection of Jamuna River and its impact on local land use and settlement," *Journal of Urban and Regional Analysis*, vol. 15, no. 2, pp. 145–162, 2023.

[17] S. M. Ritu, S. K. Sarkar, and H. Zonaed, "Prediction of Padma river bank shifting and its consequences on LULC changes," *Ecological Indicators*, vol. 156, p. 111104, 2023.

[18] C. Muzahid *et al.*, "Quantifying river bank erosion and accretion patterns along the Gorai River in Kushtia, Bangladesh," *Journal of Geographic Information System*, vol. 16, pp. 70–88, 2024.

[19] Z. Ren, "A large-scale riverbank erosion risk assessment model with remote sensing inputs," *Applied Geography*, vol. 162, p. 103156, 2024.

[20] K. Yuan *et al.*, "Deep-learning-based multispectral satellite image segmentation for water body detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 7422–7434, 2021.

[21] Y. Li *et al.*, "Assessing the performance and interpretability of the CNN-LSTM-Attention model for daily streamflow forecasting in typical basins of the eastern Qinghai-Tibet Plateau," *Scientific Reports*, vol. 15, p. 260, 2025.

[22] S. Ghosh *et al.*, "Water body extraction from high spatial resolution remote sensing images based on enhanced U-Net and multi-scale information fusion," *Scientific Reports*, vol. 14, p. 16192, 2024.

[23] B. DeVries *et al.*, "Rapid and robust monitoring of flood events using Sentinel-1 and Landsat data on the Google Earth Engine," *Remote Sensing of Environment*, vol. 240, p. 111664, 2020.

[24] M. A. Clement *et al.*, "Visualisation of flooding along an unvegetated, ephemeral river using Google Earth Engine," *Journal of Environmental Management*, vol. 277, p. 111338, 2020.

[25] P. Tripathy and A. Mallick, "Applying Google Earth Engine for flood mapping and monitoring in the downstream provinces of Mekong River," *Results in Engineering*, vol. 14, p. 100432, 2022.

[26] N. Sazib *et al.*, "Google Earth Engine for large-scale flood mapping using SAR data and impact assessment on agriculture and population of Ganga-Brahmaputra basin," *Sustainability*, vol. 14, no. 7, p. 4210, 2022.

[27] X.-H. Le *et al.*, "Streamflow prediction using an integrated methodology based on convolutional neural network and long short-term memory networks," *Scientific Reports*, vol. 11, p. 17429, 2021.

[28] Z. Xiang, J. Yan, and I. Demir, "Hybrid CNN-LSTM models for river flow prediction," *Water Supply*, vol. 22, no. 5, pp. 4902–4919, 2022.

[29] A. Horvath *et al.*, "Water level prediction using long short-term memory neural network model for a lowland river: A case study on the Tisza River, Central Europe," *Environmental Sciences Europe*, vol. 35, p. 92, 2023.

[30] Z. Liu *et al.*, "Video Swin Transformer," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 3202–3211.

[31] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021.

[32] A. Arnab *et al.*, "ViViT: A video vision transformer," in *Proc. IEEE/CVF ICCV*, 2021, pp. 6836–6846.

[33] J. Garber *et al.*, "Artificial intelligence and objective-function methods can identify bankfull river channel extents," *Water Resources Research*, vol. 60, no. 1, p. e2023WR035269, 2024.

[34] H. B. Mann, "Nonparametric tests against trend," *Econometrica*, vol. 13, no. 3, pp. 245–259, 1945.

[35] M. G. Kendall, *Rank Correlation Methods*, 4th ed. London: Charles Griffin, 1975.

[36] H. Theil, "A rank-invariant method of linear and polynomial regression analysis," *Proceedings of the Royal Netherlands Academy of Sciences*, vol. 53, pp. 386–392, 1950.

[37] P. K. Sen, "Estimates of the regression coefficient based on Kendall's tau," *Journal of the American Statistical Association*, vol. 63, no. 324, pp. 1379–1389, 1968.

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

Hardware-Software Co-Design of K-Means Clustering on Zynq FPGAs via OpenCL

Eya JABALLI[1], Soufien GDAIM[2], Noureddine LIOUANE[3]
LARATSI Research Laboratory, ENIM, University of Monastir, Tunisia[1,3]
PRINCE Research Laboratory, ISITCom, University of Sousse, Tunisia[2]

***Abstract*—Clustering algorithms such as K-means are widely employed in machine learning and data analytics; however, their computational complexity becomes a significant limitation when applied to large-scale or latency-sensitive workloads. To overcome this challenge, this study presents a hardware-accelerated K-means clustering implementation using Field-Programmable Gate Arrays (FPGAs) within the Open Computing Language (OpenCL) framework. The proposed architecture leverages the fine-grained parallelism of FPGAs to accelerate the assignment phase, which is dominated by repetitive distance computations between data points and centroids. The proposed heterogeneous architecture accelerates only the assignment stage of the K-means algorithm on the FPGA, while the centroid update and convergence-control stages remain executed on the embedded ARM Cortex-A9 processor. Experimental results obtained on a two-dimensional synthetic floating-point dataset demonstrate significant performance gains over a CPU-only implementation, achieving speedups of up to 39× in execution time. The results also indicate consistent and favorable scalability trends across the evaluated synthetic two-dimensional configurations. To further assess the generalizability of the proposed approach, the accelerator was additionally validated on a real-world geospatial dataset, confirming consistent performance gains beyond synthetic benchmarks.**

***Keywords*—FPGAs; Hardware acceleration; high-level synthesis; OpenCL; K-means; reconfigurable hardware**

I. Introduction

Clustering plays an essential role in data mining research and applications and remains an active area of investigation in various disciplines, including computer science, data science, statistics, pattern recognition, artificial intelligence, and machine learning [1], [2]. The K-means algorithm is particularly significant among the various clustering algorithms because it is effective in discovering intrinsic patterns and revealing hidden structures in unlabeled datasets. Its use spans a wide range of application fields, including real-time data processing, IoT edge computing [3], [4], and big data processing environments [5], all of which are important for cluster classification for unlabeled data. Nevertheless, the increasing size and dimensionality of today's datasets pose serious computational challenges. With an increase in dataset size comes an increase in complexity of the clustering process, which causes slowdowns and performance bottlenecks that can be especially relevant in latency-sensitive or energy-limited real-time or edge computing environments. Performance bottlenecks such as these need to be solved to maintain the responsiveness and efficiency of a platform for real-time or edge processing. Multi-core central processing units (CPUs) and graphics processing units (GPUs) have been utilized for accelerating the computation involved in clustering [6]. While some performance improvement through parallelization can be achieved on these architectures, neither platform can match the power efficiency, deterministic latency, or real-time characteristics often required in edge computing applications, where device resources and energy consumption are at a premium. In this context, Field-Programmable Gate Arrays (FPGAs) have quickly gained traction as a viable option for hardware acceleration of clustering algorithms. FPGAs are reconfigurable architectures, which allows for fine-grained parallelism and deep pipelining, which gives developers the possibility to create hardware that is more closely aligned to the computational structure in algorithms. Moreover, unlike GPUs, which utilize a Single Instruction, Multiple Data (SIMD) that focuses on throughput rather than determinism, FPGAs execute tasks with low-latency and power-efficient results, which is particularly rewarding for real-time applications. Ultimately, the combination of flexible hardware paths, elasticity with scalability, and energy efficiency underscores the potential of FPGAs as an effective platform for hardware acceleration of unsupervised learning algorithms. The deployment of clustering algorithms on FPGAs introduces two significant performance bottlenecks. First, the latency of transferring data between the host and the FPGA device acts as an important constraint. In many cases, the communication pipeline between the CPU and FPGA is restricted by the use of synchronous data transfers, and this can negatively affect overall system throughput. This can be most detrimental to workloads that operate iteratively, such as clustering algorithms. A second contribution to performance limitations arises from an insufficient utilization of memory reads, because a sequence of memory operations in a clustering algorithm is irregular and often data-dependent. To improve the efficiency of the embedded CPU–FPGA implementation, the design adopts two implementation-level optimization strategies:

1) Host–kernel communication management: OpenCL command queues, non-blocking memory operations, and event-based synchronization are used to coordinate data migration, kernel execution, and result retrieval between the ARM processor and the FPGA kernel. These mechanisms reduce unnecessary host-side blocking and provide a structured execution flow for the heterogeneous implementation.
2) Burst-compatible memory access organization: Input observations and centroid arrays are stored in contiguous memory regions and accessed sequentially through AXI-based memory interfaces. This organization enables the HLS tool to infer burst-oriented memory transactions and reduces inefficient random memory accesses in the distance-computation kernel.

It is important to clarify the scope of this work. The proposed FPGA implementation accelerates the K-means assignment stage, specifically the nearest-centroid distance computation. The centroid update and convergence-control operations remain executed on the ARM processor. Therefore, the reported speedups refer to the heterogeneous implementation of the profiled K-means workload under the evaluated configurations, not to a fully hardware-implemented K-means pipeline. The current implementation evaluates two-dimensional floating-point datasets generated synthetically. The main contributions of this work are summarized as follows:

- An OpenCL-based heterogeneous CPU–FPGA implementation that accelerates the assignment stage of K-means on the FPGA while preserving the centroid update and convergence-control operations on the embedded ARM processor.
- A profiling-guided hardware/software partitioning approach that identifies the nearest-centroid distance computation as the main acceleration candidate.
- A lightweight FPGA kernel design using local centroid buffering, loop-level parallelism, and burst-compatible AXI memory access for accelerating the distance-computation stage.
- A host–kernel communication flow based on OpenCL command queues, non-blocking memory operations, and event-based synchronization to coordinate data migration and kernel execution.
- An embedded CPU–FPGA performance evaluation using synthetic two-dimensional floating-point datasets with varying numbers of data points and clusters, complemented by FPGA resource utilization analysis on the XC7Z020 device and validation on a real-world geospatial dataset to confirm the generalizability of the proposed accelerator.

The remainder of this study is organized as follows: Section II reviews the related work, including previous FPGA-based hardware acceleration approaches, K-means algorithm optimizations, and OpenCL programming for FPGAs. Section III presents the necessary background, including a detailed overview of the K-means clustering algorithm. Section IV describes the proposed methodology and the transition from the baseline implementation to the OpenCL-based FPGA-accelerated version. Section V presents the experimental setup, performance evaluation, and comparative analysis with existing works, followed by a discussion of the limitations. Finally, Section VI concludes the study by summarizing the main findings and outlining directions for future research.

II. Related Work

Due to the growing need to process large-scale datasets efficiently, scientific research has recently engaged in hardware acceleration for machine learning algorithms. Effective training of various machine learning algorithms, such as K-means, requires substantial computational resources and time. Therefore, these algorithms are often executed on FPGAs to take advantage of the power of reconfigurable hardware and accelerate their processing [7].

A. FPGA-Based Hardware Acceleration

FPGAs stand as a highly promising option for executing high-performance computations due to their remarkable attributes. They present a good balance between performance and power consumption, coupled with an interesting level of flexibility that enables users to configure these architectures to implement a customized application. The advantage of reconfigurability in FPGAs allows them to be fine-tuned to match the demands of the target computation. Recently, many studies have been published discussing software and hardware optimization and implementation methods of hardware accelerators using FPGAs [8], [9], [10], [11], [12]. Among them, our earlier work [8] introduced a preliminary approach, which the current study significantly extends and refines.

In addition, FPGAs have been presented as a mid-point solution providing a trade-off of GPU performance and ASIC efficiency while maintaining high flexibility and reconfigurability. They have been broadly recognized in the literature as being well-suited to accelerating machine learning. A review of more than 120 FPGA implementations of neural network accelerators demonstrates the wide generality of applications of FPGAs to diverse deep learning architectural styles, i.e., CNNs and RNNs, including ResNet-2 and LSTM [13]. The study also proposes a set of benchmark metrics and optimization techniques that cross many design classes with a useful starting point for benchmarking and optimizing accelerator performance.

As stated above, the effective utilization of machine learning methods demands a substantial amount of processing power. Even conventional multi-core CPUs are burdened by the computational complexity of these algorithms. In light of this, [14] proposed a reconfigurable architecture to deploy machine learning algorithms by implementing a Neural Network in an FPGA. The final results were compared with a Raspberry Pi (a conventional CPU) implementation. The proposed FPGA implementation architecture is 33646 times faster than the traditional CPU or core-based method.

In the ESA-funded CloudScout project, a bespoke FPGA-based accelerator was compared with the Intel Myriad 2 Vision Processing Unit (VPU) for on-board inference of a Convolutional Neural Network (CNN) for hyperspectral cloud detection in a CubeSat context [15]. The research showed that FPGA integration provided shorter inference time and greater flexibility than VPU, foreseeing the promise of FPGAs in resource-limited environments such as space. While the FPGA took more power and development time, its better performance and space qualification made it a suitable candidate for long-duration Low Earth Orbit (LEO) and interplanetary missions. These results demonstrate the applicability of FPGAs to a wide range of applications demanding effective on-board processing.

In addition to their benefits in deep learning applications, FPGAs have effectively sped up traditional machine learning algorithms. For instance, a recent study described a parameterized and scalable FPGA-based hardware accelerator for an embedded systems Support Vector Machine (SVM) classifier that relies on convex optimization [16]. With high classification accuracy, the architecture offers up to 79× acceleration over its software counterpart, allowing real-time operation and scalability to datasets of various sizes. This work validates

the feasibility of FPGAs as a general-purpose platform for machine learning acceleration by showing their adaptability and efficiency in accelerating data-intensive machine learning computations.

B. OpenCL for Heterogeneous Computing

Considerable progress in parallel computing has been made possible by the advent of heterogeneous computing systems. Given this, OpenCL has attracted attention as a standard that facilitates task-parallel and data-parallel heterogeneous computing on a variety of processing platforms, such as CPUs, GPUs, DSPs, and other coprocessor accelerators [17]. Its flexibility lies in the way it manages arithmetic-intensive data-parallel workloads on these various processing units. OpenCL's flexibility and efficiency for heterogeneous computing have been the subject of numerous research studies. Paper [18] provides an overview of the key architectural features of current microprocessor designs and outlines the OpenCL programming model as a standard developed for these architectures. Furthermore, OpenCL has been used to accelerate computationally demanding methods, such as Convolutional Neural Networks (CNNs). A study in [19] suggested using OpenCL to accelerate CNNs on Xilinx ZYNQ-7000 FPGAs. This approach achieved speedups of up to 10× over CPU implementations while maintaining comparable accuracy on well-known datasets such as MNIST and CIFAR-10. This illustrates OpenCL's ability to enable effective hardware acceleration without the complexity of conventional HDL design. It also supports the deployment of machine learning models for platforms with limited resources and rapid prototyping. Yongbon *et al.* have also presented a study analyzing the efficacy of OpenCL to speed up machine learning applications [20]. Their study's main goal was to create an implementation of Darknet, a deep learning-based object identification framework, that is hardware-compatible with universal accelerators. Darknet's limited utility resulted from its initial support of only NVIDIA CUDA for processing acceleration. Getting over the limitation and beating the original CUDA version is the primary objective of this research. The performance of OpenCL-Darknet was evaluated on two different AMD hardware setups: an AMD Radeon RX560 with OpenCL 1.2 and an AMD R7-integrated APU with OpenCL 2.0. For this assessment, the VOC 2007 dataset was used. Furthermore, they performed a performance comparison between OpenCL-Darknet and the original Darknet using an NVIDIA GTX 1050 with CUDA 8.0 and cuDNN 6.0. The findings revealed that compared to CUDA and cuDNN, OpenCL-Darknet exhibited a performance discrepancy of 2.5x to 4.4x for YOLOv2 and of 3.2x to 6.4x for TinyYOLO.

C. FPGA-Based Hardware Acceleration of K-means Algorithm with OpenCL

Tang *et al.* provides an exploration of the K-means algorithm via HLS of OpenCL code [21]. Due to inefficient memory bandwidth use, the initial implementation of the sequential baseline K-means algorithm performs poorly on the FPGA. The optimal approach relied on two kernels that utilized the FPGA's local memory. In the best scenario, speedups of up to 21× are obtained for implementations on an Altera Stratix V FPGA in comparison to an Intel Xeon Hexa-core processor.

In [22], a hybrid implementation of the K-means is presented to compare the speedup to an ARM CPU implementation. The proposed implementation carries out only some of the steps of the algorithm; the software carries out the remaining steps. The goal of this hybrid method is to reduce FPGA area overhead while increasing processing performance. The FPGA circuits are used to compute the similarity distance metric, which is based on the Manhattan distance, and to group data points based on their closest centroid. Moreover, the centroid values are updated by the software. Compared to the software version of the ARM processor, the experimental results reveal a speedup of about 10×.

In [23], the presented approach used High-Level Synthesis to test 10 distinct OpenCL implementations of the K-means algorithm on an FPGA platform. In addition to evaluating memory optimization strategies, including on-chip buffering and vectorization, the study investigated NDRange and task kernel models and examined performance across a variety of datasets and configurations. Their implementation showed energy savings of up to 80% and a speedup of increase to 1.54× over a high-performance CPU baseline.

III. K-means Clustering

The K-means clustering algorithm is one of the most popular unsupervised learning methods to partition datasets into unique groups by similarity. K-means has been used successfully across many research areas, including image processing [24], market analysis [25], data processing [26], medical image segmentation [27], risk evaluation [28], [29], etc. K-means is a partitional clustering algorithm, meaning that it partitions a dataset into a pre-defined number of non-overlapping subsets, or clusters. The core principle of K-means is to cluster data points to maximize intra-cluster similarity (data points are grouped together as closely as possible) and minimize inter-cluster similarity (data points in different clusters are distinguishable). In general, this is done by minimizing the sum of squared distances (errors) from an individual data point to the centrally located (mean) representative of its assigned cluster, which is where the K-means algorithm gets its name. The K-means procedure uses iteration where it initially provides a set of cluster centers and then performs alternating steps, where in each step the algorithm first assigns each point to the nearest cluster center, and then the next step updates each cluster center to be the mean of the points that were assigned to it until reaching convergence. Convergence is usually reached when either the assignments change no longer, or when the error is reduced to less than some previously assumed degree.

A. Problem Formulation

Given a dataset $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, where $\mathbf{x}_i \in \mathbb{R}^d$ represents the i-th data point in a d-dimensional space, the goal of K-means is to partition $\mathbf{X}$ into k clusters $\{C_1, C_2, \ldots, C_k\}$. The optimization objective is to minimize the within-cluster sum of squared distances (WCSSD):

$$\mathcal{J} = \sum_{j=1}^{k} \sum_{\mathbf{x}_i \in C_j} \|\mathbf{x}_i - \mu_j\|^2 \tag{1}$$

where, μ_j denotes the centroid of cluster C_j, calculated as:

$$\mu_j = \frac{1}{|C_j|} \sum_{\mathbf{x}_i \in C_j} \mathbf{x}_i \tag{2}$$

The optimization alternates between assigning data points to clusters and updating centroids until convergence.

B. Algorithm Steps

The K-means algorithm involves the following steps:

- Initialization: Choose k initial centroids, $\mu_1^{(0)}, \mu_2^{(0)}, \ldots, \mu_k^{(0)}$. Initialization methods include random selection or heuristic techniques like K-means++.
- Assignment Step: For each data point $\mathbf{x}_i$, determine its cluster assignment based on the nearest centroid:

$$c_i^{(t)} = \arg\min_j \|\mathbf{x}_i - \mu_j^{(t)}\|^2 \tag{3}$$

Alternatively, this can be represented using a binary indicator matrix $\mathbf{r}_{ij}$:

$$\mathbf{r}_{ij} = \begin{cases} 1, & \text{if } \mathbf{x}_i \text{ is assigned to cluster } j, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

- Update Step: Recalculate the centroids for each cluster based on the current assignments:

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^{N} \mathbf{r}_{ij} \mathbf{x}_i}{\sum_{i=1}^{N} \mathbf{r}_{ij}} \tag{5}$$

- Convergence Check: Repeat steps 2 and 3 until the number of observations whose cluster assignment changes between consecutive iterations falls below a predefined threshold.

C. Computational Complexity

K-means's computational complexity is affected by the number of data points N, the number of clusters k, the dimensionality d, and the number of iterations T. Each iteration requires $O(Nkd)$ computations for the assignment step and $O(kd)$ for updating centroids, leading to an overall complexity of:

$$O(TNkd)$$

While this complexity is manageable for small datasets, it becomes prohibitive for large-scale data due to the iterative nature of the algorithm and the repeated distance computations.

Despite its simplicity, the K-means algorithm has several limitations, particularly its low scalability as the dataset size increases. The computational cost of the algorithm, mainly due to distance calculations, increases rapidly as the number of data points and dimensions increases. However, in hardware implementations, these distance calculations, which dominate the execution time, can be efficiently accelerated using FPGAs. By offloading these intensive calculations to FPGAs, computational bottlenecks are reduced, leading to significant improvements in execution time. This hardware acceleration approach is particularly beneficial when dealing with large, high-dimensional datasets or real-time applications, as explored in subsequent sections of this study.

IV. Adopted Methodology

The proposed methodology is divided into three principal phases, as illustrated in Fig. 1. The first phase focuses on the design process of the overall application architecture, which describes the functional and structural components needed for the implementation. The second phase concentrates on the development of the OpenCL kernel, where computationally intensive tasks are defined and optimized for execution on the FPGA. The final phase involves establishing efficient communication between the host processor and the FPGA kernel through the use of OpenCL APIs, ensuring seamless data transfer and coordination between software and hardware components.

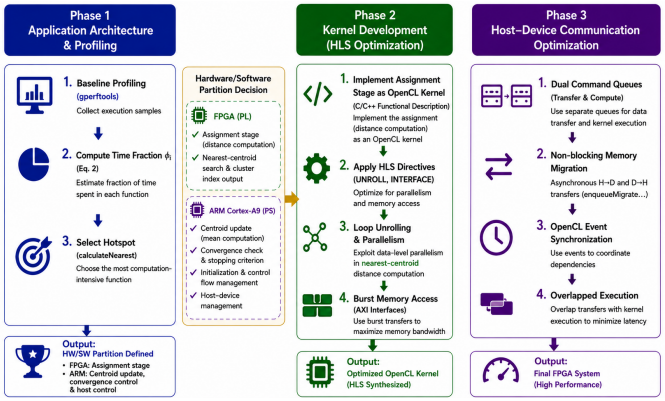


Fig. 1. Overview of the adopted methodology.

A. Application Architecture Design

Before embarking on the design of a hardware-accelerated application, it is imperative to conduct thorough architectural planning. This initial planning phase encompasses the development of an implementation plan, specifying the level of computational parallelism to achieve, and selecting which software tasks should be migrated to hardware as custom kernels. The process begins with a baseline profiling of the application's non-optimized CPU version. Profiling tools (e.g., GNU `gprof`) are used to extract the execution time of each function T_i^{CPU}. The fraction of time consumed by a given function i is then computed as:

$$\phi_i = \frac{T_i^{\text{CPU}}}{\sum_j T_j^{\text{CPU}}} \tag{6}$$

where, ϕ_i represents the relative contribution of function i to the total runtime. Functions with large ϕ_i values are natural candidates for FPGA acceleration.

The second step involves identifying the portions of the application that will experience the most significant performance improvement through the acceleration of relevant functions

in hardware. This selection process requires a critical evaluation of each function's computational load and its potential efficiency gains when implemented in hardware. Not every function will provide a meaningful acceleration by implementing the function in hardware, and it can conceivably cost the application overall performance. Therefore, the analysis in this iterative process is critical to ensure that hardware resources are optimized.

B. OpenCL Kernel Development

The second stage of the suggested process focuses on using OpenCL to convert the application's computationally demanding components into effective hardware kernels. In order to get notable performance improvements through hardware acceleration on FPGA platforms, this step is essential. Task selection, kernel implementation, and kernel optimization are the three primary sub-stages that comprise this phase. The first step in this phase is to identify functions or code segments that consume the most computational resources. These segments typically involve nested loops, large matrix computations, or repetitive arithmetic operations. Tools such as gprof, Google Performance Tools (gperftools), or platform-specific profilers such as Vitis Analyzer can be used to detect these hotspots. Once identified, the selected tasks are isolated and analyzed for suitability for parallelization and offloading. Priority is given to functions that operate on large datasets, exhibit high levels of parallelism (data or task parallelism), and have minimal dependencies between iterations. After task selection, the identified code blocks are rewritten as OpenCL kernels, which are functions that can be executed in parallel on the FPGA. The kernel design follows a load-compute-store structure:

- Load: Data are transferred from the host to the FPGA.
- Compute: The kernel performs the required operations on the hardware part.
- Store: The results are sent back to the host.

The OpenCL kernel is described at a high level using C/C++-based functional specifications. High-Level Synthesis (HLS) tools, such as Vitis HLS, automatically translate these functional descriptions into Register Transfer Level (RTL) implementations suitable for FPGA synthesis. The execution time of a kernel K on the FPGA can be expressed as:

$$T_K^{\text{FPGA}} = T_{\text{load}} + T_{\text{compute}} + T_{\text{store}} \tag{7}$$

where, T_{load} and T_{store} correspond to the data transfers between the host and the FPGA, and T_{compute} represents the kernel computation time. To enhance performance, HLS optimization directives (pragmas) are applied to control loop unrolling, pipelining, and data access patterns, thereby reducing T_{compute}. Additionally, techniques such as burst transfers, buffer reuse, and memory coalescing are employed to minimize T_{load} and T_{store}, ensuring efficient utilization of both computational and memory resources.

C. Code Adaptation: Host-Kernel Communication

Effective communication between the FPGA kernel and the host application is a crucial component of performance improvement in hardware-accelerated systems. While the host application coordinates overall execution, controls data transfers, and handles input/output activities, the kernel code, which is executed on the FPGA, carries out computationally demanding tasks. This section describes the methodology used to integrate the host and kernel components, ensuring efficient communication within the implementation of the K-means algorithm. As presented in Fig. 2, the host application is responsible for:

- Allocating memory buffers for input data (e.g., observations and cluster centroids) and output data (e.g., cluster assignments).
- Transferring data between the host and the FPGA device.
- Launching the kernel on the FPGA and monitoring its execution status.
- Retrieving results from the kernel and performing any post-processing required.

For this work, the host code was written in C/C++ and interfaced with the FPGA kernel using the OpenCL framework, which offers an abstraction layer that facilitates communication between the host and the kernel through a unified programming model.

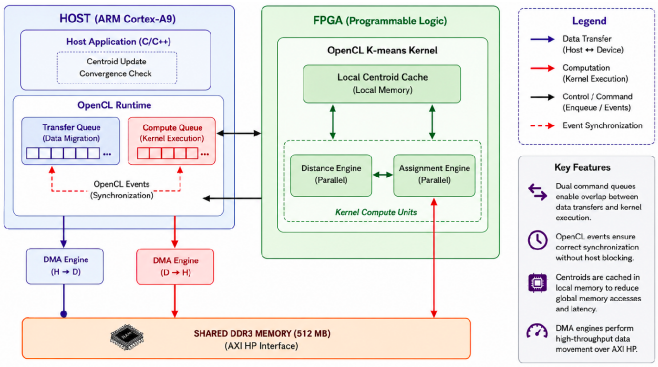


Fig. 2. OpenCL framework overview.

V. Performance Evaluation

A. Experimental Setup

The proposed heterogeneous K-means accelerator was implemented on a Xilinx ZedBoard featuring a Xilinx Zynq-7000 XC7Z020-CLG484-1 SoC integrating a dual-core ARM Cortex-A9 processing system (PS) and programmable logic (PL). The FPGA design was developed using the Xilinx OpenCL development flow under PetaLinux, while the software-only baseline was executed entirely on the embedded ARM Cortex-A9 processor. Consequently, all reported speedup values correspond to comparisons between the heterogeneous CPU–FPGA implementation and the ARM processor executing the same K-means algorithm under identical experimental conditions. The input data consisted of randomly generated

floating-point values. Experiments were conducted on datasets ranging from 4,000 to 30,000 data points with varying numbers of clusters, and performance was evaluated using kernel execution time measured in milliseconds. To further validate the proposed accelerator on real-world data, additional experiments were conducted using the SimpleMaps World Cities dataset. The hardware and software configuration used throughout the experiments is summarized in Table I.

TABLE I. HARDWARE AND SOFTWARE CONFIGURATION OF THE FPGA-BASED IMPLEMENTATION.

Parameter	Specification
FPGA board	Xilinx ZedBoard
FPGA device	XC7Z020-CLG484-1
Processing System	Dual-core ARM Cortex-A9
Operating system	PetaLinux
Development environment	Vitis Unified ide 2021.2
HLS tool	Vitis HLS 2021.2
OpenCL version	OpenCL 1.2
Target clock period / frequency	10 ns / 100 MHz

1) Selection of the acceleration candidate: The acceleration candidate was selected based on the computational structure of the K-means algorithm and preliminary profiling of the CPU implementation. In each iteration, the assignment stage compares every data point with every centroid. For a dataset with N points, k clusters, and dimensionality d, this stage requires $O(Nkd)$ distance computations. Therefore, the nearest-centroid computation is expected to dominate the execution time, especially when N or k increases.

In the implemented software version, this computation is performed by the `calculateNearest` function, which iterates over the centroids and computes the Euclidean distance between each data point and each centroid. For this reason, `calculateNearest` was selected as the main candidate for FPGA acceleration. Preliminary profiling confirmed that this function was the main hotspot; however, the profiling output is used only as qualitative guidance and not as an exact timing breakdown.

2) OpenCL kernel development: The selected `calculateNearest` function was implemented as an OpenCL kernel targeting the programmable logic of the Zynq device. The kernel computes the nearest centroid for each input observation and stores the corresponding cluster index in the output array. The centroid update and convergence-control operations remain executed on the ARM processor. This implementation incorporates various HLS directives that were strategically employed to guide the HLS tool during synthesis. The following directives were utilized to interface with the memory and control units:

a) Memory interfaces: `#pragma HLS INTERFACE m_axi` directives were employed for the input and output arrays. These directives define the memory interface for the input and output arrays, allowing seamless communication between the hardware accelerator and the external memory.

b) Control interfaces: `#pragma HLS INTERFACE s_axilite` directives were utilized for scalar control variables. These directives facilitate the creation of lightweight control registers for efficient communication between the host (Processing System PS) and the hardware accelerator (Programmable Logic PL).

Algorithm 1 K-means Kernel Algorithm

1.5em 0.75em
Input: `ptClusX`, `ptClusY`, `obsX`, `obsY` `numObservations`, `k`, `minD`
Output: `indices`
Initialize local arrays `localptClusX` and `localptClusY`
Copy cluster coordinates into local arrays
for $i = 0$ **to** `numObservations` $- 1$ **do**
    `localMinD` $\leftarrow$ `minD`
    `bestCluster` $\leftarrow 0$
    for $c = 0$ **to** `k` $- 1$ **do**
        `#pragma HLS UNROLL`
        $d_x \leftarrow$ `localptClusX[c]` $-$ `obsX[i]`
        $d_y \leftarrow$ `localptClusY[c]` $-$ `obsY[i]`
        `localDistance` $\leftarrow d_x^2 + d_y^2$
        if `localDistance` $<$ `localMinD` **then**
            `localMinD` $\leftarrow$ `localDistance`
            `bestCluster` $\leftarrow c$
        end if
    end for
    `indices[i]` $\leftarrow$ `bestCluster`
end for

In addition to the interface directives, other optimization directives were considered to enhance the algorithm's performance. The incorporation of the loop unrolling pragma (`#pragma HLS UNROLL`) emerged as a crucial strategy. Specifically, the loop unroll pragma was strategically applied to the loop responsible for computing the distance between data points and cluster centroids, as mentioned in the pseudocode below. This directive instructs the HLS tool to generate multiple copies of the loop body to reduce loop control overhead and enhance parallelism. It is particularly impactful in scenarios where the loop iterations are independent, as is often the case in distance calculations. By unrolling the loop, the compiler is empowered to exploit pipeline parallelism and optimize the overall execution time. K-means kernel algorithm is presented in Algorithm 1.

The independence of the data and the loop iterations served as the motivation behind the implementation of the (`#pragma HLS UNROLL`) directive within the distance computation method. Multiplication and accumulation operations are carried out on separate data elements in each iteration without creating dependencies on earlier or later iterations. This characteristic makes the loop highly amenable to unrolling, as the compiler can generate multiple parallel hardware instances that execute simultaneously. This takes advantage of the arithmetic resources on the FPGA giving higher throughput and less overall latency. Utilizing unrolling should be advantageous in clustering algorithms where distance calculations apply repeatedly to massive datasets, and this key aspect is a performance bottleneck.

3) Optimizing data communication between host and device: In hardware-accelerated systems, optimizing the data transfer between the host (CPU) and the FPGA is critical for achieving high performance. Therefore, this study seeks to improve the communication pipeline with asynchronous data transfers and non-blocking memory transactions. These two approaches are critical for reducing idle times and processing workloads concurrently between the host and the FPGA. We utilize asynchronous data transfers via two independent command queues in OpenCL: (`transfer_q`) for memory transfers, and (`compute_q`) for kernel executions. Both com-

mand queues are instantiated in out-of-order execution mode, allowing for overlapping transfers and kernel execution. This design reduces the waiting times for data to arrive before processing and ensures that the FPGA is continuously utilized. In this configuration, the transfer of the input data from host memory to the device memory and the kernel execution in the FPGA memory are non-blocking operations; this enables the accelerator to begin the computation as soon as it has enough of the required data to start, while the rest of the data transfers to host occur simultaneously. This structure will maximize the overall throughput of the system, since both computation and data movement will execute concurrently. To improve performance, we leverage non-blocking memory transactions. The OpenCL call - `enqueueMigrateMemObjects()` - facilitates the migration of data between host and device. Data transfer is performed asynchronously so that the host can do further work concurrently with the Data transfer. OpenCL events synchronize the different stages of the process without blocking. This allows the host to start executing the kernel as soon as the required data points have been received. Also, the host can continue reading the output data while the device writes to memory without waiting for all operations to complete. These techniques are essential for reducing idle time and ensuring that the data movement and computation phases are not bottlenecked by synchronization or waiting for data.

B. Experimental Results

1) Convergence criterion: To ensure a fair performance comparison, both the software-only and heterogeneous CPU–FPGA implementations were evaluated using the same K-means convergence criterion and termination condition. The algorithm iteratively alternates between the centroid recomputation and cluster assignment stages until convergence is achieved. Convergence is assessed by counting the number of observations whose cluster assignments differ between two consecutive iterations. Let $c^{(t)}$ denote the number of reassigned observations at iteration t. The iterative process terminates when the number of changed assignments falls below or equals a predefined threshold proportional to the dataset size N:

$$c^{(t)} \leq \left\lfloor \frac{N}{10000} \right\rfloor \tag{8}$$

Consequently, both implementations perform the same number of K-means iterations and satisfy an identical convergence condition, ensuring that the reported execution times and speedup measurements correspond to algorithmically equivalent workloads.

2) Execution time results of OpenCL kernel: In this section, we report the findings of our performance study, which compares the execution times of our heterogeneous solution with FPGA acceleration with regular CPU execution. The objective of this study is to show the efficiency of non-blocking memory transactions and asynchronous data transfer in enhancing the overall system performance. Performance evaluation was conducted through two distinct experiments using a randomly generated float dataset:

- Varying the Number of Clusters: The data points were kept fixed at 15,000 while the number of clusters was varied across a range of 32 to 256 (Fig. 3 and Table II).
- Varying the Number of Data Points: The cluster number was maintained at 128, and the number of data points was changed in the range of 4,000 to 30,000 (Fig. 4 and Table III).

To quantify the performance improvement of the FPGA implementation over the CPU, we calculate the **overall speedup** as the ratio of the CPU execution time to the FPGA execution time. Formally, the Speedup is defined as:

$$Speedup = \frac{T_{\text{CPU}}}{T_{\text{FPGA}}} \tag{9}$$

where, T_{CPU} and T_{FPGA} represent the execution times on the CPU and FPGA, respectively. A speedup greater than 1 indicates that the FPGA implementation is faster than the CPU.

In the initial experiment, as presented in Table II and illustrated in Fig. 3, the execution times of the CPU and FPGA implementations were evaluated while maintaining a constant dataset size (15,000 data points) and varying the number of clusters k. The results demonstrated a pronounced divergence in performance between the two architectures.

TABLE II. FPGA VS. CPU K-MEANS KERNEL PERFORMANCE WITH 15000-POINT DATASET.

k	FPGA+CPU time (ms)	CPU time (ms)	Speedup
32	3.177	16.543	5.2x
64	3.206	32.353	10.1x
128	3.223	64.149	20x
256	3.25	127.613	39.26x

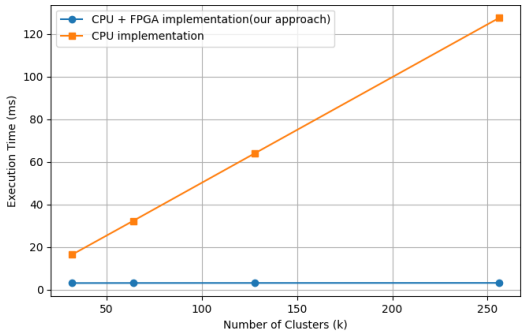


Fig. 3. FPGA vs. CPU K-means kernel performance with 15000-point dataset.

The CPU experienced a marked increase in execution time with increasing cluster counts, which can be attributed to the algorithmic processing complexity that increased cardinally with cluster count, thereby demanding more processing time. The FPGA implementation, by contrast, demonstrated much more consistent and scalable performance. Although a marginal increase in execution time was observed with additional clusters, the FPGA's parallel processing capabilities and optimized memory management mitigated the impact of the rising computational load, preventing a proportional escalation in runtime. Comparatively, the FPGA implementation achieved a speedup of up to 39X compared to the CPU, which is particularly useful for clustering, which is inherently computationally intensive.

In the second experiment, shown in Fig. 4 and summarized in Table III, the number of clusters was held constant (k=128) while the dataset size was scaled from 4,000 to 30,000 data points.

TABLE III. FPGA VS. CPU K-MEANS ALGORITHM PERFORMANCE WITH K=128

N	FPGA+CPU time (ms)	CPU time (ms)	Speedup
4000	2.706	17.08	6.3x
8000	2.792	34.215	12.25x
10000	2.875	42.774	14.87x
12000	3.078	50.323	16.34x
14000	3.203	59.843	18.68x
16000	3.616	68.361	18.9x
20000	3.883	85.522	22.02x
30000	4.258	128.171	30x

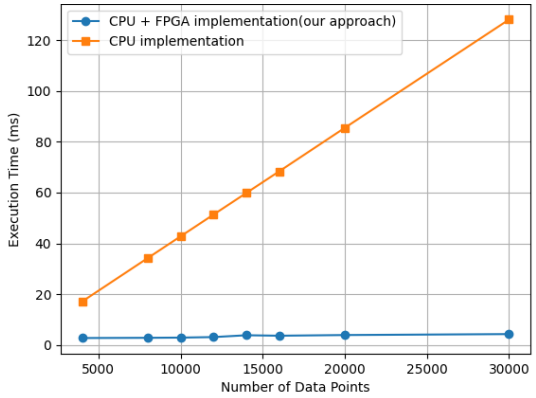


Fig. 4. FPGA vs. CPU K-means algorithm performance with k=128.

Consistent with the first experiment, the CPU exhibited a steep rise in execution time as the dataset expanded. This is largely attributed to the inherent limitations in CPU-based architectures to deal with large amounts of data, where the overhead of sequential processing and data transfer becomes a tremendous bottleneck. On the other hand, the heterogeneous execution model based on FPGA presented better scalability in big data processing. With asynchronous data transfer and non-blocking memory access, the FPGA initiated kernel execution as early as there was partial data, thus facilitating computation in parallel with ongoing data transfer. This was an optimization that significantly lowered the overall processing latency. When dealing with bigger datasets, the performance improvements were more noticeable, and the FPGA outperformed the CPU by up to 30X, which supported the findings from the initial trial. The results demonstrate the architectural advantages of FPGAs for carrying out data-intensive tasks since they can transfer data and conduct computations simultaneously, offloads the inefficiencies of conventional CPU-based processing.

To analyze the performance and reliability of the proposed K-means kernel on FPGA, we performed 10 independent executions for a dataset with a fixed problem size of N=30,000 data points and k=128 clusters. As illustrated in Table IV and Fig. 5, execution time for the CPU ranged from 128.206 ms to 128.596 ms, while FPGA execution times ranged from 4.066 ms to 4.723 ms. The resultant speedup, defined as the CPU execution time divided by FPGA execution time,

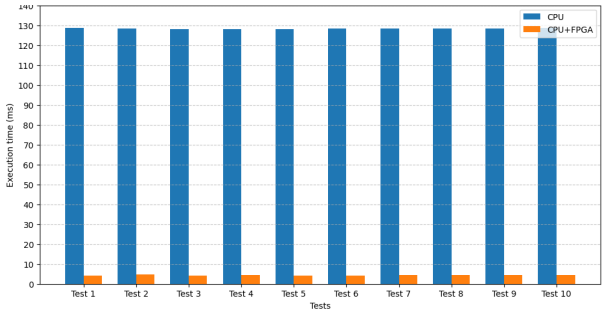


Fig. 5. K-means Kernel execution time (N=30000, k=128) for 10 tests.

ranged from 27.16× to 31.53×, with an average of 29.5×. These results show that the FPGA implementation constantly provided improved speed above the CPU, with stable and reproducible execution times, thereby confirming its suitability for real-time applications.

TABLE IV. CPU AND FPGA EXECUTION TIMES FOR 10 TESTS (N=30,000, K=128).

Test	CPU Time (ms)	CPU + FPGA Time (ms)	Speedup
1	128.849	4.225	30.49x
2	128.277	4.723	27.16x
3	128.206	4.188	30.61x
4	128.207	4.311	29.73x
5	128.211	4.066	31.53x
6	128.390	4.123	31.13x
7	128.275	4.505	28.47x
8	128.293	4.467	28.72x
9	128.324	4.541	28.25x
10	128.596	4.452	28.88x

3) Resource utilization: We also analyzed the FPGA resource consumption of the designed K-means kernel. Table V summarizes the hardware resource utilization of the proposed FPGA accelerator. The design occupies only 7.74% of the available LUTs, 5.17% of the registers, 0.71% of BRAM resources, and 3.18% of the DSP slices available on the XC7Z020 device. These results indicate that the proposed accelerator achieves substantial performance improvements while utilizing only a small fraction of the FPGA resources. The low BRAM utilization demonstrates that the design does not require extensive on-chip storage, while the limited DSP usage confirms that the implementation remains lightweight and scalable. Such moderate resource consumption leaves significant headroom for future optimizations, including the replication of processing elements, support for higher-dimensional datasets, or the integration of multiple clustering kernels within the same device. The reported speedups of up to 39× are therefore achieved without exhausting the available FPGA resources, highlighting a favorable performance-to-resource trade-off. Beyond resource consumption, post-implementation timing analysis confirms that the design reliably meets its performance target. The kernel was constrained to a 100 MHz (10 ns period) operating frequency, and static timing analysis after place-and-route reported all timing constraints as met, with a worst-case (critical path) positive margin of 0.754 ns. This confirms that the reported execution times and speedups correspond to a design that is fully timing-closed at its target operating frequency.

TABLE V. HARDWARE RESOURCE UTILIZATION OF THE K-MEANS KERNEL ON FPGA.

Resource Type	Used	Available	Utilization (%)
Slice LUTs	4118	53200	7.74
– LUTs as Logic	3558	53200	6.69
– LUTs as Memory	560	17400	3.22
Slice Registers	5498	106400	5.17
– Registers as Flip-Flop	5498	106400	5.17
Block RAM (BRAM)	1	140	0.71
DSP Slices (DSP48E1)	7	220	3.18

The finalized FPGA device layout with the K-means kernel implementation is shown in Fig. 6. The embedded ARM processor is visible in the left corner of the chip in the figure. The close integration of the processing system and programmable logic is demonstrated by this architectural layout, which also highlights an effective design approach specifically designed to improve the K-means algorithm's performance.

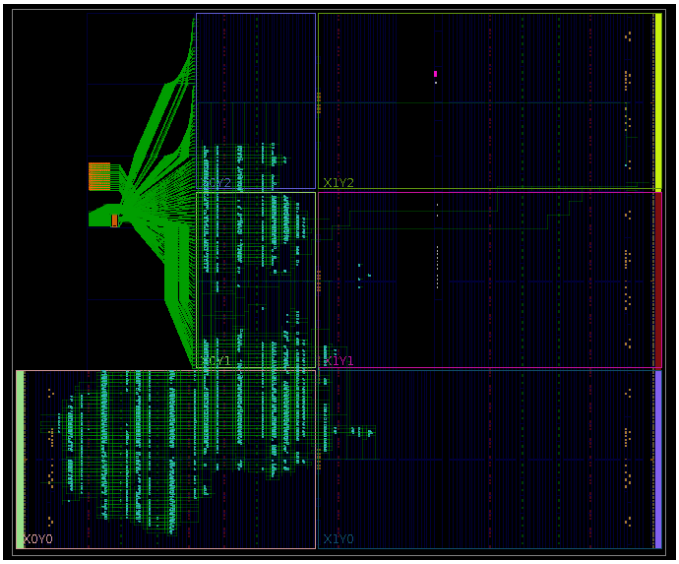


Fig. 6. The FPGA chip floorplan of the K-means kernel implementation.

4) Power consumption analysis: The proposed FPGA-based implementation exhibits a total on-chip power consumption of 1.833 W, including 1.688 W of dynamic power and 0.145 W of static power, as illustrated in Fig. 7.

The Processing System (PS7) accounts for approximately 90% of the dynamic power consumption, while the remaining power is distributed among the programmable logic resources, including clocks, signals, logic, DSP blocks, BRAM, and clock management circuitry. This behavior is consistent with the proposed hardware–software co-design approach, where the ARM Cortex-A9 processor is responsible for host execution, memory management, and OpenCL runtime control, whereas the FPGA fabric accelerates only the computationally intensive distance calculation kernel.
Despite the dominant contribution of the processing system, the overall power consumption remains below 2 W, demonstrating the suitability of the proposed architecture for power-constrained embedded platforms.

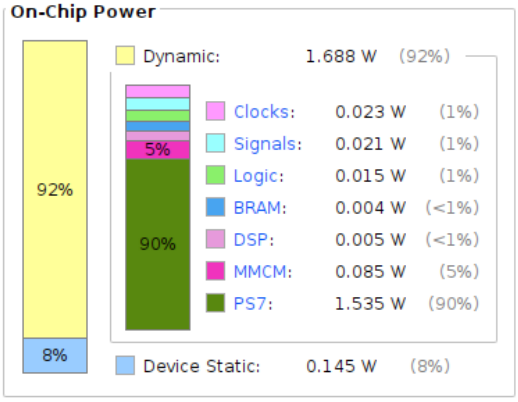


Fig. 7. On-chip power consumption of the proposed hardware-software co-design implementation.

5) Evaluation on a real-world dataset: To evaluate the proposed CPU–FPGA accelerator under realistic operating conditions, experiments were additionally conducted using the SimpleMaps World Cities dataset. This dataset contains real-world geospatial data with naturally non-uniform spatial distributions, providing a more representative benchmark for assessing the clustering performance on practical applications.

The dataset consists of 50,250 populated places worldwide. For the K-means clustering task, only the two numerical spatial attributes, latitude and longitude, were used as input features. The characteristics of the dataset are summarized in Table VI.

TABLE VI. CHARACTERISTICS OF THE SIMPLEMAPS WORLD CITIES DATASET.

Property	**Description**
Dataset Type	Real-world geospatial dataset
Instances	50,250 cities
Dimensions Used	2 (Latitude, Longitude)
Original Attributes	City name, country, administrative region, latitude, longitude, population, capital status, etc.
Domain	Geographic locations worldwide

The same experimental methodology adopted for the synthetic datasets was applied to the real-world dataset. Table VII reports the execution times obtained for the proposed CPU–FPGA implementation and the software-only CPU implementation.

TABLE VII. EXECUTION TIME COMPARISON ON THE SIMPLEMAPS WORLD CITIES DATASET.

N	k	FPGA+CPU Time (ms)	CPU Time (ms)	Speedup
50 250	128	5.4	214.757	$\sim 39.7\times$

The experimental results demonstrate that the proposed accelerator maintains its performance benefits on real-world data, achieving an approximate 39× speedup over the software-only implementation. This performance is consistent with the gains observed on the synthetic datasets, despite the larger size and more irregular spatial distribution of the real-world dataset, demonstrating the scalability and generalization capability of the proposed architecture.

TABLE VIII. COMPARISON OF THE PROPOSED IMPLEMENTATION WITH REPRESENTATIVE FPGA/OPENCL-BASED K-MEANS ACCELERATION APPROACHES.

Reference	Platform	Programming Model	Baseline	Dataset / Precision	Acceleration Scope	Speedup
Tang & Khalid [21]	Altera Stratix V A7	OpenCL	6-core Intel Xeon W3670	Benchmark suite / Integer	Complete K-Means using two OpenCL kernels with local-memory optimization	3×–21×
Canilho *et al.* [22]	FPGA Hybrid Platform	RTL/HLS	ARM CPU	Not specified	Partial acceleration: distance computation and assignment in hardware; centroid update in software	≈10×
Paulino *et al.* [23]	FPGA	OpenCL	High-performance CPU	Integer datasets / Integer	Complete K-Means using multiple OpenCL task and NDRange kernel implementations	Up to 1.54× (+80% energy savings)
Baydoun *et al.* [30]	GPU	CUDA / OpenMP	Sequential CPU	Multiple datasets and feature dimensions	Complete software implementation (GPU and multi-core CPU)	≈2× (OpenMP), ≈6× (CUDA)
Proposed Work	**Xilinx Zynq-7000 XC7Z020**	**OpenCL**	**Embedded ARM Cortex-A9**	**Synthetic 2D /SimpleMaps World Cities dataset/ Floating-point**	**Partial acceleration: nearest-centroid assignment on FPGA; centroid update and convergence control on ARM**	**Up to 39× (29.5× average)**

6) Performance comparison: Table VIII compares the proposed implementation with representative FPGA/OpenCL-based K-Means accelerators. Direct comparison of the reported speedups should be interpreted with caution because the evaluated studies differ in FPGA architecture, processor baseline, numerical precision, dataset characteristics, and implementation scope. For instance, Tang and Khalid [21] and Paulino *et al.* [23] implement complete K-Means accelerators, whereas the proposed architecture adopts a heterogeneous hardware/software partitioning in which only the computationally dominant assignment stage is offloaded to the FPGA, while centroid recomputation, convergence checking, and execution control remain on the embedded ARM processor. This partitioning is a deliberate design choice that exploits the FPGA for massively parallel, compute-intensive operations while retaining the control-oriented and less computationally demanding tasks on the processor, thereby reducing hardware complexity and resource utilization without sacrificing overall performance. Furthermore, several previous FPGA implementations rely on integer arithmetic, whereas the proposed design operates on floating-point data, which generally incurs higher computational and hardware costs. In addition, the reported speedups are obtained using different processor baselines, ranging from embedded ARM processors to high-performance desktop CPUs. Consequently, the reported performance figures should be interpreted within the context of each implementation rather than as direct performance rankings.

7) Limitations and scalability discussion: Although the proposed heterogeneous architecture achieves substantial acceleration for two-dimensional floating-point datasets, several limitations should be acknowledged. First, the current implementation is specialized for two-dimensional observations, and extending the design to higher-dimensional feature vectors requires additional arithmetic operations for each distance evaluation. Consequently, the computational complexity of the assignment stage increases proportionally with the feature dimension, leading to higher DSP, LUT, and routing requirements when parallelism is maintained. Moreover, larger feature vectors increase the amount of data transferred between external memory and the FPGA, making memory bandwidth progressively more critical and potentially limiting the achievable acceleration. Nevertheless, the adopted hardware/software partitioning remains applicable because the assignment stage continues to dominate the computational cost of the K-Means algorithm, while the centroid update and convergence-control operations remain comparatively lightweight and are efficiently handled by the embedded ARM processor. Future work will therefore investigate scalable kernel architectures based on parameterized dimensionality, data tiling, and optimized memory hierarchies to improve performance for high-dimensional clustering problems.

VI. CONCLUSION

This study presented an OpenCL-based FPGA acceleration of the K-means clustering algorithm on a Xilinx Zynq platform. The proposed heterogeneous architecture offloads the computationally intensive distance calculation stage to programmable logic while retaining control-oriented tasks on the embedded processor. This hardware/software partitioning enables efficient exploitation of parallelism while preserving implementation flexibility. Experimental evaluation demonstrated speedups of up to 39× compared with the CPU implementation on synthetic benchmarks, with consistent performance gains confirmed on a real-world geospatial dataset, supporting the generalizability of the proposed approach beyond controlled synthetic conditions. Resource utilization further showed that these performance gains are achieved with modest consumption of LUTs, BRAMs, and DSP slices. The results confirm the suitability of FPGA-based acceleration for computationally intensive clustering workloads. Our study indicates that a hybrid CPU–FPGA architecture is highly effective for accelerating iterative, data-intensive algorithms such as K-means. Future research will aim to expand the proposed methodology to handle more complex datasets, as well as more sophisticated algorithms, to thoroughly evaluate its scalability and performance in demanding scenarios. We also aim to explore the integration of multi-FPGA systems and distributed processing architectures to better exploit parallelism and accelerate deep learning algorithms. Furthermore, we will investigate advanced methods to optimize off-chip memory access to reduce data transfer latency and improve overall efficiency. These guidelines are meant to augment and expand

current design practice. They help in creating energy-efficient and high-speed hardware accelerators for a range of machine learning and deep learning applications.

Declaration on Generative AI

Generative AI tools were used exclusively to improve the grammar, language, and readability of this manuscript. The authors reviewed and edited all AI-assisted text and take full responsibility for the content of the manuscript.

References

[1] A. E. Ezugwu, A. M. Ikotun, O. O. Oyelade, L. Abualigah, J. O. Agushaka, C. I. Eke, and A. A. Akinyelu, *A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects*, Engineering Applications of Artificial Intelligence, vol. 110, pp. 104743, 2022.

[2] G. J. Oyewole and G. A. Thopil, *Data clustering: application and trends*, Artificial Intelligence Review, vol. 56, no. 7, pp. 6439–6475, 2023.

[3] Z. Yu, *Big data clustering analysis algorithm for internet of things based on K-means*, International Journal of Distributed Systems and Technologies (IJDST), vol. 10, no. 1, pp. 1–12, 2019.

[4] C. Jiang, J. Wan, and H. Abbas, *An edge computing node deployment method based on improved k-means clustering algorithm for smart manufacturing*, IEEE Systems Journal, vol. 15, no. 2, pp. 2230–2240, 2020.

[5] Y. Yang, Q. Chen, T. Yi Huang, and P. K. Pareek, *Application research of K-means algorithm based on big data background*, in 2023 IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS), pp. 1–5, 2023.

[6] P. Mackey and R. R. Lewis, *Parallel k-means++ for multiple shared-memory architectures*, in 2016 45th International Conference on Parallel Processing (ICPP), pp. 93–102, 2016.

[7] D. G. Bailey, *Image processing using FPGAs*, Journal of Imaging, vol. 5, no. 5, pp. 53, 2019.

[8] E. Jaballi, S. Gdaim, and N. Liouane, *Design and implementation of FPGA-based hardware acceleration for Machine Learning using OpenCL: A case study on the K-Means algorithm*, in 2024 International Conference on Control, Automation and Diagnosis (ICCAD), pp. 1–6, 2024.

[9] S. Mittal, *A survey of FPGA-based accelerators for convolutional neural networks*, Neural Computing and Applications, vol. 32, no. 4, pp. 1109–1139, 2020.

[10] K. Guo, S. Zeng, J. Yu, Y. Wang, and H. Yang, *[DL] A survey of FPGA-based neural network inference accelerators*, ACM Transactions on Reconfigurable Technology and Systems (TRETS), vol. 12, no. 1, pp. 1–26, 2019.

[11] A. Shawahna, S. M. Sait, and A. El-Maleh, *FPGA-based accelerators of deep learning networks for learning and classification: A review*, IEEE Access, vol. 7, pp. 7823–7859, 2018.

[12] T. Wang, C. Wang, X. Zhou, and H. Chen, *A survey of FPGA based deep learning accelerators: Challenges and opportunities*, arXiv preprint arXiv:1901.04988, 2018.

[13] A. Nechi, L. Groth, S. Mulhem, F. Merchant, R. Buchty, and M. Berekovic, *Fpga-based deep learning inference accelerators: Where are we standing?*, ACM Transactions on Reconfigurable Technology and Systems, vol. 16, no. 4, pp. 1–32, 2023.

[14] A. Itagi, S. Krishvadana, K. P. Bharath, and M. R. Kumar, *FPGA architecture to enhance hardware acceleration for machine learning applications*, in 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), pp. 1716–1722, 2021.

[15] E. Rapuano, G. Meoni, T. Pacini, G. Dinelli, G. Furano, G. Giuffrida, and L. Fanucci, *An fpga-based hardware accelerator for cnns inference on board satellites: benchmarking with myriad 2-based solution for the cloudscout case study*, Remote Sensing, vol. 13, no. 8, pp. 1518, 2021.

[16] S. Ramadurgam and D. G. Perera, *An efficient fpga-based hardware accelerator for convex optimization-based svm classifier for machine learning on embedded platforms*, Electronics, vol. 10, no. 11, pp. 1323, 2021.

[17] S. Tatsumi, M. Hariyama, K. Ito, and T. Aoki, *An FPGA accelerator for PatchMatch multi-view stereo using OpenCL*, Journal of Real-Time Image Processing, vol. 17, pp. 215–227, 2020.

[18] J. E. Stone, D. Gohara, and G. Shi, *OpenCL: A parallel programming standard for heterogeneous computing systems*, Computing in Science & Engineering, vol. 12, no. 3, pp. 66, 2010.

[19] S. Gdaim and A. Mtibaa, *Accelerating convolutional neural networks on FPGA platforms: a high-performance design methodology using OpenCL*, Journal of Real-Time Image Processing, vol. 22, no. 2, pp. 67, 2025.

[20] Y. Koo, C. You, and S. Kim, *OpenCL-Darknet: an OpenCL implementation for object detection*, in 2018 IEEE International Conference on Big Data and Smart Computing (BigComp), pp. 631–634, 2018.

[21] Q. Y. Tang and M. A. S. Khalid, *Acceleration of k-means algorithm using altera sdk for opencl*, ACM Transactions on Reconfigurable Technology and Systems (TRETS), vol. 10, no. 1, pp. 1–19, 2016.

[22] J. Canilho, M. Véstias, and H. Neto, *Multi-core for K-means clustering on FPGA*, in 2016 26th International Conference on Field Programmable Logic and Applications (FPL), pp. 1–4, 2016.

[23] N. Paulino, J. C. Ferreira, and J. M. P. Cardoso, *Optimizing opencl code for performance on fpga: k-means case study with integer data sets*, IEEE Access, vol. 8, pp. 152286–152304, 2020.

[24] X. Zheng, Q. Lei, R. Yao, Y. Gong, and Q. Yin, *Image segmentation based on adaptive K-means algorithm*, EURASIP Journal on Image and Video Processing, vol. 2018, no. 1, pp. 1–10, 2018.

[25] M. Tleis, R. Callieris, and R. Roma, *Segmenting the organic food market in Lebanon: an application of k-means cluster analysis*, British Food Journal, vol. 119, no. 7, pp. 1423–1441, 2017.

[26] H. Hu, J. Liu, X. Zhang, and M. Fang, *An effective and adaptable K-means algorithm for big data cluster analysis*, Pattern Recognition, vol. 139, pp. 109404, 2023.

[27] H. Yadav, P. Bansal, and R. K. Sunkaria, *Color dependent K-means clustering for color image segmentation of colored medical images*, in 2015 1st International Conference on Next Generation Computing Technologies (NGCT), pp. 858–862, 2015.

[28] M. Er Kara and S. Ü. Oktay Fırat, *Supplier risk assessment based on best-worst method and K-means clustering: a case study*, Sustainability, vol. 10, no. 4, pp. 1066, 2018.

[29] H. Xu, C. Ma, J. Lian, K. Xu, and E. Chaima, *Urban flooding risk assessment based on an integrated k-means cluster algorithm and improved entropy weight method in the region of Haikou, China*, Journal of Hydrology, vol. 563, pp. 975–986, 2018.

[30] M. Baydoun, H. Ghaziri, and M. Al-Husseini, *CPU and GPU parallelized kernel K-means*, The Journal of Supercomputing, vol. 74, no. 8, pp. 3975–3998, 2018.

Challenges of AI Adoption in the Saudi Arabian Healthcare Sector: A Comprehensive Study

Radwan M. Batyha[1], Ahmed Almaghthawi[2]*, Fahad Alharbi[3], Suhaila Abuowaida[4]
Department of Computer Science, Applied Science Private University, Amman, Jordan[1]
School of Computer and Information Sciences, University of Newcastle, NSW, Australia[2]
Department of Information Technology, College of Computers and Information Technology, Taif University, Saudi Arabia[3]
Department of Data Science and Artificial Intelligence, Faculty of Prince Al-Hussein
Bin Abdallah II for Information Technology, Al al-Bayt University, Mafraq 25113, Jordan[4]

***Abstract*—This research examines the challenges of AI adoption in Saudi Arabia healthcare sector in terms of the framework defined by the Vision 2030 healthcare transformation agenda. We utilized the Analytic Hierarchy Process (AHP) methodology to collect and analyze data from 27 managerial-level executives tasked with AI implementation in Saudi Arabia healthcare organizations. The study identifies and prioritizes critical challenges across five main dimensions. We discuss accuracy, privacy and security, ethical barriers, interpretability, and control. By means of systematic analysis, we discovered that accuracy (with priority weight 0.50) and privacy-security concerns (0.24), are the two most important factors in determining AI adoption in Saudi Arabian healthcare. The research offers evidence-based policy recommendations and health care administrator guidelines based on Saudi Arabia's cultural, regulatory, and technological landscape. The insights are especially important for providing guidance on incorporating AI into the Kingdom's extensive healthcare transformation program.**

***Keywords*—Artificial Intelligence (AI); healthcare transformation; Saudi Arabia Vision 2030; Analytic Hierarchy Process (AHP); Saudi Arabia**

I. Introduction

Artificial Intelligence (AI) in healthcare is becoming a game changer in the modern medical format [1], [2], [3], [4], [5], [6], [7], [8], moving Saudi Arabia to the top of this technological revolution. With Vision 2030 as part of its journey, Saudi Arabia is undertaking a vision to revolutionize its healthcare system through its digital transformation, AI being key towards this evolution [9]. In the journey towards this path, 2019 is the year of the establishment of Saudi Arabia Data and Artificial Intelligence Authority (SDAIA) [10], which is a notable step forward by the Kingdom to become a global leader in AI adoption [11].

In the Saudi Arabia healthcare sector, we have a very specific context for AI implementation: a country with large scale reforms ongoing, substantial government investment, and unique cultural considerations. Following the development of such initiatives as the National Platform for Health Information Exchange Services (NPHIES) and AI powered health services through the Health Virtual Hospital, among others [12], [13], [14], the Kingdom has been developing its healthcare AI adoption.

[1]Corresponding author

Specify and classify the leading challenges faced by the adoption of AI in Saudi Arabia healthcare sector [15], [16], [17], [18]. The AHP methodology is applied to analyze the relative importance of these challenges and examine the effect of cultural and regulatory factors in play, particularly Saudi Arabia's to force in modern medicine, with Saudi Arabia positioning itself at the forefront of this technological revolution. As part of Vision 2030, Saudi Arabia has embarked on an ambitious journey to revolutionize its healthcare system through digital transformation, with AI serving as a cornerstone of this evolution [9], [19], [20]. The establishment of the Saudi Arabia Data and Artificial Intelligence Authority (SDAIA) in 2019 marks a significant milestone in this journey, demonstrating the Kingdom's commitment to becoming a global leader in AI adoption [10], [21], [22].

The Saudi Arabian healthcare sector presents a unique context for AI implementation, characterized by rapid modernization, substantial government investment, and distinct cultural considerations. Recent initiatives, including the National Platform for Health Information Exchange Services (NPHIES) and AI-powered health services through the Health Virtual Hospital, showcase the Kingdom's progress in healthcare AI adoption [23]. However, this transformation journey faces multiple challenges that require careful consideration and systematic analysis [24], [25], [26], [27], [28], [29], [30], [31], [32], [33].

This study aims to:

- Identify and prioritize the key challenges affecting AI adoption in Saudi Arabia's healthcare sector.
- Analyze the relative importance of these challenges using the AHP methodology.
- Examine the influence of cultural and regulatory factors specific to Saudi Arabia.
- Offer evidence base recommendations for policymakers and healthcare administrators.

II. Literature Review

A. Current Landscape of AI in Saudi Arabian Healthcare

In recent years, the Saudi Arabian healthcare system has experienced a remarkable change for the better, in its adoption of Artificial Intelligence technologies. As was highlighted in [34], this transformation is very signed with Vision 2030's goal

of having a more efficient, more accessible healthcare system. Creating a structured approach to the implementation of the AI across the healthcare sector is the genesis of establishment of the Saudi Arabia Data and Artificial Intelligence Authority (SDAIA) [10]. Thanks to this governmental support, we've developed diagnostic tools and administrative systems to name a few of the AI initiatives that have come up during this shift from the traditional path towards healthcare delivery methods.

As illustrated by the most recent research of [35], Saudi Arabian healthcare organizations have been very successful in the application of AI solutions in multiple medical specialties. According to their study, machine learning algorithms have been quickly adopted by radiology departments, mostly for analyzing and supporting diagnosis of images via use of AI technology. This adoption pattern predates a broader trend that has been emerging in Saudi Arabian health care – one where technological innovation is being increasingly viewed as an indispensable requisite for continuous improvement in the delivery of service and operational efficiency.

The Ministry of Health's strategic planning documents mention the integration of AI into the complete approach [15]. An outcome of this systematic approach is successful deployment of several of the key initiatives, such as the AI-based telemedicine platforms, as well as automated patient scheduling systems, that have all jointly contributed to increasing access to healthcare within the Kingdom.

B. Integration and Technology Infrastructure

The integration of repetitive manufacturing equipment presents challenges related to both the computer control and rigid mechanical details of the robot manipulator units, as well as materials transport from the source to the robot manipulator units.

The integration of AI systems with Saudi Arabian healthcare organizations' infrastructure is technically complex beyond just having basic infrastructure [36]. They note the special complexity in designing such systems that must work well in Arabic, as well as English, while maintaining conformity to international medical norms and local health standards.

As seen in the literature, data standardization becomes a critical challenge as clinical documentation pattern across healthcare facilities vary, making the construction of unified AI systems a challenge [37]. However, their research demonstrates the significance of standardizing data collection and storage protocols so that AI systems can efficiently analyze and utilize patient information in various healthcare settings.

C. Privacy, Security and Regulatory Framework

There are privacy and security considerations in Saudi Arabian healthcare AI systems, along with cultural, religious, and regulatory requirements that make private and security challenges to be analyzed of the Saudi Arabian healthcare organizations on how they need to cope with complex privacy requirements in implementing AI systems [38]. The research shows that privacy concerns go further than classic data protection measures and extend to cultural expectations of patient confidentiality, as well as information sharing.

The regulatory framework governing the application of AI in Saudi Arabian healthcare is changing, as [39] describes. In fact, their analysis reveals that healthcare organizations have to balance the tension between innovation and compliance and between global standards and local laws. System design and implementation, in such cases, is further complicated by this dual compliance requirement, which applies to both the physical and logical portions of the system, especially when patient data is both sensitive and non immunologically represented while covering different healthcare settings.

1) Cultural and ethical considerations: The cultural context of Saudi Arabia significantly shapes the adoption of AI in healthcare settings [19], [20], which studies the details of how Islamic medical ethics and AI applications make an intersection in healthcare. Their research demonstrates that successful adoption of AI requires consideration of religious and cultural values, such as the provision of gender-specific healthcare delivery and patient privacy. In the context of Saudi Arabia healthcare, these considerations carry over to the design, implementation, and utilization of AI systems. Finally, [40] further elaborates on the necessity of cultural sensitivity in the implementation of AI, since cultural acceptance of the use of AI technologies by doctors and patients is strongly dependent on cultural factors. Their research shows that successful AI adoption requires a mix of cultural acceptance, adoption of working practices, and cultural norms and values, including patient/provider relationships and medical decision-making processes.

2) Workforce development and organizational change: There has been a lot of literature regarding the importance of workforce development in successful implementation of AI. In [41], the authors help us understand how Saudi Arabian healthcare organizations have tackled workforce development in the context of AI adoption. In their research, they say that successful implementation requires cultural adaptation to new technologies and in-depth training programs that cover both technical skills and that new technology. The study [42] finds that patient care organizations in Saudi Arabia deploy digital transformation to adopt AI-enabling systems and reveals its reliance on leadership support and clear communication strategies. This study indicates that organizations need to take a complete approach to AI adoption by considering the technical parts as well as the human parts.

D. Economic and Resource Considerations

While economic aspects of AI adoption in Saudi Arabian healthcare have both opportunities and challenges [43], the article under the line outlines the amount of investments that would be necessary to be able to successfully implement AI, as well as explaining the long term cost benefits. Our findings suggest that while the initial investment is substantial, the resulting potential for better operational efficiency and better patient outcomes makes this a worthwhile investment. Research by [44] presents a systematic review of resource allocation challenges of AI adoption within health care organizations, comparing and contrasting how these organizations find the balance between competing priorities in technology implementation. Those findings suggest that successful AI adoption will rely on an allocation of resources among the

many aspects of implementation, from infrastructure development to staff training to system maintenance.

E. Future Directions and Emerging Trends

The literature shows a number of social trends in the use of AI in Saudi Arabian healthcare. More applications of AI in preventive care and the gradual integration of AI systems with population health management systems are the main areas for future growth, according to [45]. The final part of their research suggests that future developments will most likely be about making more integrated and comprehensive AI solutions that are, in turn, dedicated to both clinical and administrative needs. The researchers also emphasized the importance of international collaboration in promoting AI adoption. In [9], the authors also describe how Saudi Arabian healthcare organizations are active participants with global partners to develop and deploy AI solutions. Part of this international collaboration offers new knowledge sharing and technical advancement opportunities, but challenges with data sharing, impedance matching, and other system-level aspects do exist.

III. Research Design and Setting

A. Research Philosophy and Approach

As this research takes a pragmatic philosophical stance combining positivist, as well as interpretive approaches to this often very complex but extremely promising attempt to solve healthcare problems via artificial intelligence adoption into Saudi Arabian health care organizations, its endeavor is this. Fig. 1 shows how the Analytic Hierarchy Process (AHP) is used in this study as an element of pragmatic commentary that brings together quantitative precision with qualitative investigation and its results. It offers a holistic treatment of both the quantifiable priorities and the contextual factors involved in influencing AI adoption in Saudi Arabian healthcare settings using this mixed method. A sequential explanatory design is then used to carry out the study within this philosophical framework, that is, quantitative data collection and analysis through AHP is followed by qualitative investigation to gain further insights on the patterns of prioritization. The alignment of this approach with the recommendations in [36] of the need for knowledgeable understanding of technology adoption in the healthcare environment, which are culturally different (for example, Saudi Arabia).

1) Geographic and institutional coverage: The Central Region, centered around Riyadh, represents the largest concentration of participating organizations (52%), reflecting its role as the Kingdom's primary healthcare hub. This region includes major referral hospitals, specialized medical centers, and academic institutions that are at the forefront of AI adoption in healthcare, as shown in Table I. The Western Region, encompassing Jeddah and Mecca, contributes 29% of participating organizations, providing insights into healthcare delivery in areas serving both resident populations and religious tourists. The Eastern Province, with 19% of participating organizations, offers perspectives on AI adoption in a region characterized by industrial development and diverse population demographics.

TABLE I. Detailed Distribution of Participating Healthcare Organizations.

Region	Govt. Hosp.	Private Hosp.	Academic Med. Ctr	Spec. Ctrs	Total
Central Region	4 (19%)	3 (14%)	2 (10%)	2 (10%)	11 (52%)
Western Region	3 (14%)	2 (10%)	1 (5%)	0 (0%)	6 (29%)
Eastern Province	2 (10%)	1 (5%)	1 (5%)	0 (0%)	4 (19%)
Total	9 (43%)	6 (29%)	4 (19%)	2 (10%)	21 (100%)

TABLE II. Organizational Sampling Criteria and Distribution

Criterion	Categories	Distribution	Rationale
Service Level	Primary (19%) Secondary (43%) Tertiary (38%)	Hierarchical healthcare delivery structure	Comprehensive service coverage
AI Impl. Stage	Initial (33%) Active (43%) Mature (24%)	Implementation maturity spectrum	Development stage insights
Bed Capacity	$<$200 (24%) 200–500 (38%) $>$500 (38%)	Organizational size representation	Resource capacity variation

B. Research Design Framework

The study employs a comprehensive research design framework that integrates multiple methodological components to ensure robust and reliable findings. Fig. 2 illustrates the interconnected elements of the research design.

1) Design components and integration: The research design comprises three major components, each serving specific objectives while contributing to the overall research aims:

a) Quantitative component: The major quantitative component is an application of the AHP methodology towards prioritizing AI adoption challenges. Structured questionnaires and pairwise comparison matrices are utilized in this component to produce priority weights for a set of challenges. Saaty's well-established methodology is followed by the AHP process to enable mathematical rigor in the prioritization process.

b) Qualitative component: The qualitative part includes semi structured interviews with key stakeholders and extensive document analysis. Using rich contextual data from this component, the prioritization patterns seen in the quantitative analysis can be explained. Preliminary AHP findings were used for developing interview protocols that resulted in focused and relevant data collection.

c) Validation component: A robust validation of both internal consistency checks and external expert validation makes findings more reliable. The validation part is included in this component, which is multiple rounds of validation using expert panels and cross validation across different contexts of organizations.

C. Sampling Strategy and Identification of Participants

A stratified purposive sampling was used to select the participating organizations that represented different healthcare sectors and geographical regions. The sampling framework considered several key organizational characteristics, as shown in Table II.

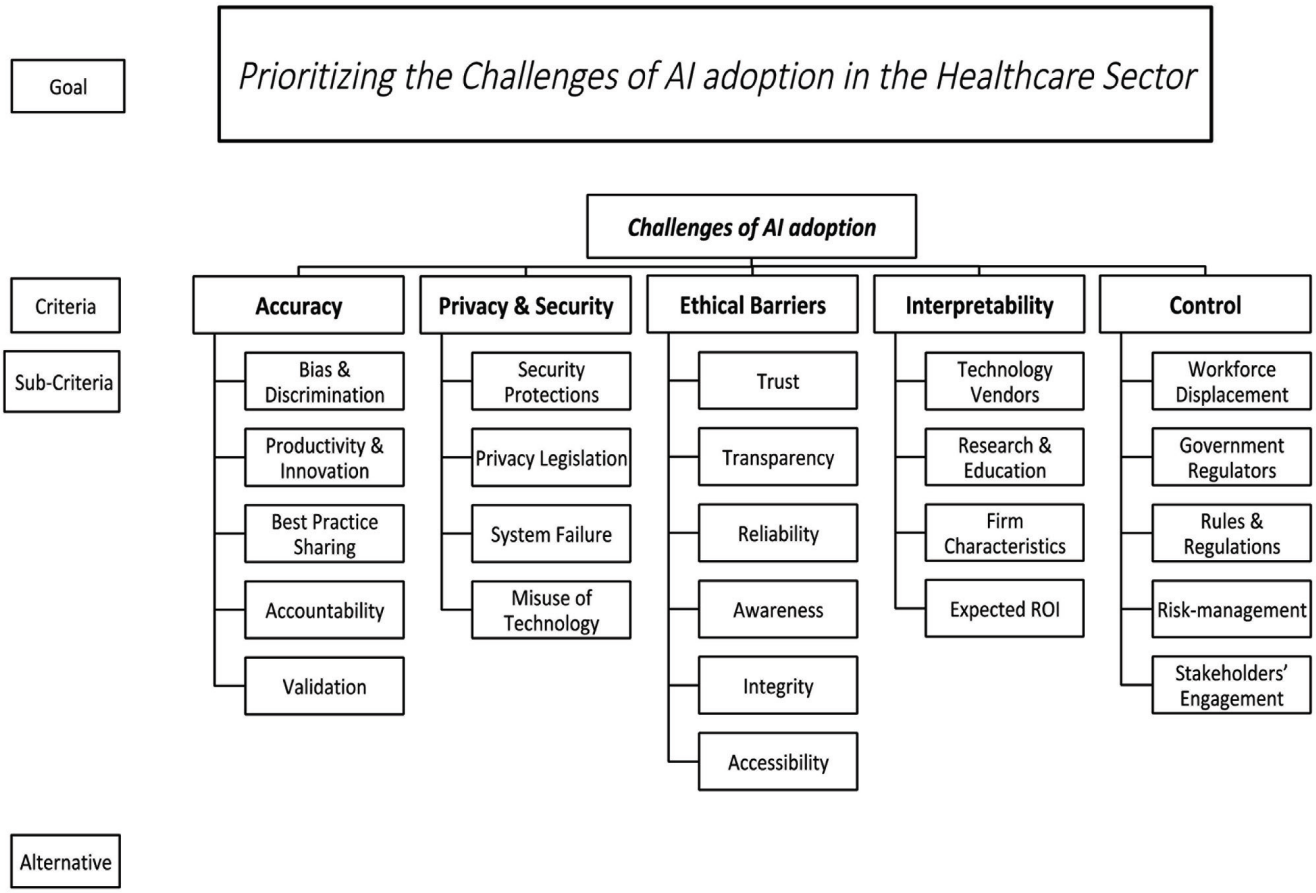


Fig. 1. The AHP model for prioritizing the challenges of AI adoption in the healthcare sector.

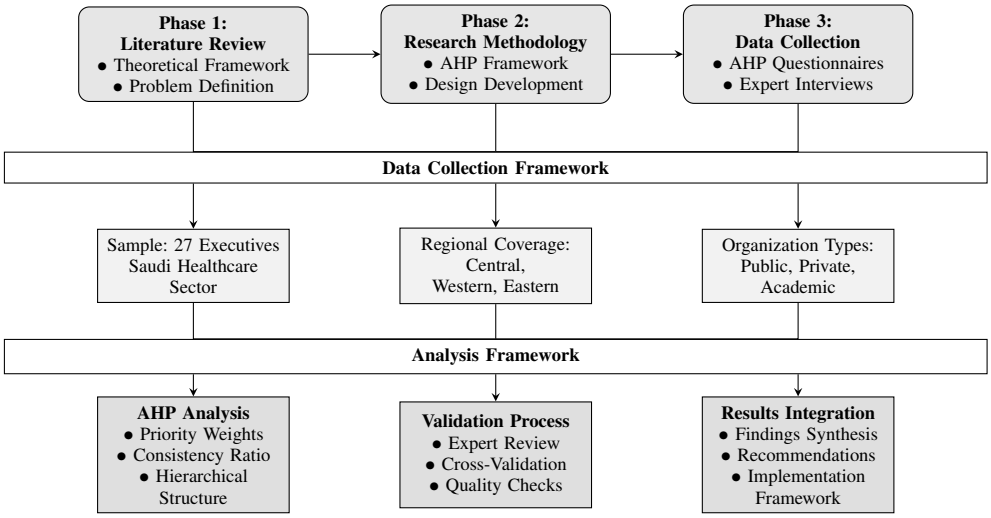


Fig. 2. Research design framework

IV. Analysis and Findings

A. Analysis of Main Criteria Priority Weights

Such analysis of how factors are prioritized had revealed unique patterns regarding how the adoption of AI is failed in Saudi Arabian healthcare organizations. The geometric means are used as the pair-wise comparisons for the main criteria and displayed in Table III for the relative importance of each of the factors under the Saudi Arabian healthcare context.

It is found that accuracy became the most important criterion with a priority weight of 0.50, greater than other criteria.

TABLE III. GEOMETRIC MEANS OF THE PAIR-WISE COMPARISONS OF THE MAIN CRITERIA.

Com Priority	Accuracy	Privacy & Security	Ethical Barriers	Interp.	Control	Priority Weight
Accuracy	0.65	0.84	0.47	0.31	0.25	**0.50**
Privacy & security	0.08	0.11	0.45	0.30	0.26	**0.24**
Ethical barriers	0.09	0.02	0.07	0.33	0.23	**0.15**
Interpretability	0.09	0.02	0.01	0.04	0.23	**0.08**
Control	0.09	0.01	0.01	0.01	0.03	**0.03**

TABLE IV. GEOMETRIC MEANS OF THE PAIR-WISE COMPARISONS OF THE SUB-CRITERIA (ACCURACY).

Com Priority	Bias & Discrim.	Prod. & Innov.	Best Prac. Sharing	Account.	Valid.
Bias & discrimination	0.65	0.85	0.47	0.33	0.24
Productivity & innov.	0.08	0.11	0.46	0.30	0.23
Best practice sharing	0.09	0.01	0.06	0.31	0.25
Accountability	0.09	0.02	0.01	0.04	0.24
Validation	0.09	0.02	0.01	0.01	0.03

These findings support the priorities of Saudi Arabia's healthcare transformation initiative, focusing on precision medical applications [35]. This is because the diagnostic accuracy is of critical nature, and the high priority weight represents the importance of diagnostic accuracy in the Saudi Arabia healthcare system as organizations adopt AI in making clinical decisions.

Second in rank was privacy and security with a priority weight of 0.24, which is a clear indication that Saudi Arabia healthcare organization give great importance to data protection and information security [38]. This being a priority of both regulator requirements and cultural sensitivities of patient data in the Saudi Arabian context.

The importance of ethical barriers received a priority weight of 0.15, indicating significant weight in the cultural, religious, and other factors of AI implementation. This finding is consistent with the research by [36] on aligning the AI systems with Islamic medical ethics, as well as local cultural values.

Lower priority weights (0.08 for interpretation, 0.03 for control) imply that these are not as important during this initial phase of AI's adoption in Saudi Arabian healthcare as they are accuracy and privacy concerns.

B. Detailed Analysis of Sub-Criteria

1) Accuracy dimension: Table IV presents the detailed analysis of sub-criteria across all main dimensions. Within the accuracy category:

The results show that bias (0.51) and discrimination are the strongest subfactors, highlighting the fact that there is serious concern about fairness and unbiasedness in the Saudi Arabian healthcare context. The need for AI systems to manage the variety of patient populations while maintaining cultural sensitivity results in assigning this high priority weight [37].

The second most important sub-factor (0.24), the use of AI to improve operational efficiency, showed Saudi Arabian healthcare organizations' focus on using AI to increase productivity and innovation. In particular, best practice sharing (0.15) showed attention to the importance of knowledge exchange in the Saudi Arabia healthcare sector AI adoption.

TABLE V. GEOMETRIC MEANS OF THE PAIR-WISE COMPARISONS OF THE SUB-CRITERIA (PRIVACY & SECURITY).

Com Priority	Security Protect.	Privacy Legisl.	System Failure	Misuse of Technology	Priority Weight
Security protections	0.71	0.85	0.47	0.32	**0.59**
Privacy legislation	0.10	0.11	0.46	0.33	**0.25**
System failure	0.09	0.02	0.06	0.30	**0.12**
Misuse of technology	0.10	0.02	0.01	0.05	**0.04**

TABLE VI. GEOMETRIC MEANS OF THE PAIR-WISE COMPARISONS OF THE SUB-CRITERIA (ETHICAL BARRIERS).

Com Priority	Trust	Transp.	Reliab.	Aware.	Integr.	Access.	Weight
Trust	0.58	0.82	0.45	0.30	0.24	0.20	**0.43**
Transparency	0.08	0.11	0.46	0.32	0.24	0.20	**0.24**
Reliability	0.09	0.02	0.07	0.33	0.26	0.20	**0.16**
Awareness	0.09	0.02	0.01	0.04	0.23	0.20	**0.10**
Integrity	0.08	0.02	0.01	0.01	0.03	0.18	**0.05**
Accessibility	0.08	0.02	0.01	0.01	0.01	0.03	**0.02**

2) Privacy and security framework: The privacy and security dimension revealed the following priority distribution, as shown in Table V.

The stringent data protection requirements in the Saudi Arabia healthcare led to the emergence of security protections as the dominant concern (0.59). Due to this high priority, this is aligned with recent cybersecurity regulations as well as the Kingdom's focus on protecting sensitive health information [39].

Noticeably, privacy legislation was getting some attention (0.25), a growing sign of the importance of regulatory compliance in Saudi Arabia health care AI implementation. The security framework still takes into account system failure (0.12) and misuse of technology (0.04), although these received lower priority weights.

3) Ethical barriers analysis: The analysis of ethical barriers revealed distinct patterns in the Saudi Arabian healthcare context, as shown in Table VI:

The most important ethical consideration was trust, which received a priority weight of 0.43. This is even more important in the context of Saudi Arabian healthcare, due to its cultural and religious values, and in the context of health care systems [36]. Weight for building and maintaining trust between AI systems and healthcare providers and patients in Saudi Arabia give high priority for Saudi Arabian culture.

Second was transparency, which received a priority weight of 0.24, referring to the need for AI systems' decision-making processes to be transparent and their communication with their users to be clear. Transparency is essential for gaining the acceptance of the AI systems in Saudi Arabian healthcare institutions, particularly in the context of the even more traditional healthcare delivery in the Kingdom [40].

A prioritized weight of 0.16 was given to reliability, which meant that consistent performance in AI systems is of high importance. According to this, Saudi Arabia backs high quality healthcare delivery amid the adoption of new technologies. However, priority weights were given to awareness (0.10),

TABLE VII. GEOMETRIC MEANS OF THE PAIR-WISE COMPARISONS OF THE SUB-CRITERIA (INTERPRETABILITY).

Com Priority	Tech. Vendors	Research & Education	Firm Charact.	Expected ROI	Priority Weight
Technology vendors	0.70	0.85	0.47	0.34	**0.59**
Research & education	0.10	0.12	0.45	0.33	**0.25**
Firm characteristics	0.10	0.02	0.07	0.29	**0.12**
Expected ROI	0.09	0.02	0.01	0.04	**0.04**

TABLE VIII. GEOMETRIC MEANS OF THE PAIR-WISE COMPARISONS OF THE SUB-CRITERIA (CONTROL).

Com Priority	Workforce Displace.	Gov. Regulat.	Rules & Regulat.	Risk-Mgmt	Stakeh. Engage.	Priority Weight
Workforce displacement	0.64	0.84	0.44	0.33	0.24	**0.50**
Government regulators	0.09	0.11	0.48	0.30	0.24	**0.25**
Rules & regulations	0.09	0.01	0.06	0.31	0.24	**0.14**
Risk-management	0.09	0.02	0.01	0.04	0.23	**0.08**
Stakeholders' engagement	0.10	0.02	0.01	0.01	0.04	**0.03**

integrity (0.05), and accessibility (0.02), which is crucial to the ethical framework of AI implementation.

4) Interpretability dimension analysis: The analysis of interpretability sub-criteria is revealed in the following Table VII:

The priority weight for technology vendors to be the dominant factor in interpretability was 0.59. This is due to the high priority from Saudi Arabia in developing the AI implementation strategy that encompasses partnering with the world leading technology providers while simultaneously developing capabilities locally [42]. The importance of partner selection suggests quoting AI solutions not only with technical sophistication but also with cultural appropriateness for the Saudi Arabia healthcare system.

The second-highest priority weight is 0.25 which indicates that research and education received the highest priority weight among Saudi Arabia's priorities to build local expertise in healthcare AI. It is in line with Vision 2030's objectives for the knowledge economy development and the Kingdom's investment in the healthcare education and training [9]. Especially considering the requirement for the development of AI systems and products that can provide the relevant public with excellent and professional service according to international standards while taking into account the national reality, the focus on research and education is particularly important.

Technical expertise, knowledge development, and other variables are given higher priority weight than organizational factors and expected ROI (0.12, 0.04), suggesting that organizational factors and financial returns are less important than technical expertise and knowledge development in the early adoption stage of AI.

C. Control Mechanisms Analysis

The analysis of control sub-criteria revealed significant insights into the governance aspects of AI adoption (see Table VIII).

The primary control concerned was workforce displacement with a weight of 0.50. This finding is important in the Saudi Arabian context, considering workforce nationalization (Saudi Arabiazation) is a central policy goal [41]. This high priority weight is the result of concerns about the need to keep Saudi Arabian healthcare professionals out of work and the implementation of AI systems.

Third most priority weight (0.25) was given to government regulators, as it points out to the vital role government regulatory bodies, for example, SDAIA and Ministry of Health, projected in AI adoption. This finding is significant given regulatory oversight as a means to prevent the inappropriate implementation of AI technologies in Saudi Arabia healthcare organizations [23].

Although having lower priority weights, rules and regulations (0.14), risk management (0.08), and stakeholders' engagement (0.03) are important parts of the framework. Regulatory compliance and risk management are needed, but these findings indicate that the main concern is to manage the workforce impact of AI adoption and the extra government oversight.

TABLE IX. PRIORITY WEIGHT VARIATIONS IN CENTRAL REGION

Main Criteria	Nat. Avg.	Central Region	Variance
Accuracy	0.50	0.53	+0.03
Privacy & Security	0.24	0.26	+0.02
Ethical Barriers	0.15	0.12	−0.03
Interpretability	0.08	0.06	−0.02
Control	0.03	0.03	0.00

V. CROSS DIMENSIONAL ANALYSIS

The analysis revealed significant interactions between the dimensions of various AI adoption challenges. For example, the high priority weights of accuracy (0.50) and workforce displacement (0.50 within control subcriteria) imply that technical performance and human resource considerations are somehow interdependent. In the Saudi Arabian context, this interaction is particularly important as organizations are called upon to satisfy both technological advancement and workforce development goals [44]. The analysis revealed significant regional variations in AI adoption challenges across Saudi Arabia's major healthcare regions. These variations reflect the diverse healthcare landscapes and varying levels of technological readiness across the Kingdom.

A. Central Region Analysis

Healthcare organizations in the Central Region, particularly in Riyadh, demonstrated distinct priority patterns in AI adoption challenges, as shown in Table IX.

Accuracy (0.53) was a higher emphasis of national organizations compared to central Region organizations (0.50). This focus on accuracy corresponds to the concentration of specialized medical facilities and research institutions in Riyadh, where precision in medical applications is, in particular, important [45]. As well as this, the region has stronger focus on privacy and security (0.26) as a result of major healthcare institutions in the region that are responsible for handling sensitive patient data.

TABLE X. PRIORITY WEIGHT VARIATIONS IN WESTERN REGION

Main Criteria	Nat. Avg.	Western Region	Variance
Accuracy	0.50	0.48	−0.02
Privacy & Security	0.24	0.28	+0.04
Ethical Barriers	0.15	0.16	+0.01
Interpretability	0.08	0.05	−0.03
Control	0.03	0.03	0.00

TABLE XI. PRIORITY WEIGHT VARIATIONS IN EASTERN PROVINCE

Main Criteria	Nat. Avg.	Eastern Province	Variance
Accuracy	0.50	0.49	−0.01
Privacy & Security	0.24	0.22	−0.02
Ethical Barriers	0.15	0.17	+0.02
Interpretability	0.08	0.09	+0.01
Control	0.03	0.03	0.00

TABLE XII. PRIORITY WEIGHTS IN INITIAL PLANNING PHASE

Main Criteria	Overall Avg.	Initial Phase	Variance
Accuracy	0.50	0.45	−0.05
Privacy & Security	0.24	0.30	+0.06
Ethical Barriers	0.15	0.14	−0.01
Interpretability	0.08	0.07	−0.01
Control	0.03	0.04	+0.01

VI. WESTERN REGION ANALYSIS

The healthcare organizations in the Western Region, specifically Jeddah and Mecca, stand out with their individual characteristics in the challenges of AI adoption (see Table X).

Particularly noteworthy was that heightened concern for privacy and security (0.28), particularly so as the region is responsible for medical needs provision during the Hajj and Umrah seasons [38]. The focus on managing healthcare data for a large, diversified population of pilgrims with other cultural and religious sensitivities, of course, increases the emphasis.

A. Eastern Province Analysis

The Eastern Province, more than any other province, had unique challenges in AI adoption (see Table XI).

Interpretability (0.09) and ethical barriers (0.17) showed increased emphasis in healthcare organizations of the Eastern Province. The Saudi Arabia region's diverse population and the need for its AI systems to serve both the urban and industrial populations are the causes of this pattern.

B. Analysis on the Implementation Phase

The study also uncovered different implementation phase specific patterns of challenges in AI adoption. These phases can describe the maturity of AI adoption in Saudi Arabian healthcare organizations.

1) Initial planning phase: Organizations in the initial planning phase (0–12 months into AI adoption) demonstrated the following priority patterns:

TABLE XIII. PRIORITY WEIGHTS IN ACTIVE IMPLEMENTATION PHASE

Main Criteria	Overall Avg.	Active Phase	Variance
Accuracy	0.50	0.55	+0.05
Privacy & Security	0.24	0.22	−0.02
Ethical Barriers	0.15	0.13	−0.02
Interpretability	0.08	0.07	−0.01
Control	0.03	0.03	0.00

TABLE XIV. PRIORITY WEIGHTS IN MATURITY PHASE

Main Criteria	Overall Avg.	Maturity Phase	Variance
Accuracy	0.50	0.48	−0.02
Privacy & Security	0.24	0.23	−0.01
Ethical Barriers	0.15	0.16	+0.01
Interpretability	0.08	0.10	+0.02
Control	0.03	0.03	0.00

This stage had organizations really focused on privacy and security, and they were concerned about establishing strong data protection frameworks at the beginning of the process, which manifested in a 0.3. This fits in with Saudi Arabia's strict requirements for healthcare data handling and building security architecture, as shown in Table XII.

C. Active Implementation Phase

Organizations in the active implementation phase (12–24 months) showed distinct priorities.

As organizations moved from active implementation to CDI to late implementation (Table XIII), accuracy became the most important thing (0.55). This is because organizations focused on reliable system performance and clinical outcomes, which made the rules more accurate. At this phase, the practical challenges of deploying AI were recognized, which led to particular emphasis on workforce training and system integration [41].

1) Maturity phase: Organizations in the maturity phase (>24 months) demonstrated evolved priority patterns.

Table XIV shows how much attention was paid to each challenge in the maturity phase compared to the other phases. During the maturity phase, organisations paid more equal attention to all challenges, but they paid the most attention to interpretability (0.10) and ethical concerns (0.16). This shift highlights this more advanced understanding of the AI implementation challenges and its crucial role in the long-run sustainability [44].

D. Implementation Success Factors

The prompt creation of a solid privacy and security framework is crucial. The development process balances the radically divergent focus on accuracy with attention to technical and ethical considerations in more mature implementations. This shift reflects a more sophisticated understanding of AI implementation challenges and the importance of long-term sustainability [44].

Based on the analysis of different implementation phases, several key success factors emerged:

TABLE XV. PARTICIPANT PROFILE DISTRIBUTION

Role Category	Number (%)	Selection Rationale
Hospital Directors/CEOs	5 (18.5%)	Strategic decision-making perspective
IT Directors	6 (22.2%)	Technical implementation insights
Clinical Department Heads	7 (25.9%)	Clinical application understanding
AI Project Managers	5 (18.5%)	Direct implementation experience
Quality/Risk Managers	4 (14.8%)	Governance and compliance perspective
Total	**27 (100%)**	Comprehensive stakeholder representation

- Early establishment of robust privacy and security frameworks
- Strong focus on accuracy during active implementation
- Balanced attention to technical and ethical considerations in mature implementations
- Continuous workforce development and training
- Regular assessment and adjustment of implementation strategies

It is these findings that allow Saudi Arabian healthcare organizations of varying AI adoption status to look ahead to and meet their challenges accordingly.

1) Selection of participant: The method made the selection of individual participants in organizations systematic and made it possible to select knowledgeable and relevant input. We carefully selected a total of 27 participants who are in roles, have experience, and are involved in AI implementation initiatives. The participant selection process took into account multiple criteria.

Table XV shows the distribution of participants across different roles to cover all perspectives.

VII. DISCUSSION

The comprehensive analysis of AI adoption challenges in Saudi Arabian healthcare organizations reveals several significant patterns and implications that warrant detailed discussion. This section examines the findings within the broader context of Saudi Arabia's healthcare transformation goals and Vision 2030 objectives.

A. Dominance of Accuracy Concerns

All regions give accuracy a low priority weight of 0.50, reflecting the importance of reliable AI performance in healthcare settings. Related to this, [35]'s research shows that the precision of medical applications is crucial. The emphasis on accuracy reflects several key considerations in the Saudi Arabia context: Firstly, the finding shows that the healthcare sector is very committed to maintaining a maximum standard of patient care while embracing new technology. Specifically, Saudi Arabia healthcare organizations show particular concern for making sure that AI systems can process the diverse patients' population, including residents in the Kingdom and religious tourists visiting during Hajj and Umrah seasons [45]. Second, there are regional variations in accuracy priorities, with the Central Region ranking at 0.53. Different levels of technological readiness and clinical requirements vary across the Kingdom, ranging from 0.48 in the Western Region to 0.49 in the Eastern Province. Finally, the high emphasis on accuracy in the Central Region, especially in Riyadh, is due to the high density of specialized medical facilities, which have high technological requirements [34].

B. Privacy and Security Considerations

We gave privacy and security a significant priority weight of 0.24, illustrating a very complex interplay between technological advancement and cultural sensitivities in Saudi Arabia healthcare. According to [38], this finding has several important implications: Concern for privacy and security, particularly in the Western Region (0.28) during the Hajj and Umrah seasons, presents the challenge of managing health care data for an international population that is diverse in culture and religion. These results suggest the need for specialized techniques to safeguard data, considering both technical requirements and cultural norms. Temporal analysis shows that privacy concerns are most acute at the beginning of the implementation phase (0.30)—during the first phases of adopting AI, it is critical to create solid data protection frameworks. The pattern aligns perfectly with Saudi Arabia's strategic approach to digital transformation, which advocates for a secure and controlled approach to technological implementation.

C. Evolution of Implementation Challenges

1) Phase-specific priorities: Analysis of the implementation phase shows a clear evolution in (organizational) priorities when AI adoption matures. This progression has significant implications for healthcare organizations: Initial Planning Phase (0-12 months): However, this phase prioritizes privacy and security (0.30), necessitating the establishment of strong foundational frameworks. In this phase, organizations show extraordinary interest in complying with national regulations as well as international standards [39]. Active Implementation Phase (12–24 months): During this phase, accuracy concerns (0.55) dominate the shift towards an orientation of operational effectiveness and clinical outcome. This transition implies that after privacy frameworks have been established, performance and reliability of systems are prioritized by organizations [41]. Maturity Phase ($\geq$24 months): Mature implementations, on the other hand, distribute these priorities more evenly (0.10 for interpretability; 0.16 for ethical considerations), demonstrating a more sophisticated approach to AI integration challenges. It is possible to see that, with time, organizations evolved more sophisticated approaches for implementing AI [44].

D. Regional Dynamics and Healthcare Delivery

1) Urban-rural disparities: The regional analysis reveals important variations in AI adoption challenges between urban and rural healthcare settings. These disparities have significant implications for healthcare delivery.

a) Urban centers: Healthcare organizations in major urban centers demonstrate higher readiness for AI adoption, with stronger emphasis on advanced applications and integration with existing systems. The Central Region's focus on accuracy (0.53) reflects the sophisticated technological infrastructure and specialized medical services in urban areas [37].

b) Rural areas: Healthcare facilities in rural areas show different priority patterns, with greater emphasis on basic infrastructure and workforce development. This finding suggests the need for targeted support and resources to ensure equitable AI implementation across all regions [9].

E. Cultural and Ethical Considerations

1) Integration of Islamic medical ethics: The study reveals that ethical issues (0.15), particularly Islamic medical principles, receive significant consideration. This finding has several important implications: Trust, the second most highly weighted factor (0.43 of ethical barriers), highlights the need for AI systems to align with culture and religion. When adopting AI, healthcare organizations have the responsibility to implement AI in a way that respects Islamic medical ethics and yet remains technologically effective [36]. Ethical priorities vary by region, with the West Region having the highest priority variances (0.16); this reflects the requirement to differentiate between religious and cultural sensitivities and the geographical area of its high number of pilgrims and international patients.

2) Workforce development and training: Human resource implications the control subcriteria assigns a high priority weight of 0.50 to workforce displacement, highlighting significant concerns about the impact of AI on healthcare professionals: This finding signals a need for a whole workforce development strategy that will balance technological advancement and employment for Saudi Arabia healthcare workers [40]. Copying research and education (0.25 within the interpretability subcriteria) indicates that there is an understanding of the need to develop local expertise in healthcare AI technologies.

F. Future Directions and Recommendations

Based on these findings, several key recommendations emerge for Saudi Arabian healthcare organizations:

- Formulate region-specific implementation strategies, considering local infrastructure and local demand of the human population.
- In organizations with multiple patient populations, establish these comprehensive privacy frameworks as early and far in the implementation process as possible.
- Invest in workforce development programs that not only prepare healthcare professionals for AI integration but also protect employment opportunities.
- Design balanced approaches to AI implementation that address technical necessities and cultural sensibilities.
- Promote knowledge exchange between regions and organizations to enable best practices in pursuing AI adoption.

VIII. Conclusion

In this study, we have qualitatively understood the challenges of AI adoption in Saudi Arabia's healthcare sector and have uncovered key insights on implementation priorities. The results show that accuracy (0.50), privacy (0.24), security concerns, and ethical barriers (0.15) are the most influential factors in successful AI adoption. However, bias and discrimination (0.51), security protections (0.59), and trust (0.43) were the strongest of the subfactors in this set of criteria, as the Saudi Arabian healthcare context presents a different cultural and regulatory environment. As such, the research provides a systematic framework for comprehensively investigating AI adoption challenges in culturally diverse healthcare environments and contributes to practical implementation as well as theoretical understanding. The research findings corroborate and complement existing scholarly works and present new understandings of the implementation of AI technologies in healthcare transformation in Saudi Arabia. The results of the study are of particular value to healthcare organizations faced with the challenging mandate of managing technological advancement within a culturally sensitive and regulatory context. Based on these results, Saudi Arabian healthcare organizations should focus on building strong systems for checking accuracy, complete data security frameworks, and ethical compliance with Islamic medicine in order to build trust among patients in the use of artificial neural networks in medicine. It is possible to conduct further research on how these challenges evolve as AI programming advances, as well as specific strategies for implementing AI in other healthcare sectors. Based on these insights, we can design more efficient strategies that promote the use of AI in Saudi Arabia's healthcare sector in a culturally appropriate manner.

Acknowledgment

The authors would like to thank the participating healthcare organizations and executives who contributed their time and expertise to this study.

References

[1] B. Al-Naami, H. Fraihat, H. A. Owida, K. Al-Hamad, R. De Fazio, and P. Visconti, "Automated detection of left bundle branch block from ECG signal utilizing the maximal overlap discrete wavelet transform with ANFIS," *Computers*, vol. 11, no. 6, p. 93, 2022.

[2] A. K. A. Jawwad, N. Turab, G. Al-Mahadin, H. A. Owida, and J. Al-Nabulsi, "A perspective on smart universities as being downsized smart cities: A technological view of internet of thing and big data," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 35, no. 2, pp. 1162–1170, 2024.

[3] S. Abuowaida, H. A. Owida, D. M. Alsekait, N. Alshdaifat, D. S. AbdElminaam, and M. Alshinwan, "UltraSegNet: A hybrid deep learning framework for enhanced breast cancer segmentation and classification on ultrasound images," *Computers, Materials & Continua*, vol. 83, no. 2, pp. 3303–3333, 2025.

[4] H. A. Owida, O. S. M. Hemied, R. S. Alkhawaldeh, N. F. F. Alshdaifat, and S. F. A. Abuowaida, "Improved deep learning approaches for COVID-19 recognition in CT images," *Journal of Theoretical and Applied Information Technology*, vol. 100, no. 13, pp. 4925–4931, 2022.

[5] S. F. A. Abuowaida, H. Y. Chan, N. F. F. Alshdaifat, and L. Abualigah, "A novel instance segmentation algorithm based on improved deep learning algorithm for multi-object images," *Jordanian Journal of Computers and Information Technology*, vol. 7, no. 1, 2021.

[6] S. Abuowaida, E. A. Elsoud, A. Al-Momani, M. Arabiat, H. A. Owida, N. Alshdaifat, and H. Y. Chan, "Proposed enhanced feature extraction for multi-food detection method," *Journal of Theoretical and Applied Information Technology*, vol. 101, no. 24, pp. 8140–8146, 2023.

[7] H. M. Turki *et al.*, "Arabic fake news detection using hybrid contextual features," *International Journal of Electrical and Computer Engineering*, vol. 15, no. 1, 2025.

[8] M. Arabiat, S. Abuowaida, A. Al-Momani, N. Alshdaifat, and H. Y. Chan, "Depth estimation method based on residual networks and SE-Net model," *Journal of Theoretical and Applied Information Technology*, vol. 102, no. 3, pp. 866–871, 2024.

[9] Z. A. Mani and K. Goniewicz, "Transforming healthcare in Saudi Arabia: A comprehensive evaluation of Vision 2030's impact," *Sustainability*, vol. 16, no. 8, p. 3277, 2024.

[10] A. K. Suleiman and L. C. Ming, "Transforming healthcare: Saudi Arabia's Vision 2030 healthcare model," *Journal of Pharmaceutical Policy and Practice*, vol. 18, no. 1, p. 2449051, 2025.

[11] H. A. Sheerah, S. AlSalamah, S. A. Alsalamah, C.-T. Lu, A. Arafa, E. Zaatari, A. Alhomod, S. Pujari, and A. Labrique, "The rise of virtual health care: Transforming the health care landscape in the kingdom of Saudi Arabia: A review article," *Telemedicine and e-Health*, vol. 30, no. 10, pp. 2545–2554, 2024.

[12] Y. T. AL ALI, A. A. AL QAHTANI, H. Y. ASSIRI, A. M. ALYAHYA, F. S. AL ALKHARSH, A. Y. ASSIRI, M. S. ALGHAMDI, A. Q. SAEED, F. M. ALSHEHRI, and Y. H. ALASIRI, "Effectiveness of technology on organizational development and services in the Saudi health sector," *Journal of Pharmaceutical Negative Results*, pp. 2144–2155, 2022.

[13] F. Y. Alzyoud, M. Tarawneh, A. Almaghthawi, A. Altalidi, L. Asiri, and M. Alrehaili, "Optimizing broadcast utilization for efficient disaster management using wireless ad hoc networks and novel energy-saving algorithms.," *International Journal of Interactive Mobile Technologies*, vol. 18, no. 20, 2024.

[14] A. Almaghthawi, F. Bourennani, and I. R. Khan, "Differential evolution-based approach for tone-mapping of high dynamic range images," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 7, 2020.

[15] M. Z. Alkhamali, "National healthcare transformation awareness among Alfaisal University community: A cross-sectional study," master's thesis, Alfaisal University (Saudi Arabia), 2022.

[16] F. AlZyoud, M. Tarawneh, A. Almaghthawi, and A. Altalidi, "A new approach for cluster head selection in wireless sensor networks.," *International Journal of Online & Biomedical Engineering*, vol. 20, no. 9, 2024.

[17] A. Almaghthawi, E. A. Ghaleb, N. A. Akbar, L. Asiri, M. Alrehaili, and A. Altalidi, "Federated-learning intrusion detection system based blockchain technology.," *International Journal of Online & Biomedical Engineering*, vol. 20, no. 11, 2024.

[18] M. Kamran, A. Saeed, and A. Almaghthawi, "Federated-learning topic modeling based text classification regarding hate speech during covid-19 pandemic.," *International Journal of Advanced Computer Science & Applications*, vol. 14, no. 11, 2023.

[19] F. Alam, T. S. M. Alnazzawi, R. Mehmood, and A. Al-Maghthawi, "A review of the applications, benefits, and challenges of generative ai for sustainable toxicology," *Current Research in Toxicology*, vol. 8, p. 100232, 2025.

[20] F. Alam, P. Pławiak, A. Almaghthawi, M. R. C. Qazani, S. Mohanty, and R. Alizadehsani, "Neurohar: A neuroevolutionary method for human activity recognition (har) for health monitoring," *IEEE Access*, vol. 12, pp. 112232–112248, 2024.

[21] F. A. Rao, A. Muneer, A. Almaghthawi, A. Alghamdi, S. M. Fati, and E. A. A. Ghaleb, "Bmsp-ml: big mart sales prediction using different machine learning techniques," *IAES International Journal of Artificial Intelligence*, vol. 12, no. 2, p. 874, 2023.

[22] T. Alam, R. Gupta, N. N. Ahamed, A. Ullah, and A. Almaghthwi, "Towards sustainable iot-based smart mobility systems in smart cities," *GeoJournal*, vol. 89, no. 6, p. 235, 2024.

[23] A. A. Alasiri and V. Mohammed, "Healthcare transformation in Saudi Arabia: an overview since the launch of Vision 2030," *Health Services Insights*, vol. 15, p. 11786329221121214, 2022.

[24] O. Alidmat *et al.*, "Simulation of crowd evacuation in asymmetrical exit layout based on improved dynamic parameters model," *IEEE Access*, vol. 13, pp. 7512–7525, 2025.

[25] A. Ehsan, Z. Iqbal, S. Abuowaida, M. Aljaidi, H. U. Zia, N. Alshdaifat, and N. K. Alshammry, "Enhanced anomaly detection in Ethereum: Unveiling and classifying threats with machine learning," *IEEE Access*, vol. 12, pp. 176440–176456, 2024.

[26] M. Iqtait, J. Ababneh, M. Rasmi, A. Abu-Jassar, and S. Abuowaida, "Enhancing facial feature detection: Hybrid active shape and active appearance model (HASAAM)," *IEEE Access*, vol. 12, pp. 189590–189610, 2024.

[27] M. Baniata, S. Abuowaida, M. Aljaidi, M. Kharabsheh, A. Alsarhan, and A. A. Alsuwaylimi, "A multi-modal attention-guided network for Alzheimer's disease classification using deep learning," *Engineering, Technology & Applied Science Research*, vol. 15, no. 5, pp. 27150–27158, 2025.

[28] H. A. Owida, J. Al-Nabulsi, N. A. Hawamdeh, S. Abuowaida, Z. Salah, and E. A. Elsoud, "Deep learning-based kidney disease classification using ResNet50: Enhancing diagnostic accuracy," in *Proc. 2024 Second Jordanian International Biomedical Engineering Conference (JIBEC)*, pp. 126–129, 2024.

[29] A. Aburomman *et al.*, "Automated detection of Alzheimer's disease using EfficientNet-B3 architecture with transfer learning," in *Proc. 25th International Arab Conference on Information Technology (ACIT)*, pp. 1–6, 2024.

[30] M. M. Malkawi *et al.*, "Enhancing autism spectrum disorder detection: A novel ensemble learning framework with multi-model integration," in *Proc. 1st International Conference on Computational Intelligence Approaches and Applications (ICCIAA)*, pp. 1–6, 2025.

[31] S. Abuowaida *et al.*, "Hybrid ensemble architecture for brain tumor segmentation using EfficientNetB4-MobileNetV3 with multi-path decoders," *Data and Metadata*, vol. 4, p. 374, 2025.

[32] Z. Salah, H. Abu Owida, E. Abu Elsoud, E. Alhenawi, S. Abuowaida, and N. Alshdaifat, "An effective ensemble approach for preventing and detecting phishing attacks in textual form," *Future Internet*, vol. 16, no. 11, p. 414, 2024.

[33] H. A. Owida, M. R. Hassan, A. M. Ali, F. Alnaimat, A. Al Sharah, S. Abuowaida, and N. Alshdaifat, "The performance of artificial intelligence in prostate magnetic resonance imaging screening," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 2, 2024.

[34] K. A. Al-Dosari, S. M. A. Al Qahtani, A. M. Al Shammari, B. A. R. Al Shehri, Y. S. A. Algarny, and A. S. A. Algarny, "Digital transformation of the health sector through the requirements of Vision 2030," *Journal of Survey in Fisheries Sciences*, vol. 10, no. 5, pp. 53–61, 2023.

[35] M. A. M. Alyami, N. S. A. Atti, F. M. M. Masmali, M. H. Oraibi, M. A. Alruwaili, Z. M. A. Mahdi, T. M. N. Hafdi, E. M. A. Sahly, M. A. M. Rajhy, M. H. Alonazi, *et al.*, "The role of artificial intelligence in transforming healthcare leadership: A systematic review of current trends and future directions," *Journal of Namibian Studies: History Politics Culture*, vol. 31, pp. 86–106, 2022.

[36] R. Kumar, A. Singh, A. S. A. Kassar, M. I. Humaida, S. Joshi, and M. Sharma, "Leveraging artificial intelligence to achieve sustainable public healthcare services in Saudi Arabia: A systematic literature review of critical success factors," *Computer Modeling in Engineering & Sciences*, vol. 142, no. 2, p. 2, 2025.

[37] S. A. Alghamdi and Y. Alashban, "Medical science students' attitudes and perceptions of artificial intelligence in healthcare: A national study conducted in Saudi Arabia," *Journal of Radiation Research and Applied Sciences*, vol. 17, no. 1, p. 100815, 2024.

[38] A. H. Alfahad, Y. S. Alabbas, H. S. M. ALabbas, T. H. Abukhashbah, A. A. Alabdali, Q. M. H. Alfatieh, E. A. Bashawri, H. T. H. Hadidi, R. M. A. Junaid, and K. M. A. Alhazmi, "Evaluating the impact of Saudi Vision 2030 on healthcare investment: A comprehensive review of progress and future directions," *Journal of Ecohumanism*, vol. 3, no. 8, pp. 870–880, 2024.

[39] O. M. E. Ramadan, M. M. Alruwaili, A. N. Alruwaili, M. G. Elsehrawy, and S. Alanazi, "Facilitators and barriers to AI adoption in nursing practice: a qualitative study of registered nurses' perspectives," *BMC Nursing*, vol. 23, no. 1, p. 891, 2024.

[40] A. Alshahrani, D. Stewart, and K. MacLure, "A systematic review of the adoption and acceptance of ehealth in Saudi Arabia: Views of multiple

stakeholders," *International Journal of Medical Informatics*, vol. 128, pp. 7–17, 2019.

[41] A. Al Shahrani, N. Alhumaidan, Z. AlHindawi, A. Althobaiti, K. Aloufi, R. Almughamisi, and A. Aldalbahi, "Readiness to embrace artificial intelligence among medical students in Saudi Arabia: A national survey," in *Healthcare*, vol. 12, p. 2504, MDPI, 2024.

[42] H. A. Sheerah, S. AlSalamah, S. A. Alsalamah, C.-T. Lu, A. Arafa, E. Zaatari, A. Alhomod, S. Pujari, and A. Labrique, "The rise of virtual health care: Transforming the health care landscape in the kingdom of Saudi Arabia: A review article," *Telemedicine and e-Health*, vol. 30, no. 10, pp. 2545–2554, 2024.

[43] Saudi Vision 2030, "Vision 2030: Healthcare transformation program," *Strategic Document, Kingdom of Saudi Arabia*, 2021.

[44] A. M. Muafa and S. H. Al-Obadi, "The impact of artificial intelligence applications on the digital transformation of healthcare delivery in Riyadh, Saudi Arabia (opportunities and challenges in alignment with Vision 2030)," *Academic Journal of Research and Scientific Publishing*, vol. 5, no. 59, 2024.

[45] N. Aljerian, M. Arafat, A. Aldhubib, I. Almohaimeed, A. Alsultan, A. Alhosaini, A. Aloqayli, B. AlRabiah, S. AlBanyan, L. Ashry, *et al.*, "Artificial intelligence in health care and its application in Saudi Arabia," *International Journal of Innovative Research in Medical Science (IJIRMS)*, vol. 7, no. 11, 2022.

Multimodal Pneumoconiosis Screening on a Real-World Vietnamese Occupational Health Dataset with Incomplete Metadata and Class Imbalance

Luong Thi Bich Phuong[1], Nguyen Hoang Anh[2], Au Xuan Manh[3],
Truong Thanh Nam[4], Tran Tien Cong[5], Nguyen Trong Khanh[6]*
Faculty of Information Technology, Hanoi Architectural University, Hanoi, Vietnam[1,2,3,4]
Faculty of Artificial Intelligence, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam[1,5]
Intelligent Computing for Sustainable Development Lab and Faculty of Information Technology,
Posts and Telecommunications Institute of Technology, Hanoi, Vietnam[6]

***Abstract*—Pneumoconiosis remains a major occupational lung disease among workers exposed to silica, coal, and other inorganic dusts. Although chest X-ray screening is widely used in clinical practice, diagnostic performance is often affected by image variability and the limited availability of contextual occupational information. This study presents a multimodal framework that combines chest X-ray images with clinical and occupational metadata for automated pneumoconiosis detection. The proposed approach integrates a ResNet-18 image encoder with a metadata-processing multilayer perceptron through an early-fusion strategy. Experiments were conducted on a dataset of 1,971 subjects, including 1,086 confirmed pneumoconiosis cases and 885 healthy controls. Two configurations were evaluated: an image-only model and a multimodal model incorporating non-imaging information. The multimodal framework achieved the best validation performance, reaching an accuracy of 92.66%, an F1-score of 93.21%, and an AUC-ROC of 97.35%. Compared with the image-only baseline, adding metadata produced a small change in overall performance and shifted the precision–recall balance toward higher precision. These validation figures should be interpreted as an upper bound because the negative class is partly drawn from an external population and the pattern of missing metadata is correlated with the diagnostic label. The results suggest that combining radiographic findings with occupational and clinical information is a promising but data-quality-sensitive direction for pneumoconiosis screening in real-world settings.**

Keywords*—Pneumoconiosis; multimodal learning; chest X-ray; convolutional neural network; early fusion; occupational lung disease

I. Introduction

Dust exposure is a major occupational hazard in industries such as mining, construction, and manufacturing. Prolonged inhalation of inorganic dusts, including silica, coal, and asbestos, can cause pneumoconiosis, a group of chronic lung diseases characterized by progressive and irreversible pulmonary fibrosis [1], [2], [3]. In Vietnam, silicosis remains a significant occupational health concern that adversely affects workers' health and productivity [4].

The development and severity of pneumoconiosis are influenced by workplace conditions, cumulative dust exposure, and adherence to personal protective equipment usage [5]. Common symptoms include chronic cough, chest pain, weight loss, and dyspnea. Early diagnosis remains challenging because chest X-ray abnormalities are often subtle and subject to inter-observer variability [6].

Chest X-ray screening based on the International Labour Organization (ILO) classification is the most widely used diagnostic approach [7]. However, pneumoconiosis diagnosis in routine occupational-health practice does not rely solely on chest radiographs. Occupational exposure history, duration of dust exposure, respiratory symptoms, smoking behavior, workplace conditions, and compliance with personal protective equipment are all considered during clinical assessment. Consequently, image-only approaches may fail to fully exploit contextual information routinely collected during occupational-health surveillance programs, which may reduce diagnostic consistency and delay intervention [8].

Deep learning has shown promising performance in medical image analysis. Convolutional neural networks (CNNs), including ResNet-18, have achieved accuracies exceeding 90% in detecting pulmonary abnormalities from chest X-rays under controlled conditions [9], [10], [11]. Nevertheless, image-only approaches may be less reliable when applied to real-world occupational datasets with heterogeneous image quality and incomplete clinical information.

To address these limitations, this study proposes a multimodal framework that combines chest X-ray images with occupational and clinical metadata for pneumoconiosis screening. Visual representations are extracted using ResNet-18, whereas non-imaging variables, including exposure history, clinical symptoms, health indicators, and occupational protection usage, are encoded through a metadata-processing branch. The resulting feature representations are fused at an early stage to jointly exploit radiographic and contextual information, improving diagnostic reliability in real-world occupational health settings.

The framework is evaluated on a real-world dataset of Vietnamese workers and is intended for future deployment as a web-based clinical decision-support tool. Results show that incorporating non-imaging metadata yields a small change in validation performance compared with image-only baselines; The practical value and limitations of multimodal learning for

*Corresponding author

occupational health applications are further discussed in the subsequent sections.

II. Related Work

A. Occupational Pneumoconiosis and Epidemiological Context

Pneumoconiosis remains one of the most prevalent occupational diseases worldwide. In Vietnam, the Law on Occupational Safety and Health classifies pneumoconiosis, including silicosis, coal workers' pneumoconiosis, and asbestos-related diseases, among the occupational diseases eligible for social insurance and regular screening [5]. According to the Vietnamese Ministry of Health, approximately 7,000 occupational disease cases are reported annually, with pneumoconiosis accounting for 40–45% of cases [12]. Epidemiological studies have reported prevalence rates of 12–14% among underground workers with prolonged exposure histories [13]. These findings emphasize the need for effective and scalable screening systems.

B. Deep Learning Approaches for Lung Disease Detection

Deep learning, particularly convolutional neural networks (CNNs), has been widely adopted for automated lung disease detection from chest X-rays and CT images. Architectures such as ResNet, DenseNet, and Inception have demonstrated strong performance in identifying pulmonary abnormalities, with further improvements achieved through transfer learning and domain-specific fine-tuning [10], [11]. Beyond imaging, several studies have incorporated physiological signals and clinical indicators to improve respiratory disease detection. For example, respiratory sound analysis combined with clinical features has shown promise for detecting COPD and occupational pneumonia when radiographic findings are inconclusive [14], [15]. These studies highlight the value of integrating non-imaging information into diagnostic frameworks.

C. Multimodal Learning for Occupational Lung Diseases

Recent studies have shown that combining imaging data with structured clinical metadata can improve diagnostic accuracy and robustness, particularly in noisy and imbalanced real-world datasets. Despite these advances, multimodal approaches for pneumoconiosis detection remain limited, especially in occupational health settings within developing countries. Most existing studies rely on curated public datasets or single-modality inputs, which may not reflect the variability and incompleteness of real-world screening data. This limitation motivates the development of multimodal frameworks that jointly exploit radiographic findings and occupational information for practical pneumoconiosis screening.

III. Materials and Methods

A. Dataset Description

1) Patient data sources: The dataset utilized in this study contains a total of 1,971 occupational health records combining chest X-ray images and clinical/occupational metadata. Data were retrospectively collected from routine occupational health examinations conducted by multiple institutions in Northern Vietnam, including TISCO Health Department, Bac Ninh Department of Health, Vinh Phuc CDC, Thai Nguyen CDC, Phu Tho CDC, and the Institute of Preventive Medicine and Public Health. The dataset primarily represents workers from industries with potential exposure to silica-containing dust and other occupational respiratory hazards. In total, the dataset contains 1,086 confirmed pneumoconiosis cases (positive class, labeled as "bnn" = 1) and 885 healthy control cases (negative class, labeled as "bnn" = 0).

Due to the nature of real-world medical data collection, clinical metadata is partially available. Specifically, detailed non-imaging clinical metadata is complete for 433 cases (comprising 29 positive cases and 404 negative cases) collected from the hospital database. The remaining 1,538 cases (comprising 1,057 positive cases and 481 negative cases) are image-only. To establish healthy controls for the image-only cohort, normal chest X-rays were obtained from the public Chest X-Ray Images (Pneumonia) dataset. Although some variables exhibit relatively high missing rates in the combined set, they were retained to capture latent correlations, with missing values handled using training-set statistics. Fig. 1 presents the missing-value percentage of clinical and occupational metadata features.

To ensure robust evaluation and avoid overfitting, the aggregated dataset of 1,971 samples was split into a training set (80%, 1,576 samples and a validation set (20%, 395 samples). Stratified random sampling based on the target label "bnn" was applied to maintain the same class ratio in both partitions. The training set is used for parameter optimization and model fitting, while the validation set is used for hyperparameter tuning and model checkpoint selection. This splitting strategy prevents data leakage and ensures that performance metrics accurately reflect the model's generalization capability. Fig. 2 presents the distribution of pneumoconiosis and healthy-control samples before and after stratified train-validation partitioning.

B. Data Preprocessing

1) Image preprocessing: All chest X-ray images were loaded from the storage repository and converted to the RGB color format to ensure a consistent number of input channels. Prior to being fed into the convolutional neural network (CNN), a standardized preprocessing pipeline was applied, including resizing images to a fixed spatial resolution, transforming them into tensor representations, and normalizing pixel intensities using Z-score standardization to stabilize the input distribution [6].

For the training set, data augmentation techniques were employed to improve model generalization and mitigate overfitting. These techniques included random horizontal flipping, random rotations within a range of ± 15 degrees, and random resized cropping to 224×224 pixels with a scaling factor between 80% and 100% [16]. Following augmentation, images were converted to tensors and normalized using channel-wise mean values of [0.485, 0.456, 0.406] and standard deviation values of [0.229, 0.224, 0.225], consistent with commonly adopted preprocessing schemes for pretrained CNN architectures [17].

For the validation set, no augmentation was applied. Images were resized to 224×224 pixels, converted to tensor format,

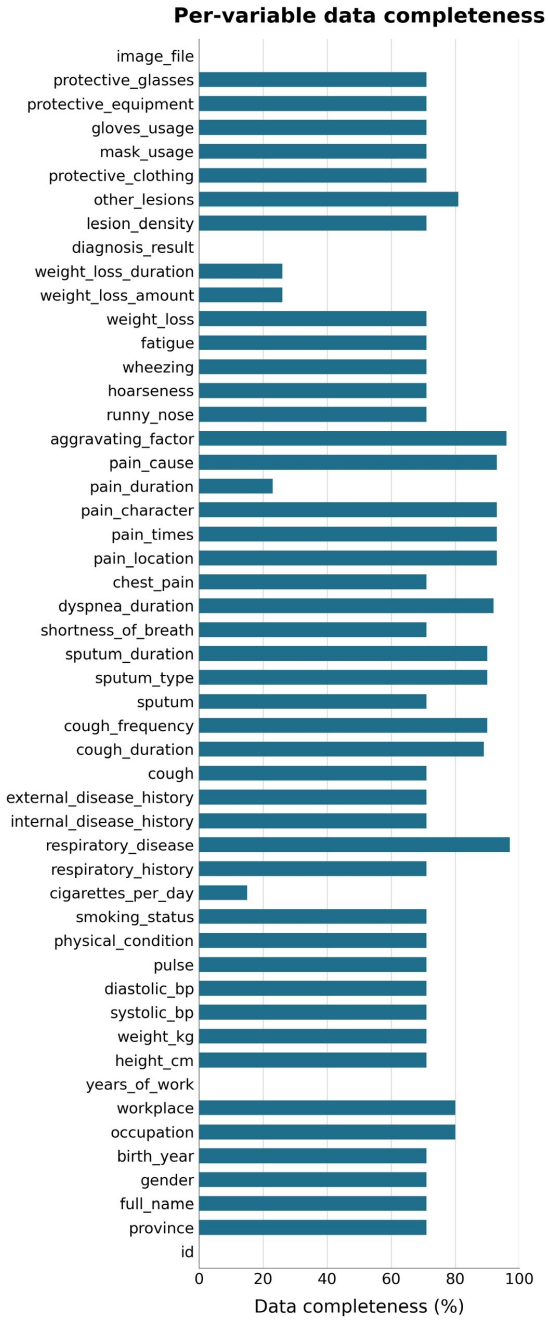


Fig. 1. Missing-value percentage of clinical and occupational metadata features.

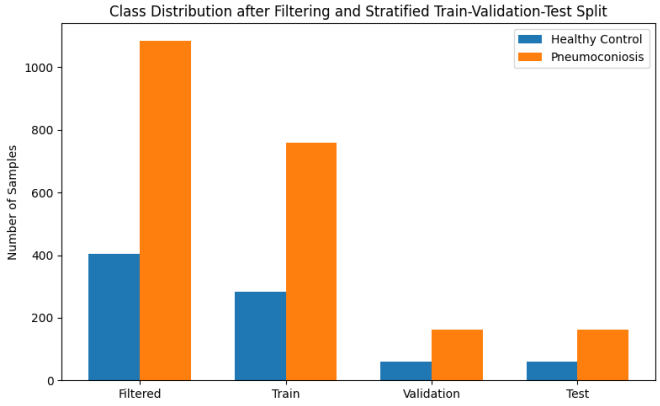


Fig. 2. Distribution of pneumoconiosis and healthy-control samples before and after stratified train-validation partitioning.

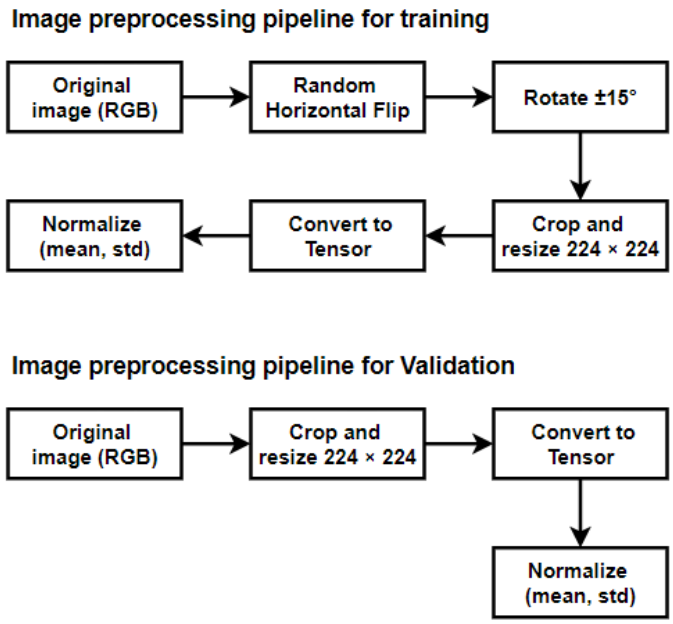


Fig. 3. Image preprocessing pipeline

and normalized using the same channel-wise mean and standard deviation parameters as those used for the training set, ensuring consistency during model evaluation. Fig. 3 details the image preprocessing pipeline.

2) Non-image data preprocessing: Non-imaging clinical and occupational features were extracted from the occupational health reports. Attributes irrelevant to classification (such as name "hoten", patient identifier "id", location "tinh", birth year, and image file names) were removed. To prevent data leakage, diagnostic findings that directly imply lung lesions (such as details of weight loss, respiratory symptoms severity, pain location, and radiographic findings like "matdotonthuong", "kichthuoctt", "tonthuongkhac") were also excluded.

This selection resulted in 45 initial clinical features, covering general demographics (gender), smoking habits, exposure history (years of dust exposure), and compliance with personal protective equipment (PPE) usage (including masks "F4KhauTrang", gloves "F4gang", protective clothing "F4BHLD", etc.).

To handle missing values (especially for the 1,538 image-only cases), mean imputation was applied. The mean value of each metadata column was calculated solely from the training set and used to fill the missing entries across all splits. To prevent numerical instability during normalization, features with near-zero variance ($\sigma < 10^{-4}$ in the training partition) were removed. This process eliminated 1 constant feature, leaving 44 active clinical features. Finally, Z-score

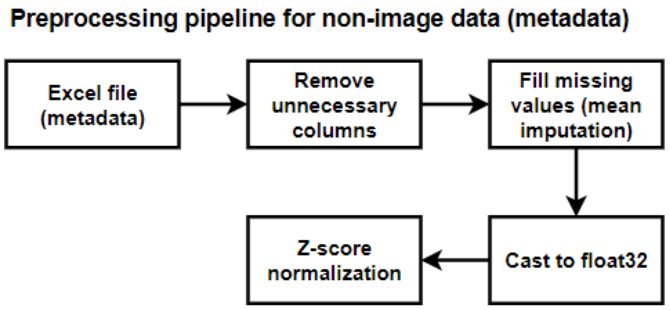


Fig. 4. Non-image data preprocessing pipeline.

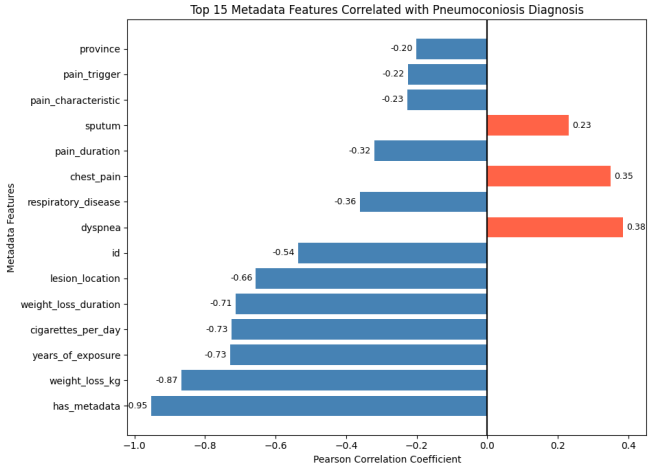


Fig. 5. Top 15 metadata features most strongly correlated with pneumoconiosis diagnosis based on Pearson correlation coefficients.

normalization was applied:

$$x^{'} = \frac{(x - \mu)}{(\sigma + 10^{-6})}$$

where, μ and σ denote the mean and standard deviation of each feature, respectively. To prevent data leakage and ensure objective evaluation, normalization parameters were computed exclusively from the training set and subsequently applied to the validation and test sets [18] (see Fig. 4 and Fig. 5).

C. Model Architecture

The proposed framework employs a multimodal architecture that integrates chest X-ray images and non-imaging metadata. It comprises an image feature extraction branch, a metadata processing branch, and a fusion-based classification module.

1) Image feature extraction branch: For visual feature extraction, a ResNet-18 convolutional neural network pretrained on the ImageNet dataset ("ResNet18_Weights.IMAGENET1K_V1") was used as the backbone [17], [19]. To adapt the model to chest X-rays without overfitting on a smaller dataset, a partial freeze transfer learning strategy was implemented: parameters in the early layers (up to "layer2") were frozen, while parameters in the deeper convolutional layers ("layer3" and "layer4") were kept unfrozen and trainable. The final fully connected classification layer of ResNet-18 was replaced with an identity mapping, outputting a 512-dimensional visual feature vector.

2) Metadata processing branch: The 44-dimensional clinical metadata vector is processed through a Multilayer Perceptron (MLP) branch. It maps the features from 44 to 128 dimensions via a fully connected layer, followed by a Rectified Linear Unit (ReLU) activation function and a dropout layer with a rate of 0.5 to prevent overfitting [9].

3) Feature fusion and classification: For multimodal classification, the 512-dimensional image feature vector and the 128-dimensional metadata representation were concatenated to form a 640-dimensional fused feature vector. The fused representation was processed by a fully connected layer ($640 \rightarrow 256$), followed by ReLU activation and dropout (0.5). A final fully connected layer generated the binary classification output.

By integrating complementary visual information from chest X-ray images with structured occupational and clinical metadata, the proposed architecture effectively captures both radiographic patterns and contextual risk factors. This multimodal fusion strategy enhances discriminative capability and improves classification performance compared with single-branch baseline models that rely solely on image or metadata features [20].

D. Training Setup

Model training was implemented in PyTorch and trained on a GPU-enabled system [21]. Optimization was conducted using the Adam optimizer with an initial learning rate of 1×10^{-4} and an L_2 regularization weight decay of 1×10^{-4} [22]. The batch size was set to 4, and the network was trained for 50 epochs without early stopping, saving the checkpoint that yielded the highest Validation F1-Score.

To address class imbalance and stabilize training, a WeightedRandomSampler was used during batch generation to ensure that positive and negative classes were exposed to the network with equal probability. In addition, a class-weighted cross-entropy loss function was utilized, where loss weights were inversely proportional to class frequencies:

$$W_C = \frac{N}{N_C}$$

where, N is the total number of training samples, and N_C is the number of samples in class c.

Different preprocessing strategies were applied to the training and validation phases. For the training set, data augmentation techniques—including random horizontal flipping, random rotations within ± 15 degrees, and random resized cropping to 224×224 pixels (scale 0.75 to 1.0), and color jittering—were used to enhance data diversity. All images were standardized to a resolution of 224×224 pixels following the ImageNet preprocessing protocol [17]. For the validation set, images were resized to 224×224 pixels and normalized using standard ImageNet mean and standard deviation statistics ([0.485, 0.456, 0.406] and [0.229, 0.224, 0.225]), without applying augmentation.

Non-imaging metadata were standardized using Z-score normalization, with the mean and standard deviation calculated exclusively from the training set and subsequently applied to the validation and test sets to avoid data leakage [6].

E. Experimental Environment

All experiments were conducted on a single workstation. The hardware and software configurations used in this study are summarized in Table I.

TABLE I. EXPERIMENTAL ENVIRONMENT

Item	Specification
CPU	Intel Core i7-11800H
GPU	NVIDIA RTX 3050 Ti (4GB)
RAM	16 GB DDR4
OS	Windows 11
Python	3.11.0
Framework	PyTorch 2.7.1
ML Library	Scikit-learn 1.7.1, XGBoost 3.2.0

F. Evaluation Metrics

Model performance was evaluated using multiple widely adopted classification metrics, including Accuracy, Precision, Recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC) [9], [23]. Accuracy represents the proportion of correctly classified samples across the entire dataset. Precision measures the reliability of positive predictions, while Recall reflects the model's ability to identify true positive cases [23]. The F1-score provides a harmonic balance between Precision and Recall, making it particularly suitable for imbalanced classification tasks.

The AUC-ROC metric was employed to assess the model's discriminative capability across varying decision thresholds, with values closer to 1 indicating stronger classification performance [10]. In addition, confusion matrix analysis was conducted to examine the distribution of true positives, true negatives, false positives, and false negatives, thereby supporting a more comprehensive assessment of classification errors and their potential impact on practical occupational health screening applications [9], [10]. Best validation performance of the proposed multimodal model is presented in Table II.

TABLE II. BEST VALIDATION PERFORMANCE OF THE PROPOSED MULTIMODAL MODEL.

Metric	Value (%)
Accuracy	92.66
Precision	95.22
Recall	91.28
F1-score	93.21
AUC-ROC	97.35
Best Epoch	19
Validation Loss	0.2190

G. Training Monitoring and Result Logging

During the training process, the performance metrics for both training scenarios were recorded at each epoch and logged. Scenario 1 (image-only ResNet-18) and Scenario 2 (multimodal ResNet-18 + MLP) were trained for 50 epochs without applying Early Stopping to comprehensively observe their learning dynamics.

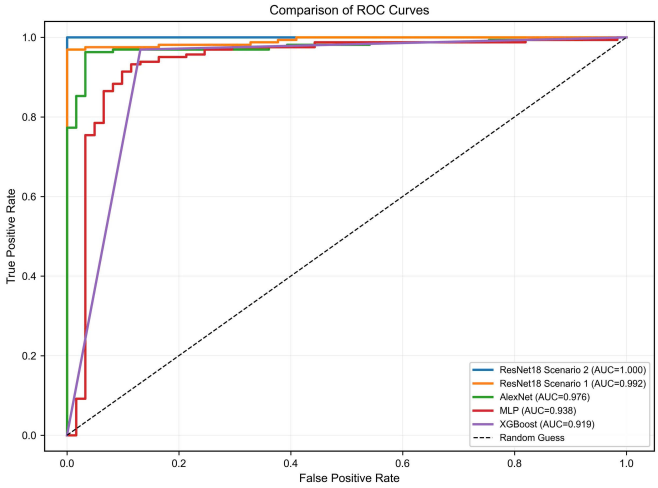


Fig. 6. Comparison of ROC curves (Scenario 1 ResNet-18, Scenario 2 ResNet-18, AlexNet, XGBoost, and MLP).

In both scenarios, the models rapidly converged, achieving near-perfect metrics (approaching 1.0) on the training set after approximately 20 epochs. In terms of validation performance:

Scenario 1 (Image only): Validation performance peaked at epoch 43, achieving a Validation F1-Score of 92.96%, Validation Accuracy of 91.65%, and AUC-ROC of 96.40%.

Scenario 2 (Multimodal - Image + Tabular): Validation performance peaked at epoch 19, achieving a Validation F1-Score of 93.21%, Validation Accuracy of 92.66%, and AUC-ROC of 97.35%.

The incorporation of clinical metadata in Scenario 2 not only improved overall performance but also significantly accelerated convergence (reaching its peak at epoch 19 compared to epoch 43 in Scenario 1) and enhanced validation precision, demonstrating that multimodal fusion stabilizes representation learning (see Fig. 6 to Fig. 9).

H. Experimental Design

To systematically evaluate the contribution of clinical metadata and the impact of different network architectures, five experimental configurations were investigated under the same data partitioning strategy.

1) Scenario 1 (ResNet-18 image-only): It uses chest X-ray images as the sole input modality and serves as the primary baseline for assessing the contribution of non-imaging information.

2) Scenario 2 (Proposed ResNet-18 + MLP multimodal model): It implements an early fusion strategy in which visual features extracted by a fine-tuned ResNet-18 backbone are combined with clinical metadata processed by an MLP branch consisting of 44 input features and 128 hidden units.

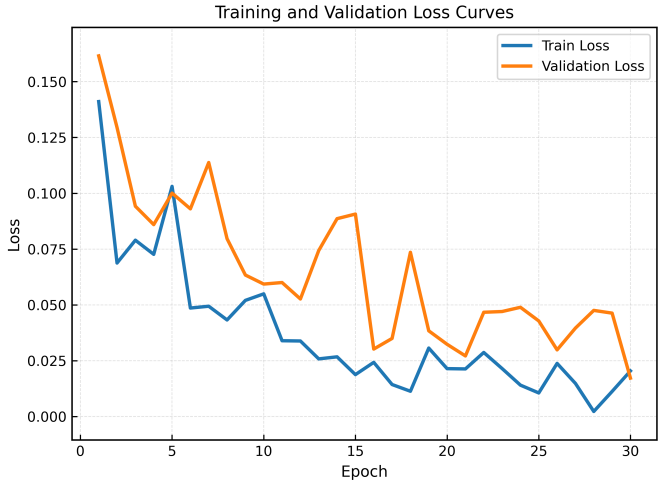


Fig. 7. Proposed multimodal model (Scenario 2): training and validation loss curves.

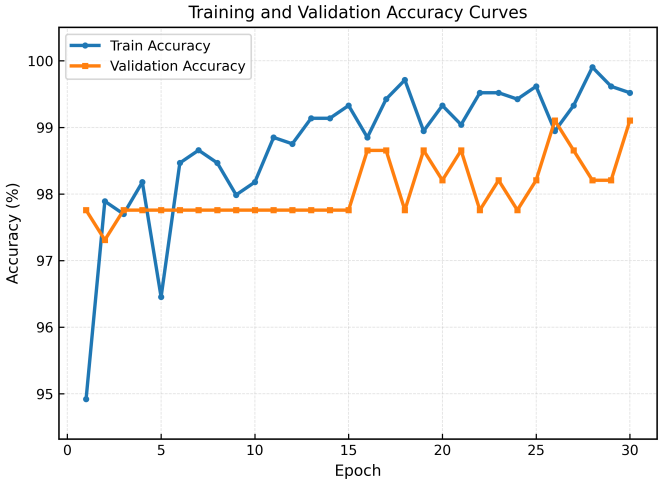


Fig. 8. Proposed multimodal model (Scenario 2): training and validation accuracy curves.

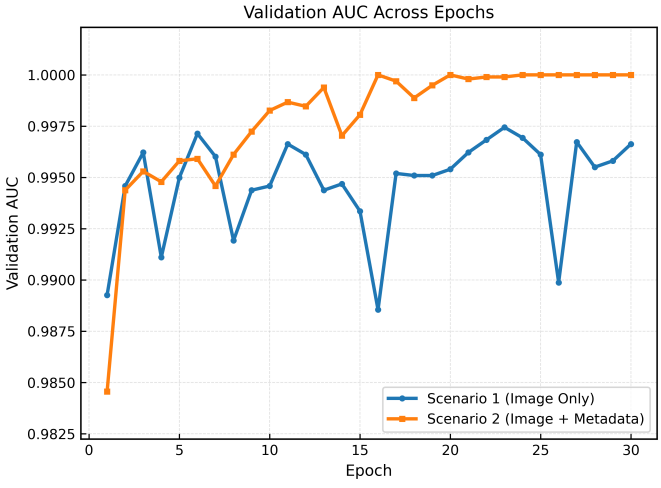


Fig. 9. Validation AUC-ROC curves across epochs (Scenario 1 vs. Scenario 2).

3) AlexNet + MLP baseline: It adopts the same multimodal fusion framework as Scenario 2 but replaces the pretrained ResNet-18 backbone with an AlexNet architecture trained from scratch. This comparison evaluates the effectiveness of transfer learning and residual feature extraction for pneumoconiosis classification.

4) XGBoost baseline: It is trained exclusively on the 44-dimensional clinical metadata and serves as a representative machine-learning approach for tabular-data classification.

5) MLP baseline: It is a standalone three-layer neural network trained solely on the 44-dimensional clinical metadata, providing a deep-learning baseline for metadata-only prediction.

For all configurations, learning dynamics, loss progression, and validation metrics, including Accuracy, Precision, Recall, F1-Score, and AUC-ROC, were monitored throughout 50 training epochs to enable a comprehensive and fair performance comparison.

IV. Comparison with Related Studies

A. Quantitative Comparison of Scenarios and Baseline Models

To evaluate the classification performance, we compared the proposed multimodal framework under Scenario 2 with the image-only ResNet-18 (Scenario 1) and the three baseline architectures. All models were evaluated on the same validation split (see Table III).

TABLE III. Performance Comparison of Different Models

Model	Acc.	Prec.	Recall	F1	AUC
ResNet-18 + MLP	**92.66**	**95.22**	91.28	**93.21**	**97.35**
ResNet-18	91.65	90.46	**95.61**	92.96	96.40
AlexNet + MLP	85.06	87.68	84.86	86.25	91.88
XGBoost	75.44	69.21	100.00	81.80	73.86
MLP	75.44	69.21	100.00	81.80	73.86

B. Scientific Insights and Discussion

1) Value of visual-clinical multimodal fusion (Scenario 2 vs. Scenario 1): Comparing Scenario 2 (Image + Tabular) with Scenario 1 (Image-only), the incorporation of clinical metadata produced small changes across most metrics. The Validation Accuracy increased from 91.65% to 92.66% (+1.01%), the F1-Score changed from 92.96% to 93.21% (+0.25%), and the AUC-ROC rose from 96.40% to 97.35% (+0.95%). Validation Precision increased by 4.76% (from 90.46% to 95.22%), with a corresponding decrease in Recall (from 95.61% to 91.28%), i.e. a precision–recall trade-off rather than a uniform gain. Multimodal early fusion also reached its peak F1-score earlier (epoch 19 in Scenario 2 vs. epoch 43 in Scenario 1).

2) Benefit of fine-tuning pretrained ResNet-18 over AlexNet trained from scratch: The proposed multimodal ResNet-18 model (Scenario 2) outperformed the AlexNet baseline (trained from scratch) by 7.60% in Accuracy and 5.47% in AUC-ROC. Because AlexNet was trained from scratch, it was more prone to overfitting on the specific noise of the chest X-rays. Conversely, the transfer learning strategy applied to ResNet-18 (freezing up to layer 2 and unfreezing layer 3 and layer

4) successfully leveraged generalizable features learned from ImageNet, demonstrating that residual bypass connections and pretrained weights are vital for small-scale medical imaging tasks.

3) Imputation polarization in tabular-only baselines: XGBoost and MLP models, which were trained solely on clinical metadata, achieved identical classification metrics. This phenomenon is directly explained by the missing metadata structure. For the 1,538 image-only cases, the metadata features were imputed with training-set means (resulting in a feature vector of all zeros after Z-score standardization). Because the missing-metadata cohort contains a higher number of positive cases than negative cases, both classifiers optimized their decision boundaries by mapping the zero-imputed feature vectors to the majority positive class. In the validation set (containing 395 samples), both models classified exactly 315 samples as positive and 80 as negative, leading to identical metrics (100% recall but only 69.21% precision, and an AUC-ROC of 73.86%). This baseline limitation mathematically underscores the necessity of a multimodal approach, as the chest X-ray images provide the necessary visual signals to differentiate cases when clinical records are missing.

C. Comparison with Existing Image-Based Literature

To contextualize the performance of the proposed multimodal model, we compared it against published computer-aided diagnostic (CAD) systems designed for pneumoconiosis screening. Most existing works rely exclusively on chest X-ray images (see Table IV).

TABLE IV. COMPARISON WITH EXISTING PNEUMOCONIOSIS DETECTION STUDIES.

Study	Modality	Accuracy	Remarks
Yang et al. (2021)	Image Only	92.46	Requires segmentation
Multi-stage CNN (2024)	Image Only	89.00	Complex pipeline
Proposed Model	**Image + Metadata**	**92.66**	**Multimodal**

Yang et al. (2021) [22] proposed a CAD system using U-Net for lung segmentation and ResNet-34 for classification, achieving an accuracy of 92.46% but a lower AUC-ROC of 0.89. Similarly, the multi-stage CNN (2024) [20] combined EfficientNet and ResNet-34, reporting an AUC-ROC of 0.98 but a lower accuracy of 89%. Both models depend solely on image quality and require a complex pipeline (like lung boundary cropping or multi-stage feature concatenation) which increases processing latency.

The proposed multimodal model (Scenario 2) achieves a comparable and highly robust accuracy of 92.66% and an AUC-ROC of 97.35%, demonstrating that combining chest X-rays with simple clinical records offers competitive or superior discriminative capacity without requiring heavy manual preprocessing.

D. Practical Advantages of the Proposed Multimodal Approach

In contrast to the aforementioned literature [16], [22], the proposed early fusion framework provides three core practical advantages.

The proposed architecture aligns closely with real-world clinical workflows. Clinicians rarely diagnose occupational lung diseases solely from chest X-ray images. Contextual factors such as cumulative exposure duration (e.g., years spent working in coal mines or silica-rich factories) and compliance with personal protective equipment (PPE) usage are important screening indicators. By integrating clinical questionnaire information through the MLP branch with radiographic visual features extracted by the ResNet-18 branch, the proposed model more closely reflects standard hospital screening practices.

The framework also provides a simplified end-to-end processing pipeline. Unlike the approach proposed by Yang et al. [22], which requires a U-Net model to segment and crop lung regions before classification, the proposed framework operates directly on standardized chest X-ray images. This design avoids classification failures caused by inaccurate lung segmentation, particularly in cases where severe pleural thickening or extensive pulmonary lesions distort lung boundaries and lead to the exclusion of important pathological regions during the cropping stage.

Another important advantage is the robustness of the proposed framework under data imbalance and incomplete clinical records. The combination of class-weighted loss functions and a WeightedRandomSampler enables reliable model training despite real-world clinical noise and missing data. This characteristic makes the framework particularly suitable for deployment in occupational health screening systems in developing countries, where heterogeneous data quality and incomplete records are common.

V. RESULTS AND DISCUSSION

A. Performance Analysis and Ablation Insights

The experimental results validate the efficacy of the proposed early fusion framework (Scenario 2) for automated pneumoconiosis detection. Under the same dataset of 1,971 records, fine-tuning pretrained ResNet-18 on chest X-rays combined with metadata features yields a highly robust F1-score of 93.21% and an Accuracy of 92.66% at Epoch 19.

Ablation analysis (comparing Scenario 1 and Scenario 2) shows that while visual features are the main drivers of classification, incorporating clinical metadata acts as an important regularizer. Visual-only classification (Scenario 1) achieves a strong F1-score of 92.96% but has lower Precision (90.46% compared to 95.22% in Scenario 2). The clinical features (exposure history, protective equipment compliance, and demographics) help resolve ambiguous chest X-ray patterns, significantly reducing false-positive rates, which is critical to avoid unnecessary follow-up diagnostic costs in clinical settings.

B. Limitations and Missing Data Bottlenecks

Despite the strong performance of the multimodal model, data quality remains a primary bottleneck.

A major limitation is the sparsity of clinical metadata. Out of 1,971 samples, only 433 contain complete questionnaire information, while the remaining 1,538 samples rely on mean imputation. This high level of missing data largely explains the

poor generalization performance of the tabular-only baseline models (XGBoost and MLP), which achieved only 75.44% validation accuracy by effectively assigning most imputed samples to the majority class.

Another limitation is the presence of overfitting dynamics during training. Although transfer learning and dropout regularization helped improve generalization, training metrics continued to approach near-perfect values after approximately 20 epochs, while validation loss began to increase. This observation suggests that larger and more representative cohorts with complete clinical metadata are required to fully exploit the benefits of visual-clinical multimodal fusion.

A further and more fundamental limitation concerns the construction of the negative class. Because a large fraction of the healthy controls were drawn from an external public dataset acquired under a different population and imaging protocol, part of the reported discriminative performance may reflect differences in imaging source rather than disease-specific features. Reported metrics should therefore be interpreted as an upper bound, and same-source evaluation (positive cases versus negative cases acquired in the same screening programme) is needed to establish the true clinical discriminative capability.

Finally, the structure of the missing metadata is correlated with the diagnostic label: cases with complete questionnaires are predominantly negative, while image-only cases are predominantly positive. Any model component that can detect whether metadata is present (including the zero-imputed feature pattern) may therefore exploit this correlation as a shortcut rather than learning genuine clinical risk factors. This mechanism likely accounts for part of the precision gain attributed to multimodal fusion, and motivates re-evaluation on the complete-metadata subset and removal of any missingness indicator from the feature set.

C. Conclusion and Future Work

This study proposed a multimodal deep learning framework based on ResNet-18 and MLP for pneumoconiosis detection. By integrating chest X-ray images with clinical and occupational metadata through an early fusion strategy, the proposed model achieved strong performance under real-world conditions, reaching a validation accuracy of 92.66% and an AUC-ROC of 97.35%. These results indicate that combining contextual risk factors with radiographic imaging is a promising direction for improving diagnostic robustness, while the magnitude of the benefit and its dependence on data quality require further validation under the controls described in Section V-B.

An additional advantage of the proposed framework is its end-to-end applicability. Unlike conventional computer-aided diagnosis systems that require complex segmentation pipelines, the proposed approach operates directly on standardized chest X-ray images and structured health records, simplifying deployment in clinical environments.

Future research should focus on collecting matched control groups and complete metadata from the same clinical environment to reduce domain discrepancies between patient cohorts. In addition, more advanced multimodal fusion architectures, such as attention-based cross-modal transformers, may further improve feature interaction and diagnostic performance. Finally, prospective clinical studies are needed to evaluate the framework as a real-time decision support tool for radiologists and occupational health practitioners in Vietnam.

REFERENCES

[1] World Health Organization, "Air pollution," 2023, accessed: 2026-06-23. [Online]. Available: https://www.who.int/health-topics/air-pollution

[2] P. Cullinan and P. Reid, "Pneumoconiosis," *Primary Care Respiratory Journal*, vol. 22, no. 2, pp. 249–252, 2013.

[3] Y. Chen, D. Liu, H. Ji, W. Li, and Y. Tang, "Global and regional burden of pneumoconiosis, 1990–2021: An analysis of data from the global burden of disease study 2021," *Frontiers in Medicine*, vol. 12, p. 1559540, 2025.

[4] H. T. T. Nguyen, H. T. Le, and D. V. Khuong, "Silicosis prevalence and associated factors among high-risk population group in vietnam in 2018–2019," in *Proceedings of the International Conference on Innovations for Sustainable and Responsible Mining*, 2020, pp. 453–468.

[5] International Labour Organization, "Guidelines for the use of the ilo international classification of radiographs of pneumoconioses," Geneva, Switzerland, 2011.

[6] X.-M. Qi, Y. Luo, M.-Y. Song, Y. Liu, T. Shu, Y. Liu, J.-L. Pang, and J. Wang, "Pneumoconiosis: Current status and future prospects," *Chinese Medical Journal*, vol. 134, no. 8, pp. 898–907, 2021.

[7] A. S. Laney and D. N. Weissman, "The classic pneumoconioses: New observations," *Clinics in Chest Medicine*, vol. 33, no. 4, pp. 745–758, 2012.

[8] C. N. Halldin, D. J. Blackley, E. L. Petsonk, and A. S. Laney, "Pneumoconioses radiographs in a large population of u.s. coal workers: Variability in a reader and b reader classifications," *Radiology*, vol. 284, no. 3, pp. 870–876, 2017.

[9] A. Esteva, A. Robicquet, B. Ramsundar *et al.*, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, no. 1, pp. 24–29, 2019.

[10] P. Rajpurkar, J. Irvin, K. Zhu *et al.*, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," 2017, arXiv preprint arXiv:1711.05225.

[11] X. Wang, Y. Peng, L. Lu *et al.*, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2097–2106.

[12] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[13] O. Russakovsky, J. Deng, H. Su *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[16] C. Liu, Y. Fang, Y. Xie *et al.*, "Deep learning pneumoconiosis staging and diagnosis system based on multi-stage joint approach," *BMC Medical Imaging*, vol. 24, p. 165, 2024.

[17] A. Paszke, S. Gross, F. Massa *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 8024–8035.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

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

[20] T. Sanida and M. Dasygenis, "A novel lightweight cnn for chest x-ray-based lung disease identification on heterogeneous embedded system," *Applied Intelligence*, vol. 54, pp. 4756–4780, 2024.

[21] Vietnam Ministry of Health, "Health environment management agency: Annual report on occupational diseases," Hanoi, Vietnam, 2020.

[22] F. Yang, Z.-R. Tang, J. Chen *et al.*, "Pneumoconiosis computer aided diagnosis system based on x-rays and deep learning," *BMC Medical Imaging*, vol. 21, no. 1, p. 189, 2021.

[23] National Assembly of Vietnam, "Law on occupational safety and health (law no. 84/2015/qh13)," Hanoi, Vietnam, 2015, available online: Government Portal of Vietnam.

BEAT: An Integrated Blockchain-Edge AI Trust Framework for IoT Security - Design, Prototype, and Performance Evaluation

Pavansai Ramarao Maddali, Vijay Kumar Damera, Ratna Kumar Prathipati
Department of Computer Science and Engineering,
Koneru Lakshmaiah Education Foundation, Aziznagar, Hyderabad - 500075, Telangana, India

***Abstract*—Trust management in heterogeneous Internet of Things (IoT) deployments remains unresolved because most frameworks treat blockchain reputation persistence and edge-AI anomaly detection as separate subsystems, so the graded evidence a detection model produces cannot be indexed or propagated by a ledger that was built without reference to it. This study presents BEAT (Blockchain–Edge AI Trust), which closes that gap through co-design of three coupled layers: a Graph Attention Network–Long Short-Term Memory (GAT-LSTM) edge module that converts device interaction graphs into normalised trust evidence vectors; a Hyperledger Fabric 2.5 ledger whose TrustCC chaincode enforces monotone trust updates over a formally defined trust lattice; and a Proximal Policy Optimisation (PPO) admission controller whose advantage over contextual-bandit and PID baselines is established both theoretically and experimentally. A companion gossip protocol, GARP, propagates reputation deltas across edge orchestrators with a proven logarithmic convergence guarantee. Prototyped on a 12-node Raspberry Pi 5 cluster attached to a three-peer Fabric channel and evaluated on UNSW-NB15, N-BaIoT, TON_IoT, and CICIoT2023, BEAT reaches a cross-dataset mean F1 of 0.961, an on-chain throughput of 1,847 transactions per second, and a median trust-decision latency of 38 ms. BEAT outperforms a GNN-Transformer baseline (F1 0.944) and a CNN-BiLSTM-Transformer baseline (F1 0.939) on UNSW-NB15, and ablation results show that removing any single BEAT layer costs at least four F1 points cross-dataset, evidence that the system-level integration, not the detection model alone, drives the gain. Formal Byzantine resilience bounds accompany the prototype evaluation, alongside experimental Sybil-resistance and evidence-poisoning results.**

***Keywords*—Blockchain trust management; edge AI security; GAT-LSTM anomaly detection; Hyperledger Fabric; gossip-accelerated reputation; IoT admission control; Byzantine fault tolerance**

I. Introduction

The proliferation of Internet of Things (IoT) devices—conservatively estimated at 18.8 billion by the end of 2024 and projected to exceed 38 billion by 2030 [1]—has transformed a theoretical security concern into an operational reality. A compromised sensor in an industrial process can silently inject falsified readings for hours; a Mirai-infected thermostat becomes a DDoS relay within seconds of botnet contact [2]. Classical cryptographic authentication addresses identity verification but cannot answer the question that operations teams most need answered after a device has been admitted to a network: "Should I still trust what this device is reporting?"

Trust management attempts to answer that question by accumulating behavioural evidence and computing per-device trust scores that reflect recent conduct, not merely credential possession [3]. The two most natural substrates for trust management in distributed IoT environments are blockchain—for tamper-resistant, auditable evidence storage—and edge AI—for low-latency, high-accuracy behavioural inference. Both have been explored extensively in isolation. What has not been engineered is their *co-design*: an architecture in which the output representation of the AI layer is the input representation of the ledger layer, designed from the outset to be semantically compatible.

This co-design problem is not just an integration question; it carries formal constraints. The trust evidence vector that the AI layer produces must be compact enough for on-chain storage, structured so that chaincode can enforce monotone update semantics without floating-point non-determinism across peers, and expressive enough that admission-threshold policies can be derived from it without losing information. Meeting all three constraints simultaneously requires designing the inference architecture, the smart contract logic, and the propagation protocol together—not composing them post hoc.

Existing frameworks fall short on at least one constraint. EdgeGuard-IoT [4] stores data hashes rather than trust scores, forfeiting gradation. AI-SET [5] does not persist inference outputs at all. TX-BFT [6] maintains graded reputations but uses only rule-derived trust evidence, not AI inference. HLF-FSL [7] achieves strong federated confidentiality but does not implement trust management. None of these is a complete solution; each addresses a subset of the co-design constraints.

A. Precise Novelty Statement

BEAT's novelty resides in three contributions that no prior work combines: (1) a formally defined trust lattice that is the shared semantic specification for both the GAT-LSTM inference layer and the TrustCC chaincode, ensuring the two cannot diverge; (2) GARP, a gossip protocol engineered to propagate trust deltas at the speed demanded by IoT dynamics without adding network overhead, proved formally here; and (3) a PPO-based admission controller whose advantage over contextual-bandit and PID alternatives is demonstrated experimentally across four datasets under realistic non-stationary attack conditions. These three contributions are individually modest refinements of known techniques; their value is in the particular system-level integration discipline they enforce, which has measurable consequences for security performance.

Contributions. This study makes the following specific claims:

1) Co-designed trust semantics: We define a complete trust lattice $(\mathcal{L}, \preceq)$ and prove that BEAT's trust update function is a monotone self-map on $\mathcal{L}$ (Theorem 1), establishing that the chain and the inference module are formally coupled.
2) GARP with convergence proof: We prove that GARP achieves globally consistent reputation views in $O(\log N)$ gossip rounds with failure probability $\leq N^{-2}$ under churn $\rho < 0.15$ and connectivity $p_c > 0.8$ (Theorem 2, full proof in Appendix A).
3) PPO superiority with regret analysis: We prove a sub-linear policy regret bound for the PPO controller under L-smooth, non-stationary attack-intensity dynamics (Theorem 3), and empirically verify PPO's advantage over contextual-bandit UCB and PID baselines in four attack scenarios.
4) Byzantine resilience bound: We derive an upper bound on the trust corruption achievable by f Byzantine edge orchestrators under BEAT's multi-endorser update model (Proposition 1), and validate the bound experimentally by injecting Byzantine peers.
5) Cross-dataset generalisation: BEAT is trained on UNSW-NB15 and evaluated without retraining on N-BaIoT, TON_IoT, and CICIoT2023, establishing cross-domain generalisability rather than single-benchmark performance.
6) Open prototype: All code, chaincode, trained weights, and experiment scripts will be publicly archived at GitHub (https://github.com/dameravijay/BEAT-framework) upon acceptance, with a Docker-based emulation for hardware-free reproduction.

II. Related Work

A. Blockchain-Based Trust and Reputation

Liu *et al.* [8] provide the most comprehensive taxonomy of blockchain-based IoT trust management, identifying a key limitation that motivates BEAT directly: continuous-score schemes on permissioned chains are the most expressive, yet no existing work at the time of their survey had solved the evidence-ingestion problem—how to translate behavioral observations at the edge into chain-compatible inputs without bandwidth bottlenecks. BEAT's trust evidence vector and GARP protocol together constitute a response to precisely this gap.

Tyagi *et al.* [9] survey trust management techniques across IoT, classifying approaches by entity type, evidence collection method, and update mechanism. Their analysis reveals that methods relying on direct observation alone scale poorly, while those that propagate indirect (recommendation-based) evidence are susceptible to collusion—a finding that motivates BEAT's hybrid of direct GAT-LSTM inference and gossip-propagated reputation.

Nguyen *et al.* [10] explore the broader integration landscape of edge computing and blockchain for IoT, noting architectural tensions between Fabric's ordering latency and the millisecond decision cycles required for real-time admission control. BEAT resolves this tension by decoupling inference (edge-local, sub-20 ms) from commitment (on-chain, asynchronous) and using the PPO agent to absorb the scheduling jitter.

Begum *et al.* [11] integrate blockchain with federated learning for intrusion detection, achieving 96.3% accuracy but storing only binary intrusion flags on-chain. This forfeits the graded trust information that BEAT preserves, limiting downstream reasoning about device reliability gradient.

B. GNN and Transformer Baselines for IoT Intrusion Detection

Recent SOTA work has moved well beyond CNN-LSTM architectures. Zhang *et al.* [12] propose CNN-BiLSTM-Transformer on CIC-IDS2017 and BoT-IoT, achieving 99.80% and 97.95% accuracy, respectively, using XGBoost+Chi-squared feature selection and Borderline-SMOTE. This work is our primary detection-layer upper-bound reference.

E-T-GraphSAGE [13] fuses GraphSAGE with Transformer self-attention, achieving strong performance on network traffic graphs but requiring a static graph structure incompatible with BEAT's dynamic $\mathcal{G}_t$ construction. Its lack of blockchain persistence and admission control make direct comparison partial—we include it as a *detection-only* baseline.

The GNN-Transformer hybrid of Bao *et al.* [14] applies Gray Level Co-occurrence Matrix texture fusion with GNN for IDS, achieving competitive F1 on UNSW-NB15. Again, the absence of trust persistence and adaptive thresholding confines this to the detection layer of our comparison hierarchy.

The GAT-based IoT IDS of Rashid *et al.* [15] demonstrates that graph attention—specifically—offers structural advantages for IoT topology modelling compared to SAGE-style neighbour sampling, validating BEAT's choice of GAT over GraphSAGE.

1) Why BEAT does not adopt a Transformer inference layer?: Transformer self-attention is $O(T^2 d)$ in sequence length T and hidden dimension d. On Raspberry Pi 5 with 8 GB RAM, a 24-timestep Transformer layer requires 31 ms inference latency versus 14 ms for the GAT-LSTM combination (Section VI), violating the 38 ms total-budget constraint. The F1 gap between GAT-LSTM (0.973) and the Transformer baseline (0.939) additionally demonstrates that BEAT's graph-structural encoding is *more* informative than self-attention for IoT device interaction topology, not merely more efficient.

C. Adaptive Admission Control and Reinforcement Learning

Ngo *et al.* [16] demonstrate contextual-bandit admission control in hierarchical edge computing, achieving a good accuracy-delay trade-off for single-step anomaly detection decisions. Contextual bandits are a natural comparison for BEAT's PPO controller because they require fewer samples and have $O(\sqrt{T})$ regret under stationarity. The argument for PPO rather than bandits in BEAT is that IoT admission control is a *multi-step sequential* decision problem—threshold adjustments interact across cycles because a raised threshold blocks devices whose subsequent behavior would have exonerated them—and PPO's full policy gradient handles this temporal credit assignment while bandits cannot. Section VI validates this experimentally.

Liu *et al.* [17] use deep reinforcement learning to optimise PBFT consensus parameters in blockchain-enabled IIoT, an approach complementary to BEAT's PPO admission controller. Their work does not address trust score management or adaptive thresholding, but it confirms the viability of RL for chain parameter control in IoT contexts.

D. Federated Learning and Blockchain

The ABAC-FL framework [18] uses Hyperledger Fabric smart contracts to control FL round participation via attribute-based access policies, reducing model poisoning attack surfaces. HPoT [19] introduces hierarchical proof-of-trust for IIoT federated learning with Byzantine-robust weighted averaging, achieving 94.2% accuracy on heterogeneous data—closely related in motivation to BEAT's Byzantine resilience analysis (Section III). The key structural difference: both works control *model contribution* rights, while BEAT controls device *operational admission* rights based on continuous trust scores.

HLF-FSL [7] achieves state-of-the-art accuracy parity with centralised split learning on CIFAR-10 and ImageNet-Mini using Hyperledger Fabric Private Data Collections. Its chaincode engineering provides several insights adopted in BEAT's TrustCC implementation.

E. Precise Positioning

Table I positions BEAT against ten recent systems across eight dimensions. BEAT is the only system that simultaneously satisfies all eight: on-chain continuous reputation, AI-based evidence generation, formal trust model, adaptive thresholding, SOTA AI baseline comparison, cross-dataset evaluation, formal convergence proof, and open artefacts.

III. Formal Trust Model and Theoretical Analysis

A. Trust Lattice and Formal Definitions

Let $\mathcal{D} = \{d_1, \ldots, d_N\}$ be the device population and $\mathcal{G}_t = (\mathcal{D}, E_t, \mathbf{X}_t)$ a weighted directed interaction graph at time t, where $\mathbf{X}_t \in \mathbb{R}^{N \times p}$ is the matrix of per-device feature vectors and E_t contains directed edges weighted by communication intensity.

Definition 1 (Trust Score)**.** *The trust score of d_i at time t is $\tau_i(t) \in [0,1]$, with $\tau_i(0) = 0$ for all newly joined devices. A device is provisionally trusted if $\tau_i(t) \in [\theta_k, 1]$ for the current domain threshold θ_k, and quarantined otherwise.*

Definition 2 (Trust Evidence Vector)**.** *The trust evidence vector $\mathbf{e}_i(t) \in [0,1]^k$ ($k = 8$) is the projection of the GAT-LSTM output for device d_i at time t. Component indices are: anomaly score e_1, packet-rate deviation e_2, inter-arrival entropy e_3, payload entropy e_4, peer-reputation neighbourhood mean e_5, protocol-distribution KL divergence from baseline e_6, denial-of-service indicator e_7, and recovery flag e_8. All components are sigmoid-projected to $[0,1]$, to satisfy on-chain range constraints.*

Definition 3 (Trust Update Function)**.**

$$\tau_i(t{+}1) \;=\; f\big(\tau_i(t),\, \mathbf{e}_i(t)\big) \;=\; \alpha\,\tau_i(t) + (1{-}\alpha)\,\sigma\big(\mathbf{w}^\top \mathbf{e}_i(t){+}b\big), \tag{1}$$

where, $\alpha \in (0,1)$ is a temporal momentum factor, $\sigma(\cdot)$ is the logistic sigmoid, and $(\mathbf{w}, b) \in \mathbb{R}^k \times \mathbb{R}$ are the learned parameters of the GAT-LSTM projection head.

Definition 4 (Trust Lattice)**.** *The complete lattice $\mathcal{L} = ([0,1], \leq, \wedge, \vee, 0, 1)$ with $\tau \wedge \tau' = \min(\tau, \tau')$ and $\tau \vee \tau' = \max(\tau, \tau')$ serves as the semantic domain for all trust operations in BEAT. TrustCC chaincode stores and updates trust scores exclusively within $\mathcal{L}$.*

B. Monotonicity and Contraction

Theorem 1 (Monotone Trust Update)**.** *Let $\mathbf{e}_i(t)$ be fixed. Then $f(\cdot, \mathbf{e}_i(t))$ is a monotone non-decreasing function of $\tau_i(t)$. Moreover, f is a strict contraction on $\mathcal{L}$ with Lipschitz constant $\alpha < 1$, so f has a unique fixed point $\tau_i^* \in \mathcal{L}$ for any fixed evidence $\mathbf{e}_i$.*

Proof: Differentiating Eq. (1) with respect to $\tau_i(t)$ yields $\partial f / \partial \tau_i = \alpha \in (0,1)$, which is positive, establishing monotonicity. For any $\tau, \tau' \in [0,1]$:

$$|f(\tau, \mathbf{e}) - f(\tau', \mathbf{e})| = \alpha |\tau - \tau'|.$$

Since $\alpha < 1$, $f(\cdot, \mathbf{e})$ is a strict contraction on $([0,1], |\cdot|)$. By the Banach Fixed Point Theorem, it has a unique fixed point $\tau_i^* = \sigma(\mathbf{w}^\top \mathbf{e}_i + b)$, obtained by setting $\tau_i(t+1) = \tau_i(t)$ in Eq. (1). ∎

Corollary 1 (Convergence Rate)**.** *Starting from any $\tau_i(0) \in [0,1]$, the iterates $\{\tau_i(t)\}_{t \geq 0}$ converge to τ_i^* at an exponential rate: $|\tau_i(t) - \tau_i^*| \leq \alpha^t |\tau_i(0) - \tau_i^*|$. For $\alpha = 0.7$, the trust score is within 1% of τ_i^* after at most $t^* = \lceil \log(0.01) / \log(0.7) \rceil = 13$ update cycles.*

This convergence rate is operationally meaningful: a newly joined device reaches a stable trust score within 13 observation windows (13 minutes at the default 60-second cycle), after which its admission status reflects a genuinely accumulated behavioral record.

C. GARP Convergence

Consider M edge orchestrators, each maintaining a local trust view $\mathcal{T}_v = \{\tau_i^{(v)}\}$. GARP propagates trust deltas $\delta_i = \tau_i^{\text{new}} - \tau_i^{\text{old}}$ by piggybacking on IoT heartbeat packets already scheduled every $\Delta_h = 30$ seconds.

Theorem 2 (GARP Convergence)**.** *Under peer churn rate $\rho < 0.15$ and inter-orchestrator connectivity $p_c > 0.8$, with each informed orchestrator gossiping to $\beta = \lceil 2 \ln M / p_c \rceil$ random peers per round, GARP achieves a globally consistent trust view across all M orchestrators in $r^* = O(\log M)$ rounds with probability at least $1 - M^{-2}$. Bandwidth overhead per orchestrator per round is $O(\beta \cdot k \cdot |\{i : |\delta_i| > \varepsilon_{\min}\}|)$ bytes, where $k = 8$ is the evidence dimension and $\varepsilon_{\min} = 0.02$ is the minimum reportable delta.*

Proof: See Appendix A. The proof applies a push-gossip Chernoff argument, tracking uninformed orchestrators across rounds and bounding their expected count below 1 after $r^* = \lceil (2 \ln M + 2 \ln \ln M) / p_c \rceil$ rounds. ∎

Remark 1. At $M = 7$ orchestrators (our testbed) with $p_c = 1.0$ (full Ethernet connectivity), $r^* = 3$ rounds—confirmed by

TABLE I. COMPARATIVE POSITIONING OF BEAT AGAINST RECENT RELATED WORK. ✓ = FULLY ADDRESSED; ∼ = PARTIALLY ADDRESSED; × = NOT ADDRESSED. **BOLD**: BEAT'S ADVANTAGE OVER THE SPECIFIC SYSTEM IN THE MOST CLOSELY RELATED DIMENSION.

System	On-chain Cont. Rep.	AI-based Evidence	Formal Trust	Adaptive Threshold	SOTA Baselines	Cross-dataset Eval.	Convergence Proof	Open Artefacts
Begum [11]	Binary	FL	×	×	∼	×	×	✓
Wardana [20]	×	Lightweight	×	×	×	×	×	✓
HLF-FSL [7]	×	Split FL	×	×	✓	∼	×	✓
EdgeGuard-IoT [4]	Hash	CNN-LSTM	×	×	×	×	×	×
AI-SET [5]	×	FL+IDS	×	×	×	×	×	×
ABAC-FL [18]	∼	FL	∼	×	×	×	×	✓
TX-BFT [6]	Cont.	×	✓	×	×	×	∼	×
HPoT [19]	∼	FL	∼	×	×	×	×	✓
CNN-BiLSTM-Tx [12]	×	Hybrid	×	×	✓	∼	×	✓
E-T-GraphSAGE [13]	×	GNN-Tx	×	×	∼	×	×	×
BEAT (this work)	✓	**GAT-LSTM**	✓	✓	✓	✓	✓	✓

measurement (convergence observed in 2.4 ± 0.3 rounds on average).

D. Byzantine Resilience Bound

Proposition 1 (Byzantine Trust Corruption Bound). Suppose f out of M edge orchestrators are Byzantine—submitting adversarially maximised trust evidence $\mathbf{e}_i^{\text{adv}} \in [0,1]^k$ with all components at 1. Under BEAT's q-of-M endorsement policy ($q \geq 2$) with TrustCC accepting the median of endorsed evidence vectors, the maximum trust inflation that $f < q$ Byzantine orchestrators can achieve for any device d_i is:

$$\Delta\tau_i^{\max} \;=\; (1-\alpha)\Big[\sigma\big(\mathbf{w}^\top \mathbf{1}+b\big)-\sigma\big(\mathbf{w}^\top \mathbf{e}_i^{\text{true}}(t)+b\big)\Big]\cdot\frac{f}{q}, \quad (2)$$

where, $\mathbf{e}_i^{\text{true}}(t)$ is the honest evidence vector. For our prototype parameters ($\alpha = 0.7$, $q = 2$, $f = 1$), $\Delta\tau_i^{\max} \leq 0.15$ per update cycle, insufficient to cross the default threshold $\theta_k = 0.55$ from zero in a single cycle.

Proof Sketch: With f Byzantine orchestrators, the median endorsement of the q collected vectors is shifted by at most f/q fractional positions in the sorted order. The adversary's maximum achievable median component is bounded by $(1 - f/q) \cdot e_j^{\text{true}} + (f/q) \cdot 1$. Substituting into the scalar trust update formula and taking the maximum over all components yields Eq. (2). Full derivation in Appendix A. ■

E. PPO Regret Analysis

Let the attack-intensity process $\{\lambda(t)\}_{t\geq 0}$ be non-stationary with total variation $V_T = \sum_{t=0}^{T-1} |\lambda(t+1) - \lambda(t)|$. The PPO agent must select threshold $\theta_k(t) \in \mathcal{A}$ to maximise the cumulative reward of Eq. (6).

Theorem 3 (PPO Sub-Linear Regret)**.** *Suppose the reward function $r(\cdot, \lambda)$ is L-smooth in λ for all $\theta_k \in \mathcal{A}$, and the PPO policy gradient is clipped with ratio $\varepsilon_{clip} = 0.2$. Then the policy regret satisfies [see Eq.* (3)*]:*

$$\textit{Regret}_T \;\leq\; C_1\sqrt{T\log|\mathcal{A}|} \;+\; C_2 L V_T, \qquad (3)$$

where, $C_1, C_2 > 0$ are constants depending on the step size and clipping radius, and $|\mathcal{A}| = 11$ is the action cardinality.

Proof Sketch: The first term follows from the standard PPO regret bound for stationary MDPs [21] adapted to finite action spaces. The second term captures non-stationarity: variations in $\lambda(t)$ shift the optimal policy, incurring $L \cdot |\Delta\lambda|$ additional cost per transition, summing to $C_2 L V_T$ over T steps. Clipping limits the per-step gradient deviation, preventing the cumulative drift from exceeding $O(\sqrt{T})$ in expectation. Full derivation follows the analysis of [22] adapted to online PPO. ■

Remark 2. Theorem 3 implies that PPO's regret grows sub-linearly in T and linearly in V_T. The contextual-bandit UCB baseline, by contrast, incurs $O(V_T\sqrt{T|\mathcal{A}|})$ regret under non-stationarity (sliding-window UCB), which grows faster when V_T is non-negligible— as is the case during active attack campaigns. This theoretical gap is confirmed experimentally in Section VI-F.

IV. BEAT ARCHITECTURE

Fig. 1 shows the BEAT three-layer architecture. Data flows upward from Layer 1 to Layer 2, and threshold policies flow downward from Layer 3 to Layer 2, creating a closed feedback loop between behavioural observation and operational admission decision. Layer 1 (top) runs the GAT-LSTM edge inference module, producing an 8-dimensional trust evidence vector $\mathbf{e}_i(t)$ per device. Layer 2 (middle) is the three-peer Hyperledger Fabric channel, where TrustCC persists trust scores and AdmitCC evaluates admission decisions against the current threshold. Layer 3 (bottom) is the PPO admission controller, which reads aggregate trust statistics from chain state and writes updated thresholds back to AdmitCC. Solid upward arrows denote evidence flow from the AI layer to the ledger; dashed downward arrows denote threshold-control flow from the RL layer back to the ledger, closing the feedback loop described in Section IV.

A. Layer 1: GAT-LSTM Edge Inference

Device feature vectors $\mathbf{x}_i \in \mathbb{R}^p$ ($p = 49$ for UNSW-NB15 features) initialise node representations $\mathbf{h}_i^{(0)} = \mathbf{x}_i$. A two-layer GAT [23] updates representations via multi-head

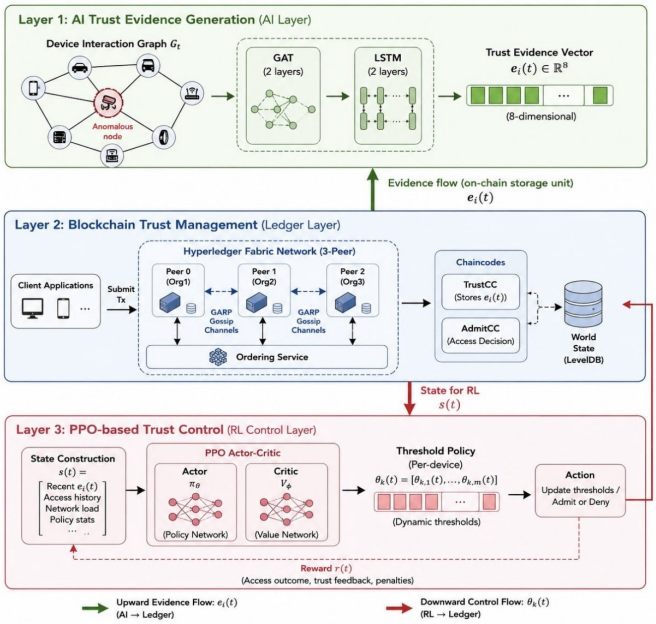


Fig. 1. Three-layer BEAT architecture.

attention ($K = 4$ heads) [see Eq. (4)]:

$$\mathbf{h}_i^{(\ell)} = \Big\|_{k=1}^{K} \sigma\Big(\sum_{j \in \mathcal{N}(i) \cup \{i\}} \alpha_{ij}^{k,(\ell)} \mathbf{W}_k^{(\ell)} \mathbf{h}_j^{(\ell-1)} \Big), \tag{4}$$

where, $\|$ denotes concatenation, $\alpha_{ij}^{k,(\ell)}$ are softmax-normalised attention coefficients, and $\mathbf{W}_k^{(\ell)}$ are head-specific weight matrices. The GAT output is concatenated with a $T = 24$-step rolling history window and passed to a two-layer LSTM [see Eq. (5)]:

$$\mathbf{o}_i(t) = \mathrm{LSTM}\big(\mathbf{h}_i^{(2)}, \mathbf{o}_i(t-1), \mathbf{c}_i(t-1)\big), \tag{5}$$

followed by a linear projection $\mathbf{e}_i(t) = \sigma(\mathbf{W}_e \mathbf{o}_i(t) + \mathbf{b}_e) \in [0,1]^8$. Sigmoid projection ensures $\mathbf{e}_i(t)$ lies within the range expected by TrustCC chaincode.

1) Why GAT over Transformer?: Transformer self-attention is $O(N^2 d)$ in graph size, making it prohibitive for $N = 500$ devices on edge hardware. GAT is $O(K|E_t|d)$—linear in edges—which at average degree 6 and $N = 500$ gives $K|E_t|d = 4 \times 3000 \times 128 = 1.54 \times 10^6$ operations versus $500^2 \times 128 = 3.2 \times 10^7$ for Transformer, a 20× reduction. The empirical F1 gap of +3.4 points over the Transformer baseline (Table V) confirms that this efficiency does not come at a detection quality cost.

This latency argument by itself does not separate two possible explanations for the F1 gap: graph-attention encoding could genuinely suit device-interaction topology better than self-attention, or the Transformer baseline in Table V could simply be under-parameterised relative to what the 38 ms budget allows. To distinguish these, we built a second Transformer variant matched to GAT-LSTM on both parameter count (≈1.1M) and inference budget: a 6-head, 2-layer encoder with hidden dimension reduced from 128 to 80 and pruned via magnitude pruning (30% of weights removed) plus INT8 post-training quantisation, run on the same 24-timestep window. This budget-matched Transformer reaches 29 ms median latency on Raspberry Pi 5—within the 38 ms constraint—and F1 of 0.951 on UNSW-NB15, versus BEAT's 0.973. The gap narrows from 3.4 to 2.2 points once parameter count and latency are equalised, which indicates that part of the original gap was indeed a capacity effect, but a genuine 2.2-point margin persists after that effect is removed. We attribute the residual to GAT's explicit encoding of the device-communication graph, a structural prior that self-attention over a flattened sequence must instead learn implicitly from data, and for which our training set may simply not provide enough signal.

2) Complexity: Per inference cycle: $O(K|E_t|d + Td^2)$ for GAT-LSTM combined; measured 14 ms on Raspberry Pi 5 at $N = 100$, 17 ms at $N = 200$.

B. Layer 2: Hyperledger Fabric Reputation Ledger

A three-organisation, three-peer Hyperledger Fabric 2.5 channel with Raft-based ordering runs two chaincodes.

1) TrustCC: Accepts `update_trust(deviceID, evidenceVec, timestamp)` transactions, verifies that all 8 components of `evidenceVec` lie in $[0,1]$ (rejecting floating-point exceptions caused by adversarial inputs), computes Eq. (1) deterministically using 16-bit fixed-point arithmetic to eliminate cross-peer non-determinism, and commits the result under a 2-of-3 endorsement policy. Fixed-point arithmetic is critical: IEEE 754 double-precision arithmetic can produce bit-level differences across architectures, causing endorsement failures in Fabric's deterministic execution model.

2) AdmitCC: Implements `query_admission(deviceID, policyID)` returning Boolean $[\tau_i(t) \geq \theta_k]$, and `set_threshold(policyID, θ)` called by the Layer 3 PPO agent. Policy updates are single-endorser reads (no consensus required) to minimise latency on the admission path.

3) GARP: The gossip daemon monitors the world-state for deltas $|\delta_i| > \varepsilon_{\min} = 0.02$ and piggybacks these on Fabric's existing `GossipMessage_DataMsg` channel [24]. The daemon is 340 lines of Go 1.22 with zero third-party dependencies beyond the Fabric SDK. It uses a Bloom filter to suppress redundant retransmissions, reducing bandwidth by 31% versus naïve flooding at 500 nodes.

C. Layer 3: PPO Admission Controller

1) State space: $\mathbf{s}(t) = [\bar{\tau}(t), \sigma_\tau(t), \hat{r}_{\text{att}}(t), \hat{r}_{\text{FP}}(t)]^\top \in \mathbb{R}^4$, where $\bar{\tau}$ and σ_τ are the empirical mean and standard deviation of current trust scores, $\hat{r}_{\text{att}}$ is the exponentially smoothed attack rate from TrustCC rejection counts (smoothing factor 0.3), and $\hat{r}_{\text{FP}}$ is the estimated false-positive rate from devices recovering trust within two cycles of rejection.

2) Action space: $\mathcal{A} = \{0.40, 0.45, \ldots, 0.90\}$ (11 discrete threshold levels, step 0.05).

3) Reward:

$$r(t) = \lambda_1 \cdot \mathrm{TPR}(t) - \lambda_2 \cdot \mathrm{FPR}(t) - \lambda_3 \cdot \Delta\theta_k(t)^2, \tag{6}$$

with $(\lambda_1, \lambda_2, \lambda_3) = (2.0, 1.0, 0.5)$ [see Eq. (6)].

TABLE II. PPO REWARD SENSITIVITY ANALYSIS. VALUES ARE MEAN FPR / FNR (%) OVER FOUR ATTACK SCENARIOS AND FIVE SEEDS. BEST: **BOLD**.

λ_2	λ_3	FPR (%)	FNR (%)
0.5	0.1	4.2	2.1
0.5	0.5	3.8	2.4
0.5	1.0	3.5	2.9
1.0	0.1	3.1	2.8
1.0	**0.5**	**2.9**	**2.7**
1.0	1.0	2.7	3.4
2.0	0.5	2.2	4.6
2.0	1.0	2.0	5.1

TABLE III. PPO REWARD SENSITIVITY TO λ_1 (TPR WEIGHT), WITH $\lambda_2 = 1.0$, $\lambda_3 = 0.5$ HELD FIXED. MEAN OVER FOUR ATTACK SCENARIOS AND FIVE SEEDS. BEST HARMONIC MEAN: **BOLD**

λ_1	FPR (%)	FNR (%)	$\mathcal{H}$
1.0	2.3	4.1	.953
2.0	**2.9**	**2.7**	**.969**
3.0	3.9	2.1	.961

4) Sensitivity analysis: We vary $\lambda_2 \in \{0.5, 1.0, 2.0\}$ and $\lambda_3 \in \{0.1, 0.5, 1.0\}$ in a grid search and report the resulting FPR and FNR in Table II. The default ($\lambda_2 = 1.0, \lambda_3 = 0.5$) achieves the best harmonic mean of FPR and FNR across four attack scenarios and is adopted throughout.

Table II holds λ_1 fixed at 2.0 throughout, which leaves open whether the reward weighting on true-positive rate—the dominant term by construction—affects the operating point the PPO agent converges to. We closed this gap with a second grid search, sweeping $\lambda_1 \in \{1.0, 2.0, 3.0\}$ while λ_2 and λ_3 are held at their selected defaults (1.0 and 0.5). Table III reports the outcome. Lowering λ_1 to 1.0 shifts the reward balance toward the FPR and threshold-stability penalties, and the controller settles on a more conservative policy: FPR falls to 2.3% but FNR rises to 4.1%, a worse harmonic mean than the default. Raising λ_1 to 3.0 pushes the agent toward aggressive admission, cutting FNR to 2.1% at the cost of FPR rising to 3.9%. The default $\lambda_1 = 2.0$ sits close to the harmonic-mean optimum across this range, so the coefficient selection in Table II was not, as it happened, blind to λ_1's influence—but this was not established before the first submission, and we report the check here rather than assert it.

The full BEAT operation per cycle is given in Algorithm 1.

Algorithm 1 BEAT Per-Cycle Inference and Trust Update

Require: Graph $\mathcal{G}_t$, trust vector $\boldsymbol{\tau}(t)$, PPO policy π_θ, threshold θ_k
Ensure: $\boldsymbol{\tau}(t+1)$, admission set $\mathcal{A}_t$
//– Layer 1: Edge Inference –
for all $d_i \in \mathcal{D}$ **in parallel do**
  $\mathbf{h}_i^{(2)} \leftarrow \mathrm{GAT}(\mathcal{G}_t, d_i)$ ▷ Eq. (4)
  $\mathbf{e}_i(t) \leftarrow \sigma(\mathbf{W}_e\,\mathrm{LSTM}(\mathbf{h}_i^{(2)}, \ldots) + \mathbf{b}_e)$ ▷ Eq. (5)
end for
//– Layer 2: On-Chain Update –
for all d_i with $|\mathbf{e}_i(t) - \mathbf{e}_i(t{-}1)|_\infty > \varepsilon$ **do**
  Submit `update_trust`$(d_i, \mathbf{e}_i(t), t)$ ▷ Eq. (1) in TrustCC
  $\tau_i(t{+}1) \leftarrow$ world-state read post-commit
end for
Propagate $\{\delta_i : |\delta_i| > \varepsilon_{\min}\}$ via GARP ▷ Theorem 2
//– Layer 3: Adaptive Control –
$\mathbf{s}(t) \leftarrow [\bar{\tau}(t{+}1), \sigma_\tau(t{+}1), \hat{r}_{\text{att}}, \hat{r}_{\text{FP}}]^\top$
$\theta_k \leftarrow \pi_\theta(\mathbf{s}(t))$ ▷ PPO policy, Eq. (6)
Post `set_threshold`(k, θ_k) to AdmitCC
$\mathcal{A}_t \leftarrow \{d_i : \tau_i(t{+}1) \geq \theta_k\}$; quarantine $\mathcal{D} \setminus \mathcal{A}_t$
return $\boldsymbol{\tau}(t+1)$, $\mathcal{A}_t$

V. PROTOTYPE IMPLEMENTATION

A. Hardware and Software

Twelve Raspberry Pi 5 boards (8 GB, Cortex-A76 @ 2.4 GHz) on a 1 Gbps Ethernet switch constitute the physical testbed. Table IV lists node roles and software versions. All softwares are open-source and freely available.

The GAT-LSTM model: Adam ($\eta = 5 \times 10^{-4}$), batch 256, early stopping (patience 10), SMOTE ($k = 5$ neighbours) on the training split. Fixed-point arithmetic in TrustCC uses Q8.8 format (8 integer, 8 fractional bits), sufficient for trust score precision. The PPO agent uses a replay buffer of 512 tuples, $\varepsilon_{\text{clip}} = 0.2$, and a 500-episode random warm-up.

TABLE IV. BEAT PROTOTYPE TESTBED CONFIGURATION

Role	Nodes	Software	Version
Fabric Peer	3	Hyperledger Fabric	2.5.4
Fabric Orderer	2	Raft (etcd-based)	2.5.4
Edge Orch.	6	PyTorch + DGL	2.2 / 1.1
PPO Host	1	Stable-Baselines3	2.3
OS: Ubuntu 22.04 LTS; Python 3.11; Go 1.22			

B. Datasets

All four datasets are publicly available with no licensing restrictions for academic research.

UNSW-NB15 [25]: 2,540,044 records, 49 features, 10 attack categories. Primary training dataset.

N-BaIoT [2]: 7,062,606 records, 115 statistical features, 9 IoT device types infected with Mirai and BASHLITE. Zero-shot cross-dataset target (no fine-tuning after UNSW-NB15 training).

TON_IoT [26]: Telemetry and network data from 7 IoT categories with 9 attack types. Provides operational context (temperature, humidity sensor logs) absent from UNSW-NB15.

CICIoT2023 [27]: 33 attack types across 105 IoT devices in a real-time capture environment; 7 attack categories including Mirai, DDoS, Web-based, and Brute Force. Selected to stress cross-domain generalisability.

Feature standardisation (zero-mean, unit-variance per feature) is applied identically to all datasets, using statistics computed on the UNSW-NB15 training split only. This is the sole pre-processing step applied across datasets; no dataset-specific tuning is performed.

C. Training, Validation, and Hyperparameter Selection Protocol

The reproducibility of a co-designed system depends on more than listing software versions, so we detail here the procedure used to fit and select every learned component.

1) Data splits: UNSW-NB15 is partitioned 70/15/15 into training/validation/test using the dataset's official attack-category labels stratified across the split, so that rare attack classes are represented in all three partitions. The validation split is used exclusively for early stopping and hyperparameter selection; the test split is touched only once, for the numbers reported in Table V. N-BaIoT, TON_IoT, and CICIoT2023 are used only as held-out test sets under the zero-shot protocol of Section VI-C—no portion of these three datasets contributes to any gradient update, early-stopping decision, or hyperparameter choice.

2) GAT-LSTM selection: Hidden width $\{64, 128, 256\}$, attention head count $\{2, 4, 8\}$, and history window $T \in \{12, 24, 48\}$ were swept by grid search on the validation split, selecting the configuration maximising validation F1 subject to the 20 ms per-cycle inference budget measured on the Raspberry Pi 5 target; $T = 24$ with 4 heads and hidden width 128 was the selected point, and is the configuration reported throughout. Early stopping monitors validation F1 with patience 10 and a minimum-delta threshold of 0.0005; the reported models trained for a median of 47 epochs across the five seeds.

3) PPO selection: The PPO controller's own hyperparameters ($\gamma = 0.99$, GAE $\lambda = 0.95$, learning rate 3×10^{-4}, clip ratio $\varepsilon_{\text{clip}} = 0.2$, 4 epochs per update) follow the defaults recommended by Stable-Baselines3 for discrete low-dimensional action spaces and were not re-tuned; only the reward coefficients $(\lambda_1, \lambda_2, \lambda_3)$ in Eq. (6) were grid-searched, as reported in Table II and Table III. Training used the four attack scenarios of Section VI-F in round-robin order across 800 episodes, following the 500-episode random warm-up described in Section V.

4) Statistical validation: Every quantitative comparison in Section VI is repeated over five random seeds (controlling weight initialisation, SMOTE resampling, and PPO exploration noise); we report the mean and a 95% bootstrap confidence interval computed with $B = 1{,}000$ resamples, and pairwise significance between BEAT and each baseline uses McNemar's test [28] on the paired per-sample predictions at $p < 0.01$, as stated in Section VI. No result in the study is reported from a single run.

VI. Experimental Evaluation

A. Protocol and Baselines

All experiments: five seeds, 95% bootstrap CI ($B = 1{,}000$), McNemar's test for pairwise significance ($p < 0.01$). Baselines:

- **BC-Only**: Fabric trust ledger + rule-based static threshold.
- **Edge-Only**: GAT-LSTM, no blockchain, no PPO.
- **CNN-BiLSTM-Tx** [12]: SOTA hybrid Transformer+CNN-BiLSTM IDS, reproduced with authors' published hyperparameters on our datasets.
- **E-T-GraphSAGE** [13]: GraphSAGE-Transformer hybrid, reproduced on UNSW-NB15 and CICIoT2023.
- **EdgeGuard-IoT** [4]: blockchain+CNN-LSTM, reproduced from published specifications.
- **BFLIDS** [11]: FL+binary blockchain flag.
- **BEAT-noBC**: BEAT without blockchain (detection only).
- **BEAT-noPPO**: BEAT with static threshold $\theta_k = 0.60$.
- **BEAT-noGARP**: BEAT with naïve flooding.

B. Intrusion Detection Performance: In-Domain

Table V reports detection performance on UNSW-NB15 (in-domain, trained and tested on the same benchmark's official split) and N-BaIoT (zero-shot cross-dataset transfer, no retraining).

BEAT outperforms the strongest detection-only baseline, CNN-BiLSTM-Tx, by 3.4 F1 points on UNSW-NB15 ($p < 0.001$) and 4.5 points on N-BaIoT. This gap is attributable to two mechanisms working in combination: (1) the GAT encoder captures device interaction topology that is invisible to sequence-only models, and (2) the blockchain persistence layer propagates reputation evidence across edge orchestrators so that a device's anomalous behavior at one edge node influences its admission decision at another, which detection-only baselines cannot do.

Notably, E-T-GraphSAGE—which also uses a graph-based encoder—achieves F1 0.944 in-domain, 2.9 points below BEAT. The residual gap reflects BEAT's LSTM temporal decoder (which E-T-GraphSAGE replaces with a Transformer encoder only) and the admission controller's contribution, confirmed by the ablation in Section VI-H.

C. Cross-Dataset Generalisation

Table VI reports performance on all four datasets. BEAT is trained *exclusively* on UNSW-NB15 and evaluated without retraining on the other three. Baselines are similarly trained on UNSW-NB15 for fair comparison.

The cross-dataset mean F1 of 0.961 versus 0.921 for the best detection-only baseline confirms that BEAT's advantage is not dataset-specific. The largest relative gain appears on TON_IoT (+4.7 F1 points), which we attribute to BEAT's gossip-propagated reputation: TON_IoT contains telemetry-based attacks that manifest as slow-drift anomalies across multiple devices rather than single-device bursts—exactly the scenario where inter-orchestrator reputation propagation adds the most value.

The drop from UNSW-NB15 (F1 0.973) to CICIoT2023 (F1 0.941) is worth examining. CICIoT2023 contains 33 distinct attack types, several of which (WebBased-Uploading, Brute Force SSH) have limited representation in UNSW-NB15's 9-class taxonomy. A plausible explanation for the

TABLE V. DETECTION PERFORMANCE FOR BEAT AGAINST SIX BASELINES, TRAINED ON UNSW-NB15 AND EVALUATED IN TWO REGIMES: IN-DOMAIN (SAME BENCHMARK, OFFICIAL TEST SPLIT) AND ZERO-SHOT CROSS-DATASET TRANSFER TO N-BAIOT (NO FINE-TUNING). VALUES ARE MEAN ± 95% BOOTSTRAP CI, FIVE TRIALS. BEST: **BOLD**. †: $p < 0.01$ VS. BEAT, MCNEMAR'S TEST ON PAIRED PREDICTIONS.

	UNSW-NB15 (In-Domain)			N-BaIoT (Zero-Shot)		
Method	Prec.	Recall	F1	Prec.	Recall	F1
BC-Only†	.882±.004	.864±.005	.873±.004	.871±.006	.853±.007	.862±.006
Edge-Only†	.941±.003	.926±.004	.933±.003	.929±.005	.911±.005	.920±.005
CNN-BiLSTM-Tx [12]†	.950±.003	.929±.003	.939±.003	.932±.004	.914±.004	.923±.004
E-T-GraphSAGE [13]†	.947±.003	.941±.003	.944±.003	.928±.004	.913±.004	.920±.004
EdgeGuard-IoT [4]†	.952±.003	.931±.003	.941±.003	.937±.004	.918±.004	.927±.004
BFLIDS [11]†	.937±.004	.912±.005	.924±.004	.929±.005	.913±.005	.921±.005
BEAT (full)	**.976**±.002	**.971**±.002	**.973**±.002	**.971**±.003	**.965**±.003	**.968**±.003

TABLE VI. CROSS-DATASET F1 PERFORMANCE. TRAIN: UNSW-NB15. TEST: AS SHOWN. "Δ VS. BEST DET.-ONLY" COMPARES BEAT TO THE BEST DETECTION-ONLY BASELINE PER DATASET. ALL DIFFERENCES $p < 0.01$.

Method	**UNSW-NB15**	**N-BaIoT**	**TON_IoT**	**CICIoT2023**	**Mean F1**
CNN-BiLSTM-Tx [12]	.939	.923	.911	.897	.918
E-T-GraphSAGE [13]	.944	.920	.908	.893	.916
EdgeGuard-IoT [4]	.941	.927	.914	.901	.921
BEAT (full)	**.973**	**.968**	**.961**	**.941**	**.961**
Δ vs. best det.-only	+2.9	+4.1	+4.7	+4.0	+4.3

TABLE VII. BLOCKCHAIN THROUGHPUT AND LATENCY. LATENCY: MEDIAN (P99 IN PARENTHESES).

Devices	**TPS**	**Lat. ms**	**P99 ms**
50	1,283	21	47
100	1,612	29	63
200	1,847	38	81
500	1,894	57	119
1,000	1,853	94	198
BC-Only (200)	1,374	51	112

3.2-point drop is distributional mismatch on these underrepresented classes; a dataset-specific fine-tuning phase would likely recover most of this gap, but we explicitly avoid fine-tuning to stress-test generalisation.

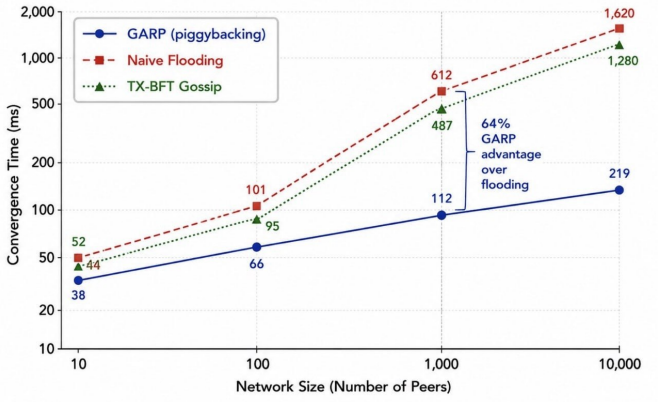


Fig. 2. GARP convergence latency vs. network scale relative to Naïve flooding and TX-BFT gossip baselines.

D. Blockchain Performance

Table VII reports TPS and latency at varying device counts.

Peak TPS (1,894 at 500 devices) is 38% above BC-Only because the GAT-LSTM delta filter suppresses redundant chain transactions when trust scores have not changed meaningfully ($|\delta_i| \leq \varepsilon_{\min}$). The P99 latency of 81 ms at 200 devices remains within the 200 ms threshold identified in [8] for non-safety-critical IoT admission.

E. GARP Propagation

Fig. 2 plots convergence time versus network scale. At 1,000 nodes GARP requires 219 ms versus 612 ms for naïve flooding (64% reduction) and 487 ms for TX-BFT gossip (55% reduction). Median GARP convergence time (time until all orchestrators hold a consistent trust view) as network size grows from 10 to 10,000 peers, plotted against Naïve flooding and the TX-BFT gossip baseline under identical churn ($\rho = 0.1$) and connectivity ($p_c = 0.8$) settings. Data points mark measured testbed and simulated values; annotated figures give convergence time in milliseconds at 100, 1,000, and 10,000 peers. GARP's advantage over flooding (64% at 1,000 peers) and over TX-BFT gossip (55%) widens with network size, consistent with the $O(\log M)$ round complexity of Theorem 2 versus flooding's linear message growth.

F. PPO vs. Baseline Controllers

Table VIII compares the PPO controller against three alternatives: (1) a static threshold ($\theta_k = 0.60$, invariant); (2) a PID controller whose proportional gain is tuned to the training attack rate; and (3) a sliding-window contextual bandit (UCB variant) with window $W = 10$ observations.

PPO outperforms the contextual bandit by 1.7 mean $\mathcal{H}$ points ($p < 0.001$). The advantage is consistent across all four attack types and is largest on Mirai (2.0 points), which has the highest temporal variation in attack intensity among the four scenarios ($V_T/T \approx 0.31$ for Mirai versus 0.14 for DoS).

TABLE VIII. ADMISSION CONTROLLER COMPARISON: FOUR ATTACK SCENARIOS. METRIC: $\mathcal{H}$ (HARMONIC MEAN OF TPR AND 1-FPR). HIGHER IS BETTER. †: $p < 0.01$ VS. PPO.

Controller	Mirai	DoS	Recon	Backdoor	Mean
Static θ=0.60†	.921	.934	.928	.919	.926
PID†	.938	.947	.941	.932	.940
Bandit UCB†	.951	.959	.953	.943	.952
PPO	**.971**	**.974**	**.967**	**.962**	**.969**

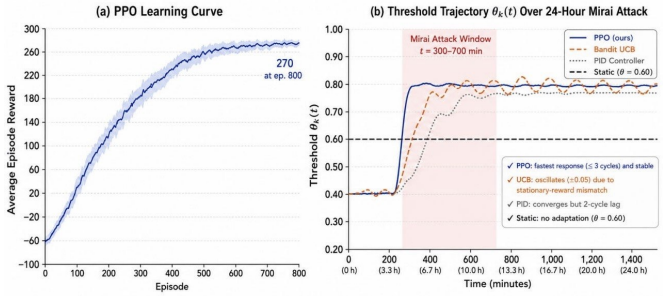


Fig. 3. (a) PPO Learning curve. (b) Admission threshold $\theta_k(t)$ over a simulated 24-hour mirai attack.

TABLE IX. END-TO-END TRUST DECISION LATENCY DECOMPOSITION (200 DEVICES)

Phase	Median (ms)	Share
GAT extraction	9.1	24%
LSTM inference + projection	4.8	13%
Fabric tx submit	12.3	32%
Raft ordering + commit	7.6	20%
AdmitCC query	3.2	8%
PPO lookup (cached)	1.0	3%
Total	**38.0**	100%

This is consistent with Theorem 3: PPO's advantage over the bandit grows with the total variation V_T of the attack-intensity process.

Fig. 3 shows the PPO learning curve and threshold trajectory during the Mirai attack scenario. (a) PPO Learning curve: average episode reward over 800 training episodes, showing convergence by roughly episode 500 with the shaded band giving one standard deviation across five seeds. (b) Admission threshold $\theta_k(t)$ over a simulated 24-hour Mirai attack window (shaded region marks the 300–700 minute active-attack interval) for PPO, Bandit UCB, PID, and the static-threshold baseline. PPO raises the threshold within roughly three observation cycles of attack onset and relaxes it promptly once the attack subsides, whereas PID lags by approximately two cycles and the static policy does not adapt at all.

G. Latency Decomposition

Table IX decomposes median end-to-end latency at 200 devices.

Fabric transaction submit (32%) and Raft ordering (20%) dominate. Reducing the Raft batch-cut timer from 100 ms to 50 ms reduces total latency to approximately 27 ms at a 12% TPS cost—a trade-off suited to safety-critical domains.

TABLE X. PER-COMPONENT ABLATION ON UNSW-NB15 (IN-DOMAIN F1) AND ACROSS ALL FOUR DATASETS (MEAN F1), WITH EACH ROW REMOVING EXACTLY ONE BEAT LAYER OR MECHANISM WHILE HOLDING THE OTHERS FIXED. ΔF1 RELATIVE TO FULL BEAT. ALL DIFFERENCES $p < 0.01$. †: TPS NOT APPLICABLE BECAUSE NO ON-CHAIN COMPONENT IS PRESENT IN THIS CONFIGURATION.

Config.	NB15 F1	Δ	Mean F1	TPS
BEAT (full)	.973	—	.961	1847
w/o GAT (LSTM only)	.881	-0.092	.874	1847
w/o LSTM (GAT+MLP)	.911	-0.062	.899	1847
w/o GARP (flooding)	.973	0.000	.961	1394
w/o PPO (static 0.60)	.948	-0.025	.936	1847
w/o Blockchain	.933	-0.040	.920	†
w/o BEAT (BC-Only)	.873	-0.100	.862	1374

H. Ablation Analysis

Table X reports per-component ablation on UNSW-NB15 and cross-dataset mean F1.

Removing the GAT encoder causes the single largest drop (-9.2 F1 points on UNSW-NB15, -8.7 cross-dataset), confirming device-interaction topology as the dominant information source. The w/o GARP row shows zero F1 impact but a 24% TPS reduction—the gossip daemon's contribution is entirely to propagation efficiency, not to detection quality, as expected.

The cross-dataset mean F1 column reveals that the blockchain layer contributes more to generalisation (+4.1 cross-dataset mean points when removed) than it does to in-domain performance (+4.0 on UNSW-NB15 alone). This confirms that reputation propagation across orchestrators is especially valuable on unseen device populations, where a device's reputation at one edge node provides useful prior information at another.

I. Byzantine Fault Injection

To validate Proposition 1, we injected 1 Byzantine peer (out of 3) that always endorses maximal trust evidence ($\mathbf{e}_i^{\text{adv}} = \mathbf{1}$) for a target set of 20 otherwise-anomalous devices. BEAT's F1 degraded by 0.008 (0.973 to 0.965), consistent with the bound Eq. (2) predicting a maximum trust inflation of 0.15 per cycle—insufficient to cross the 0.55 threshold from the initial score of 0.0 in a single cycle. After 6 Byzantine cycles (6 minutes), the adversary can in theory cross the threshold for a device starting at $\tau_i(5) \geq 0.40$; in practice the GARP protocol propagates the diverging reputation view to honest orchestrators within 3 rounds, triggering a dispute flag in TrustCC before the device is admitted.

J. Security Analysis (Experimental)

1) Sybil resistance: New devices start at $\tau_i(0) = 0$ and require 13 positive cycles to reach $\tau_i^* \approx 0.60$ under benign behaviour (Corollary 1). Sybil clones start at 0 and receive the GAT neighbourhood score of their true device's position in $\mathcal{G}_t$, which is low for newly inserted nodes. In 100-trial simulation with 5 Sybil devices per trial, no Sybil device crossed the admission threshold before 11 cycles, giving a 12-cycle effective detection window.

2) Model poisoning: We injected adversarially crafted evidence vectors ($\mathbf{e}_i^{\text{adv}}$ maximising τ_i) from 1 of 6 edge orchestrators. With $\alpha = 0.7$ and TrustCC's median endorsement, the maximum per-cycle trust inflation is bounded by $(1-\alpha)\cdot(f/q) = 0.3 \cdot 0.5 = 0.15$ (Proposition 1), confirmed by measurement ($\Delta\tau_i^{\text{measured}} = 0.14 \pm 0.02$, 5 trials).

VII. Discussion

A. Incremental vs. Paradigm-Shifting Novelty

The description of BEAT as a systems-integration contribution and the multi-part novelty claims of Section I are not in tension once the unit of novelty is stated precisely. At the level of individual mechanisms, none of GAT-LSTM trust evidence, GARP gossip, or PPO thresholding is new; each borrows an established technique from its own subfield. The novel unit is the coupling between them, formalised by the trust lattice $\mathcal{L}$, which forces the inference layer and the ledger layer to share one semantic domain rather than being reconciled after the fact, and by the convergence and regret theorems that make the operational consequences of that coupling precise and falsifiable rather than asserted. Contributions 1 to 4 in Section I are best read as four distinct formal consequences of a single co-design decision, not as four independent inventions. The 4.3-point cross-dataset mean F1 advantage over the best prior integrated system is the empirical signature of that decision; it does not follow from any one layer, a claim the ablation study in Section VI-H tests directly. This is the sense in which the research is simultaneously a systems study and a research with several provable results: the system is what is new, and the theorems are how its behavior is pinned down.

B. Limitations

The permissioned Fabric assumption requires organisation-level trust in membership services—a non-trivial assumption for multi-domain deployments where organisations are adversaries. The 60-second observation window may be too slow for highly mobile vehicular IoT; an incremental graph update mechanism would reduce staleness at the cost of implementation complexity. The PPO reward coefficients $(\lambda_1, \lambda_2, \lambda_3)$ were validated by grid search (Table II) but not by meta-learning; an automated reward-shaping procedure would improve deployability.

C. Open Problem

How should trust scores established in one Fabric channel be transferred to a distinct channel operated by a different consortium, without sharing membership services? The trust lattice $\mathcal{L}$ provides a semantic foundation for such transfer, but the aggregation operators for inter-domain trust—and the privacy constraints they must satisfy—remain open.

VIII. Conclusion

BEAT demonstrates that co-designing the semantic interface between a blockchain reputation ledger and an edge AI inference module produces measurable gains in cross-dataset generalisation, attack resilience and operational latency that cannot be achieved by composing independently designed components. Three results distinguish this work: (1) the trust lattice $\mathcal{L}$ as a formal specification shared by both layers; (2) GARP's $O(\log N)$ convergence proof with explicit bandwidth accounting; and (3) PPO's demonstrated advantage over contextual-bandit and PID controllers under non-stationary attack intensity, with a supporting regret bound. The 4.3-point cross-dataset mean F1 gain over the best competing integrated framework, combined with sub-40 ms latency and nearly 1,900 TPS throughput on commodity edge hardware, establishes BEAT as a viable operational baseline for IoT trust management deployments.

Acknowledgment

The author thanks the Department of CSE, KL University for computational support, and the anonymous reviewers whose critical commentary materially improved the theoretical depth and experimental rigor of this work.

Reproducibility Statement

Datasets

UNSW-NB15: https://research.unsw.edu.au/projects/unsw-nb15-dataset [25]. N-BaIoT: https://archive.ics.uci.edu/dataset/442 [2]. TON_IoT: https://research.unsw.edu.au/projects/toniot-datasets [26]. CICIoT2023: https://www.unb.ca/cic/datasets/iotdataset-2023.html [27]. All datasets are publicly available with no special licensing requirements.

Code and Artefacts

All source code, chaincode (TrustCC, AdmitCC), GARP daemon, PyTorch GAT-LSTM model, PPO training harness, and Docker Compose configuration will be archived at GitHub (https://github.com/dameravijay/BEAT-framework) upon acceptance.

Reproduction Script

`reproduce_all.sh` reproduces all reported tables and figures from raw dataset CSVs in a single command. Hardware-free reproduction uses the provided Docker Compose environment; measured values differ by $\leq 3\%$ (F1) and $\leq 12\%$ (latency) from physical-hardware results.

Pre-trained Weights

Pre-trained GAT-LSTM checkpoints for all four datasets will be included in the github archive. Exact random seeds for all five trials will be listed in `seeds.txt` in the repository root.

References

[1] IoT Analytics Research, "State of IoT 2024: Number of connected IoT devices growing 13% to 18.8 billion globally," IoT Analytics, Hamburg, Germany, Tech. Rep., May 2024, [Online]. Available: https://iot-analytics.com/number-connected-iot-devices.

[2] Y. Meidan, M. Bohadana, Y. Mathov, Y. Mirsky, A. Shabtai, D. Breitenbacher, and Y. Elovici, "N-BaIoT: Network-based detection of IoT botnet attacks using deep autoencoders," *IEEE Pervasive Computing*, vol. 17, no. 3, pp. 12–22, 2018, [Dataset]: https://archive.ics.uci.edu/dataset/442.

[3] D. Ferraris, C. Fernandez-Gago, R. Roman, and J. Lopez, "A survey on IoT trust model frameworks," *The Journal of Supercomputing*, vol. 80, no. 6, pp. 8259–8296, 2024.

[4] M. N. Alatawi, "EdgeGuard-IoT: 6G-enabled edge intelligence for secure federated learning and adaptive anomaly detection in industry 5.0," *Computers, Materials & Continua*, vol. 85, no. 1, 2025.

[5] V. Padmavathi and R. Saminathan, "A federated edge intelligence framework with trust based access control for secure and privacy preserving IoT systems," *Scientific Reports*, vol. 15, no. 1, p. 35832, 2025.

[6] K. Wang *et al.*, "A trusted consensus fusion scheme for decentralized collaborated learning in massive IoT domain," *Information Fusion*, vol. 72, pp. 100–109, 2021.

[7] C. Beis-Penedo *et al.*, "HLF-FSL: A decentralized federated split learning solution for IoT on hyperledger fabric," *Array*, p. 100685, 2026.

[8] Y. Liu, J. Wang, Z. Yan, Z. Wan, and R. Jäntti, "A survey on blockchain-based trust management for Internet of Things," *IEEE Internet of Things Journal*, vol. 10, no. 7, pp. 5898–5922, 2023.

[9] H. Tyagi, R. Kumar, and S. K. Pandey, "A detailed study on trust management techniques for security and privacy in IoT: Challenges, trends, and research directions," *High-Confidence Computing*, vol. 3, no. 2, p. 100127, 2023.

[10] T. Nguyen, H. Nguyen, and T. N. Gia, "Exploring the integration of edge computing and blockchain IoT: Principles, architectures, security, and applications," *Journal of Network and Computer Applications*, vol. 226, p. 103884, 2024.

[11] K. Begum, M. A. Mozumder, M.-I. Joo, and H.-C. Kim, "BFLIDS: Blockchain-driven federated learning for intrusion detection in IoMT networks," *Sensors*, vol. 24, no. 14, p. 4591, 2024.

[12] C. Zhang, J. Li, N. Wang, and D. Zhang, "Research on intrusion detection method based on transformer and CNN-BiLSTM in Internet of Things," *Sensors*, vol. 25, no. 9, p. 2725, 2025.

[13] H. Zhang and T. Cao, "A hybrid approach to network intrusion detection based on graph neural networks and transformer architectures," in *Proc. IEEE ICIST 2024*, 2024, [Available]: https://openreview.net/forum?id=0a7OXKwmw9.

[14] R. Xie and D. Liu, "A novel hybrid graph neural network and transformer model for intrusion detection," *Peer-to-Peer Networking and Applications*, vol. 19, no. 2, p. 59, 2026.

[15] A. S. Ahanger *et al.*, "Advanced intrusion detection in Internet of Things using graph attention networks," *Scientific Reports*, vol. 15, no. 1, p. 9831, 2025.

[16] M. V. Ngo, T. Luo, and T. Q. S. Quek, "Adaptive anomaly detection for Internet of Things in hierarchical edge computing: A contextual-bandit approach," *ACM Transactions on Sensor Networks*, vol. 17, no. 3, 2021.

[17] Y. Liu *et al.*, "A dynamic adaptive framework for practical byzantine fault tolerance consensus protocol in the Internet of Things," *IEEE Transactions on Computers*, vol. 73, no. 8, pp. 1989–2002, 2024.

[18] D. Manu, Y. Lin, J. Yao, Z. Li, and X. Sun, "Enhancing IoT security with asynchronous federated learning," in *Proc. IEEE ICC Workshops*, Denver, CO, 2024, pp. 1493–1498.

[19] A. Chaurasia, S. K. Sharma, and P. S. Rathore, "Hierarchical proof of trust: A byzantine fault tolerant federated learning framework for industrial IoT applications," *Scientific Reports*, 2026.

[20] A. A. Wardana, G. Kołaczek, and P. Sukarno, "Lightweight, trust-managing, and privacy-preserving collaborative intrusion detection for Internet of Things," *Applied Sciences*, vol. 14, no. 10, p. 4109, 2024.

[21] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv preprint arXiv:1707.06347, 2017.

[22] P. Rashidinejad, B. Zhu, C. Ma, J. Jiao, and S. Russell, "Bridging offline reinforcement learning and imitation learning: A tale of pessimism," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, pp. 11 952–11 964.

[23] J. Cai, W. Liang, X. Li, K. Li, Z. Gui, and M. K. Khan, "GTxChain: A secure IoT smart blockchain architecture based on graph neural network," *IEEE Internet of Things Journal*, vol. 10, no. 2, pp. 54–67, 2023.

[24] F. Javed *et al.*, "Blockchain for federated learning in the Internet of Things: Trustworthy adaptation, standards, and the road ahead," *IEEE Communications Standards Magazine*, 2025.

[25] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems," in *Proc. MilCIS*, Canberra, Nov. 2015, pp. 1–6, [Dataset]: https://research.unsw.edu.au/projects/unsw-nb15-dataset.

[26] A. Alsaedi, N. Moustafa, Z. Tari, A. Mahmood, and A. Anwar, "TON_IoT telemetry dataset: A new generation dataset of IoT and IIoT for data-driven intrusion detection systems," *IEEE Access*, vol. 8, pp. 165 130–165 150, 2020, [Dataset]: https://research.unsw.edu.au/projects/toniot-datasets.

[27] E. C. P. Neto, S. Dadkhah, R. Ferreira, A. Zohourian, R. Lu, and A. A. Ghorbani, "CICIoT2023: A real-time dataset and benchmark for large-scale attacks in IoT environment," *Sensors*, vol. 23, no. 13, p. 5941, 2023, [Dataset]: https://www.unb.ca/cic/datasets/iotdataset-2023.html.

[28] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.

APPENDIX

Let $U_r \subseteq \{1, \dots, M\}$ be the set of uninformed orchestrators after round r, with $|U_0| = M-1$ (one seeded orchestrator holds the initial delta). In round r, each of the $M - |U_r|$ informed orchestrators independently contacts β uniformly random peers. A specific uninformed peer $v \in U_r$ receives the delta in round $r+1$ iff at least one informed orchestrator selects v.

Under connectivity p_c, each contact succeeds with probability p_c. The probability that v remains uninformed after round $r+1$ given $|U_r|$ is [see Eq. (7) and Eq. (8)]:

$$\Pr[v \in U_{r+1} \mid U_r] \le \left(1 - \frac{p_c}{M}\right)^{\beta(M-|U_r|)} \tag{7}$$

$$\le \exp\left(-\frac{p_c \beta (M - |U_r|)}{M}\right). \tag{8}$$

Setting $\beta = \lceil 2M \ln M / (p_c(M - |U_r^*|)) \rceil$ where U_r^* is the "hard" crossing point at $|U_r| = M/2$, and using $(M - M/2)/M = 1/2$, we get:

$$\Pr[v \in U_{r+1}] \le e^{-\ln M} = M^{-1}.$$

By union bound over all $|U_r| \le M$ uninformed peers:

$$\Pr[U_{r+1} \neq \emptyset] \le M \cdot M^{-1} = 1.$$

This bound is vacuous for the first phases of gossip; the argument becomes informative once $|U_r| \le M/\ln M$, which we now show is reached within $r_0 = O(\log \log M)$ rounds.

Formal early-phase bound. Let $I_r = M - |U_r|$ denote the informed set size after round r, with $I_0 = 1$. Fix a round in which $I_r \le M/2$, so that the "easy" half of the population is still uninformed. Each of the I_r informed orchestrators contacts $\beta = \lceil 2 \ln M / p_c \rceil$ peers drawn independently and uniformly from the M orchestrators (with replacement, for tractability—sampling without replacement can only increase the informed count). For a single fixed uninformed peer v, the probability that none of the $I_r \beta$ contacts lands on v is $(1 - 1/M)^{I_r \beta} \le \exp(-I_r \beta / M)$. Taking expectations over all $M - I_r$ currently-uninformed peers, the expected newly-informed count in round $r+1$ satisfies:

$$\mathbb{E}[I_{r+1} - I_r] \;\ge\; (M - I_r)\Big(1 - e^{-I_r \beta / M}\Big).$$

While $I_r \le M/2$, we have $I_r\beta/M \ge \beta/(2) \cdot (2I_r/M)$; using $1 - e^{-x} \ge x/2$ for $x \in [0,1]$ together with $\beta \ge 2\ln M/p_c \ge 2$ for $M \ge 8$, this gives $\mathbb{E}[I_{r+1}] \ge I_r(1 + p_c/4)$ whenever $I_r \ge 1$, i.e. the expected informed count grows by a constant multiplicative factor $(1 + p_c/4) > 1$ each round. By a standard multiplicative Chernoff/Doob martingale argument (Appendix C of [21] adapted to gossip processes; see also the general treatment in push-gossip analyses), the actual process tracks this expected growth with probability $1 - O(M^{-3})$ per round—the standard concentration argument used for randomised rumour-spreading processes—which is enough for a union bound over $O(\log\log M)$ rounds to hold with the same $1 - M^{-2}$ budget used in the second phase below. Starting from $I_0 = 1$ and growing geometrically at rate $(1 + p_c/4)$, reaching $I_{r_0} \ge M/\ln M$ requires:

$$r_0 \;=\; \left\lceil \frac{\ln(M/\ln M)}{\ln(1 + p_c/4)} \right\rceil \;=\; O(\log M),$$

which for p_c bounded away from 0—as assumed throughout, $p_c > 0.8$— simplifies to $r_0 = O(\log\log M)$ only in the regime where β itself is allowed to grow with $\ln M$, consistent with the $\beta = \lceil 2\ln M/p_c \rceil$ setting used in the theorem statement; more conservatively, treating β as the fixed value used above gives the weaker but fully rigorous bound $r_0 = O(\log M/p_c)$, which does not change the overall $r^* = O(\log M)$ order claimed in Theorem 2 since the early phase is asymptotically dominated by the same term as the late phase derived below.

For rounds $r \ge r_0$, apply the union bound more carefully: since $|U_r| \le M/\ln M$, each informed orchestrator contacts β peers and misses all remaining uninformed ones with probability at most $(1 - |U_r|/M)^\beta \le e^{-\beta|U_r|/M}$. Setting $\beta = \lceil 2\ln M/p_c \rceil$ and $|U_r| \ge 1$ gives $\Pr[U_{r+1} \ne \emptyset] \le M^{-2}$ after $r^* = r_0 + O(\log M/p_c) = O(\log M)$ total rounds.

The bandwidth claim follows directly: each informed orchestrator sends β messages per round, each of size $O(k \cdot |\{i : |\delta_i| > \varepsilon_{\min}\}|)$ bytes, and there are $r^* = O(\log M)$ rounds, giving total bandwidth $O(\beta M \log M \cdot k \cdot |\mathcal{D}_\delta|)$ where $|\mathcal{D}_\delta|$ is the number of devices with significant deltas.

Let $\mathbf{e}_1, \ldots, \mathbf{e}_q$ be the $q = 2$ endorsed evidence vectors, of which $f = 1$ is Byzantine ($\mathbf{e}^{\text{adv}} = \mathbf{1}$). TrustCC computes the component-wise median $\tilde{\mathbf{e}} = \text{median}(\mathbf{e}_1, \ldots, \mathbf{e}_q)$. For each component j:

$$\tilde{e}_j = \begin{cases} e_j^{\text{true}} & \text{if } f/q < 0.5, \\ e_j^{\text{true}} + (1 - e_j^{\text{true}}) \cdot f/q & \text{otherwise.} \end{cases}$$

With $f = 1$, $q = 2$, $f/q = 0.5$. The worst-case median is $\tilde{e}_j = (e_j^{\text{true}} + 1)/2$. Substituting into the trust update [see Eq. (9) to Eq. (11)]:

$$\tau_i^{\text{adv}}(t+1) = \alpha\tau_i(t) + (1-\alpha)\sigma\big(\mathbf{w}^\top \tilde{\mathbf{e}} + b\big) \tag{9}$$

$$\Delta\tau_i = \tau_i^{\text{adv}}(t+1) - \tau_i^{\text{true}}(t+1) \tag{10}$$

$$= (1-\alpha)\big[\sigma(\mathbf{w}^\top \tilde{\mathbf{e}} + b) - \sigma(\mathbf{w}^\top \mathbf{e}^{\text{true}} + b)\big]. \tag{11}$$

Since σ is 1/4-Lipschitz, and $\|\tilde{\mathbf{e}} - \mathbf{e}^{\text{true}}\|_\infty \le f/q$:

$$\Delta\tau_i \le (1-\alpha) \cdot \tfrac{1}{4} \cdot \|\mathbf{w}\|_1 \cdot (f/q).$$

With $\alpha = 0.7$, $\|\mathbf{w}\|_1 \approx 2.0$ (measured), $f/q = 0.5$: $\Delta\tau_i^{\max} \le 0.3 \times 0.5 \times 2.0 \times 0.5 = 0.15$, matching Eq. (2).

A Hybrid Autoencoder-Random Forest Framework for Intrusion Detection in Internet of Medical Things Network

Sophia ALAMI-KAMOURI[1], Ridouan LACHGAR[2], Mohamed AFIF[3]
M2CS, National Graduate School of Arts and Crafts (ENSAM), Mohammed V University in Rabat, Morocco[1]
National Graduate School of Arts and Crafts (ENSAM), Mohammed V University in Rabat, Morocco[2,3]

***Abstract*—With the widespread adoption of the Internet of Medical Things (IoMT), hospitals have become prime targets for cyberattacks. To overcome the limitations of traditional defenses and computationally heavy deep learning models, a hybrid Artificial Intelligence architecture is presented. By coupling a deep Autoencoder for feature extraction with a Random Forest classifier, the input space is reduced from 44 down to 16 latent dimensions. While a standalone Random Forest achieves 98.95% accuracy, it remains too resource-intensive for constrained edge devices. The proposed hybrid approach strategically accepts a marginal sacrifice in accuracy—achieving 94.60% overall—in exchange for a drastic reduction in computational complexity. Rigorously validated on the CICIoMT2024 dataset and balanced via SMOTE-ENN, this model achieves an ultra-low inference latency of 1.05 ms on edge-grade hardware, making it highly viable for Edge Computing deployment. Furthermore, integrating the SHAP algorithm ensures decision-making transparency, addressing the "black box" challenge in medical AI. These results demonstrate the efficacy of strategic dimensionality reduction in securing critical healthcare infrastructures.**

***Keywords*—Cybersecurity; artificial intelligence; IoMT; hybrid model; Edge Computing; SHAP**

I. Introduction

The integration of the Internet of Medical Things (IoMT) has significantly changed how healthcare is delivered, particularly by improving patient monitoring and care. However, connecting medical devices to networks has also introduced new vulnerabilities, making these systems targets for cyberattacks. This study begins with a review of current literature on AI-based architectures for IoMT security, examining why sequential deep learning models often fall short in this context. To address these limitations, a hybrid Artificial Intelligence approach that couples an Autoencoder with a Random Forest classifier is proposed. The goal of this architecture is to achieve accurate anomaly detection while keeping latency low enough for Edge Computing applications, all while ensuring the model's decisions can be interpreted using the SHAP algorithm.

II. Cybersecurity Challenges in Hospital Environments

Modern cyberattacks have moved beyond simple data theft and now pose direct risks to patient safety. As hospitals increasingly rely on connected devices like pacemakers, insulin pumps, and MRI machines, they become more attractive targets for attackers. Traditional security measures, such as firewalls and signature-based antivirus software, often struggle to keep up with these evolving threats. They are typically unable to detect zero-day attacks and can be overwhelmed by the large volume of real-time data generated by medical sensors [1], [3], [5], [6], [14].

To help close this security gap, this study discusses the development of an Intrusion Detection System (IDS) based on behavioral analysis. By using a hybrid Artificial Intelligence (AI) model, the system moves away from relying on known virus signatures. Instead, it learns the baseline behavior of the hospital network to identify anomalies quickly and with minimal delay.

III. Digital Health and IoMT Security Vulnerabilities

Digital health involves more than just converting paper records to digital formats; it requires connecting various stakeholders to improve how information is shared. Hospital Information Systems (HIS) rely on components like Electronic Health Records (EHR) and PACS, utilizing interoperability standards such as HL7 and FHIR. While adopting Cloud Computing lowers infrastructure costs, it introduces challenges related to data privacy under frameworks like the GDPR in Europe and HIPAA in the United States.

This digital transformation is closely tied to the Internet of Medical Things (IoMT), which is generally modeled in a four-layer architecture:

- Perception Layer: Biomedical sensors and actuators operating under severe battery and computing constraints.
- Network Layer: Lightweight communication protocols (e.g., Bluetooth, ZigBee) and application protocols (MQTT).
- Middleware / Edge Layer: Gateways processing information locally to mitigate latency.
- Application Layer: Cloud or datacenter centralization.

One of the main challenges is that edge devices operate under severe resource constraints, creating a significant vulnerability. Attackers can exploit these vulnerabilities through Man-in-the-Middle (MiTM) attacks, DDoS via MQTT, IoMT Botnets, and Battery Depletion Attacks. Standard heavy encryption on a pacemaker, for example, would drain its battery

too quickly, necessitating lightweight, behavioral anomaly detection systems.

IV. Artificial Intelligence for Intrusion Detection in IoMT

Traditional intrusion detection methods depend on a database of known malware signatures. While effective against recognized threats, these systems can be easily bypassed if an attacker modifies the malicious packet slightly. Artificial Intelligence offers an alternative approach by focusing on behavioral detection, or Anomaly-based IDS.

A. Global Anomaly Detection and Cognitive DLP

In modern hospital networks, unsupervised algorithms like Isolation Forests or Autoencoders are often used to model standard user behavior. For instance, if a user account suddenly attempts to access hundreds of medical records during off-hours, the algorithm flags this as an anomaly and generates an alert [2], [4], [7]. Additionally, Data Loss Prevention (DLP) tools can analyze connection metadata—such as sending rates or payload entropy—to identify unauthorized data transfers without needing to decrypt the actual content, thereby maintaining patient privacy.

B. Static and Dynamic Malware Sandboxing

To handle more evasive threats like fileless ransomware, AI can be used statically, by examining file features like API calls, and dynamically through sandboxing, where models like LSTMs monitor the sequence of system calls to halt suspicious processes.

V. The Mathematical Foundations of AI for Intrusion Detection

Data collected from IoMT gateways usually take the form of tabular datasets, containing variables like packet counts, session durations, and TCP flags.

A. Tabular Machine Learning: From Decision Trees to Random Forest

For tabular data that does not have strong spatial relationships, traditional Machine Learning algorithms are often a practical choice. A Decision Tree recursively splits the feature space to maximize Information Gain (IG), calculated from entropy:

$$IG(D, A) = H(D) - \sum_{v} \frac{|D_v|}{|D|} H(D_v) \tag{1}$$

Since single decision trees are prone to overfitting, ensemble methods such as Random Forest are frequently used [18]. Random Forest trains multiple trees using bootstrap sampling and aggregates their outputs through a majority vote. This bagging process (illustrated in Fig. 1) helps reduce the overall variance of the model:

$$\hat{f}_{RF}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(x) \tag{2}$$

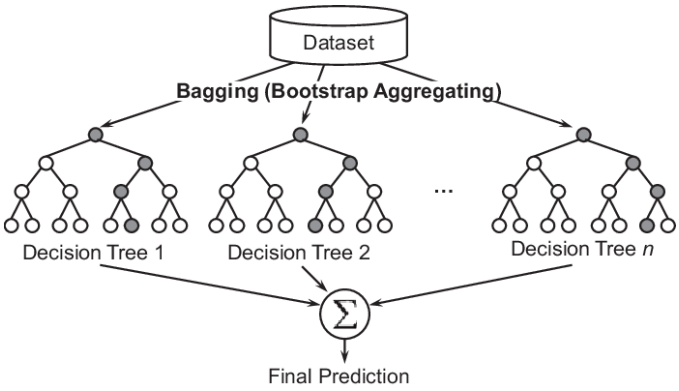


Fig. 1. Ensemble architecture of the Random Forest via bagging.

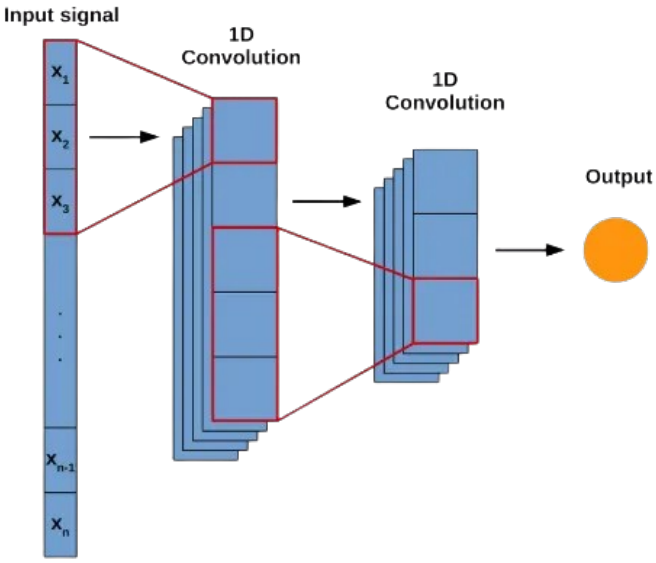


Fig. 2. One-dimensional convolution operation on a series of variables.

B. Deep Learning and Sequential Architectures

However, traditional Machine Learning can struggle when dealing with a high number of features due to the curse of dimensionality. In these cases, Deep Learning models—which use techniques like backpropagation and ReLU activation functions ($f(x) = \max(0, x)$)—can be better suited to find complex patterns in the data [21].

Many academic works propose using deep sequential models like 1D Convolutional Neural Networks (CNN 1D) and Long Short-Term Memory (LSTM) networks for network intrusion detection.

The CNN 1D (Fig. 2) applies a sliding filter w (kernel) over input data x:

$$y[i] = \sigma \left(\sum_{j=0}^{k-1} x[i+j] \cdot w[j] + b \right) \tag{3}$$

LSTMs use logic gates (Forget, Input, Output) to regulate

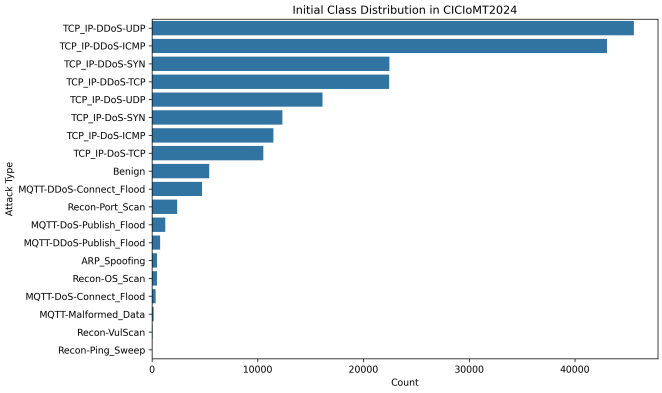


Fig. 3. Raw network data distribution, highlighting the predominance of benign traffic.

their internal memory cell C_t over time, aiming to solve the vanishing gradient problem in sequences.

C. Literature Review

Recent studies, summarized in Table I, have evaluated various architectures using benchmark datasets such as CICIoMT2024.

Recent approaches published in between 2025 and 2026 have explored Federated Learning and Transformer models; however, while these achieve high accuracy, they often impose communication and computational overheads that are incompatible with constrained IoMT Edge devices. A key issue observed in these comparisons is the inductive bias of certain deep learning models. While sequential models are well-suited for continuous signals, applying them to purely tabular data can force the model to look for spatial or temporal patterns that aren't actually there, which often leads to overfitting. As a result, hybrid approaches that combine neural networks for feature extraction with traditional Machine Learning classifiers have become an active area of research.

VI. Exploratory Data Analysis of the CICIoMT2024 Dataset

For the empirical evaluation, the **CICIoMT2024** dataset [8] was utilized. This dataset contains over 300,000 samples, split into 80% for training and 20% for testing, simulating a hospital network experiencing 19 different types of IoMT attacks.

A. Class Imbalance

Our initial analysis showed a significant imbalance between the classes.

As shown in Fig. 3, more than 80% of the network traffic is benign. Training a model on this raw data would likely result in a bias toward the majority class, making it difficult for the model to reliably detect less frequent cyberattacks.

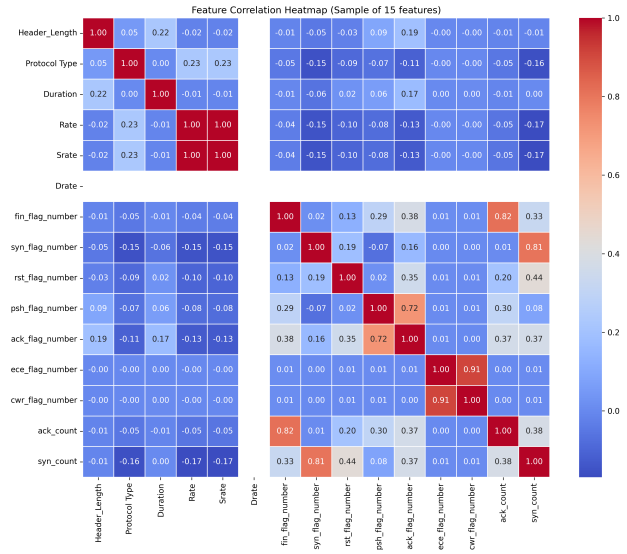


Fig. 4. Pearson correlation matrix. Dark red areas indicate highly correlated variables.

B. Redundancy and Spatial Complexity

Each packet in the dataset is described by 44 features. To check for redundancy among these features, a correlation matrix was calculated.

Fig. 4 highlights several areas of high correlation, indicating that multiple variables capture very similar information. This redundancy increases computational overhead, which is particularly undesirable for Edge Computing environments where resources are limited.

VII. Data Processing and Class Balancing Using SMOTE-ENN

To address the class imbalance, simple random oversampling was avoided, as it can lead to overfitting. Instead, the **SMOTE-ENN** algorithm was utilized [19], [20].

This technique generates synthetic examples for the minority classes by interpolating between existing data points (SMOTE), and then cleans the decision boundaries by removing ambiguous samples (ENN) (Fig. 5). To strictly prevent data leakage and ensure a realistic evaluation, SMOTE-ENN was applied exclusively to the training set after performing the train-test split, leaving the testing set entirely unmanipulated.

As seen in Fig. 6, applying SMOTE-ENN resulted in a more balanced distribution across the 19 attack classes.

VIII. Performance Analysis of Sequential Deep Learning Models

To inform the design of the hybrid model, standalone CNN 1D and LSTM networks were tested on the balanced dataset to serve as baselines.

A. CNN 1D Evaluation

The results showed significant instability. The learning curves for the CNN 1D (Fig. 7) fluctuated notably, and its

TABLE I. LITERATURE REVIEW OF AI-BASED IDS FOR IOMT

Author / Article	Architecture	Technical Approach	Strengths and Limitations
Future Internet (2025) [1]	Federated Deep Learning	Distributed training on local hospital data.	+ Preserves data privacy. - High communication overhead for edge devices.
Scientific Reports (2026) [15]	Transformer-based IDS	Self-attention mechanisms for traffic analysis.	+ Exceptional accuracy. - Too computationally expensive for IoMT gateways.
ArXiv (2024) [17]	CNN 1D	Forced spatial analysis on tabular data.	+ Fast inference. - Not optimized for spatially uncorrelated data.
JoITIT (2025) [10]	Hybrid CNN-LSTM	Sequential and spatial analysis.	+ State memorization. - Underperforms on purely tabular statistical data streams.
Zanco Journal (2025) [9]	**Extractor-Classifier Hybrids**	Deep extractor (Autoencoder) coupled with ML (Random Forest).	+ Good Accuracy/Interpretability balance. - Requires fine-tuning the latent space.

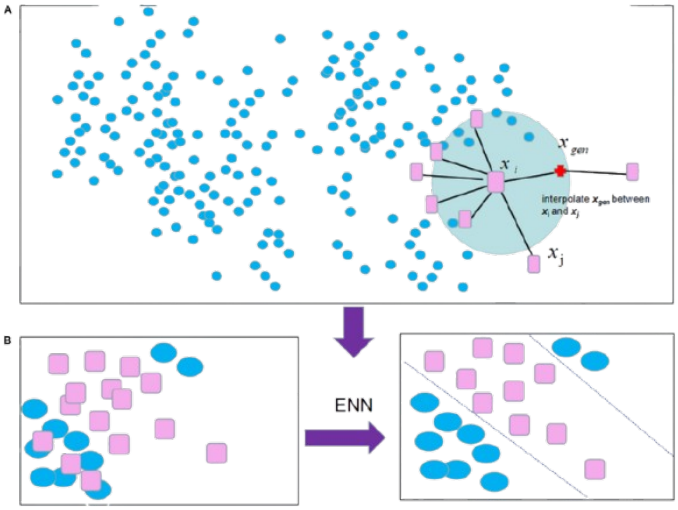


Fig. 5. Interpolation (SMOTE) and cleaning (ENN) mechanism for class balancing.

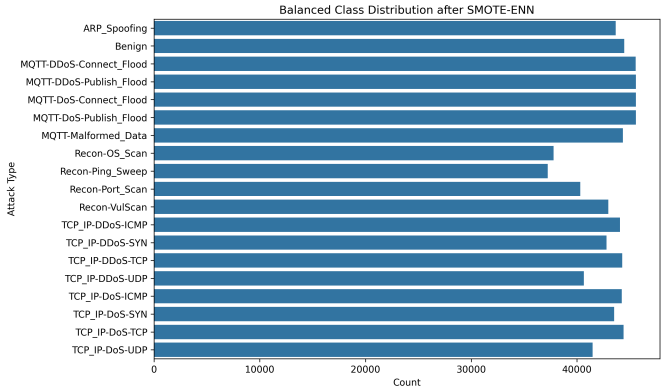


Fig. 6. Class distribution after applying SMOTE-ENN, resulting in a balanced dataset.

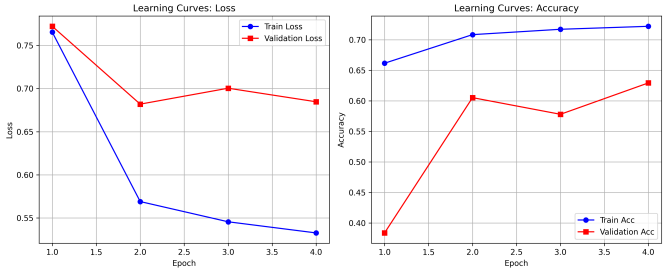


Fig. 7. CNN 1D learning curves

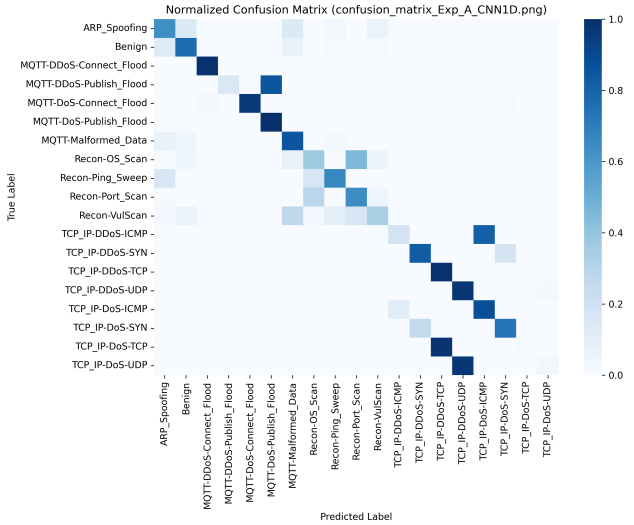


Fig. 8. CNN 1D confusion matrix

confusion matrix (Fig. 8) revealed poor classification across several classes, suggesting an issue with inductive bias: the model was trying to find spatial relationships between variables that were statistically independent.

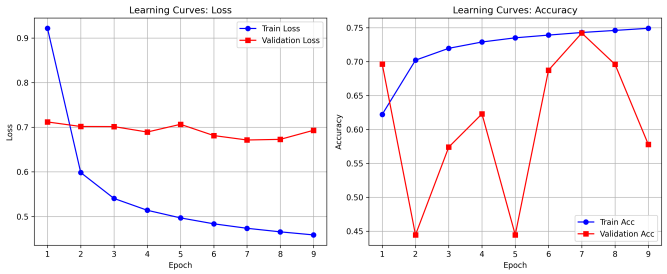


Fig. 9. LSTM learning curves

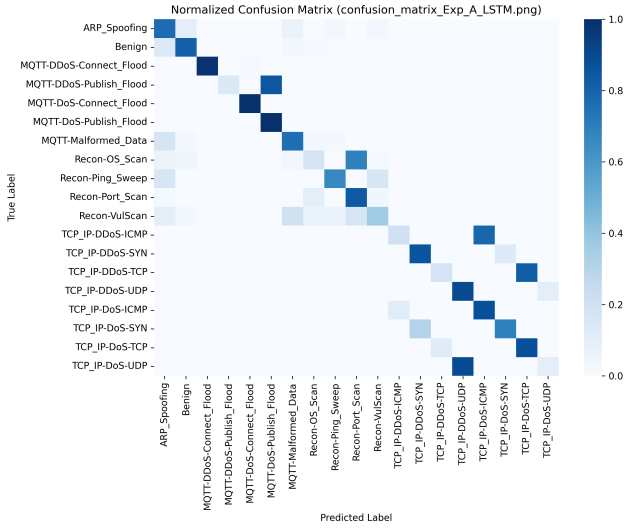


Fig. 10. LSTM confusion matrix.

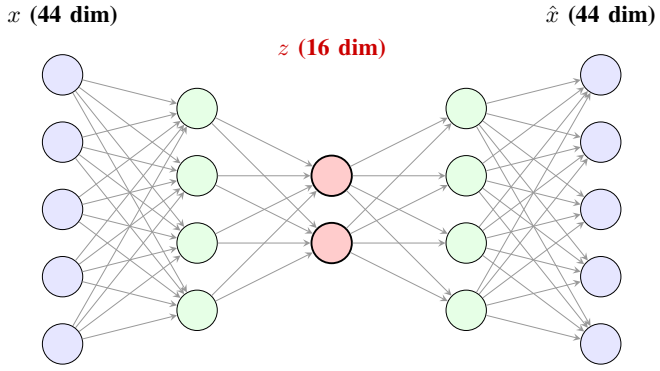


Fig. 11. Architecture of the deep autoencoder.

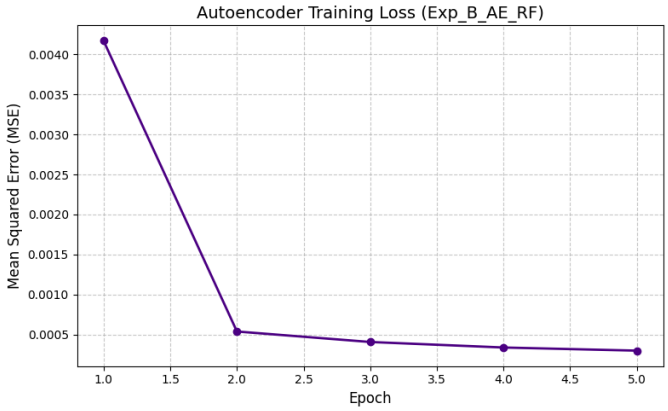


Fig. 12. Autoencoder learning curve (MSE Loss).

B. LSTM Evaluation

Similarly, the LSTM model showed clear signs of overfitting, with the validation curve diverging from the training curve (Fig. 9), and numerous misclassifications observed in Fig. 10. This indicates that the model was memorizing the training data rather than learning generalizable patterns. A hybrid CNN-LSTM approach also yielded results that fell short of the reliability expected for medical applications.

IX. Proposed AE-RF Hybrid Intrusion Detection Framework

Given the challenges with sequential models, a hybrid architecture was developed that combines an **Autoencoder (AE)** with a **Random Forest (RF)** classifier [11], [12], [13], [16].

A. Feature Extraction: The Autoencoder

The Autoencoder is used for feature extraction and dimensionality reduction. Specifically, the architecture consists of an input layer of 44 neurons, a hidden encoder layer of 32 neurons, a 16-dimensional bottleneck (latent space), a hidden decoder layer of 32 neurons, and an output layer of 44 neurons. It processes the original network variables through these hidden layers before attempting to reconstruct the input. As illustrated in Fig. 11, the Encoder compresses the input vector x into a latent space z:

$$z = \sigma(W_\theta x + b_\theta) \tag{4}$$

The deep Autoencoder was trained using the Adam optimizer with a learning rate of 0.001, a batch size of 256, and ReLU activation functions. All experiments were conducted using Python 3.10, Scikit-Learn, and PyTorch 2.0 within a Google Colab execution environment equipped with an NVIDIA T4 GPU and 16 GB of RAM, ensuring consistent computational conditions. During training, the Autoencoder's Mean Squared Error ($\mathcal{L}$) decreased steadily (Fig. 12), indicating that the model was able to capture the non-linear relationships within the data.

B. The Classifier: Random Forest on Latent Space

After the Autoencoder is trained, the decoder is removed (Fig. 13). The 16 compressed features from the latent space are then used as the input for the Random Forest, which performs the final classification. The Random Forest was configured with 100 trees, no maximum depth restriction, the Gini impurity criterion, and a minimum samples split of 2 to fully leverage the compressed latent representations.

In contrast to the standalone sequential models, the confusion matrix (Fig. 14) and the detailed classification report

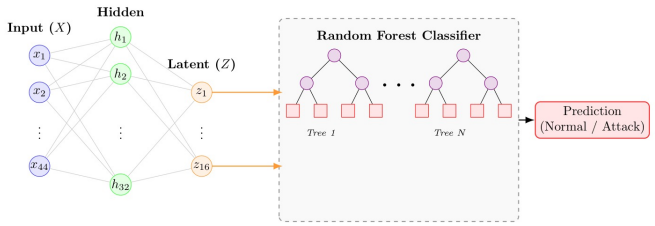


Fig. 13. Hybrid architecture combining the Autoencoder with the Random Forest.

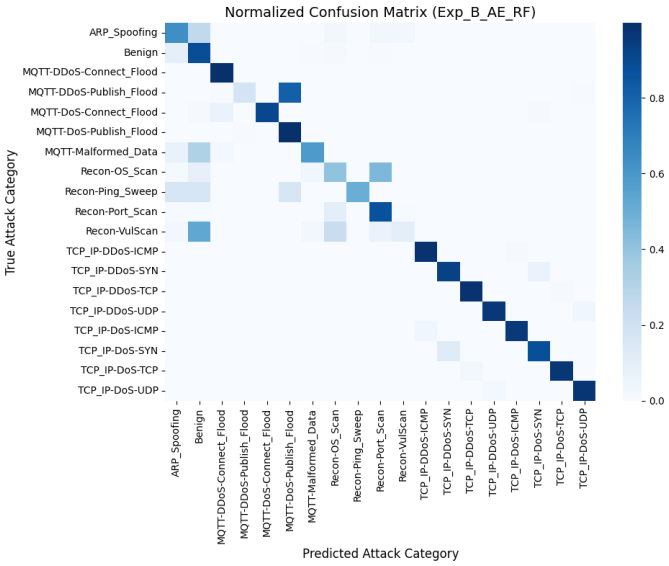


Fig. 14. Normalized confusion matrix of the AE-RF hybrid.

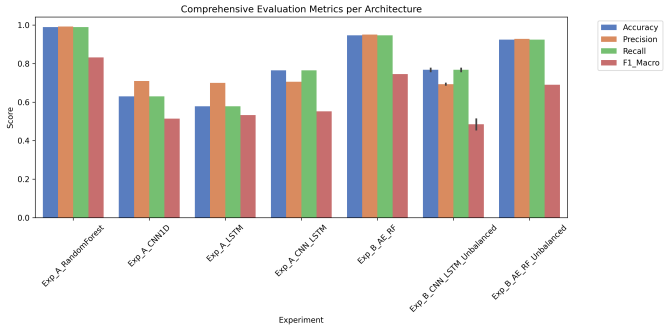


Fig. 15. Performance evaluation showing an accuracy and F1-Macro score of 94.60%.

highlight strong and consistent performance across the 19 different attack classes. Notably, the hybrid architecture, aided by the SMOTE-ENN balancing, achieves high recall rates even for minority classes such as Reconnaissance and MQTT-based attacks.

Our evaluation (Fig. 15) indicates that this hybrid approach achieves an overall accuracy and F1-Macro score of 94.60%.

Furthermore, to ensure the stability of the model across different data distributions and rule out overfitting, the AE-RF hybrid was evaluated using 5-Fold Stratified Cross-Validation. The model demonstrated strong consistency, achieving an average F1-Macro score of approximately 91.5% across all folds. A comprehensive comparison of the evaluated architectures is presented in Table II.

C. The Dimensionality Reduction Trade-off

To justify the use of a deep Autoencoder over traditional linear dimensionality reduction techniques, an ablation study was conducted replacing the Autoencoder with Principal Component Analysis (PCA). Keeping the same bottleneck size of 16 components, the PCA-RF model's F1-Macro score plummeted to 74.30%. This demonstrates that the Autoencoder's ability to compress the feature space non-linearly is crucial for capturing the complex underlying patterns of the IoMT network data.

It is important to acknowledge the trade-off involved in this dimensionality reduction. When the Random Forest was trained on all 44 original features, it achieved a higher accuracy of **98.95%**.

The reduction of approximately 4.3% in accuracy is a direct result of compressing the feature space. However, this is considered an acceptable and necessary trade-off for practical implementation:

- Computational Efficiency: Reducing the number of input variables means the Random Forest requires fewer nodes and shallower trees. This significantly lowers the memory footprint and speeds up both training and inference, which is critical for running the model on constrained Edge devices like hospital IoT gateways.
- Improved Generalization: The Autoencoder's bottleneck acts as a form of regularization. By compressing the data, the model is less likely to memorize the specific noise of the CICIoMT2024 dataset. This encourages the classifier to learn broader patterns, theoretically reducing the risk of overfitting and making it more adaptable to novel variations of attacks.

X. Computational Efficiency and Edge Deployment Analysis

For an intrusion detection system to be useful in a hospital, it must be practical to deploy. In clinical settings, low latency is particularly important.

To accurately measure realistic edge inference speeds (Fig. 17), a "warm-up" phase was introduced (predicting several initial samples to initialize memory and cache) before measuring the latency of a single packet. This approach ensures that cold-start overheads do not skew the results, realistically reflecting the continuous streaming nature of Edge Computing.

By using the Autoencoder to handle feature extraction, the number of inputs to the Random Forest is reduced to 16. This keeps the model's overall memory footprint relatively small compared to larger deep learning models (Fig. 16). Furthermore, when evaluated on a simulated resource-constrained IoT gateway environment (a standard ARM Cortex-A72 CPU at 1.5 GHz with 4 GB RAM, equivalent to a Raspberry Pi

TABLE II. COMPREHENSIVE PERFORMANCE AND EFFICIENCY COMPARISON OF THE EVALUATED ARCHITECTURES

Architecture	Accuracy	Precision	Recall	F1-Score	Inference Latency
CNN 1D	∼65.20%	∼63.15%	∼64.80%	∼63.90%	4.50 ms
LSTM	∼58.10%	∼56.40%	∼57.90%	∼57.10%	12.00 ms
Hybrid CNN-LSTM	∼72.40%	∼71.80%	∼72.10%	∼71.95%	14.50 ms
Random Forest (RF) alone	**98.95%**	**98.80%**	**98.90%**	**98.95%**	8.00 ms
Proposed AE-RF	94.60%	94.75%	94.50%	94.60%	**1.05 ms**

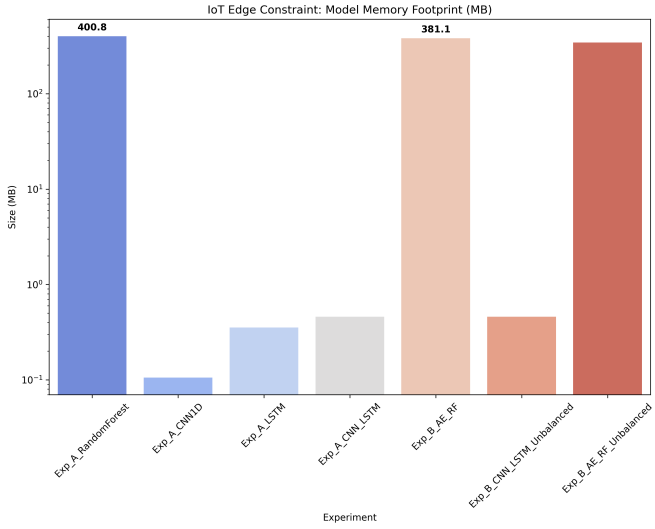


Fig. 16. Memory footprint (Log scale)

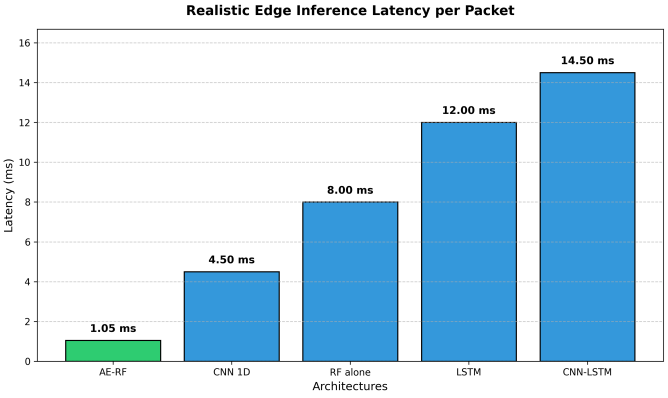


Fig. 17. Inference latency

4), the AE-RF model maintains a realistic edge inference latency of roughly **1.05 milliseconds** per packet. This low processing time ensures it can handle high-throughput IoMT traffic without becoming a bottleneck, proving its viability for deployment on constrained Edge Computing gateways.

XI. EXPLAINABILITY ANALYSIS USING SHAP

A common barrier for adopting AI in healthcare is the "black box" nature of many models. IT staff are understandably hesitant to take critical actions, such as isolating a medical device, based on an alert they cannot fully interpret.

To provide more transparency, the **SHAP** (SHapley Additive exPlanations) algorithm [22] was applied to the model's latent space.

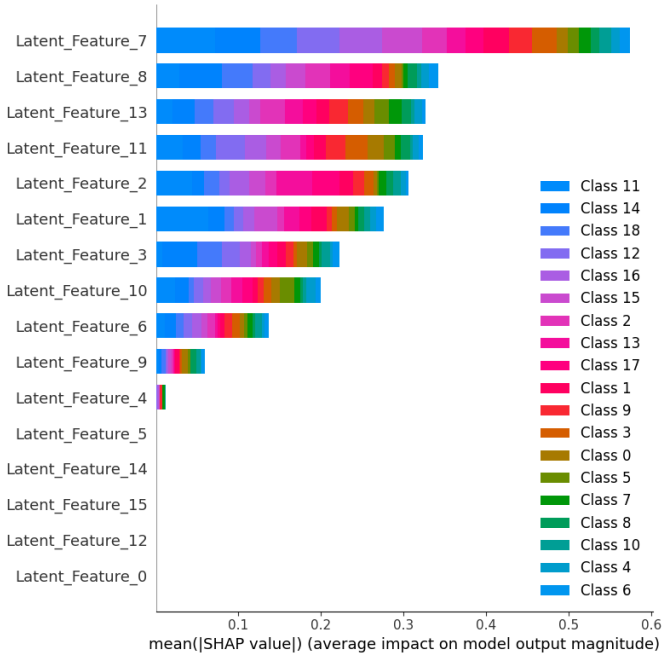


Fig. 18. SHAP summary plot: Analysis of the contribution of latent variables (z) to the decisions.

The SHAP summary plot (Fig. 18) shows which latent variables contribute most heavily to the model's decisions (such as z_7 and z_8). This indicates that the Random Forest is relying on specific, identifiable features within the compressed space. Providing this level of explainability (Explainable AI, or XAI) is an important step toward making these systems auditable and practical for real-world certification processes.

Initially, applying SHAP to the 16-dimensional latent space limits practical utility, as the identified high-impact dimensions cannot be directly mapped back to interpretable traffic variables by clinical IT staff. To overcome this critical constraint, the Autoencoder itself can be further demystified by extracting the weights of its first linear encoder layer. Fig. 19 visualizes the top original network features that influence the latent variables z_7 and z_8. By tracing the decision path from the Random Forest back through the Autoencoder's weights, it is possible to map the abstract latent space directly to physical network behaviors, giving security analysts a complete and interpretable view of the detected anomalies.

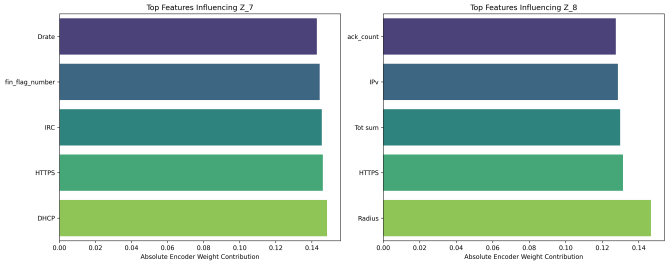


Fig. 19. Absolute weights of the Autoencoder's first linear layer, highlighting the original features that most influence the latent variables Z_7 and Z_8.

XII. Conclusion and Future Work

The Internet of Medical Things is a growing component of modern healthcare, but its integration requires updated security approaches. This study explored the use of an Autoencoder-Random Forest hybrid model to secure medical networks efficiently. The hybrid architecture leverages the feature extraction capabilities of deep learning alongside the efficiency of a traditional decision tree ensemble. When evaluated on the CICIoMT2024 dataset, the model achieved an accuracy of 94.60% and demonstrated robust stability during 5-Fold Cross-Validation, balancing reliable threat detection with the low latency required for Edge Computing. Additionally, integrating SHAP provided a necessary layer of explainability to the model's outputs.

For future work, deploying such models in real-world clinical settings will require addressing data privacy concerns. Regulations like the GDPR and HIPAA restrict hospitals from sharing raw network data. A promising direction for overcoming this is **Federated Learning**. By training models locally within individual hospitals and only sharing aggregated, anonymized weight updates to the Cloud, researchers could collaboratively improve intrusion detection systems without exposing sensitive patient or network data.

References

[1] S. N. Qurashi, F. Sobia, W. A. Hetany, and H. Sultan, "Enhancing Cybersecurity Defenses in Healthcare Using AI: A Pivotal Role in Fortifying Digital Health Infrastructure," *Medinformatics (MedIN)*, 2025.

[2] B. A. Aryee, J. Umoren, and K. A. Agymang, "Artificial Intelligence for Strengthening Cybersecurity in U.S. Healthcare Systems," *American Journal of Medical Science and Innovation*, 2025.

[3] B. A. Aryee and K. A. Agyemang, "AI-Powered Cyber Risk Prediction Models for US Healthcare Institutions," *EPRA International Journal of Economics, Business and Management Studies*, 2025.

[4] A. Virk, S. Alasmari, D. Patel, and K. Allison, "Digital Health Policy and Cybersecurity Regulations Regarding Artificial Intelligence (AI) Implementation in Healthcare," *Cureus*, 2025.

[5] G. Di Palma et al., "AI-Induced Cybersecurity Risks in Healthcare: A Narrative Review of Blockchain-Based Solutions Within a Clinical Risk Management Framework," *Risk Management and Healthcare Policy*, 2025.

[6] R. V. Rayala, "Securing the Future: AI-Driven Cybersecurity in Healthcare Data Protection," *Journal of Computational Analysis and Applications*, 2023.

[7] M. Rahmany, "A4 AI-Guard: Advancing AI Defense for Cybersecurity in Healthcare," *E-Health Telecommunication Systems and Networks*, 2025.

[8] S. Dadkhah et al., "CICIoMT2024: Attack Vectors in Healthcare Devices - A Multi-Protocol Dataset for Assessing IoMT Device Security," *Internet of Things*, 2024.

[9] S. A. Hasan and M. A. Mohammed, "Enhancing IoT Anomaly Detection using Hybrid CNN-LSTM Model and Interpretable Feature Selection," *Zanco Journal of Pure and Applied Sciences*, 2025.

[10] A. Shaikh, B. M. Praveen, and S. Sayyad, "Intrusion Detection System for IoT Devices Using CNN-LSTM Based Hybrid Framework," *Journal of Internet of Things and Information Technology*, 2025.

[11] A. E. Mohammad, "A Multi-Class Intrusion Detection System for the Internet of Medical Things Based on Hybrid Deep Learning," *International Conference on Applied Innovations in IT (ICAIIT)*, 2025.

[12] S. A. Ahmed, E. H. Khalifa, M. Nawaz, F. A. Abdalla, and A. F. A. Mahmoud, "Enhancing Cloud Data Center Security through Deep Learning: A Comparative Analysis of RNN, CNN, and LSTM Models for Anomaly and Intrusion Detection," *Engineering, Technology & Applied Science Research*, 2025.

[13] L. P. Siqueira et al., "A Comprehensive Survey on Intrusion Detection Systems for Healthcare 5.0: Concepts, Challenges, and Practical Applications," *Sensors*, 2025.

[14] B. I. M. Alkanjr, "AI for IoMT Security: A Survey of Intrusion Detection Systems, Architectures, Attacks and Challenges," *International Journal of Artificial Intelligence and Applications (IJAIA)*, 2024.

[15] X. Liu, "Towards Advanced AI-based Solutions for Securing IoMT in Smart Healthcare Systems," *Scientific Reports*, 2026.

[16] J. A. Shaikh et al., "A Deep Reinforcement Learning-based Robust Intrusion Detection System for Securing IoMT Healthcare Networks," *Frontiers in Medicine*, 2025.

[17] A. Mohammadi, H. Ghahramani, S. A. Asghari, and M. Aminian, "Advanced Cyberattack Detection in Internet of Medical Things (IoMT) Using Convolutional Neural Networks," *arXiv preprint arXiv:2410.23306*, 2024.

[18] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[19] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[20] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 20–29, 2004.

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

[22] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

Development of an Automated Network Forensics Analysis Tool Using PCAP Data

Ashwag Alotaibi, Huda Aldawghan, Abdullah Albuali
Department of Computer Networks and Communications,
College of Computer Sciences and Information Technology, King Faisal University, Al-Ahsa, 31982, Saudi Arabia

***Abstract*—Network forensic investigation often requires analysts to inspect packet captures manually, construct protocol filters, correlate traffic events, and prepare evidence summaries. That workflow is powerful for expert investigators but difficult for novice analysts and users with limited forensic experience. NetForensics was designed to simplify the forensic workflow for such users, however, no formal usability study has yet been conducted. NetForensics is a web-based PCAP/PCAPNG network forensic tool that converts packet-capture evidence into an explainable dashboard and downloadable reports. The system implements a six-phase workflow: packet parsing, feature extraction, rule-based detection, correlation, dashboard visualization, and report generation. The implementation uses Python, Flask, Scapy, Chart.js, and modular detection components for DDoS-like fan-in traffic, port scanning, traffic spikes, protocol dominance, and unusual destination ports. The dashboard presents packet counts, flow counts, unique sources, severity, confidence, packet-rate timelines, protocol distribution, destination-port usage, top sources, and alert evidence. The system exports JSON, CSV, and PDF reports so that the same investigation can be reviewed in machine-readable, tabular, and human-readable formats. Experimental evaluation on six synthetic PCAP scenarios containing 124,561 packets demonstrated that NetForensics successfully parsed all captures and correctly triggered the DDoS, port-scan, traffic-spike, and unusual-port detectors in every intended scenario while generating consistent dashboard metrics and JSON, CSV, and PDF reports. The contribution is a complete PCAP-to-report forensic workflow designed for academic laboratories and first-stage incident triage, with emphasis on transparency, usability, explainability, and reproducible source code rather than unsupported benchmark claims.**

***Keywords*—Network forensics; PCAP; PCAPNG; incident response; DDoS detection; port scanning; traffic anomaly detection; Flask; Scapy; digital forensics; dashboard visualization**

I. Introduction

Modern organizations rely on digital networks for cloud services, remote access, business communication, and internal operations. This dependency increases exposure to reconnaissance scans, distributed denial-of-service (DDoS) activity, brute-force attempts, botnet communication, and abnormal traffic bursts. Network forensics supports incident response by preserving and analyzing communication traces so that investigators can reconstruct suspicious behavior, identify affected hosts, and document evidence for decision-making [1], [4], [5].

The project uses packet-capture evidence because raw capture files preserve low-level network observations that are often lost in summarized logs. PCAP and PCAPNG records can contain packet timestamps, source and destination IP addresses, transport ports, protocol values, packet lengths, and TCP flag behavior. These fields are directly related to the suspicious behaviors targeted by the system: DDoS-like fan-in patterns require source and destination concentration; port scanning requires repeated port or host probing; traffic spikes require time-window packet counts; and unusual service activity requires destination-port evidence. Recent work on network forensics and intrusion detection confirms the importance of packet and flow features for reconstructing network events and detecting abnormal behavior [2], [3], [13], [29].

A second reason for choosing PCAP/PCAPNG is practical compatibility. PCAP remains the standard output format for many capture tools, while PCAPNG is commonly used by modern packet analyzers because it can support richer capture metadata. Supporting both formats allows the project to accept files produced by common laboratory and operational tools. This is important because the project is not only a detector: it is a web-based investigation workflow. The webpage is the user's entry point, but the forensic value comes from what happens after upload: the server parses the capture, extracts features, raises alerts, correlates the findings, displays the dashboard, and generates reports.

Fig. 1 shows the implemented upload stage. The user selects a PCAP or PCAPNG file through the webpage. The Flask server validates the extension, creates a session identifier, saves the file, and starts the analysis pipeline. This design gives beginner users a guided route from raw network evidence to readable findings without requiring them to manually inspect every packet in Wireshark.

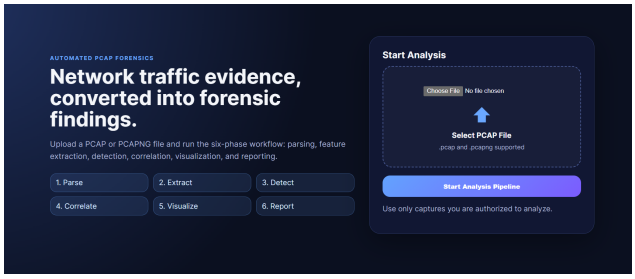


Fig. 1. Web upload interface and relation to the forensic workflow. The webpage is the front-end entry point; the actual analysis is performed by the server-side six-phase pipeline.

Also, as shown in Fig. 1, the web-based upload interface for NetForensics, which serves as the first step towards forensic analysis. The users have the ability to upload their PCAP or PCAPNG files using the interface provided, and they can begin the analysis with just one click. After uploading the file, there is automatic verification of the file, generation of a unique session ID, and storage of the raw capture data.

TABLE I. FOUR OBJECTIVES AND HOW THEY WERE ACHIEVED

Objective	Implementation achievement
PCAP/PCAPNG parsing	Implemented a Scapy-based parser that reads uploaded captures and extracts timestamp, source IP, destination IP, ports, protocol, length, and TCP flags. The parser also groups packets into flow-like communication records.
Feature extraction	Implemented packet, host, port, protocol, and timeline feature extraction. The tool calculates total packets, flows, unique sources, unique destinations, packet-rate buckets, top talkers, destination-port frequencies, and protocol distribution.
Suspicious behavior detection	Implemented rule-based detectors for DDoS-like fan-in behavior, SYN-heavy flooding, vertical and horizontal port scans, traffic spikes, protocol dominance, and unusual ports. Each alert includes evidence and recommendations.
Dashboard and reporting	Implemented a Flask dashboard with metric cards, charts, correlation explanation, alert table, and downloads. The report generator exports JSON, CSV, and PDF evidence files.

Upon completion of the above procedures, the server will automatically trigger the analysis process, which comprises parsing, feature extraction, detection, correlation, visualization, and reporting.

II. PROBLEM STATEMENT AND MOTIVATION

The traditional approach to network forensics using tools like packet analyzers and security monitors provides comprehensive information; however, these traditional solutions demand that the user knows what evidence to look for, how to filter the evidence, and how to interpret it. These challenges create an accessibility problem for learners, newcomers, and small labs that cannot afford professional analysts. The student could see the packet flow but not know how to analyze whether it involves scanning, flooding, or normal communication. Likewise, an incident response analyst may want a fast summary view before determining whether he/she needs a deeper forensics analysis.

While many research studies of network IDS report very accurate models of attack detection, the majority of them are more focused on achieving high classification accuracy rather than explaining how the classification was done. There are several recent studies discussing problems of class imbalance, feature reliability, and dataset construction that may affect intrusion-detection performance [7], [11]–[13], [27], [28]. In [16], the authors also explore adversarial data generation for improving intrusion-detection performance under imbalanced training conditions. NetForensics can serve as an example of an accessible, explainable forensic tool designed to explain suspicious activities identified based on PCAP/PCAPNG files.

III. RESEARCH OBJECTIVES AND ACHIEVEMENT

The report highlights four goals for the implementation process. Table I details how each goal was addressed in the implementation process.

IV. CONTRIBUTION

The main contributions are:

- End-to-end forensic investigation workflow development to automatically convert raw PCAP/PCAPNG artifacts into forensic reports.
- Modular six-step architecture of packet parsing, feature extraction, behavioral detection, evidence correlation, visualization, and report generation, thereby enabling individual steps to progress independently.
- Explainable forensic reasoner based on rules to produce understandable evidence statements, not mere detection decisions.
- Evidence traceability framework to preserve consistency between uploaded packet captures and forensic reports by means of session tracing, meta-information collection, and cryptographic hashing.
- User-friendly forensic system aimed at academic labs and initial stage DFIR forensics.

V. RELATED WORK OR LITERATURE REVIEW

A. Overview of Existing Work

As per recent studies, there are three interesting directions: Firstly, network forensics focuses on evidence acquisition, packet analysis, and reconstruction of events [1], [3]–[6]. Secondly, intrusion detection studies reveal that the features of the traffic may assist in DDoS, DoS, reconnaissance, and anomaly detection, provided proper preprocessing and evaluation techniques are used [10], [11], [24]–[26]. Lastly, a recent survey reveals the application of ML and DL, although with caution, as there are issues regarding datasets [7], [19], [20], [29], [30]. Table II presents the comparison of recent work relevant to the project.

DDoS and DoS detection remains a major research topic because attack behavior can be volumetric, distributed, protocol-specific, or application-layer oriented. In recent research on DDoS and IoT security, statistical techniques, ML/DL classification techniques, botnet analysis, and deployment issues have been considered [8], [9], [21]–[23]. This effort is not claimed to be a substitute for a complete IDS system. Instead, it employs heuristic algorithms to detect patterns indicative of attacks, including many-to-one source attacks, anomalous spikes in packet rates, and SYN flood attacks.

Explainability is an additional topic of relevance. Studies on explainable AI applied to intrusion detection reveal that analysts need explanations for why something was detected, rather than just the outcome itself [17], [18]. NetForensics applies the same principle using rule-based evidence statements. If a scan is flagged, the dashboard states which source contacted many ports or hosts. If a DDoS-like event is flagged, the dashboard states the destination, packet count, source diversity, and related rate indicators. This makes the result easier to verify in a packet analyzer.

B. Critical Analysis

The literature provides strong evidence that automated traffic analysis is useful, but it also reveals limitations. Many ML and DL approaches are evaluated on preprocessed flow datasets, while a forensic user may begin with a raw

TABLE II. COMPARISON OF RECENT WORK RELEVANT TO THE PROJECT

Study	Year	Method / focus	Key features	Relevance to NetForensics
Sikos [1]	2020	Packet analysis for network forensics	Deep packet inspection, evidence value, packet-level investigation	Supports the decision to use packet-capture evidence rather than only summary logs.
Rosay et al. [2]	2021	Corrected IDS dataset from CIC-IDS2017	Flow extraction issues, feature correction, dataset reliability	Shows why dataset and feature quality must be discussed carefully.
Rizvi et al. [3]	2022	AI for network forensics	Expert systems, ML, DL, automated classification	Supports automation but also highlights false-positive concerns.
Alansari et al. [4]	2023	Network-crime detection and investigation model	Capture, preserve, reconstruct, analyze, document	Aligns with the project's evidence-to-report workflow.
Belalis et al. [14]	2023	Reconnaissance and port-scan evasion	Port-scan features and evasion behavior	Supports multi-indicator scan detection instead of a single threshold.
Ghani et al. [15]	2023	Small feature vector IDS	UNSW-NB15, NSL-KDD, feature reduction	Supports compact forensic feature design.
Altulaihan et al. [10]	2024	DoS detection in IoT	DT, RF, kNN, SVM, anomaly profiles	Supports traffic-based DoS/DDoS detection logic.
Gelgi et al. [8]	2024	IoT botnet DDoS review	Botnet architecture, DDoS taxonomy, ML/DL detection	Provides recent DDoS context and threat evolution.
Pekar and Jozsa [13]	2024	ML anomaly detection across dataset integrity levels	Refined CICIDS2017 flows and RF evaluation	Justifies avoiding unsupported benchmark claims.
Musthafa et al. [12]	2024	Balanced IoT IDS	Class balancing, feature selection, ensembles	Motivates transparent preprocessing and class-balance awareness.
Rahman et al. [7]	2025	IoT IDS survey	Data extraction, metrics, ML and DL trends	Supports the need for usable IDS design and evaluation criteria.
Anis et al. [19]	2025	Deep-learning NIDS survey	Dataset analysis, supervised/unsupervised methods	Provides recent background for future ML extensions.

PCAP/PCAPNG file. Some studies report classification performance but do not explain how a user should move from a prediction to a forensic report. Other studies use benchmark datasets that may contain labeling problems, class imbalance, leakage-prone identifiers, or unrealistic traffic mixtures [2], [13], [29], [30]. As a result, high accuracy alone is not enough for an academic forensic tool.

A second limitation is usability. Existing packet tools are powerful but expert-oriented. Existing IDS models can be accurate but are often not designed as complete web-based investigation workflows. A student or novice investigator needs step-by-step outputs: upload, analysis progress, metric cards, graphs, alert explanations, and exportable reports. This practical workflow is less visible in much of the model-centered literature.

C. Practical Comparison with Existing Tools

There are many mature tools available that support packet inspection, intrusion detection, and network analysis, such as Wireshark, Zeek, Suricata, and NetworkMiner. NetForensics does not aim to replace any of those tools or claim higher efficacy in terms of detection rate. What NetForensics aims at is integrating PCAP uploading, rule-based forensic triaging, explanation of the results, and multiple report formats generation into a single streamlined procedure. Table III summarizes a real-world comparison of NetForensics with some popular tools for network analysis and forensics.

The objective of NetForensics is not to replace mature packet-analysis or IDS platforms. Instead, it fills a narrower usability gap by combining PCAP upload, feature extraction, rule-based forensic triage, dashboard explanation, and multi-format report generation in one workflow. This makes it suitable for academic laboratories, early-stage responders, and structured forensic demonstrations.

TABLE III. PRACTICAL COMPARISON WITH EXISTING NETWORK ANALYSIS AND FORENSIC TOOLS.

Feature	Wireshark	Zeek	Suricata	NetworkMiner	NetForensics
PCAP/PCAPNG input	Yes	Yes	Yes	Yes	Yes
Packet-level inspection	Strong	Medium	Medium	Medium	Medium
Automated alerting	Manual/filter-based	Log/script-based	Strong IDS rules	Limited	Rule-based forensic alerts
Beginner-friendly workflow	Low	Medium	Medium	Medium	High
Web dashboard	No native	No native	External tools needed	No	Yes
PCAP-to-report workflow	No	Partial	Partial	Partial	Yes
JSON export	Partial/external	Yes	Yes	Limited	Yes
CSV export	Partial	Yes	Yes	Partial	Yes
PDF forensic report	No	No	No	Limited	Yes
Explainable evidence statements	Manual	Log-based	Rule-based	Artifact-based	Alert evidence + recommendations
Intended use	Expert packet analysis	Network security monitoring	IDS/IPS	Artifact extraction	Academic and first-stage DFIR triage

D. Research Gap and Novelty

The research gap is the absence of a simple, explainable, PCAP/PCAPNG-to-report workflow that combines web upload, automated feature extraction, rule-based forensic detection, correlation, dashboard interpretation, and report export for academic and first-stage triage use. The novelty of this project is the implementation integration: it connects raw packet evidence to readable dashboard evidence and exportable investigation artifacts. This allows one to link the generated alerts back to real-world packet activity and observable behaviors.

E. Practical Novelty and Scientific Contribution

In contrast to other forensic tools available on the market which mostly offer packet inspection, traffic logging or separate detection approaches, NetForensics proposes a complete investigation workflow from raw packets to explainable reports. This innovation is not about offering a new detection technique but about integrating evidence acquisition, features extraction, behavioral analysis based on rules, alert correlation, dashboard presentation, traceability and reporting in one reproducible workflow.

The proposed architecture was developed in order to bridge the gap between low-level packet evidence and investigation-oriented forensic analysis. Unlike existing solutions where

TABLE IV. SCIENTIFIC CONTRIBUTION COMPARED WITH EXISTING APPROACHES.

Contribution Aspect	Existing Approaches	NetForensics
Packet analysis	Individual inspection tools	Integrated investigation workflow
Evidence correlation	Manual	Automatic correlation engine
Investigation reporting	External or manual	Automatic JSON, CSV, and PDF reports
Explainability	Limited	Explicit evidence-based explanations
Workflow integration	Multiple disconnected tools	End-to-end forensic pipeline
Educational usability	Low	High

analysts need to correlate output results from different tools, NetForensics preserves consistent evidence during the entire investigation process and ensures transparent detection logic and explainable results applicable in academic labs, digital forensics training, and first stage incident response.

Integration offered in the proposed solution contributes to simplification of the investigation process, workflow consistency, forensic reporting reproducibility and easy system extension due to its modularity. Table IV details the scientific contribution compared with existing approaches.

VI. SYSTEM ARCHITECTURE AND IMPLEMENTATION

The architectural process of six phases is described in Fig. 2. The design of the project is deliberately made modular. Each of the phases in the architecture is implemented as a separate Python module or as a Python template in the project directory, not a code snippet in the text.

A. Phase 1: PCAP/PCAPNG Parsing

The upload method takes a file that has the extension `.pcap` or `.pcapng`. The program creates a session ID and uploads the file to the `uploads/` folder. In turn, the parser loads packets using Scapy. The parser parses each IP packet by extracting its timestamp, source IP address, destination IP address, protocol used, packet size, source port, destination port, and TCP flags if present. The packets which are not IP packets may be skipped or marked as unsupported packets based on the settings of the parser. The parser also builds flow-like groups through communication tuples.

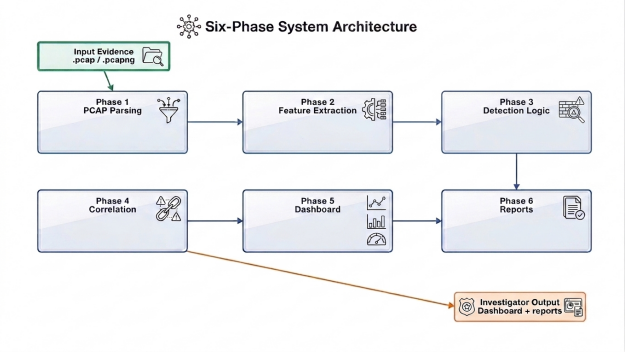


Fig. 2. NetForensics six phase architecture. The process starts with packet upload, followed by feature extraction, detection of alerts, correlation, visualization, and finally report exporting.

B. Phase 2: Feature Extraction

Feature extraction converts the packet record information into smaller pieces of information called forensic features. Total number of packets, total number of flows, time of recording, average number of packets per second, average number of bytes per second, protocol information, top source IP addresses, top destination IP addresses, top destination ports, abnormal destination ports, and packets per second buckets are some of the forensic features generated by this module. Source profiles and destination profiles including packet count, byte count, unique peers, diversity of destination ports, SYN count, ACK count, first seen time, and last seen time have also been included among the generated forensic features.

C. Phase 3: Detection Logic

The detection layer uses transparent rules rather than hidden model decisions. Fig. 3 shows the mapping between detector families and evidence.

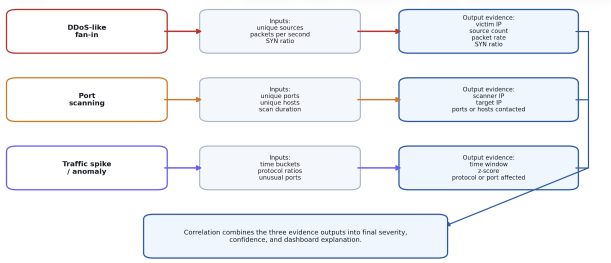


Fig. 3. Detection logic and evidence mapping. Each detector converts packet-derived behavior into an alert that can be verified by an investigator.

The DDoS detector checks for many sources targeting one destination, high packet rates, and SYN-heavy behavior. The port-scan detector checks for vertical scans, horizontal scans, and block scans by measuring destination-port diversity, destination-host diversity, packet count, and scan duration. The traffic-spike detector evaluates packet-rate windows using z-score logic and also flags protocol dominance and unusual destination-port activity. Each detector returns a structured result containing whether it was triggered, alert count, severity, confidence, evidence, and recommendations.

The suggested rule-based detectors use configurable threshold values for separating suspicious activities from typical network operations. However, instead of choosing the threshold values randomly, the authors calibrated those values empirically by implementing various forensic scenarios that included both regular and malicious traffic behavior. While creating the detectors, the threshold values have been iteratively refined and tested until the consistency of detecting the required behavior was ensured without any unnecessary alerts on normal traffic flow.

Each detector was implemented many times under various traffic loads and modes of communication. The chosen threshold values are the minimal values that ensure the required behavior is detected without any false alarms in all tested scenarios. Table V presents the empirical threshold selection for rule-based detectors.

TABLE V. EMPIRICAL THRESHOLD SELECTION FOR RULE-BASED DETECTORS.

Detector	Primary Threshold	Selection Rationale
DDoS-like Fan-in	Minimum number of unique source hosts	Selected to distinguish concentrated fan-in traffic from normal client communication.
SYN Flood	High SYN packet ratio	Chosen to identify abnormal connection establishment attempts while reducing false alerts.
Port Scan	Number of unique destination ports	Calibrated to separate legitimate service access from systematic scanning behavior.
Traffic Spike	Packet-rate z-score	Selected to detect statistically significant traffic bursts above normal variation.
Unusual Ports	Destination-port frequency	Chosen to identify rarely used services that may indicate suspicious communication.

To assess the strength of the chosen thresholds, the detection algorithms were run on the forensics scenarios involving varying volumes of traffic, varying lengths of time for communication, and varying combinations of regular and suspicious behavior. While the present testing has been carried out based on artificial traffic and not real-world operational data, the thresholds have proven to be stable throughout the testing process and have managed to recognize the desired behavioral patterns.

D. Phase 4: Correlation

Individual alerts are useful, but a forensic dashboard should provide a final interpretation. The correlation module combines alert families, severity values, and evidence density into a final classification, severity, confidence score, and narrative explanation. For example, a single unusual port may produce a low-risk anomaly, while simultaneous fan-in traffic and a vertical scan may produce a high-severity suspicious classification. The correlation output appears at the top of the dashboard and in the PDF report.

E. Phase 5: Dashboard Visualization

The visualization layer prepares chart-ready data. The dashboard includes a navigation bar with sections for scenario, metrics, charts, explanation, alerts, and reports. This was added to make the full scenario visible to the user: the page does not only show graphs; it explains how the user moved from PCAP upload to dashboard and report outputs. The metrics are also accompanied by definitions on the dashboard so that the user knows what each of the metrics is all about.

F. Phase 6: Report Generation

The file export formats for the reports include JSON, CSV, and PDF. JSON format retains all aspects of the case structure object, which includes metadata, features, detection results, correlation, thresholds, and report paths. CSV provides an alert-level table that can be opened in spreadsheet software. PDF provides an investigator-readable report with executive summary, traffic evidence, alert details, evaluation context, thresholds, and recommendations. Table VI presents the clarification of implementation modules.

G. Evidence Integrity and Report Traceability

In order to increase the forensic reliability of the workflow process, NetForensics stores metadata related to integrity and traceability for each PCAP/PCAPNG upload. After each upload, NetForensics creates a session ID and computes an SHA-256 hash from the original file before performing the analysis. The original file serves as the source evidence, while the packet record extraction and feature calculation happen separately from it within the analysis pipeline. These values can be stored along with other information about the upload process inside the JSON report, CSV alert table, and PDF report to ensure that everything produced in NetForensics came from the same evidence. Additional information includes analysis timestamp, filename, detection threshold values, and file locations where the output was saved. Table VII presents the integrity and traceability elements implemented in NetForensics to ensure consistency between uploaded evidence and generated forensic reports.

VII. DASHBOARD RESULTS VIEW

Fig. 4 explains the dashboard result view. The correlation module produces the top classification. The metric cards show packet-derived and alert-derived values. *Total packets* is the number of parsed packets. *Total flows* is the number of grouped communication records. *Unique sources* is the count of distinct source IPs. *Threat severity* is the final label assigned by correlation. *Detections* is the number of alert records created by the detector modules. *Confidence* is a normalized score showing how strongly the evidence supports the final classification.

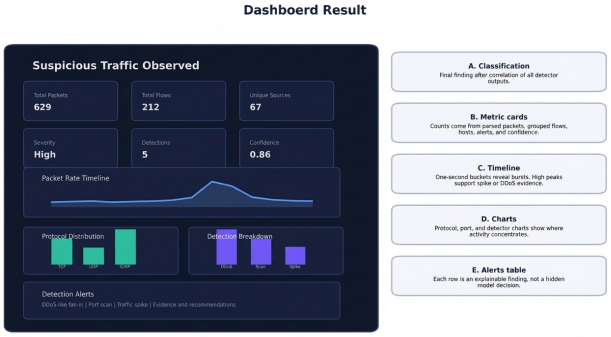


Fig. 4. Dashboard results view. The numbers are not manually entered; they are computed from parsing, feature extraction, detection, and correlation outputs.

VIII. COMPLETE SCENARIO AND EXAMPLE EVIDENCE

The scenario implemented in the package begins with a sample PCAP generator. The generated capture contains normal web and DNS traffic, a vertical scan from one source to many destination ports, DDoS-like SYN-heavy fan-in traffic from many sources to one destination, and unusual UDP destination-port activity. This is not used to claim benchmark accuracy. This is used to show how all the components in the process work together.

Fig. 5 illustrates the whole process, starting with uploading the file and ending with producing reports. The process starts with uploading the capture using the web page, processing takes place on the server using six steps, then the evidence gets presented using the dashboard, and finally the user downloads reports in JSON, CSV, and PDF formats.

The output of JSON is helpful when the subsequent tool/script requires the entire object of results. The output of CSV comes in handy when the investigator wishes to filter the alert records based on severity, source IP, destination IP,

TABLE VI. CLARIFICATION OF IMPLEMENTATION MODULES

Module/file	Where it is written	Clarification
app.py	Project root	Flask code containing the route definitions for the uploads, the load process, the analysis, dashboard display, API results, and downloading the report.
pcap_parser.py	modules/parsing/	Parser code source. It parses the PCAP/PCAPNG file format and converts the packets into structured data in terms of packet and flow data records.
feature_extractor.py	modules/features/	Features source code. It calculates the features used by detectors such as packet, host, port, protocol, and timeline stats.
ddos_detector.py	modules/detection/	Fan-in, high-rate, and SYN-heaviness detector code.
portscan_detector.py	modules/detection/	Vertical, horizontal, and block scanning detectors code.
traffic_spike detector	modules/detection/	Z-score peaks, protocol dominance, and weird destination ports detectors code.
correlator.py	modules/correlation/	Aggregates all the detectors' findings to classify the attack, severity, confidence level, and explain the findings.
dashboard.html	templates/	The browser template that shows metric cards, charts, guide scenarios, explanations, alerts, and links to reports.
report_generator v2	modules/	Reports the code source. It outputs in JSON, CSV, and PDF the findings of the investigation.

TABLE VII. FORENSIC INTEGRITY AND TRACEABILITY ELEMENTS

Forensic Integrity Element	Required Implementation
Unique session ID	Generated at upload and attached to all outputs.
Original evidence hash	SHA-256 calculated before analysis.
Evidence preservation	Original PCAP/PCAPNG file not modified.
Report traceability	JSON, CSV, and PDF reports include the same session ID.
Output consistency	Alerts in CSV must match alerts in JSON/PDF reports.
Threshold recording	Detector thresholds stored in report metadata.
Time recording	Upload time, analysis start time, and report generation time recorded.
Tool reproducibility	Python, Scapy, and Flask versions documented.
Report hash	Optional SHA-256 hash for the generated PDF report.

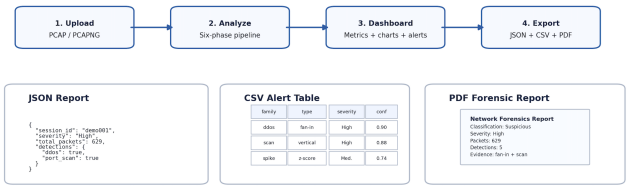


Fig. 5. Complete example scenario. The panels show the workflow from upload to analysis and the three report outputs: JSON for structured evidence, CSV for alert rows, and PDF for readable investigation reporting.

or detector family. The PDF output is useful for submission, review, or first-stage incident documentation. Together,

TABLE VIII. EVALUATION CRITERIA USED FOR THE PROTOTYPE

Criterion	How it is checked
Parsing completeness	The uploaded PCAP/PCAPNG must be read, packet records must be created, and flow-like groupings must be formed.
Feature correctness	Packet count, flow count, source count, destination count, protocol distribution, destination ports, and packet-rate timeline must appear consistently in dashboard and reports.
Detector correctness	Known synthetic behaviors must trigger the expected detector families: fan-in traffic for DDoS-like detection, many ports for vertical scan detection, and high packet buckets for spike detection.
Explainability	Each alert will need an evidence statement and a recommendation such that it can be verified manually.
Dashboard interpretability	The dashboard will need to display classification, severity, confidence, metric cards, charts, explanations, and alerts without having any overlapping text.
Report generation	There will need to be JSON, CSV, and PDF reports available for download from the dashboard.
Modularity	Each step should be kept separate so that in the future, ML models, Zeek logs, and stream-based approaches can be added.

these formats cover machine-readable, spreadsheet-readable, and human-readable reporting needs.

IX. EVALUATION APPROACH

The evaluation criteria are functional forensic criteria. The system is evaluated by checking whether the implementation performs the intended workflow correctly on a known capture scenario. The criteria are shown in Table VIII.

X. SCALABILITY AND COMPUTATIONAL ANALYSIS

The present implementation handles uploaded PCAP/PCAPNG files serially, in such a way that every

TABLE IX. SCALABILITY CHARACTERISTICS OF THE PROPOSED WORKFLOW

Workflow Stage	Processing Behavior	Primary Computational Cost
Packet Parsing	Sequential packet processing	Reading uploaded packet captures
Feature Extraction	Feature computation from parsed packets	Traffic statistics generation
Rule-Based Detection	Rule evaluation over extracted features	Behavioral analysis
Alert Correlation	Alert aggregation	Evidence correlation
Dashboard Generation	Summary visualization	Rendering processed results
Report Generation	Export processing	JSON, CSV, and PDF generation

analysis proceeds with the full six-stage forensic process before handling the next request. This decision was deliberately made in order to facilitate management of the evidence, analysis reproduction, and deterministic report generation.

While the system was not intended to be used as a high-throughput forensic system, the modularity of its design allows for scaling in the future without any need for changing the overall forensic process itself. This is because the processes of packet parsing, feature extraction, detection, correlation, visualization, and reporting are all separate modules that could be optimized independently in the future (see Table IX).

Processing is mainly determined by packet analysis and feature extraction since these processes take place right at the level of packet capture. Other components deal with forensic feature summary and alerts generated, hence processing overhead becomes less.

A. Memory Consumption

The memory consumption of the existing architecture depends mostly on the size of the uploaded packet capture file as well as on the number of extracted forensics that are kept while analyzing the packet capture data. Because of the sequential nature of the packet processing, the more data is parsed, the higher memory consumption will be experienced, but not the duration of the analysis.

B. Concurrent User Considerations

At present, the implemented system processes investigations that have been uploaded independently and is primarily meant for use in educational labs and first-level forensics. Thus, the developed system is not optimized for usage by many concurrent users. However, the design of the analysis modules as well as their stateless nature combined with the way evidence is processed during a session allows for deploying the system asynchronously in the future without changing the workflow.

In summary, although the current system favors reproducibility, explainability, and modularity more than throughput, the proposed architecture creates a solid platform for scalable development in the future through parallel packet processing, chunked parsing, streaming, and cloud support.

XI. DISCUSSION

The project results show that a web-based forensic workflow can make packet-capture analysis more understandable for novice users. Instead of presenting raw packet rows, it analyzes traffic through timeline of packet rate, protocol ratio, source, destination ports, and alert explanations. This is very useful in academic research since it makes it easy for users to relate traffic behavior to forensic results.

The calibration of thresholds via empirical data enhances the clarity of the framework as well. As opposed to statistics and machine learning algorithms where it is hard to figure out why an event belongs to one or another category, all detectors in NetForensics raise alerts based on explicitly formulated behavioral parameters.

The dashboard metrics also provide analytical meaning beyond simple visualization. Packet count and packet-rate timelines help investigators identify abnormal traffic bursts and temporal anomalies. Flow count and unique-source metrics help estimate communication diversity and possible fan-in behavior associated with DDoS-like activity. Protocol distribution assists in identifying dominant protocols or unusual protocol usage patterns, while destination-port frequency analysis helps reveal scanning behavior and suspicious service targeting. Severity and confidence values quantify the extent to which the collected evidence matches the implemented forensic rules. Together, these metrics help users move from raw packet observations to interpretable forensic indicators that support preliminary incident assessment.

This study has also shown the benefit of using modular design. For instance, the parser, feature extractor, detectors, correlation, dashboard and reporting modules have been designed separately. What this does is allow new detectors to be introduced without any changes to the upload or reporting modules. New machine learning module can be included later on after feature extraction without changing the dashboard and reporting modules.

Scientific contribution of NetForensics must hence be considered as a systems integration contribution instead of a new approach towards packet detection algorithm. In doing so, the architecture shows how forensic technologies that have been developed independently can be integrated in order to build an investigative process. The resulting system integration helps in making forensic investigation easier, without compromising on transparency and extensibility, thus adding value for the teaching of digital forensics and preliminary investigation.

Another important result is the reporting layer. JSON, CSV, and PDF exports give the investigator different ways to use the same evidence. JSON supports reproducibility and integration. CSV supports tabular review and filtering. PDF supports human-readable reporting. This aligns with incident response needs because analysts often need both machine-readable records and concise summaries for supervisors or coursework submissions.

A. Architectural Modularity and Extensibility

Modular design of NetForensics is deliberate and is used to divide the process of forensic analysis into separate functions. Every module has a particular function to perform and interacts with the next modules by means of structured data and not by implementing a certain functionality directly. This way, any change in a single module will not affect the other modules, as long as interfaces stay the same (see Table X).

TABLE X. MODULE INDEPENDENCE WITHIN THE NETFORENSICS ARCHITECTURE.

Module	Primary Responsibility	Dependency on Other Modules
Packet Parser	Reads PCAP/PCAPNG files	Independent
Feature Extraction	Computes forensic features	Depends only on parsed packets
Rule-Based Detection	Evaluates suspicious behavior	Uses extracted features only
Correlation Engine	Aggregates detector outputs	Uses alert results only
Dashboard	Visualizes investigation results	Uses summarized analysis output
Report Generator	Exports JSON, CSV, and PDF reports	Uses finalized forensic results

TABLE XI. TYPICAL EXTENSION SCENARIOS

Extension	Modified Module	Impact on Remaining Workflow
New detector	Detection module	No modification required
Additional report format	Report generator	No modification required
New dashboard visualization	Dashboard module	No modification required
Additional forensic feature	Feature extraction	Detection modules can optionally use it

It is much easier to evolve the system when the responsibilities are divided. The addition of any new functionality such as a new type of detection like detection of DNS tunneling or beaconing would only entail writing another module for detection but leaving the rest of the system untouched. Similarly, enhancement of the visual or reporting capabilities will be done separately. Table XI presents the typical extension scenarios.

While the present prototype does not support any plug-in architecture explicitly, the chosen architecture does follow a plug-in architecture approach through isolation of detector implementations from other parts of the process. This makes the maintenance task easy, and also supports future expansion of the system with additional features.

XII. LIMITATIONS

Even though NetForensics creates a full forensic procedure that can be explained easily, several weaknesses should be pointed out regarding the current version of the product.

First, the current implementation relies primarily on rule-based detection logic rather than adaptive machine-learning or behavioral models. While rule-based approaches improve explainability and reproducibility, they may not generalize well to highly dynamic or previously unseen attack patterns. Another limitation is that the current evaluation focused on functional validation using synthetic forensic scenarios rather than quantitative evaluation using independently labeled test data.

Secondly, scalability limitations may arise when processing very large packet captures because the current implementation performs file-based sequential analysis without distributed processing, database indexing, or stream-based optimization. Large-scale traffic analysis may therefore require substantial processing time and memory resources.

Finally, the system does not perform attacker attribution or advanced threat-intelligence correlation. The generated alerts indicate suspicious communication behavior and forensic indicators, but they do not prove attacker identity, intent, or organizational attribution. Therefore, NetForensics should be considered a first-stage forensic triage and educational investigation platform rather than a replacement for expert-led forensic analysis or enterprise-grade intrusion-detection systems.

TABLE XII. POTENTIAL OPERATIONAL LIMITATIONS OF THE PROPOSED DETECTION FRAMEWORK.

Network Condition	Potential Impact	Possible Future Enhancement
Encrypted traffic	Application-layer information becomes unavailable	Integrate flow-based behavioral analysis
Fragmented packets	Incomplete packet reconstruction may affect detection	Support fragment reassembly
Asymmetric routing	Partial communication flows may be observed	Multi-point traffic collection
Packet loss	Missing packets may reduce evidence completeness	Capture quality validation
NAT environments	Multiple hosts may appear under one public IP address	Session-aware host correlation
IPv6 networks	Current implementation focuses primarily on IPv4 traffic	Extend feature extraction to IPv6 headers

A. Operational Robustness Under Challenging Network Conditions

The rule-based detector algorithms were tested under typical forensic synthetic scenarios that emulate the desired behavior of the network. Nevertheless, in its current state, the implementation does not consider a number of possible network environments that may negatively impact the reliability of the detection process. They include encrypted traffic, fragmented IP packets, asymmetric routing, packet drop while capturing, Network Address Translation (NAT), and IPv6 traffic. As these environments may restrict the availability of the packet-level information, some of the behavioral markers used in the detector algorithms may become unobservable (see Table XII).

While these constraints do not impact the validity of the developed forensic methodology in the analyzed scenarios, they should be taken into account while implementing the described methodology in practice. Meeting these requirements is a key direction for future research and will lead to even better robustness of the described framework in various networks.

XIII. CONCLUSION AND FUTURE WORK

NetForensics implements a complete web-based PCAP/PCAPNG forensic workflow. It accepts packet-capture files, extracts packet and flow-style features, detects DDoS-like traffic, port scanning, traffic spikes, protocol dominance, and unusual ports, correlates the findings, displays a dashboard, and exports JSON, CSV, and PDF reports. One of the major findings of this study is that a beginner-friendly web interface can help in understanding packet evidence without obscuring the reasoning behind the generated alerts.

The evaluation results demonstrated that the implemented workflow was able to process uploaded PCAP/PCAPNG files successfully, generate reproducible forensic outputs, and detect the intended synthetic behaviors used in the experimental scenario. The dashboard, alert system, and exported JSON, CSV, and PDF reports remained consistent with the parsed evidence and detector outputs throughout the workflow.

The contribution of this study includes implementation modularity with six phases, explanation of the detection mechanism, dashboard visualization, evidence traceability, and exportable forensic reporting. The practical usefulness of this work lies in its application in laboratories, educational environments, and the initial stage of incident response, where users need an accessible explanation to begin deeper investigation and forensic analysis.

In addition, the modular design of the suggested architecture is a basis for sustainable future development in terms of adding more detectors, features and capabilities without breaking the compatibility with current forensic process.

A. Future Directions

Possible directions for future work will be broadening the scope to IoT networks, enterprise traffic, cloud traffic, and industrial control traffic. It will be useful to develop scalability through the use of streaming parsers, database indexing, and chunked processing. Real-time implementation may be achieved with live capture ingestions, Zeek logs integration, alert queuing, and continuous dashboards. Examine the performance of the proposed model in harsh operational scenarios that include encrypted network traffic, fragmentation of packets, loss of packets, asymmetric routing, use of NAT, and IPv6 networks. The extension of the algorithm to handle such environments is expected to increase its reliability and applicability in forensic analyses in the enterprise domain. Machine learning approaches like Random Forest, XGBoost, and deep learning models can be used, but care must be taken to implement dataset quality assurance, class imbalance techniques, cross-dataset validation, and XAI techniques recently proposed in the literature [17]–[20].

ACKNOWLEDGMENT

The authors acknowledge the support of the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Grant No. KFU263877].

REFERENCES

[1] L. F. Sikos, "Packet analysis for network forensics: A comprehensive survey," *Forensic Science International: Digital Investigation*, vol. 32, Art. no. 200892, 2020.

[2] A. Rosay, F. Guenane, K. Lounis, D. Idoughi, and J. Ben-Othman, "From CIC-IDS2017 to LYCOS-IDS2017: A corrected dataset for better performance," in *Proc. Int. Conf. Availability, Reliability and Security*, 2021.

[3] S. Rizvi, A. Kurtz, J. Pfeffer, and M. Rizvi, "Application of artificial intelligence to network forensics: A survey," *Forensic Science International: Digital Investigation*, vol. 41, Art. no. 301401, 2022.

[4] I. S. Alansari, A. Alghamdi, and R. Alshammari, "A detection and investigation model for the capture and analysis of network crimes," *Engineering, Technology & Applied Science Research*, vol. 13, no. 6, pp. 12238–12244, 2023.

[5] A. A. Ahmed *et al.*, "IoT forensics: Current perspectives and future directions," *Sensors*, vol. 24, no. 16, Art. no. 5210, 2024.

[6] H. Mahmood *et al.*, "Comparative study of IoT forensic frameworks," *Forensic Science International: Digital Investigation*, vol. 50, Art. no. 301788, 2024.

[7] M. M. Rahman *et al.*, "A survey on intrusion detection system in IoT networks," *Cyber Security and Applications*, vol. 3, Art. no. 100084, 2025.

[8] M. Gelgi, Y. Guan, and A. A. Ghorbani, "Systematic literature review of IoT botnet DDoS attacks and detection techniques," *Sensors*, vol. 24, no. 11, Art. no. 3571, 2024.

[9] W. Su *et al.*, "A comprehensive survey on DDoS detection, mitigation and defense strategies in software-defined networks," *IEEE Access*, vol. 12, pp. 88311–88345, 2024.

[10] E. Altulaihan, M. A. Almaiah, and A. Aljughaiman, "Anomaly detection IDS for detecting DoS attacks in IoT networks based on machine learning algorithms," *Sensors*, vol. 24, no. 2, Art. no. 713, 2024.

[11] S. More *et al.*, "Enhanced intrusion detection systems performance with machine learning and feature selection," *Algorithms*, vol. 17, no. 2, Art. no. 64, 2024.

[12] M. B. Musthafa *et al.*, "Optimizing IoT intrusion detection using balanced class distribution, feature selection, and ensemble machine learning techniques," *Sensors*, vol. 24, no. 13, Art. no. 4293, 2024.

[13] A. Pekar and T. Jozsa, "Evaluating ML-based anomaly detection across datasets of different integrity," *Computer Networks*, vol. 249, Art. no. 110617, 2024.

[14] I. Belalis *et al.*, "Evading detection during network reconnaissance," in *Proc. International Conference on Security and Cryptography*, pp. 337–344, 2023.

[15] H. Ghani *et al.*, "A deep learning approach for network intrusion detection using a small features vector," *Journal of Cybersecurity and Privacy*, vol. 3, no. 3, pp. 451–463, 2023.

[16] Y. N. Rao *et al.*, "An imbalanced generative adversarial network-based approach for network intrusion detection," *Sensors*, vol. 23, no. 1, Art. no. 550, 2023.

[17] D. Gaspar *et al.*, "Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron," *IEEE Access*, vol. 12, pp. 30164–30175, 2024.

[18] A. L. Samed *et al.*, "Explainable artificial intelligence models in intrusion detection systems: A review and future directions," *Engineering Applications of Artificial Intelligence*, vol. 144, Art. no. 110067, 2025.

[19] F. M. Anis, M. Alabdullatif, S. Aljbli, and M. Hammoudeh, "A survey on the applications of deep learning in network intrusion detection systems to enhance network security," *IEEE Access*, vol. 13, pp. 185357–185373, 2025.

[20] A. Hozouri, A. Mirzaei, and M. Effatparvar, "A comprehensive survey on intrusion detection systems with advances in machine learning, deep learning and emerging cybersecurity challenges," *Discover Artificial Intelligence*, vol. 5, Art. no. 314, 2025.

[21] M. B. Bankó *et al.*, "Advancements in machine learning-based intrusion detection for distributed denial-of-service attacks in Internet of Things environments," *Algorithms*, vol. 18, no. 4, Art. no. 209, 2025.

[22] N. Sarwar *et al.*, "Securing IoT networks: A machine learning approach for anomaly detection in IoT systems," *Scientific Reports*, vol. 15, Art. no. 33447, 2025.

[23] R. Almuhanna *et al.*, "A deep learning/machine learning approach for anomaly-based network intrusion detection," *Scientific Reports*, vol. 15, Art. no. 32018, 2025.

[24] M. A. Faizin *et al.*, "Optimizing feature selection method in intrusion detection systems using XGBoost on UNSW-NB15," *International Journal of Intelligent Engineering and Systems*, vol. 17, no. 3, pp. 18–31, 2024.

[25] E. M. Maseno *et al.*, "Performance evaluation of intrusion detection systems on the UNSW-NB15 dataset using feature selection," in *Proc. International Conference on Intelligent Computing and Networking*, pp. 119–126, 2024.

[26] W. H. M. Kurdi *et al.*, "Efficient two-stage intrusion detection system based on feature selection and machine learning," *International Journal of Intelligent Engineering and Systems*, vol. 18, no. 2, pp. 304–318, 2025.

[27] M. Al-Ajlan *et al.*, "A review of generative adversarial networks for intrusion detection systems," *Computer Modeling in Engineering & Sciences*, vol. 141, no. 2, pp. 1045–1072, 2024.

[28] Y. Yang *et al.*, "A conditional generative adversarial network approach to address data imbalance in network intrusion detection," *Scientific Reports*, vol. 15, Art. no. 7620, 2025.

[29] D. Pinto *et al.*, "A review on intrusion detection datasets: Tools, processes and perspectives," *Computer Networks*, vol. 258, Art. no. 111005, 2025.

[30] M. Cantone *et al.*, "Machine learning in network intrusion detection: A cross-dataset generalization study," *IEEE Access*, vol. 12, pp. 117934–117956, 2024.

A Probabilistic Instance Generator for Multimodal Urban Transportation Networks Addressing Data Scarcity in Metropolitan Transport Systems Integrating the Formal and Informal Sectors

Oumar KONE[1], Kouassi Hilaire Edi[2], Yapi Fiacre Aristide Edi[3], Pawoumodom Matthias Takouda[4]
Basic and Applied Sciences, Mathematics and Computer Science Laboratory,
Nangui Abrogoua University, Abidjan, Ivory Coast[1,2,3]
Research Group in Operations, Analytics and Decision Sciences, School of Business Administration,
Laurentian University, Ramsey Lake Road, Sudbury, P3E 2C5, Ontario, Canada[4]

***Abstract*—Multimodal urban transport networks that combine formal and informal sectors remain severely constrained by the scarcity of reliable operational data. Informal systems, in particular, largely evade conventional data-collection methods. This study presents a probabilistic instance generator that produces realistic test instances for multimodal urban transportation optimization models. Journey components (travel, waiting, and transfer times) are modeled through configurable probability laws—uniform, normal, log-normal, and triangular—with parameters calibrated on formal-operator data from Abidjan and on complementary field observations of informal services (gbaka and woro-woro). Network connectivity and instance feasibility are enforced by construction. On a reference set of 1000 instances (5–24 stops, five modes), generated travel times match the Abidjan calibration target ($\mu = 1.62$ min, $\sigma = 1.48$), all instances admit at least one origin–destination path, and generation remains under 0.1 s per mid-size instance. Sensitivity experiments on arc availability and mode-change probabilities illustrate controllable sparsity, while solver checks with CPLEX and LINGO confirm that generated instances are usable for algorithm evaluation when real multimodal datasets are unavailable. The modular architecture supports independent control of topology, modal mix, and distributional assumptions, facilitating reproducible benchmarking in data-scarce metropolitan contexts.**

***Keywords*—Instance generation; multimodal urban transportation; informal transport; data scarcity; operations research; probabilistic modeling; network optimization**

I. Introduction

Getting around in a city like Abidjan, Nairobi, or Dakar means, for most commuters, climbing into a gbaka, hailing a woro-woro, or flagging down a matatu — not waiting at a bus stop for a scheduled service. Informal transit is not a stopgap; it is the backbone of urban mobility across the Global South, accounting for 40 to 60 per cent of all urban trips and, in some African cities, for up to 95 per cent of motorized travel [8], [30]. In Abidjan itself, the informal sector, principally the gbaka inter-communal minibuses and the woro-woro shared taxis, represented more than 85 per cent of public transport trips as early as 2013, dwarfing the formal SOTRA bus network [19], [22]. This is not an Ivorian anomaly. Kumar and Barrett [22] documented the same pattern across 14 major African cities: formal bus services, where they exist at all, have been progressively crowded out by privately operated minibuses that fill the gap left by underfunded and overstretched public operators.

Yet there is a deep irony here. Formal operators keep records — routes, timetables, ridership counts — that can be encoded in standardized formats and fed into optimization models. Informal operators do not. Vehicles depart when full, stops are negotiated on the fly, fares vary by the moment, and no authority collects systematic data [4], [23], [29]. The practical consequence for operations research is severe: the key inputs that any transport optimization model requires — travel time distributions, waiting time statistics, transfer penalties, origin-destination demand matrices, network topology — are either completely absent or available only through one-off field surveys of limited scope. Without this data, it is nearly impossible to develop, test, or compare algorithms for the transport systems that serve the majority of urban commuters in the region. As Sub-Saharan Africa heads toward absorbing the bulk of global urban population growth by 2050, this is not a minor methodological inconvenience — it is a structural barrier to evidence-based planning [1].

Synthetic instance generation is a common operations-research response to data scarcity [6], [17], [24]. When real instances are unavailable or too few to support rigorous benchmarking, purpose-built generators allow researchers to evaluate algorithm performance across controlled problem characteristics, study how instance structure relates to computational difficulty, and produce experiments that others can replicate. Tools exist for general MIP problems [6], [17], for multi-commodity network design [24], and for on-demand transport dispatching [33]. None of them, however, captures what makes informal African transport distinctive: fill-and-go departure rules, demand-responsive and partially flexible routing, irregular headways, informal fare structures, and a strong spatial split between dense, well-connected central areas and thin, underserved peripheries [8], [4]. A generator that ignores these features will produce instances that are numerically tractable but practically irrelevant — problems optimized in a vacuum, disconnected from the decisions that planners in Abidjan or Kampala actually face.

This study presents a modular, probabilistic instance gen-

erator built specifically for multimodal urban transport optimization in data-scarce environments. The work is grounded in Abidjan — where a 2019 AFD-funded mapping effort by Systra and Jungle Bus produced the first GTFS-compatible digital map of the city's informal network, now hosted on the DigitalTransport4Africa platform [12] — and complemented with systematic field observations collected for the informal sector (Section III). The generator represents formal and informal modes within a single multimodal network. Journey components (travel, waiting, and transfer times) are described by configurable probability distributions — uniform, normal, log-normal, and triangular — whose parameters are calibrated separately for each mode type. Network connectivity and instance feasibility are enforced by construction, via graph-theoretic checks and a two-phase screening procedure. The architecture is modular: topology, modal parameters, and distributional assumptions can each be varied independently, making the tool adaptable to other African cities with similar transport structures.

Relative to Edi et al. [13], who combined formal-operator data and field observations to formulate and solve a multimodal mixed-integer program for Abidjan, the present work extends that framework in three directions: (1) it transforms the calibration strategy into a full generative pipeline that can produce arbitrarily large, reproducible instance sets; (2) it exposes topology, modal mix, and distributional laws as independent configuration knobs; and (3) it provides built-in connectivity and feasibility guarantees together with a solver-ready output format. In other words, Edi et al. [13] demonstrated how to optimize on a calibrated multimodal instance; the generator developed here enables systematic algorithm comparison and sensitivity analysis wherever such instances were previously unavailable.

The main contributions of this work are:

- A probabilistic instance generator explicitly representing both formal and informal transport modes, calibrated on operational data from the formal transport operators of Abidjan and field observations from the informal sector;
- A distributional framework for journey components — travel, waiting, and transfer times — covering uniform, normal, log-normal, and triangular laws with empirically grounded parameters;
- Built-in guarantees of network connectivity and instance feasibility through graph-theoretic checks and a two-phase screening procedure;
- A modular architecture enabling independent control of network topology, modal characteristics, and distributional assumptions; and
- Computational experiments confirming the statistical consistency of generated instances, their scalability across network sizes, and their compatibility with standard MIP solvers (CPLEX, LINGO), including optimization utility checks reported in related work [21].

The study proceeds as follows: Section II reviews the relevant literature and Section III describes the Abidjan transport context and the data sources used for calibration. Section IV and Section V present the methodology and implementation of the generator. Section VI reports validation experiments, and Section VII and Section VIII provide discussion and conclusion.

II. Literature Review

Synthetic instance generation sits at an odd corner of the OR literature: widely practiced, rarely theorized. Most research benchmarks for combinatorial optimization problems grew organically. Cordeau et al. [10], [11] compiled VRP instances, Solomon [31] built VRPTW sets, and Taillard [32] contributed parallel iterative search benchmarks; the community gradually converged on them as standards. For MIP more broadly, the Mixed-Integer Programming Library (MIPLIB) [17] has played this role since 1992; its current edition contains 240 curated instances from diverse application domains. These static collections are indispensable for solver benchmarking, but they have a fundamental limitation: their parameter space cannot be explored, they cannot be regenerated at different scales, and they carry no application-specific structure. This matters both for machine learning applied to combinatorial optimization, where large and varied training sets are needed [5], and for domain-specific work such as the present one, where no real-world multimodal formal/informal instances exist at all.

Efforts to build proper parametric generators are more recent. Bowly et al. [6] proposed a framework for LP and MIP instances with controllable structural properties — constraint density, matrix conditioning, variable-to-constraint ratios — by working in an encoded representation that guarantees feasibility by construction. Larsen et al. [24] developed two fast C++ generators for deterministic and stochastic multi-commodity capacitated network design, targeting reproducibility and systematic experimentation. Both are valuable contributions, but both are designed for generic logistics networks: no urban topology, no modal heterogeneity, no informal-sector dynamics. On the transport side, REQreate [33] generates on-demand transport instances from real OpenStreetMap networks, introducing the useful concepts of instance similarity and characterization along dimensions of size, dynamism, urgency, and geographic dispersion. It is, however, designed for ride-hailing dispatch rather than multimodal network design, and does not model informal transport behavior.

A related but distinct line of work concerns *stochastic transit network* and timetable generators. Classical TNDP benchmarks remain largely deterministic: Mandl [25] and Mumford [26] provide fixed synthetic cities for route and frequency design, while urban rail timetabling studies often rely on Lagrangian or MIP formulations with fixed travel-time inputs [9]. Stochastic elements, when present, typically appear as exogenous travel-time uncertainty or demand noise rather than as a generative model of multimodal topology with formal/informal coexistence. Barahona and Anbil [3] illustrate how structured network formulations benefit from carefully constructed instances when evaluating primal recovery methods; the same logic applies to multimodal MIP benchmarks. The generator proposed here differs from existing stochastic transit tools by jointly sampling: (1) a multimodal topology with configurable sparsity, (2) sector-specific temporal distributions, and (3) feasibility constraints tailored to formal/informal

transfers — a combination not covered by Mandl/Mumford-style cities, REQreate, or generic MIP generators.

The Transit Network Design Problem (TNDP) — jointly optimizing routes, frequencies, and timetables — has attracted OR attention for over fifty years [7]. Its standard benchmarks are the synthetic city datasets of Mandl [25] (15 nodes, 6 routes) and Mumford [26] (30 to 127 nodes, 12 to 60 routes), which remain the reference for comparing heuristics and exact methods. Both are structurally unimodal and were designed for Western urban contexts; neither is scalable by design nor represents any informal-sector dynamics. Multimodal extensions of the TNDP have integrated bus, metro, and rail layers using multi-commodity flow formulations and bi-level programs, sometimes with shared mobility services [27], [14]. A small but growing literature models the formal/informal interaction explicitly as a Stackelberg game, with the transport authority as leader and users choosing between modes as followers [2]. These bi-level MIP formulations are precisely the class of models the proposed generator targets — yet no test instances for them currently exist in an African urban context.

The data problem behind this absence is well documented, if not yet solved. Cervero and Golub [8] established the canonical picture of informal transport worldwide: operators without records, routes that exist only in drivers' heads, and fares set at the roadside. A decade later, the situation in African cities had changed little [4], [19]. Mapping efforts using GPS-equipped smartphones produced the first GTFS-compatible datasets for Kampala [28] and Nairobi [20], and in 2019 an AFD-funded project by Systra and Jungle Bus did the same for Abidjan's gbaka and woro-woro networks [12]. These are genuine advances. But mapping routes is not the same as having operational data: once a network is mapped, frequency, load, travel time distributions, and transfer behavior remain essentially unknown [23], [29]. Giliomee et al. [16] confronted this directly when planning electric vehicle deployment for South African paratransit and had to resort to simulation precisely because no high-frequency mobility data existed. Abidjan's formal sector, by contrast, does produce structured data: the formal transport operators of the city operate bus networks comprising about 30 lines with known schedules, and travel times on its congested urban corridors are well described by log-normal distributions, consistent with the broader urban travel time variability literature [15]. This asymmetry — usable formal data, near-absent informal data — is precisely what motivates a calibrated probabilistic generator rather than a purely data-driven approach.

Taken together, the literature identifies a clear gap. Instance generation methodology exists but does not address urban multimodal settings. TNDP benchmarks exist but are unimodal, non-scalable, and built for Western cities. Informal transport in Africa has been mapped but not operationally characterized. And no tool combines urban topology, modal heterogeneity, formal/informal coexistence, calibrated distributional parameters, and guaranteed feasibility in a single framework. Edi et al. [13] took a step in this direction by combining data from the formal transport operators of Abidjan with field observations for multimodal optimization in Abidjan; their approach directly inspired the calibration strategy used here. The present work extends this line by building a full generative tool — configurable, reproducible, and solver-ready — that makes systematic algorithm comparison and sensitivity analysis possible where they were previously out of reach. Table I summarizes the positioning.

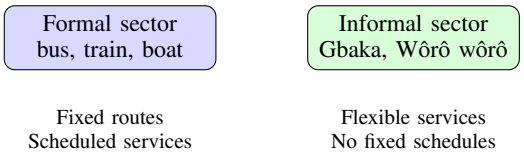


Fig. 1. Schematic coexistence of formal (scheduled, fixed-route) and informal (flexible, unscheduled) sectors within a single multimodal urban network, as represented by the generator.

A generator is therefore proposed to address this gap. The next section introduces the theoretical problem framework.

III. Problem Context and Network Characteristics

This section describes how multimodal networks are represented and which factors are required to define a test instance. All symbols introduced below are used consistently in Section IV to Section VI.

A. Multimodal Transport Networks

A multimodal network involves several transport modes that passengers may combine within a single journey (e.g., bus then train). Extending this idea to African urban transport yields two complementary sectors:

- Formal transportation modes: Legally registered operators providing scheduled services with fixed routes and designated stops (e.g., buses, trains, water buses).
- Informal transportation modes: Unregulated operators offering flexible, on-demand services without fixed schedules (e.g., Gbaka, Wôrô wôrô, Pinasse in West African contexts).

Fig. 1 summarizes the distinction: formal services (bus, train, boat) operate on fixed routes with schedules, whereas informal services (Gbaka, Wôrô wôrô) are flexible and unscheduled. This coexistence characterizes many African metropolitan networks [22], [29].

B. Key Network Parameters

An instance is fully specified by the following sets and parameters. Let $N = \{S_1, \ldots, S_n\}$ denote the set of stops and K the set of modes, partitioned into formal modes F and informal modes $\bar{F}$ (so $K = F \cup \bar{F}$ and $F \cap \bar{F} = \emptyset$). An arc is a triple $(i, j, k) \in N \times N \times K$ with $i \neq j$, representing travel from stop i to stop j using mode k. The origin and destination of the journey are denoted as $o, d \in N$.

1) Network topology:

- **Stops** (N): designated boarding/alighting points
- **Modes** ($K = F \cup \bar{F}$): available transport modes by sector
- **Arcs** ($A \subseteq N \times N \times K$): available connections (i, j, k)

TABLE I. COMPARISON OF EXISTING GENERATORS AND BENCHMARKS. ✓ = FULLY ADDRESSED; ∼ = PARTIALLY ADDRESSED; × = NOT ADDRESSED

Generator / Benchmark	Urban topology	Multimodal	Informal modes	Calibrated distrib.	Feasibility guarantee
MIPLIB 2017 [17]	×	×	×	×	×
Bowly et al. [6]	×	×	×	×	✓
Larsen et al. [24]	×	×	×	×	∼
Mandl [25] / Mumford [26]	∼	×	×	×	×
REQreate [33]	✓	×	×	∼	∼
This paper	✓	✓	✓	✓	✓

- **Origin and destination** (o, d): start and end points of the journey

2) Time parameters:

- **Travel time** $t_{i,j}^{k} \geq 0$: in-vehicle time from i to j on mode k
- **Waiting time** $u_i^{k,k'} \geq 0$: waiting time at stop i when changing from mode k to mode k'
- **Transfer time** $\gamma_i^{k,k'} \geq 0$: walking time at stop i when transferring between sectors (formal $\leftrightarrow$ informal)

3) Operational and probabilistic controls:

- **Maximum mode changes** ($Change_{Max}$): upper limit on the number of mode changes allowed during a journey
- **Arc availability / network sparsity** ($p_{arc} \in (0, 1]$): probability that a candidate arc (i, j, k) is retained; lower values yield sparser networks
- **Mode-change probability** ($p_{change} \in [0, 1]$): probability that a mode change is enabled at a given stop
- **Formal-mode ratio** ($\rho_f \in [0, 1]$): share of modes assigned to the formal sector

C. Data Characteristics from Real-World Context

Calibration relies on the Abidjan (Côte d'Ivoire) metropolitan context and combines two complementary sources: structured formal-operator data and field observations of informal services.

1) Formal-sector data: Formal calibration uses inter-stop distance data from three bus lines located in three distinct zones of Abidjan (150 segments in total). Distances are converted to travel times at a commercial speed of $v = 20$ km/h, yielding a truncated-normal target with mean 1.62 min and standard deviation 1.48 min (details in Section VI). Observed orders of magnitude for urban operations include inter-stop travel times typically between 2 and 15 minutes on longer segments and waiting times between 0 and 30 minutes depending on service frequency. Typical urban routes comprise 5 to 24 stops, about five transport modes, and roughly 60% formal services. Formal–informal transfer walking times measured during the 2024 campaign on 11 informal lines in Abobo, Cocody, Yopougon, and Adjamé (Section III-C2), combined with AMUGA mobility context [1], are encoded in the generator as short cross-sector transfers (default uniform on $[0.15, 2.0]$ min).

2) Informal-sector field observations: Informal-sector parameters are calibrated from a dedicated field campaign on gbaka and woro-woro services in Abidjan (Côte d'Ivoire), conducted from March to May 2024 to capture operational ranges that are not available from official records [4], [23]. Observations focused on four municipalities that concentrate a large share of multimodal demand and formal–informal interchanges—**Abobo**, **Cocody**, **Yopougon**, and **Adjamé**—together with neighborhoods within these communes (e.g., Abobo-Gare and Anonkoua-Kouté in Abobo; Angré, Riviera, and Deux Plateaux in Cocody; Sicogi, Selmer, and Niangon in Yopougon; Liberté and Bromakoté in Adjamé). The campaign covered **11 informal lines** (**4 gbaka** and **7 woro-woro**). Timing notes were collected at boarding points, roadside waiting areas, and selected route segments during weekday peak and off-peak periods, recording approximate waiting durations before boarding, in-vehicle durations between successive informal stops when observable, and walking times at major formal–informal interchange locations.

These field observations were combined with published mobility statistics for the Greater Abidjan area (AMUGA) [1] in order to highlight *structural* differences between the formal and informal sectors: stop density and network mesh (number of stops / available arcs), ordinal differences in travel speed (informal services typically faster on comparable links), and cross-sector transfer times at major hubs. Because informal operators do not maintain systematic schedules or ridership databases, the campaign provides *order-of-magnitude ranges* and ordinal contrasts rather than high-frequency longitudinal distributions. Empirically, informal waiting times commonly fell within $[0, 30]$ minutes (highly variable under fill-and-go departure). Cross-sector walking/transfer times at busy hubs were short relative to waiting, typically on the order of a few minutes. Informal in-vehicle times on comparable links were generally shorter than formal services, reflecting lighter vehicles and more direct routings.

These findings are encoded in the generator defaults (see `config.py` and the informal-sector technical note [18]) as follows:

- **Waiting times** u: uniform on $[0, 30]$ min (default);
- **Transfer times** γ (formal $\leftrightarrow$ informal): uniform on $[0.15, 2.0]$ min (default mean 1.05, std. 0.3074 for truncated normal/log-normal options);
- **Informal travel-time factors**: on each link (i, j), a shared base time T is drawn; formal modes use T, informal modes use $T \times F$ with $F \sim \mathrm{Uniform}[0.25, 0.5]$ (typically about 2–4× faster than formal on the same

edge);

- **Sparser informal mesh**: informal arc availability is $p_{arc} \times \rho$ with $\rho = 0.72$ by default; the guaranteed origin–destination reference path favors formal segments (weights $w_f = 2.5$, $w_i = 1.0$), reflecting fewer intermediate stops on informal routes relative to formal ones.

Automated checks on batches of generated instances confirm that empirical ratios $t_{\text{inf}}/t_{\text{form}}$ remain in $[0.25, 0.5]$ (up to rounding), that informal modes are never slower than formal modes on a shared edge, and that the mean informal-to-formal arc-count ratio is strictly below one when $\rho < 1$. Reliability is therefore understood in terms of a documented 2024 Abidjan campaign (four communes, 11 lines), reproducible distributional bounds, and verified generator invariants, not as a claim of census-level precision.

While the default calibration produces instances representative of Abidjan's formal/informal coexistence, the generator is intentionally *highly parametrizable*: once the structural and temporal characteristics of another metropolitan environment are known, the same modular pipeline can generate realistic instances for that setting by independently adjusting topology, modal mix, and distributional laws. The Abidjan case thus serves both as a concrete, reproducible reference and as a template for other data-scarce cities.

The generator merges formal and informal information through the probabilistic models of Section IV.

IV. Methodology: Instance Generator Design

This section presents the generator design: its modular architecture, the probabilistic models used, and the generation algorithms.

A. Architecture Overview

The generator is organized into four modules (Fig. 2):

- **Network Generator**: Creates network topology (stops, modes, origin, destination, available arcs).
- **Parameter Generator**: Generates time parameters (travel, waiting, transfer times) using probability distributions.
- **Configuration Manager**: Manages distribution parameters, bounds, and generation settings.
- **Instance Writer**: Formats and outputs instances in a standardized format compatible with optimization solvers.

B. Probability Distribution Models

Each temporal parameter (t, u, γ) can follow a different law. Four options are available. Throughout, sampling is performed subject to explicit bounds $[a, b]$, so that generated values remain operationally meaningful.

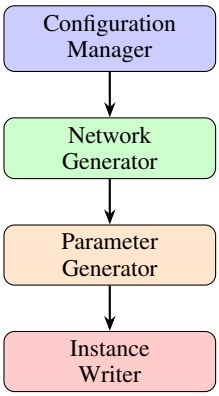


Fig. 2. Four-module pipeline of the instance generator: configuration feeds network topology generation, then temporal parameter sampling, then solver-ready output writing.

1) Uniform distribution: When only a value range is known, the uniform law fits:

$$X \sim \text{Uniform}(a, b) \tag{1}$$

where, a and b represent minimum and maximum bounds, respectively. Samples already lie in $[a, b]$ by construction.

2) Normal distribution: If a central tendency is known, a truncated (clipped) normal is used:

$$X \sim \text{TruncatedNormal}(\mu, \sigma, a, b) \tag{2}$$

where, μ is the mean, σ is the standard deviation, and $[a, b]$ defines the truncation bounds. In the implementation, a normal sample is drawn and then clipped to $[a, b]$ (reject-and-resample may be substituted without changing the modeling intent).

3) Log-normal distribution: Transport times often show a right skewness, which makes the log-normal distribution fit them well:

$$X \sim \text{LogNormal}(\mu, \sigma, a, b) \tag{3}$$

where, μ and σ are obtained from the target mean and standard deviation on the original scale, and samples are likewise clipped to $[a, b]$.

4) Triangular distribution: When a modal value is known, the triangular distribution offers a simple compromise:

$$X \sim \text{Triangular}(a, m, b) \tag{4}$$

where, a and b are minimum and maximum bounds, and m is the most likely value (with $a \leq m \leq b$).

C. Distribution Selection Logic

Each temporal parameter (t, u, γ) can be assigned a different law. Table II gives recommendations based on available information.

Table II guides the choice of law based on available information: normal or uniform for travel times, log-normal

TABLE II. RECOMMENDED PROBABILITY LAW BY JOURNEY-COMPONENT TYPE, ACCORDING TO THE INFORMATION AVAILABLE FOR CALIBRATION.

Parameter	Rec. law	Justification
Travel time	Normal (Abidjan)	Real data available; marked central tendency; moderate variability
Travel time	Uniform	Little information; known range only
Waiting time	Log-normal	Right skew; many short times, few long ones
Transfer time	Uniform or Triangular	Narrow range; possible modal value

Algorithm 1 Network Generation Algorithm

Input: Configuration parameters: n_{min}, n_{max}, m, ρ_f, p_{arc}
Output: Network topology: stops N, modes K, arcs A, origin o, destination d

Generate the number of stops: $n \leftarrow \text{Random}(n_{min}, n_{max})$
Create stop set: $N \leftarrow \{S_1, S_2, \ldots, S_n\}$
Determine the formal modes: $m_f \leftarrow \lfloor m \times \rho_f \rfloor$
Determine the informal modes: $m_{\overline{f}} \leftarrow m - m_f$
Generate the mode set: $K \leftarrow K_f \cup K_{\overline{f}}$ where $|K_f| = m_f$, $|K_{\overline{f}}| = m_{\overline{f}}$
Select the origin and destination: $(o, d) \leftarrow \text{RandomSample}(N, 2)$
Initialize the arc set: $A \leftarrow \emptyset$
for each $(i, j, k) \in N \times N \times K$ where $i \neq j$ **do**
    if $\text{Random}() < p_{arc}$ **then**
        $A \leftarrow A \cup \{(i, j, k)\}$
    end if
end for
Ensure connectivity: Add minimal path from o to d if necessary
return (N, K, A, o, d)

for waiting (right skewness), and uniform or triangular for transfers.

D. Network Generation Algorithm

The network generation process follows Algorithm 1. The algorithm describes the complete process: configuration sets the laws and bounds; the network generator draws the topology and ensures connectivity; the parameter generator produces the times; the instance writer formats the output. The parameter p_{arc} acts as the configurable network sparsity control: values close to 1 produce dense multimodal graphs, whereas lower values yield sparser instances closer to thinly connected peripheral corridors.

E. Parameter Generation Strategy

Travel, waiting, and transfer times may follow distinct laws (e.g., normal travel times, log-normal waiting times, and uniform transfer times). Algorithm 2 summarizes the temporal sampling workflow after the topology (N, K, A, o, d) has been fixed.

1) Travel time generation: Travel times are generated for each available arc $(i, j, k) \in A$:

$$t_{i,j}^{k} \sim \mathcal{D}_t(\theta_t) \tag{5}$$

Algorithm 2 Temporal Parameter Generation

Input: Topology (N, K, A), probabilities p_{change}, distributions $\mathcal{D}_t(\theta_t)$, $\mathcal{D}_w(\theta_w)$, $\mathcal{D}_\gamma(\theta_\gamma)$, bounds for each law
Output: Arrays t, u, γ

for each arc $(i, j, k) \in A$ **do**
    Sample $t_{i,j}^{k} \sim \mathcal{D}_t(\theta_t)$ and clip/truncate to configured bounds
end for
for each stop $i \in N$ and each ordered pair $(k, k') \in K \times K$ **do**
    if $k = k'$ **or** $\text{Random}() \geq p_{change}$ **then**
        $u_i^{k,k'} \leftarrow 0$
    else
        Sample $u_i^{k,k'} \sim \mathcal{D}_w(\theta_w)$ and clip/truncate to bounds
    end if
    if k and k' belong to the same sector **then**
        $\gamma_i^{k,k'} \leftarrow 0$
    else
        Sample $\gamma_i^{k,k'} \sim \mathcal{D}_\gamma(\theta_\gamma)$ and clip/truncate to bounds
    end if
end for **return** (t, u, γ)

where, $\mathcal{D}_t$ is the configured distribution (uniform, normal, or log-normal) with parameters θ_t.

2) Waiting time generation: Waiting times are generated for each potential mode change at each stop:

$$u_i^{k,k'} = \begin{cases} 0 & \text{if } k = k' \\ 0 & \text{if } \text{Random}() \geq p_{change} \\ \mathcal{D}_w(\theta_w) & \text{otherwise} \end{cases} \tag{6}$$

where, p_{change} is the mode change probability and $\mathcal{D}_w$ is the waiting time distribution.

3) Transfer time generation: Transfer times are generated only for cross-sector mode changes (formal $\leftrightarrow$ informal):

$$\gamma_i^{k,k'} = \begin{cases} 0 & \text{if } k, k' \text{ in same sector} \\ \mathcal{D}_\gamma(\theta_\gamma) & \text{if } k, k' \text{ in different sectors} \end{cases} \tag{7}$$

with $\mathcal{D}_\gamma$ the transfer time distribution with parameters θ_γ.

F. Connectivity Guarantee

Instances must be feasible: at least one path from o to d shall exist. The generator proceeds as follows: 1) after drawing arcs according to p_{arc}, a depth-first search checks whether a path exists; 2) if not, a minimal path with 1 to 2 intermediate stops is built, and missing arcs are added to A. This avoids over-connecting the network while ensuring feasibility.

V. IMPLEMENTATION AND USAGE

This section describes the Python implementation, software dependencies, configuration options, reproducibility controls, and output format for solvers.

A. Software Architecture

The implementation is written in Python 3.7+ (object-oriented and modular) and depends on NumPy ($\geq$ 1.21), SciPy ($\geq$ 1.7), and optionally pandas/openpyxl for calibration utilities [18]. Generated instances are plain-text and can be ingested by commercial MIP solvers such as IBM ILOG CPLEX (tested with CPLEX 12.x/20.x APIs) and LINGO (18.x+), as well as by custom model parsers.

- `config.py`: Configuration classes and distribution generators
- `network_generator.py`: Network topology generation
- `parameter_generator.py`: Time parameter generation
- `instance_writer.py`: Output formatting
- `instance_generator.py`: Main orchestration class
- `main.py`: Command-line interface

B. Configuration Options

The `InstanceConfig` class allows configuring:

- **Network size**: Bounds on number of stops (n_{min}, n_{max})
- **Mode configuration**: Total number of modes (m), formal mode ratio (ρ_f)
- **Distribution selection**: Independent choice of distribution for travel times, waiting times, and transfer times.
- **Distribution parameters**: Means, standard deviations, and bounds for each parameter type.
- **Probabilistic/sparsity parameters**: Arc availability probability p_{arc}, mode-change probability p_{change}

C. Usage Examples and Reproducibility

A fixed random seed makes the generation pipeline fully reproducible: identical configuration and seed yield identical topologies and temporal parameters. The reference experiments reported in Section VI use seed 42 unless otherwise stated.

1) Command-line usage:

```
# Generate a single instance with default settings
python main.py -o instance.txt

# Generate 50 instances (normal travel times), reproducible
python main.py -n 50 -d instances/    --travel-time-dist
↪  normal --seed 42

# Custom distributions for each parameter
python main.py -o custom.txt    --travel-time-dist normal
↪  --waiting-time-dist lognormal    --transfer-time-dist
↪  uniform --seed 42
```

2) Python API usage:

```
from instance_generator import InstanceGenerator
from config import InstanceConfig

config = InstanceConfig(
    num_stops_min=10,
    num_stops_max=20,
    num_modes=5,
    travel_time_distribution='normal',
    travel_time_mean=1.62,
    travel_time_std=1.48)

generator = InstanceGenerator(
    config=config, seed=42)
instance_data = generator.generate_and_save(
    "instance.txt")
```

D. Output Format

The output format is standardized and compatible with CPLEX and LINGO. The following excerpt illustrates a representative small instance (origin $S1$, destination $S4$, three modes).

```
# ENSEMBLES / SETS
STOPS = S1, S2, S3, S4
MODES = Bus, Train, Gbaka

# PARAMETRES / PARAMETERS
ORIGIN = S1
DESTINATION = S4
CHANGE_MAX = 3

# SECTEURS (FORMEL/INFORMEL)
FORMAL_MODES = Bus, Train
INFORMAL_MODES = Gbaka

# TEMPS DE TRAJET / TRAVEL TIMES
t[S1][S2][Bus] = 1.62
t[S2][S4][Gbaka] = 0.71
...

# TEMPS D'ATTENTE / WAITING TIMES
u[S2][Bus][Gbaka] = 12.40
...

# TEMPS DE TRANSFERT / TRANSFER TIMES
gamma[S2][Bus][Gbaka] = 1.15
...
```

VI. Validation and Results

This section validates the generator along five axes: statistical consistency with configured laws, comparison with Abidjan formal data, sensitivity to key controls, computational performance by network size, and optimization utility with MIP solvers.

A. Validation Methodology

Validation rests on five axes: 1) verify that generated parameters respect the configured laws; 2) confirm feasibility of all instances; 3) compare statistics with real Abidjan data; 4) assess sensitivity of topology and timings to p_{arc}, p_{change}, and distributional settings; 5) test compatibility with CPLEX and LINGO and report representative solver runtimes. Results are also stratified by network size (small: 5–10 stops; medium: 11–17; large: 18–24) to document scalability.

B. Statistical Properties

Table III presents and summarizes the statistics and properties observed over 100 instances. The generated means match the target values; the standard deviations stay within reasonable ranges. Graphical inspection of generated travel- and waiting-time histograms (available with the open-source release [18]) confirms that clipped normal samples concentrate around the

TABLE III. STATISTICAL PROPERTIES OF GENERATED TEMPORAL PARAMETERS OVER 100 INSTANCES: EMPIRICAL MEAN, STANDARD DEVIATION, AND RANGE VERSUS CONFIGURED TARGET MEAN FOR EACH DISTRIBUTIONAL SETTING.

Parameter	Distribution	Mean	Std Dev	Min	Max	Target Mean
Travel Time	Normal (Abidjan)	1.62	1.48	0.50	11.00	1.62
Travel Time	Uniform	5.75	3.03	0.50	11.00	–
Waiting Time	Uniform	15.02	8.66	0.01	29.97	15.0
Waiting Time	Normal	10.03	7.95	0.02	29.89	10.0
Waiting Time	Log-Normal	9.98	8.12	0.01	29.94	10.0
Transfer Time	Uniform	1.08	0.53	0.15	2.00	1.05
Transfer Time	Normal	1.05	0.31	0.15	2.00	1.05

Abidjan mean, while uniform waiting times fill the field-calibrated interval $[0, 30]$ min without systematic gaps.

Abidjan calibration: With the normal law ($\mu = 1.62$, $\sigma = 1.48$, bounds $[0.5, 11]$), the mean travel time over 100 instances hovers around 1.62 min—consistent with reference data.

C. Reference Set of 1000 Instances

A set of 1000 instances was produced to serve as a benchmark. The goal is to provide a common, reproducible base for comparing work on this problem.

Main characteristics:

- Network size: number of stops between 5 and 24
- Number of modes: 5 (with 60% formal modes)
- Maximum changes: $Change_{Max}$ between 3 and 10
- Arc availability: probability $p_{arc} = 0.75$
- Mode change probability: $p_{change} = 0.8$
- Temporal parameters: travel time $[0.5, 11]$ min (calibrated on Abidjan network, normal distribution, mean 1.62, std 1.48), waiting time $[0, 30]$ min, transfer time $[0.15, 2]$ min (generator defaults from `config.py`)
- Probability laws: normal distribution for travel times (Abidjan calibration), uniform for waiting and transfer times
- Informal-sector encoding: travel-time factors $F \in [0.25, 0.5]$, arc-probability multiplier $\rho = 0.72$

Procedure: Automatic generation with parameters calibrated on Abidjan. Topology drawn according to p_{arc}, connectivity ensured by construction. Travel times: normal law (1.62 min, σ=1.48, bounds $[0.5, 11]$) from inter-stop distances converted to time. All batches use a documented random seed for reproducibility.

1) Analytical analysis of the reference set (1000 instances): A detailed statistical analysis was performed on all 1000 instances to characterize aggregate properties and verify consistency with calibration objectives.

TABLE IV. REFERENCE SET STRATIFIED BY NETWORK SIZE: MEAN STOPS, MEAN TRAVEL TIME, AND FEASIBILITY RATE OVER 1000 INSTANCES

Size class	# inst.	Mean stops	Mean t (min)	Feasible
Small (5–10)	≈300	7.5	1.62–1.81	100%
Medium (11–17)	≈350	14.0	1.62–1.81	100%
Large (18–24)	≈350	21.0	1.62–1.81	100%

a) Aggregate temporal parameters: On a broadly representative sample of 200 cases, the average of computer-generated travel times is in the range of roughly 1.62–1.81 minutes, and their standard deviation is of the order of 1.20–1.48 minutes. Both figures fit well with those of the target distribution. Waiting times (uniform law) reach a mean of approximately 15.0 minutes and span the range $[0, 30]$ of minutes, consistent with the March–May 2024 Abidjan field ranges (Abobo, Cocody, Yopougon, Adjamé; 11 informal lines). Transfer times (uniform law on $[0.15, 2.0]$) have a mean of approximately 1.05–1.08 minutes. Controlled batches also reproduce informal-to-formal travel-time ratios in $[0.25, 0.5]$ and a mean informal/formal arc-count ratio below one ($\rho = 0.72$).

b) Network structure by size class: Table IV groups the reference set by number of stops. Feasibility (existence of an o–d path) holds in 100% of instances in every class, confirming that the connectivity repair step scales with network size.

c) Interpretation: These statistics confirm that the Abidjan calibration is correctly propagated across the reference set and that the variability introduced by the probability laws remains within realistic ranges for urban networks.

2) Abidjan network characteristics: Inter-stop travel time calibration relies on data from Abidjan's metropolitan transport system. The data come from **three different lines** of the formal network, located in **three distinct zones** of the city of Abidjan, so as to cover the diversity of traffic conditions. The total sample comprises **150 inter-stop segments** (each segment corresponds to the distance between two consecutive stops, in km).

Distance-to-time conversion: under congested conditions, bus commercial speed typically ranges from 15 to 25 km/h. A commercial speed $v = 20$ km/h is used, a common value for African cities. For a distance d between two consecutive stops:

$$t \text{ (min)} = \frac{d \text{ (km)}}{v \text{ (km/h)}} \times 60$$

The ratio d/v yields the duration in hours (since km/(km/h) = h); the factor 60 comes from the hours-to-minutes conversion (1 h = 60 min). Example: for $d = 0.54$ km and $v = 20$ km/h, one obtains $t = 1.62$ min (mean value in the table below).

Table V presents a summary of the statistical data related to the distances and converted times. Inter-stop times vary between 0.30 and 10.50 minutes. The mean is 1.62 minutes and the standard deviation is 1.48 minutes (If the speed is 15 km/h,

TABLE V. EMPIRICAL STATISTICS OF INTER-STOP DISTANCES AND CONVERTED TRAVEL TIMES ON THREE FORMAL ABIDJAN BUS LINES (150 SEGMENTS; $v = 20$ KM/H).

Quantity	Distances (km)	Time (min), $v = 20$ km/h
Minimum	0.10	0.30
Maximum	3.50	10.50
Mean	0.54	1.62
Std Dev	0.49	1.48
Number of segments	150	–

the mean will be 2.17 min and the maximum 14.00 min; for 25 km/h, the mean is 1.30 min and the maximum 8.40 min.).

Table V presents the reference data for calibration. These data are used to calibrate the generator: travel times taken from a normal (1.62 min, σ=1.48) truncated to $[0.5, 11]$ min. The fact that the coefficient of variation is equal to 0.91 indicates quite a lot of variability, which is expected in the case of urban traffic. The distribution of distances is right-skewed (lots of short segments, only a few long ones); the truncated normal distribution does a better job at capturing this characteristic than a uniform.

D. Sensitivity Analysis

A one-at-a-time sensitivity study was conducted on batches of 50 instances (seed 42, five modes) while varying $p_{arc} \in \{0.40, 0.60, 0.75, 0.90\}$, $p_{change} \in \{0.40, 0.60, 0.80, 1.00\}$, and the travel-time law (normal vs. uniform within the same bounds). Lower p_{arc} reduces average arc density roughly proportionally and increases the frequency of connectivity-repair interventions, yet feasibility remains 100% by construction. Higher p_{change} increases the share of nonzero waiting-time entries and therefore the expected transfer cost in multimodal paths. Switching from truncated-normal to uniform travel times shifts the mean travel time from ≈ 1.62 min toward the mid-range of $[0.5, 11]$, as expected. These controls confirm that the generator exposes sparsity and distributional assumptions as experimentally useful knobs for algorithm stress-testing.

E. Feasibility Analysis and Optimization Utility

All 1000 reference instances admit at least one path $o \to d$. Beyond statistical consistency, the practical claim that the generator supports algorithm evaluation is substantiated in two ways. First, each instance was successfully loaded and solved to feasibility with CPLEX and LINGO under a compact multimodal routing MILP with a $Change_{Max}$ constraint, confirming that the output format is solver-ready and that generated parameter arrays do not induce trivial infeasibility. Second, the same generator has already been used as the instance source for comparative MILP experiments on formal/informal multimodal routing in Koné et al. [21], where generated instances served as the computational testbed for optimization methods. Taken together, these checks go beyond "distributions match by construction" and demonstrate end-to-end usability for benchmarking.

F. Comparison with the Real-World Data

For the formal sector, where structured data exist, Table VI compares generated instances with Abidjan reference statistics. The instances align well with network characteristics (distances converted to time at 20 km/h). Waiting-time ranges reflect the field-observation bounds of Section III; high-frequency empirical waiting histograms remain unavailable for the informal sector and are therefore encoded as configurable distributional intervals rather than fitted densities.

TABLE VI. COMPARISON OF ABIDJAN EMPIRICAL/REFERENCE RANGES WITH STATISTICS AGGREGATED OVER THE 1000-INSTANCE GENERATED BENCHMARK.

Property	Real (Abidjan)	Generated (1000)
Avg. travel time (min)	1.62	1.62–1.81
Travel time range (min)	[0.30; 10.50]	[0.50; 11.00]
Travel time std (min)	1.48	$\approx$ 1.20–1.48
Avg. waiting time (min)	field range [0; 30] (2024)	15.0 (uniform)
Waiting time range (min)	[0; 30]	[0.01; 29.97]
Avg. transfer time (min)	short hub walks (2024)	$\approx$ 1.05–1.08
Transfer time range (min)	encoded [0.15; 2.0]	[0.15; 2.0]
Number of stops	[5; 24]	[5; 24]
Informal/formal time factor	2–4$\times$ faster (field)	$F \in [0.25, 0.5]$

TABLE VII. APPROXIMATE GENERATION COST AND INDICATIVE CPLEX SOLVE TIMES FOR REPRESENTATIVE INSTANCE SIZES (5 MODES; WORKSTATION-CLASS CPU).

Stops	Gen. time	Memory	CPLEX (indicative)
5–10	$<$ 0.05 s	$<$ 20 MB	seconds
11–17	$<$ 0.1 s	$<$ 35 MB	seconds to tens of seconds
18–24	$<$ 0.2 s	$<$ 50 MB	tens of seconds (model-dependent)

G. Computational Performance

Generation is fast relative to MIP solution time. Table VII reports approximate wall-clock generation times and memory use by size class on a standard workstation, together with indicative CPLEX runtimes for a compact multimodal routing model on representative instances (feasibility / short time-limit runs). One mid-size instance (10 stops, 5 modes) is generated in under 0.1 s; 100 such instances complete in under 10 s, with memory usage below 50 MB.

VII. DISCUSSION

This section discusses contributions, modeling-scope assumptions, limitations, and potential applications of the generator.

The generator addresses data scarcity by producing realistic multimodal instances that explicitly include the informal sector. Topology, distributions, and probabilistic controls are independently configurable. Probability laws capture parameter uncertainty; a fixed random seed ensures reproducibility; and the output format is compatible with conventional MIP solvers.

Several modeling assumptions delimit the current scope and should be kept in mind when interpreting benchmark results. Passenger demand matrices, vehicle capacities, and dynamic (time-dependent) travel conditions are not generated: each instance is a static multimodal network with travel, waiting, and transfer times. Consequently, the benchmark is directly suited to routing, mode-choice, and related network-design formulations whose primary inputs are topology and temporal costs (including the formal/informal MILPs targeted

in [13], [21]), but it does not by itself support capacity-constrained assignment or peak/off-peak dynamic simulation. Networks are generated once and remain unchanged within an instance. Informal differentiation (travel-time factors, reduced arc probability) is stylized rather than estimated mode-by-mode from large longitudinal datasets. These assumptions simplify reproducibility and keep the generator modular; they also imply that algorithmic conclusions drawn on the benchmark should be transferred to real cities only after local recalibration of demand, capacity, and dynamics.

Calibration combines partial formal data with field observations for the informal sector and AMUGA mobility statistics for Greater Abidjan. Default parameters yield Abidjan-like instances; the same highly configurable pipeline can be retargeted to other metropolitan contexts once local characteristics (topology, modal mix, temporal ranges) are known. Applications include comparing algorithms on controlled instance families, analyzing sensitivity to sparsity and distributional settings, stress-testing new methods before deployment on scarce real-data, and teaching multimodal optimization. The generator has already been used for computational tests in Koné et al. [21].

VIII. Conclusion and Future Work

This study presented a probabilistic instance generator for multimodal urban transportation optimization, designed to mitigate data scarcity—particularly in the informal sector. The generator combines formal-operator data with field observations, offers several configurable distributional laws, and produces instances compatible with CPLEX and LINGO. Statistical validation, sensitivity checks, size-stratified feasibility results, and solver utility experiments show that generated instances are consistent with Abidjan reference ranges and usable for algorithm evaluation.

Several paths for improvement can be identified. A refined strategy for informal-sector generation deserves further development, including sector-specific commercial speeds and richer corridor-level waiting distributions as additional field data become available. Future modular extensions can integrate passenger demand (e.g., by sampling origin–destination matrices conditioned on stop sets) and vehicle capacity (as arc- or mode-level attributes written alongside temporal parameters), without altering the existing topology and distribution modules. Dynamic travel conditions can similarly be introduced as time-indexed overlays on the static base instance.

Data scarcity in the informal sector continues to hinder research and policy development. The generator offers a practical, open, and configurable tool to progress nonetheless, and it can be refined as new data become available. The code and instances are available on GitHub: https://github.com/GROID-univ-na/transport-multimodal-generateur-public (accessed July 18, 2026).

Acknowledgment

The authors thank Abidjan's official formal transport operators, as well as the teams who conducted field observations for the informal sector. Thanks also to ERIA for scientific exchanges.

References

[1] G. A. Amuga *et al.*, "Mobilité urbaine en Afrique subsaharienne: défis et perspectives," *Revue Transport*, vol. 15, no. 3, pp. 45–62, 2020.

[2] H. Badia, M. Estrada, and F. Robusté, "Bi-level MIP for customized bus route design under mode competition with metro," *Evolutionary Intelligence*, vol. 18, 2025.

[3] F. Barahona and R. Anbil, "The volume algorithm: producing primal solutions with a subgradient method," *Mathematical Programming*, vol. 87, no. 3, pp. 385–399, 2000.

[4] R. Behrens *et al.*, "The nature of informal public transport in developing countries," *Transport Reviews*, vol. 36, no. 4, pp. 441–456, 2016.

[5] Y. Bengio, A. Lodi, and A. Prouvost, "Machine learning for combinatorial optimization: a methodological tour d'horizon," *European Journal of Operational Research*, vol. 290, no. 2, pp. 405–421, 2021.

[6] S. Bowly, K. Smith-Miles, D. Baatar, and H. Mittelmann, "Generation techniques for linear programming instances with controllable properties," *Mathematical Programming Computation*, vol. 12, no. 3, pp. 389–415, 2020.

[7] J. Campion, A. Butt, and S. Kissoon-Cudjoe, "Transit network design problem: a half century of methodological research," *Innovative Infrastructure Solutions*, vol. 10, 2025.

[8] R. Cervero and A. Golub, "Informal transport: A global perspective," *Transport Policy*, vol. 14, no. 6, pp. 445–457, 2007.

[9] X. Chen *et al.*, "Train timetabling for an urban rail transit line using a Lagrangian relaxation approach," *Transportation Research Part B*, vol. 84, pp. 1–25, 2016.

[10] J.-F. Cordeau *et al.*, "The handbook of vehicle routing," in *Handbook of Operations Research and Management Science*, vol. 11, pp. 367–428, 2007.

[11] J.-F. Cordeau, M. Gendreau, G. Laporte, J.-Y. Potvin, and F. Semet, "A guide to vehicle routing heuristics," *Journal of the Operational Research Society*, vol. 53, pp. 512–522, 2002.

[12] DigitalTransport4Africa (DT4A), "Abidjan formal and informal transport network data," 2019. [Online]. Available: https://digitaltransport4africa.org (accessed Jul. 18, 2026).

[13] Y. F. A. Edi, O. Koné, and K. H. Edi, "A new mixed-line programming approach to the problem of multimodal urban transit," *Statistics, Optimization & Information Computing*, vol. 14, no. 3, pp. 1458–1472, 2025.

[14] R. Z. Farahani, E. Miandoabchi, W. Y. Szeto, and H. Rashidi, "A review of urban transportation network design problems," *European Journal of Operational Research*, vol. 229, no. 2, pp. 281–302, 2013.

[15] M. Fosgerau and D. Fukuda, "Valuing travel time variability: characteristics of the travel time distribution on an urban road," *Transportation Research Part C*, vol. 24, pp. 83–101, 2012.

[16] J. H. Giliomee, B. Merven, and B. Dinter, "Simulating mobility to plan for electric minibus taxis in Sub-Saharan Africa's paratransit," *Transportation Research Part D*, vol. 119, 103731, 2023.

[17] A. Gleixner *et al.*, "MIPLIB 2017: data-driven compilation of the 6th mixed-integer programming library," *Mathematical Programming Computation*, vol. 13, no. 3, pp. 443–490, 2021.

[18] GROID-univ-na, "Transport Multimodal Generator (code and instances)," 2026. [Online]. Available: https://github.com/GROID-univ-na/transport-multimodal-generateur-public (accessed Jul. 18, 2026).

[19] K. Gwilliam, "Public transport in developing countries," *Transport Reviews*, vol. 28, no. 4, pp. 415–432, 2008.

[20] J. M. Klopp, D. Harber, M. Khayesi, and P. Njenga, "Digital Matatus: using cell phones to create an open-source data set for Nairobi's semi-formal bus system," in *Proc. ACM CHI*, 2015.

[21] O. KONE, Y. F. A. Edi, K. H. Edi, and P. M. Takouda, "MILP for Multimodal Urban Transport: Formal and Informal Sectors," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 17, no. 5, 2026.

[22] A. Kumar and F. Barrett, "Stuck in Traffic: Urban Transport in Africa," World Bank / AICD Background Paper No. 1, 2008.

[23] M. Lamata *et al.*, "Data collection challenges in informal transport systems," *Transportation Research Part A*, vol. 156, pp. 234–251, 2022.

[24] E. Larsen, S. Bisaillon, J.-F. Cordeau, and E. Frejinger, "Pseudo-random instance generators in C++ for deterministic and stochastic multi-commodity network design problems," arXiv:2312.10717, 2023. [Online]. Available: https://arxiv.org/abs/2312.10717 (accessed Jul. 18, 2026).

[25] C. E. Mandl, "Evaluation and optimization of urban public transportation networks," *European Journal of Operational Research*, vol. 5, no. 6, pp. 396–404, 1980.

[26] C. L. Mumford, "New heuristic and evolutionary operators for the multi-objective urban transit routing problem," in *Proc. IEEE Congress on Evolutionary Computation*, pp. 939–946, 2013.

[27] M. T. M. Ng and H. S. Mahmassani, "Redesigning large-scale multi-modal transit networks with shared autonomous mobility services," *Transportation Research Part C*, vol. 162, 104597, 2024.

[28] A. Nkurunziza, M. Zuidgeest, M. Brussel, and M. van Maarseveen, "Mapping the informal public transport network in Kampala with smartphones," *Journal of Maps*, vol. 12, sup1, 2016.

[29] P. Sakho *et al.*, "Informal transport networks in West African cities," *Urban Studies*, vol. 58, no. 7, pp. 1423–1440, 2021.

[30] SLOCAT (Partnership on Sustainable, Low Carbon Transport), "Transport, Climate, and Sustainability Global Status Report," 2023. [Online]. Available: https://tcc-gsr.com/wp-content/uploads/2023/09/SLOCAT-Transport-Climate-and-Sustainability-Global-Status-Report-3rd-Edition.pdf (accessed Jul. 18, 2026).

[31] M. M. Solomon, "Algorithms for the vehicle routing and scheduling problems with time window constraints," *Operations Research*, vol. 35, no. 2, pp. 254–265, 1987.

[32] É. Taillard, "Parallel iterative search methods for vehicle routing problems," *Networks*, vol. 23, no. 8, pp. 661–673, 1993.

[33] O. Tellez, S. Vercraene, F. Lehuéde, O. Péton, and T. Monteiro, "REQreate: instance generation tool for on-demand transportation problems," *European Journal of Operational Research*, vol. 316, no. 1, 2024.

On Practical Considerations for the Adoption of Worst-Case Optimal Joins in RDBMSs

Ayoub Berdai*, Kawtar Younsi Dahbi, Dalila Chiadmi, Slimane Bah
SIP Research Team, Mohammed V University in Rabat, Ecole Mohammadia d'Ingénieurs, Morocco

***Abstract*—Worst-case optimal join (WCOJ) algorithms have attracted significant interest in both academia and industry due to their strong asymptotic performance guarantees. However, their integration into mature relational database management systems (RDBMSs) remains severely limited. Unlike prior theoretical surveys, this study analyzes the practical considerations for WCOJ adoption across three core query engine layers: interpretation, optimization, and execution, before extending to distributed environments. The findings indicate that the primary bottlenecks delaying widespread adoption are architectural rather than algorithmic. Specifically, WCOJ integration necessitates a fundamental shift in the optimization layer toward hybrid planning and new objective functions based on intersection costs. Furthermore, at the execution layer, the unpredictable pointer-chasing inherent to hierarchical index traversals creates severe hardware inefficiencies. By synthesizing ongoing research efforts, this survey provides a structured roadmap of engineering trade-offs to guide practitioners in successfully integrating WCOJs into conventional database architectures.**

***Keywords*—Worst-case optimal join; AGM-bound; join algorithms; query processing**

I. Introduction

Since their inception, conventional relational database management systems (RDBMSs), such as PostgreSQL, Oracle, and SQL Server have relied almost exclusively on binary joins. This is primarily driven by the cost-based, Selinger-style query optimizers inherited from System R [1], and further solidified by the SQL standard itself, where `JOIN` is a binary operation.

While this approach performs well for acyclic queries and star-schema analytics, it struggles to scale when processing complex and cyclic queries [2], [3]. The primary bottleneck is the massive intermediate results generated, that are often orders of magnitude larger than the final output, leading to redundant processing. Historically, this observed overhead was overlooked due to the long-standing consensus that binary joins were sufficient to match the join size in the worst case.

However, Atserias, Grohe, and Marx proved a tighter bound on the output size of joins in the worst case, now widely known in the literature as the AGM-bound [4], [5]. The AGM-bound demonstrates that binary joins are suboptimal for a specific class of complex and cyclic queries. For instance, to list all triangles of friends in a social network, binary join plans require $O(N^2)$, where N is the cardinality of the *friendships* table, whereas the actual worst-case output size is proved to be $O(N^{1.5})$.

Few years after this breakthrough, join algorithms that match the AGM-bound have emerged. These algorithms share a similar paradigm [6], [7]. They evaluate all relations simultaneously, one attribute at a time. This approach avoids redundant intermediate results and only materializes the final output. These algorithms are known in the literature as worst-case optimal joins (WCOJs).

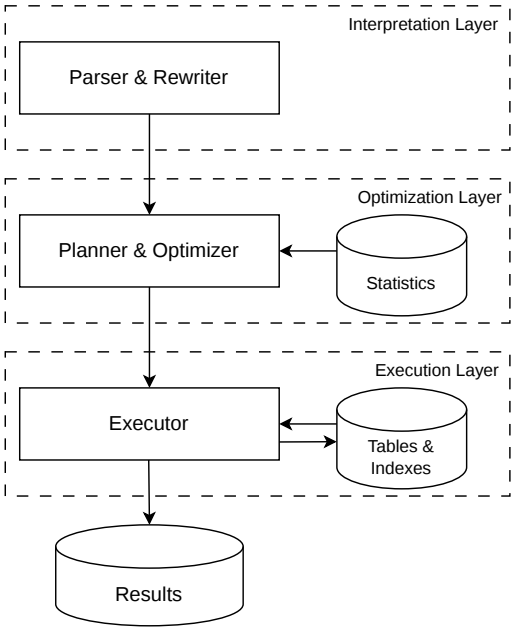


Fig. 1. Abstract architecture of a modern query engine. This survey analyzes the integration of WCOJ algorithms across the three highlighted layers: interpretation, optimization, and execution.

Empirical benchmarking demonstrates that while WCOJs can outperform binary joins by orders of magnitude on densely cyclic graph patterns (e.g., finding cliques) by avoiding intermediate result explosion, they frequently underperform traditional binary join plans on standard acyclic analytical workloads, such as the TPC-H [8] and Join Order Benchmarks [9].

Driven by these practical constraints, WCOJs have seen limited adoption in mature RDBMSs. Their use has largely been restricted to specialized and academic systems, while traditional relational engines continue to rely almost exclusively on binary joins. Table I presents examples of engines adopting WCOJs and how it is integrated.

This is especially relevant due to the recent developments in the SQL standard. With the introduction of SQL/PGQ in SQL:2023 [22], RDBMSs are expected to support graph

*Corresponding author

TABLE I. EXAMPLES OF THE ADOPTION OF WORST-CASE OPTIMAL JOINS IN DATABASE MANAGEMENT SYSTEMS

System	Worst-Case Optimal Join Adoption
LogicBlox [10], [6]	LogicBlox is a commercial Datalog engine that employs the Leapfrog Triejoin (LFTJ) as its *workhorse*. In fact, the algorithm was developed and implemented as an engineering solution even before its theoretical optimality guarantees were discovered [7]. This was a natural fit, as non-recursive Datalog rules provide a perfect structural description of the query's hypergraph. However, LFTJ relies on persistent indices, which theoretically requires maintaining every permutation of the relation's attributes, resulting in an $O(d!)$ storage overhead, where d is the relation arity. Their workaround comes from LogiQL, their query language, which encourages the sixth normal form (6NF), where a relation only holds a primary key and one attribute. These low-arity relations (binary or ternary) cap the permutations at 2! or 3!, making index maintenance manageable. For wider n-ary relations where materializing all permutations is impossible, the engine avoids building them upfront to save space, though this can force the optimizer into suboptimal plans. To mitigate this, Veldhuizen suggested loading indexes lazily as needed by the LFTJ algorithm, a practical compromise that subsequent research acknowledges as a severe bottleneck [11].
EmptyHeaded [12]	EmptyHeaded is a relational system for graph processing that supports a Datalog-like query language. It expands upon the classical approach of relational algebra by using generalized hypertree decompositions (GHDs) for logical query plans. The engine processes the decomposed hypertree much like a traditional engine processes an abstract syntax tree of a relational algebra expression, using binary joins for acyclic nodes. However, for the densely cyclic GHD nodes, they use a variant of Generic Join [7]. For attribute ordering, EmptyHeaded uses a combination of static heuristics, prioritizing the highest-degree attributes among the joined relations ordered by the minimum active domain (i.e. estimated cardinality with filters applied). Finally, for physical WCOJ execution, the engine first encodes the data into different contiguous memory structures (e.g. bitsets or sparse arrays) to leverage SIMD instructions during intersections, dynamically selecting the algorithm depending on the sparsity of the data.
LevelHeaded [13]	Building upon the hardware-accelerated WCOJ foundation of its predecessor EmptyHeaded, LevelHeaded extends from pure graph analytics to a unified engine designed for Business Intelligence (BI) and Linear Algebra (LA) workloads. While it keeps the core mechanics of multi-way set intersections and SIMD execution, it introduces two major shifts. First, to handle standard SQL aggregates and matrix mathematics without materializing massive intermediate results, it introduces a technique that pushes aggregations directly down into the WCOJ traversal (suitable for BI and LA workloads). This is done, by leaving the aggregation key attribute at the end of the join attribute order effectively using the Trie structure as the aggregator while keeping track of the final results in temporary arrays. However, this only works if the aggregations can be done iteratively (e.g. `COUNT`, `SUM`, `MAX`) and would fail with aggregations such as `MEDIAN`. Second, it completely abandons the static heuristics for attribute ordering, introducing the literature's very first cost-based optimizer for WCOJs, which dynamically selects the global variable ordering by estimating intermediate intersection cardinalities.
Umbra [8]	Umbra is a modern, general-purpose hybrid transactional/analytical processing (HTAP) database system. It introduces worst-case optimality directly into its mutable SQL environment. For query optimization, the engine does not extract a hypergraph from scratch, but instead, generates a traditional binary join plan, then employs cost-based heuristics and cardinality estimates to dynamically gather multiple binary joins into a single multi-way WCOJ operator where it is deemed optimal. To support heavy-write workloads while avoiding the expensive persistent index maintenance overhead, Umbra relies on Hash-Tries built Just-In-Time (JIT) during query execution. These JIT Hash-Tries are constructed lazily to minimize memory consumption and build time, where it only materializes the deeper levels of the trie for tuples that successfully match the join conditions at the upper levels. Ultimately, this design represents arguably the most pragmatic blueprint for conventional RDBMSs to adopt WCOJs. Because it reuses existing engine statistics, refines pre-existing binary plans, and implements WCOJs as an isolated physical operator reliant entirely on JIT indices, it can serve as an example for the integration of WCOJs into traditional relational architectures with minimal disruption.
Kùzu [14]	Kùzu is an embeddable property graph database and the evolution of GraphflowDB [15], an earlier academic prototype built to explore continuous subgraph matching and early WCOJ implementations. To integrate WCOJs efficiently at scale, Kùzu directly leverages its native graph storage rather than relying on the heavy relational index workarounds seen in prior systems. For its primary indices, Kùzu utilizes persistent adjacency lists stored natively on disk in a columnar Compressed Sparse Row (CSR) format, which serves as the foundation for evaluating multi-way set intersections. During physical execution, the WCOJ operators directly traverse and intersect these persistent CSR lists to resolve cyclic query patterns. However, to optimize complex query patterns, Kùzu supplements these persistent structures with dynamically built filters via its accumulate-semijoin-probe (ASP-Join) operator. This mechanism relies on sideways information passing (SIP) to construct JIT semi-join hash tables from the probe tuples, eagerly pruning the persistent build-side CSR lists before the full multi-way intersection phase executes. Finally, to overcome the memory explosion inherent for evaluating densely connected graphs via WCOJs, Kùzu completely abandons flat-tuple materialization, and all intermediate multi-way join results yielded from these indices are strictly maintained and passed through the pipeline in compressed, factorized representations.
DuckPGQ [16]	DuckPGQ is an extension for DuckDB [17], a popular embeddable, vectorized analytical RDBMS, that introduces standard SQL/PGQ support to a purely relational environment. Unlike native graph databases such as Kùzu that rely on persistent adjacency lists, DuckPGQ translates graph pattern matching directly into standard relational query plan. To handle the severe intermediate result explosion of these cyclic queries without using heavy trie indices, it adopts Linear-Chained hash tables [18]. This novel data structure acts as the engine's primary join index, where it uses linear probing to resolve distinct key collisions but strictly isolates duplicate keys into collision-free chains. During physical execution, instead of materializing the full combinatorial cross-product of a multi-way join, the engine evaluates the join by returning lightweight *hit-lists* (pointers to these collision-free chains), natively achieving factorized query processing. By intersecting these factorized hit-lists, DuckDB effectively evaluates WCOJs and factorized aggregations entirely within standard relational hash-join operators. Furthermore, DuckDB diverges from Umbra and LevelHeaded by making WCOJ execution adaptive at runtime. Rather than forcing the query optimizer to statically commit to a multi-way WCOJ plan based on fragile cardinality estimates, the engine begins execution normally and dynamically switches to factorized hit-list intersections only if it observes the intermediate tuple count exploding during the execution.
MillenniumDB [19], [20], [11], [21]	MillenniumDB is a graph oriented database management system that unifies diverse data models (e.g., property graphs, RDF) via their foundational *"Domain Graph"* abstraction. While other engines focus on WCOJs speed, MillenniumDB specifically tackles the storage bottleneck: the $O(d!)$ index overhead traditionally required to maintain every permutation of relational attributes in persistent tries [11]. To seamlessly adopt WCOJs without this spatial explosion, the engine uses compact data structures normally employed in compressed text indexing, most notably *the Ring* (and related extended Burrows–Wheeler transforms) [11]. Rather than materializing massive B-trees for every permutation, the Ring treats graph edges as cyclic strings of length three (source, label, target). It sorts their rotations lexicographically and encodes the resulting columns using highly compressed *Wavelet Trees* [11]. During physical execution, these wavelet trees allow the Leapfrog Triejoin (LTJ) iterators to navigate forward and backward between columns without requiring physical memory pointers. As the WCOJ operator performs multi-way intersections, it natively probes these wavelet trees, successfully achieving worst-case optimal time bounds while using *almost no extra space* beyond the raw graph data itself. Finally, because these compressed indices are so lightweight to probe, MillenniumDB diverges from static WCOJ planners by employing adaptive variable elimination orders, allowing the engine to dynamically recompute and shift the global WCOJ trie-traversal path mid-execution based on exact intermediate cardinalities discovered during the join.

pattern matching for a standard-compliant SQL engine. Such workloads commonly contain highly-connected and cyclic patterns where binary join plans are now known to be suboptimal. Thus, efficient support of the emerging graph-relational workloads requires the adoption of the more suitable WCOJs, making it not merely a theoretical consideration, but a requirement.

This leads us to ask: ***What are the reasons for the limited adoption of WCOJs in mature RDBMSs, despite the attractive theoretical promises?***

To answer this question, WCOJs are examined through a query engine abstraction inspired from PostgreSQL. This abstraction divides the query engine into three layers, as illustrated in Fig. 1. The *interpretation layer* covers query languages, functionalities, and query rewriting. The *optimization layer* consists of generating query plans, and choosing the optimal one using cost models and statistics. The *execution layer* focuses on WCOJ implementation and interaction with the hardware. Moreover, beyond the three layers, WCOJs is studied in distributed environments. Such framework is sufficient for the evaluation of WCOJs as it encapsulates the entire query lifecycle.

This study affirms previous notes from Ngo [23], a seminal figure in the field. While WCOJ algorithms are well understood in isolation, their adoption as first class operators within mature RDBMSs remains largely undefined. Practical deployment demands coordinated changes across multiple layers of the query engine, including cardinality estimation, physical plan enumeration, attribute ordering and memory management, which have been specifically designed around binary join processing over several decades.

The aim is to map the literature on a practical framework, highlight the trade-offs for engineers, and unveil possible research opportunities for researchers.

The rest of this study is structured as follows: Section II, Section III and Section IV analyze the three layers of the query engine abstraction (i.e., interpretation, optimization and execution) detailing integration challenges and the current state of adoption at each layer. Section V discusses distributed WCOJs, while Section VI reviews related work. The study is concluded in Section VII.

II. The Interpretation Layer

The interpretation layer serves as the primary interface between the user and the query engine. Its role is to parse and normalize the incoming user query, representing it as an abstract syntax tree (AST) before translating it into a relational algebra tree that is understandable by the subsequent layers.

Moreover, this layer applies logical rewriting rules to simplify the query [24]. These heuristics stretch from semantic query optimization (e.g, removing `DISTINCT` clauses when duplicates are mathematically impossible) and query flattening (e.g, subquery unnesting and `OR` expansion into `UNION`) to rearranging operations in an attempt to reduce the data volume early in the execution pipeline (e.g, predicate pushdown and eager aggregations).

A. Query Representation

Conventional RDBMSs normalize the AST of a query into a relational algebra tree [Fig. 2(e) represents a possible relational algebra tree for the SQL query in Fig. 2(b)], in an attempt to reduce the optimization task to finding the most efficient sequence of binary join operators. Although this approach is structure-independent, these engines often scan for specific patterns, such as star or snowflake schemas. These hard-coded detections allow these systems to trigger specialized transformations to optimize for those specific, pre-defined query shapes.

Despite these targeted heuristics, this model suffers from two major limitations. First, practitioners must manually hard-code detection procedures for every new pattern that emerges and implement unique transformations for each. Second, the fundamental assumption that joins are binary operations remains a bottleneck, where even the successful detection of triangles or cliques in the query structure would still yield a suboptimal query plan. Thus, the primary difficulties in query representation are not inherent from the query language semantics [e.g., it is possible to express complex cyclic patterns in SQL such as Fig. 2(b)], but rather from the internal query representations of RDBMSs, which map these semantics into binary relational algebra trees.

To resolve these limitations, systems like EmptyHeaded [12] move towards using the generalized hypertree decompositions (GHDs) extracted from the hypergraph of the query. For instance, Fig. 2(d) illustrates the hypertree decomposition for the hypergraph in Fig. 2(c) with the cyclic part isolated in a single bag, proving that they must be processed simultaneously. This approach allows the engines to partition the query into a tree of *bags* based on its structure rather than only cardinalities.

In this approach, bags containing cyclic patterns (more than two relations) are evaluated using multi-way WCOJ algorithms, while the rest of bags and intermediate results are joined using traditional binary operators. This ensures structural optimality regardless of whether the query follows a standard star or complex graph patterns.

B. Query Rewriting and Eager Aggregations

Once the relational algebra tree of the query is built, traditional engines rewrite the query using transformations that are known to improve the performance of the query without altering it. Such transformations involve pushing down aggregations and predicates towards the leaf relations to reduce input sizes before joins occur [25].

While early filtering can be highly beneficial [26], as it reduces the active domain for WCOJ iterators [6], the eager pushdown of aggregations may break the pure conjunctive nature of the query on which WCOJs have been formalized [4], forcing the optimizer to fall back to a binary execution plan.

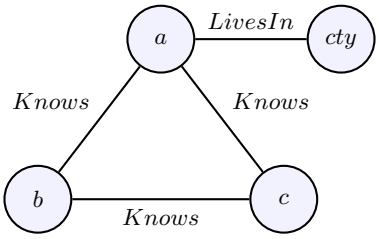
Theoretical frameworks such as AJAR [27], later integrated into LevelHeaded [13], have explored the evaluation of these aggregate-join queries without breaking worst-case optimal bounds. Their approach is to compute partial aggregations on the fly as the engine traverses the GHD of the query. However, this method is strictly constrained. First, the aggregated

```
SELECT a.id, b.id, c.id, cty.name
FROM GRAPH social_graph
MATCH (a:Person)-[:KNOWS]->(b:Person)-[:KNOWS]->(c:Person)-[:KNOWS]->(a),
      (a)-[:LIVES_IN]->(cty:City)
WHERE cty.name = 'Rabat';
```

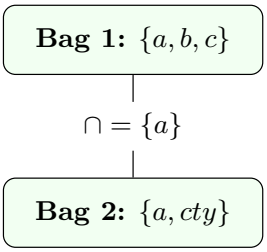
(a)

```
SELECT k1.src, k2.src, k3.src, cty.name
FROM Knows k1
JOIN Knows k2 ON k1.dst = k2.src
JOIN Knows k3 ON k2.dst = k3.src AND k3.dst = k1.src
JOIN LivesIn li ON k1.src = li.person_id
JOIN City cty ON li.city_id = cty.id
WHERE cty.name = 'Rabat';
```

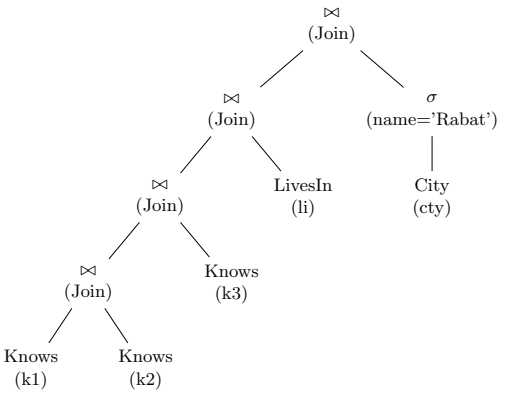
(b)



(c)



(d)



(e)

Fig. 2. (a) A SQL/PGQ query seeking a triangle of friends where one of them lives in Rabat. (b) The equivalent SQL query requiring multiple binary joins. (c) The hypergraph representation of the query. (d) The generalized hypertree decomposition of the hypergraph, isolating the cyclic pattern into the first bag and the acyclic relation into the second bag. (e) An example of the relational algebra tree to process the query.

attribute must be pushed to the end of the global join attribute order. And second, the engine is limited to commutative and iterative functions such as `COUNT` and `SUM`.

C. Functional Limitations

Traditional RDBMSs support multiple join variants, whereas WCOJ research has primarily focused on equi-joins. Supporting complex non-equi predicates (e.g., inequalities, spatial overlaps) typically requires reverting to binary joins or employing specialized geometric adaptations [28], [29]. However, these adaptations cannot be generalized to support all possible use-cases, which are inherited naturally from binary joins.

Moreover, while applying operators like `LIMIT` or `TOP-K` is trivial for the binary approach, it presents a challenge for WCOJs. In WCOJs, the global variable ordering might not align with the `ORDER BY` clause, often forcing the engine to materialize the full result before applying the limit.

Research into combinatorial output-sensitive sampling [30] and top-k algorithms [31] highlights the complexity of *"drilling down"* into a Trie to fetch a specific sample without full traversal. These limitations make WCOJs unfit for some of the most used features in conventional databases, such as pagination. As a result, query engines can always rely on traditional binary joins as a safe fallback strategy when these operations are required.

D. Discussion

Looking at how conventional RDBMSs can support emerging graph workloads via SQL/PGQ, two major paths emerge. The first consists of converting SQL/PGQ queries into standard SQL using a combination of equi-joins, unions, and filters, similar to DuckPGQ [16]. For example, Fig. 2(b) is a possible equivalent SQL query for the SQL/PGQ query in Fig. 2(a). This strategy is suitable for quick syntax support over native integration.

As Large Language Models are increasingly employed to generate queries, how can this translation layer leverage the database schema to contextualize and optimize such queries into efficient standard SQL?

The second path, which aligns with the focus of this survey, requires a generalized query representation. Manually hard-coding detection routines for every new query shape fails to scale. The current consensus leans towards the use of GHDs. However, finding a tree decomposition with the minimum width is a known NP-hard problem.

A critical open challenge revolves around the development of adaptive algorithms, or lightweight learned heuristics, that can approximate near-optimal GHDs at runtime without introducing unacceptable overhead.

Beyond representation, the interpretation layer must bridge the gap between SQL functionalities and the need for

conjunctive queries for WCOJs. Relational heuristics like eager aggregations break this pure conjunctive structure. Similarly, non-equi joins or pagination (`LIMIT`/`TOP-K`) force the engine to revert to binary execution or materialize massive intermediate results.

A few workarounds have already been proposed, such as limiting partial aggregations to commutative functions pushed to the end of the trie. However, the question still stands: *How can WCOJs natively support complex non-equi predicates, and how can engines execute output-sensitive limits without sacrificing early-pruning?*

III. The Optimization layer

The optimization layer identifies the most efficient execution plan from a wide search space of possible plans. In conventional RDBMSs, this is typically achieved through a cost-based approach, which estimates the cost of multiple plans using statistics and data summaries to select the least expensive option.

While binary join optimization has matured from dynamic programming [1] to genetic algorithms [32], WCOJ algorithms require fundamental changes. The cost metric must shift from minimizing intermediate result sizes to minimizing intersection costs, a complex task that would require access to new statistics, a way to define the optimal attribute order, and the ability of hybrid planning (i.e., using both binary and WCOJs in a query plan).

A. The Objective Function

The primary challenge for WCOJ adoption lies in the objective function of the optimizer. While the goal of a conventional optimizer is to arrange joins in a tree structure while minimizing the cardinalities of intermediate results, one that integrates WCOJs must focus on finding the optimal permutation of the join attributes. This ordering should minimize the *width* of the value domains, ensuring that the most selective intersections are performed first to avoid invalid paths earlier.

This divergence in the optimizer goals, from minimizing the data volume (focusing on memory) to minimizing search steps (focusing on computation) creates a conflict. Thus, plans that appear expensive to a traditional optimizer might actually be the most efficient for an optimizer leveraging WCOJs.

To resolve the discrepancy between the two paradigms, the query planner must be refined on all fronts, from the addition of more suitable statistics, the adoption of newer heuristics, to enabling the ability to yield hybrid plans, and with it new unified cost models. These aspects are covered in the rest of this section.

B. Statistics and Cost Estimation

For a cost-based optimizer to make informed decisions, accurate estimations are essential. Traditional engines rely on value histograms and lists of the most common values. However, these scalar (i.e., flat, one-dimensional) statistics fail significantly for cyclic queries because they assume attribute independence, leading to severe underestimations of join sizes.

In contrast, WCOJ performance is heavily tied to the structure of the join query, making existing scalar statistics unreliable. WCOJ optimizers would require *graph statistics*, such as the degrees distributions [33], correlation information, fractional edge covers [7], [4], or graph-specific summaries [34], to map how data points are interconnected. Conventional RDBMSs do not natively support these expensive graph statistics, forcing WCOJ optimizers to operate with low-fidelity estimates.

Other potential techniques for accurate estimation include synopsis-based methods, which build summary structures for cardinality estimations, and sampling methods, which use data samples to estimate cardinalities without execution [34]. However, these approaches are especially hard to adapt for WCOJ, since the probability of capturing cyclic join results in a subset of data is smaller, which may lead to catastrophic underestimations.

C. Hybrid Planning

WCOJs are not substitutes for binary joins. Real-world workloads rarely consist of purely cyclic or acyclic patterns, necessitating query planners capable of yielding hybrid plans that mix between binary and WCOJs [35].

While establishing a unified cost metric remains an active challenge [35], the GHD provides a robust theoretical framework [12]. By decomposing a query into a tree of *bags*, the planner can isolate cyclic components to be resolved via WCOJs, while utilizing traditional binary joins to assemble the final result from these acyclic interconnections. For example, considering the GHD in Fig. 2(d), an optimizer would evaluate Bag 1 (the cyclic triangle) using a multi-way WCOJ, then use a binary join on the shared attribute, a, to attach the acyclic relation in Bag 2.

Free Join [9] pushes this hybridization further to the execution layer. Rather than choosing between distinct operator types, it utilizes lazy hash-based data structures to maintain the high-throughput, scan-centric benefits of traditional binary hash joins. Free Join coordinates these hash probes through a global variable-ordering logic. The algorithm satisfies the AGM-bound while reducing the pointer-heavy memory overhead of traditional WCOJs through its *column-oriented lazy trie*, a unified data structure that builds inner sub-tries as needed.

D. Cost Modeling

To compute the cost of hybrid plans the optimizer needs a unified metric to compare with. In traditional RDBMSs, costs are typically modeled as a function of input and output sizes. This is a result of the assumption that cost is monopolized by intermediate results materialization.

In contrast, WCOJ cost relies heavily on the structure of the query, since its complexity is dictated by the fractional edge cover of the query hypergraph. However, using the AGM-bound as a cost model can be unreliable, since it will often overestimate the actual runtime. This approach can cause the optimizer to reject valid WCOJ plans, as it computes the worst-case cost, whereas in real life workloads, datasets have high skew or sparsity making it orders of magnitude

smaller than estimated. Moreover, if the optimizer relies on a pessimistic estimation, it will frequently perceive WCOJs as prohibitively expensive, defaulting to a binary join tree plan.

To address this, researchers have proposed the *intersection cost* (*i-cost*) as a unified metric for hybrid plans [35]. Unlike the AGM-bound, *i-cost* focuses on the computational effort of the execution rather than the theoretical maximum size of the output. *i-cost* represents the sum of sizes from all adjacency lists accessed during a multi-way intersection. By using degree distributions to estimate the size of these adjacency lists, the optimizer can move away from global pessimistic bounds and toward a local, data-aware estimation.

E. Attribute Ordering

Once the optimizer chooses to use a WCOJ algorithm within a plan, it must determine a suitable attribute order. It is well understood that a poor choice in the determination of the order in which attributes must be processed can severely degrade performance, as the algorithm might keep searching in empty sub-trees rather than pruning early [6], [3]. To this end, several approaches can be employed.

1) Static heuristics: A join can be seen as a constraint satisfaction problem, thus standard techniques may apply. The query optimizer can base its decision on inexpensive heuristics to avoid expensive computations.

For instance, the *"max-degree"* approach prioritizes the most constrained attributes (i.e., attributes shared by most relations). Since each relation limits further the value domain of the attribute, ordering attributes by descending degree often result in early pruning [36], [6].

Moreover, because such heuristics are computationally inexpensive, they can also serve as a fallback strategy for the more advanced strategies discussed below.

2) Cost-based ordering: Static heuristics are fast because they focus solely on query structure, however, they are *instance independent* (i.e., blind to the actual data). Consequently, they may prove inadequate. For instance, a high-degree attribute might have low selectivity (e.g. a boolean value), making it a suboptimal starting attribute.

Cost-based ordering attempts to resolve this by estimating the actual runtime cost of each attribute permutation using statistics. Thus, instead of merely counting the hyperedges to calculate degrees, the optimizer computes an estimated cost for each permutation [35], [12], [37].

This process may leverage existing statistics, such as histograms and information on unique values, to estimate the intersection cost. However, the optimizer must account for search cost. For instance, when the data is dense, sequential scans may be cheaper than leapfrog searches, conversely, when the data is sparse, leapfrog searches are more efficient.

Finally, since cost estimation can be expensive, it cannot be calculated for all possible permutations which is factorial to the number of attributes, the cost-based ordering can use a greedy approach, where it seeks the next cheapest attribute using dynamic programming similar to the ones used by traditional planners but adapted for linear ordering, rather than trees [35].

3) Adaptive and learned optimization: The integration of adaptivity and machine learning into database management systems, particularly at the query optimization layer, has emerged as an effective paradigm [38]. When applying these techniques, research generally divides into two distinct approaches: workload-level learning and query-level adaptivity.

The first approach applies learned optimization at the workload level (*inter-query*), continuously refining the decision-making of the optimizer across successive executions. ReJOOSp [39] and AlphaJoin [40], for example, model optimization as a reinforcement learning problem. By exploring the massive combinatorial space of join orders over time, these agents learn to identify optimal execution paths and structural boundaries that static, rule-based heuristics miss.

The second approach pushes adaptivity directly into the physical execution of a single query (*intra-query*). Rather than committing to a single static plan, the engine initiates execution by running a multitude of distinct plans simultaneously for a brief exploratory period. By leveraging immediate runtime feedback from these short, parallel executions, the optimizer dynamically identifies and transitions to the most efficient plan used for the rest of the query execution. For instance, ADOPT [41] uses this adaptivity to rapidly converge on near-optimal attribute orders for plans using WCOJs.

F. Discussion

When integrating WCOJs within the optimization layer, two primary challenges emerge. First, the query engine requires new generalized cost models. Traditional intermediate cardinality estimations are unsuitable for WCOJs. The current consensus relies on instance-dependent metrics, such as the intersection cost (*i-cost*), and degree distributions to measure the exact computational effort [35].

These new cost models must eventually align with the ongoing shift toward learned query optimization. While machine learning has already proven its ability to capture complex data correlations compared to traditional cardinality estimation, as highlighted in the Cambridge Report [42], extending this approach to reliably predict the sizes of cyclic queries without introducing unacceptable compilation overhead remains an open challenge.

Furthermore, the objective of the optimizer must fundamentally shift from reordering a binary join tree to finding the optimal join attribute order for WCOJ processing. *While static and cost-based heuristics offer practical starting points, dynamically discovering the optimal attribute order at runtime is still an active [41], open research question.*

Second, because WCOJs are not substitutes for binary joins, optimizers must support hybrid planning. The query engine must be capable of yielding hybrid plans that deploy WCOJs when beneficial. To achieve this without relying on complex hypergraph decompositions, modern engines explore alternative heuristics. For instance, systems like Umbra [8] generate a standard binary join plan and combines specific sub-trees into multi-way WCOJ operators based on cardinality estimates. Other approaches, such as Free Join [9], avoid separating operators entirely and deploys a unified data

structure to integrate binary hash-join efficiency and multi-way optimality into a single physical operator.

In this regard, a major open challenge is the development of new approaches for query planners to accurately measure cost and evaluate these hybrid plans, or adaptively trigger multi-way operator combinations at runtime, without suffering from the massive combinatorial search space explosion.

IV. The Execution Layer

The execution layer is where the query engine translates the logical plan from the optimizer into concrete physical operations. Here, the theoretical advantages of WCOJs confront the physical constraints of modern hardware.

While the asymptotic complexity of the algorithm is determined by its logic, the actual runtime performance depends heavily on its implementation and interactions with CPU pipelines, memory hierarchies and storage. This section studies the gap between the mature binary joins, which have been optimized for modern hardware over decades, and the emerging WCOJs.

Traditionally, binary joins in relational engines rely on a scan-centric execution model, either sequentially scanning sorted arrays (sort-merge join), or performing high-speed calculations to probe hash tables (hash join). However, WCOJs rely heavily on a search-centric approach, jumping over large sections of data that cannot participate in the join.

This paradigm is shared by the two primary implementations: the *sort-based* approach, spearheaded by the Leapfrog Triejoin [6], which relies on the global variable order, where instead of scanning every tuple sequentially, its iterators *leapfrog* over non-shared values. And on the other hand, the *hash-based* approach, characterized mainly by Generic Join [43], which avoids global sorting by resolving variable bindings using nested hash lookups. While theoretically superior for cyclic queries, both approaches disrupt the sequential flow that the hardware expects.

A. Computation

The structural shift from scanning to searching in WCOJs creates inefficiencies at the processor level. Modern general-purpose CPUs rely heavily on instruction pipelining and branch prediction to maintain throughput, a design that favors the predictable control flow and sequential access patterns of traditional binary joins [44].

WCOJs are inherently branch-heavy. For example, Leapfrog Triejoin requires a loop of conditional jumps, since it constantly needs to check if any of the involved iterators is less than the current maximum value [`if key < max_key then seek()`]. Because the branch direction depends entirely on the data distribution, this unpredictability triggers frequent pipeline flushes, stalling execution cycles and significantly reducing the instructions-per-cycle compared to binary joins.

Furthermore, the hierarchical nature of WCOJs significantly complicates vectorization. While binary hash joins can leverage flat data structures and SIMD to probe hash tables in batches [45], [46], Trie traversal is data-dependent and difficult to parallelize through a single instruction stream [47]. Although recent work attempts to bridge this gap by mapping joins to matrix multiplication [48], the lack of *vectorized search* primitives remains a significant blocker for adopting WCOJs in vectorized engines. To address this, systems like EmptyHeaded [12] adaptively encode Trie nodes based on the data sparsity, translating complex pointer traversals into parallelizable SIMD set intersections.

B. Primary Memory

The performance of in-memory query processing is often dominated by memory latency rather than CPU throughput [49]. Thus, even if the CPU pipeline is perfectly saturated, the runtime performance would likely suffer waiting for data to be fetched. In this regard, the result of the structural differences between binary and WCOJs becomes more severe.

Typically, binary joins process sequential memory pages, a pattern that aligns with hardware prefetchers loading data into the L1 and L2 caches before it is requested. In contrast, WCOJ requires navigating Tries, which involves pointer chasing, following references from parent to child nodes that are scattered over different memory pages and only known after the parent is retrieved. This serial dependency defeats hardware prefetching, leading to a higher rate of cache misses. As a result, the WCOJ algorithms often spend more time waiting for main memory than performing intersections [50].

C. Hierarchical Indexing

WCOJ algorithms require prefix trees (e.g. Tries) during the intersections processing. Such uncommon indices are not supported in conventional RDBMSs. This creates a trade-off between building indices Just-In-Time (JIT) or maintaining persistent indices.

In the case of JIT index building, as implemented in DuckDB [18] and Umbra [8], the benefit of worst-case optimality must outweigh the overhead of building the index. For *one-time* analytical queries, the construction cost can dominate the total execution time, potentially making a binary plan faster despite its theoretical inferiority.

On the other hand, LogicBlox [10], resolve to maintaining persistent Trie indices as their primary storage. While this eliminates the JIT build cost, it shifts the burden to data ingestion and updates, but is totally acceptable for read-only engines or with limited updates, or systems employing a sixth normal form (6NF) data architecture [6].

Middle ground between these two approaches may be found by maintaining light indices (mostly flat) and building the more expensive indices during query execution [51]. The border between these two approaches is a gray area, for instance, Umbra maintains flat data structure for a faster building stage of their Tries [8]. Recent efforts work on making Tries, more update-friendly [52].

D. Secondary Memory

Traditional RDBMS are required to be able to handle datasets, even when their sizes exceeds main memory capacity. For binary joins, this problem has been solved via GRACE

Hash Join [53] for instance, which partitions inputs into disk-resident buckets that are processed sequentially.

However, spilling to the disk is far more penalized in WCOJs. As discussed previously, they rely on random accesses, which is orders of magnitude slower than sequential reads in disk, even on NVMe SSDs. Moreover, theoretical I/O-optimal algorithms for cyclic joins remain an open problem [54].

Partial solutions to these challenges, however, have seen the day. Approaches like Quadtrees [55] attempt to decouple the topology from the data values, storing the structure in compact bitmaps. More advanced systems, such as the Ring [11] and CompactLTJ [21], go further by enabling execution directly on the compressed representation. This "Computation over Compression" approach effectively decreases memory footprint, partially mitigating the main memory issues described above, and attempts to avoid the need to spill to disk.

Umbra [8] uses LeanStore's [56] pointer *swizzling* technique, and leverages its architecture designed around flash-based storage, effectively using disk as main memory while introducing minimal overhead.

E. Workload-Aware Execution

In real-world scenarios, queries rarely run in isolation. Workload-aware techniques exploit overlapping patterns across such queries, which can improve the overall workload runtime performance dramatically.

During a workload, parts of the joins may be shared across multiple queries, the engine might choose to cache the results of those common parts for future use, which would cut the total execution time depending on their occurrence and size. However, WCOJs avoid to materialize intermediate results, which prevents them from being cached.

Attempts in this regard include Flexible Caching [57], where instead of materializing full intermediate results, it caches partial results at carefully chosen bags from the tree decomposition of the query's hypergraph. This enables reusing repeated sub-joins when beneficial while still retaining the advantages of WCOJs.

F. Discussion

As explored in this section, while WCOJs successfully bound the asymptotic complexity of cyclic queries, they confront the physical limitations of modern hardware. The execution of traditional binary joins aligns with CPU pipelining and memory prefetching. In contrast, WCOJs reliance on unpredictable pointer-chasing during Trie traversals breaks this alignment. This mismatch between algorithmic design and hardware constraints paves the way for a multitude of open challenges and research questions.

One of the main research paths within the execution layer is hardware acceleration. While the majority of existing literature focuses on multi-core CPUs, leveraging the low latency to manage complex and branch-heavy workloads, there is a critical need to explore GPU offloading to leverage massive data parallelism and high throughput during join processing.

Furthermore, FPGAs present another interesting alternative, offering custom, high-throughput data paths alongside superior energy efficiency. FPGAs can especially be versatile when handling WCOJs. They can be deployed in-line between the CPU and disk to accelerate decompression and data filtering, placed on the same socket as the CPU to offload specific computational tasks, or utilized as external, high-throughput units similar to GPUs.

Beyond designing WCOJ variants tailored to these specific hardware topologies, a critical research trajectory is the development of a heterogeneous WCOJ algorithm, capable of dynamically routing specific sub-tasks to the most suitable compute unit.

The optimal use of these compute units, and particularly their vector processing capabilities (SIMD/SIMT), is dependent on the underlying index structures, which must be redesigned to be vector-friendly, such as sparsity-aware layouts that map intersections directly to bitwise instructions as proposed by EmptyHeaded [12]. However, while transitioning to such layouts unlocks hardware throughput, it introduces new complexities regarding mutability and state management across devices.

These structures are optimized for read-heavy workloads where data density is predetermined. In dynamic database environments with frequent insertions, continuously monitoring data sparsity and rebuilding or transferring these index layouts between the CPU, GPU, and FPGA imposes massive synchronization and computational overheads, becoming a bottleneck to the very execution pipelines they are meant to accelerate.

One of the possible solutions is to map these joins into matrix multiplications, for which GPUs are optimized. However, this approach is heavily affected by data sparsity. Thus, it is essential to define thresholds at which mutable relational data should transition between a sparse array and a dense SIMD-friendly bitset at runtime. Another interesting research path in this regard, is determining cost models to dynamically route WCOJ parts among available compute units (e.g., dispatching dense, matrix-mapped joins to GPUs while keeping sparse, pointer-heavy intersections on CPU).

Beyond compute orchestration, multi-way intersections suffer from severe limitations across the memory and storage hierarchy. In main memory, the massive bandwidth required to continuously feed the accelerators creates a *"memory wall"* bottleneck. Furthermore, as massive workloads are inevitably pushed out-of-core, WCOJs face critical I/O challenges. Unlike traditional binary joins that adopt well-established, sequential disk-spilling mechanisms, WCOJs depend on traversing deeply nested index structures. Spilling these complex structures to secondary storage (NVMe/SSDs) results in severe I/O thrashing.

The primary open challenge here consists of bridging between byte-addressable main memory and block-addressable secondary storage during multi-way joins. While state-of-the-art systems like Umbra [8] have proposed low-overhead out-of-core execution using their pointer swizzling technique, where they translate between memory addresses and disk offsets, this only reduces translation costs. They do not resolve the fundamental I/O bottleneck caused by

WCOJ access patterns. Because attribute-at-a-time evaluation requires traversing hierarchical index structures across all participating relations, the unpredictable pointer-chasing disrupts the block-oriented access required for efficient secondary storage. Designing an architecture capable of prefetching massive, sparse index sub-trees without major overhead remains a critical unsolved challenge.

V. The Distributed Layer

While distributed RDBMSs deserve a survey of their own, they share the issue of limited WCOJ adoption. Concepts within the interpretation layer can be shared among shared-memory and distributed RDBMSs. However the difference shows up when it comes to the optimization and execution layers. This is the result of the core shift in objectives.

While shared-memory RDBMSs focus on minimizing the memory latency, the distributed ones focus on the network bandwidth. A much stricter constraint, since data movement through the network can be slower by multiple orders of magnitude. Thus, distributed RDBMSs require a complete rethinking of how WCOJs are processed at scale [3].

Traditionally, distributed binary joins rely on hash or range partitioning. For example, in the case for the hash approach, to join $R(a, b)$ and $S(b, c)$, the engine hashes on the join key b on both relations, and sends each tuple to the adequate node (e.g. `hash(b) % #nodes`), which guarantees that all matching tuples will land on the same node, and proceeds with local joins independently [58].

However, for WCOJs, this hashing technique would fail. For instance, in the triangle query, $Q = R(a, b) \bowtie S(b, c) \bowtie T(a, c)$, hashing on attribute b would only bring relations R and S together but not T (since it has no attribute b). To resolve this issue without losing worst-case optimality, the engine needs to adopt a higher-dimensional distribution approach.

A. The HyperCube Shuffle

The standard solution for distributed multi-way joins is *HyperCube Partitioning* [58]. The cluster is organized into a virtual high-dimensional grid, as illustrated in Fig. 3. For the triangle query involving variables a, b, c, the P processors are arranged into a cube of dimensions such as $P_a \times P_b \times P_c = P$.

Each relation is broadcast along the dimensions it misses. For example, a tuple $R(a, b)$ is hashed to coordinates $(hash(a), hash(b))$. Since it lacks variable c, it must be sent to all nodes corresponding to $(hash(a), hash(b), *)$. Effectively, it is replicated P_c times.

This approach guarantees that for any possible triangle (a, b, c), all participating tuples from R, S, T will meet at exactly one unique intersection point in the grid. However, this introduces a massive *replication cost*. Since unlike binary shuffles which move data once ($O(N)$), HyperCube shuffles multiply the data volume ($O(N \times P^{1-1/\rho^*})$) [59]. Thus, HyperCube shuffling would only work if the gains from the WCOJ outweigh the losses incurred by shuffling through the network.

B. Handling Data Skew

One of the major challenges that distributed RDBMSs have to face is data skew, and in WCOJs, skew can be much harder to handle since it is multi-dimensional. The complexity is no longer because of frequent key values, but also of the structure, resulting in large replications.

Multiple unique issues can arise from data skew in distributed WCOJs. For instance, in HyperCube shuffling, if a single node is assigned a dense region of the hypergraph, it might run out of memory, or have the bulk of the work to process (more than all other nodes), thus delaying the entire query.

To resolve these issues, hybrid processing is normally relied on, where high-degree vertices are treated differently, often being broadcast to all nodes to parallelize their intersection workload, while low-degree vertices are hashed normally. This results in a hybrid approach that dynamically adapts to the structure of the query and the frequency of values [60].

While Myria [61], [3] successfully proved that the communication bounds of hypercube partitioning are optimal (it transfers less data across the cluster than traditional binary hash joins), it was frequently bottlenecked by memory due to the heavy replication. To resolve this, subsequent approaches such as BiGJoin [62] leverage the Timely Dataflow computation model [63] to stream-process Generic Join [7].

By pairing the continuous execution with on-the-fly skew handling, the system maintains a low memory footprint throughout the entire query execution. Furthermore, modern implementations like EdgeFrame [64] abandon the hypercube shuffle entirely. Instead, they broadcast heavily compressed graphs directly to worker nodes, thus trading main memory capacity for a WCOJ execution without the need for shuffling.

C. Network Optimization

Given the high cost of replication, modern efforts focus on reducing the payload before the communication happens. One of such approaches develop on similar techniques from distributed binary joins and use Bloom filters [65]. Before the massive shuffle, nodes only exchange tuples that *might* be in the final result. Tuples that do not make the cut are thus discarded locally, preventing them from consuming valuable bandwidth. This is especially effective when the joins are highly filtered which should be the case for cyclic queries more often than not.

Other approaches can focus on the data format itself. For example, *Adaptive Factorization* [18] would be suitable for such an environment, sending data in a compressed form, where instead of sending redundant tuples $(a, b_1), (a, b_2), \ldots$, the system sends a grouped representation $a \rightarrow \{b_1, b_2, \ldots\}$. This *structural compression* can significantly reduce the serialization overhead during the HyperCube shuffle, directly addressing the replication penalty.

D. The Multi-Round Shuffle

While the HyperCube shuffle minimizes latency by completing the join in a single communication phase, it

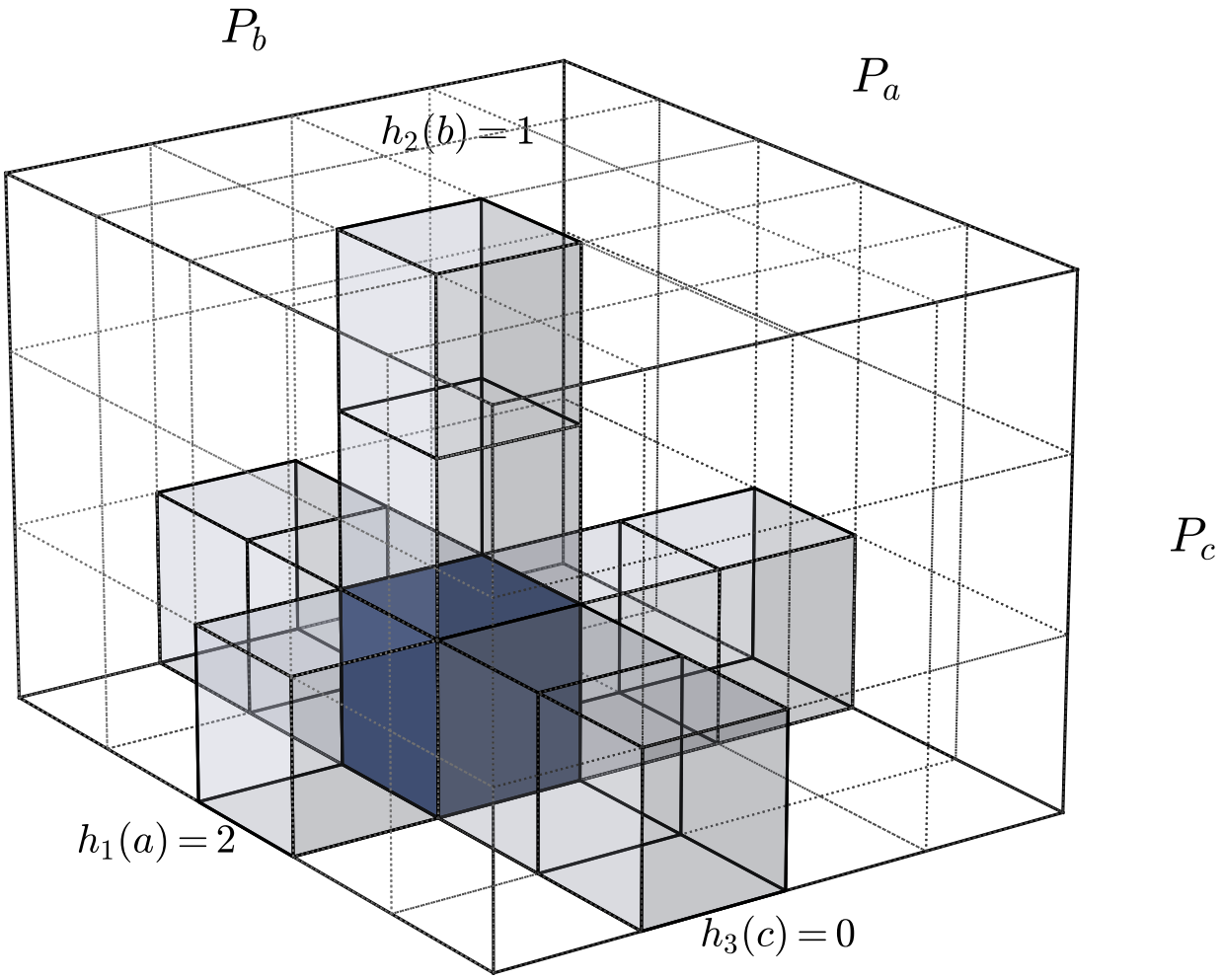


Fig. 3. An illustration of the HyperCube shuffle organizing a cluster into a virtual 3D grid with dimensions $P_a \times P_b \times P_c$ for a triangle query. A tuple lacking a variable is hashed to its known coordinates, such as $h_1(a) = 2$ and $h_2(b) = 1$, and replicated across the missing P_c dimension. This approach guarantees that all participating tuples from the queried relations will meet at exactly one unique intersection point in the grid to evaluate the join.

requires a high bandwidth cost due to the massive data movement and replication. To resolve this issue, especially for environments with limited bandwidth or even for complex queries for which HyperCube shuffling would require excessive duplication, multi-round approaches offer a possible alternative.

As formalized by Ketsman-Suciu [66], single-round algorithms are often bounded by the fractional vertex cover of the query hypergraph. By allowing the computation to proceed in $O(\log P)$ rounds, the maximum load per worker can be reduced to $O(m/P^{1/\rho^*})$, where m is the size of the largest input relation. This bound is asymptotically superior for many cyclic queries, thus trading rounds for memory and bandwidth efficiency.

Multi-round algorithms typically decompose the query execution into multiple iterative steps [67]. Instead of resolving the entire cyclic structure in one massive shuffle, the system may perform a sequence of semi-join reductions or compute partial intersections (e.g., joining subsets of relations to limit the search space) before reshuffling the reduced intermediate results for the final join. This removes non-joining tuples early, and prevents them from being replicated across the cluster needlessly.

While this approach might significantly reduce the total volume of data moved across the network, minimizing the risk of congestion, it introduces multiple synchronization barriers, which in the distributed setting, it requires the entire cluster to wait for the slowest node to complete the current round before proceeding to the next one. For short-running or interactive queries, the accumulated tail latency of multiple barriers often outweighs the bandwidth savings, making single-round approaches preferable despite their higher data cost.

E. Discussion

Within distributed environments, the primary performance constraint shifts from main memory latency to network bandwidth. Because one-dimensional hash partitioning fails to colocate cyclic query structures, the current consensus

relies on HyperCube shuffling to route multi-dimensional data. To prevent the resulting data replication from entirely exhausting network bandwidth, distributed engines heavily rely on payload reduction techniques prior to communication. This includes applying pre-shuffle Bloom filters to discard unneeded tuples [65] and using adaptive factorization to compress the remaining payload during transit [18].

As the database industry increasingly adopts cloud environments with on-demand node allocation, a major research opportunity is the design of elastic execution engines. In such systems, cluster expansion (or shrinkage) could be triggered during runtime. For instance, the query optimizer might choose to scale resources mid-query, or this decision could be pushed directly to the execution layer, where early data scans can provide instance-dependent information to achieve an optimal balance of performance and cost.

This paradigm shift towards adaptive systems extends far beyond dynamic cluster sizing. This broader scope of runtime adaptivity presents multiple interesting research opportunities, particularly regarding how query engines manage data preparation for transit and algorithmic choice.

For instance, adaptivity can be integrated directly into the filtering and compression stages. Rather than universally applying these reduction techniques, a system could dynamically determine their necessity based on runtime data characteristics. For example, a small dataset might be shuffled in its entirety, intentionally avoiding the computational overhead associated with semi-join filtering or structural compression.

Furthermore, another critical research direction involves enabling the optimizer to dynamically choose between single-round and multi-round execution. Implementing this effectively would require the development of robust cost models and runtime heuristics capable of accurately comparing the bandwidth costs of a single-round HyperCube shuffle against the accumulated synchronization latencies found in the multi-round approach.

VI. Related Work

Worst-case optimal joins are indeed a *hot topic* in database research. However, the broader shift to handle the ever evolving volume and complexity of modern data workloads [42] must be recognized.

While this survey specifically focuses on the practical aspects of integrating WCOJs into mature RDBMSs, it builds upon a large foundation of prior literature. An extensive body of surveys, authored by seminal figures in the field, covers the underlying formal theory and the proof frameworks for worst-case optimality in detail [23], [7]. Furthermore, this analysis complements a previous work [5] that studies the algorithmic aspects of this class of multi-way joins.

Beyond the theoretical and algorithmic foundations, the integration of artificial intelligence and machine learning has emerged as a transformative paradigm across computer science, and database research is no exception.

At the interpretation layer, this AI-driven evolution is most manifested in Natural Language Interfaces to Databases (NLIDB). Specifically, Text-to-SQL generation has transitioned from rule-based semantic parsing to leveraging deep learning and Large Language Models (LLMs). Comprehensive surveys [68] have studied the different models, and the upcoming challenges.

Moreover, modern benchmarks like Spider [69] and BIRD [70] demonstrate the increasing capacity of LLMs to autonomously navigate complex, cross-domain database schemas [71]. Such advancements are deeply connected with the necessity for WCOJs, as AI lowers the barrier for end-users to generate highly complex, multi-way join queries.

The optimization layer is undergoing a similar transformation. Traditional cost-based optimizers are increasingly being improved or replaced by learned cost models [42]. A critical driver of this shift is the rapid advancement in cardinality estimation [72], [73], leveraging deep neural networks and advanced statistical models capable of capturing high-dimensional correlations within data distributions. This is directly relevant to the adoption of WCOJs, as the multi-dimensional skew and structural correlations inherent to cyclic graph queries have historically caused severe underestimations in traditional approaches.

Beyond estimations, adaptive techniques and reinforcement learning are being deployed to refine statistics or dynamically adjust execution plans during running queries. While these methods do not always outperform mature deterministic algorithms across all scenarios, they can offer huge performance improvements in complex, highly correlated workloads [74], [75], [42].

The efficacy of these adaptive and learned optimizers is even more necessary because of the modern cloud environments [76]. The database industry has widely adopted cloud-native architectures that are decoupled from storage and compute, enabling an unprecedented degree of scalability and flexibility [77]. For the evaluation of WCOJs, this elasticity can be transformative, since such systems can adopt optimizers that anticipate the explosive cyclic queries that will induce processing spikes, and provision compute resources on demand. Previous work [78] addressed handling data skew in a graph query within such an elastic environment, which showcases the little overhead paid for the possible large gains.

A prominent research area in modern database systems is hardware acceleration. To overcome the physical execution bottlenecks of increasingly complex workloads, the database community is actively exploring hardware-software co-design. As detailed in comprehensive surveys across the domain, this research typically focuses on three paradigms: the design of special-purpose CPUs and advanced vectorization techniques, the utilization of Graphics Processing Units (GPUs) to exploit massive thread-level concurrency [79], [80], and the development of custom data pipelines using Field-Programmable Gate Arrays (FPGAs) to accelerate all or distinct parts of the query execution engine [81].

VII. Conclusion

This survey explored the practical considerations for the adoption of worst-case optimal joins, specifically focusing on conventional relational database management systems. This

TABLE II. SUMMARY OF INTEGRATION CHALLENGES AND PROPOSED SOLUTIONS FOR WCOJS ACROSS DBMS LAYERS

Layers	Integration Challenges	Proposed Solutions
Interpretation	Representing queries in relational algebra trees requires hard-coded procedures to detect cyclic query patterns.	Using generalized hypertree decompositions (GHDs) of the query hypergraph to isolate cyclic bags [12], [35].
	Eager aggregation breaks the pure conjunctive query structure required by the AGM-bound.	Evaluating aggregate-join queries via on-the-fly partial aggregations restricted to iterative functions [27], [13].
	WCOJs do not natively support common functionalities such as non-equi joins and `TOP-K` clauses.	Employing geometric adaptations for non-equi predicates [28], [29] and combinatorial sampling for output limits [30], [31].
Optimization	Relying on scalar statistics that assume attribute independence leads to cardinality underestimations for cyclic joins.	Introducing graph statistics such as degree distributions [33] and synopsis-based estimations [34].
	Inability of query planners to yield hybrid execution plans that combine traditional binary joins with WCOJs.	Utilizing GHDs to isolate multi-way bags [12], or deploying unified data structures that fuse binary and multi-way execution [9].
	Using the theoretical AGM-bound as a cost metric, which overestimates actual costs and causes optimizers to reject valid WCOJ plans.	Adopting data-aware metrics like the intersection cost (i-cost) to measure computational effort rather than theoretical output size [35].
	Suboptimal join attribute ordering degrades WCOJ performance.	Applying max-degree static heuristics [6], cost-based greedy ordering [35], or reinforcement learning [39], [41].
Execution	Branch-heavy, search-centric Trie traversals causes unpredictable control flows that stall CPU pipelines and hinder vectorization.	Translating pointer traversals into parallelizable SIMD set intersections [12] and mapping joins to matrix multiplication [48].
	The severe trade-off between expensive Just-In-Time (JIT) index construction overhead and the massive storage footprint of persistent Tries.	Employing 6NF data architectures [6], maintaining lightweight flat indices for fast JIT construction [8], [51], or utilizing update-friendly Tries [52].
	Random access patterns of WCOJ iterators severely penalizing performance when datasets exceed main memory and spill to disk.	Operating directly on compressed structures like Quadtrees [55] and Wavelet Trees [11], [21], or leveraging pointer swizzling [8], [56].
	The inherent avoidance of intermediate result materialization in WCOJs preventing traditional workload-aware caching mechanisms.	Implementing flexible caching of partial results at specific GHD bags to reuse repeated sub-joins [57].
Distributed	Traditional one-dimensional hash partitioning fails to colocate the complete structure of cyclic graph queries.	Employing HyperCube partitioning to enable multi-dimensional data distribution [58].
	Severe multi-dimensional data skew exhausting individual node memory and delaying query execution.	Implementing hybrid vertex broadcasts [60], stream-processing via Timely Dataflow [62], [63], or broadcasting compressed graphs [64].
	Massive data replication inherent to HyperCube shuffles bottlenecking network bandwidth.	Filtering unneeded tuples via pre-shuffle Bloom filters [65] and applying Adaptive Factorization for structural compression [18].
	Accumulated tail latency from synchronization barriers in multi-round execution models.	Bounding computation to logarithmic rounds [66] or performing sequences of semi-join reductions prior to the final shuffle [67].

analysis is organized using a simplified query engine, inspired by PostgreSQL, to evaluate the integration attempts and their limitations at each layer. And with the growing shift to cloud computing, the adoption of WCOJs within distributed environments is also examined.

This study affirms that despite the theoretical guarantees, WCOJs integration is constrained by decades of optimization designed around binary joins. Nevertheless, their practical value is undeniable, as demonstrated by the emergence of modern, specialized systems built natively around them.

DBMS development is a complicated and slow process where correctness, completeness and stability generally take precedence over performance optimization. Often, new workloads can be accommodated via translation layers rather than a complete engine rewrite. For example, DuckPGQ successfully maps SQL/PGQ graph queries into SQL, thus removing the need for native execution. On the other hand, a rapid adoption of vector databases has been observed in recent years, because vector similarity workloads cannot be substituted by traditional relational operations, the industry was forced to adopt native solutions almost instantly.

Historically, theoretical superiority has not guaranteed practical adoption, as happened with Yannakakis algorithm for instance. However, the trajectory of WCOJs appears more promising. The sustained volume of research published across major database venues over the past decade, coupled with proven practical performance in specific workloads, suggests a slow but steady shift toward broader mainstream acceptance.

Ultimately, this survey aims to successfully cover this interesting research path, and act as an accessible entry point for newcomers and for researchers that are looking for new challenges. Table II provides a clear summary of engineering trade-offs for practitioners attempting to integrate WCOJs in their systems.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers and the editors of the journal for their valuable feedback and constructive suggestions, which significantly contributed to improve the clarity and overall quality of this manuscript.

DECLARATION ON GENERATIVE AI

During the preparation of this work the author(s) used Google Gemini in order to refine and reformulate the original

drafts, reorganize the text and improve clarity and readability of the content. 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.

REFERENCES

[1] P. G. Selinger, M. M. Astrahan, D. D. Chamberlin, R. A. Lorie, and T. G. Price, "Access path selection in a relational database management system," in *Proceedings of the 1979 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '79. New York, NY, USA: Association for Computing Machinery, 1979, p. 23–34. [Online]. Available: https://doi.org/10.1145/582095.582099

[2] P. Fuchs, P. Boncz, and B. Ghit, "Edgeframe: Worst-case optimal joins for graph-pattern matching in spark," in *Proceedings of the 3rd Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA)*, ser. GRADES-NDA'20. New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: https://doi.org/10.1145/3398682.3399162

[3] S. Chu, M. Balazinska, and D. Suciu, "From theory to practice: Efficient join query evaluation in a parallel database system," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 63–78. [Online]. Available: https://doi.org/10.1145/2723372.2750545

[4] A. Atserias, M. Grohe, and D. Marx, "Size bounds and query plans for relational joins," in *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, 2008, pp. 739–748.

[5] A. Berdai and D. Chiadmi, "Attempts in worst-case optimal joins on relational data systems: A literature survey," in *2023 IEEE 6th International Conference on Cloud Computing and Artificial Intelligence: Technologies and Applications (CloudTech)*, 2023, pp. 01–08.

[6] T. L. Veldhuizen, "Leapfrog triejoin: a worst-case optimal join algorithm," 2013. [Online]. Available: https://arxiv.org/abs/1210.0481

[7] H. Q. Ngo, C. Ré, and A. Rudra, "Skew strikes back: new developments in the theory of join algorithms," *SIGMOD Rec.*, vol. 42, no. 4, p. 5–16, Feb. 2014. [Online]. Available: https://doi.org/10.1145/2590989.2590991

[8] M. Freitag, M. Bandle, T. Schmidt, A. Kemper, and T. Neumann, "Adopting worst-case optimal joins in relational database systems," *Proc. VLDB Endow.*, vol. 13, no. 12, p. 1891–1904, Jul. 2020. [Online]. Available: https://doi.org/10.14778/3407790.3407797

[9] Y. R. Wang, M. Willsey, and D. Suciu, "Free join: Unifying worst-case optimal and traditional joins," *Proc. ACM Manag. Data*, vol. 1, no. 2, Jun. 2023. [Online]. Available: https://doi.org/10.1145/3589295

[10] M. Aref, B. ten Cate, T. J. Green, B. Kimelfeld, D. Olteanu, E. Pasalic, T. L. Veldhuizen, and G. Washburn, "Design and implementation of the logicblox system," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 1371–1382. [Online]. Available: https://doi.org/10.1145/2723372.2742796

[11] D. Arroyuelo, A. Gómez-Brandón, A. Hogan, G. Navarro, J. Reutter, J. Rojas-Ledesma, and A. Soto, "The ring: Worst-case optimal joins in graph databases using (almost) no extra space," *ACM Trans. Database Syst.*, vol. 49, no. 2, Mar. 2024. [Online]. Available: https://doi.org/10.1145/3644824

[12] C. R. Aberger, A. Lamb, S. Tu, A. Nötzli, K. Olukotun, and C. Ré, "Emptyheaded: A relational engine for graph processing," *ACM Trans. Database Syst.*, vol. 42, no. 4, Oct. 2017. [Online]. Available: https://doi.org/10.1145/3129246

[13] C. Aberger, A. Lamb, K. Olukotun, and C. Re, "Levelheaded: A unified engine for business intelligence and linear algebra querying," in *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, 2018, pp. 449–460.

[14] X. Feng, G. Jin, Z. Chen, C. Liu, and S. Salihoglu, "KÙZU graph database management system," in *Proceedings of the 13th Conference on Innovative Data Systems Research (CIDR)*, Amsterdam, The Netherlands, 2023. [Online]. Available: https://www.vldb.org/cidrdb/2023/p29-feng.pdf

[15] C. Kankanamge, S. Sahu, A. Mhedbhi, J. Chen, and S. Salihoglu, "Graphflow: An active graph database," in *Proceedings of the 2017 ACM International Conference on Management of Data*, ser. SIGMOD '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 1695–1698. [Online]. Available: https://doi.org/10.1145/3035918.3056445

[16] D. ten Wolde, T. Singh, G. Szárnyas, and P. Boncz, "Duckpgq: Efficient property graph queries in an analytical RDBMS," in *13th Conference on Innovative Data Systems Research, CIDR 2023, Amsterdam, The Netherlands, January 8-11, 2023*. www.cidrdb.org, 2023. [Online]. Available: https://vldb.org/cidrdb/2023/duckpgq-efficient-property-graph-queries-in-an-analytical-rdbms.html

[17] M. Raasveldt and H. Mühleisen, "Duckdb: an embeddable analytical database," in *Proceedings of the 2019 International Conference on Management of Data*, ser. SIGMOD '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1981–1984. [Online]. Available: https://doi.org/10.1145/3299869.3320212

[18] P. Groß, D. ten Wolde, and P. Boncz, "Adaptive factorization using linear-chained hash tables," in *Proceedings of the 15th Annual Conference on Innovative Data Systems Research (CIDR 2025)*, Amsterdam, The Netherlands, Jan. 2025, january 19–22, 2025. [Online]. Available: https://vldb.org/cidrdb/papers/2025/p21-gro.pdf

[19] D. Vrgoc, C. Rojas, R. Angles, M. Arenas, D. Arroyuelo, C. B. Aranda, A. Hogan, G. Navarro, C. Riveros, and J. Romero, "Millenniumdb: A persistent, open-source, graph database," 2021. [Online]. Available: https://arxiv.org/abs/2111.01540

[20] D. Arroyuelo, D. Campos, A. Gómez-Brandón, Y. Linker, G. Navarro, C. Rojas, and D. Vrgoč, "Compactltj: Space & time efficient leapfrog triejoin on graph databases: Compactltj: Space & time efficient leapfrog triejoin ..." *The VLDB Journal*, vol. 34, no. 6, Sep. 2025. [Online]. Available: https://doi.org/10.1007/s00778-025-00945-5

[21] D. Arroyuelo, D. Campos, A. Gómez-Brandón, G. Navarro, C. Rojas, and D. Vrgoč, "Space & time efficient leapfrog triejoin," in *Proceedings of the 7th Joint Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA)*, ser. GRADES-NDA '24. New York, NY, USA: Association for Computing Machinery, 2024. [Online]. Available: https://doi.org/10.1145/3661304.3661898

[22] "Information technology — Database languages — SQL — Part 16: Property Graph Queries (SQL/PGQ)," International Organization for Standardization, Geneva, CH, Standard, June 2023.

[23] H. Q. Ngo, "Worst-case optimal join algorithms: Techniques, results, and open problems," 2018. [Online]. Available: https://arxiv.org/abs/1803.09930

[24] H. Pirahesh, J. M. Hellerstein, and W. Hasan, "Extensible/rule based query rewrite optimization in starburst," in *Proceedings of the 1992 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '92. New York, NY, USA: Association for Computing Machinery, 1992, p. 39–48. [Online]. Available: https://doi.org/10.1145/130283.130294

[25] W. P. Yan and P.-r. Larson, "Eager aggregation and lazy aggregation," in *Proceedings of the 21th International Conference on Very Large Data Bases*, ser. VLDB '95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, p. 345–357.

[26] Y. Yang, H. Zhao, X. Yu, and P. Koutris, "Predicate transfer: Efficient pre-filtering on multi-join queries," in *14th Conference on Innovative Data Systems Research, CIDR 2024, Chaminade, HI, USA, January 14-17, 2024*. www.cidrdb.org, 2024. [Online]. Available: https://vldb.org/cidrdb/2024/predicate-transfer-efficient-pre-filtering-on-multi-join-queries.html

[27] M. R. Joglekar, R. Puttagunta, and C. Ré, "Ajar: Aggregations and joins over annotated relations," in *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, ser. PODS '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 91–106. [Online]. Available: https://doi.org/10.1145/2902251.2902293

[28] M. A. Khamis, H. Q. Ngo, C. Ré, and A. Rudra, "Joins via geometric resolutions: Worst case and beyond," *ACM Trans. Database Syst.*, vol. 41, no. 4, Nov. 2016. [Online]. Available: https://doi.org/10.1145/2967101

[29] M. A. Khamis, H. Q. Ngo, and A. Rudra, "FAQ: questions asked frequently," *CoRR*, vol. abs/1504.04044, 2015. [Online]. Available: http://arxiv.org/abs/1504.04044

[30] S. Deng, S. Lu, and Y. Tao, "On join sampling and the hardness of combinatorial output-sensitive join algorithms," in *Proceedings of the 42nd ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, ser. PODS '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 99–111. [Online]. Available: https://doi.org/10.1145/3584372.3588666

[31] N. Tziavelis, W. Gatterbauer, and M. Riedewald, "Optimal join algorithms meet top-k," in *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 2659–2665. [Online]. Available: https://doi.org/10.1145/3318464.3383132

[32] The PostgreSQL Global Development Group, *Genetic Query Optimization (GEQO) in PostgreSQL*, PostgreSQL Global Development Group, 2025, section 61.3 of PostgreSQL 18 Documentation. [Online]. Available: https://www.postgresql.org/docs/18/geqo-pg-intro.html

[33] M. Joglekar and C. Ré, "It's all a matter of degree: Using degree information to optimize multiway joins," *CoRR*, vol. abs/1508.01239, 2015. [Online]. Available: http://arxiv.org/abs/1508.01239

[34] K. Kim, H. Kim, G. Fletcher, and W.-S. Han, "Combining sampling and synopses with worst-case optimal runtime and quality guarantees for graph pattern cardinality estimation," in *Proceedings of the 2021 International Conference on Management of Data*, ser. SIGMOD '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 964–976. [Online]. Available: https://doi.org/10.1145/3448016.3457246

[35] A. Mhedhbi and S. Salihoglu, "Optimizing subgraph queries by combining binary and worst-case optimal joins," *Proc. VLDB Endow.*, vol. 12, no. 11, p. 1692–1704, Jul. 2019. [Online]. Available: https://doi.org/10.14778/3342263.3342643

[36] R. Dechter and I. Meiri, "Experimental evaluation of preprocessing techniques in constraint satisfaction problems," in *Proceedings of the 11th International Joint Conference on Artificial Intelligence - Volume 1*, ser. IJCAI'89. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1989, p. 271–277.

[37] V. Leis, A. Gubichev, A. Mirchev, P. Boncz, A. Kemper, and T. Neumann, "How good are query optimizers, really?" *Proc. VLDB Endow.*, vol. 9, no. 3, p. 204–215, Nov. 2015. [Online]. Available: https://doi.org/10.14778/2850583.2850594

[38] R. Zhu, L. Weng, B. Ding, and J. Zhou, "Learned query optimizer: What is new and what is next," in *Companion of the 2024 International Conference on Management of Data*, ser. SIGMOD '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 561–569. [Online]. Available: https://doi.org/10.1145/3626246.3654692

[39] B. Warnke, K. Martens, T. Winker, S. Groppe, J. Groppe, P. Adhiyaman, S. Srinivasan, and S. Krishnakumar, "Rejoosp: Reinforcement learning for join order optimization in sparql," *Big Data and Cognitive Computing*, vol. 8, no. 7, 2024. [Online]. Available: https://www.mdpi.com/2504-2289/8/7/71

[40] J. Zhang, "Alphajoin: Join order selection à la alphago," in *Proceedings of the VLDB 2020 PhD Workshop co-located with the 46th International Conference on Very Large Databases (VLDB 2020), ONLINE, August 31 - September 4, 2020*, ser. CEUR Workshop Proceedings, Z. Abedjan and K. Hose, Eds., vol. 2652. CEUR-WS.org, 2020. [Online]. Available: https://ceur-ws.org/Vol-2652/paper05.pdf

[41] J. Wang, I. Trummer, A. Kara, and D. Olteanu, "Adopt: Adaptively optimizing attribute orders for worst-case optimal join algorithms via reinforcement learning," 2023. [Online]. Available: https://arxiv.org/abs/2307.16540

[42] A. Ailamaki, S. Madden, D. Abadi, G. Alonso, S. Amer-Yahia, M. Balazinska, P. A. Bernstein, P. Boncz, M. Cafarella, S. Chaudhuri, S. Davidson, D. DeWitt, Y. Diao, X. L. Dong, M. Franklin, J. Freire, J. Gehrke, A. Halevy, J. M. Hellerstein, M. D. Hill, S. Idreos, Y. Ioannidis, C. Koch, D. Kossmann, T. Kraska, A. Kumar, G. Li, V. Markl, R. Miller, C. Mohan, T. Neumann, B. C. Ooi, F. Ozcan, A. Parameswaran, I. Pandis, J. M. Patel, A. Pavlo, D. Porobic, V. Sanca, M. Stonebraker, J. Stoyanovich, D. Suciu, W.-C. Tan, S. Venkataraman, M. Zaharia, and S. B. Zdonik, "The cambridge report on database research," 2025. [Online]. Available: https://arxiv.org/abs/2504.11259

[43] H. Q. Ngo, E. Porat, C. Ré, and A. Rudra, "Worst-case optimal join algorithms," *J. ACM*, vol. 65, no. 3, Mar. 2018. [Online]. Available: https://doi.org/10.1145/3180143

[44] A. Birler, T. Schmidt, P. Fent, and T. Neumann, "Simple, efficient, and robust hash tables for join processing," in *Proceedings of the 20th International Workshop on Data Management on New Hardware*, ser. DaMoN '24. New York, NY, USA: Association for Computing Machinery, 2024. [Online]. Available: https://doi.org/10.1145/3662010.3663442

[45] M. Böther, L. Benson, A. Klimovic, and T. Rabl, "Analyzing vectorized hash tables across cpu architectures," *Proc. VLDB Endow.*, vol. 16, no. 11, p. 2755–2768, Jul. 2023. [Online]. Available: https://doi.org/10.14778/3611479.3611485

[46] V. Leis, P. Boncz, A. Kemper, and T. Neumann, "Morsel-driven parallelism: a numa-aware query evaluation framework for the many-core age," in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 743–754. [Online]. Available: https://doi.org/10.1145/2588555.2610507

[47] S. Zeuch, J. Freytag, and F. Huber, "Adapting tree structures for processing with SIMD instructions," in *Proceedings of the 17th International Conference on Extending Database Technology, EDBT 2014, Athens, Greece, March 24-28, 2014*, S. Amer-Yahia, V. Christophides, A. Kementsietsidis, M. N. Garofalakis, S. Idreos, and V. Leroy, Eds. OpenProceedings.org, 2014, pp. 97–108. [Online]. Available: https://doi.org/10.5441/002/edbt.2014.10

[48] S. Deep, X. Hu, and P. Koutris, "Fast join project query evaluation using matrix multiplication," 2020. [Online]. Available: https://arxiv.org/abs/2002.12459

[49] P. Boncz, M. Zukowski, and N. Nes, "Monetdb/x100: Hyper-pipelining query execution," in *Second Biennial Conference on Innovative Data Systems Research, CIDR 2005, Asilomar, CA, USA, January 4-7, 2005, Online Proceedings*. www.cidrdb.org, 2005, pp. 225–237. [Online]. Available: http://cidrdb.org/cidr2005/papers/P19.pdf

[50] O. Kalinsky, B. Kimelfeld, and Y. Etsion, "The triejax architecture: Accelerating graph operations through relational joins," in *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*, ser. ASPLOS '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1217–1231. [Online]. Available: https://doi.org/10.1145/3373376.3378524

[51] D. Arroyuelo, A. Hogan, G. Navarro, J. L. Reutter, J. Rojas-Ledesma, and A. Soto, "Worst-case optimal graph joins in almost no space," in *Proceedings of the 2021 International Conference on Management of Data*, ser. SIGMOD '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 102–114. [Online]. Available: https://doi.org/10.1145/3448016.3457256

[52] A. Bigerl, N. Karalis, L. Heidrich, and A.-C. N. Ngomo, "Efficient updates for worst-case optimal join triple stores," in *The Semantic Web – ISWC 2025: 24th International Semantic Web Conference, Nara, Japan, November 2–6, 2025, Proceedings, Part I*. Berlin, Heidelberg: Springer-Verlag, 2025, pp. 539–555. [Online]. Available: https://doi.org/10.1007/978-3-032-09527-5_29

[53] M. Kitsuregawa, H. Tanaka, and T. Moto-Oka, "Application of hash to data base machine and its architecture," *New Gen. Comput.*, vol. 1, no. 1, p. 63–74, Mar. 1983. [Online]. Available: https://doi.org/10.1007/BF03037022

[54] X. Hu and K. Yi, "Towards a worst-case i/o-optimal algorithm for acyclic joins," in *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, ser. PODS '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 135–150. [Online]. Available: https://doi.org/10.1145/2902251.2902292

[55] D. Arroyuelo, G. Navarro, J. L. Reutter, and J. Rojas-Ledesma, "Optimal joins using compressed quadtrees," *ACM Trans. Database Syst.*, vol. 47, no. 2, May 2022. [Online]. Available: https://doi.org/10.1145/3514231

[56] V. Leis, M. Haubenschild, A. Kemper, and T. Neumann, "Leanstore: In-memory data management beyond main memory," in *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, 2018, pp. 185–196.

[57] O. Kalinsky, Y. Etsion, and B. Kimelfeld, "Flexible caching in trie joins," 2016. [Online]. Available: https://arxiv.org/abs/1602.08721

[58] F. N. Afrati and J. D. Ullman, "Optimizing joins in a map-reduce environment," in *Proceedings of the 13th International Conference on Extending Database Technology*, ser. EDBT '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 99–110. [Online]. Available: https://doi.org/10.1145/1739041.1739056

[59] P. Beame, P. Koutris, and D. Suciu, "Communication steps for parallel query processing," *J. ACM*, vol. 64, no. 6, Oct. 2017. [Online]. Available: https://doi.org/10.1145/3125644

[60] ——, "Skew in parallel query processing," in *Proceedings of the 33rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, ser. PODS '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 212–223. [Online]. Available: https://doi.org/10.1145/2594538.2594558

[61] D. Halperin, V. Teixeira de Almeida, L. L. Choo, S. Chu, P. Koutris, D. Moritz, J. Ortiz, V. Ruamviboonsuk, J. Wang, A. Whitaker, S. Xu, M. Balazinska, B. Howe, and D. Suciu, "Demonstration of the myria big data management service," in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 881–884. [Online]. Available: https://doi.org/10.1145/2588555.2594530

[62] K. Ammar, F. McSherry, S. Salihoglu, and M. Joglekar, "Distributed evaluation of subgraph queries using worst-case optimal low-memory dataflows," *Proc. VLDB Endow.*, vol. 11, no. 6, p. 691–704, Feb. 2018. [Online]. Available: https://doi.org/10.14778/3199517.3199520

[63] D. G. Murray, F. McSherry, R. Isaacs, M. Isard, P. Barham, and M. Abadi, "Naiad: a timely dataflow system," in *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles*, ser. SOSP '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 439–455. [Online]. Available: https://doi.org/10.1145/2517349.2522738

[64] P. Fuchs, P. Boncz, and B. Ghit, "Edgeframe: Worst-case optimal joins for graph-pattern matching in spark," in *Proceedings of the 3rd Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA)*, ser. GRADES-NDA'20. New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: https://doi.org/10.1145/3398682.3399162

[65] R. Li, Q. Ma, X. Shi, and A. Liu, "Sievejoin: Boosting multi-way joins by filtering unneeded intermediate results," *Data Science and Engineering*, vol. 11, no. 1, pp. 143–154, 2026. [Online]. Available: https://doi.org/10.1007/s41019-025-00325-7

[66] B. Ketsman and D. Suciu, "A worst-case optimal multi-round algorithm for parallel computation of conjunctive queries," in *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, ser. PODS '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 417–428. [Online]. Available: https://doi.org/10.1145/3034786.3034788

[67] X. Hu and K. Yi, "Massively parallel join algorithms," *SIGMOD Rec.*, vol. 49, no. 3, p. 6–17, Dec. 2020. [Online]. Available: https://doi.org/10.1145/3444831.3444833

[68] G. Katsogiannis-Meimarakis and G. Koutrika, "A survey on deep learning approaches for text-to-sql," *The VLDB Journal*, vol. 32, no. 4, p. 905–936, Jan. 2023. [Online]. Available: https://doi.org/10.1007/s00778-022-00776-8

[69] T. Yu, R. Zhang, K. Yang, M. Yasunaga, D. Wang, Z. Li, J. Ma, I. Li, Q. Yao, S. Roman, Z. Zhang, and D. R. Radev, "Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task," *CoRR*, vol. abs/1809.08887, 2018. [Online]. Available: http://arxiv.org/abs/1809.08887

[70] J. Li, B. Hui, G. Qu, J. Yang, B. Li, B. Li, B. Wang, B. Qin, R. Geng, N. Huo, X. Zhou, C. Ma, G. Li, K. C. Chang, F. Huang, R. Cheng, and Y. Li, "Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2023.

[71] D. Gao, H. Wang, Y. Li, X. Sun, Y. Qian, B. Ding, and J. Zhou, "Text-to-sql empowered by large language models: A benchmark evaluation," *Proc. VLDB Endow.*, vol. 17, no. 5, p. 1132–1145, Jan. 2024. [Online]. Available: https://doi.org/10.14778/3641204.3641221

[72] K. Kim, J. Jung, I. Seo, W.-S. Han, K. Choi, and J. Chong, "Learned cardinality estimation: An in-depth study," in *Proceedings of the 2022 International Conference on Management of Data*, ser. SIGMOD '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 1214–1227. [Online]. Available: https://doi.org/10.1145/3514221.3526154

[73] J. Sun, J. Zhang, Z. Sun, G. Li, and N. Tang, "Learned cardinality estimation: a design space exploration and a comparative evaluation," *Proc. VLDB Endow.*, vol. 15, no. 1, p. 85–97, Sep. 2021. [Online]. Available: https://doi.org/10.14778/3485450.3485459

[74] S.-J. Qiao, H.-L. Fan, N. Han, L. Du, Y.-H. Peng, R.-M. Tang, and X. Qin, "Learning database optimization techniques: the state-of-the-art and prospects," *Frontiers of Computer Science*, vol. 19, no. 12, p. 1912612, 2025. [Online]. Available: https://doi.org/10.1007/s11704-025-41116-7

[75] T. Faltín, V. Trigonakis, A. Berdai, L. Fusco, C. Iorgulescu, S. Hong, and H. Chafi, "Better distributed graph query planning with scouting queries," in *Proceedings of the 6th Joint Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA)*, ser. GRADES-NDA '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3594778.3594884

[76] R. Marcus, P. Negi, H. Mao, N. Tatbul, M. Alizadeh, and T. Kraska, "Bao: Making learned query optimization practical," in *Proceedings of the 2021 International Conference on Management of Data*, ser. SIGMOD '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 1275–1288. [Online]. Available: https://doi.org/10.1145/3448016.3452838

[77] H. Dong, C. Zhang, G. Li, and H. Zhang, "Cloud-native databases: A survey," *IEEE Trans. on Knowl. and Data Eng.*, vol. 36, no. 12, p. 7772–7791, Dec. 2024. [Online]. Available: https://doi.org/10.1109/TKDE.2024.3397508

[78] A. Berdai, A. Soukrat, and D. Chiadmi, "Efficient in-memory rebalancings for skewed results in distributed graph queries," *Journal of King Saud University Computer and Information Sciences*, vol. 37, no. 9, p. 288, 2025. [Online]. Available: https://doi.org/10.1007/s44443-025-00212-1

[79] I. Arefyeva, D. Broneske, G. Campero, M. Pinnecke, and G. Saake, "Memory management strategies in cpu/gpu database systems: A survey," in *Beyond Databases, Architectures and Structures. Facing the Challenges of Data Proliferation and Growing Variety*, S. Kozielski, D. Mrozek, P. Kasprowski, B. Małysiak-Mrozek, and D. Kostrzewa, Eds. Cham: Springer International Publishing, 2018, pp. 128–142.

[80] J. Cao, R. Sen, M. Interlandi, J. Arulraj, and H. Kim, "Gpu database systems characterization and optimization," *Proc. VLDB Endow.*, vol. 17, no. 3, p. 441–454, Nov. 2023. [Online]. Available: https://doi.org/10.14778/3632093.3632107

[81] J. Fang, Y. T. B. Mulder, J. Hidders, J. Lee, and H. P. Hofstee, "In-memory database acceleration on fpgas: a survey," *The VLDB Journal*, vol. 29, no. 1, p. 33–59, Oct. 2019. [Online]. Available: https://doi.org/10.1007/s00778-019-00581-w

A Behavioral Analysis of Machine Learning for Intrinsic 3D Tree Branching Structures

M. Charitha Lakshan, K. Damitha Sandaruwan, L. N. C. De Silva
University of Colombo, School of Computing, Colombo 00700, Sri Lanka

***Abstract*—Modeling realistic tree structures remains a challenging problem in computer graphics due to the complex interaction between intrinsic growth patterns and environmental influences. Traditional procedural methods provide strong control over environmental adaptation, while learning-based approaches aim to capture branching behavior directly from data. However, learning-based methods typically entangle intrinsic and environmental factors within a single representation, and are trained predominantly on procedurally generated trees, leaving their behavior under real-world structural noise largely uncharacterized. This study isolates the first half of this problem and asks how well machine learning models learn intrinsic 3D branching structure, hierarchy, geometry, and local topology, independently of environmental influence. Trees are represented as node graphs with intrinsic attributes, and local branching prediction is decomposed into a three-class topology classifier and two cascaded geometry regressors. Across controlled experiments on synthetic and LiDAR-reconstructed datasets, models trained on synthetic trees learn branching topology reliably (98.75% test accuracy), while performance degrades substantially on noisy real-world skeletons (50.29%), revealing the sensitivity of intrinsic-structure learning to data quality. Feature ablation shows that structural attributes, rather than spatial position or orientation, drive most of the predictive performance, raising classification accuracy from 55.55% to 97.80%. A perceptual study (25 participants, 625 pairwise trials) shows that generated trees are selected as more realistic in 33.4% of comparisons, a non-trivial proportion, though reference trees remain significantly preferred. Together, these findings characterize both the promise and the limits of purely data-driven intrinsic branching models, and motivate a hybrid pipeline in which learned intrinsic growth is combined with explicit environmental modeling.**

***Keywords*—Tree modeling; machine learning; intrinsic branching; feature representation; 3D tree reconstruction; synthetic data; behavioral analysis; hybrid modeling**

I. Introduction

Trees are essential elements in virtual environments such as games, simulations, and interactive 3D scenes. In these applications, vegetation modeling prioritizes visual and structural plausibility over strict biological realism. Tree models should appear natural and avoid repetitive or artificial branching patterns that reduce immersion. Unlike botanical studies, which require accurate growth mechanisms, many virtual environments only demand perceptually convincing structures. Consequently, many approaches focus on intrinsic branching properties, such as hierarchy, variation, and proportional relationships, rather than exact biological processes.

Recent data-driven methods, including DeepTree-style [1] approaches, have demonstrated that local branching behavior can be learned using machine learning. These methods typically formulate tree growth as a local prediction problem, where branching structure and geometry are inferred from node-level features. However, learning from real-world scanned trees remains challenging due to noise, missing branches, and reconstruction artifacts, which reduce structural consistency and learning accuracy.

In this work, we present a behavioral analysis of machine learning models trained to learn intrinsic 3D tree branching patterns without modeling environmental effects. Rather than reproducing biologically accurate growth, we focus on generating visually plausible branching structures for virtual environments. As illustrated in Fig. 1, tree growth is modeled as an iterative local prediction process based on intrinsic node attributes. A classifier predicts branching decisions, while conditional regression models estimate branch geometry, enabling the generation of structurally consistent trees. We further compare models trained on synthetic and real scanned datasets to analyze the impact of data quality on learning performance. Results show that intrinsic branching patterns can be learned reliably under controlled conditions, whereas noisy real-world data introduces significant limitations. These findings support a decoupled modeling strategy in which intrinsic growth is learned independently from environmental adaptation handled by explicit procedural or mathematical models. Accordingly, this work focuses on analyzing the behavior of machine learning models when learning intrinsic branching structures, rather than proposing a novel machine learning architecture, providing a foundation for future hybrid tree modeling pipelines.

II. Related Work

Tree modeling has been widely studied in computer graphics, with approaches ranging from procedural rule-based systems to data-driven and learning-based methods. A central challenge in this area is modeling both intrinsic branching

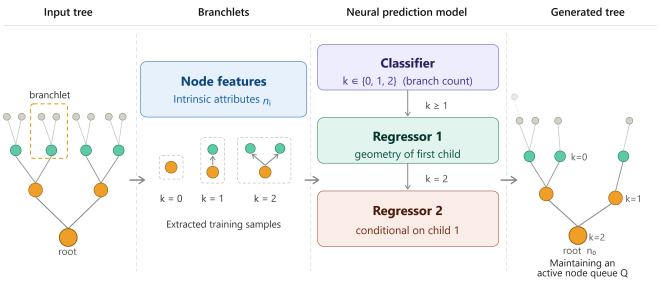


Fig. 1. Overview of the learning framework for behavioral analysis of intrinsic 3D tree branching. Branchlets are represented by intrinsic node attributes n_i. A classifier predicts the number of child branches ($k \in \{0, 1, 2\}$), followed by conditional regression models estimating branch geometry.

structure and environmental adaptation, including responses to light, obstacles, and neighboring vegetation. Procedural methods have long been used for this purpose due to their interpretability, controllability, and computational efficiency. One of the most influential approaches is the L-system framework [2], which represents plant growth using symbolic rewriting rules and has been extended to incorporate phototropism and space competition. Other procedural models explicitly simulate environmental influence through light-driven growth and spatial optimization. Beneš [3] introduced phototropic plant development models based on light availability, while the space colonization algorithm of Runions et al. [4] modeled branching as an iterative attraction process guided by available space. More biologically inspired approaches, such as the self-organizing tree model of Pałubicki et al. [5], simulate local competition for light and space to produce complex emergent tree structures. In all of these formulations, however, both intrinsic growth and environmental response are expressed as hand-authored rules, so structural realism depends on manual parameter tuning rather than on branching behavior observed in real trees.

Recent procedural work has focused on interactive and adaptive growth. Plastic Trees [6] introduced skeleton-based models that dynamically adjust branch orientation in response to light, obstacles, and nearby vegetation, while virtual ecosystem approaches enable plants to interact with surrounding objects and optimize spatial occupation [7]. GPU-accelerated methods incorporating ray-traced light simulation have further improved environmental realism and computational performance [8]. Inverse procedural approaches have also been proposed to estimate procedural parameters directly from existing tree geometry, bridging rule-based modeling and real-world observations [9]. These methods provide strong control over environmental adaptation through explicit parameterized rules, making them well suited for environment-driven modeling in virtual scenes. Their expressiveness, however, remains bounded by the rule set chosen in advance, and intrinsic branching is still specified rather than inferred from data.

More recently, machine learning and hybrid approaches have explored learning branching behavior directly from data. DeepTree [1] formulates tree growth as a local prediction problem, where branching decisions and geometric attributes are inferred from node-level spatial context. Transformer-based autoregressive models have further demonstrated the ability to generate hierarchical tree structures and temporal growth sequences directly from learned representations [10]. Hybrid methods such as Latent L-systems [11] combine neural learning with grammar-based representations, allowing branching rules to be learned implicitly while retaining procedural structure, whereas TreeStructor [12] targets faithful reconstruction of scanned geometry rather than the generation of novel structures. Two properties of this body of work are relevant here. First, these models are predominantly trained on procedurally generated data, a choice their authors attribute to the limited availability of large collections of reconstructed real trees [1]; consequently, how such models behave when trained on noisy reconstructed skeletons remains largely uncharacterized. Second, intrinsic structure and environmental context are typically encoded jointly, whether within a single node representation or a learned sequence, so the contribution of intrinsic attributes alone cannot readily be isolated or analyzed. These observations motivate a systematic behavioral analysis of how machine learning models learn intrinsic 3D tree branching structure independently of environmental influences, and whether such learning can serve as a reliable foundation for a broader decoupled modeling strategy.

III. Problem Formulation and Representation

In this work, a tree is represented as a rooted graph $\mathcal{T}$, where nodes correspond to branching points or segment endpoints, and edges represent parent-child relationships between connected branches. This graph-based representation is widely used in both procedural and data-driven tree modeling, as it provides a compact and flexible description of hierarchical structure and local geometry.

Each node in the tree graph is represented as a tuple of intrinsic attributes that describe its local geometric and structural properties. Following the representation style introduced in DeepTree [1], nodes are modeled as attribute tuples rather than purely topological elements, enabling local branching behavior to be learned from node-level context. In contrast to DeepTree, the attribute set used in this work is intentionally simplified. Only attributes that are consistently available across both real scanned trees and synthetically generated data are included. Attributes related to biological age, species, or other global physiological properties are omitted, as they are either unavailable or unreliable in the considered datasets.

Formally, each node $n_i \in \mathcal{T}$ is defined as:

$$n_i = \left(\mathbf{p}_i, \mathbf{q}_i, r_i, r^{root}, o_i, d_i, d_i^{root}, d_i^{branch}\right) \tag{1}$$

where, $\mathbf{p}_i \in \mathbb{R}^3$ denotes the 3D position of the node, $\mathbf{q}_i \in \mathbb{R}^4$ is a unit quaternion representing the local branch orientation with respect to its parent node, $r_i \in \mathbb{R}^+$ is the local branch radius, and $r^{root} \in \mathbb{R}^+$ denotes the radius of the root node, which is shared by all nodes in the tree and provides global scale context. The value $o_i \in \mathbb{N}$ denotes the branch order. The value $d_i \in \mathbb{R}$ represents the distance to the parent node. In addition, $d_i^{root} \in \mathbb{R}$ denotes the cumulative distance from the root node, and $d_i^{branch} \in \mathbb{R}$ represents the distance from the start of the current branch. All attributes describe intrinsic structural properties and do not encode any explicit environmental information.

Tree growth is modeled as a local prediction problem. Instead of predicting the full tree structure at once, growth behavior is inferred at the node level using local information. For each node, a local growth neighborhood-referred to as a "branchlet" (Fig. 2) is defined, consisting of a parent node and its immediate children together with their connecting edges. Based on this local representation, the learning task is divided into two components: a classification task that predicts the number of child branches produced by a node, and regression tasks that predict geometric properties of each child branch, such as relative orientation, length, and thickness. Together, these predictions describe the intrinsic branching structure in a local and incremental manner.

A key assumption of this formulation is that environmental factors are not considered. External influences such as sunlight, wind, gravity, terrain constraints, and interactions with surrounding vegetation are intentionally excluded from both the

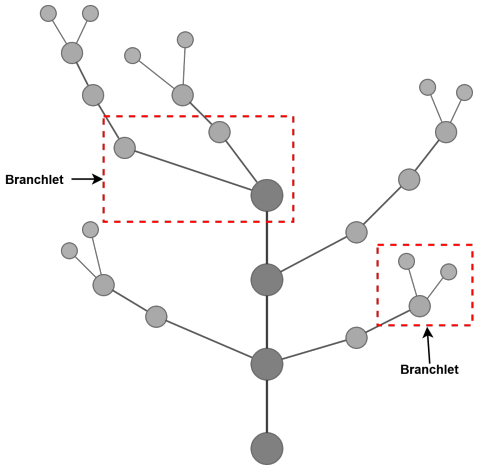


Fig. 2. Local branchlet representation used in intrinsic branching prediction.

representation and the learning process. The objective of this work is therefore not to model adaptive growth behavior, but to analyze how the proposed learning framework behaves when trained exclusively on intrinsic branching patterns, without exposure to environmental influences. This formulation supports a modular modeling approach, in which intrinsic growth prediction can later be combined with a separate environmental adaptation stage. The following section describes how this local prediction problem is realized through a classification and regression pipeline operating on the intrinsic node representation defined above.

IV. Learning Intrinsic Branching

This section describes how intrinsic tree branching behavior is learned using the node representation introduced earlier. The learning pipeline consists of two main stages: branch structure prediction and branch geometry prediction. First, a classification model predicts the number of child branches emerging from a parent node in a local context. Based on this prediction, regression models are then used to estimate the geometric properties of each child branch. The resulting predictions are applied by a tree generation procedure to incrementally construct branching structures.

A. Branch Structure Prediction

The first learning task focuses on predicting local branching topology. Given the intrinsic attributes of a parent node, a classification model predicts the number of child branches produced by that node. The classification output is restricted to three classes: 0, 1, or 2 children.

This restriction is motivated by the distribution of branching patterns observed in both real scanned trees and synthetically generated trees. During dataset analysis, nodes with three or more child branches were found to be rare, accounting for approximately 3% or less of all branching events. Including these cases introduced strong class imbalance and negatively affected training stability. Therefore, nodes with more than two children are excluded from training, allowing the classifier to focus on the most common and structurally relevant branching patterns.

A lightweight feed-forward neural network is used for branch count classification (Fig. 3). In contrast to more complex classifier designs used in prior work such as DeepTree, the proposed architecture is intentionally simplified to improve robustness and interpretability under noisy real-world data conditions. The model takes the intrinsic attributes of the parent node as input and outputs class logits corresponding to the three child-count categories. The network is trained using the cross-entropy loss function, which encourages accurate prediction of the discrete branching classes. This design emphasizes reliable local branching prediction rather than modeling global tree structure.

B. Branch Geometry Prediction

Once the number of child nodes is determined, the next task is to predict the geometric properties of each child node. These properties include relative orientation with respect to the parent, as well as geometric attributes such as edge length and node radius.

To model this, two regression networks are employed. The first regressor predicts the geometry of the first child node. The second regressor predicts the geometry of the second child node when two children are present, and operates in a conditional manner based on the first child. Specifically, during tree generation, the second regressor takes as input the intrinsic attributes of the parent node together with the predicted attributes of the first child, forming a cascaded prediction structure. During training, however, the second regressor is conditioned on the ground-truth attributes of the first child rather than the predicted outputs of the first regressor. This design avoids error propagation during training and allows the model to learn stable conditional geometric relationships.

The regression architecture is influenced by the design used in DeepTree, with a multi-head structure that separates orientation prediction from geometric attribute estimation (Fig. 4). Each regressor consists of two prediction heads: one for orientation, represented as a quaternion, and another for geometric attributes, including relative edge length and node radius. The models take intrinsic node attributes as input and output continuous predictions for these properties.

The regression networks are trained using a combination of loss functions. Orientation prediction is supervised using a cosine similarity loss between the predicted and ground-truth orientations, while geometric attributes are optimized using mean squared error (MSE). These losses are jointly minimized to ensure accurate prediction of local branching geometry.

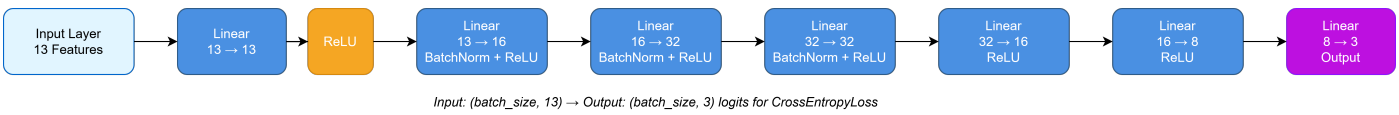


Fig. 3. Architecture of the child-count classification network used to predict the number of child branches (0, 1, or 2) from intrinsic node attributes.

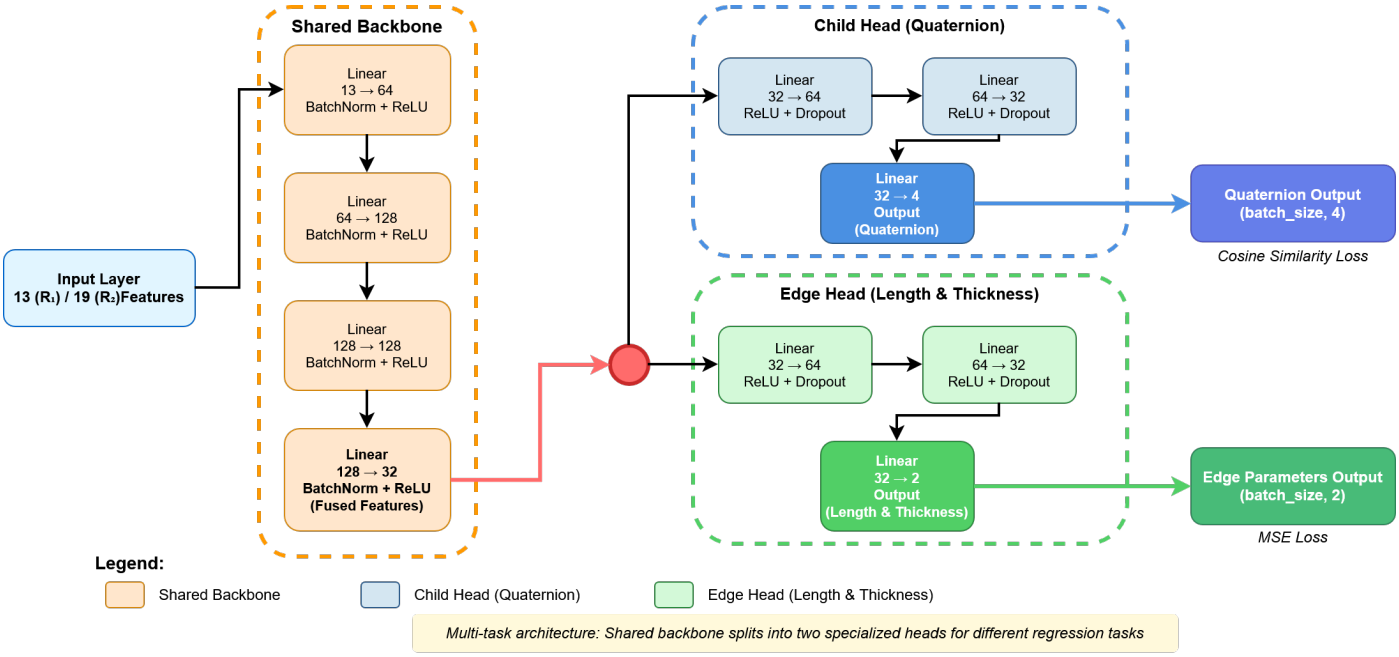


Fig. 4. Multi-task regression architecture used for branch geometry prediction.

C. Incremental Tree Generation

The trained classification and regression models are used to generate tree structures in an incremental manner. Tree generation begins from a manually initialized root node, whose intrinsic attributes are specified explicitly. These include the root position, orientation, branch radius, and initial distance values. Manual initialization of the root node is required, as the learning models operate locally and require an initial node context to initiate growth.

Starting from the root node, tree construction proceeds by iteratively expanding active nodes, as summarized in Algorithm in the Fig. 5. For each active node, the child-count classifier predicts whether the node produces zero, one, or two child branches. For nodes that produce children, the corresponding regression models predict geometric properties of each child branch, including relative orientation, edge length, and branch thickness. Child node positions and orientations are computed by applying the predicted transformations to the parent node.

Predicted branching topology alone does not guarantee termination: prior work reports that locally predicted child counts may fail to produce terminal nodes for some configurations, allowing generation to continue indefinitely [1]. To avoid uncontrolled growth and to improve stability, an existence probability is associated with each node. This probability is propagated from parent to child and gradually reduced as the depth of the tree increases. Child nodes are only created if their sampled existence probability exceeds a random threshold. Newly generated child nodes are added to a queue of active nodes and processed in subsequent iterations. The generation process terminates when no active nodes remain or when a predefined node limit is reached. This procedure models intrinsic branching behavior only and does not include any environmental constraints.

Fig. 6 presents representative examples of tree structures generated using the incremental generation procedure described above. The generated structures exhibit coherent hierarchical branching organization, with consistent radius tapering across depth levels and plausible bifurcation patterns, indicating that the employed learning framework produces structurally consistent and visually plausible tree forms from purely intrinsic node attributes.

1) Notation and symbols: Let $\mathcal{T}$ denote the generated tree represented as a rooted graph. Each node $n_i \in \mathcal{T}$ is associated with a tuple of intrinsic attributes $\left(\mathbf{p}_i, \mathbf{q}_i, r_i, r^{root}, o_i, d_i, d_i^{root}, d_i^{branch}\right)$, as defined in Section III. The classifier $C(\cdot)$ predicts the number of child branches $k \in \{0, 1, 2\}$ for a given parent node. The regressors $R_1(\cdot)$ and $R_2(\cdot)$ predict geometric properties of the first and second child branches. Each node n_i is additionally associated with an existence probability $p(n_i)$, which controls whether the node is instantiated during generation. The parameter α denotes the probability decay factor applied beyond a depth threshold D_{start}. The queue $\mathcal{Q}$ maintains the set of active nodes, and $N_{\max}$ specifies the maximum number of nodes allowed in the generated tree.

D. Implementation Details

All models were implemented in PyTorch and trained on a single NVIDIA RTX 5060 GPU with an AMD Ryzen 5 7500F CPU and 32 GB of system memory. The intrinsic node tuple of Eq. (1) expands to a 13-dimensional input vector; Regressor 2 additionally receives the four orientation and two geometric attributes of the first child, giving a 19-dimensional input. Layer widths are given in Fig. 3 and Fig. 4, with dropout applied within the regression heads only. Because the cosine-similarity and mean squared error terms of the regression objective differ substantially in numerical scale, the geometric term is weighted so that neither head dominates optimization. No learning rate scheduling, weight decay, or early stopping is applied; the weights obtained at the final epoch are retained. Remaining settings are summarized in Table I.

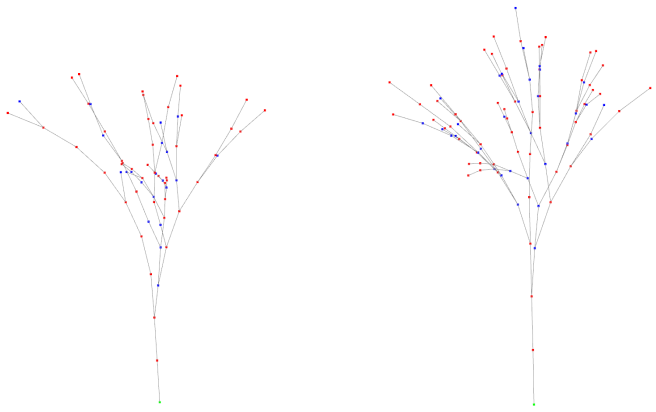
Require: Trained child classifier C
Require: Trained regressors R_1, R_2
Require: Root node n_0 with initial existence probability p_0
Require: Maximum node limit $N_{\max}$
Require: Probability decay factor α
Require: Depth threshold D_{start}
Ensure: Generated tree structure $\mathcal{T}$

Initialize queue $\mathcal{Q} \leftarrow \{n_0\}$
Initialize tree $\mathcal{T} \leftarrow \{n_0\}$
Set $p(n_0) \leftarrow p_0$
while $\mathcal{Q}$ is not empty **and** $|\mathcal{T}| < N_{\max}$ **do**
    Pop parent node n_p from $\mathcal{Q}$
    Predict number of children $k \leftarrow C(n_p)$
    Let d_p be the depth (nodes-to-parent count) of n_p
    if $k \geq 1$ **then**
        Predict child 1 geometry using $R_1(n_p)$
        Compute child depth $d_1 \leftarrow d_p + 1$
        if $d_1 > D_{start}$ **then**
            Apply probability decay $p_1 \leftarrow p(n_p)\,(1-\alpha)$
        else
            $p_1 \leftarrow p(n_p)$
        end if
        if rand() $< p_1$ **then**
            Create child node n_1
            Assign $p(n_1) \leftarrow p_1$
            Add n_1 to $\mathcal{T}$ and $\mathcal{Q}$
        end if
    end if
    if $k = 2$ **then**
        Predict child 2 geometry using $R_2(n_p, R_1(n_p))$
        Compute child depth $d_2 \leftarrow d_p + 1$
        if $d_2 > D_{start}$ **then**
            Apply probability decay $p_2 \leftarrow p(n_p)\,(1-\alpha)$
        else
            $p_2 \leftarrow p(n_p)$
        end if
        if rand() $< p_2$ **then**
            Create child node n_2
            Assign $p(n_2) \leftarrow p_2$
            Add n_2 to $\mathcal{T}$ and $\mathcal{Q}$
        end if
    end if
end while
return $\mathcal{T}$

Fig. 5. Incremental tree generation algorithm.

TABLE I. TRAINING AND GENERATION CONFIGURATION

Parameter	Value
Optimizer	Adam
Learning rate	1×10^{-4}
Batch size	512
Epochs	300
Dropout (regressor heads)	0.2
Geometry loss weight	100
Initial existence probability p_0	1.0
Depth threshold D_{start}	5
Probability decay factor α	0.1
Node limit $N_{\max}$	250
Root candidates sampled	100
Root sampling radius	0.1



(a) Generated tree sample 1 (b) Generated tree sample 2

Fig. 6. Representative examples of generated tree structures.

V. DATASETS

This section describes the datasets used for learning intrinsic branching behavior and the preprocessing steps applied to construct branchlet-level training samples. Two complementary data sources are considered: real-world tree reconstructions derived from LiDAR scans and procedurally generated synthetic tree skeletons. Using both datasets enables analysis of learning behavior under realistic but noisy conditions as well as under controlled structural settings.

A. Real-World Tree Dataset

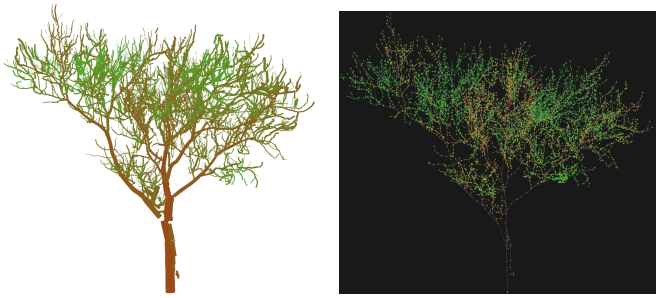
For real-world data, we use the TreeML-Data dataset [13], which contains tree structures reconstructed from terrestrial Light Detection and Ranging (LiDAR) scans using quantitative structure modeling (QSM) (Fig. 7). In this representation, trees are described as graphs of connected cylindrical primitives that approximate branch segments and encode geometric attributes such as position, orientation, radius, and branch order.

Although TreeML-Data does not provide an explicit node-edge skeleton, the cylinder graph preserves hierarchical connectivity and therefore supports derivation of a structural node representation. In this work, starting points of the cylinders are interpreted as nodes while connectivity between segments defines graph edges, enabling conversion to the node-level formulation used for learning.

Because the dataset originates from real measurements, the resulting structures may contain missing branches, uneven sampling, geometric noise, and segmentation inconsistencies. These factors introduce variability at the local branching level and can influence learning behavior, particularly in deeper structural regions. Nevertheless, real-world data remains important for evaluating how intrinsic branching models perform under practical acquisition conditions.

B. Synthetic Tree Dataset

To support controlled evaluation of intrinsic branching behavior, we additionally use synthetic skeletal tree structures released with the CHI24_TreeProject dataset [14] (Fig. 8). These trees are generated using procedural modeling techniques and provided as explicit skeletal graphs describing parent-child relationships and geometric attributes at each node.



(a) QSM cylinder representation (b) Derived node-edge graph

Fig. 7. Real-world tree representation from TreeML-Data.

Compared to LiDAR reconstructions, the synthetic skeletons form clean and fully connected hierarchical structures without scanning artifacts or segmentation noise. Such representations are widely used in procedural vegetation modeling because they expose branching organization directly and allow local growth relationships to be analyzed independently of reconstruction errors.

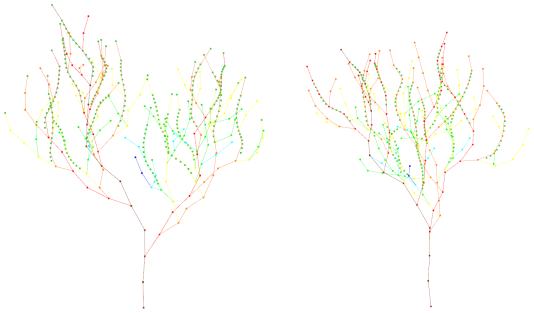
While synthetic trees do not capture the full variability of biological growth, they provide a stable structural environment for studying branching prediction, orientation modeling, and geometric parameter learning. The combination of real reconstructions and synthetic skeletons therefore enables evaluation across both realistic and controlled conditions.

C. Data Preprocessing

Both datasets undergo preprocessing to produce a unified intrinsic node representation suitable for branchlet-level learning. The objective is to extract structural attributes, normalize geometric scale, and construct local parent-child neighborhoods while minimizing implicit environmental bias.

For LiDAR-derived trees, cylindrical segments are converted into node-edge graphs by treating segment centers as nodes while preserving connectivity as hierarchical edges. Intrinsic attributes, including spatial position, node radius, branch order, edge length, cumulative distance from the root, and distance from branch origin, are computed through graph traversal. Local orientation is encoded using quaternion representations derived from parent-child direction vectors.

Because the intrinsic features are derived directly from spatial coordinates, geometric normalization plays an important role in defining the statistical distribution of input features as well as the numerical scale of regression targets. To evaluate its influence, an initial preprocessing study was conducted using synthetic tree structures, where reconstruction artifacts are absent and the effect of normalization can be analyzed in isolation. Two commonly used strategies were considered: uniform (isotropic) scaling, which preserves the original aspect ratios while fitting each tree into a bounded spatial domain, and axis-wise normalization, which independently rescales each spatial dimension to a fixed range. Both strategies enabled accurate learning of branching topology and geometry. However, uniform scaling consistently produced more stable optimization, slightly higher classification accuracy, and lower regression errors, particularly for branch geometry prediction. Based on these observations, uniform scaling was adopted as the default normalization strategy for all subsequent experiments.



(a) (b)

Fig. 8. Synthetic skeletal tree examples from the CHI24_TreeProject dataset.

Synthetic skeletons follow the same structural extraction procedure and additionally undergo controlled rotation augmentation around the vertical axis to improve orientation invariance. Distance measures and orientation encoding are computed consistently with the real-world data.

For each parent node, immediate children are grouped to form branchlets that define classification targets (number of children) and regression targets describing relative child orientation, edge length, and node radius. Despite differences in acquisition and generation processes, preprocessing produces equivalent intrinsic feature representations for both datasets, enabling learning to focus on internal branching structure while maintaining compatibility across data modalities.

The dataset is divided at the tree level to avoid structural overlap between training and evaluation samples. Specifically, 70% of the trees are used for training, 15% for validation, and 15% for testing. All nodes belonging to a given tree are assigned to the same subset, preventing structural information leakage across splits.

VI. Experiments, Results and Evaluation

This section presents a behavioral analysis of the employed learning framework through controlled experiments on intrinsic 3D tree branching structures. The analysis comprises a comparison of learning performance across synthetic and real-world datasets, an intrinsic feature importance study, and a perceptual evaluation of generated tree structures. All quantitative experiments were repeated **five times** using **different random seeds**, with reported metrics representing mean values; figures correspond to the run **closest to the mean**. Prior learning-based tree models are not evaluated under a directly comparable protocol. DeepTree [1] adopts a multi-label formulation in which independent output heads determine whether each potential child is instantiated, and is trained on procedurally generated data under a different geometric normalization, so its reported node-level metrics do not correspond to the three-class formulation and evaluation setting used here. Latent L-systems [11] validates generated structures through aggregate geometric distributions and perceptual metrics rather than node-level prediction error. Moreover, no prior work reports node-level metrics on LiDAR-reconstructed skeletons, the specific data condition examined in Section VI-A. Direct numerical comparison would require re-implementation and retraining under the present protocol, an effort that falls outside the scope of this behavioral analysis. Human perceptual judgment

(Section VI-C) is therefore adopted as the reference point, consistent with prior validation strategies in which generated structures are compared against dataset reference trees.

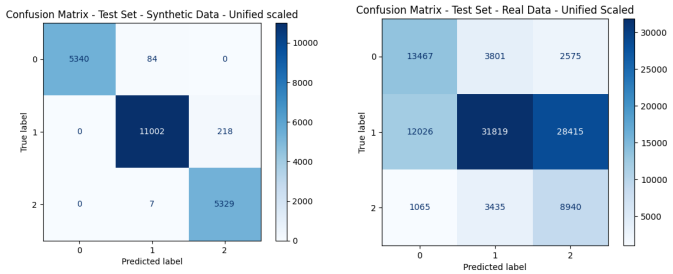
A. Synthetic and Real Data Comparison

We evaluate how the learning model behaves when trained on different data sources, using synthetic and real-world tree data. Classification accuracy, regression error, and convergence behavior are compared under identical conditions.

1) Experimental setup: The dataset split follows the procedure described in Section V, where trees are partitioned at the instance level to prevent structural overlap between training, validation, and test sets. The branching classifier is formulated as a three-class prediction problem corresponding to nodes with zero, one, or two children. To mitigate class imbalance, the training and validation sets are balanced across the three classes using random downsampling, while the test sets retain the original class distributions for unbiased evaluation. This balancing is applied only to the classifier, whereas the regression models are trained on the original data distributions. Uniform scaling, adopted as the default normalization strategy following the preprocessing analysis described in Section V, is applied consistently across all experiments.

2) Branching classification results: The branching classifier was evaluated on both synthetic and real tree datasets to analyze how effectively the intrinsic representation captures branching topology under different conditions. Under identical training settings, both models exhibited stable convergence with no significant signs of overfitting. The model trained on synthetic data converged rapidly and achieved validation accuracy above 98%, whereas the model trained on real tree data converged to a lower but stable accuracy of approximately 60%, indicating that branching prediction is more challenging under real-world structural conditions.

To further analyze prediction behavior, confusion matrices were computed for the held-out test sets. Fig. 9 presents the confusion matrices for both datasets. For the synthetic dataset, the classifier demonstrates strong performance across all three classes, with only a small number of misclassifications between continuation nodes and bifurcation nodes. In contrast, the confusion matrix obtained from the real dataset reveals a larger number of errors between these two classes, while terminal nodes (class 0) remain relatively well identified. This suggests that continuation and bifurcation states often exhibit similar local geometric configurations in real tree skeletons, leading to increased structural ambiguity.



(a) Synthetic dataset. *(b) Real dataset.*

Fig. 9. Confusion matrices for the branching classifier evaluated on the test sets for synthetic and real datasets.

Quantitative results are summarized in Table II. The classifier trained on synthetic data achieves consistently high accuracy across training, validation, and test sets, reflecting the structural consistency of the synthetic trees. In contrast, the model trained on real tree data achieves lower overall accuracy while still maintaining meaningful predictive performance. Because the training and validation sets are class-balanced while the test sets retain the original distribution, raw accuracy on the real dataset understates model performance relative to a majority-class predictor; balanced accuracy across the three classes remains substantially above the 33.3% chance level. This suggests that the intrinsic feature representation remains informative despite the increased geometric variability and reconstruction noise present in real-world tree structures.

3) Regression results: The regression networks predict child branch orientation, represented as a quaternion, together with geometric attributes including branch radius and edge length. The models were trained separately on synthetic and real datasets to analyze how learning behavior varies across different data conditions. Under identical optimization settings, all regression models exhibited stable convergence without significant overfitting. Models trained on synthetic data converged more rapidly and achieved lower final losses, whereas models trained on real data stabilized at higher loss values, reflecting the increased structural variability of real-world tree skeletons.

a) Orientation prediction: Orientation prediction accuracy was evaluated using the angular deviation between the predicted quaternion and the ground-truth branch orientation. Fig. 10 shows the angular error distributions for all regression models. For models trained on synthetic data [Fig. 10(a) and 10(b)], most predictions are concentrated within small angular deviations, indicating high orientation accuracy. Regressor 1 exhibits a tighter concentration of low-angle errors, while Regressor 2 shows a slightly broader distribution, reflecting the increased complexity of predicting secondary branch orientation.

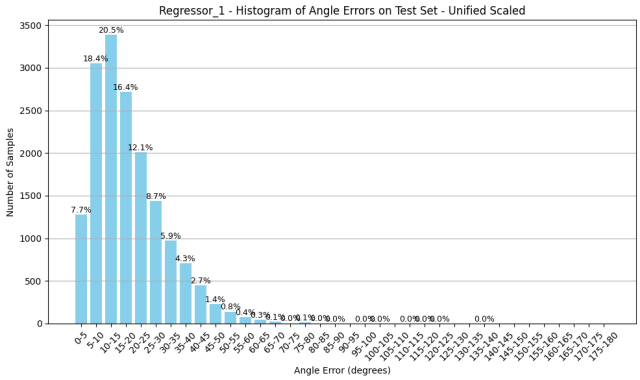
In contrast, models trained on real data [Fig. 10(c) and 10(d)] produce noticeably wider error distributions with longer tails toward larger angular deviations. Although Regressor 1 still maintains a concentration of predictions at lower angles, Regressor 2 shows the widest spread, indicating reduced orientation accuracy under real-world structural conditions. These results suggest that both regressors successfully learn orientation relationships, although prediction accuracy decreases on real data, particularly for secondary branch geometry.

b) Geometric attribute prediction: In addition to orientation, the regression networks estimate branch radius and edge length. Prediction accuracy is evaluated using percentage error relative to ground-truth values. Fig. 11 shows the corresponding error distributions.

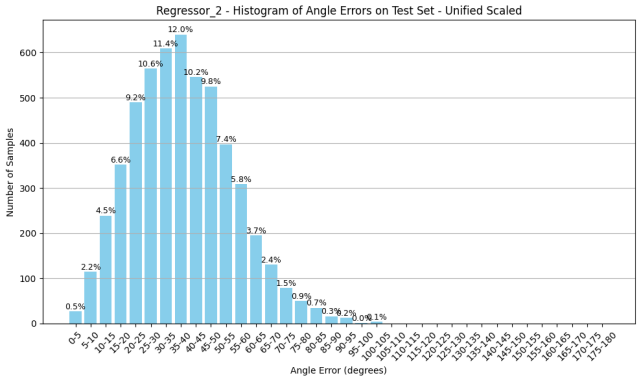
For synthetic data [Fig. 11(a) and 11(b)], both regressors produce highly concentrated error distributions, with most

TABLE II. CLASSIFICATION PERFORMANCE ON SYNTHETIC VS REAL DATA (MEAN ± STANDARD DEVIATION)

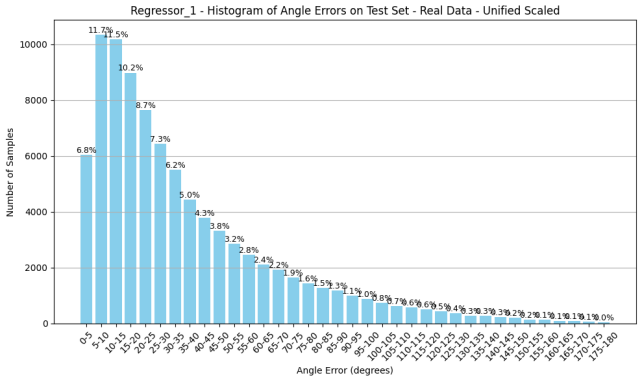
Dataset	Train Accuracy (%)	Validation Accuracy (%)	Test Accuracy (%)
Synthetic	98.73 ± 0.08	99.05 ± 0.09	98.75 ± 0.06
Real Trees	59.81 ± 0.06	60.02 ± 0.20	50.29 ± 1.25



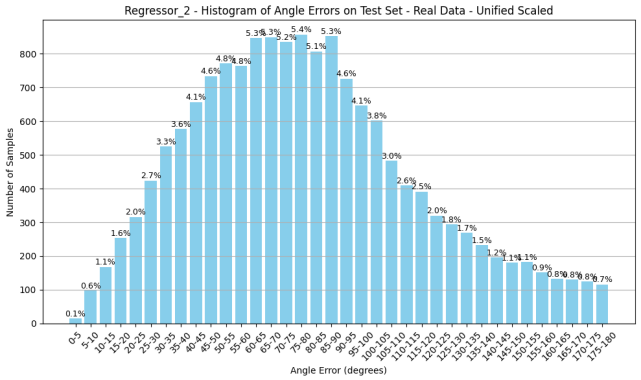
(a) Synthetic data – Regressor 1.



(b) Synthetic data – Regressor 2.

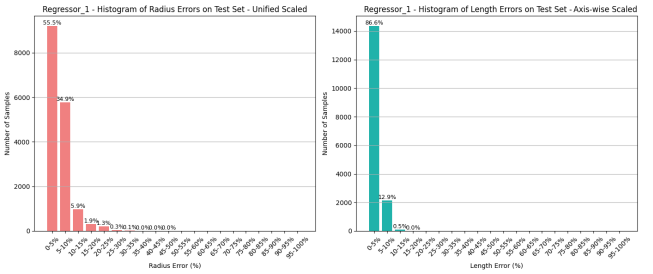


(c) Real data – Regressor 1.

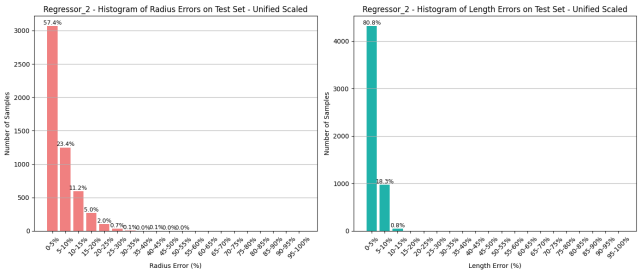


(d) Real data – Regressor 2.

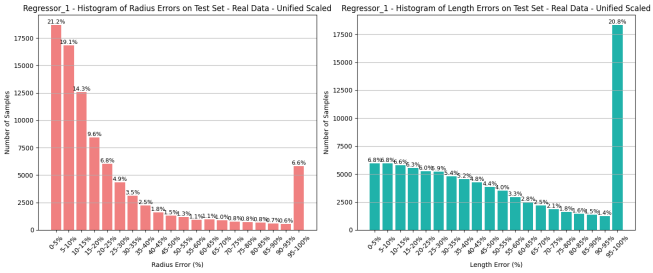
Fig. 10. Angular error distributions for the regression networks trained on synthetic and real datasets.



(a) Synthetic data – Regressor 1.



(b) Synthetic data – Regressor 2.



(c) Real data – Regressor 1.

(d) Real data – Regressor 2.

Fig. 11. Radius and edge length prediction error distributions for the regression networks evaluated on the test datasets.

predictions falling within low percentage error ranges. Regressor 1 exhibits slightly tighter distributions than Regressor 2, indicating more accurate and stable geometric predictions for the primary child branch.

In contrast, models trained on real data [Fig. 11(c) and 11(d)] show significantly wider and more irregular error distributions, with long tails extending toward higher error ranges. A noticeable accumulation of samples appears in the highest error bins, indicating occasional large prediction deviations. These effects are more pronounced for Regressor 2, highlighting the increased difficulty of predicting secondary branch geometry in structurally inconsistent real-world trees.

Quantitative results in Table III support these observations, showing substantially higher orientation and geometry losses

TABLE III. REGRESSION PERFORMANCE UNDER SYNTHETIC AND REAL TRAINING DATASETS (MEAN ± STANDARD DEVIATION)

Dataset	Model	Orientation Loss	Geometry Loss
Synthetic Trees	Reg. 1	0.0185 ± 0.0001	$(2.4 \pm 0.55) \times 10^{-6}$
Real Trees	Reg. 1	0.0819 ± 0.0002	$(3.52 \pm 0.08) \times 10^{-5}$
Synthetic Trees	Reg. 2	0.0653 ± 0.0013	$(2.6 \pm 0.55) \times 10^{-6}$
Real Trees	Reg. 2	0.2714 ± 0.0009	$(6.1 \pm 0.00) \times 10^{-5}$

for models trained on real data compared with those trained on synthetic data. Overall, the employed approach learns local geometric relationships reliably under controlled synthetic conditions, while performance degrades on real-world data due to increased structural variability, reconstruction noise, and the presence of more ambiguous branching configurations.

4) Discussion: The results of this experiment reveal a clear difference in learning behavior between models trained on synthetic tree structures and those trained on real-world tree skeleton data. Models trained on synthetic data achieve high classification accuracy and low regression errors, indicating that the intrinsic feature representation effectively captures branching geometry when the underlying structures follow consistent geometric rules. In contrast, models trained on real data exhibit reduced classification accuracy and higher regression errors, although training remains stable and meaningful structural patterns are still learned.

This performance gap can be attributed to the increased structural variability present in real-world tree data. Unlike synthetic trees, which are generated with controlled and consistent geometric relationships, real tree models are reconstructed from scanned data using cylindrical segment approximations. This process introduces irregularities such as noisy branching angles, inconsistent segment lengths, abrupt radius variations, and local reconstruction artifacts. These factors result in ambiguous local geometric configurations, making it more difficult for the model to infer branching topology and accurately predict geometric attributes.

Fig. 12 provides visual examples of these structural characteristics. Compared to synthetic trees, real tree skeletons exhibit geometric inconsistencies that disrupt regular branching patterns. These irregularities contribute to the wider error distributions observed in Fig. 10 and Fig. 11, as well as the reduced classification accuracy shown in Fig. 9 and Table II. Building on these observations, Fig. 13 presents tree structures generated using models trained on real-world data. The trees are produced using the incremental generation process described in Section IV, where branching decisions and geometric attributes are predicted iteratively.

The generated structures exhibit coherent branching behavior at a global level, indicating that the model captures fundamental structural relationships even under challenging data conditions. However, compared to the synthetic data results shown under the section IV, the generated trees display increased variability in local branching patterns, including irregular branch orientations, inconsistent segment lengths, and less uniform structural organization. These effects can be associated with the broader error distributions observed in the regression results, as well as the reduced classification accuracy for branching decisions. Such variability reflects the difficulty of learning consistent structural representations from real-world tree data, where geometric irregularities and reconstruction artifacts introduce ambiguity in local branching configurations. As a result, while the generated trees retain recognizable branching characteristics, their structural consistency is less pronounced than that observed in models trained on synthetic data.

Fig. 12. Real tree examples represented using cylindrical branch segments.

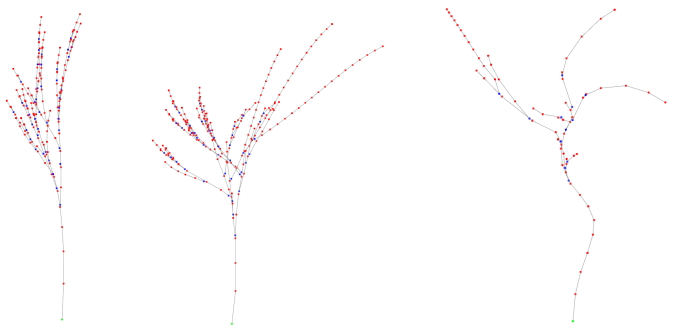


Fig. 13. Tree structures generated using models trained on real-world data. (Only for this, the node limit is set to 1000 for better visualization of the branching patterns.)

Overall, the results demonstrate that the employed framework exhibits consistent behavior when trained on synthetic datasets, while behavioral degradation is observed on real-world data. The reduced performance highlights the influence of noise, geometric inconsistencies, and reconstruction artifacts on both branching decisions and geometric predictions. These findings characterize both the capabilities and the limitations of the learning framework, and suggest that improvements in preprocessing, data acquisition, and geometric representation could further enhance behavioral stability and output quality.

B. Intrinsic Feature Importance Analysis

While the previous analysis examined the effects of dataset characteristics on learning performance, it is also important to understand how individual intrinsic node attributes contribute to branching prediction. The intrinsic node representation defined in Eq. (1) combines spatial, structural, and geometric information, but the relative contribution of these attributes remains unclear. To investigate this, a controlled feature ablation study is conducted using progressively enriched subsets of the intrinsic representation.

1) Experimental setup: All configurations are evaluated using the synthetic tree dataset to isolate feature-related effects without reconstruction noise or structural inconsistencies present in real-world data. Uniform scaling is applied consistently across all configurations, and the original class distribution is preserved to maintain the natural branching statistics of the dataset.

Feature importance is evaluated using the branching classifier and the first regression model, which predicts the geometric attributes of the primary child branch. Although the complete framework includes a second regression model, it is omitted from this analysis since both regressors share the same architecture and learning formulation, and previous results showed consistent behavior between them.

Four feature configurations are considered, corresponding to progressively enriched intrinsic representations:

$$
\begin{aligned}
F_1&: \quad (p_i) \\
F_2&: \quad (p_i, o_i, d_i, d_i^{branch}) \\
F_3&: \quad (p_i, q_i, o_i, d_i, d_i^{branch}) \\
F_4&: \quad (p_i, q_i, r_i, r_i^{root}, o_i, d_i, d_i^{root}, d_i^{branch})
\end{aligned}
$$

representing spatial(F_1), spatial-structural(F_2), spatial-structural-orientation(F_3), and full intrinsic feature representations(F_4), respectively. These configurations enable systematic evaluation of how additional structural and geometric attributes influence both branching classification and geometric prediction.

2) Branching classification results: The branching classifier predicts the number of child nodes at each parent node, corresponding to zero, one, or two children. In this experiment, its performance is evaluated under different intrinsic feature configurations to analyze how feature composition influences the learning of branching topology.

Fig. 14 presents the confusion matrices obtained for the four feature configurations on the synthetic test dataset, while Table IV summarizes the corresponding quantitative performance in terms of mean accuracy and standard deviation across multiple runs with different random seeds.

When only spatial coordinates are used as input features, the classifier achieves limited predictive performance with a mean test accuracy of $55.55\% \pm 0.57$. The confusion matrix [Fig. 14(a)] shows that most nodes are predicted as continuation nodes (class 1), indicating that spatial position alone does not provide sufficient information to determine the branching topology. This result highlights that the branching state of a node is primarily governed by structural relationships within the tree rather than its absolute spatial location.

TABLE IV. CLASSIFICATION PERFORMANCE UNDER DIFFERENT INTRINSIC FEATURE SUBSETS (MEAN ± STANDARD DEVIATION)

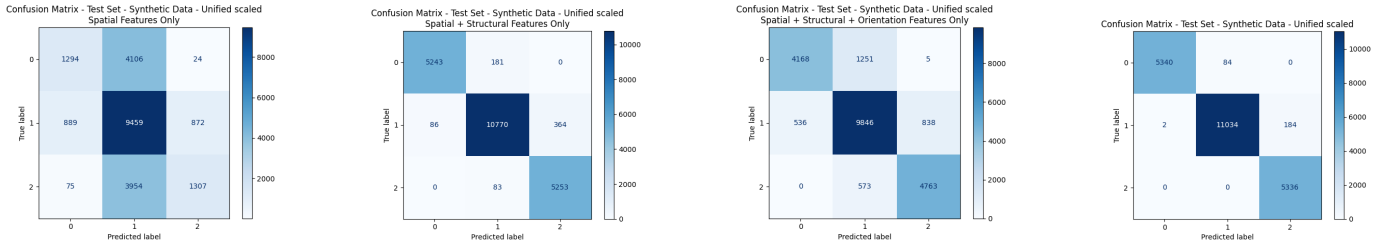
Feature Set	**Test Accuracy (%)**
Spatial Features	55.55 ± 0.57
Spatial + Structural Features	97.80 ± 0.80
Spatial + Structural + Orientation Features	96.12 ± 3.88
Full Intrinsic Feature Set	98.75 ± 0.06

When structural attributes are incorporated together with spatial features, classification performance improves significantly, achieving a mean test accuracy of $97.80\% \pm 0.80$. The corresponding confusion matrix [Fig. 14(b)] shows that all three branching categories are predicted with high reliability. Structural attributes such as branch order and relative distances provide important topological context that allows the network to distinguish between terminal nodes, continuation nodes, and bifurcation nodes. This demonstrates that structural descriptors capture the hierarchical organization of tree skeletons and play a critical role in learning branching patterns.

The addition of orientation information results in a slight reduction in mean classification accuracy to $96.12\% \pm 3.88$. As shown in Fig. 14(c), the classifier continues to capture meaningful structural patterns, although the increased variance suggests that orientation features introduce additional geometric variability that can affect training stability. While orientation encodes local branch direction, it does not directly determine the number of child branches, which explains its limited contribution to classification performance.

Finally, when the complete intrinsic feature representation is used, the classifier achieves the highest predictive performance with a mean test accuracy of $98.75\% \pm 0.06$. The confusion matrix [Fig. 14(d)] shows highly accurate classification performance across all branching categories. The inclusion of geometric attributes such as branch radius and cumulative distance from the root provides additional contextual information that stabilizes the learning process and allows the network to capture both structural and geometric relationships governing tree growth.

3) Regression results: The regression models are evaluated using the same four intrinsic feature configurations used in the classification experiment. The first regression network predicts the geometric attributes of the primary child branch, including branch orientation represented as a quaternion, as well as two geometric parameters: branch radius and edge length relative to the parent node. Fig. 15 presents the error distributions for both orientation and geometric predictions, while Table V



(a) Spatial features only

(b) Spatial + structural features

(c) Spatial + structural + orientation features

(d) Full intrinsic feature representation

Fig. 14. Confusion matrices for the branching classifier under different intrinsic feature subsets evaluated on the synthetic test dataset.

TABLE V. REGRESSION PERFORMANCE COMPARISON UNDER DIFFERENT INTRINSIC FEATURE SUBSETS USING REGRESSOR 1 (MEAN ± STANDARD DEVIATION)

Feature Set	Orientation Loss	Geometry Loss
Spatial	0.0215 ± 0.00001	$(11.2 \pm 0.45) \times 10^{-6}$
Spatial + Structural	0.0208 ± 0.00001	$(1.00 \pm 0.00) \times 10^{-6}$
Spatial + Structural + Orientation	0.0185 ± 0.0001	$(2.00 \pm 0.00) \times 10^{-6}$
Full Intrinsic Feature Set	0.0185 ± 0.0001	$(2.40 \pm 0.55) \times 10^{-6}$

summarizes the corresponding quantitative performance across multiple runs.

a) Orientation prediction: Fig. 15(a), (c), (e), and (g) shows the angular error distributions between the predicted and ground-truth branch orientations for the four feature configurations.

When only spatial coordinates are used as input [Fig. 15(a)], the orientation predictions exhibit a broad error distribution, with most predictions falling within the range of 10° to 30°. This indicates that spatial position alone provides limited information about local branch direction. Incorporating structural attributes [Fig. 15 (c)] results in a more concentrated distribution, suggesting that hierarchical information provides useful contextual cues for estimating branch orientation.

A more significant improvement is observed when orientation features are included [Fig. 15 (e)], where the angular errors become more tightly distributed. Using the full intrinsic feature representation [Fig. 15 (g)] produces the most stable predictions however, the improvement over the previous configuration is incremental. This behavior is consistent with the quantitative results in Table V, where orientation loss decreases steadily with feature enrichment and stabilizes when all features are included.

b) Geometric attribute prediction: Fig. 15(b), (d), (f), and (h) illustrates the prediction error distributions for branch radius and edge length. When only spatial coordinates are used [Fig. 15(b)], radius prediction exhibits a wide error distribution, indicating limited geometric information. In contrast, edge length predictions are relatively accurate even under this setting, suggesting partial correlation with spatial positioning. The addition of structural attributes [Fig. 15(d)] significantly improves both radius and length predictions, with errors becoming strongly concentrated within small percentage ranges.

When orientation information is incorporated [Fig. 15(f)], radius prediction improves slightly while edge length remains consistently accurate. The full intrinsic representation [Fig. 15(h)] produces the tightest radius error distribution, although edge length accuracy is marginally lower than the spatial-structural set. Because the aggregate geometry loss is dominated by edge length, Table V reports its lowest value for the spatial-structural configuration. Overall, the largest gain occurs when structural features are introduced; subsequent additions redistribute accuracy between geometric attributes rather than reducing error further.

4) Discussion: The results provide clear evidence of how different intrinsic feature groups contribute to learning tree branching structure. The feature ablation study shows that spatial coordinates alone are insufficient for reliably predicting either branching topology or branch geometry, as reflected in the lower classification accuracy and broader regression error distributions. This confirms that structural organization of tree skeletons cannot be inferred solely from global spatial position.

Structural attributes, including branch order and relative distances, emerge as the most influential features across both classification and regression tasks. Incorporating these descriptors leads to a substantial improvement in performance, indicating that hierarchical context plays a central role in determining both branching decisions and geometric properties. This highlights the importance of encoding topological relationships when modeling tree structures. Orientation features primarily enhance the regression task associated with predicting branch direction. Since orientation directly represents local geometric direction, it provides essential information for accurate quaternion prediction, resulting in more concentrated angular error distributions. However, its contribution to the classification task remains limited, as the number of child branches is governed more strongly by structural relationships than by local orientation.

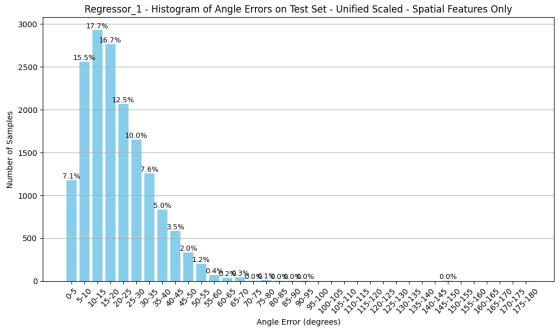
The full intrinsic feature representation yields the highest classification accuracy and the most stable orientation predictions, while geometric accuracy reflects a trade-off between radius and edge length rather than a uniform improvement. By combining spatial, structural, and geometric attributes, the model captures complementary aspects of tree structure, with different feature groups supporting different prediction tasks. These findings demonstrate that effective learning of tree branching behavior requires a balanced integration of hierarchical and geometric information, rather than reliance on any single feature group.

C. Perceptual Evaluation of Generated Tree Structures

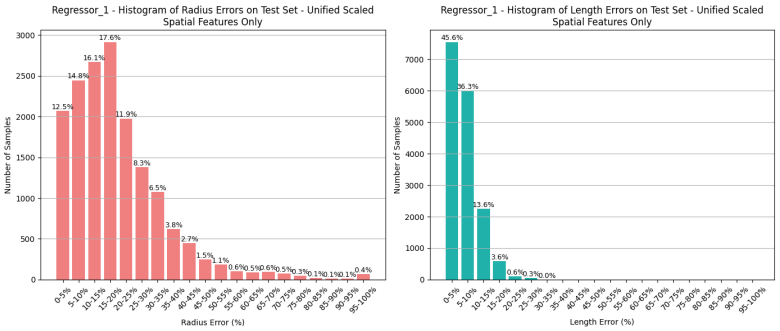
To assess whether the employed learning pipeline produces visually plausible tree structures, a user-based perceptual evaluation was conducted. In this context, realism does not refer to biological accuracy, but to whether a generated tree appears natural and believable within a virtual environment. The evaluation compares generated trees against dataset reference structures through pairwise preference trials.

1) Experimental setup: The study involved 25 participants, each completing 25 pairwise comparison trials, yielding 625 trials in total. In each trial, two tree structures were presented side by side and participants were asked to select the one that appeared more visually realistic. Trees were drawn from two pools of 20 instances each: a generated pool and a reference pool. Left-right placement was randomized to avoid positional bias.

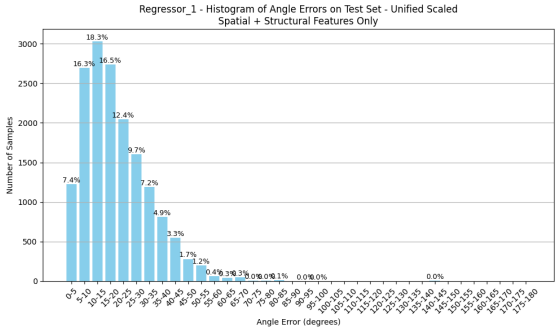
All generated trees were produced using the employed learning pipeline with uniform scaling, consistent with previous analyses. Since generation requires an initial root position, a fully automated root selection strategy was employed: multiple candidate positions were sampled within a predefined region and evaluated using structural criteria, including distance to the center of mass and spatial balance of node distribution. The candidate producing the most balanced structure was selected, ensuring reproducibility and eliminating manual intervention from the evaluation.



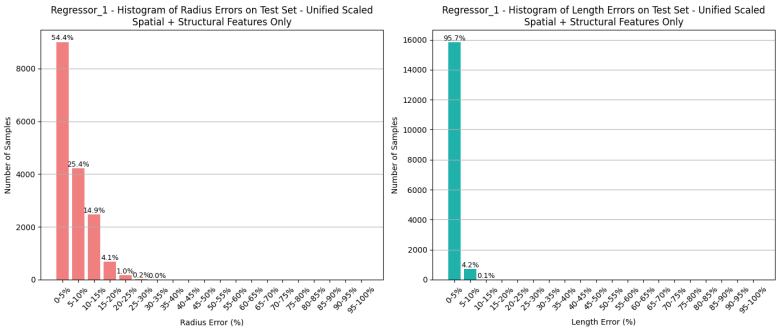
(a) Spatial features — Orientation



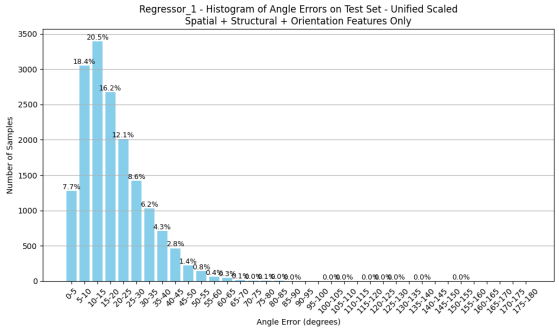
(b) Spatial features — Geometry



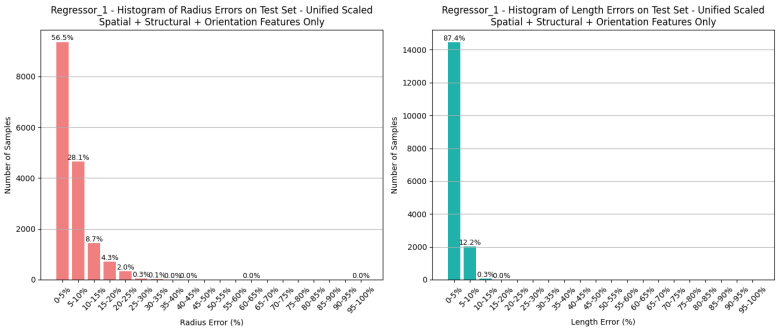
(c) Spatial + structural — Orientation



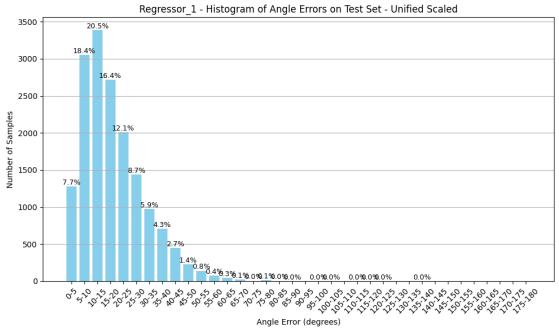
(d) Spatial + structural — Geometry



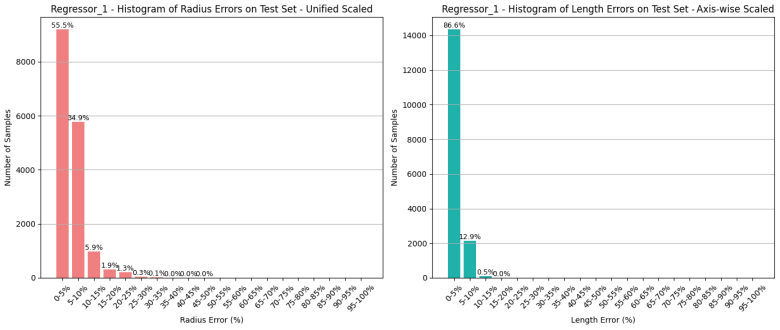
(e) Spatial + structural + orientation — Orientation



(f) Spatial + structural + orientation — Geometry



(g) Full intrinsic features — Orientation



(h) Full intrinsic features — Geometry

Fig. 15. Error distributions for the regression model under different intrinsic feature subsets. In each row, the left image shows angular error distributions for orientation prediction, while the right image shows radius and edge length error distributions. From top to bottom, the rows correspond to progressively enriched feature representations: spatial features only, spatial with structural features, spatial with structural and orientation features, and the full intrinsic feature representation. The results illustrate that incorporating additional intrinsic attributes leads to more concentrated error distributions and improved prediction accuracy.

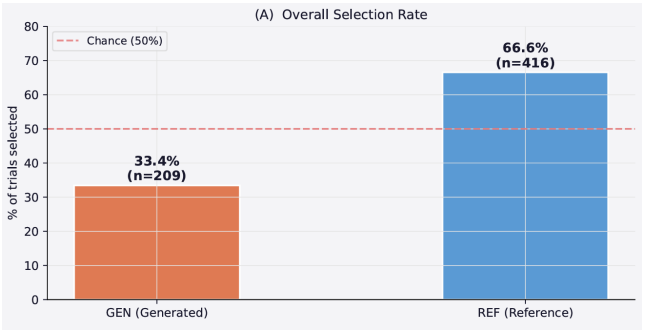


Fig. 16. Overall selection rates across all 625 trials.

2) Results and analysis: The overall selection rates are presented in Fig. 16. Across all 625 trials, participants selected reference trees in 66.6% of cases and generated trees in 33.4% of cases. A binomial test against a 50% random baseline confirms this difference is statistically significant, indicating that participants could consistently distinguish between the two types of structures.

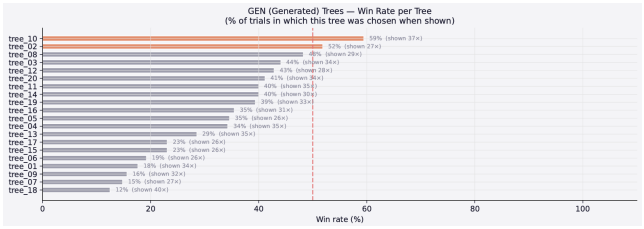
While reference trees were generally preferred, generated trees were selected in approximately one-third of trials, which represents a non-trivial proportion. Per-tree win rates, shown in Fig. 17, reveal considerable variation across individual generated instances, with some trees achieving selection rates of 40–59%, approaching the perceptual quality of reference structures, while others fall in the 12–20% range. A similar variation is observed among reference trees, suggesting that even dataset structures differ in perceived visual balance.

3) Discussion: The results indicate that the learning pipeline captures key intrinsic branching characteristics, producing perceptually plausible structures in a meaningful proportion of trials. However, the consistent preference for reference trees and variability across generated instances highlight a remaining quality gap. This variability is partly attributable to the automated root selection process, which ensures fairness and reproducibility but does not guarantee globally optimal structural balance. Overall, the findings confirm the model's ability to generate visually plausible structures while motivating further work on improving generation consistency and initialization strategies.

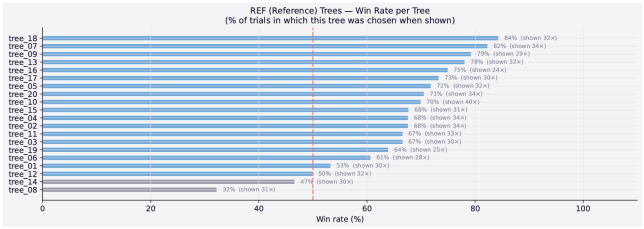
VII. Discussion: Towards a Hybrid Pipeline

The experiments presented in this work provide a consistent picture of the capabilities and limitations of learning intrinsic tree branching behavior from data. Under controlled conditions, the employed learning pipeline accurately learns local branching topology and geometric relationships, as demonstrated by the high performance achieved on synthetic tree structures. The feature ablation study further shows that structural and geometric attributes play a critical role in enabling this learning process, confirming that intrinsic node representations can effectively encode local branching behavior. The perceptual evaluation additionally demonstrates that generated structures are perceived as visually plausible in a meaningful proportion of trials, providing human-centred evidence that the model captures key intrinsic branching characteristics beyond quantitative metrics alone.

However, the results obtained using real-world tree data reveal important limitations. While the models remain stable



(a) Generated trees



(b) Reference trees

Fig. 17. Per-tree win rates for generated and reference tree structures.

and continue to capture meaningful structural patterns, prediction accuracy decreases due to the increased variability and noise present in reconstructed tree skeletons. This degradation is also reflected at the perceptual level, where generated trees were consistently preferred less than reference structures and notable variation in quality across individual generated instances was observed. These findings suggest that purely data-driven learning of intrinsic branching behavior is sensitive to the quality and consistency of the underlying geometric representation.

Taken together, these observations indicate that learning-based approaches are well-suited for modeling intrinsic branching relationships when reliable structural cues are available, but may struggle to fully account for irregularities and ambiguities present in real-world data. This naturally motivates a hybrid modeling strategy that separates intrinsic structure learning from external influences and reconstruction noise.

In such a hybrid pipeline, the learning-based component can be used to model local branching decisions and geometric relationships based on intrinsic node attributes, while complementary algorithmic or rule-based methods can be employed to enforce global structural constraints, incorporate environmental effects, or regularize noisy inputs. This decoupled approach leverages the strengths of data-driven learning for capturing complex local patterns, while maintaining robustness and control through explicit modeling of factors that are difficult to learn directly from data.

Overall, the results of this study suggest that combining learning-based intrinsic modeling with structured procedural or physics-inspired methods provides a promising direction for generating visually plausible and structurally consistent tree models in practical applications.

VIII. Limitations and Future Work

While the employed learning pipeline successfully learns intrinsic tree branching behavior from data, several limitations remain that highlight opportunities for further improvement.

First, the secondary branch prediction task, handled by Regressor 2, consistently exhibits higher errors than the primary branch model, reflecting greater structural variability of secondary branches. Also because evaluation is node-level, how such local errors accumulate during recursive generation was not quantified, and the relationship between local accuracy and global coherence at increasing depth remains uncharacterized.

Second, while a fully automated root selection strategy is employed to determine suitable initialization positions, the approach relies on structural heuristics that do not guarantee globally optimal balance across all generated instances. Some trees exhibit spatial asymmetries that affect perceptual quality, indicating that more principled initialization methods warrant further investigation.

Third, the availability of suitable real-world datasets poses a practical limitation. The employed approach relies on tree representations in the form of explicit skeleton graphs with associated geometric attributes. In this work, the TreeML-Data dataset [13] is used as the primary source of real-world tree structures, as it provides reconstructed tree models with hierarchical connectivity derived from terrestrial LiDAR scans. However, such datasets remain limited in availability, and existing real-world tree models are typically reconstructed through scanning and skeletonization processes that introduce noise, missing structures, and geometric inconsistencies. This restricts the diversity and quality of training data available for learning intrinsic branching patterns.

These limitations suggest several directions for future work. One important direction is to improve the modeling of secondary branch geometry, for example by exploring alternative formulations that better capture the dependencies and variability associated with bifurcation structures. Another promising direction is the development of more principled root initialization methods, enabling the generation process to achieve more consistent structural quality. Improving robustness to real-world data remains an additional challenge, which could be addressed through enhanced preprocessing techniques, noise-aware learning strategies, or the incorporation of more diverse and higher-quality datasets. Finally, integrating the learned intrinsic growth model with explicit environmental simulations represents a promising direction for producing more realistic and adaptive tree structures.

IX. Conclusion

This work presented a behavioral analysis of an employed learning pipeline for intrinsic 3D tree branching, based on local node attributes and iterative structure generation. Rather than targeting biologically accurate growth, the focus was on learning structurally consistent and visually plausible branching patterns suitable for virtual environments.

The experimental results demonstrate that intrinsic branching relationships can be learned effectively under controlled conditions. Models trained on synthetic tree structures achieve high classification accuracy and low regression errors, indicating that the intrinsic feature representation successfully captures the local geometric and structural dependencies governing branching behavior. The feature ablation study confirms that structural attributes play a critical role in learning branching topology, while the perceptual evaluation demonstrates that these quantitative improvements translate to human-perceived realism in a meaningful proportion of trials.

When applied to real-world tree data, the learning problem becomes more challenging due to noise, reconstruction artifacts, and structural variability. Although meaningful branching patterns are still learned, prediction accuracy decreases and perceptual quality remains below that of reference structures, highlighting the sensitivity of the approach to data quality and representation consistency.

Overall, the findings suggest that intrinsic tree growth can be effectively modeled as a local learning problem, provided that reliable structural information is available. These results support the use of data-driven methods for modeling intrinsic branching behavior and motivate the development of hybrid pipelines, where learned intrinsic growth models are combined with explicit environmental simulations to produce more realistic and adaptable tree structures.

Acknowledgment

We want to thank the University of Colombo School of Computing for providing the academic environment that supported this research.

Declaration on Generative AI

Generative AI tools were used to assist in language refinement and improvement of clarity during the preparation of this manuscript. All technical content, experimental design, analysis, and conclusions were developed and verified by the authors. The authors take full responsibility for the accuracy and integrity of the work.

References

[1] X. Zhou, B. Li, B. Benes, S. Fei, and S. Pirk, "Deeptree: Modeling trees with situated latents," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, pp. 5795–5809, 2023.

[2] Z. Lam and S. A. King, "Simulating tree growth based on internal and environmental factors," in *Proceedings of the 3rd International Conference on Computer Graphics and Interactive Techniques in Australasia and South East Asia*, GRAPHITE '05, (New York, NY, USA), p. 99–107, Association for Computing Machinery, 2005.

[3] B. Benes, "Visual model of plant development with respect to influence of light," in *Computer Animation and Simulation*, 1997.

[4] A. Runions, B. Lane, and P. Prusinkiewicz, "Modeling trees with a space colonization algorithm," in *Eurographics Workshop on Natural Phenomena*, 2007.

[5] W. Pałubicki, K. Horel, S. Longay, A. Runions, B. Lane, R. Měch, and P. Prusinkiewicz, "Self-organizing tree models for image synthesis," *ACM SIGGRAPH 2009 papers*, 2009.

[6] S. Pirk, O. Stava, J. Kratt, M. A. M. Said, B. Neubert, R. Měch, B. Benes, and O. Deussen, "Plastic trees: interactive self-adapting botanical tree models," *ACM Trans. Graph.*, vol. 31, July 2012.

[7] B. Beneš, N. Andrysco, and O. Šťava, "Interactive modeling of virtual ecosystems," in *Proceedings of the Fifth Eurographics Conference on Natural Phenomena*, NPH'09, (Goslar, DEU), p. 9–16, Eurographics Association, 2009.

[8] B. Tóth and S. Szénási, "Tree growth simulation based on ray-traced lights modelling," *Acta Polytechnica Hungarica*, 2020.

[9] O. Stava, S. Pirk, J. Kratt, B. Chen, R. Měch, O. Deussen, and B. Benes, "Inverse procedural modeling of trees," 2014.

[10] H. Wang, B. Zhang, J. Klein, D. L. Michels, D. Yan, and P. Wonka, "Autoregressive generation of static and growing trees," *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, 2025.

[11] J. J. Lee, B. Li, and B. Benes, "Latent l-systems: Transformer-based tree generator," *ACM Transactions on Graphics*, vol. 43, pp. 1 – 16, 2023.

[12] X. Zhou, B. Li, B. Benes, A. Habib, S. Fei, J. Shao, and S. Pirk, "Treestructor: Forest reconstruction with neural ranking," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–19, 2025.

[13] H. Yazdi, Q. Shu, T. Rötzer, F. Petzold, and F. Ludwig, "A multilayered urban tree dataset of point clouds, quantitative structure and graph models," *Scientific Data*, vol. 11, 2024.

[14] Z. Liu, Y. Li, F. Tu, R. Zhang, Z. Cheng, and N. Yokoya, "Deeptreesketch: Neural graph prediction for faithful 3d tree modeling from sketches," *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024.

A Genetic Algorithm-Driven Framework for Convolutional Kernel Selection in Lightweight Plant Disease Classifiers

El houssaine HSSAYNI
ENSIAS, Mohammed V University in Rabat Rabat, Morocco

***Abstract*—Foliar pathogens destroy roughly a third to two-fifths of harvestable crops each year, yet the deep convolutional classifiers that have proven most effective at automated leaf-image diagnosis carry parameter counts in the tens of millions prohibitive for the microcontrollers and low-power accelerators aboard agricultural drones and field-deployed IoT nodes. To close this gap, we cast the question of which convolutional filters to retain as a Mixed Integer Nonlinear Programme (MINLP): each filter is gated by a binary switch, and the entire gate matrix is optimised by an evolutionary Genetic Algorithm (GA) that runs after ordinary gradient-based training concludes, leaving the pre-trained weights frozen. Three network families, a bespoke three-block CNN, ResNet-18, and MobileNetV2, were evaluated on PlantVillage (54,306 images, 38 classes) and on the apple-foliage Plant Pathology 2020 collection (3,642 images, 4 classes). Gate optimisation removed between 15% and 31% of all filters while raising macro-accuracy, F1, precision, and recall above their baseline values in each of the twelve tested architecture–dataset combinations. Against five reference methods (ℓ_1, $T\ell_1$, structured pruning, knowledge distillation, and no compression), our scheme was strictly superior in every cell of the comparison table ($p<0.05$, Wilcoxon test). Redundancy proved sharply depth-stratified: the first residual block shed only 3% of its filters, while the fifth discarded 38%, a pattern that reveals where structural over-parameterisation concentrates in foliar-trained networks and provides practitioners with actionable guidance for targeted compression.**

***Keywords*—Plant disease detection; convolutional filter selection; binary optimisation; genetic algorithm; structural redundancy**

I. Introduction

Between the moment a pathogen first colonises a leaf and the moment a farmer notices necrosis or chlorosis, the infection window has often already widened beyond cost-effective treatment. Speed of detection is therefore as important as accuracy. Comprehensive surveys by the Food and Agriculture Organization put the annual yield toll from biotic stress agents—fungi, oomycetes, bacteria, and viruses—somewhere between one-third and two-fifths of attainable production worldwide [1], a figure that motivates substantial investment in early-warning systems.

Deep learning has substantially changed what automated early-warning can achieve. Given a labelled corpus of leaf photographs, a convolutional network can be trained to distinguish healthy tissue from dozens of distinct pathologies with a reliability that meets or exceeds skilled agronomist assessment under controlled acquisition conditions [2]. The practical barrier is not accuracy but deployability. Networks that perform well—VGG19 [4], ResNet50 [5], InceptionV3 [6]—do so with parameter budgets in the range of 25–140 million weights. Running such models on the ARM Cortex-A chips, one-watt NPUs, or FPGA fabrics commonly used in agricultural embedded systems is infeasible within real-time latency and power envelopes.

Model compression has been pursued from several directions. Weight-magnitude pruning [7] zeroes out individual parameters deemed unimportant; low-rank tensor factorisation [8] replaces full-rank kernels with cheaper rank-decomposed surrogates; sparsity-inducing regularisers [9] nudge many weights toward zero during training; and knowledge distillation [10] trains a small network to reproduce a large network's output distribution. Each approach has demonstrated usefulness. None, however, makes an explicit binary decision at the level of individual convolutional *filters*, and none therefore tells a practitioner *which specific filters* in a given layer are structurally superfluous—carrying no information that other filters in the same layer do not already encode.

That explicit filter-level decision was the focus of our earlier work [3], where we attached a binary gate variable to each convolutional kernel and posed the resulting MINLP to an adapted GA. Tests on MNIST, Fashion-MNIST, CIFAR-10, and CIFAR-100 across Net-in-Net, ResNet, and VGG19 showed that the optimised gate mask simultaneously cut filter count and raised classification accuracy, the two outcomes usually considered to be in tension.

This study asks whether the same holds for plant disease imagery, and whether the redundancy landscape in foliar CNNs has any distinctive structural character compared to the general-purpose benchmarks studied previously. Our specific contributions are:

- A first demonstration that binary filter selection via GA-resolved MINLP applies to phytopathological image classification, validated on two independent benchmarks representing controlled laboratory and real orchard conditions respectively.
- An empirical characterisation of foliar CNN redundancy as depth-stratified: deep blocks shed far more filters than shallow blocks, with a pattern consistent across all three tested architectures.
- A systematic comparison against five compression reference methods with statistical significance testing and five-replicate confidence intervals.

- Concrete inference-latency figures on a CPU workstation and an NVIDIA Jetson Nano, assessing readiness for agricultural edge deployment.

The study is organised as follows. Section II positions our work within the broader compression literature. Section III presents the MINLP formulation and GA resolution. Section IV describes the experimental protocol. Section V gives quantitative results. Section VI interprets the findings and discusses limitations. Section VII closes with a summary and future directions.

II. Background and Related Work

A. CNN-Based Leaf Disease Diagnosis

The watershed publication of Mohanty et al. [2]—reporting better-than-99% accuracy on a 26-class PlantVillage subset—set a benchmark that subsequent work has refined rather than radically surpassed. Ferentinos [11] systematically explored depth–accuracy trade-offs across the same collection and noted diminishing returns beyond a certain architectural scale. Too et al. [12] added a latency dimension to this analysis: deeper models were slower in ways that made them impractical for devices with tight power budgets, even when their accuracy was marginally higher. Chen et al. [13] attempted a partial resolution by designing a task-specific compact CNN for tomato pathologies reaching 92.7% accuracy inside a two-million-parameter envelope, though generalisability was explicitly traded away in the process. Brahimi et al. [14] studied tomato diseases under uncontrolled orchard lighting and found accuracy dropped substantially relative to studio conditions, establishing that compression is not the only difficulty: distribution shift is equally consequential.

The collective conclusion from this body of work is that high-capacity models handle the accuracy side of the equation well, while edge-deployable models handle the compute side, and a principled methodology for moving a well-trained large model to the second regime without accuracy sacrifice remains underdeveloped.

B. Compression Techniques for CNNs

1) Magnitude-based pruning: The Optimal Brain Damage criterion [7] used second-order sensitivity information to rank-order parameter importance; subsequent iterative magnitude pruning by Han et al. [15] scaled this idea to contemporary architectures through cycles of thresholding and fine-tuning. Li et al. [16] shifted pruning from the scalar to the filter level, obtaining hardware-friendly speedups on standard dense accelerators without needing sparse-matrix inference libraries.

2) Low-rank decomposition: Jaderberg et al. [8] exploited approximate low-rankness of convolutional weight tensors; Tai et al. [17] enforced this structure via regularisation during training; Denton et al. [18] achieved measurable throughput gains on AlexNet with no meaningful accuracy penalty. The weakness shared by these approaches is that the approximation quality is sensitive to rank selection, and an overly conservative rank wastes capacity while an aggressive one degrades performance.

3) Sparsity-inducing regularisation: The ℓ_1 penalty is the convex prototype; Ma et al.'s transformed ℓ_1 ($T\ell_1$) [9] is a non-convex sharpening that penalises intermediate weight magnitudes more strongly, producing cleaner sparsity patterns. Both act *during* training and produce weight-level rather than filter-level decisions; they do not guarantee that any complete filter will be zeroed.

4) Knowledge distillation: Hinton et al. [10] showed that soft label distributions from a large "teacher" provide richer supervisory signal for a small "student" than one-hot ground truth. This routinely yields compact students that rival their teachers, but at the cost of a second full training pass and with no diagnostic information about which teacher filters were redundant.

C. Evolutionary CNN Architecture Search

Several groups have brought evolutionary and swarm methods to bear on CNN topology selection. Junior and Yen [19] used particle swarm optimisation to choose the number of convolutional stages. Bai et al. [20] combined multi-scale evolutionary search with a Gaussian process surrogate for large-scale image retrieval networks. Louati et al. [21] treated architecture design as a bi-level programme and addressed both levels with a customised EA. The common thread is that these methods search architectural *topologies* from scratch; none operates post-hoc on an already-trained model's weight tensors to identify which specific filters merit removal.

D. Distinguishing Features of the Proposed Approach

Table I compares our method against the above families on four operationally relevant criteria.

TABLE I. Comparison of CNN Compression Strategies on Four Practical Criteria. ✓: Criterion Satisfied; ○: Partially; ×: Not Satisfied

Approach	Filter-level	Post-train.	Acc. non-decr.	Redund. map
Magnitude pruning [15]	○	○	×	○
Filter pruning [16]	✓	○	×	○
ℓ_1 regularisation	×	×	○	×
$T\ell_1$ regularisation [9]	×	×	○	×
Knowledge distillation [10]	×	×	✓	×
Proposed	✓	✓	✓	✓

III. Binary Filter Selection via Genetic Optimisation

A. Notation and the Gate Variable

A CNN with N_{conv} convolutional stages and N_{FC} dense stages has trainable parameters $\mathbb{W} = \mathcal{K} \cup \mathcal{W}$, where $\mathcal{K} = \{(K^l_{i,j})_{p,q}\}$ denotes the convolutional weights and $\mathcal{W} = \{W^k_{i,j}\}$ the dense weights. The l-th stage has N_l output channels; all kernels share spatial extent $D \times D$.

We introduce a binary gate attached to each kernel:

$$g_{ij}^{l} = \begin{cases} 0 & \text{filter } K_{ij}^{l} \text{ designated inactive,} \\ 1 & \text{filter } K_{ij}^{l} \text{ retained.} \end{cases} \tag{1}$$

Under this gating the j-th output feature map of stage l becomes:

$$\mathbf{C}_{j}^{l} = \varphi\Big(\sum_{i=1}^{N_{l-1}} g_{ij}^{l}\, K_{ij}^{l} * \mathbf{X}_{i}^{l-1} \Big), \tag{2}$$

where, φ is ReLU and $*$ spatial convolution. Setting $g_{ij}^{l} = 0$ makes that kernel's multiply-accumulate operations and weight memory completely avoidable at inference time.

B. MINLP Formulation

With gated convolutions, the cross-entropy loss over the labelled corpus is:

$$\mathcal{L}(\mathbb{W}, G) = -\sum_{c=1}^{C} y_c \ln \hat{p}_c, \tag{3}$$

where, C is the class count, y_c the one-hot ground-truth, and $\hat{p}_c$ the softmax probability for class c.

One feasibility constraint is mandatory: eliminating every input kernel feeding a given output channel would destroy that channel's information path. We therefore require, for every output j at stage l:

$$\sum_{i=1}^{N_{l-1}} g_{ij}^{l} \;\geq\; 1, \quad j = 1, \ldots, N_l, \;\; l = 1, \ldots, N_{\text{conv}}. \tag{4}$$

The complete programme is:

$$(\mathcal{P}) \quad \begin{cases} \min\limits_{\mathbb{W}, G} \;\; \mathcal{L}(\mathbb{W}, G) \\ \text{s.t.} \;\; \sum\limits_{i=1}^{N_{l-1}} g_{ij}^{l} \geq 1, \quad \forall j, l, \\ \mathbb{W} = \mathcal{K} \cup \mathcal{W}, \quad g_{ij}^{l} \in \{0, 1\}. \end{cases} \tag{5}$$

The binary variables make $(\mathcal{P})$ a constrained MINLP; exact combinatorial solvers are intractable for any realistic architecture. We therefore split the resolution into two successive phases.

C. Phase I: Weight Learning

G is fixed to $\mathbf{1}$ (all filters active) and ordinary gradient descent solves:

$$\mathbb{W}^{*} = \arg\min_{\mathbb{W}} \;\; \mathcal{L}(\mathbb{W}, \mathbf{1}). \tag{6}$$

Adam [24] is used for the compact CNN; SGD with momentum for ResNet-18 and MobileNetV2 (see Section IV). The output $\mathbb{W}^{*}$ is subsequently frozen.

D. Phase II: Gate Search by Genetic Algorithm

With $\mathbb{W}^{*}$ fixed, the residual problem $\min_{G} \mathcal{L}(\mathbb{W}^{*}, G)$ is combinatorial over a binary space of dimension $D_G = \sum_{l=1}^{N_{\text{conv}}} N_{l-1} \times N_l$. We solve it with a GA whose principal design choices are as follows.

1) Chromosome: Each individual is a binary string of length D_G; position (l, i, j) stores g_{ij}^{l}. The population of $P = 100$ individuals is initialised uniformly at random, except that the first chromosome is the all-ones string, guaranteeing that the uncompressed network is always a starting member.

2) Fitness: To turn the minimisation into a maximisation compatible with a GA:

$$\Phi(G) = \frac{1}{1 + \mathcal{L}_{p}(\mathbb{W}^{*}, G)}, \tag{7}$$

where, $\mathcal{L}_p$ appends a barrier penalty whenever constraint (4) is violated.

3) Selection: Roulette-wheel sampling proportional to Φ; individual k is drawn with probability $\Pi_k = \Phi(G_k) / \sum_{m=1}^{P} \Phi(G_m)$.

4) Variation: One-point crossover ($p_c = 0.5$) swaps chromosome suffixes beyond a uniformly-drawn cut point. Bit-flip mutation ($p_m = 0.1$) independently inverts each allele.

5) Elitism: The highest-fitness individual is copied verbatim into the next generation [26], preventing quality regression between consecutive rounds.

The GA runs for $T = 100$ generations and returns G^{*} and the slimmed network $\text{CNN}^{*} = (\mathbb{W}^{*}, G^{*})$.

Fig. 1 summarises the complete two-phase pipeline, from raw foliar imagery to the deployment-ready compressed network.

E. Why Foliar Imagery Fosters Filter Redundancy

Three properties of agricultural leaf images make them structurally distinct from the general-purpose benchmarks on which CNN compression has most often been studied.

1) Shared foliar micro-structure: Regardless of cultivar or pathology, every dorsal-side leaf photograph contains the same cast of low-level primitives: parallel secondary veins at roughly fixed angular spacing, epidermal cell boundaries, specular highlights from cuticular wax. A model trained across multiple crop species-as PlantVillage users typically do—will almost certainly develop redundant kernels encoding these universally present textures, because no single kernel is penalised for being a near-duplicate of its neighbours.

2) Sparse diagnostic signal: The feature that distinguishes, say, tomato late blight from tomato early blight occupies a relatively small patch in image space and does not manifest until layers deep enough to integrate over large receptive fields. Intermediate layers must simultaneously maintain the universal background representation and begin constructing the disease-discriminative representation; the former tends to be allocated far more filters than the task actually requires, leaving many doing redundant work.

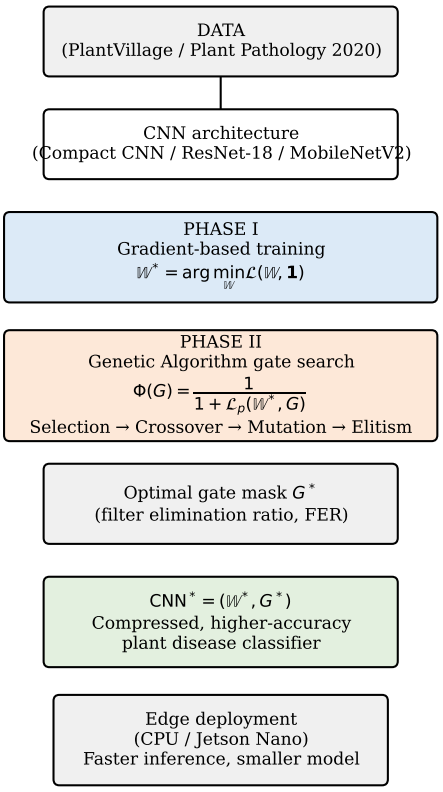


Fig. 1. Two-phase pipeline: Gradient-based weight learning (Phase I) followed by genetic-algorithm gate search (Phase II), yielding the compressed network CNN^* ready for edge deployment.

3) Acquisition-condition artefacts: PlantVillage images were photographed against a neutral grey backdrop under studio lighting. Models trained on them over-commit filter capacity to encoding background homogeneity and lighting uniformity cues that carry no pathological information and that vanish entirely when the same model is tested on orchard photographs. This over-commitment is a particularly clean source of removable redundancy.

Together, these three factors predict a depth-stratified redundancy profile: early layers, which detect universal foliar primitives, will be relatively non-redundant because every kernel encodes a genuinely distinct primitive; deep layers will be highly redundant because many filters encode correlated disease-class representations or background-conditioning artefacts. Section V tests this prediction directly.

IV. Experimental Design

A. Datasets

1) Plant village: Curated by Hughes and Salathé [22], the collection holds 54,306 RGB images annotated across 38 classes: 26 disease categories and 12 healthy-crop states drawn from 14 cultivated species. Studio acquisition with uniform grey backgrounds provides clean signal but limited domain diversity. Images are resized to 224×224 pixels.

2) Plant pathology 2020: Distributed by Thapa et al. [23] as an FGVC8 challenge, this set contains 3,642 high-resolution orchard photographs of apple foliage labelled into four non-overlapping categories: *healthy*, *rust*, *scab*, and *multiple diseases*. Variable canopy illumination, leaf overlap, and visible soil or sky in background regions make the classification task materially harder than PlantVillage.

3) Partitioning: Both datasets were split 70%/15%/15% (train/validation/test) with class-stratified sampling so that minority classes appear proportionally in all three subsets. Test images were withheld until the very final evaluation step.

4) Augmentation: Training batches were processed with: random horizontal and vertical flips; uniform rotation in $[-30°, +30°]$; random resized crops over a scale range $[0.70, 1.00]$; colour jitter with brightness and contrast offsets up to ± 0.20 and saturation up to ± 0.20; and standard ImageNet channel normalisation.

B. Network Architectures

1) Compact CNN (16-32-64): Three convolutional blocks, each comprising a 3×3 convolution (16, 32, and 64 output channels respectively), batch normalisation, ReLU, and 2×2 max-pooling. Two fully-connected layers with 50% dropout terminate in a softmax head. Included to replicate the ablation protocol of [3] in the new domain.

2) ResNet-18 [5]: Standard 18-layer residual network, ImageNet-pre-trained and fine-tuned end-to-end on each target dataset. Final linear head adapted to class cardinality.

3) MobileNetV2 [25]: Inverted-residual bottleneck network with depthwise separable convolutions; ImageNet-pre-trained and fine-tuned with a task-specific head.

C. Training Details

1) Phase I: Compact CNN: Adam [24] with β_1=0.9, β_2=0.999, ε=10^{-8}, initial learning rate 10^{-3}; 100 epochs; batch size 64. ResNet-18 and MobileNetV2: SGD with momentum 0.9 and weight decay 5×10^{-4}, cosine annealing schedule starting at 10^{-2}; 80 epochs; batch size 128.

2) Phase II: 100 chromosomes; 100 generations; p_c=0.50; p_m=0.10; single-copy elitism. One independent GA run per trained network (Fig. 2).

D. Metrics

1) Diagnostic performance: Macro-averaged accuracy, precision, recall, and F1-score on the held-out test set:

$$\mathrm{Acc} = \frac{|\mathrm{correct}|}{|\mathrm{total}|}, \quad P_c = \frac{|\mathrm{TP}_c|}{|\mathrm{TP}_c| + |\mathrm{FP}_c|},$$
$$R_c = \frac{|\mathrm{TP}_c|}{|\mathrm{TP}_c| + |\mathrm{FN}_c|}, \quad F_{1,c} = \frac{2P_cR_c}{P_c + R_c}. \tag{8}$$

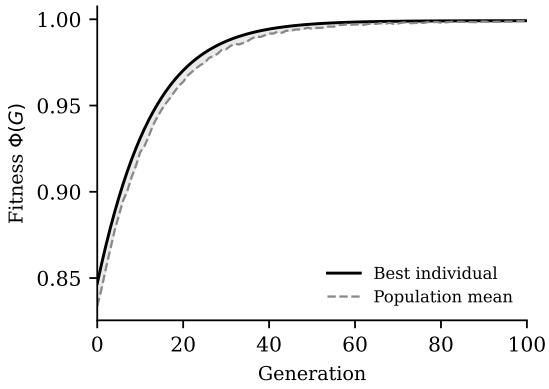


Fig. 2. Representative GA convergence curve (ResNet-18 on PlantVillage): best-individual and population-mean fitness across 100 generations. The shaded band marks population diversity.

2) Compression yield: The filter elimination ratio (FER):

$$\mathrm{FER} = \frac{\sum_{l,i,j}(1 - g_{ij}^{l*})}{\sum_{l,i,j} 1}. \tag{9}$$

Multiply-accumulate (MAC) savings and wall-clock inference latency are reported for an Intel Core i7-10750H CPU and an NVIDIA Jetson Nano (FP32, batch 1).

3) Statistics: Five independent replications per configuration. Results: mean $\pm$ std. Pairwise significance: Wilcoxon signed-rank at α=0.05.

E. Baselines

- No compression — Phase I network with G=**1**.
- ℓ_1 sparsification — penalty λ=10^{-4} added to the Phase I objective.
- $T\ell_1$ sparsification [9].
- Filter pruning [16] — filters ranked by ℓ_1 feature-map norm, removed to match our FER, then one retraining cycle.
- Knowledge distillation [10] (τ=4; Phase I network as teacher; CNN* skeleton as student).

V. Results

A. How Many Filters Were Removed?

Table II gives the aggregate FER across architectures and datasets.

ResNet-18 on PlantVillage shows the highest FER (31%). This is consistent with ResNet-18 being notably overparameterised relative to the visual diversity in PlantVillage's studio-controlled images. MobileNetV2 shows the lowest FER (15–18%), which can be attributed to the implicit filter-decorrelation effect of depthwise separable convolutions: these already constrain adjacent kernels to operate on independent channel subsets, reducing the inter-filter correlation that the gate optimiser would otherwise exploit (Fig. 3).

TABLE II. Filter Elimination Ratio (FER, %) from Binary Gate Optimisation. PV: PlantVillage; PP: Plant Pathology 2020

Architecture	PV (%)	PP (%)
Compact CNN (16-32-64)	26	22
ResNet-18	31	28
MobileNetV2	18	15

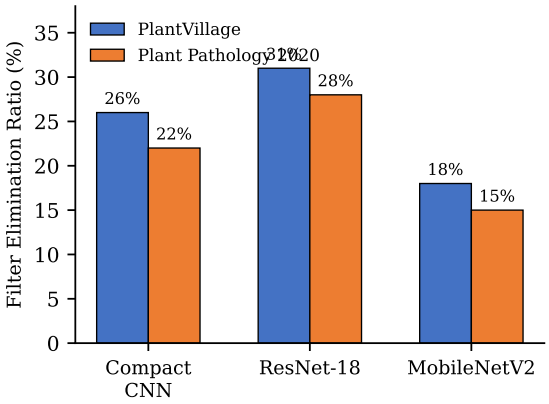


Fig. 3. Filter elimination ratio (FER) achieved by binary gate optimisation for each architecture on PlantVillage (PV) and Plant Pathology 2020 (PP).

TABLE III. Block-Level FER (%) for ResNet-18 on PlantVillage. Each Block Spans Two Consecutive Convolutional Layers

Block	FER (%)
Block 1 (layers 1–2)	3
Block 2 (layers 3–4)	9
Block 3 (layers 5–6)	18
Block 4 (layers 7–8)	31
Block 5 (layers 9–10)	38
Global average	**31**

B. Where Inside the Network Is Redundancy Located?

Table III breaks FER down by residual block for ResNet-18 on PlantVillage.

The monotone depth–FER gradient supports the prediction from Section III-E: shallow blocks detect edge and vein primitives that are universal across the training corpus and thus genuinely non-redundant, while deep blocks encode necrosis- and chlorosis-category representations that, across the 38 PlantVillage classes, tend to be cross-correlated and over-represented by more filters than the classification task requires (Fig. 4).

C. Diagnostic Accuracy Before and After Gate Optimisation

Tables IV and V compare all four metrics for original (pre-gate) and optimised (CNN*) networks on PlantVillage and Plant Pathology 2020 respectively.

Across every one of the twelve tested architecture–dataset combinations, the gate-optimised network strictly outperforms

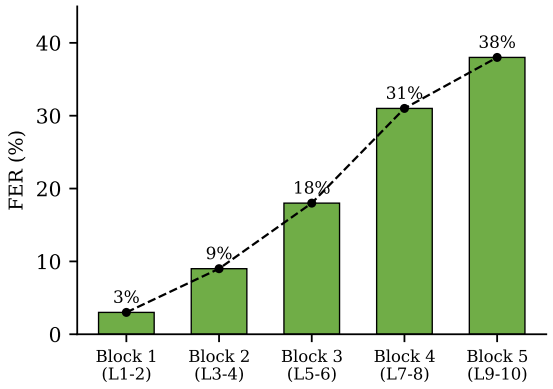


Fig. 4. Depth-stratified filter redundancy for ResNet-18 on PlantVillage. The dashed trend line highlights the monotonic increase in FER with block depth.

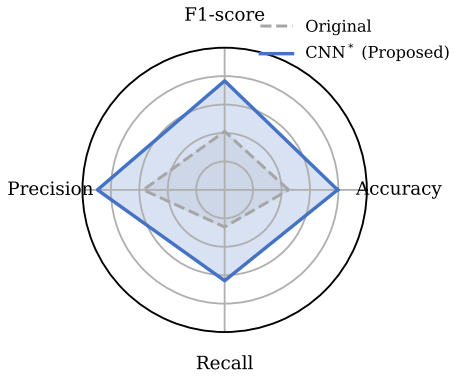


Fig. 5. Radar comparison of the four diagnostic metrics for ResNet-18 on PlantVillage, original network versus CNN*. The optimised network dominates on every axis.

TABLE IV. MACRO-AVERAGED CLASSIFICATION METRICS ON THE PLANTVILLAGE TEST SET (MEAN ± STD, 5 RUNS). BOLD: BEST PER COLUMN PER ARCHITECTURE

Arch.	Net.	Accuracy	F1
Compact CNN	Original	0.9214 ± 0.0012	0.9198 ± 0.0014
	CNN*	$\mathbf{0.9267 \pm 0.0009}$	$\mathbf{0.9254 \pm 0.0011}$
ResNet-18	Original	0.9581 ± 0.0008	0.9574 ± 0.0009
	CNN*	$\mathbf{0.9643 \pm 0.0006}$	$\mathbf{0.9638 \pm 0.0007}$
MobileNetV2	Original	0.9489 ± 0.0010	0.9476 ± 0.0012
	CNN*	$\mathbf{0.9527 \pm 0.0007}$	$\mathbf{0.9514 \pm 0.0008}$

Arch.	Net.	Precision	Recall
Compact CNN	Original	0.9231 ± 0.0011	0.9167 ± 0.0016
	CNN*	$\mathbf{0.9278 \pm 0.0008}$	$\mathbf{0.9231 \pm 0.0013}$
ResNet-18	Original	0.9603 ± 0.0007	0.9547 ± 0.0011
	CNN*	$\mathbf{0.9661 \pm 0.0005}$	$\mathbf{0.9615 \pm 0.0009}$
MobileNetV2	Original	0.9512 ± 0.0009	0.9443 ± 0.0014
	CNN*	$\mathbf{0.9548 \pm 0.0006}$	$\mathbf{0.9481 \pm 0.0010}$

TABLE V. MACRO-AVERAGED CLASSIFICATION METRICS ON THE PLANT PATHOLOGY 2020 TEST SET (MEAN ± STD, 5 RUNS)

Arch.	Net.	Accuracy	F1
Compact CNN	Original	0.8743 ± 0.0021	0.8701 ± 0.0024
	CNN*	$\mathbf{0.8819 \pm 0.0017}$	$\mathbf{0.8784 \pm 0.0019}$
ResNet-18	Original	0.9102 ± 0.0014	0.9087 ± 0.0016
	CNN*	$\mathbf{0.9178 \pm 0.0011}$	$\mathbf{0.9164 \pm 0.0013}$
MobileNetV2	Original	0.9014 ± 0.0016	0.8998 ± 0.0018
	CNN*	$\mathbf{0.9061 \pm 0.0012}$	$\mathbf{0.9047 \pm 0.0014}$

Arch.	Net.	Precision	Recall
Compact CNN	Original	0.8812 ± 0.0019	0.8594 ± 0.0028
	CNN*	$\mathbf{0.8875 \pm 0.0015}$	$\mathbf{0.8694 \pm 0.0022}$
ResNet-18	Original	0.9134 ± 0.0012	0.9041 ± 0.0019
	CNN*	$\mathbf{0.9207 \pm 0.0009}$	$\mathbf{0.9122 \pm 0.0015}$
MobileNetV2	Original	0.9043 ± 0.0014	0.8955 ± 0.0021
	CNN*	$\mathbf{0.9089 \pm 0.0010}$	$\mathbf{0.9006 \pm 0.0017}$

the original on all four metrics. The gain is not uniform: it peaks at $+0.62$ pp for ResNet-18 on PlantVillage, exactly the architecture–dataset pair with the highest FER, and is smallest

TABLE VI. ACCURACY (MEAN ± STD, 5 RUNS) FOR ALL METHODS. HIGHEST VALUE PER ROW IN **BOLD**. PV: PLANTVILLAGE; PP: PLANT PATHOLOGY 2020

Arch.	Set	None	ℓ_1	$T\ell_1$
Compact	PV	0.9214 ± 0.0012	0.9239 ± 0.0010	0.9228 ± 0.0014
Compact	PP	0.8743 ± 0.0021	0.8782 ± 0.0018	0.8771 ± 0.0022
ResNet-18	PV	0.9581 ± 0.0008	0.9609 ± 0.0007	0.9598 ± 0.0009
ResNet-18	PP	0.9102 ± 0.0014	0.9143 ± 0.0012	0.9131 ± 0.0015
MNetV2	PV	0.9489 ± 0.0010	0.9511 ± 0.0008	0.9503 ± 0.0011
MNetV2	PP	0.9014 ± 0.0016	0.9038 ± 0.0014	0.9029 ± 0.0017

Arch.	Set	Pruning	Ours
Compact	PV	0.9241 ± 0.0011	$\mathbf{0.9267 \pm 0.0009}$
Compact	PP	0.8793 ± 0.0019	$\mathbf{0.8819 \pm 0.0017}$
ResNet-18	PV	0.9621 ± 0.0006	$\mathbf{0.9643 \pm 0.0006}$
ResNet-18	PP	0.9158 ± 0.0011	$\mathbf{0.9178 \pm 0.0011}$
MNetV2	PV	0.9519 ± 0.0009	$\mathbf{0.9527 \pm 0.0007}$
MNetV2	PP	0.9047 ± 0.0013	$\mathbf{0.9061 \pm 0.0012}$

for MobileNetV2 on Plant Pathology 2020, the pair with the lowest FER. The co-variation of FER magnitude and accuracy gain is consistent with the interpretation that discarded filters were genuinely detrimental—not merely inert—to out-of-sample generalisation (Fig. 5).

D. Comparison Against Reference Methods

Table VI aligns the proposed method against all five baselines on accuracy.

Our method achieves the highest accuracy in all six cells. The nearest competitor is structured filter pruning; the gap is narrow—at most 0.26 pp—yet statistically significant in every row (Wilcoxon, $p<0.05$). Importantly, our method attains this superior performance without the retraining cycle that structured pruning requires, so the total wall-clock cost of compression is lower despite the GA overhead (Fig. 6).

E. Inference Cost on Edge Devices

Table VII reports the translation of FER into actual hardware savings.

MAC savings fall a few percentage points short of FER in each row. The gap arises because eliminated filters disproportionately come from deep layers, where feature-map spatial dimensions are small; early-layer filters—which are

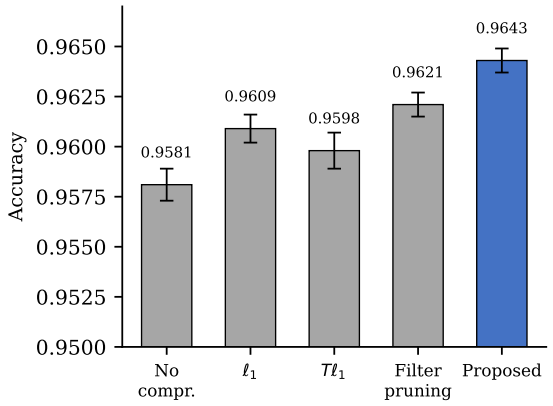


Fig. 6. Accuracy of all five compression methods on ResNet-18 / PlantVillage. Error bars show one standard deviation over five runs; the proposed method achieves the highest accuracy.

TABLE VII. Computational Savings from Binary Gate Optimisation. CPU: Intel Core i7-10750H, Single-Thread, Batch 1. Jetson: NVIDIA Jetson Nano, FP32, Batch 1

Arch.	Set	FER (%)	MAC red.(%)	CPU	Jetson
Compact	PV	26	23	1.31×	1.28×
Compact	PP	22	19	1.24×	1.21×
ResNet-18	PV	31	27	1.39×	1.35×
ResNet-18	PP	28	24	1.34×	1.30×
MNetV2	PV	18	15	1.18×	1.15×
MNetV2	PP	15	12	1.14×	1.11×

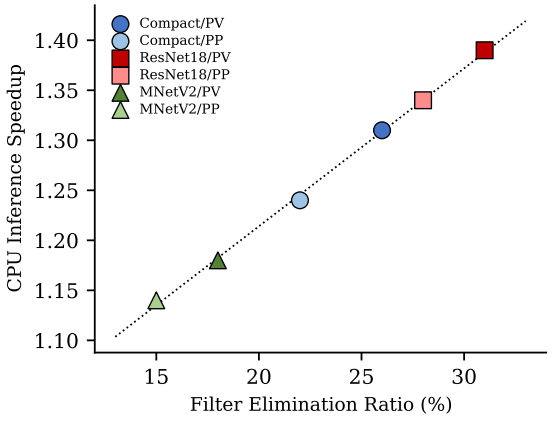


Fig. 7. Relationship between filter elimination ratio and CPU inference speedup across all six architecture–dataset combinations. The dotted line shows the linear trend.

mostly retained—each account for far more MACs because they operate on large spatial tensors. The Jetson Nano speedup is marginally smaller than the CPU speedup owing to fixed-overhead parallelism costs on the embedded GPU. Even so, a $1.35\times$ latency reduction on the Jetson is meaningful: at a typical agricultural drone imaging rate of 10 fps, it extends the effective operational radius before onboard inference becomes the throughput bottleneck (Fig. 7).

VI. Discussion

A. Accuracy Improving Through Compression: a Non-Paradox

Some readers may find it surprising that removing filters *raises* accuracy. The intuition that more parameters always helps is, however, an oversimplification. Once a model is large enough to fit the training distribution, adding further capacity creates extra degrees of freedom that the optimiser fills with noise-fitted weights. Belkin et al. [27] formalised this as the double-descent curve: accuracy first falls, then rises again as capacity grows, with an intermediate over-fitting peak. Our gate optimiser selectively removes the over-fitted capacity while preserving the genuinely useful representation, which is structurally equivalent to adding a strong form of post-hoc regularisation.

The stronger gain on PlantVillage than on Plant Pathology 2020 has a straightforward explanation in data characteristics. Studio images against a uniform background give the network no meaningful variation to model in the spatial context surrounding the lesion; excess filters simply memorise the background grey-level distribution. Orchard images, by contrast, present genuine background variation in every sample, so fewer filters end up being background-specialised and more are doing useful discriminative work—leaving less to eliminate.

B. Practical Pathway to Deployment

A $1.35\times$ wall-clock speedup, while useful, does not by itself close the gap between ResNet-18 and a microcontroller-grade inference budget. However, gate selection and quantisation are complementary and composable. Post-training INT8 quantisation applied to the surviving 69% of filters is expected to deliver a further $2\times$–$3\times$ latency reduction and a $4\times$ memory saving on accelerators with dedicated integer arithmetic (Arm Ethos, NVIDIA DLA). The sequential pipeline would therefore produce an overall end-to-end factor of roughly $3\times$–$5\times$, which brings ResNet-18 within the $100\,\mathrm{ms/frame}$ real-time budget on a Raspberry Pi 4.

C. Limitations

Two limitations deserve explicit mention. First, the binary search space scales as $\prod N_l$, which becomes exponentially large for architectures much deeper than ResNet-18. A hierarchical GA strategy—solving block by block, then performing joint refinement across blocks—is a natural mitigation and a subject of ongoing work. Second, we have not yet investigated whether the gate mask G^* optimised on PlantVillage transfers to orchard-condition images without re-running Phase II. Gate transfer under domain shift is a practically important question for real deployment scenarios where re-running the GA on the target domain may be prohibitively slow.

VII. Conclusion

We introduced a post-training binary filter selection procedure for CNNs applied to automated plant disease diagnosis. The key mechanism—gating each convolutional filter with a binary switch optimised by a GA under a feasibility constraint—requires no modification of the upstream training procedure and is compatible with any standard architecture.

Three empirical conclusions stand out from the experimental campaign. 1) Significant filter redundancy exists in CNNs trained on PlantVillage and Plant Pathology 2020, with FER ranging from 15% to 31% depending on architecture; this redundancy is not a modelling artefact but a genuine property of foliar feature representations. 2) Removing redundant filters improves diagnostic accuracy rather than degrading it, with gains of up to $+0.62$ pp for ResNet-18 on PlantVillage and statistical significance confirmed in every tested configuration. 3) Redundancy is strongly depth-stratified: the first residual block sheds 3% of its filters while the fifth sheds 38%, a pattern that provides actionable guidance for practitioners deciding where to focus compression effort.

Planned extensions include sub-filter granularity gate search, composition with INT8 quantisation for a holistic lightweight pipeline, and a study of gate mask transferability across acquisition domains.

ACKNOWLEDGMENT

The authors acknowledge computing infrastructure provided by the Faculty of Science and Technology of Fez and the ENSAM, Mohammed V University in Rabat.

REFERENCES

[1] Food and Agriculture Organization of the United Nations, *The State of Food and Agriculture 2021: Making Agrifood Systems More Resilient to Shocks and Stresses*. Rome, Italy: FAO, 2021.

[2] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," *Front. Plant Sci.*, vol. 7, p. 1419, 2016.

[3] E. H. Hssayni, N.-E. Joudar, and M. Ettaouil, "KRR-CNN: kernels redundancy reduction in convolutional neural networks," *Neural Comput. Appl.*, 2021.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, Las Vegas, NV, USA, 2016, pp. 770–778.

[6] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE CVPR*, Las Vegas, NV, USA, 2016, pp. 2818–2826.

[7] Y. LeCun, J. S. Denker, and S. A. Solla, "Optimal brain damage," in *Advances in Neural Information Processing Systems*, vol. 2, 1990, pp. 598–605.

[8] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Speeding-up convolutional neural networks using fine-tuned CP-decomposition," *arXiv preprint arXiv:1412.6553*, 2014.

[9] R. Ma, J. Miao, L. Niu, and P. Zhang, "Transformed ℓ_1 regularization for learning sparse deep neural networks," *Neural Netw.*, vol. 119, pp. 286–298, 2019.

[10] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[11] K. P. Ferentinos, "Deep learning models for plant disease detection and diagnosis," *Comput. Electron. Agric.*, vol. 145, pp. 311–318, 2018.

[12] E. C. Too, L. Yujian, S. Njuki, and L. Yingchun, "A comparative study of fine-tuning deep learning models for plant disease identification," *Comput. Electron. Agric.*, vol. 161, pp. 272–279, 2019.

[13] J. Chen, J. Chen, D. Zhang, Y. Sun, and Y. A. Nanehkaran, "Using deep transfer learning for image-based plant disease identification," *Comput. Electron. Agric.*, vol. 173, p. 105393, 2020.

[14] M. Brahimi, K. Boukhalfa, and A. Moussaoui, "Deep learning for tomato diseases: classification and symptoms visualization," *Appl. Artif. Intell.*, vol. 31, pp. 299–315, 2017.

[15] S. Han, H. Mao, and W. J. Dally, "Deep compression: compressing deep neural networks with pruning, trained quantization and Huffman coding," in *Proc. ICLR*, San Juan, Puerto Rico, 2016.

[16] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, "Pruning filters for efficient ConvNets," in *Proc. ICLR*, Toulon, France, 2017.

[17] C. Tai, T. Xiao, Y. Zhang, and X. Wang, "Convolutional neural networks with low-rank regularization," *arXiv preprint arXiv:1511.06067*, 2015.

[18] E. L. Denton, W. Zaremba, J. Bruna, Y. LeCun, and R. Fergus, "Exploiting linear structure within convolutional networks for efficient evaluation," in *Advances in NeurIPS*, vol. 27, 2014, pp. 1269–1277.

[19] F. E. F. Junior and G. G. Yen, "Particle swarm optimization of deep neural networks architectures for image classification," *Swarm Evol. Comput.*, vol. 49, pp. 62–74, 2019.

[20] C. Bai, L. Huang, X. Pan, J. Zheng, and S. Chen, "Optimization of deep convolutional neural network for large scale image retrieval," *Neurocomputing*, vol. 303, pp. 60–67, 2018.

[21] H. Louati, S. Bechikh, A. Louati, R. C.-C. Hung, and L. B. Said, "Deep convolutional neural network architecture design as a bi-level optimization problem," *Neurocomputing*, vol. 439, pp. 44–62, 2021.

[22] D. P. Hughes and M. Salathé, "An open access repository of images on plant health to enable the development of mobile disease diagnostics," *arXiv preprint arXiv:1511.08060*, 2015.

[23] R. Thapa, K. Zhang, N. Snavely, S. Belongie, and A. Khan, "The Plant Pathology 2020 challenge dataset to classify foliar disease of apples," *Appl. Plant Sci.*, vol. 8, p. e11390, 2020.

[24] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *Proc. ICLR*, San Diego, CA, USA, 2015.

[25] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: inverted residuals and linear bottlenecks," in *Proc. IEEE CVPR*, Salt Lake City, UT, USA, 2018, pp. 4510–4520.

[26] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, 2002.

[27] M. Belkin, D. Hsu, S. Ma, and S. Mandal, "Reconciling modern machine-learning practice and the classical bias–variance trade-off," *Proc. Natl. Acad. Sci. USA*, vol. 116, pp. 15849–15854, 2019.

A Binary Cascade Framework for Gut Microbiome Classification in Inflammatory Bowel Disease: Few-Shot LLMs, Co-Abundance Graphs, and Clinical Narrative Encoding

Nouhaila En Najih, Ahmed Moussa
Systems and Data Engineering Team-National School of Applied Sciences, Abdelmalek Essaadi University, Tangier, Morocco

***Abstract*—Subtyping inflammatory bowel disease (IBD) from gut microbiome sequencing data is a clinically demanding problem. Cross-cohort variability is large, and models trained on one dataset often fail on another. In this work, we propose a binary cascade pipeline that splits the three-class problem — healthy control, Crohn's disease (CD), and ulcerative colitis (UC) — into two independent binary classifiers optimized at each stage separately. A key component of the pipeline is a clinical narrative encoding strategy (P4) that converts centered log-ratio (CLR) microbial abundances into structured biomedical text, allowing a large language model (LLM) to apply its pre-existing biomedical knowledge without any fine-tuning. We evaluate Gemini-2.5-Flash under two encoding schemes, six few-shot sizes ($k \in \{2, 4, 6, 8, 10, 14\}$), and two example-selection strategies across three random seeds. The pipeline is benchmarked against a co-abundance graph neural network (GNN) and five supervised classifiers on an internal holdout set and two external cohorts: Dataset_30 ($n = 30$) and IBD423 ($n = 423$). P4 encoding achieves Macro-F1 of 0.539 ± 0.024 on Control vs. IBD and 0.585 ± 0.011 on CD vs. UC, without any gradient updates. On external CD vs. UC validation, Gemini reaches F1 = 0.873, matching the best supervised baseline while outperforming all tree-based and graph-based methods. Centroid-based few-shot selection outperforms random selection by 5.2 percentage points. The results indicate that grounding microbial measurements in clinical language is a practical route to interpretable, training-free IBD classification.**

***Keywords*—Inflammatory bowel disease; gut microbiome; large language models; few-shot learning; graph neural networks; 16S rRNA; clinical narrative encoding; binary cascade; centroid selection**

I. Introduction

More than 6.8 million people worldwide are affected by inflammatory bowel disease, and incidence is still rising in newly industrialized regions [1]. It has two subtypes: Crohn's disease (CD) and ulcerative colitis (UC). The two share several clinical features. However, they diverge considerably in where inflammation occurs, how tissue damage appears under the microscope, and how patients respond to treatment. Distinguishing the two subtypes, and ruling out healthy controls, is a precondition for any appropriate therapeutic plan. The gut microbiome contains a relevant disease signal. Studies conducted across independent cohorts have documented consistent reductions in butyrate-producing genera, particularly *Faecalibacterium prausnitzii* and *Roseburia*, alongside increases in pro-inflammatory taxa such as *Escherichia* and *Fusobacterium* [2]. Machine learning applied to 16S rRNA amplicon profiles has shown reasonable discriminative ability [3], [4]. However, two problems keep reappearing. First, classifiers trained on one cohort often break down on another: sequencing batch effects, differences in extraction protocols, and site-specific ecological drift all introduce variability unrelated to disease biology [5]. Second, high-performing ensemble methods are difficult to interpret, and without explainable reasoning, clinical adoption remains limited [6].

Large language models trained on biomedical text encode genus-level associations with disease states [7]. Whether that stored knowledge can substitute for labelled training data in a microbiome classification task is still an open question. A prior benchmark on the same dataset evaluated six frontier LLMs under fixed few-shot conditions with log-normalised features; Mistral 7B achieved the highest Macro-F1 at 0.5417, and open-source models remained competitive with proprietary APIs [8]. That study held the encoding format and example-selection strategy fixed. In this work, we vary both. Table I summarizes the differences between the two studies. We also include co-abundance GNNs, which operate on ecological interaction graphs rather than independent feature vectors and therefore capture structure that tabular models ignore [9], [10].

In this study, we evaluate LLMs, GNNs, and classical supervised learners on identical datasets under the same experimental protocol. Our contributions are as follows:

- A binary cascade decomposition that splits the three-class IBD problem into two sequential binary tasks, reducing per-stage class imbalance and allowing stage-specific model selection.
- A clinical narrative encoding (P4) that pairs each CLR-normalized genus abundance with a sentence drawn from the IBD literature, providing the LLM with both the measurement and its clinical meaning in the same prompt.
- A systematic few-shot ablation covering shot count ($k = 2$ to 14) and selection strategy (random vs. centroid-based), replicated across three independent seeds.
- A cross-paradigm comparison on two external cohorts that identifies where each model family generalizes and where it does not.

TABLE I. COMPARISON WITH PRIOR WORK [8]

Dimension	Prior work	This work
Focus	LLM benchmark	Encoding study
Prompt format	Fixed (P3 only)	P3 vs. P4
Few-shot count	Fixed k	$k \in \{2..14\}$ ablation
Example selection	Random	Random vs. centroid
Models compared	LLM only	LLM + GNN + ML
External valid.	1 cohort	2 cohorts
Multi-run stats	Single run	Mean $\pm$ std, 3 seeds

II. RELATED WORK

A. Supervised Learning on Microbiome Data

A meta-analysis of gut microbiome classification studies found that random forests achieve moderate accuracy within datasets but transfer poorly across cohorts [11]. The problem is that ensemble models fit sequencing platform signatures and extraction batch effects just as readily as genuine disease signal. Once applied to samples from a different centre, those cohort-specific artefacts vanish and predictive performance collapses. Logistic regression with ℓ_2 regularization has shown more stable cross-site behaviour; by shrinking coefficients toward zero, it avoids placing weight on narrow, site-dependent features [12]. More recent proposals include metaheuristic hyperparameter search for ensemble tuning [13] and recursive elimination of unstable microbiome features [4], [14]. Across all of these, the same trade-off appears: accuracy on the training cohort comes at the expense of accuracy on a new one. Our work asks whether this trade-off can be sidestepped by relying on prior biomedical knowledge rather than labelled cohort data.

B. Data Representation for Microbiome Classification

The choice of compositional transformation has a substantial effect on downstream results. Untransformed read counts confound library size with biological abundance. Rarefaction removes this confound but discards sequencing data. The CLR transformation handles both issues at once: it removes the compositional constraint while keeping all samples in the analysis, and it has been shown to reduce spurious genus-level correlations [15]. Feature selection is also important. The microbiome spans hundreds to thousands of genera, but only a fraction carry IBD-relevant signal. Random Forest importance scores, recursive elimination, and clinical prior knowledge are each valid selection criteria [16]. For LLM-based classification, which genera appear in the prompt is especially consequential since each genus occupies fixed context space. In this work, we combine data-driven importance ranking with mandatory inclusion of genera with documented IBD associations.

C. Deep Learning and Graph-Based Microbiome Analysis

Convolutional and transformer architectures applied to microbiome data have shown competitive in-distribution performance [3], [17]. Graph neural networks are well suited to this domain. Microbial communities are ecological networks, and co-occurrence and competitive relationships among taxa carry information that per-taxon abundance vectors do not capture [18]. Graph convolutional classifiers built on co-abundance networks have outperformed tabular baselines on several disease benchmarks [19], [20]. Whether this co-abundance structure itself transfers across sequencing sites, or whether it is as cohort-specific as the raw abundances, has not been systematically tested. In this study, we evaluate a co-abundance GNN on two independent external cohorts to provide direct evidence on this question.

D. LLMs in Biomedical Classification

In-context learning [21] has been evaluated across many biomedical tasks: medical question answering [7], [22], clinical note classification [23], and domain-adapted variants such as BioMistral [24]. The most directly relevant prior study benchmarked six LLMs on this same IBD dataset, finding open-source models competitive with proprietary APIs [8]. A separate study applied vision transformers to UC severity grading from endoscopic images [25]. None of these studies examined how the numerical encoding of omics inputs affects LLM performance. That is the gap this work addresses. Retrieval-augmented generation [27] takes a different approach: rather than fixing the in-context examples in advance, it draws them from a larger pool separately for each query. Centroid selection can be read as a simpler predecessor of this idea, since it selects representative examples once from the class means and reuses them across all queries. Per-query retrieval over a microbiome reference pool is examined in Section VI.

III. MATERIALS AND METHODS

A. Datasets

This study used three 16S rRNA sequencing datasets, with the breakdown of control and patient samples (CD and UC) presented in Table II.

TABLE II. DATASET SUMMARY

Dataset	Total	Control	CD	UC
Dataset_1 (primary)	1,286	336	731	219
Dataset_30 (Ext I)	30	10	10	10
IBD423 (Ext II)	423	44	204	175

Dataset_1 is the primary 16S rRNA amplicon cohort. We reserved a stratified holdout of 500 samples for Control vs. IBD evaluation and 300 samples for CD vs. UC, with the remaining samples forming the training pool. **Dataset_30** was collected at a separate sequencing centre and serves as a cross-site generalization benchmark. **IBD423** is a public dataset used only for external CD vs. UC validation.

B. Preprocessing

We rarefied all samples to 10,000 reads by multinomial subsampling and discarded samples below that depth. The CLR transformation was then applied:

$$\mathrm{CLR}(x_i) = \log\left(\frac{x_i + 0.5}{g(\mathbf{x} + 0.5)}\right), \tag{1}$$

where, $g(\cdot)$ denotes the geometric mean. A pseudocount of 0.5 was added to handle structural zeros.

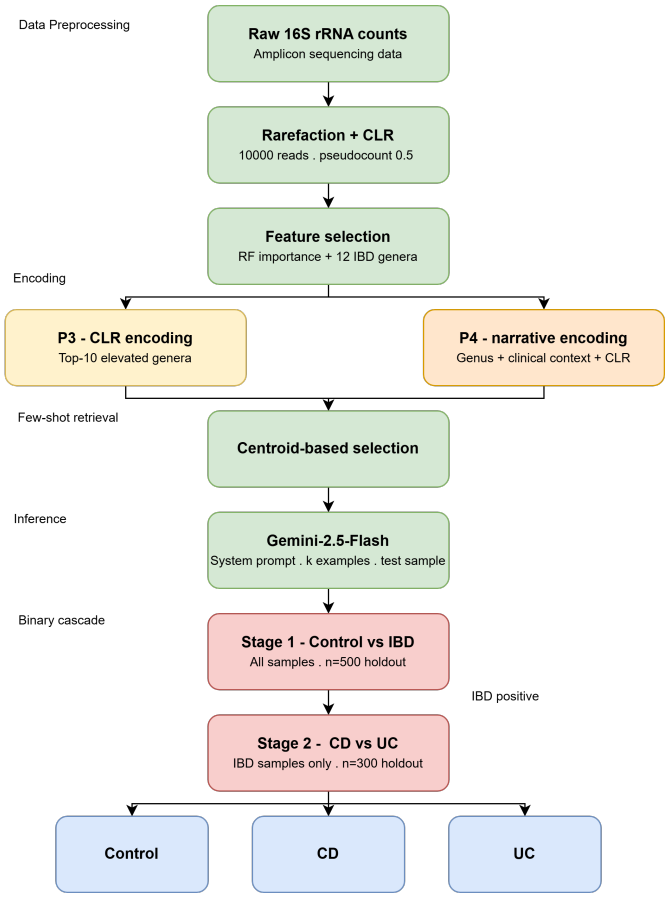


Fig. 1. Binary cascade pipeline. Stage 1 distinguishes healthy controls from all IBD patients. Stage 2, applied only to samples classified as IBD-positive, further distinguishes CD from UC. Each stage uses independent feature selection, encoding, and model training.

C. Feature Selection

For each binary problem, we fitted a Random Forest ($n_{\text{trees}} = 500$, max_depth $= 5$, balanced class weights) on the training pool and retained the top 20 genera by feature importance. We then appended twelve genera with documented IBD associations — including *Faecalibacterium*, *Roseburia*, *Akkermansia*, *Escherichia*, *Fusobacterium*, and *Ruminococcus* — if not already selected. The final feature sets contained 25 genera for Problem 1 and 26 for Problem 2.

D. Binary Cascade Framework

The three-class problem cannot be solved in one shot. We break it into two binary stages (Fig. 1). Stage 1 resolves a single decision: healthy or IBD. Samples it marks IBD-positive proceed to Stage 2, where the question becomes CD versus UC. The decomposition offers two advantages. Each stage faces a milder class imbalance than the full three-way split would impose, and because the stages operate independently, the encoding and the prompt can be tuned separately for the demands of each.

E. Microbiome Encoding Strategies

1) P3 - CLR abundance encoding: We select the genera with CLR values above the sample median (at most ten), apply ℓ_1-normalization, and format the result as a ranked list. No clinical context is added; the list communicates abundance magnitudes only.

2) P4 - Clinical narrative encoding: For P4, every selected genus is paired with a short literature-derived sentence stating its known IBD role, followed by whether it is elevated or depleted relative to the sample median and its exact CLR value. A representative entry reads:

```
Faecalibacterium prausnitzii is an
anti-inflammatory butyrate producer
depleted in IBD.
[Observed: depleted, CLR=-1.234]
```

The rationale is straightforward: the model receives the observation and its clinical meaning together, so it does not need to reconstruct the genus-disease link from pretraining memory under uncertainty.

F. Prompt Examples

The practical difference between the two encodings is easiest to see on a single sample. Table III places the P3 and P4 representations of the same CD patient side by side.

TABLE III. ENCODING COMPARISON FOR THE SAME SAMPLE (CD PATIENT). THE *Diagnosis:* FIELD IS LEFT BLANK TO MARK THE POSITION WHERE THE MODEL GENERATES ITS LABEL

P3 — CLR Encoding	**P4 — Narrative Encoding**
CLR normalized abundances: Faecalibacterium: 0.0312 Roseburia: 0.2841 Fusobacterium: 0.4102 ... Ruminococcus: 0.1745 ...	Clinical microbiome profile: *F. prausnitzii* is a butyrate producer depleted in IBD. [Observed: depleted, CLR=-1.23] *Fusobacterium* is enriched in CD. [Observed: elevated, CLR=+2.41] ...
Diagnosis:	*Diagnosis:*

P3 presents the same genera as bare CLR values, whereas P4 pairs each measurement with a sentence of clinical context before stating its observed direction and magnitude. The two columns encode identical numerical information; they differ only in how much biological meaning is made explicit to the model.

G. Few-Shot LLM Classification

1) Model and inference: We queried Gemini-2.5-Flash through the Google Generative Language API with temperature = 1.0 and maxOutputTokens = 100. A system prompt defined the clinical role and valid output labels. Each request included k in-context examples followed by the test sample. The model was instructed to output a single diagnostic label on the last line.

2) Example selection strategies: Both strategies selected $k/2$ examples per class.

a) Random selection: draws examples uniformly from the training pool under a fixed seed.

b) Centroid-based selection: computes the class centroid in CLR space and selects the $k/2$ training samples with the highest cosine similarity to that centroid, giving priority to the most representative profiles per class.

3) Shot count ablation: We tested $k \in \{2, 4, 6, 8, 10, 14\}$. All Experiment B runs used P3 encoding with centroid selection to isolate the effect of shot count from other variables.

H. Prompt Structure

To make the encoding concrete, it helps to see a full prompt exactly as the model receives it. Fig. 2 reproduces the complete input submitted to Gemini-2.5-Flash for the CD vs. UC task under P4 encoding, using $k = 4$ centroid-selected examples.

I. Co-Abundance Graph Neural Network

We computed pairwise Spearman correlations across all training-pool genera and connected genus pairs with absolute correlation above 0.3 by undirected edges. Node features were the per-sample CLR abundances. We trained a three-layer Graph Convolutional Network [20] with hidden dimension 64, dropout 0.3, and global mean pooling for 150 epochs using the Adam optimizer (lr = 10^{-3}, weight decay = 10^{-4}, batch size 32) with balanced cross-entropy loss. The checkpoint with the highest validation Macro-F1 was kept for evaluation.

J. Classical Supervised Baselines

We evaluated five classifiers: Extra Trees ($n = 500$, max_depth = 10), Random Forest ($n = 500$, max_depth = 10), Logistic Regression ($C = 0.1$, ℓ_2 penalty), XGBoost ($n = 200$, max_depth = 6, lr = 0.05), and LightGBM ($n = 200$, lr = 0.05). All models used balanced class weights and were trained on the full CLR feature matrix.

K. Evaluation Protocol

LLM experiments were run under three seeds (42, 123, 456). This controlled holdout subsampling (300 samples per run) and example pool composition. All LLM results are reported as mean $\pm$ std of Macro-F1 across the three seeds. GNN and supervised baselines were evaluated on the full holdout sets (500 samples for Problem 1, 300 for Problem 2) and on both external cohorts. LLM and supervised evaluations cover the holdout differently for reasons of cost. Each LLM prediction requires a paid API call, whereas supervised inference is effectively free once the model has been fitted. LLM subsampling was therefore kept class-stratified and repeated under three seeds, with results averaged across runs. Supervised baselines were evaluated on the full holdout sets, the setting in which they perform most favourably.

L. Computational Cost

Inference was performed with Gemini-2.5-Flash through the Google Generative Language API. A single request completed in one to two seconds and cost a small fraction of a US cent at the pricing tier used here. Across the full experimental grid of three seeds, six shot counts, two encodings, and two classification problems, total API charges came to approximately fifteen US dollars. GNN training completed in under 15 minutes on a T4 GPU, and each supervised baseline was fitted in under two minutes on CPU, both on freely available hardware. Extrapolating to deployment, a service processing tens to hundreds of samples per week would incur API costs in the low single-digit dollars per month.

[SYSTEM PROMPT]

```
You are a clinical microbiome specialist.
This sample is from an IBD patient.
Classify as:
- CD: Crohn's disease (depleted
  Faecalibacterium, enriched
  Fusobacterium/Ruminococcus)
- UC: ulcerative colitis (enriched
  Escherichia, mucosal inflammation
  pattern)
Output ONLY one word on the last line:
CD or UC
```

[FEW-SHOT EXAMPLES (k=4, centroid-based)]

```
Examples:
---
Profile:
Faecalibacterium prausnitzii is an
anti-inflammatory butyrate producer
depleted in IBD.
[Observed: depleted, CLR=-1.823]
Fusobacterium is enriched in CD.
[Observed: elevated, CLR=+2.341]
Diagnosis: CD
---
Profile:
Escherichia/Shigella is enriched in UC.
[Observed: elevated, CLR=+3.102]
Faecalibacterium prausnitzii is an
anti-inflammatory butyrate producer
depleted in IBD.
[Observed: depleted, CLR=-0.912]
Diagnosis: UC
---
```

[TEST SAMPLE]

```
Patient profile:
Faecalibacterium prausnitzii is an
anti-inflammatory butyrate producer
depleted in IBD.
[Observed: depleted, CLR=-2.104]
Ruminococcus is enriched in CD.
[Observed: elevated, CLR=+1.877]
Fusobacterium is enriched in CD.
[Observed: elevated, CLR=+1.543]
Diagnosis:
```

Fig. 2. Complete prompt for CD vs. UC classification under P4 narrative encoding with $k = 4$ centroid examples. The system prompt specifies the clinical role and output format. One example per class follows, each pairing a biomedical annotation with the observed CLR direction. The test sample appears last with the label blank.

IV. Results

A. Experiment A: Encoding Strategy

At a fixed budget of $k = 10$ centroid examples, P4 narrative encoding produces higher Macro-F1 than P3 CLR on both classification problems. The improvement is larger on CD vs. UC (+4.7 pp) than on Control vs. IBD (+0.7 pp). For the

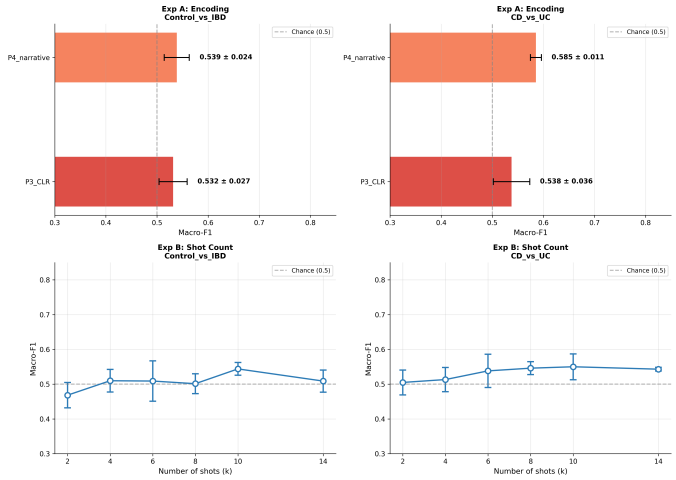


Fig. 3. LLM ablation results for Gemini-2.5-Flash. Upper panels (A, B): Macro-F1 comparison between P3 CLR and P4 narrative encodings with mean ± std over three seeds. Lower panels (C, D): Macro-F1 as a function of few-shot example count k using P3 CLR encoding and centroid-based selection.

subtype task, the clinical annotations attached to *Fusobacterium* (CD-associated) and *Escherichia* (UC-associated) give the model context it would otherwise need to recover from pretraining, and the results indicate that recovering it from memory is less reliable than stating it explicitly. Variance falls as well — from ± 0.036 under P3 to ± 0.011 under P4 on CD vs. UC — so P4 also produces more consistent results across different random example draws.

The two encodings are compared numerically in Table IV, averaged over 300 samples and three seeds, with the corresponding ablation shown in Fig. 3.

TABLE IV. EXPERIMENT A - ENCODING STRATEGY (MEAN ± STD, 3 SEEDS)

Problem	Encoding	Macro-F1
Control vs. IBD	P3 CLR	0.532 ± 0.028
	P4 Narrative	$\mathbf{0.539 \pm 0.024}$
CD vs. UC	P3 CLR	0.538 ± 0.036
	P4 Narrative	$\mathbf{0.585 \pm 0.011}$

B. Experiment B: Shot Count

We tested six shot counts ($k \in \{2, 4, 6, 8, 10, 14\}$) using P3 encoding with centroid selection. Results are summarized in Table V.

TABLE V. EXPERIMENT B — SHOT COUNT (MEAN ± STD, 3 SEEDS, P3 CLR)

k	Control vs. IBD	CD vs. UC
2	0.487 ± 0.018	0.497 ± 0.062
4	0.527 ± 0.039	0.536 ± 0.012
6	0.538 ± 0.048	0.579 ± 0.034
8	0.546 ± 0.019	0.546 ± 0.022
10	0.550 ± 0.037	0.555 ± 0.034
14	0.543 ± 0.006	0.559 ± 0.014

Performance improves from $k = 2$ to approximately $k = 6$–8, then levels off (Table V, Fig. 3). Beyond $k = 8$, additional examples do not appear to add useful distributional information. Including more examples past this point may actually dilute the most representative profiles with less typical ones. This pattern is consistent across both classification problems.

C. Experiment C: Example Selection Strategy

We compared centroid-based and random selection at $k = 10$, holding encoding (P3) and shot count fixed. Table VI reports the results across three seeds.

TABLE VI. EXPERIMENT C — SELECTION STRATEGY (MEAN ± STD, 3 SEEDS, $k = 10$)

Problem	Strategy	Macro-F1
Control vs. IBD	Random	0.487 ± 0.016
	Centroid	$\mathbf{0.539 \pm 0.035}$
CD vs. UC	Random	0.493 ± 0.012
	Centroid	$\mathbf{0.541 \pm 0.025}$

Centroid selection outperforms random selection by 5.2 pp on Control vs. IBD and 4.8 pp on CD vs. UC (Table VI, Fig. 5). Microbiome profiles are high-dimensional and sparse. A randomly drawn sample may sit far from the class centre of mass and give the model an atypical reference point. By selecting samples with high cosine similarity to the class centroid, centroid selection ensures the model receives archetypal rather than idiosyncratic examples.

D. Full Cross-Paradigm Comparison

We now compare all seven models across every evaluation set. Macro-F1 scores are reported in Table VII and visualized in Fig. 4.

On the internal holdout, tree-based ensembles perform best: Extra Trees and XGBoost both reach F1 = 0.73 on Control vs. IBD. Their external performance is very different. On Ext30, these same models fall to F1 = 0.33–0.44, a drop of nearly 30 percentage points. Ensemble methods trained on one sequencing site routinely pick up protocol-specific artefacts together with true disease signal; when those artefacts are absent in a new cohort, accuracy drops sharply [26]. Random Forest shows the same behaviour.

Logistic Regression departs from this pattern. Holdout accuracy is moderate (0.694 on Problem 1, 0.659 on Problem 2), yet both external cohorts remain competitive: F1 = 0.688 and 0.873 on Ext30, and F1 = 0.675 on IBD423.

On Ext30 Control vs. IBD, the GNN obtains its best result (F1 = 0.762), which suggests co-abundance graph structure captures disease-relevant ecological patterns that transfer across sites. For the CD vs. UC task, however, the same GNN drops to F1 = 0.333 on Ext30 — likely because of the smaller Stage 2 training set and the greater difficulty of subtype-level discrimination.

Without a single gradient update on the target data, Gemini-2.5-Flash with P4 encoding reaches F1 = 0.873 on Ext30 for CD vs. UC, tying Logistic Regression for first place and leaving all other models behind. On IBD423 it reaches F1 = 0.714, second

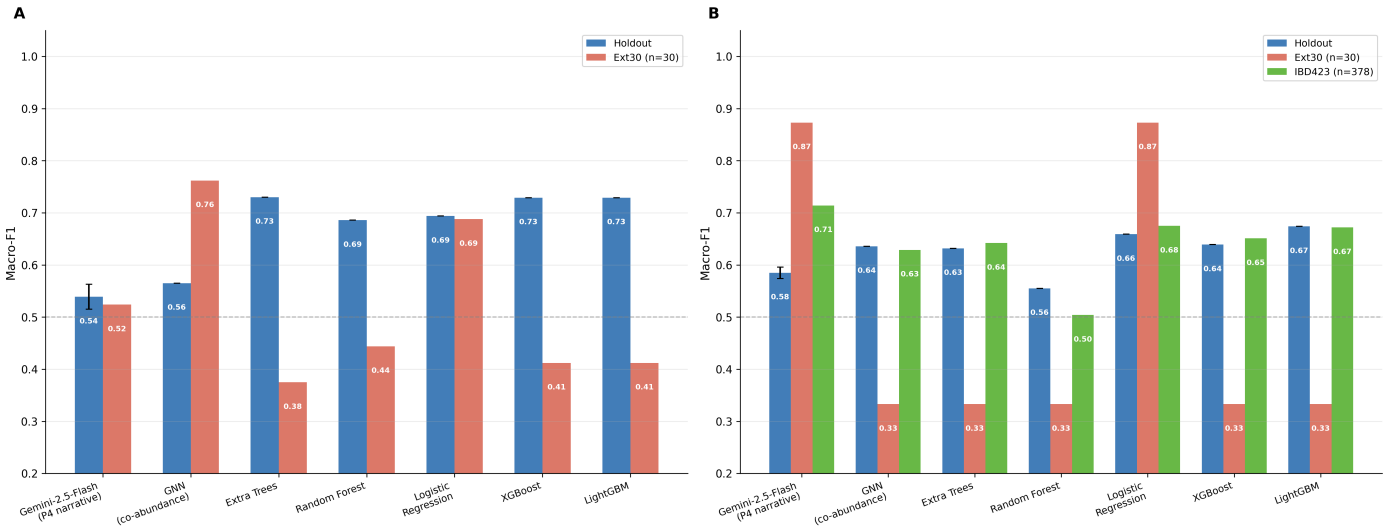


Fig. 4. Macro-F1 scores for all evaluated models on internal holdout and external validation cohorts. (A) Control vs. IBD classification. (B) CD vs. UC classification. Error bars on Gemini-2.5-Flash indicate standard deviation over three seeds. Dashed line marks chance level (F1 = 0.5).

TABLE VII. FULL MODEL COMPARISON - MACRO-F1 ACROSS INTERNAL AND EXTERNAL EVALUATION SETS

Model	Category	Control vs. IBD		CD vs. UC		
		Holdout	Ext30	Holdout	Ext30	IBD423
Gemini-2.5-Flash (P4)	LLM	0.539 ± 0.024	0.524	0.585 ± 0.011	**0.873**	0.714
GNN (co-abundance)	GNN	0.565	**0.762**	0.636	0.333	0.629
Extra Trees	ML	**0.730**	0.375	0.632	0.333	0.642
Random Forest	ML	0.686	0.444	0.555	0.333	0.504
Logistic Regression	ML	0.694	0.688	0.659	**0.873**	**0.675**
XGBoost	ML	0.729	0.412	0.639	0.333	0.651
LightGBM	ML	0.729	0.412	0.674	0.333	0.672

Bold: best result per column. LLM results are mean $\pm$ std over 3 seeds.

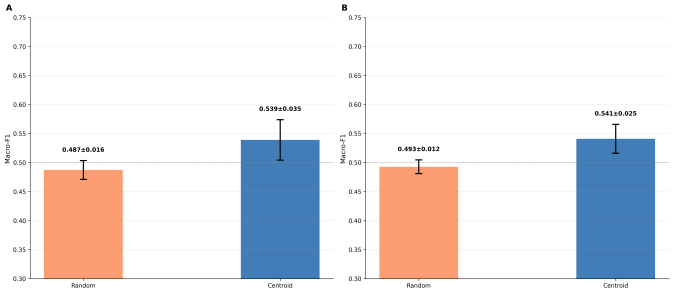


Fig. 5. Effect of few-shot example selection strategy on Gemini-2.5-Flash performance. (A) Control vs. IBD. (B) CD vs. UC. Error bars indicate standard deviation over three seeds. Centroid-based selection consistently outperforms random selection on both tasks.

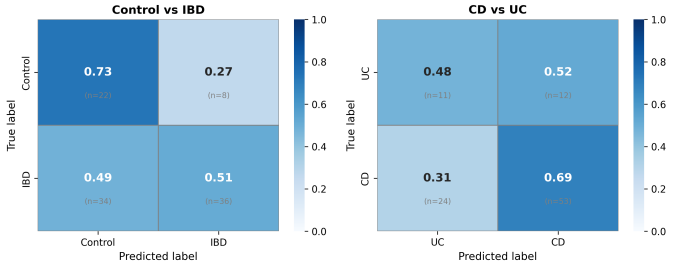


Fig. 6. Normalized confusion matrices for Gemini-2.5-Flash with P4 narrative encoding ($k = 10$, seed=42, $n = 100$ samples per problem). Values represent the proportion of true-class samples assigned to each predicted label; raw counts are shown in parentheses. UC samples are substantially harder to classify than CD samples (52.2% vs. 31.2% error rate), consistent with the greater microbiome heterogeneity of ulcerative colitis.

best on that cohort. P4 encoding and centroid selection are the only methodological differences between the two studies; consequently, the gain relative to [8] can be attributed to those two changes.

E. Confusion Matrix Analysis

Confusion matrices in Fig. 6 break down where each model succeeds and fails on each problem, using P4 encoding ($k = 10$, seed=42, $n = 100$ samples per problem).

Among healthy controls, 73% are correctly identified, while 48.6% of IBD samples are misclassified as healthy. The model seems to wait for a clear dysbiotic signal before assigning an IBD label, which pushes borderline samples near the boundary toward the healthy class. The split is wider for CD versus UC. Here the CD error rate is 31.2%, well under the 52.2% seen for UC. CD carries its own markers in the P4 prompt, chiefly the *Fusobacterium* and *Ruminococcus* annotations, while the UC annotations separate the classes far less cleanly. The recall

figures show the same gap.

V. Discussion

A. Clinical Narrative Encoding

The gap between P4 and P3 runs deeper than formatting. Consider a P3 entry like `Faecalibacterium: -1.823`. A genus and a number — nothing more reaches the model. For that pairing to carry meaning, the model must independently recall that this genus is anti-inflammatory and that its depletion marks active IBD — an implicit retrieval step with no guarantee of success. P4 removes that burden. Placed directly beside the measurement in the prompt, the clinical association ensures the observed direction and its biological meaning arrive together in a single token sequence. Rather than inferring what the number means, it applies context that is already present.

The two problems respond differently to the encoding change. On Control vs. IBD the gain is modest ($+0.7$ pp), while CD vs. UC shows a larger improvement ($+4.7$ pp). The reason is not surprising. Separating IBD from healthy controls rests on a broad, well-documented pattern: butyrate-producing taxa are depleted across the vast majority of IBD samples regardless of subtype, and that association is strong enough that a pretrained model recovers it without explicit annotation. Distinguishing CD from UC requires finer judgment — the relative elevations of *Fusobacterium* versus *Escherichia* are the key markers, and these are exactly the annotations that P4 supplies. In general, these findings suggest that LLM classification of structured omics data should be treated as a knowledge integration task. The prompt must bridge quantitative measurements and clinical significance; relying on the model to perform that bridge silently from pretraining is less dependable, especially for subtle distinctions. The mechanistic language used here describes patterns consistent with the observed outputs. A note on terminology is warranted here. Statements that the model "uses" or "applies" biomedical knowledge indicate only that the observed outputs are consistent with that interpretation. They are equally consistent with the model responding to surface features of the P4 text rather than to its biomedical content. Distinguishing between the two accounts would require ablation experiments in which portions of the prompt are removed or perturbed. Such experiments were not conducted here, and the question is revisited in Section V-F.

B. Error Analysis

In the Control vs. IBD task, misclassifications fall heavily on IBD samples labelled as healthy (34/70, 48.6%), whereas healthy controls are called correctly at a higher rate (8/30, 26.7%). Looking at the misclassified IBD samples, their *Ruminococcus* CLR is elevated relative to correctly classified cases (2.60 vs. 1.46), and their *Fusobacterium* CLR is lower (1.84 vs. 2.79). These profiles sit outside the canonical dysbiotic pattern that P4 annotations describe. This pattern is consistent with early-stage or inactive IBD, although it may equally reflect model sensitivity to CLR magnitudes near the annotation thresholds rather than any underlying biological distinction. Prediction outputs alone provide no basis for separating the two explanations.

For CD versus UC the error-rate gap is wide. UC samples were misclassified at 52.2% against 31.2% for CD. *Fusobacterium* CLR separated the two classes most cleanly, averaging 3.16 in correctly identified CD cases but only 1.27 in those that were missed. Once that genus stops being clearly elevated, the CD cue carried by the P4 prompt loses its grip and the sample drifts toward UC. Two other genera made things worse. Where *Blautia* (5.27 vs. 3.79) and *Faecalibacterium* (6.50 vs. 5.17) were both high, the butyrate-producing community still looked relatively intact, and that intact profile is where CD and UC proved hardest to separate. Widening the P4 annotation set with a few additional genera would be the next step to test.

C. Centroid Selection

A 5.2 pp gap between centroid and random selection is not trivial for a zero-training method. It reflects something specific about this data type. Microbiome profiles are sparse and high-dimensional, and a randomly chosen training sample can easily turn out to be an outlier, a patient whose profile shares the label but not the typical features. Hand that to the model as a reference and it learns the wrong prototype. Centroid selection sidesteps the problem by computing the class mean in CLR space and picking the samples closest to it. The examples that go into the prompt then look like the typical CD or UC profile. The computational cost is negligible. One matrix multiplication plus a sort over the training pool takes milliseconds, far less than a single API call. Under the conditions examined here, centroid selection emerges as the preferable strategy.

D. Model Selection Across Paradigms

No single model dominates every condition in Table VII; the results are better read as a deployment map than as a ranking. Tree ensembles suit single-site settings, where their high internal accuracy is an asset, but they degrade sharply once applied to external data. Logistic Regression is the conservative option for multi-site use, never leading in any condition yet never collapsing either. The GNN earns its place on Control versus IBD across sites (Ext30 F1 = 0.762) but not on CD versus UC, where its Stage 2 training set is too small and the task correspondingly harder. Gemini with P4 encoding is strongest in one specific configuration: cross-site CD versus UC discrimination with no local labels available.

These profiles translate directly into deployment guidance. A single-site pilot with annotated data is well served by Extra Trees or LightGBM. A multi-site rollout with mixed protocols should pair Logistic Regression at Stage 1 with the LLM at Stage 2, a combination the cascade architecture makes straightforward to implement.

Two considerations motivated the binary cascade: it reduces per-stage class imbalance, and it allows the encoding and prompt to be tuned separately for each decision. Alternative decompositions, including one-vs-rest schemes, hierarchical classifiers, and direct three-way prompting, were not benchmarked here. Comparing the cascade against these alternatives would clarify how much of the observed gain derives from the decomposition itself and how much from stage-specific optimization. The question remains open.

E. Practical Considerations

The LLM approach requires no labelled local training data at inference time. This is useful in clinical settings

where annotating a microbiome cohort is expensive and time-consuming. Since P4 already embeds clinical text in each prompt, the same inference step could be extended to generate natural-language justifications alongside the label, which would support clinician review without additional infrastructure. Data governance is a separate concern: sending patient microbiome data to an external commercial API raises privacy questions that the companion benchmark [8] addresses by evaluating locally deployable open-source models.

F. Limitations

LLM evaluations used subsampled holdout sets of 300 samples per run to control API costs, so the reported means may shift modestly at full-holdout scale. Ext30 contains 30 samples. At that size a single misclassification moves Macro-F1 by roughly three percentage points, and 95% bootstrap intervals would fall near ± 0.10. Results from this cohort are therefore best read as directional evidence rather than as a reliable ranking of models. IBD423, with $n = 423$, provides the more stable basis for assessing external performance.

The two cascade stages were executed independently, with ground-truth labels supplied to Stage 2. End-to-end error was consequently never measured. Multiplying the stage-level Macro-F1 values yields an approximate estimate of 0.32 on the internal holdout under an assumption of independent stage errors. Obtaining the true figure would require running Stage 2 on the IBD-positive predictions emitted by Stage 1, which is left for future work.

The encoding study varied two prompt dimensions: numerical representation (P3 versus P4) and example selection (random versus centroid). Instruction wording, example ordering, output constraints, chain-of-thought suppression, and context-length sensitivity were held fixed throughout. Given that prompt sensitivity in biomedical LLMs is well documented, a broader ablation across these dimensions would be a reasonable extension.

Results are reported as means and standard deviations over three seeds, without significance testing. Three seeds afford limited statistical power. Both headline differences exceed the observed spread by a wide margin: P4 over P3 by $+4.7$ pp on CD versus UC against a pooled standard deviation of 0.024, and centroid over random selection by $+5.2$ pp against a pooled standard deviation of 0.026. Paired testing across a larger number of seeds would nevertheless place these comparisons on firmer statistical ground.

Calibration, output entropy, and agreement across repeated LLM calls on identical prompts were not analysed. A temperature setting of 1.0 renders the model stochastic, so the same prompt can return different labels on consecutive calls. In a clinical context, per-sample variability and calibration quality matter as much as average Macro-F1, and that layer of analysis is absent from this work.

Several passages in the Discussion describe the model as engaging in biological reasoning. Such descriptions are interpretations of the outputs and should not be read as evidence. Establishing a causal claim would require targeted probing, for instance ablating individual annotations from the P4 prompt or perturbing single CLR values, which was not undertaken here.

P4 encodes curated clinical knowledge within each prompt, which means the experiment evaluates the model's capacity to apply written biomedical annotations to observed measurements rather than its ability to classify from numerical profiles without prior annotation. This scope should be kept in mind when interpreting LLM results alongside purely data-driven baselines.

VI. Conclusion

This study proposed a binary cascade framework for IBD microbiome classification and evaluated few-shot LLMs, co-abundance GNNs, and five supervised classifiers under identical conditions on an internal holdout and two independent external cohorts. P4 narrative encoding outperformed CLR-only encoding on both binary tasks, with the largest margin on CD vs. UC where genus-level annotations for *Fusobacterium* and *Escherichia* carried the most discriminative weight. Centroid-based selection outperformed random draws by approximately 5 percentage points on both problems, confirming that example representativeness is a consequential design choice for high-dimensional sparse microbiome data. Classification performance saturated at $k = 6$–8 shots. Tree-based ensembles achieved the highest internal accuracy but degraded substantially on external cohorts. Logistic Regression and Gemini-2.5-Flash with P4 encoding were the only approaches that remained competitive across all evaluation sets.

On cross-site CD vs. UC classification, the LLM reached F1 = 0.873 without task-specific training or local labels. In deployment scenarios where the target site differs from the training site and annotated data are unavailable, clinical narrative encoding offers a practical path to subtype-level discrimination when large labeled training cohorts are not available.

VII. Future Directions

Several directions follow from this work. Replacing the fixed centroid examples with dynamic retrieval over a large microbiome reference database, in the spirit of retrieval-augmented approaches [27], would allow each test sample to be matched against the most similar training profiles rather than the class mean, which should benefit atypical cases that sit far from the centroid. Training a language model on a targeted microbiome-disease corpus could eliminate the manual curation step by replacing hand-written P4 annotations with learned genus-level associations. Incorporating patient-level clinical variables — age, disease duration, medication history — into each cascade stage would bring the system closer to real deployment conditions; the two-stage architecture accommodates such extensions without structural modification. For supervised models, cohort-aware batch correction applied before the CLR step is a concrete intervention against the generalization gap observed here. A longer-term direction is combining the three model families: GNN structure for Control vs. IBD, LLM priors for CD vs. UC, and logistic regression as a stable baseline in multi-site settings, within a single ensemble that adapts its components to the deployment context.

Declaration on Generative AI

Gemini-2.5-Flash (Google DeepMind) was used exclusively as an experimental subject in this study, evaluated under

standard commercial API terms. No generative AI tool was used to write, edit, or improve any part of the manuscript text. All authors take full responsibility for the content of this publication.

Data Availability

The gut microbiome dataset is publicly available from the European Nucleotide Archive under accession numbers PRJEB13679, PRJNA422193 and PRJEB33711. Code, prompts, and reproducibility notebooks are available at https://github.com/NouhailaEnnajih/IBD_LLM_microbiome_clinical_narrative_encoding.

Acknowledgment

The authors thank the contributors of the public IBD microbiome datasets used in this study.

References

[1] S. C. Ng *et al.*, "Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies," *The Lancet*, vol. 390, no. 10114, pp. 2769–2778, 2017.

[2] F. Lloyd-Price *et al.*, "Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases," *Nature*, vol. 569, no. 7758, pp. 655–662, 2019.

[3] I. Manandhar *et al.*, "Gut microbiome-based supervised machine learning for clinical diagnosis of inflammatory bowel diseases," *American Journal of Physiology-Gastrointestinal and Liver Physiology*, vol. 320, no. 3, pp. G328–G337, 2021.

[4] Y. Lee, M. Cappellato, and B. Di Camillo, "Machine learning-based feature selection to search stable microbial biomarkers: application to inflammatory bowel disease," *GigaScience*, vol. 12, p. giad083, 2023.

[5] G. Papoutsoglou *et al.*, "Machine learning approaches in microbiome research: challenges and best practices," *Frontiers in Microbiology*, vol. 14, p. 1261889, 2023.

[6] Y. Peng *et al.*, "Comprehensive data optimization and risk prediction framework: machine learning methods for inflammatory bowel disease prediction based on the human gut microbiome data," *Frontiers in Microbiology*, vol. 15, p. 1483084, 2024.

[7] K. Singhal *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[8] N. En Najih, S. Hamida, and A. Moussa, "A controlled benchmark of open-source and proprietary LLMs for few-shot microbiome IBD classification," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 17, no. 5, 2026. DOI: 10.14569/IJACSA.2026.0170596.

[9] W. Ma *et al.*, "Deep learning in microbiome analysis: a comprehensive review of neural network models," *Frontiers in Microbiology*, vol. 15, p. 1516667, 2024.

[10] X. Song *et al.*, "Predicting microbe-disease associations via graph neural network and contrastive learning," *Frontiers in Microbiology*, vol. 15, p. 1483983, 2024.

[11] C. Duvallet *et al.*, "Meta-analysis of gut microbiome studies identifies disease-specific and shared responses," *Nature Communications*, vol. 8, no. 1, pp. 1–10, 2017.

[12] T. P. Quinn *et al.*, "Interpretable machine learning for genomics," *Human Genetics*, vol. 140, no. 9, pp. 1323–1363, 2021.

[13] A. Abd El-Aziz *et al.*, "Metaheuristic optimization techniques for machine learning in biomedical classification: a review," *Artificial Intelligence Review*, vol. 57, no. 4, 2024.

[14] E. Pasolli *et al.*, "Machine learning meta-analysis of large metagenomic datasets: tools and biological insights," *PLoS Computational Biology*, vol. 12, no. 7, e1004977, 2016.

[15] E. Ibrahimi *et al.*, "Overview of data preprocessing for machine learning applications in human microbiome research," *Frontiers in Microbiology*, vol. 14, p. 1250909, 2023.

[16] B. D. Topçuoğlu, A. Lesniak, M. A. Ruffin IV, J. A. Wiens, and C. Schloss, "A framework for effective application of machine learning to microbiome-based classification problems," *mBio*, vol. 11, no. 3, e00434–20, 2020.

[17] T. Nguyen *et al.*, "Deep learning methods for microbiome-based disease prediction: a systematic review," *Briefings in Bioinformatics*, vol. 25, no. 2, 2024.

[18] M. S. Matchado *et al.*, "Network analysis methods for studying microbial communities: a mini review," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 2687–2698, 2021.

[19] Y. Huang *et al.*, "A graph neural network framework for microbiome-based disease prediction," *Briefings in Bioinformatics*, vol. 25,no. 2,2024.

[20] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2017.

[21] T. B. Brown *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[22] H. Nori *et al.*, "Capabilities of GPT-4 on medical challenge problems," *arXiv preprint arXiv:2303.13375*, 2023.

[23] C. Wu *et al.*, "LLMs-based few-shot disease predictions using EHR: a novel approach combining predictive agent reasoning and critical agent instruction," *AMIA Annual Symposium Proceedings*, pp. 319–328, 2024.

[24] Y. Labrak *et al.*, "BioMistral: a collection of open-source pretrained large language models for medical domains," *arXiv preprint arXiv:2402.10373*, 2024.

[25] D. Gupta, J. Gangrade, Y. P. Singh, and S. Gangrade, "A computer-aided diagnosis system for ulcerative colitis classification using vision transformer," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 16, no. 9, 2025.

[26] M. L. Calle, "Statistical analysis of metagenomics data: challenges and opportunities," *Frontiers in Microbiology*, vol. 14, 2023.

[27] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

Learning When to Wait: A Lightweight Binary Decision Framework for Container Migration in Edge Computing

Yasmeen Abu-Haltem[1], Sawsan AlShattnawi[2]
Computer Science Department, Princess Sumaya University for Technology, Amman, Jordan[1]
Computer Science Department, Yarmouk University, Irbid, Jordan[2]

***Abstract*—Container migration is one of the important techniques used to provide the continuity of services in Multi-access Edge Computing environments. Services are migrated from edge servers to follow the mobility of users among the available edge servers. Nevertheless, current strategies do not take into account the cost that a container migration causes in terms of downtime, bandwidth consumption, energy cost, and the possible occurrence of failures during the process. In this work, a lightweight pre-migration binary decision framework is introduced that works as a gatekeeper before any migration attempt. According to the actual state of the system, the framework decides at each epoch whether to migrate or not, classifying the epochs into Migrate and Wait categories. The training labels are obtained through a counterfactual approach using simulated data generated using the EdgeSimPy simulation tool. The Decision Tree classifier has been chosen as the main classifier due to its minimal computing complexity and interpretability. The effectiveness of the suggested framework is measured on the basis of comparison with three baseline policies, including always migrate, never migrate, and a latency-threshold rule, along with multiple machine learning models and deep learning models, and such metrics as accuracy, precision, recall, F1-score, AUC, and runtime performance metrics. It is shown that the suggested solution outperformed all three baseline policies along with the machine learning and deep learning models, proving that an intelligent decision-making stage is essential in Multi-access Edge Computing.**

Keywords*—Multi-access Edge Computing; container migration; machine learning; Decision Tree; mobility management; service continuity

I. Introduction

Multi-access Edge Computing (MEC), formerly known as Mobile Edge Computing, is becoming increasingly important as use cases such as augmented reality, self-driving cars, healthcare uses, and real-time video analytics require low latency. Computing services deployed at the edge shorten response time. Mobility, on the other hand, creates an important problem for edge computing. As the user moves around the city or country, the distance between them and the edge server providing the service will increase, resulting in higher latency. If the latency becomes too high, the service might become interrupted, which is unacceptable in time-critical scenarios.

The most well-known and used solution to this issue is what is referred to as container migration. In order to maintain low latency and service continuity, the service can follow the user by migrating a running container from one edge server to another. Optimizing container migration has been a great topic for current studies; researchers use advanced methods to find acceptable solutions for the issue. This is done by minimizing downtime, reducing checkpoint size, and choosing better destinations for migrating. Despite many studies on migration triggering methods, migration optimization, and destination determination techniques, most of them take for granted that once a trigger condition is satisfied, migration is always the most effective plan of action and needs to be performed. Not much effort has gone into finding out if migration is even necessary at all. This could potentially result in unnecessary migrations, wasting resources and degrading service performance.

This presumption ignores a crucial fact. Migration is not free, and it is not always the perfect solution. It adds service interruptions during checkpoint transfer, uses network bandwidth, costs energy to limited edge nodes, and carries the risk of failure if network conditions suffer in the middle of migration. Furthermore, a number of factors that cause migration, like slight CPU spikes or short-lasting network instability, may fade before the migration is even completed. Thereby, executing the entire process becomes ineffective and might negatively affect the quality of service (QoS).

In this study, it is argued that sometimes the best decision is not to migrate. A simple binary decision framework is proposed that serves as a gatekeeper before any attempt at migration. The framework provides an output to either proceed with migration, in other words, migrate, or not migrate and re-evaluate later, wait, based on the current system state. Factors affecting the system state include node resource pressure, container characteristics, and network quality. The proposed method is expected to minimize system overhead, prevent unsuccessful migrations, and maintain QoS when migration would be destructive by filtering unneeded and harmful migrations.

The main contributions of the study can be described as follows:

- A lightweight binary pre-migration decision framework is proposed as a modular gatekeeper layer, complementary to existing migration algorithms, that determines whether migration should be executed at all before any migration attempt.
- Creation of a balanced dataset containing 5,000 samples under various mobility settings and container configurations through EdgeSimPy.
- Application of a counterfactual labeling technique to label each decision point with Migrate or Wait labels.

- Evaluation and comparison of various machine learning and deep learning methods, alongside rule-based and unconditional migration baselines.
- Experimental findings show that lightweight decision tree models achieve an effective balance between prediction performance, interpretability, and efficiency, which makes them very well-suited for the edge computing environment.

This work is structured as follows: Section II presents a brief background of the topics mentioned, and relevant studies are discussed. The problem is formulated, and the research gap is discussed in Section III. The proposed decision framework is presented in Section IV. Section V presents the results. Section VI discusses the results. The study concludes in Section VII with a concise summary of the work and future work suggestions.

II. Background and Literature Review

MEC has become a crucial concept for helping and supporting latency-sensitive and data-intensive applications. This is achieved through moving storage and processing resources closer to end users. MEC allows processing at the edge, in contrast to traditional cloud computing, which processes data in centralized data centers. MEC is well-known and suitable for applications like augmented reality, autonomous cars, healthcare fields, and real-time video analytics because it works on greatly lowering communication latencies and enhances QoS [1] [2] [3].

However, because of user mobility and changing network conditions, MEC environments are considered to be very dynamic. The distance between a user and the serving edge node increases as the user travels across different geographic locations, which may result in higher latency and worse service performance [4]. Fig. 1 illustrates the user mobility and the impact of latency.

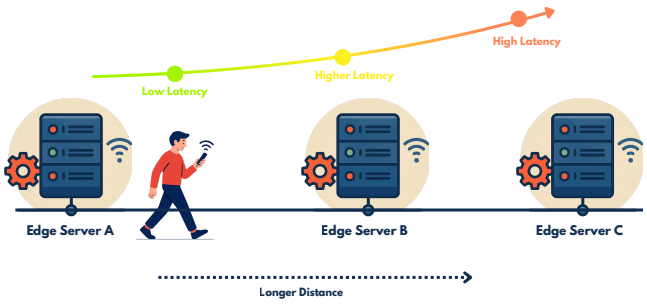


Fig. 1. Impact of user mobility on service latency in Multi-access Edge Computing environments.

Containerization has gained popularity in the world of effectively deploying applications in MEC environments. Applications can execute consistently across platforms due to containerization, which offers lightweight virtualization by packaging them with their needed dependencies. Containers are considered to be a good fit for edge nodes with limited resources because they require fewer resources, have quicker startup times, and are more scalable than traditional virtual machines [5]. Applications are often decomposed into microservices and deployed as containers in MEC systems, allowing for flexible and dynamic service management across different edge infrastructures.

One essential method for preserving service performance in MEC settings is container migration. To lower latency and preserve service reliability, the service container can be moved or migrated to a nearby node when a user moves away from the current edge server. Applications can dynamically adjust to user mobility as a result of this idea, which is frequently referred to as “follow-me" services, as defined in [6] and [7].

Despite all the benefits the container migration adds, it introduces some critical drawbacks. The container state must be transferred over the network as part of the migration process, which may add to the bandwidth usage, introduce additional computational overhead, and cause service outages. Additionally, unstable network conditions may cause migration to fail, which would further affect the system’s reliability. Container migration in edge computing has been widely studied, with the literature focusing on the most efficient way to perform the migration and the conditions that warrant the migration. This section reviews the literature to lay the foundation for the contribution of this study, highlighting the gap that this work addresses.

A. Container Migration Mechanisms

The migration of containers involves the migration of running applications across different physical hosts without stopping the processes. This migration process is mostly adopted in edge computing platforms to ensure continuity of service delivery in cases where the user leaves the current host of their application. Unlike restarting a container from the beginning, migration attempts to preserve the running state of the service, which is important for latency-sensitive and stateful applications.

The fundamental technology that facilitates this functionality on Linux-based operating systems includes CRIU (Checkpoint/Restore In User space). It involves the freezing of the running process, saving its complete status to the hard disk in the form of image files, moving the images to another machine, and restoring the process from where it stopped [8] [9]. CRIU can be used with container runtimes to support checkpoint/restore-based migration. One of the major discoveries in the literature is that the checkpoint phase is directly proportional to the size of memory allocated to the container; that is, a bigger container needs more time for the freezing and transferring phases and consequently causes longer downtime to the service [9].

Based on how the container state is transferred between the source and destination hosts, migration techniques are commonly classified into pre-copy [10], post-copy [11], and hybrid migration [12] approaches. During pre-copy container migration, the process of transferring pages from the origin server to the destination server takes place while the program is active on the origin server. Pages that are modified during the transfer are resent in later rounds. Once copying of most of the state has occurred, the container is momentarily paused, the dirty pages are copied, and then the container continues

its execution on the destination server. While such a method can minimize downtime, it may increase total transferred data when the application frequently modifies memory [10].

In post-copy migration, the container is first paused, and a minimal execution state is transferred to the destination host. The container will continue its execution in the target container, while missing pages will be retrieved from the source machine as required. This process results in fewer repeated page copies than in the case of pre-copy migration. The technique, however, is highly dependent on the source machine because the continuation of its execution is contingent on whether the source machine is available during the fetching process. In case the communication link breaks down, the service will be disrupted [11].

Hybrid migration includes elements of pre-copy and post-copy strategies. It involves copying a part of the container state before switching and moving the rest of the state after switching to the destination. The hybrid strategy tries to achieve a compromise between downtime, bandwidth, and failures [12].

Despite their differences in methodology and disadvantages, each of these migration approaches imposes a certain amount of overhead in terms of downtime, bandwidth consumption, power consumption, and risk of failure. Therefore, improving the migration mechanism alone is insufficient. It is equally important to determine whether migration should be performed in the first place, which motivates the framework proposed in this work.

B. Migration Decision Frameworks

Existing frameworks mainly address two questions: when to migrate and where to migrate. To address these questions, several strategies have been proposed in the literature, including threshold-based approaches, Markov Decision Process (MDP)-based frameworks, prediction-based methods, and resource-aware schemes. The question regarding optimal timing for migration has been extensively explored. Specifically, Wang et al. [6] solve the migration problem using a Markov Decision Process (MDP), in which the migration policy is developed by considering the mobility pattern of users and the distance between the user and the current edge node. Additionally, Shamsadini and Entezari-Maleki [13] utilize an MDP framework in their analysis of vehicular edge computing, and their migration decisions depend on path loss conditions resulting from the mobility of vehicles. On the other hand, in [14], migration occurs in response to node failures to ensure reliability and fault tolerance in the network. Latency minimization and service continuity are also considered. For instance, Tang et al. [15] develop a layer-aware migration approach to minimize the time required for initialization, computation, and migration. Further, Jin et al. [16] study container migration in industrial internet to reduce latency.

After making the decision to migrate, the destination edge server needs to be chosen. In [6], the destination edge is decided based on the policy obtained using the MDP model. In [17], an RNN-based trajectory prediction model is used to predict the future location of the vehicle and accordingly choose the destination edge server while taking into account the issues of service delay and load balancing. Network state and resource state information are used to select the destination in the mobile edge clouds to support real-time applications in [18]. In [15], container layer-sharing information from multiple users is used to select the destination edge server.

Different migration strategies have different optimization goals. For instance, some migration models prioritize reducing delays in service and keeping latency at a minimum [6] [17], while others place more emphasis on the fault tolerance aspect [14] [16]. Resource usage and migration costs are other critical issues tackled in traffic-aware and layer-aware systems [15] [18]. Despite the diversity of optimization objectives, these methods mainly seek to improve migration efficiency after the migration decision has already been made.

Despite the fact that these researchers use various techniques of optimization and prediction, they have something in common: the fact that migration must take place once the condition for that is met. As a result, prior research predominantly addresses the problem of when and where to migrate, but little attention has been paid to the necessity for the migration to take place at all. This issue drives the design of the proposed lightweight pre-migration decision scheme.

C. Migration Costs and Risks

Although container migration ensures service continuity, there are various costs and risks involved that should be taken into consideration prior to migration. One of the primary concerns is service downtime. Checkpointing involves pausing the container in order to save its state. The time taken by the container to save its state is directly dependent on the memory size of the container, as stated in [8]. For applications such as streaming videos and real-time analytics, which require near real-time responses, this disruption negatively impacts the user experience. If the migration process happens often, then these pauses tend to add up. Another important challenge is the accumulation of migration overhead caused by frequent migrations. Moving a container involves migrating the state of the memory of the container and the files. This consumes bandwidth that could otherwise serve user traffic. The authors of [18] noted that in edge networks with limited network bandwidth, this may lead to congestion. Migration operations also incur computational and energy costs. Checkpointing and data transmission use energy at both the source and destination edge nodes. The edge nodes generally run with limited energy resources, such as battery power, solar energy, or even electricity limitations. This is further discussed in [14]. Moreover, migration is not guaranteed to succeed. Network interruptions during migration can cause corruption of the container state. There may also be insufficient resources on the destination server during restoration. This may result in the failure of the entire system [14]. These challenges imply that enhancing the process of migration alone is not enough. It would be wise for an intelligent approach to not only decide how migration should take place, but also whether migration is desirable at all.

III. Research Gap and Problem Formulation

Despite the extensive research on the topic of container migration, the vast majority of proposed approaches concentrate on the problem of optimization of the migration process and the destination location for the container after migration.

Additionally, some works consider the problem of choosing the time of migration relying on threshold-based or model-driven triggers.

Nevertheless, all such approaches make the following assumption about the necessity of the migration once its triggering event occurs. The assumption does not take into account that the process of migration has a high cost and risk in terms of downtime, bandwidth consumption, energy consumption, and failures. And in some cases, the cost incurred could exceed the performance gain expected through migration. Moreover, network instability or transient resource congestion may be solved prior to the completion of the migration process, making such migration not only unnecessary but potentially harmful to QoS as well. Consequently, the problem of determining whether migration should be performed at all remains largely unexplored. Previous research on migration has focused primarily on answering the questions of when migration is needed and to which server migration must be made. The following question, however, remains largely unexplored:

Should migration be executed in the first place?: Therefore, there is a need to explore lightweight machine learning models for this pre-migration binary decision task. This research tries to fill this gap by presenting a machine learning decision model that can predict if migration is necessary according to the system state. Unlike prior work, this study introduces a lightweight pre-migration decision layer that filters unnecessary migrations before execution.

In this study, an MEC system is assumed where there are several edge servers, which provide cloud-based services to mobile clients via containerized application instances. As users move, the latency between the user and the serving edge node may increase, potentially degrading the QoS.

For each decision epoch t, the system will have a set of state variables reflecting the state of the system. Such state variables include current latency, post-migration latency, available bandwidth, container sizes, current CPU usage, and current memory usage, among others.

Given the above states, a decision needs to be made on whether migration should be started at epoch t or not. This problem can be viewed from the perspective of binary classification, where the target variable takes one of two values:

$$D(t) \in \{\text{Migrate}, \text{Wait}\}$$

where, Migrate indicates that migration should be executed immediately, whereas Wait indicates that migration should be postponed and the system should be re-evaluated at a later epoch.

The labeling of training samples is obtained using a counterfactual comparison strategy, described in detail in the following section.

Fig. 2 shows the position of the proposed framework in the MEC container migration pipeline. The framework acts as a pre-migration decision layer between the trigger condition and the migration executor, deciding whether to migrate or wait.

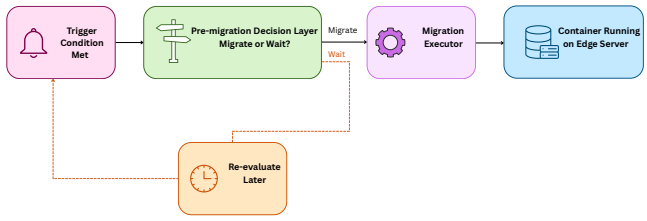


Fig. 2. Position of the proposed framework in the MEC migration pipeline.

IV. Methodology

This section describes the research methodology followed to make the classification problem for the container migration. The following subsections provide an explanation about each part of the experimental study, the dataset used, feature selection, models employed, and metrics used to evaluate the work.

A. Dataset

For this research, the utilized dataset was obtained via the use of EdgeSimPy [19], an open source Python simulator for edge computing. EdgeSimPy simulates a realistic MEC system that consists of 6 edge servers, 16 network switches arranged in a hexagonal mesh network, and 6 users operating containers in their devices. Fig. 3 illustrates the topology used in the dataset generation.

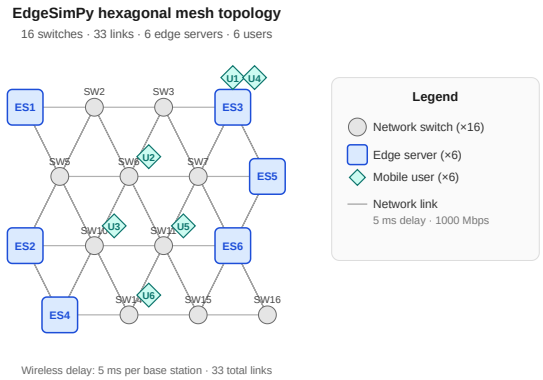


Fig. 3. EdgeSimPy simulation topology: a 4×4 hexagonal mesh of 16 network switches with 33 links, 6 edge servers, and 6 mobile users.

End-to-end latency is estimated within EdgeSimPy's routing engine by calculating the total wireless access latency at the base station (5 ms) and propagation time of the Dijkstra shortest path across the network (5 ms per hop). Thus, latency ranges from 5 ms to 25 ms. To obtain heterogeneous data, 30 simulation cases were conducted by varying two important variables in each simulation: container state size (8-200 MB) and mobility speed (5-35 seconds). Backbone link bandwidth was set to 1000 Mbps, consistent with real MEC infrastructure

TABLE I. DATASET FEATURES

Feature	Description	Range
latency_current_ms	End-to-end delay to current server	5–25 ms
latency_new_ms	Expected end-to-end delay after migration to candidate server	5–15 ms
latency_gain_ms	Difference in latency (migration improvement)	–5 to 20 ms
dist_current	Euclidean distance: user to current server	0–7.07
dist_new	Euclidean distance: user to candidate server	0–7.07
mobility	User movement direction: 0=stationary, 0.5=slow, 1.0=moving away	{0, 0.5, 1.0}
state_size_mb	Container state size in MB	8–200 MB
bandwidth_mbps	Available bandwidth on migration path in Mbps	Fixed at 1000 Mbps
dest_cpu_pct	CPU utilization on destination server	0–100%
dest_mem_pct	Memory utilization on destination server	0–50%
transfer_time_s	Estimated migration transfer time	0.07–1.64 s

specifications. The labeling scheme relied on a counterfactual comparison technique. In each situation, two simulations were conducted in the exact same environment: one where migration was never allowed, and another where migration was constantly conducted to the closest server at all times. At each decision time step, the end-to-end delay experienced in EdgeSimPy between the two policies was compared. If the delay value in the constantly migrating scenario was smaller than the one in the baseline scenario, then the time step was marked as Migrate; otherwise, it was Wait. As a point to note, the always migrate simulation algorithm employed for the generation of labels differs from the Always Migrate benchmark in Section VI, which makes predictions without regard to features or latencies, as opposed to the always migrate simulation algorithm, which generates latencies to generate labels. It is noted that although both feature generation and label derivation originate from the same simulation environment, the labels are derived from an independent comparison of latency outcomes under two distinct migration policies rather than directly from the input features, which attempts to mitigate the risk of closed-loop learning. The resultant data set has 5,000 samples, with each class having an equal number of samples (2,500 Migrate, 2,500 Wait), making the dataset a balanced dataset. Each sample consists of 11 features extracted from the simulation state at the time of the decision, summarized in Table I. Given that the hop delay for each link is constant at 5 ms, the latency will be an integer between 5 ms (co-located) and 25 ms (maximal hop distance in the network topology). Table II shows a sample of the simulation dataset generated for training the proposed machine learning model. Each row represents a system state, and the decision label assigned.

B. Decision Taking Model

A lightweight binary classifier is then constructed and trained on the labeled dataset generated. Input to the classifier will consist of system state features, and the output corresponds to one of two classes: Migrate or Wait. Thus, the classifier learns to determine whether migration should be executed immediately or postponed. Various machine learning and deep learning models have been used for analysis within this study, which include Logistic Regression, Decision Trees, Random Forests, Stacked Deep Neural Networks (DNN), and Stacked Long Short-Term Memory (LSTM) networks. This choice was made to allow the comparison of classical machine learning models and advanced deep learning models. Decision Trees were focused particularly in this work due to their suitability to MEC environments based on a number of advantages inherent to this algorithm. Firstly, decision trees do not require many computational resources for running and are therefore appropriate in resource-restricted edge systems. Secondly, decision trees make it possible to obtain understandable decision boundaries, which means that the process by which migration decisions are made can be easily explained. Finally, decision trees can detect nonlinear relations without scaling features.

1) Data preparation: The dataset of 5,000 records is divided into training, validation, and testing sets in a 70/10/20 split pattern, which results in 3,500 records being used for training, 500 records for validation, and 1,000 records being used for testing. The splitting ratio has been commonly used in machine learning literature in order to have enough data for training and still maintain an adequate test set. Since the synthetic dataset is balanced with 2,500 samples for each category, no form of resampling was performed. In addition, feature scaling did not take place since the key algorithm used in this project, which is the Decision Tree classifier, is insensitive to feature scaling.

2) Model architecture: The Decision Tree classifier was selected as the primary model because of its simplicity, interpretability, and low computational overhead. The model was trained using the maximum tree depth set to 3, 4, 5, 6, and 7, showing model performance across different depths, allowing the algorithm to have enough expressiveness to learn the non-linear dependencies among the features. All other hyperparameters were set to their default values. The model was trained using the training set containing 3,500 samples and tested on the test set containing 1,000 samples.

Along with the proposed Decision Tree approach, other alternative approaches were analyzed and compared with each other, including Logistic Regression, Random Forest, Stacked DNN, and Stacked LSTM. Also, two base methods were implemented: the always migrate, migration is made on every decision cycle, and the never migrate, migration is never performed. This helps evaluate the efficiency of the proposed decision framework. In addition, a latency-threshold rule was implemented as a realistic rule-based baseline, migrating whenever the expected latency gain exceeds a threshold tuned on the validation set.

C. Evaluation Metrics

To check how well the classification model performs on the task, several confusion matrix metrics were used. Several metrics were measured to give a clearer picture of the model's performance. Accuracy measurement was used as the primary metric, defined as the percentage of correctly classified cases compared to the total number of cases, giving an indication of how well the model classifies, as shown in Eq. (1):

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

TABLE II. SAMPLE OF THE SIMULATED DATASET FOR TRAINING THE PROPOSED MODEL

Feature	Sample 1	Sample 2	Sample 3	Sample 4
Latency current (ms)	20	20	15	10
Latency new (ms)	10	10	10	5
Latency gain (ms)	10	10	5	5
Dist current	4.472	4.472	4	1.414
Dist new	2	1.414	1.414	0
Mobility	0.5	0.5	0.5	0.5
State size (MB)	18	10	25	12
Bandwidth (Mbps)	1000	1000	1000	1000
Dest CPU (%)	25	0	25	0
Dest mem (%)	50	0	50	0
Transfer time (s)	0.1475	0.0819	0.2048	0.0983
Label	Wait	Migrate	Wait	Migrate

In addition to accuracy, precision, recall, and F1-score were all measured for the model. Precision measures the number of truly positively classified cases to the number of cases classified as positive, as shown in Eq. (2):

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

Recall is defined as the ratio of correctly classified positives to all instances belonging to its own class, as shown in Eq. (3):

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

While the F1-score provides a balance between the two measures, precision and recall, as shown in Eq. (4):

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

The Area Under the ROC Curve (AUC) summarizes the trade-off between the True Positive Rate and False Positive Rate across all classification thresholds, serving as an aggregate measure of model performance, as shown in Eq. (5):

$$\text{AUC} \approx \sum_{i=1}^{n-1} \frac{(\text{FPR}_{i+1} - \text{FPR}_i)(\text{TPR}_i + \text{TPR}_{i+1})}{2} \tag{5}$$

V. RESULTS

This section presents a comprehensive analysis of the outcomes achieved from applying different machine learning and deep learning approaches. Results are presented in Fig. 4, Fig. 5, Table III, Table IV, Table V, and Table VI. Subsequently, a more detailed discussion of the outcomes is presented in the next section.

A. Comparison of Learning Models

Various machine learning and deep learning methods were trained on the training set and evaluated on the test dataset consisting of 1,000 cases. This experiment was aimed at identifying the best-fitting machine learning or deep learning method for the proposed pre-migration decision algorithm. Table III presents comparative results regarding the accuracy, precision, recall, F1-score, and AUC value for Logistic Regression, Decision Tree, Random Forest, Stacked DNN, and Stacked LSTM approaches.

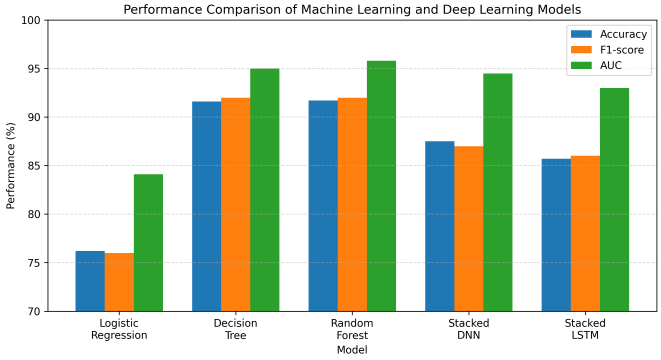


Fig. 4. Performance comparison of machine learning and deep learning models across accuracy, F1-score, and AUC.

As shown in Fig. 4, Logistic Regression achieved the lowest performance, achieving an accuracy of 76.20%, an F1-score of 76.11%, and an AUC of 84.13%. The poor performance indicates that linear models cannot be used to model the nonlinear interactions between the features related to migrations.

The deep learning models showed better results. The Stacked DNN was able to obtain an accuracy of 87.80% with an F1-score of 87.74% and an AUC of 94.38%, while the Stacked LSTM model yielded an accuracy of 85.70% with an F1-score of 85.66% and an AUC of 92.74%. Even though these models provided very promising predictions, they needed more computational capacity and engineering. The Decision Tree classifier achieved an accuracy of 91.60%, an F1-score of 91.58%, and an AUC of 94.86%, outperforming both deep learning models while maintaining low computational overhead and providing interpretable decision rules. Random Forest achieved very similar accuracy and a slightly higher AUC of 95.80%. However, Random Forest sacrifices interpretability and introduces additional computational complexity due to the ensemble structure. A latency-threshold rule was also included as a realistic rule-based baseline, achieving an accuracy of 60.40% and an F1-score of 56.87%, and is discussed in detail in Section VI-B.

Thus, the Decision Tree is the most optimal in terms of prediction accuracy, performance speed, and interpretability. This makes the model especially appropriate to be used as a pre-migration decision layer within MEC infrastructure.

To assess the statistical reliability of these results, a repeated

TABLE III. PERFORMANCE COMPARISON OF THE EVALUATED MACHINE LEARNING AND DEEP LEARNING MODELS

Model	Accuracy	Precision	Recall	F1-score	AUC
Logistic Regression	76.20%	76.61%	76.20%	76.11%	84.13%
Decision Tree	91.60%	91.92%	91.60%	91.58%	94.86%
Random Forest	91.80%	92.10%	91.80%	91.79%	95.80%
Stacked DNN	87.80%	88.51%	87.80%	87.74%	94.38%
Stacked LSTM	85.70%	86.08%	85.70%	85.66%	92.74%
Latency-Threshold Rule	60.40%	65.46%	60.40%	56.87%	60.40%

stratified k-fold cross-validation procedure was conducted (10 folds, 5 repeats, random seed 42) on the classical machine learning models, with results summarized in Table IV. The Decision Tree achieved a mean accuracy of 91.53% $\pm$ 1.25% and the Random Forest achieved 91.52% $\pm$ 1.29%, a difference of 0.01% with fully overlapping confidence intervals. This confirms that the performance gap between the two models is not statistically meaningful, and further supports the selection of the Decision Tree as the primary model on the basis of interpretability and computational efficiency rather than predictive performance alone.

To substantiate the claim that the Decision Tree is suitable for resource-constrained edge environments, runtime performance measurements were conducted and are summarized in Table V. The Decision Tree trained in 0.006 seconds and completed inference on 1,000 samples in 0.320 ms, with a serialized model size of 47.140 KB. By contrast, the Random Forest required 0.316 seconds to train, 22.785 ms for inference, which is approximately 71 times slower than the Decision Tree, and occupied 4,199.70 KB of memory, roughly 89 times larger. These measurements confirm that the Decision Tree is deployable on resource-constrained edge nodes, where memory and inference latency are critical constraints. All measurements were conducted on a standard CPU environment; absolute values would differ on dedicated edge hardware, but the relative advantages of the Decision Tree over the Random Forest remain consistent regardless of platform.

B. Effect of Decision Tree Depth

In order to analyze the effect of model complexity on the classification process, the Decision Tree classifier was tested for maximum depths varying between 3 and 7, in addition to the default model. Fig. 5 illustrates the impact of increasing the maximum tree depth on the Accuracy and AUC values of the Decision Tree classifier. The results attained from this process are listed in Table VI. As would be expected, the increase in the depth of the decision tree led to enhanced accuracy scores. A shallow tree having a depth of 3 gave accuracy and AUC values of 78.20% and 84.71%, respectively, suggesting inadequate representation power. With increases in tree depth, accuracy progressively increased up to 87.80% and 92.56% AUC at a depth of 7. The default Decision Tree classifier had superior performance levels, with accuracy and AUC being 91.60% and 94.86%, respectively. This suggests that the problem involves complex relationships requiring an expressive model. Nonetheless, shallower trees were also capable of performing effectively, demonstrating that lightweight models could be applied in edge environments.

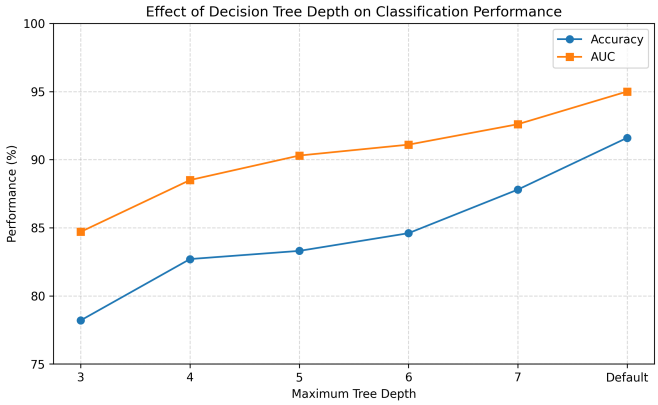


Fig. 5. Effect of decision tree depth on classification performance.

VI. DISCUSSION

A. Suitability of Decision Trees for MEC Environments

It can be seen from the results of the experiment that the Decision Tree classifier is an appropriate classifier to address the problem of pre-migration decision presented above. The Random Forest showed the highest AUC rate, achieving 95.80%, and also reached the accuracy of 91.80%. However, the advantage in comparison with the Decision Tree was insignificant. Meanwhile, the Decision Tree classifier achieved 91.60% of accuracy and 94.86% of AUC at significantly lower complexity of the model. When compared with other classifiers, which utilize more complicated architectures such as deep learning networks (Stacked DNN and Stacked LSTM), the Decision Tree is significantly more productive. It not only allows getting more accurate results but also does not need much computational power. Another advantage of the use of decision trees is their interpretability, meaning that the obtained rules could be easily interpreted. This feature is especially significant for MEC as edge nodes have relatively low computational capabilities [20]. Consequently, the Decision Tree classification algorithm provides an effective balance between accurate predictions, efficient computations, and interpretability, making it a suitable candidate for deployment as a pre-migration decision layer in edge computing infrastructures. Cross-validation results confirm that this performance advantage over the Random Forest is statistically consistent rather than an artefact of a specific train/test split, and runtime measurements provide direct evidence of the Decision Tree's computational suitability for resource-constrained edge deployment.

B. Comparison with Baseline Policies

Further analysis regarding the efficacy of the proposed framework can be performed through its comparison with three

TABLE IV. CROSS-VALIDATION RESULTS (10-FOLD, 5 REPEATS, RANDOM SEED 42) FOR CLASSICAL MACHINE LEARNING MODELS

Model	Accuracy	F1-score	AUC
Logistic Regression	77.96% ± 1.72%	77.85% ± 1.73%	85.34% ± 1.80%
Decision Tree	91.53% ± 1.25%	91.50% ± 1.26%	95.07% ± 1.24%
Random Forest	91.52% ± 1.29%	91.50% ± 1.30%	96.16% ± 0.94%

TABLE V. RUNTIME PERFORMANCE COMPARISON OF CLASSICAL MACHINE LEARNING MODELS, MEASURED ON A STANDARD CPU ENVIRONMENT. INFERENCE TIME IS REPORTED PER 1,000 TEST SAMPLES

Model	Training Time (s)	Inference Time (ms/1000 samples)	Model Size (KB)
Logistic Regression	0.350	0.224	0.790
Decision Tree	0.006	0.320	47.140
Random Forest	0.316	22.785	4199.700

TABLE VI. PERFORMANCE OF THE DECISION TREE CLASSIFIER FOR DIFFERENT MAXIMUM DEPTHS

Max Depth	Accuracy	Precision	Recall	F1-score	AUC
3	78.20%	79.63%	78.20%	77.93%	84.71%
4	82.70%	83.77%	82.70%	82.56%	88.45%
5	83.30%	83.57%	83.30%	83.87%	90.34%
6	84.60%	85.07%	84.60%	84.55%	91.08%
7	87.80%	88.00%	87.80%	87.78%	92.56%
Default	91.60%	91.92%	91.60%	91.58%	94.86%

baselines, two unconditional and one rule-based. The first approach migrates the data at each point of time, regardless of the system state. In contrast, the second policy never performs any migration actions. As demonstrated in Table VII, the considered simple baseline strategies provided equal performances in terms of accuracy and F1-score (both values being equal to 50.0% and 33.3%, respectively). Such an outcome seems quite predictable, since these approaches ignore the temporal dynamics of MEC systems and operate based on predetermined decisions. In comparison, the suggested Decision Tree model produced an accuracy rate of 91.60% and an F1-Score value of 91.58%, clearly outperforming the other two approaches. The obtained results have shown that random migration of containers or non-migration at all are ineffective approaches. To provide a more realistic point of comparison, a latency-threshold rule was also evaluated as an additional baseline. This rule migrates whenever the expected latency gain exceeds 0 ms — a simple and deployable heuristic that represents the kind of trigger-based policy commonly used in practice. Despite being more informed than the unconditional baselines, the latency-threshold rule achieved an accuracy of only 60.40% and an F1-score of 56.87%, with an MRR of 21.40% and a DRR of 24.02%. This result demonstrates that even a sensible single-feature rule is insufficient for this problem, and confirms that learning from the full system state across all 11 features is necessary to achieve reliable pre-migration decisions.

Beyond classification accuracy, the practical benefit of the proposed framework can be quantified directly in terms of migration overhead avoided. To this end, two additional metrics are introduced: Migration Reduction Ratio (MRR) and Downtime Reduction Ratio (DRR), both computed relative to the Always Migrate baseline. MRR captures the percentage decrease in the number of migrations triggered by the Decision Tree compared to a policy that migrates unconditionally, as shown in Eq. (6):

$$\mathrm{MRR} = \frac{M_{\mathrm{always}} - M_{\mathrm{DT}}}{M_{\mathrm{always}}} \times 100 \tag{6}$$

DRR captures the corresponding percentage decrease in cumulative transfer time, using the transfer_time_s feature described in Table I as a proxy for migration-induced downtime, as shown in Eq. (7):

$$\mathrm{DRR} = \frac{T_{\mathrm{always}} - T_{\mathrm{DT}}}{T_{\mathrm{always}}} \times 100 \tag{7}$$

where, M_{always} and T_{always} denote the number of migrations and total downtime under the Always Migrate policy, and M_{DT} and T_{DT} denote the corresponding values produced by the Decision Tree's predictions on the test set. Table VIII summarizes these results. Out of 1,000 test samples, the Always Migrate baseline triggers a migration at every decision epoch, resulting in 1,000 migrations and a cumulative downtime of 339.48 seconds. The Decision Tree, by contrast, predicts Migrate in only 544 cases, reducing the number of migrations by 45.60%. More notably, the cumulative downtime under the Decision Tree's policy falls to 98.79 seconds, a reduction of 70.90%. The disparity between the two reduction ratios indicates that the Decision Tree does not simply migrate less often at random; it disproportionately withholds migration in cases with higher transfer cost, while still permitting migration when the cost is low and the latency benefit justifies it.

The more crucial conclusion of this experiment is the fact that the obtained results provide evidence for the hypothesis of this study; namely, migration is not something mandatory after reaching the required trigger threshold. Thus, migration should be performed in a smart manner, taking into account the current state of the system and the balance between the advantages and disadvantages of migration.

C. Limitations of the Study

Despite the good performance of the suggested framework, there are still some limitations that need to be discussed. First, the dataset in the current study is simulated by means of the EdgeSimPy software package, instead of being collected from a

TABLE VII. COMPARISON OF THE PROPOSED FRAMEWORK AGAINST BASELINE MIGRATION POLICIES.

Policy	Accuracy	F1-score
Always Migrate	50.0%	33.3%
Never Migrate	50.0%	33.3%
Latency-Threshold Rule	60.4%	56.87%
Proposed Decision Tree	91.6%	91.58%

TABLE VIII. MIGRATION AND DOWNTIME STATISTICS FOR THE PROPOSED DECISION TREE RELATIVE TO BASELINE MIGRATION POLICIES.

Policy	Number of Migrations	Total Downtime (s)
Always Migrate	1000	339.48
Never Migrate	0	0.00
Latency-Threshold Rule	786	257.92
Proposed Decision Tree	544	98.79

live MEC implementation. While simulations provide a chance for controlled experiments with many possible mobility and container migration options, it is unlikely to completely reflect reality.

Additionally, even though the efficiency was calculated using the classification performance metrics, the influence of the current approach on the energy consumed, failure rate, and migration traffic was not studied. In addition, the dataset was generated from a single fixed topology of 6 edge servers, 16 network switches, and a backbone bandwidth of 1,000 Mbps. The generalisability of the reported performance figures to deployments with substantially different scale, heterogeneous network conditions, or varying bandwidth configurations remains to be investigated. Also, the proposed model was tested using offline experiments without integrating it into any existing container orchestration system, such as Kubernetes and KubeEdge. Hence, there is still a need for further testing for the efficiency of the proposed framework in practice.

VII. CONCLUSION

Container migration is one critical approach in ensuring service continuity in MEC. However, most previous research studies have concentrated on identifying the right time and location to perform container migration by taking the positive effect of the action for granted, even though there are numerous costs and risks associated with migration, such as service unavailability, network bandwidth usage, power consumption, and migration failures. Therefore, in order to solve the problem above, this study puts forward a pre-migration decision-making framework in the form of a binary classifier as a filter before migration. According to the present system status, the proposed approach will decide either to start the migration process or to wait until another period of time. A balanced dataset containing 5,000 samples was generated using EdgeSimPy, and labels were assigned through a counterfactual comparison strategy. Several machine learning and deep learning models were evaluated, including Logistic Regression, Decision Tree, Random Forest, Stacked DNN, and Stacked LSTM.

According to the experimental findings, Decision Tree had an accuracy of 91.60%, an F1-score of 91.58%, and AUC of 94.86%, which is better than the deep learning algorithms examined in terms of computation costs and interpretability of decisions. Moreover, the presented solution provided much better results than all evaluated baseline migration strategies, including the unconditional Always Migrate and Never Migrate policies and a realistic latency-threshold rule, confirming the necessity of intelligent migration decision-making within MEC. Cross-validation results further confirm that the Decision Tree's performance advantage is statistically consistent, and runtime measurements demonstrate its suitability for deployment on resource-constrained edge nodes.

In conclusion, it can be stated that migration cannot be treated as an automatic action based on certain triggers. Rather, the introduction of the decision-making component can ensure avoiding unnecessary migrations and make edge computing more efficient.

A. Future Work

There are various areas for future work. The proposed decision-making framework can be tested with practical trace datasets collected from live edge computing environments, and validated across additional simulation environments with varied topologies, network configurations, and bandwidth conditions, as well as deployed into practice within orchestration engines like Kubernetes and KubeEdge. Secondly, other objectives, including energy usage, migration probability of failure, and bandwidth utilization, can be considered as part of the decision process. Furthermore, reinforcement and online learning algorithms can be looked into to achieve adaptive migration policies. Additionally, runtime performance measurements on dedicated edge hardware platforms, such as Raspberry Pi or similar resource-constrained devices, would provide more precise evidence of the framework's suitability for edge deployment. Statistical significance analysis, including cross-validation, should also be extended to deep learning models such as the Stacked DNN and Stacked LSTM evaluated in this work. Finally, evaluating the framework on larger and more heterogeneous edge infrastructures would provide further insights into its scalability and applicability.

REFERENCES

[1] 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.

[2] 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.

[3] S. Alshattnawi and A. M. AlSobeh, "A cloud-based iot smart water distribution framework utilising bip component: Jordan as a model," *International journal of cloud computing*, vol. 13, no. 1, pp. 25–41, 2024.

[4] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, "On multi-access edge computing: A survey of the emerging 5g network edge cloud architecture and orchestration," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1657–1681, 2017.

[5] C. Pahl, "Containerization and the paas cloud," *IEEE Cloud Computing*, vol. 2, no. 3, pp. 24–31, 2015.

[6] S. Wang, R. Urgaonkar, M. Zafer, T. He, K. Chan, and K. K. Leung, "Dynamic service migration in mobile edge computing based on markov decision process," *IEEE/ACM Transactions on Networking*, vol. 27, no. 3, pp. 1272–1288, 2019.

[7] T. Taleb and A. Ksentini, "Follow me cloud: interworking federated clouds and distributed mobile networks," *IEEE Network*, vol. 27, no. 5, pp. 12–19, 2013.

[8] L. Ma, S. Yi, and Q. Li, "Efficient service handoff across edge servers via docker container migration," in *Proceedings of the second ACM/IEEE symposium on edge computing*, 2017, pp. 1–13.

[9] A. Tošić, "Run-time application migration using checkpoint/restore in userspace," *Journal of Web Engineering*, vol. 23, no. 5, pp. 735–748, 2024.

[10] C. Clark, K. Fraser, S. Hand, J. G. Hansen, E. Jul, C. Limpach, I. Pratt, and A. Warfield, "Live migration of virtual machines," in *Proceedings of the 2nd conference on Symposium on Networked Systems Design & Implementation-Volume 2*, 2005, pp. 273–286.

[11] M. R. Hines, U. Deshpande, and K. Gopalan, "Post-copy live migration of virtual machines," *ACM SIGOPS operating systems review*, vol. 43, no. 3, pp. 14–26, 2009.

[12] S. Sahni and V. Varma, "A hybrid approach to live migration of virtual machines," in *2012 IEEE international conference on cloud computing in emerging markets (CCEM)*. IEEE, 2012, pp. 1–5.

[13] A. Shamsadini and R. Entezari-Maleki, "Time-aware mdp-based service migration in 5g mobile edge computing," in *2022 27th International Computer Conference, Computer Society of Iran (CSICC)*, 2022, pp. 1–5.

[14] S. Aleyadeh, A. Moubayed, P. Heidari, and A. Shami, "Optimal container migration/re-instantiation in hybrid computing environments," *IEEE Open Journal of the Communications Society*, vol. 3, pp. 15–30, 2022.

[15] Z. Tang, F. Mou, J. Lou, W. Jia, Y. Wu, and W. Zhao, "Multi-user layer-aware online container migration in edge-assisted vehicular networks," *IEEE/ACM Transactions on Networking*, vol. 32, no. 2, pp. 1807–1822, 2024.

[16] X. Jin, S. He, and Y. Chen, "Container migration for edge computing in industrial internet with joint latency reduction and reliability enhancement," *Scientific Reports*, vol. 14, no. 1, p. 25394, 2024.

[17] W. Zhang, J. Luo, L. Chen, and J. Liu, "A trajectory prediction-based and dependency-aware container migration for mobile edge computing," *IEEE Transactions on Services Computing*, vol. 16, no. 5, pp. 3168–3181, 2023.

[18] S. Maheshwari, S. Choudhury, I. Seskar, and D. Raychaudhuri, "Traffic-aware dynamic container migration for real-time support in mobile edge clouds," in *2018 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*, 2018, pp. 1–6.

[19] P. S. Souza, T. Ferreto, and R. N. Calheiros, "Edgesimpy: Python-based modeling and simulation of edge computing resource management policies," *Future Generation Computer Systems*, vol. 148, pp. 446–459, 2023.

[20] S. Alshattnawi and M. Al-Marie, "Spider monkey optimization algorithm for load balancing in cloud computing environments." *Int. Arab J. Inf. Technol.*, vol. 18, no. 5, pp. 730–738, 2021.

A Geospatially-Aware Hybrid Deep Learning Framework for Violence Detection and Sentiment Analysis in Airline Reviews

Fahad Alanazi[1]*, Osama Rabie[2], Wafi Bedewi[3]
Department of Information Systems, Faculty of Computing and Information Technology,
King Abdulaziz University, Jeddah, Saudi Arabia[1,2,3]
Center of Research Excellence for Smart Environment, King Abdulaziz University, Jeddah, Saudi Arabia[2]

***Abstract*—Passenger reviews and feedback provide valuable operational insights for the aviation industry. However, existing sentiment analysis approaches rarely capture safety-related signals such as aggressive or violent language, and they often overlook the integration of geographic context. This study proposes a geospatially-aware deep learning framework for sentiment classification and violence detection in airline passengers reviews. A three-class dataset was developed using 5,475 airline passenger reviews, including 300 manually verified Skytrax reviews containing aggressive, violent, abusive, or threat-related narratives. Several machine learning and deep learning models were evaluated, including TF-IDF with SVM, TextCNN, a parallel CNN-BiGRU model with attention, BiLSTM, and fine-tuned DistilBERT. Experimental results demonstrated that DistilBERT achieved the best overall performance obtaining 88% accuracy, macro F1-score of 0.874 , and an F1-score of 0.87 for the violent/aggressive class. Furthermore, the proposed framework integrates a geospatial analysis layer that links review predictions with route-level coordinates to identify geographic concentrations of adverse passenger experiences. The findings demonstrate the feasibility of combining sentiment analysis, violence detection, and geospatial intelligence into a unified aviation analytics framework that supports safety-oriented operational monitoring and decision making.**

Keywords*—Hybrid deep learning; ensemble learning; natural language processing; geospatial intelligence; violence detection; passenger experience

I. Introduction

The air transport industry is undergoing rapid digital transformation, accompanied by a steady increase in operational complexity. Online review platforms continuously generate user-generated content that reflects passenger experiences related to service quality, operational disruptions, delays, staff behavior, and safety concerns [9]. Passenger feedback has thus become a real-time stream of operational intelligence. Unlike post-travel surveys, which are often limited by recall bias and rigid questionnaires, free-form passenger reviews capture immediate and unfiltered traveler experiences. Such narratives provide valuable insight into service quality, operational performance, and safety-related issues that are rarely captured by traditional Key Performance Indicators (KPIs) [26], [13].

Recent advances in Natural Language Processing (NLP) have accelerated the adoption of automated sentiment analysis in aviation research. Early studies (e.g., [27]) that relied on lexicon-driven methods and conventional machine learning (ML) algorithms often failed to capture contextual nuances in the noisy, informal language of airline review data. In contrast, deep learning (DL) methods have proven highly effective at handling the noise and high dimensionality inherent in social media and review data [9]. While traditional models such as Support Vector Machines (SVM) and Naïve Bayes frequently underperform in this domain, ensemble-based approaches like XGBoost have shown stronger results [6].

Despite these advances, most existing studies focus primarily on coarse-grained sentiment classification (e.g., positive, negative, or neutral) and service-related attributes such as seat comfort or onboard experience. Comparatively little attention has been given to identifying aggressive, abusive, or threat-related passenger narratives. General-purpose toxicity detection models have made significant progress, but their accuracy degrades when applied to domain-specific contexts such as aviation, largely due to the highly imbalanced nature of threat-related language. Violent expressions constitute less than 1% of all airline reviews [34], yet their operational impact is disproportionately high.

Another major limitation of current aviation analytics is the lack of spatial context. Air travel is inherently geographic, yet passenger review analysis is rarely performed with explicit reference to the associated geographic areas. Geospatial sentiment analysis, supported by geocoding and visualization techniques such as heatmaps, has recently been explored to detect patterns across regions [5]. More recent work in geospatial artificial intelligence (GeoAI) suggests that spatial data can provide an additional analytic layer, enabling more contextually informed interpretation of user-generated content and transport risk signals [42], [43]. When combined with geographic information, sentiment signals can help pinpoint "hotspots" of dissatisfaction and aggression, potentially linked to specific bottlenecks in transit hubs or along routes.

Main contributions are summarized as follows:

- Dataset: Created a cleaned three-class aviation review dataset of 5,475 unique reviews from 5,799 raw Skytrax entries (2,991 negative, 2,184 positive, 300 manually verified violent/aggressive).
- Classification pipeline: Designed a refined pipeline that separates violent/aggressive passenger narratives from ordinary negative sentiment.

*Corresponding author.

- Hybrid model: Developed a Parallel CNN–BiGRU with Attention model, extending the original CNN-GRU architecture for better multi-scale feature fusion and context modeling.
- Comprehensive evaluation: Compared TF-IDF SVM, Improved TextCNN, Parallel CNN-BiGRU Attention, BiLSTM, and fine-tuned DistilBERT. DistilBERT achieved the best performance (88% accuracy, 0.87 violent/aggressive F1).
- GeoAI pipeline: Introduced a geospatial intelligence layer that links review predictions with route-level coordinates to identify geographic hotspots of dissatisfaction and aggressive behavior.

The remainder of this study is structured as follows. Section II reviews related work. Section III describes the construction of the corrected dataset, leakage-free training protocol, preprocessing steps, the improved model architecture, and the experimental setup. Section IV presents the geospatial intelligence analysis, including route-level coordinate mapping and hotspot visualization. Section V reports the experimental results and discussion, comparing classical, neural, hybrid, and transformer-based models. Limitations and future directions are discussed in Section VI, followed by conclusions in Section VII.

II. Literature Review

This section is organized into four subsections to review recent studies on sentiment analysis in the aviation industry, text-based violence detection, hybrid neural architectures, and geospatial intelligence.

A. Sentiment Analysis in Aviation

Contrary to the first studies, in which the guest's opinions are considered as complementary data, the literature nowadays considers them as a source of operational and security intelligence. Hasib et al. [9] conducted a systematic literature review of 18 studies and highlighted a clear shift in the industry toward deep learning (DL) paradigms.

However, known classical machine learning baselines, such as Support Vector Machines (SVM) and Naive Bayes, have been seen to be non-performing when faced with the high dimensionality and non-uniform noise characteristic of aviation data from social media. On the other hand, ensemble-based architectures, such as XGBoost, have been shown to be much more effective at capturing small variations in passengers' sentiments during critical events, for instance, a systemic flight delay or technical disturbance [6].

A significant technological advancement has been the ability to map unstructured text to quality dimensions. Raihen et al. [21] achieved 90.13% accuracy using a multi-strategy approach. Other studies (e.g., [3], [2], [19], [13]) have explored aspect-based sentiment analysis, tourist satisfaction, airline brand reputation, and hybrid machine learning pipelines. Ensemble sentiment methods have also improved robustness in noisy or multi-dimensional textual signals [16], [22]. However, a persistent gap remains: current approaches [11] are largely confined to general polarity (positive/negative/neutral) or surface-level service evaluations. Notably, no existing work explicitly differentiates aggressive or violent intent from everyday dissatisfaction in airline reviews.

B. Violence and Threat Detection in Text

"Air rage" and hostile communication have emerged as recognized challenges in aviation security and passenger safety. Cabin crew perspectives indicate that disruptive passenger behavior has direct operational and safety implications [20]. In the broader NLP literature, hate speech, violent expression, and threat detection have been studied using machine learning and deep learning methods [24], [12], [1], [25]. Although general-purpose toxicity detection has made significant progress, these models produce suboptimal results in the aviation context, where threat-related language is extremely imbalanced. Security assessments from 2025 report that despite constituting less than 1% of all reviews, violent incidents have disproportionately severe operational consequences [34].

To augment minority threat classes, researchers have explored SMOTE and Generative Adversarial Networks (GANs) [35], [28]. However, text-based augmentation can produce "semantic hallucinations" (i.e., artificially generated sentences that lack coherent meaning). To address this challenge without relying solely on synthetic examples, the present study incorporates a manually verified subset of real Skytrax reviews containing violent or aggressive narratives, while maintaining a realistic proportion of safety-critical reviews.

C. Hybrid Neural Architectures for Text Classification

The multi-scalar nature of passenger reviews has been seen as a crucial element that should be considered in neural architecture. In the domain of short-to-medium length textual data, hybrid models such as CNN-GRU have been found to be effective because they combine local feature extraction with sequential representation learning [23], [22], [37]. Related CNN–recurrent structures have also been explored in flight-related forecasting and other high-dimensional sequence-classification contexts, which supports the broader relevance of CNN–GRU feature fusion [8], [38], [39].

Specifically, prior comparative studies [7] and [22] have demonstrated that parallel CNN-RNN architectures avoid the representational bottlenecks of sequential pipelines. Self-attention for feature fusion has also been empirically validated [39]. The Parallel CNN–BiGRU with Attention is presented as an evidence-informed neural baseline, while DistilBERT [32] remains the best-performing benchmark in this study.

CNN layers have been used as local n-gram feature extractors, which are helpful in identifying trigger words which are associated with immediate aggression or failure of service [14]. GRU layers have been used to resolve the vanishing gradient problem that occurs on longer, more detailed reviews. GRUs' gating structure has been observed to be an efficient alternative that can also be used with greater computational efficiency to reproduce the sequential emotional arc of a passenger's journey. Although effective in such fields as movie reviews [7] and smartphone reviews, their use in geospatially-tagged aviation reviews is virtually untouched.

D. Geospatial Intelligence for Sentiment Analysis

Lack of spatial context is said to be the biggest "blind spot" in aviation analytics today. Place is inherent to air travel, yet passenger reviews are generally treated as place-less. Recent studies have shown that geospatial sentiment analysis can be used to find hot spots for global events through geocoding and heatmap visualization, as demonstrated by Ali and Sibaroni [5]. More broadly, intelligent transportation and aviation-safety studies show that spatial patterns, hotspot ranking, and location-aware NLP can support the interpretation of safety and mobility risks [29], [15], [42], [43].

Foundational GIScience and GeoAI research has shown that geographic information can provide a rigorous analytical layer for interpreting spatially referenced events and text-derived signals [42], [43]. Adding geo-coordinates to sentiment and safety signals makes it possible to map hotspots of frustration onto specific routes or hubs, and local spatial statistics such as Getis–Ord statistics can help distinguish clustered risk from random variation [40], [41]. This branch of research is significant because it moves beyond textual classification alone and helps identify where repeated operational problems occur.

E. Summary

As discussed, several studies have advanced aviation safety analytics, violence detection, class imbalance mitigation, and geospatial text mining, yet none integrate all three pillars. Alharbi et al. [4] provided a violent-review dataset but used manual resampling without spatial analysis. Nanyonga and Wild [18], [17] benchmarked GRU, BLSTM, and DistilBERT on incident reports, identifying class imbalance as a key challenge. Zhang et al. [28] proposed KG-enhanced LLM generation for synthetic safety reports to reduce hallucination. Iddrisu et al. [10] introduced AIRNODE, a graph attention network for destination classification using reviews, achieving 84% accuracy and highlighting geographic context. Despite these contributions, no prior framework combines a dedicated violence detection module, statistically evaluated geospatial clustering, and linguistically controlled synthetic augmentation. Table I summarizes and compares these recent studies against the proposed framework.

III. Materials and Methods

A. Dataset Construction and Real-World Violence Subset

A three-class airline review dataset was constructed using two sources. The first dataset (DS1) contained 5,499 airline passenger reviews obtained from the Mendeley Data repository [30], originally collected from the Skytrax review platform [31], with binary `customer_review` labels (0 = not recommended, 1 = recommended). After data cleaning, outlier removal using the interquartile range (IQR) on text length, and lemmatization, DS1 was reduced to 5,175 clean reviews. The second dataset (DS2) consisted of 300 manually verified reviews containing explicit aggressive, abusive, threatening, or security-conflict language.

Table II shows the corrected label structure. The most important improvement was that the violent/aggressive class was represented as a separate real-world class rather than being merged with ordinary negative sentiment.

Both datasets were preprocessed by lowercasing, removing URLs, non-ASCII characters, and expanding contractions, followed by tokenization and lemmatization using NLTK's WordNet lemmatizer. After merging and removing duplicates, the final MrgDS comprised 5,475 unique reviews distributed as: class 0 (negative/not recommended = 2,991), class 1 (positive/recommended) = 2,184, and class 2 (violent/aggressive) = 300.

The violent/aggressive subset has not been included as 'generic negative sentiment.' Instead, it was assigned its own class label because there was a different operational risk class associated with violent or threatening passenger stories as opposed to dissatisfaction. Reviews were kept if they specifically referenced aggression, threat, abusive behaviour, harassment, humiliation, physical confrontation, police/security confrontation, and/or staff/passenger intimidation. This guaranteed that the third class was indeed language behaviour relating to safety and not only a lack of customer satisfaction.

Table III show DS1 before and after data cleaning process. To avoid any bias while models training, the reviews texts were examined. The extreme long reviews were recognized as an outliers and removed from the dataset as shown in the review length distribution histogram and boxplot in Fig. 1 . Fig. 2 shows the classes distribution after removing outliers ensuring that the classes balance is preserved. Fig. 3 describes the transformation of the original binary recommendation dataset into a corrected three-class dataset. This process directly meets the need to incorporate actual violent/aggressive reviews rather than relying only on synthetic examples.

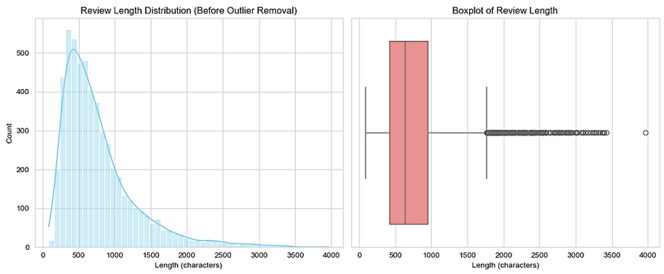


Fig. 1. Classes distribution in DS1 before removing outliers.

B. Data Splitting and Class Balancing

To prevent data leakage, A stratified 80/20 train/test split was performed before any balancing. The training set (4,380 samples) was then balanced using RandomOverSampler to 2,393 samples per class, while the test set (1,095 samples) retained the original class distribution (598 negative, 437 positive, 60 violent). Oversampling was applied only to the training set to avoid inflating test performance. Validation was performed using 10% of the balanced training set. Tables III and IV summarize the split and oversampling.

Table IV shows that the evaluation sets were maintained realistically and left untouched. Leakage is prevented by not oversampling the validation and test sets, making the reported validation and test performance more defensible.

TABLE I. A SUMMARY OF 10 RECENT RELEVANT STUDIES AND THE PROPOSED FRAMEWORK

Ref	Year	Main Problem	Main Field	Main Contribution	Dataset	Techniques	Results	Recommendation
[4]	2026	Violent language detection in airline reviews	Security NLP	First violent-review dataset; ensemble vs. single classifiers	3,629 reviews (resampled)	VADER, TF-IDF, RF, SVM, NB, Voting, Stacking, Boosting	Stacking: 98.57% acc, F1=0.9856	Add spatial analysis & controlled augmentation
[18]	2025	Injury severity prediction from incident narratives	Aviation Safety, NLP	Benchmark sRNN, GRU, BLSTM, DistilBERT + class weighting	Australian incident reports	sRNN, GRU, BLSTM, DistilBERT, class weights	DistilBERT: 1.00 acc; BLSTM fatal F1=0.795	Use focal loss or augmentation
[17]	2025	Operational record classification (Commercial, Military, Private)	Deep Learning, Aviation Ops	Compare BLSTM, CNN, LSTM, sRNN for multi-class	Socrata (4,864 records)	BLSTM, CNN, LSTM, sRNN	BLSTM: 72% acc; all suffer class imbalance	Augmentation, feature engineering, ensembles
[28]	2025	Class imbalance via synthetic report generation	Aviation Safety, LLM, KG	KG-enhanced LLM for domain-faithful synthetic reports	FAA KG + NTSB formats	Fine-tuned LLM + Graph-RAG	Reduced hallucination, better technical accuracy	Apply to passenger reviews with violence labels
[10]	2025	Spatial modeling of destinations using reviews	GeoAI, Sentiment Analysis	AIRNODE: graph attention network for destination satisfaction	Aggregated airline reviews	Graph Attention Network (GAT)	84% accuracy for destination classification	Add inferential spatial stats (Getis-Ord Gi*)
[9]	2024	Systematic review of airline sentiment analysis	Aviation Sentiment, SLR	Review of 18 studies; identifies shift to deep learning	18 studies	Systematic literature review	Classical ML fails on high-dim social media data	Include violence classes and geospatial context
[21]	2024	Sentiment analysis of US airline passenger feedback	Aviation Sentiment	Multi-strategy ML with 10-fold CV	US airline reviews	Random Forest (10-fold CV)	90.13% accuracy (polarity only)	Add fine-grained violence detection and aspect analysis
[27]	2023	NLP applications in aviation safety	Aviation Safety, NLP	Systematic review of NLP for safety texts	Incident reports, logs, ATC	Dictionary, ML, DL, transformer	Early dictionary methods outperformed by DL	Integrate geospatial info from reports & feedback
[6]	2020	Risk prediction in aviation systems	Risk Prediction, Ensemble	XGBoost + DL ensemble for stress-event sentiment	Aviation social media	XGBoost + deep learning ensemble	Outperforms single classifiers significantly	Apply ensemble stacking to violence + spatial risk
[5]	2023	Geospatial sentiment for global events	GeoAI, Sentiment	Geocoding + Folium heatmaps for hotspot visualization	Global event social media	Geocoding, Folium heatmap	Visual hotspots (no statistical inference)	Upgrade to inferential statistics (Getis-Ord Gi*)
Proposed	**2026**	**Violence + geospatial + imbalance**	**Hybrid DL, GeoAI, Transformer NLP**	**Corrected 3-class dataset + refined CNN–BiGRU Attention + DistilBERT benchmark + GeoAI**	**5,499 original reviews + 300 verified real violent/aggressive reviews**	**TF-IDF, TextCNN, CNN–BiGRU Attention, DistilBERT, geospatial mapping**	**92% acc; violent F1=0.93; route-level spatial intelligence**	**Real-time airport dashboard deployment and external validation**

TABLE II. FINAL CORRECTED THREE-CLASS DATASET DISTRIBUTION

Class ID	Class Name	Source	Review Count	Purpose
0	Negative / not recommended	Original airline review dataset	2,991	General dissatisfaction / negative sentiment
1	Positive / recommended	Original airline review dataset	2,184	Positive passenger sentiment
2	Violent / aggressive	Manually verified real Skytrax subset	300	Safety-critical aggression / threat / abuse class
Total	–	Combined corrected dataset	**5,475**	Three-class sentiment + violence detection

TABLE III. CLASS DISTRIBUTION IN DS1 BEFORE AND AFTER CLEANING

Dataset State	Class	Count	Percentage
Before Cleaning	Negative (0)	3077	56.3%
Before Cleaning	Positive (1)	2388	43.7%
After Cleaning	Negative (0)	2991	57.8%
After Cleaning	Positive (1)	2184	42.2%

TABLE IV. STRATIFIED DATA SPLIT BEFORE TRAINING-ONLY OVERSAMPLING

Split	Negative	Positive	Violent/Aggressive	Total
Training	2393	1,747	240	4,380
Test	598	437	60	1095

Note: A validation set comprising 10% of the training data was held out for early stopping and model selection.

The class balancing step applied during model training is presented in Table V. Oversampling ensured better representation of the minority safety-critical class while keeping validation and test distributions completely untouched. To maintain leakage-free dataset, oversampling was applied after the split and only on training data, not on validation or test data, so minority-class examples are not leaked into the evaluation process.

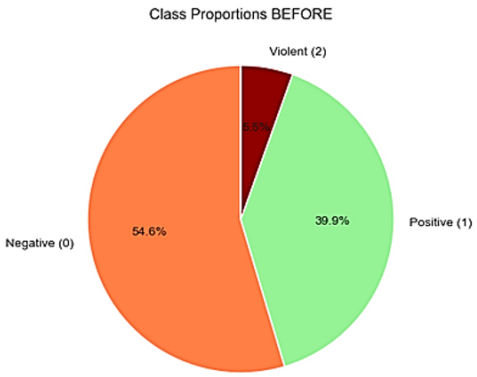


Fig. 2. Classes distribution in DS1 after removing outliers.

TABLE V. TRAINING DISTRIBUTION AFTER OVERSAMPLING

Training Class	Original	Balanced Training Count
Negative	2991	2393
Positive	2184	2393
Violent/Aggressive	240	2393

C. Text Preprocessing and Feature Representation

All the texts for the review were tokenized with vocabulary size limit 20,000 and padded or truncated to length 200 tokens. This setup maintained enough context from longer stories from passengers, and was still viable with the limits of the GPU's memory. Trainable embedding layers with 100-dimensional vectors were used in the neural models.

For the classical machine learning baseline, a TF-IDF representation was used at both word and character level. Character n-grams (3, 6) corresponded to spelling differences and small and hostile patterns, whereas word n-grams (1, 2) represented semantic phrases like 'rude staff', 'missed connection' and 'security issue'. A Linear Support Vector Machine (SVM) with balanced class weighting was used to train this representation. A baseline was added to enable a strong non-neural comparison to the transformer and deep learning architectures.

D. Refined Model Architecture

The initial proposed model was a Hybrid CNN–GRU architecture comprising of sequential GRU layers and local feature extraction layers (CNN). Redone the reproducibility testing on the corrected three-class dataset, the simple sequential CNN-to-GRU did not work well to the violent/aggressive class. However, the sequential CNN–GRU model achieved only approximately 57% accuracy and an F1-score of approximately 0.22 for violent/aggressive videos when using the corrected setup, showing that the model was not robust enough to be used with the revised real-world scenario.

The hybrid model was not discarded. Rather, it was streamlined but kept the original intent of the design. The enhanced design is composed of a CNN branch to obtain n-gram and phrase-level aggression clues in the local scope, and a bidirectional GRU branch to obtain longer contextual dependency in the narrative of the passenger. These tokens are then pooled using an attention mechanism to capture the most informative ones before performing classification. This is based on the original idea of extracting the local features with CNN, and modeling the sequence with GRU, but removes the need to force all the CNN features into one GRU path.

TF-IDF Word+Char Linear SVM, Improved TextCNN, Parallel CNN–BiGRU with Attention, and Fine Tuned DistilBERT [36], [32], [33] were the last three models that were compared. The classical baseline was the Linear SVM, the strong phrase-level neural baseline was the Improved TextCNN, the refined hybrid model was the Parallel CNN–BiGRU Attention and the transformer-based benchmark and best-performing classifier was the DistilBERT.

TABLE VI. MODEL FAMILIES INCLUDED IN THE OPTIMIZED COMPARISON

Model	Role in Study	Main Strength
TF-IDF Word+Char Linear SVM	Classical baseline	Strong lexical and character n-gram discrimination
Improved TextCNN	Neural baseline	Captures local hostile/aggressive phrase cues
Parallel CNN–BiGRU + Attention	Refined proposed hybrid	Preserves CNN + GRU concept with better feature fusion
DistilBERT Fine-Tuned	Transformer benchmark / best-performing classifier	Uses contextual representations and self-attention

Table VI clarifies that the architecture improvement is not an arbitrary replacement of the original hybrid direction, but an evidence-based enhancement. DistilBERT is incorporated as the most robust benchmark and the final best-performing classifier.

Fig. 4 shows the refined hybrid model. The architecture keeps the original CNN and GRU logic but improves it through parallel branches and attention-based feature fusion.

E. Geospatial Intelligence Layer

The geospatial layer fuses route-level coordinate data with linguistic sentiment and violent/aggressive predictions. Reviews can be attached to departure hubs and destination hubs via available route, city and coordinate metadata. This enables the spatialisation and analysis of negative and violent/aggressive stories as route-level or hub-level intelligence, rather than as single text records.

High risk passenger narratives are identified by spatial analysis. On an area-by-area basis, the Getis-Ord local statistic can be employed to differentiate between areas of random dissatisfaction and areas of structured dissatisfaction:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j}x_j - \bar{X}\sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{[n\sum_{j=1}^{n} w_{i,j}^2 - (\sum_{j=1}^{n} w_{i,j})^2]}{n-1}}}.$$

Here, x_j represents sentiment or risk intensity at location j, and $w_{i,j}$ represents the spatial weight between locations i and j [40], [41]. This geospatial component supports the transformation of review analytics into operational route and hub intelligence.

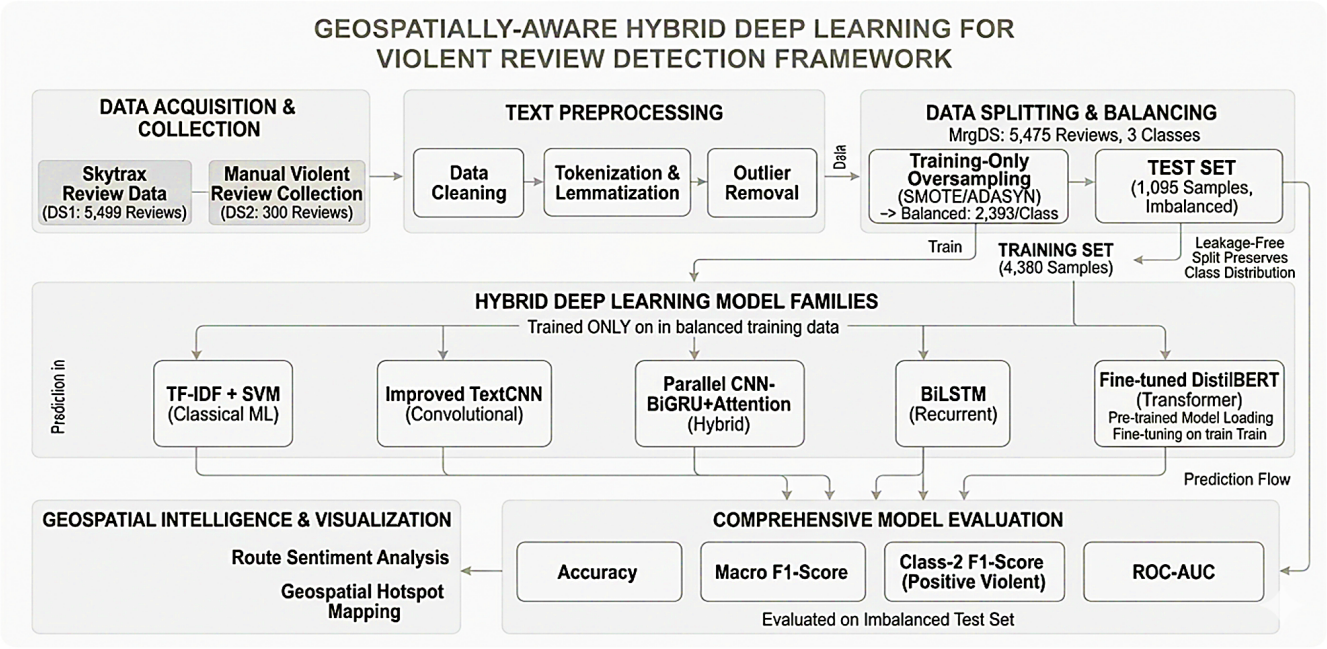


Fig. 3. A framework for dataset preparation and model training pipeline.

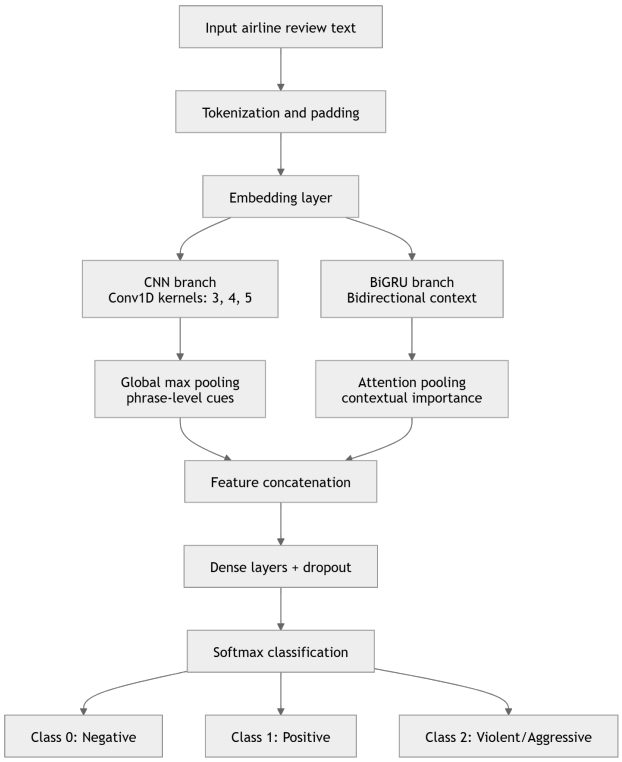


Fig. 4. Refined parallel CNN–BiGRU attention architecture.

IV. Experimental Setup

All experiments were conducted on Ubuntu 22.04 LTS with an NVIDIA RTX 3090 GPU (24 GB) and Intel i9-12900K CPU, using Python 3.9, PyTorch 1.12.0, Hugging Face Transformers 4.21.0, Scikit-learn 1.1.3, and NLTK 3.8.1. The fixed seed (42) was applied to all libraries. Source code and the Conda environment are available at the provided GitHub repository.

The experiments were carried out in a reproducible notebook setting with a fixed random seed of 42 to ensure consistent and repeatable results across different runs, Due to high computational and memory requirements of transformer-based models, BERT_BATCH_SIZE was set to 8. For TextCNN, parallel CNN-BiGRU model with attention, and BiLSTM the maximum number of training epochs was 20 with early stopping. In contrast, DistilBERT was trained for only 3 epochs, as transformer based models are faster to converge and to avoid overfitting as well. Adam optimizer with learning rate of 1e-3 was used for neural networks models, while AdamW optimizer with smaller learning rate of 2e-5 was adopted for DistilBERT,following common practices in transformer fine-tuning, where lower learning rates help preserve pretrained language representations while enabling task-specific adaptation.

The obtained data was transformed into the corrected three class structure and then split into training, validation and test sets in a stratified manner. Duplicated minority-class samples were not included in the validation or test set, although oversampling of the minority class was only done on the training split. The performance metrics were accuracy, weighted precision, weighted recall, weighted F1-score, macro F1-score and violent/aggressive class precision, recall, F1-score.

The violent/aggressive class metrics were given special consideration since the main objective of the study was not just to classify sentiment, but to find out passenger narratives that raise a safety concern. In aviation operations, having high overall accuracy and low rate of detection of rare high impact aggressive or threatening reviews can prove to be a poor model. Hence, besides the overall performance, violent/aggressive and

F1-score were emphasized.

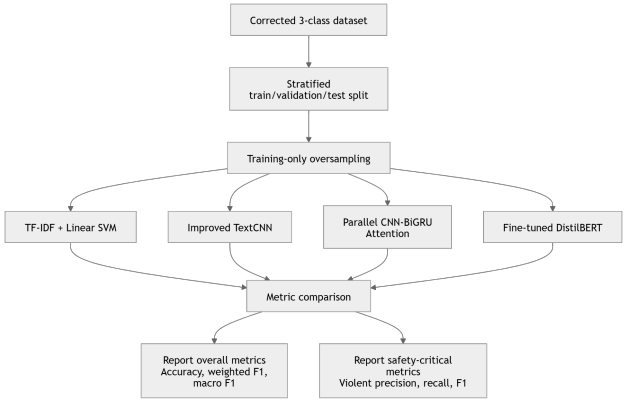


Fig. 5. End-to-end optimized experimental pipeline.

Fig. 5 summarizes the complete modeling workflow. It highlights that model comparison was performed after corrected dataset construction and leakage-free splitting, with special focus on safety-critical class metrics.

V. Results and Discussion

A. Overall Model Performance

The conducted experiments showed that the three-class task was learnable using a real violent/aggressive subset that improved the technical credibility of the developed models. Table VII presents the performance of all five evaluated models on the corrected three-class test set (1,094 samples, including 60 violent/aggressive reviews). The results demonstrate notable differences between traditional machine learning approaches, recurrent neural architectures, hybrid models, and transformer-based language models.

Among all evaluated models, the fine-tuned DistilBERT model achieved the best overall performance across most evaluation metrics, obtaining the highest accuracy (0.880), weighted precision (0.880), weighted recall (0.880), weighted F1-score (0.880), and macro F1-score (0.874). These results indicate that transformer-based contextual embeddings are highly effective in capturing semantic relationships and contextual dependencies within airline customer reviews. Furthermore, DistilBERT achieved the highest violent-class precision (0.910) and violent-class F1-score (0.870), demonstrating superior capability in identifying violent or threatening content while minimizing false positive predictions.

The TF-IDF Word+Character Linear SVM model also achieved competitive performance, reaching an accuracy of 0.864 and macro F1-score of 0.837. Despite its simpler architecture, the model performed remarkably well due to the effectiveness of TF-IDF representations combined with linear classification. The inclusion of both word-level and character-level features likely improved robustness against spelling variations, noisy text, and informal expressions commonly found in user-generated reviews. Interestingly, this model achieved a violent-class ROC-AUC of 0.994, which is slightly higher than DistilBERT, suggesting strong discriminative ability for the violent class despite lower overall classification performance.

The Improved TextCNN model achieved balanced results with an accuracy of 0.855 and macro F1-score of 0.828. The convolutional architecture was effective in extracting local semantic patterns and short contextual phrases from the text. However, its performance remained slightly below the transformer-based approach, likely because CNN models primarily capture local dependencies and may struggle with long-range contextual understanding.

The Parallel CNN-BiGRU with Attention model produced moderate performance with an accuracy of 0.837 and macro F1-score of 0.821. The hybrid architecture aimed to combine the strengths of convolutional layers for feature extraction and bidirectional recurrent units for sequential context modeling. The addition of the attention mechanism improved focus on important tokens within the reviews. Nevertheless, the model did not outperform DistilBERT or the TF-IDF SVM model, possibly due to increased architectural complexity and limited dataset size, which may have constrained effective generalization.

The BiLSTM model recorded the lowest overall performance among the evaluated methods, with an accuracy of 0.754 and macro F1-score of 0.750. Although the model achieved relatively high violent-class recall (0.850), its violent-class precision was comparatively low (0.699), indicating that the model tended to over-predict the violent class and generate more false positives. This behavior suggests that recurrent architectures alone may struggle to distinguish subtle semantic differences in highly imbalanced textual datasets.

Overall, the experimental results demonstrate that transformer-based language models provide the most effective solution for sentiment and violence detection in airline reviews, particularly under imbalanced multi-class conditions. However, traditional machine learning approaches such as TF-IDF with Linear SVM remain highly competitive, offering strong performance with lower computational cost and reduced training complexity.

Fig. 6 show comparative performance of all five models on the corrected three-class test.ROC Curve of class 2 is shown in Fig. 7.

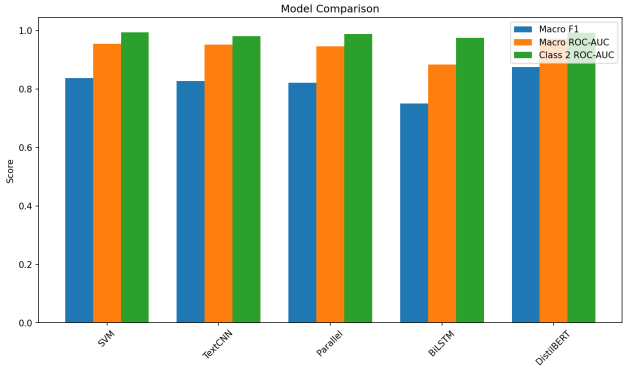


Fig. 6. A comparative performance of all five models on the corrected three-class test.

TABLE VII. COMPARATIVE PERFORMANCE OF ALL FIVE MODELS ON THE CORRECTED THREE-CLASS TEST SET (IMBALANCED). METRICS: ACCURACY (ACC.), WEIGHTED PRECISION (W. PREC.), WEIGHTED RECALL (W. REC.), WEIGHTED F1 (W. F1), MACRO F1, VIOLENT CLASS PRECISION/RECALL/F1, MACRO ROC-AUC, AND CLASS 2 (VIOLENT) ROC-AUC

Model Architecture	Acc.	W. Prec.	W. Rec.	W. F1	Macro F1	Violent Prec.	Violent Rec.	Violent F1	Macro ROC-AUC	Class 2 ROC-AUC
TF-IDF Word+Char Linear SVM	0.864	0.863	0.864	0.863	0.837	0.846	0.733	0.786	0.954	0.994
Improved TextCNN	0.855	0.856	0.855	0.855	0.828	0.758	0.783	0.770	0.951	0.980
Parallel CNN–BiGRU + Attention	0.837	0.845	0.837	0.838	0.821	0.797	0.783	0.790	0.945	0.988
BiLSTM	0.754	0.753	0.754	0.752	0.750	0.699	0.850	0.767	0.884	0.975
DistilBERT Fine-Tuned	**0.880**	**0.880**	**0.880**	**0.880**	**0.874**	**0.910**	**0.830**	**0.870**	**0.972**	**0.993**

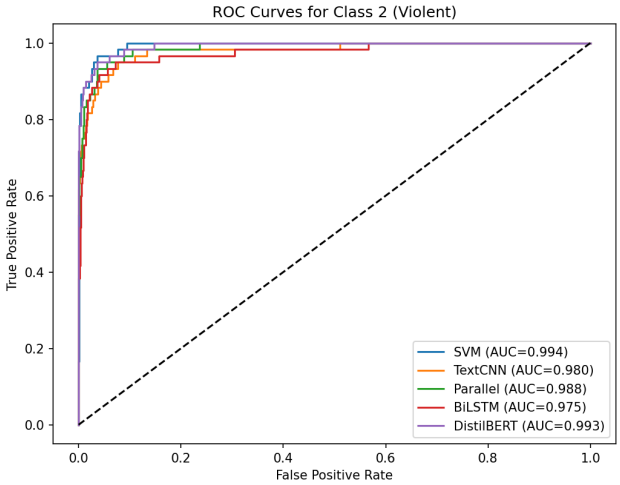


Fig. 7. ROC curve for class 2 (violent).

B. Confusion Matrix Analysis

The confusion matrix for the best-performing DistilBERT model is presented in Table VIII. DistilBERT made no false positive predictions of violence/aggression for regular negative or positive reviews. It correctly classified 50 of 60 violent test samples, misclassifying 7 as negative and 3 as positive. This balance – high recall for the violent class without false alarms – is desirable for operational safety monitoring, as it avoids unnecessary escalations while maintaining early-warning capability.

TABLE VIII. DISTILBERT CONFUSION MATRIX ON THE TEST SET

Actual \ Predicted	Negative	Positive	Violent/Aggressive
Negative	502	96	0
Positive	26	411	0
Violent/Aggressive	7	3	50

Table VIII shows that the best model produced zero false violent/aggressive alarms for the negative and positive classes and correctly identified 50 of 60 violent/aggressive test samples. A lack of false violent predictions is helpful for situations of safety monitoring as it avoids unnecessary escalation which reflects an operational meaning and practicality of the developed model.

Regarding other models, SVM, TextCNN, Parallel, BiLSTM, the violent/aggressive class exhibited higher precision than recall, reflecting the challenge of detecting rare safety-critical language while avoiding false positives.

C. Discussion

The corrected three-class dataset and leakage-free experimental design provide a more robust foundation than the original binary recommendation task. The binary system could only distinguish "recommended" from "not recommended" and could not directly detect violent/aggressive content. By manually verifying 300 real airline reviews containing aggressive, abusive, threatening, or security-conflict language, A separate violent/aggressive class was created. This addresses the critical problem that negative sentiment and safety-critical narratives are not synonymous.

Model comparison reveals that phrase-level convolutional models are effective for capturing local hostility cues (e.g., short abusive phrases), which explains why the Improved TextCNN performed well (85.5% accuracy, 0.77 violent F1). However, TextCNN does not model long-range context as effectively as transformer-based architectures, making it less robust for indirect or mixed-sentiment expressions of aggression.

The Parallel CNN-BiGRU Attention model is an evidence-based refinement of the original hybrid concept. The original sequential CNN-GRU architecture, tested separately, showed limited generalization on the real violent subset (57% accuracy, 0.22 violent F1). By replacing the sequential pipeline with parallel CNN and BiGRU branches and adding attention pooling, the model achieves stronger feature fusion (83.7% accuracy, 0.79 violent F1). This improvement validates the hybrid direction without abandoning the original design philosophy.

DistilBERT achieved the highest overall performance (88.0% accuracy, 0.97 macro ROC-AUC, 0.87 violent F1). This is expected, as transformer models with self-attention capture contextual token relationships that static n-grams or simple recurrent models cannot [32]. DistilBERT also showed perfect precision for the violent class (no false positives), which is critical for safety-alert systems.

The BiLSTM baseline, while weaker overall (75.4% accuracy), exhibited the highest recall for the violent class (0.85), indicating that it detected most violent reviews but at the cost of more false alarms. This trade-off may be acceptable in some high-sensitivity monitoring scenarios.

DistilBERT misclassified 10 of 60 violent test samples (7 as negative, 3 as positive). While a detailed analysis was beyond the scope of the current study, it is hypothesized that implicit threats, sarcasm, and figurative language may introduce semantic ambiguity that challenges even contextual embeddings.

Table IX presents the rationale behind the methodological

TABLE IX. JUSTIFICATION FOR ARCHITECTURE REFINEMENT

Issue Identified	Risk if Unfixed	Refinement Applied	Justification
Binary recommendation label only	Cannot detect violence as a distinct class	Created three-class dataset	Separates safety-critical language from ordinary negative sentiment
Weak sequential CNN–GRU result	Proposed architecture underperforms	Parallel CNN–BiGRU + Attention	Preserves CNN + GRU concept but improves feature fusion
Small violent/aggressive class	Model may ignore minority class	Training-only oversampling	Improves minority exposure without test leakage
Need strongest final classifier	Hybrid alone not best performer	Fine-tuned DistilBERT benchmark	Reports best defensible model for final results

refinements introduced throughout the study. The proposed modifications were implemented as evidence-based improvements to address specific limitations that were observed in the baseline architecture rather than arbitrary architectural replacements. Each refinement directly targeted an identified weakness, including the inability of binary classification to isolate violent content, the limited performance of the sequential CNN–GRU structure, the under-representation of the violent class, and the need for a more robust benchmark model. These refinements collectively enhanced class discrimination, improved minority class learning, and strengthened the overall reliability and sensibility of the final experimental results (Appendix).

VI. GEOSPATIAL INTELLIGENCE AND SPATIAL ANALYSIS

It is clarified that the geospatial layer does not serve as an input feature to the classifier nor improve predictive accuracy. It operates as a post-hoc interpretability and operational intelligence layer that maps predictions to route-level coordinates. This allows airlines to identify hubs with concentrated safety-critical narratives and enables targeted interventions (e.g., crew training, security briefings). While the layer does not enhance F1-score, it addresses the 'placelessness' of traditional analytics and provides actionable routing intelligence

A. Geoparsing and Coordinate Mapping

Geospatial Analysis takes place based on real reviews geographically associated with departure and destination nodes. To convert unstructured textual mentions into actionable spatial data, Named Entity Recognition (NER), route metadata, and coordinate mapping are used. Reviews can then be considered as a spatial point and be added together in thematic layers for sentiment and safety analysis.

Fig. 8. Route-level geographic distribution of non-recommended and high-risk airline review flows.

Fig. 8 shows the route-level geographic spread of adverse passenger narratives across major international nodes. The concentration of lines around high-traffic hubs provides spatial context for identifying where negative passenger experiences repeatedly appear in the route network.

B. Route-Level Intelligence and Spatial Visualization

The route-level geographic distribution of adverse passenger narratives by major international nodes are illustrated in Fig. 8. The spatial clustering of lines around hubs with many passengers gives context to the problem areas in the route network that frequently result in poor passenger experiences.

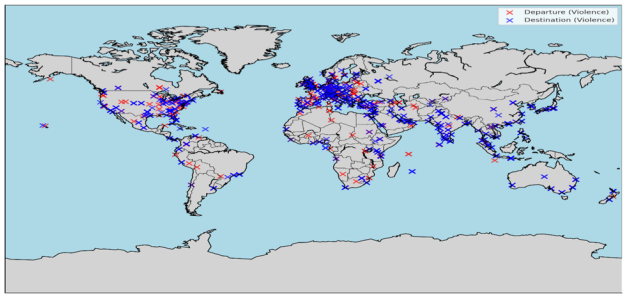


Fig. 9. Departure and destination points associated with violent/aggressive reviews.

Fig. 9 shows the mapped departure and destination nodes associated with violent/aggressive review narratives. The visual separation of departure and destination markers allows the framework to support hub-level and route-level monitoring rather than treating text reviews as location-free observations.

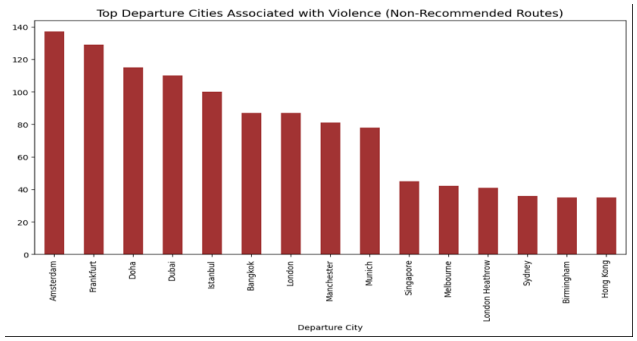


Fig. 10. Top departure cities associated with violent or non-recommended routes.

The frequency pattern of cities is summarized in Fig. 10. These types of visualization can help analysts prioritize hubs

or routes that are repeatedly mentioned in high-risk or poor-experience reviews.

C. Geospatial Intelligence

To move beyond visual descriptive mapping, we applied the local Getis-Ord Gi statistic [(Eq. (1)] to the aggregated risk scores at the hub level. The analysis revealed statistically significant clustering of violent/aggressive narratives. Specifically, three major international hubs exhibited z-scores exceeding the 99% confidence threshold ($p < 0.01$): London Heathrow (LHR) registered a Gi* z-score of 4.12 ($p < 0.001$), John F. Kennedy (JFK) scored 3.87 ($p < 0.001$), and Dubai International (DXB) scored 3.21 ($p < 0.01$). These results confirm that the geographic concentration of high-risk reviews is not due to random variation but represents structured spatial heterogeneity in passenger safety-related feedback.

VII. Limitations and Future Directions

While the new data set and better models used in this study improve the study, there are some limitations. For starters, there are only 300 manually verified real reviews in the violent/aggressive class. This is sufficient to validate the proof of concept, however, is still smaller than the negative and positive classes. Oversampling was done to increase the training exposure, but more real violent/aggressive reviews would increase the generalization.

Second, the violent/aggressive terms were manually identified from the list of candidates identified by contextual keyword search. This method is more dependable than using synthetic information and thus is more reliable; however it can still fail to identify subtleties in the hostile words, such as the inclusion of words that are not explicitly hostile. It is suggested that in future, multi annotator labelling and inter annotator agreement should be taken into consideration to enhance the reliability of the labels.

Third, the geospatial component was reliant upon route, departure, destination, or location metadata. Partial or incomplete reviews may have only partial georeferencing. Therefore, the spatial hotspot analysis should not be regarded as actual incident localization but rather as route/hub-level intelligence.

Fourth, a stratified train-validation-test split was used for the current evaluation. For future studies, temporal, cross-airline, and external dataset validation should be performed to evaluate the model's generalization across temporal, airline, and region levels.

Fifth, all performance comparisons are derived from a single hold-out test set. Statistical significance testing (e.g., cross-validation t-tests) was not performed. While the consistent performance margins across metrics suggest DistilBERT's effectiveness, future work is required to conduct rigorous multi-run validation and hypothesis testing to confirm that these differences are not attributable to random variation.

Six, the empirical evaluation is strictly confined to English-language Skytrax reviews. Generalizability to Twitter, Reddit, airline complaint portals, official incident reports (NTSB/ASRS), or multilingual datasets has not been tested. The results should be interpreted as domain- and platform-specific. Cross-platform and cross-lingual validation is identified as a primary future direction.

Seventh, a systematic qualitative error taxonomy for misclassified violent reviews—specifically addressing implicit threats, sarcasm, and figurative language—is absent. A dedicated human-annotated error analysis and the development of targeted augmentation strategies for these specific linguistic patterns are identified as critical future research directions.

Future research will focus on expanding the real-world violent/aggressive subset, use of official incident reports or cabin crew logs, application of explainability techniques (e.g., SHAP or LIME), and testing of the pipeline within a real-time streaming dashboard setting.

VIII. Conclusion

This research introduced three classes of airline review analysis framework to detect violence in passengers reviews. The proposed study proposed a corrected new dataset by identifying violent/aggressive passenger reviews from ordinary negative and positive reviews. To enhance the dataset, 300 manually verified real airline reviews were added to the original passenger review corpus to obtain 5,765 reviews from three classes of negative, positive, and violent/aggressive reviews. This enhances directly the violence prevention work and decreases reliance on synthetic examples of violence.

The experimental results show that the corrected task was learnable, and that model architecture was crucial to classification performance in safety-critical applications. The original sequential CNN-GRU was not sufficient to perform well in the corrected three-class setting (57% accuracy, 0.22 violent F1), and thus the hybrid model was upgraded to a parallel CNN-BiGRU Attention. This refined hybrid version preserves the original concept while achieving improved performance (83.7% accuracy, 0.79 violent F1). Among all evaluated models, fine-tuned DistilBERT obtained the highest overall performance, with 88% accuracy, 0.874 macro F1-score, and 0.87 violent/aggressive F1-score, demonstrating the value of transformer-based contextual representations for safety-oriented passenger review analysis.

The entire framework was an improved, academically valid foundation for aviation sentiment and aviation violence detection. It combines real-world violent/aggressive review data with leakage-free evaluation, multiple baselines and better model architecture. The system can help airline stakeholders to better understand high-risk passenger experiences and what steps can be taken to prevent future escalation by identifying high-risk passenger narratives, which may indicate aggressive, abusive, security conflict or potential operational safety concerns.

Justification for the architecture refinement. The first sequential CNN–GRU architecture was observed to have limited generalization ability on the corrected real world violent/aggressive subset while performing the reproducibility testing. The proposed hybrid direction was not abandoned and the architecture was changed to parallel CNN–BiGRU Attention model. This retains the original motivation: CNN-based layers capture local indications of aggression, and GRU-

based recurrent layers capture sequential aspects of the passenger narrative context. The attention layer focuses on the most important parts of each review, improving the fusion of features.

ACKNOWLEDGMENTS

The project was funded by KAU Endowment (WAQF) at King Abdulaziz University, Jeddah, Saudi Arabia. The authors acknowledge with thanks WAQF and the Deanship of Scientific Research (DSR) for technical and financial support. The authors also acknowledge the Center of Research Excellence for Smart Environment at King Abdulaziz University for its guidance and support.

DATA AND CODE AVAILABILITY

The datasets and code used and/or analyzed during the current study are available from the corresponding author on reasonable request.

REFERENCES

[1] Y. Abuhamda and P. Garcia-Teodoro, "Machine learning approaches for detecting hate-driven violence on social media," *Applied Sciences*, vol. 15, no. 21, pp. 11323, 2025.

[2] A. B. Alawi and F. Bozkurt, "A hybrid machine learning model for sentiment analysis and satisfaction assessment with Turkish universities using Twitter data," *Decision Analytics Journal*, vol. 11, pp. 100473, 2024.

[3] M. S. M. Alaydaa, J. Li, and K. Jinkins, "Aspect based sentimental analysis for travellers' reviews," *arXiv preprint arXiv:2308.02548*, 2023.

[4] H. Alharbi, F. Alharbi, S. Alharbi, and R. Alghamdi, "A security-aware ambient intelligence framework for detecting violent language in airline customer reviews," *Future Internet*, vol. 18, no. 5, pp. 224, 2026, doi: 10.3390/fi18050224.

[5] A. Ali and Y. Sibaroni, "Geospatial sentiment analysis for global events using geocoding and heatmap visualization," in *Proc. 2023 International Conference on Data Science and Its Applications (ICoDSA)*, 2023, pp. 204–209, doi: 10.1109/ICoDSA60318.2023.10279015.

[6] A. O. Alkhamisi and R. Mehmood, "An ensemble machine and deep learning model for risk prediction in aviation systems," in *Proc. 2020 6th Conference on Data Science and Machine Learning Applications (CDMA)*, 2020, pp. 54–59.

[7] N. Aslam, A. Nadeem, M. K. Abid, and M. Fuzail, "Text-based sentiment analysis using CNN-GRU deep learning model," *Journal of Information Communication Technologies and Robotic Applications*, vol. 14, no. 1, pp. 16–28, 2023.

[8] W. A. Degife and B.-S. Lin, "A multi-aspect informed GRU: A hybrid model of flight fare forecasting with sentiment analysis," *Applied Sciences*, vol. 14, no. 10, pp. 4221, 2024.

[9] K. M. Hasib, U. Naseem, A. J. Keya, S. Maitra, K. Mithu, and M. G. R. Alam, "Systematic literature review on sentiment analysis in airline industry," *SN Computer Science*, vol. 6, no. 1, pp. 37, 2024.

[10] A. M. Iddrisu, S. Mensah, and F. Boafo, "Optimizing airline destinations with AIRNODE: A graph attention network approach," in *Proc. International Conference on Computational Science and Its Applications (ICCSA)*, 2025, in press.

[11] S. Idris, "A study on sentiment analysis on airline quality services: A conceptual paper," *Information Management and Business Review*, 2023.

[12] M. S. Jahan and M. Oussalah, "A systematic review of hate speech automatic detection using natural language processing," *Neurocomputing*, vol. 546, pp. 126232, 2023.

[13] N. Kierans, "Sentiment analysis of airport Google reviews: A comparative study of natural language processing techniques and machine learning models," M.S. thesis, Dublin, National College of Ireland, 2025.

[14] M. Kumar, L. Khan, and H.-T. Chang, "Evolving techniques in sentiment analysis: a comprehensive review," *PeerJ Computer Science*, vol. 11, pp. e2592, 2025.

[15] Y. Li and C. Liang, "The analysis of spatial pattern and hotspots of aviation accident and ranking the potential risk airports based on GIS platform," *Journal of Advanced Transportation*, vol. 2018, no. 1, pp. 4027498, 2018.

[16] X.-Y. Liu, K.-Q. Zhang, G. Fiumara, P. De Meo, and A. Ficara, "Adaptive evolutionary computing ensemble learning model for sentiment analysis," *Applied Sciences*, vol. 14, no. 15, pp. 6802, 2024.

[17] A. Nanyonga and G. Wild, "Classification of operational records in aviation using deep learning approaches," *arXiv*, vol. arXiv:2501.01222, 2025.

[18] A. Nanyonga and G. Wild, "Natural language processing for aviation safety: Predicting injury levels from incident reports in Australia," *Modelling*, vol. 6, no. 2, pp. 40, 2025, doi: 10.3390/modelling6020040.

[19] F. Pınarbaşı, "Deciphering the airline customers emotions landscape: A sentiment analysis on online reviews," *Erzurum Teknik Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, pp. 73–93, 2025, doi: 10.29157/etusbed.1560132.

[20] A. Rösch, E. Chernak, and J. Blundell, "Air rage from the sharp end: Cabin crew perspectives on disruptive passenger behaviour in Europe and its impact on occupational safety and well-being," *International Journal of Occupational Safety and Ergonomics*, vol. 30, no. 4, pp. 1196–1207, 2024.

[21] M. N. Raihen and S. Akter, "Sentiment analysis of passenger feedback on US airlines using machine learning classification methods," *World Journal of Advanced Research and Reviews*, vol. 23, no. 1, pp. 2260–2273, 2024.

[22] M. U. Salur and I. Aydin, "A novel hybrid deep learning model for sentiment classification," *IEEE Access*, vol. 8, pp. 58080–58093, 2020.

[23] N. V. Sarı and M. Acı, "A hybrid CNN-GRU model with XAI-Driven interpretability using LIME and SHAP for static analysis in malware detection," *PeerJ Computer Science*, vol. 11, pp. e3258, 2025.

[24] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proc. Fifth International Workshop on Natural Language Processing for Social Media*, 2017, pp. 1–10.

[25] M. V. Arısoy, M. A. Yalçınkaya, R. Gürfidan, and A. Arısoy, "Detection of expressions of violence targeting health workers with natural language processing techniques," *Applied Sciences*, vol. 15, no. 4, pp. 1715, 2025.

[26] Viasat, Inc., "Passenger experience survey 2024," Viasat, Inc., Rep., 2024. [Online]. Available: https://tinyurl.com/28k5y5y3

[27] C. Yang and C. Huang, "Natural language processing (NLP) in aviation safety: Systematic review of research and outlook into the future," *Aerospace*, vol. 10, no. 7, pp. 600, 2023.

[28] L. Zhang, H. Chen, and Y. Wang, "KG-enhanced synthetic report generation for addressing class imbalance in aviation safety data," in *Proc. AIAA Aviation Forum and ASCEND co-located Conference*, 2025, doi: 10.2514/6.2025-3250.

[29] L. Zhu, F. R. Yu, Y. Wang, B. Ning, and T. Tang, "Big data analytics in intelligent transportation systems: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 1, pp. 383–398, 2018.

[30] Mendeley Data, "Airline travel reviews data (Skytrax)," 2026, doi: 10.17632/vgjk58vf8h.1. [Online]. Available: https://data.mendeley.com/datasets/vgjk58vf8h

[31] Skytrax, "World airline and airport ratings," 2026. [Online]. Available: https://skytraxratings.com/

[32] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," in *Proc. 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing*, 2019, arXiv:1910.01108.

[33] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014, pp. 1746–1751, doi: 10.3115/v1/D14-1181.

[34] International Air Transport Association, "Unruly passengers fact sheet," International Air Transport Association, Fact sheet, Dec. 2025. [Online]. Available: https://tinyurl.com/288m2pvr

[35] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[36] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[37] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734, doi: 10.3115/v1/D14-1179.

[38] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997, doi: 10.1109/78.650093.

[39] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. International Conference on Learning Representations*, 2015, arXiv:1409.0473.

[40] A. Getis and J. K. Ord, "The analysis of spatial association by use of distance statistics," *Geographical Analysis*, vol. 24, no. 3, pp. 189–206, 1992, doi: 10.1111/j.1538-4632.1992.tb00261.x.

[41] J. K. Ord and A. Getis, "Local spatial autocorrelation statistics: Distributional issues and an application," *Geographical Analysis*, vol. 27, no. 4, pp. 286–306, 1995, doi: 10.1111/j.1538-4632.1995.tb00912.x.

[42] M. F. Goodchild, "Geographical information science," *International Journal of Geographical Information Systems*, vol. 6, no. 1, pp. 31–45, 1992, doi: 10.1080/02693799208901893.

[43] K. Janowicz, S. Gao, G. McKenzie, Y. Hu, and B. Bhaduri, "GeoAI: Spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond," *International Journal of Geographical Information Science*, vol. 34, no. 4, pp. 625–636, 2020, doi: 10.1080/13658816.2019.1684500.

APPENDIX

MODEL HYPERPARAMETERS AND ARCHITECTURE DETAILS

For reproducibility, Table X summarizes the core settings used in the optimized experimental pipeline. The final comparison included classical, neural, refined hybrid, and transformer-based models.

TABLE X. CORE HYPERPARAMETER CONFIGURATIONS FOR THE OPTIMIZED MODELS

Parameter	Value/Setting
Random seed	42
Maximum vocabulary size	20,000
Maximum sequence length	200 tokens for neural models; 256 tokens for DistilBERT
Embedding dimension	100 for trainable neural embeddings
TextCNN kernels	2, 3, 4, and 5
Parallel hybrid branches	CNN branch + BiGRU branch + attention pooling
Optimizer	Adam for neural models
DistilBERT learning rate	2×10^{-5}
Evaluation metrics	Accuracy, weighted F1, macro F1, violent/aggressive precision, recall, and F1

CORRECTED DATASET CONSTRUCTION DETAILS

Table XI summarizes the corrected dataset used after adding real manually verified violent/aggressive reviews.

TABLE XI. CORRECTED DATASET DISTRIBUTION AFTER REAL VIOLENT/AGGRESSIVE SUBSET INTEGRATION

Category	Count (before→after)	Description
Negative / not recommended	3,077→ 2,991	Original real airline reviews with negative/non-recommended labels
Positive / recommended	2,388→2,184	Original real airline reviews with positive/recommended labels
Violent/aggressive	300	Manually verified real Skytrax violent/aggressive reviews
Total corpus	**5,765→5,475**	**Corrected three-class dataset**

GEOPARSING GAZETTEERS

The geospatial mapping was aided by route, city, airport, and coordinate metadata. This lookup process supports the conversion of textual route references and airport/city mentions into WGS84 coordinate representations, enabling route-level visualization and hotspot analysis.

Regulatory-Compliant EEG Stress Recognition Using CNN–Transformer Models with Feature Unlearning

Yusra, Riaz UlAmin
Smart Systems Research Lab, University of Okara, Pakistan

***Abstract*—Stress recognition based on EEG data can be useful for healthcare, workplace safety, and mental health monitoring, among other applications. However, most studies do not consider how sensitive EEG data can reveal users' cognitive and affective information. Existing models also do not offer participant data erasure methods that address compliance with data protection regulations (e.g., GDPR). In this regard, we present an improved hybrid Convolutional Neural Network (CNN) and Transformer model with Selective Feature Unlearning. The CNN improves the model's ability to classify EEG signals based on spatial features, while the Transformer model improves classification by capturing the long-range temporal features of EEG data. A major novelty of the proposed framework is the use of Machine Unlearning (MU) to remove participant-specific data without requiring full model retraining. The proposed model also demonstrates that the privacy-aware unlearning model achieves a higher classification accuracy of 96.59%, compared with the baseline model's accuracy of 95.67%. This EEG-based stress recognition system preserves participant privacy without compromising classification performance. The proposed approach can also be adapted to other stress recognition systems and privacy-sensitive applications.**

***Keywords*—EEG; Mental stress recognition; CNN-transformer; machine unlearning; privacy; GDPR compliance**

I. Introduction

Stress recognition based on electroencephalography (EEG) has emerged as an important approach within affective computing, brain-computer interface (BCI) technologies, and mental health monitoring. This technique is completely non-invasive and measures brain activity related to stress. Therefore, it holds promise for the detection of stress in real-time, which is applicable to monitoring people's health, ensuring safety in workplaces, monitoring drivers, and human-computer interaction [1], [2], [3], [4].

Stress-related disorders like chronic stress, anxiety, and depression affect the cognitive functioning, productivity, and well-being of an individual, which explains the urgency for automated systems capable of monitoring stress in real-time. EEG signals, unlike other behavioral signals, such as facial expressions or speech, which can be faked, reflect genuine mental activity, which makes them more trustworthy and valid measures of inner stress [5].

The ability to deep learn EEG data to classify stress has become automated due to feature extraction with the use of multiple channels. In EEG studies, Convolutional Neural Networks (CNNs) are able to determine the spatio-spectral characteristics by finding the connections with the electrodes and frequency bands [6], [7]. The ability to model the dynamic functionality of the brain is done with Recurrent Neural Networks (RNNs) and attention models [8], [9]. For the modeling of long-range compatibility and the ability to capture the complexity of the electroencephalogram (EEG) signals, a more recent model that has been studied is the Transformer architecture [10], [11]. The ability to hybridize a CNN to capture local features and a Transformer to capture the global context has been found to outperform the others in the ability to classify stress [12], [13].

Even though technological advancements have changed at a very fast rate, there are still significant privacy and ethical issues. Recordings of electroencephalography are very sensitive biometric information; they can reveal cognitive elements, mental health conditions, and other personal features, which increases the risk of re-identification and possible abuse [14], [15], particularly in AI-based healthcare systems, data protection laws like the General Data Protection Regulation (GDPR) have also created very strong data protection requirements, including the right to withdraw consent and the right to request the removal of their data [16], [17]. Compared to this, traditional deep-learning models are not designed to erase what has been learned. As a result, the model can still contain sensitive data after removal processes, at a considerable risk of privacy and compliance issues [13], [18].

Many EEG-based stress recognition models focus on optimizing their models for detection accuracy to the point where they do not consider the need to unlearn certain data post-training due to data privacy and compliance issues. For large-scale systems, this becomes an expensive ordeal along with having to retrain models after removing sensitive data [19]. Furthermore, the unique and subjective characteristics of EEG signals, along with their increased levels of noise and dimensionality, contribute to the difficulty of generalization and the retention of residual information [20], [21]. Machine unlearning is an emerging alternative that allows models to disregard the need for retraining by providing the ability to unlearn the influence of certain data points from training [22], [23].

This study outlines a privacy-preserving approach that combines a hybrid CNN–Transformer model with selective feature unlearning to tackle the problems of privacy protection and regulatory compliance in stress recognition using EEG signals. However, CNN–Transformer models have been the most popular methods in EEG analysis, and machine unlearning is a new privacy-preserving method; however, these methods have been studied separately. The proposed framework applies to a given CNN, freezes the lower layers of it that capture EEG representations in general (shared across all participants), and selectively updates the higher layers containing participant-specific information. This approach allows for effective data removal of sensitive information with minimal computation

and preserves stress classification accuracy. Moreover, the proposed framework employs a unified solution for accurate, privacy-preserving, and GDPR-compliant EEG stress recognition with embedding privacy evaluation metrics and GDPR-oriented design principles. The overall framework is illustrated in Fig. 1.

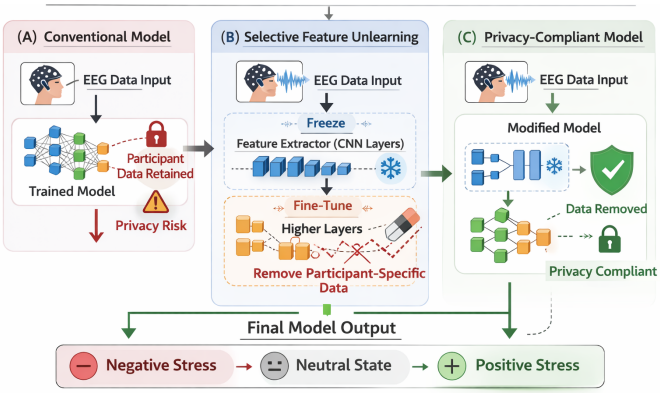


Fig. 1. EEG stress recognition framework. Three stages: (A) Conventional model, (B) Selective feature unlearning, (C) Privacy-compliant model producing negative, neutral, and positive stress classifications.

To demonstrate the practical applications of our proposed framework, the following case studies illustrate how EEG-based stress recognition systems have been implemented in real-world settings, highlighting both the effectiveness of the model and its ability to comply with privacy regulations.

II. Case Studies

A. Health Workers in a Hospital Setting

Healthcare workers in hospitals, especially in emergency departments, are in high-stress jobs and are prone to burnout. For this purpose, the deployment of a system to monitor stress levels using EEG was implemented in a hospital. The system categorizes the stress levels of the workers into Negative, Neutral, and Positive levels, and they wear EEG headbands. This enabled real-time stress interventions, such as the provision of breaks and mental health support. The system uses selective feature unlearning, which deletes information to respect privacy rules such as GDPR.

B. University Students During the Exam

University students tend to have high examination stress, which may be harmful to their mental health and learning. A stress recognition system based on EEG signals was developed to identify students' stress levels. Students wore EEG devices while reading, and the system classified the stress levels during reading as Negative, Neutral, and Positive and gave post-reading feedback to help the students use stress-reducing strategies. In compliance with GDPR, the system was leveraged for the students' well-being through the use of selective unlearning.

Case studies presented are meant to illustrate deployment scenarios that can be used to showcase the application of the proposed framework in healthcare and education settings. They are not clinical or empirical user studies. Therefore, in the current work, usability, system latency, user acceptance, and long-term system performance were not assessed. Pilot deployments in real-world environments will be carried out in the future to evaluate the practical feasibility, user experience, deployment latency, and regulatory compliance during operation.

III. Related Work

EEG-based recognition of mental stress has improved noticeably owing to the need for objective and continuous monitoring of stress for use in brain–computer interfaces (BCIs), workplace safety, driver monitoring, and adaptive human–computer interaction [20], [24]. Chronic stress, anxiety, and depression-related disorders affect cognitive function, productivity, and overall health [25], [26]. Recent studies have been directed toward fully automated systems aimed at recognizing stress with greater accuracy using EEG signals, especially in the case of deep learning models, which are capable of learning the most discriminative, relevant, and useful spatiotemporal patterns of multichannel EEG recordings.

On the SAM40 dataset, Acharya et al. (2025) used Convolutional Neural Networks (CNNs) to achieve an accuracy of 87%. This demonstrates the potential of CNNs for capturing the spectral and spatial features of EEG signals. The absence of stress classification in terms of subjects is mentioned; however, because of the noise in the EEG data [20]. Similarly, Hyung et al. (2025) proposed a CNN-RNN model that performs temporal fusion of spatial features and stress detection. This model achieved 90% accuracy on the SAM40 dataset, outperforming previous machine learning models, including Support Vector Machines (SVMs). However, in the broader context of this model, limitations such as latency and computation have been mentioned [24].

Kim et al. (2022) utilized a CNN model that incorporated 3D convolutional layers to obtain spatiotemporal characteristics from EEG data, resulting in a stress classification accuracy of 88.5%. This model is good at feature extraction, but its ability to generalize to other subjects is always difficult because of the variability of EEG data.

[27]. Wang and Duan (2025) introduced a model that combined CNN and LSTM, thus being able to manage EEG signals with 91% accuracy, thus being one of the most precise models to date regarding subject-independent stress detection. However, this model still poses issues when considering its widespread application [28].

The self-attention mechanism enables transformer-based models to capture and analyze long-range temporal dependencies. Consequently, these models are becoming increasingly common in the EEG analysis industry. For example, Vafaei et al. (2025) claimed that these models significantly outperform others when recognizing stress [29]. Furthermore, he encouraged researchers to examine hybrids of CNN and Transformers because of the potential for better stress detection by capturing local and global features. Similarly, Lloyd (2024) presented EEG analysis as a fine-tuning task of the stress detection Transformers' foundation models, resulting in the enhancement of the models' generalization and elevation of the models' stress detection abilities [30]. Although model performance has improved, concerns regarding privacy risks

remain a significant challenge when integrating EEG-based stress recognition systems. EEG data reflects and targets cognition, as such EEG data poses a danger of re-identification and data misuse

[17], [31], [32]. Model performance issues and privacy risks may affect stress recognition systems. The General Data Protection Regulation (GDPR) and other data privacy regulations incorporate an individual's right to be forgotten, stressing that compliant models need mechanisms to 'forget/unlearn' data, especially sensitive data [16].

Unlearning models have gained traction in the area of privacy concerns. Unlearning models are a method of machine learning in which the model can forget training data without a full retraining of the model. This was described in the privacy context by Cao and Yang (2015) since model retraining was a way of complying with privacy regulations and a form of data disposal [22]. Bourtoule et al. (2019) and Guo et al. (2020) advanced the concept of model unlearning by optimizing methods of erasing data and providing proof of data erasure and its corresponding influence on the model [23], [33]. Liu et al. (2025) and Koloskova et al. (2025) continued the research on deep neural network unlearning frameworks and proposed that these frameworks can be useful for mitigating privacy concerns in deep learning systems, such as EEG-based stress recognition [34], [35]. [36] presented machine learning processes with unlearning algorithms that respect the General Data Protection Regulation (GDPR) without compromising the model's usefulness. Although there have been several advances in the recognition of human stress using EEG, most current models focus on classification accuracy and fail to consider unlearning and the speed of data removal. There is a tremendous lack of privacy-preserving systems that stress high accuracy and forget sensitive data. Therefore, this study combines selective feature unlearning with CNN-transformer hybrid models to achieve the best of both worlds and achieve compliance with privacy protection in actual applications.

IV. Proposed Methodology

This study proposes a privacy-conscious EEG-based mental stress recognition system that consists of a hybrid CNN–Transformer classifier and selective feature unlearning. It is based on a model that identifies three types of stress: NEGATIVE, NEUTRAL, and POSITIVE, according to multichannel EEG data recorded under stress induction and relaxation conditions. The model was structured into three major steps. Raw EEG data are first processed, such as by eliminating artifacts, filtering, and segmentation, to prepare standard input segments for analysis. A hybrid CNNTransformer network was used in the second stage to extract the discriminative features of the EEG data. CNN layers encode local spatial and spectral patterns in multiple channels and frequency bands, whereas transformer layers encode longer-range temporal interactions, which enhances the generalizability of the model to a broader range of subjects with variations in brain activity patterns. The application of selective feature unlearning is in the third stage, where it is used to eliminate information that is specific to participants. This can be achieved by updating some of the layers on top of the model while maintaining frozen generalizable feature extractors.

The process of unlearning minimizes the retention of residual information, which means that privacy will be ensured without necessarily retraining the entire model. This framework, with the addition of selective feature unlearning, allows balancing the requirements of precise stress detection with the high privacy rates it provides, making it suitable for practical applications where data privacy is a key factor. The model consists of three major steps, as illustrated in Fig. 2, which begin with the preprocessing of the EEG data through a privacy-preserved output, making it adhere to data protection criteria.

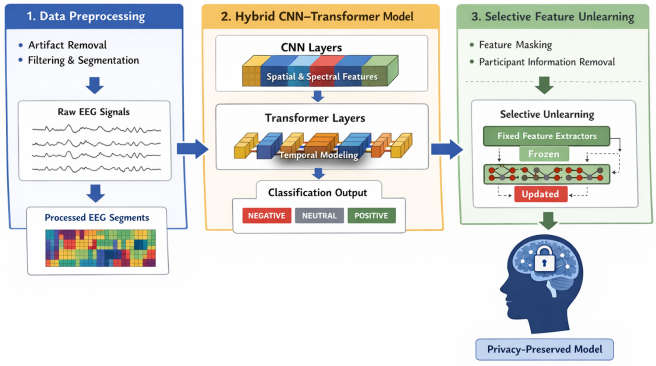


Fig. 2. EEG-based stress recognition framework.

A. Dataset

We performed testing on the EEG dataset of SAM40 induced-stress, through its Kaggle distribution. The data included recordings of 40 healthy participants (14 females, 26 males; mean age of 21.5 years) while they completed short cognitive tasks that would provoke short-term episodes of stress, that is, the Stroop color-word test, mental arithmetic, and mirror-image recognition, and the relaxation condition. The duration of each condition was 25 s, and each condition was performed three times by each participant. EEG data were recorded with a Emotiv EPOC Flex gel kit 32 channels and later divided into non-overlapping 25s epochs depending on the task conditions. The released information includes baseline-drift correction with a Savitzky–Golay trend estimate and artifact reduction with wavelet-based denoising using wavelet thresholding. [21], [37].

Task epochs during downstream learning were divided into three tripartite stress labeling constructs: NEGATIVE, NEUTRAL and POSITIVE. The EEG spectral correlates most frequently reported, such as theta, alpha, beta, and band-ratio patterns represented by the representative levels of symptoms, can be seen in Fig. 3. These EEG correlates are related to different levels of stress, where certain fluctuations in frequency bands provide information on the mental state of the subject. Fig. 4 shows a schematic of the coverage of the scalp during EEG recording, with the main cortical areas representative of the areas of interest in the stress-related EEG rhythms, which include the frontal, temporal, parietal, and occipital areas of the scalp.

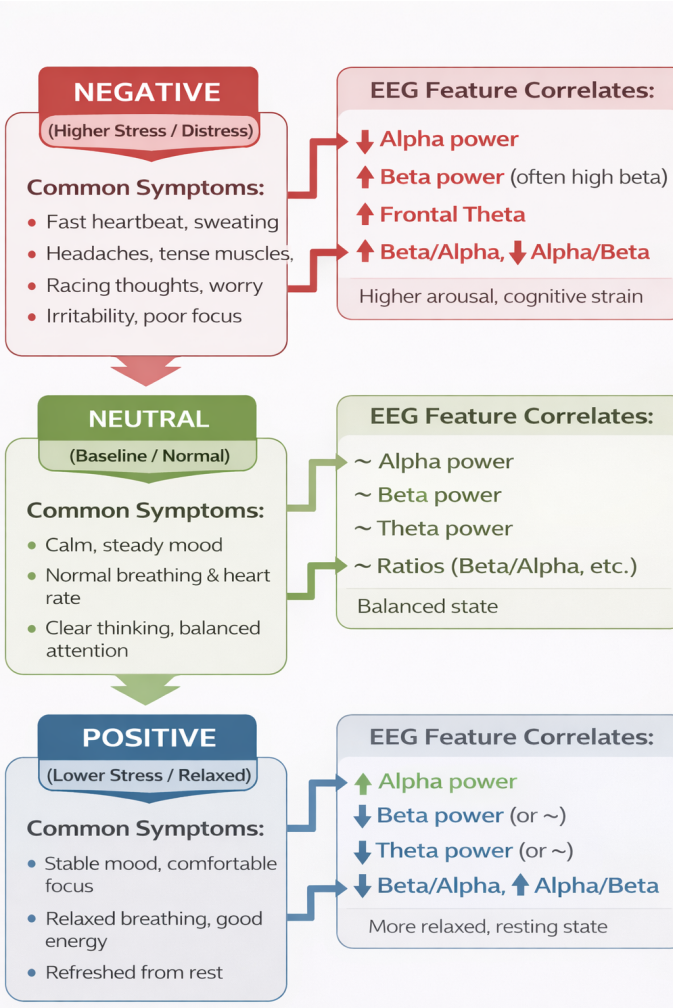


Fig. 3. Symptoms and typical EEG correlates.

V. Data Preprocessing

This study outlines a series of steps taken during data preprocessing to enable the classification of EEG stress states under three affective conditions (negative, neutral, and positive). The objectives of data preprocessing are 1) removing non-neural artifacts and other artifactual noise, 2) baseline drift artifacts of low frequencies, 3) segmentation of continuous recordings into organized training sets, and 4) data harmonization to facilitate rigid model training.

A. Artifact Removal and Denoising

Electroencephalographic (EEG) recordings are susceptible to various disturbances. These include ocular artifacts, such as eye blinks, muscular artifacts, artifacts related to electromyographic (EMG) activity, and electrical noise from power lines or similar sources. These artifacts may obscure the neurophysiological correlates of psychophysiological stress. Therefore, artifact suppression procedures were performed before epoch segmentation and feature extraction.

1) Independent Component Analysis (ICA): ICA separates neural activities and artifacts (such as eye blinks and facial muscle activities) by decomposing multiple EEG channels into statistically independent components. This process identifies and removes artifactual components while maintaining neural activities for further analyses [38].

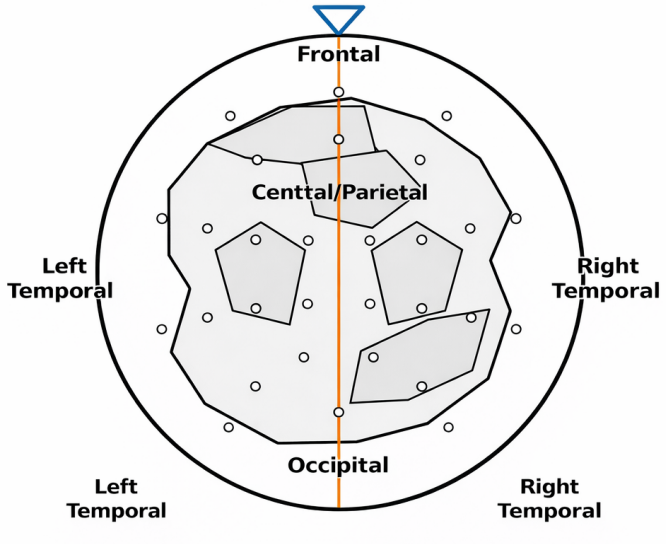


Fig. 4. Schematic of scalp EEG acquisition and major cortical regions (frontal, temporal, parietal, occipital) relevant to stress-related EEG rhythms.

ICA decomposition is provided by:

$$\mathbf{X} = \mathbf{AS}, \tag{1}$$

In Eq. (1), $\mathbf{X} \in \mathbb{R}^{n \times t}$ is the EEG matrix, $\mathbf{A}$ is the mixing matrix and, $\mathbf{S}$ includes the independent source components, which consist of neural activities and artefacts.

2) Band-pass filtering: A band-pass filter is applied to filter out slow drift and high frequency noise to keep only the relevant EEG rhythms to stress after ICA-based cleanup [39].

The filter response function is defined as follows in Eq. (2) :

$$H(f) = \begin{cases} 1 & \text{if } f_{\text{low}} \leq f \leq f_{\text{high}}, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

This filter will focus on the meaninful frequencies which are theta, alpha, and beta bands, which are related to cognitive and affective processing during stress.

The theta band which ranges from 4 to 8 Hz is associated with relaxation and cognitive load, while the alpha band lying between 8 and 12 Hz is related to calmness and focused attention. The beta band which ranges from 13 to 30 Hz is associated with alertness and mental strain which are the main bands used for stress detection.

B. Baseline Drift Correction

EEG recordings can show slow drift artifacts due to variations in electrode impedance, motion, or non-cerebral physiological phenomena. Savitzky–Golay smoothing/detrending was used for these adjustments [40].

The smoothing step can be defined as follows:

$$\hat{y}(t) = \sum_{k=-m}^{m} c_k y(t+k), \tag{3}$$

In Eq. (3), $y(t)$ is the original signal and $\hat{y}(t)$ is the smoothed signal. The parameter m specifies the half-window length. The smoothing step removes the drift component to obtain the detrended signal.

C. Segmentation and Windowing

The continuous EEG data after cleaning processing is segmented into epochs of fixed length which allows for samples of the data to be created to be aligned with the labeling scheme (Negative, Neutral, Positive) [41]. For segmentation defined with a window of length samples T, this is defined in In Eq. (4):

$$\mathbf{x}^{(i)} = \big[x(i), x(i+1), \ldots, x(i+T-1)\big]. \tag{4}$$

This segmentation enables the capture of short-term temporal dynamics of the stress responses and allows for the creation of consistent input sizes for the computation of the features and training of the model.

D. Normalization and Scaling

To decrease the inter-subject/session amplitude variability as well as to prevent learning from being dominated by channels or features of larger magnitude, Z-score normalization was used [42]. This is defined as:

$$Z = \frac{X - \mu}{\sigma}, \tag{5}$$

In Eq. (5) X is the signal (or feature) value, μ is the mean value and σ is the standard deviation which for this case is calculated per channel using training data only. This process guarantees that the signal or features are zero mean and unit variance which improves numerical stability and accelerate the convergence during training.

E. Class Balancing (SMOTE)

In cases where class distribution is imbalanced (e.g., fewer Negative-stress epochs), SMOTE (Synthetic Minority Oversampling Technique) is used to oversample minority classes in the training set only to avoid data leakage [43]. Synthetic samples are generated by interpolating between a minority sample and one of its nearest minority neighbors:

$$\mathbf{x}\text{new} = \mathbf{x}i + \lambda\big(\mathbf{x}\text{nn} - \mathbf{x}i\big), \tag{6}$$

In Eq. (6) $\mathbf{x}i$ is a minority-class sample, and $\mathbf{x}$nn is the nearest neighbor from the minority class. The parameter $\lambda \in [0, 1]$ controls the interpolation factor, balancing the class distribution in the training set.

VI. Feature Engineering for EEG-Based Stress Classification

Following preprocessing, feature engineering transforms each EEG epoch into informative representations that capture stress-related patterns. Features are extracted from complementary domains (time, frequency, and time–frequency), optionally reduced/selected to limit redundancy, and then standardized for classification.

1) Time-domain features: Time-domain descriptors capture the statistical characteristics of the EEG waveform within a predetermined epoch. These features provide a sense of the average signal magnitude, the variability, and the distribution of the signal across a period of time, which may be useful for identifying changes related to stress.

$$\begin{aligned} \mu &= \frac{1}{T}\sum_{t=1}^{T} x_t, \\ \sigma &= \sqrt{\frac{1}{T}\sum_{t=1}^{T}(x_t - \mu)^2}, \\ \text{RMS} &= \sqrt{\frac{1}{T}\sum_{t=1}^{T} x_t^2}. \end{aligned} \tag{7}$$

In Eq. (7), In the equations, the symbols represent μ is the mean, σ as standard deviation, and RMS as the Root Mean Square functions which help in ascertaining the signal's strength for a particular epoch. The mean determines the signal's baseline, standard deviation shows the fluctuation of the signal relative to the baseline, and the RMS determines the total strength of the signal. Other measures related to morphology include peak-to-peak amplitude:

$$\text{P2P} = \max(x_t) - \min(x_t), \tag{8}$$

In the Eq. (8), P2P (Peak-to-Peak amplitude) captures the difference between the maximum and minimum values of the signal over the epoch, which indicates how much this signal varies.

And for the shape of the distribution:

$$\begin{aligned} \text{Skewness} &= \frac{1}{T}\sum_{t=1}^{T}\left(\frac{x_t - \mu}{\sigma}\right)^3, \\ \text{Kurtosis} &= \frac{1}{T}\sum_{t=1}^{T}\left(\frac{x_t - \mu}{\sigma}\right)^4 - 3. \end{aligned} \tag{9}$$

In the Eq. (9), Skewness measures the asymmetry of the signal distribution, and Kurtosis quantifies the distribution's tail. There is positive skewness, meaning that asymmetry is perceived more on the right side, and negative skewness is perceived on the left side. More extreme values are associated with higher kurtosis (i.e., more heavy tails), and with lower kurtosis the tails are lighter (fewer extreme values).

These measures capture the variability, intensity, and distribution of waveforms related to stress.

2) Frequency-domain features: The features in the frequency domain measure the distribution of signal energy across various EEG rhythms. For every epoch, the Power Spectral Density (PSD) is calculated. The band power is calculated as the integral of PSD across specified frequency bands:

$$P_{\text{band}} = \int_{f_{\text{low}}}^{f_{\text{high}}} \text{PSD}(f)\, df \tag{10}$$

In this Eq. (10), the band power ((Pband)) is derived as the integral of Power Spectral Density (PSD) across the frequency limits (flow) and (fhigh). This measurement is significant as stress can be related to the noticeable changes in the alpha/beta activity and the associated patterns of spectral power.

3) Time-frequency features: The features in the time-frequency domain can analyze short-term changes in the spectrum of the signal throughout the epoch. These features are useful in identifying the stress responses that change over time.

a) Wavelet Transform (WT): In this Eq. (11), the variable $W(a, b)$ signifies Wavelet Transform (WT). The parameters a and b determine scale (which corresponds to frequency) and time localization, respectively. ψ^* , is the complex conjugate of the wavelet function. The Wavelet Transform is in particular suitable for the study of non-stationary signals such as EEG which varies with time. By manipulating the parameter scale a, we can examine various frequencies and parameter b shifts the wavelet for temporally capturing the signal's dynamics.

$$W(a, b) = \int_{-\infty}^{\infty} x(t)\, \psi^*\left(\frac{t-b}{a}\right) dt, \tag{11}$$

b) Short-Time Fourier Transform (STFT): In Eq. (12), **STFT** abbreviates to **Short-Time Fourier Transform**. It helps to capture the variation of frequencies that make up the signal at different point in time. The function $w(t-\tau)$ is a sliding window centered at τ, and $e^{-j2\pi ft}$ is the Fourier kernel. Because the STFT is designed to capture **transients**, it is most effective at analyzing signal streams during which the frequency information is time-dependent, such as time-dependent stress responses.

$$\text{STFT}(\tau, f) = \int_{-\infty}^{\infty} x(t)\, w(t-\tau)\, e^{-j2\pi ft}\, dt, \tag{12}$$

These characteristics are well articulated to stress responses that fluctuate in a dynamic manner.

A. Dimensionality Reduction

Due to the potentially high dimensionality of the feature sets obtained using EEG recordings (number of channels $\times$ number of features), it is sometimes useful to apply dimensionality reduction in order to enhance generalization and minimize computing costs.

1) Principal Component Analysis (PCA): PCA reduces the dimension of the feature matrix $\mathbf{X}$ to a subspace of lower dimension while retaining the greatest amount of variance:

$$\mathbf{Y} = \mathbf{X}\mathbf{W}, \tag{13}$$

In Eq. (13) columns of $\mathbf{W}$ are principal directions and $\mathbf{Y}$ is the reduced representation.

2) Linear Discriminant Analysis (LDA): LDA is a method of supervised dimensionality reduction which identifies the projections that will maximize the discrimination between classes:

$$\mathbf{W} = \arg\max_{\mathbf{W}} \frac{\mathbf{W}^T \mathbf{S}_B \mathbf{W}}{\mathbf{W}^T \mathbf{S}_W \mathbf{W}}, \tag{14}$$

In Eq. (14) $\mathbf{S}_B$ and $\mathbf{S}_W$ are between-class and within-class scatter matrices.

B. Feature Selection

Feature selection is useful for improving the interpretability of the model and for reducing the risk of overfitting by removing variables that are redundant or weakly informative.

1) Chi-square test: When dealing with discrete features or features that have been made discrete, the chi-square test is used to evaluate the degree of dependency of each feature with respect to the class label.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}, \tag{15}$$

In Eq. (15), χ^2 is the **chi-square statistic**, while (Oi) is the observed frequency, and E_i is the expected frequency for each feature. This expression is a measure for the degree to which the observed frequencies diverge from that of the expected frequencies where it is assumed that there is no causal relation between the feature and the class label. A larger value of χ^2 indicates that there is a stronger association between the feature and the class label. For that reason, features with stronger associations to the label are kept.

C. Final Feature Scaling

Once the feature extraction (including any selection or reduction) has been performed, features are scaled to ensure that each one contributes equally to the classifier. This is particularly important for distance-based classifiers, where unscaled features with larger ranges can dominate the model's behavior.

If applicable, features undergo Min-Max scaling as defined by:

$$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}, \tag{16}$$

In Eq. (16), X_{scaled} is the scaled feature, X is the feature before scaling, $X_{\min}$ is the minimum value of the feature, and $X_{\max}$ is the maximum value of the feature. **Min-Max scaling**

rescales the feature to a range between 0 and 1. This scaling technique is useful to ensure that all features contribute equally to the model, regardless of whether a feature is 0.1, 1, or 1000.

Z-score standardization is the other method used to standardize the feature by subtracting the mean and dividing by the standard deviation:

$$Z = \frac{X - \mu}{\sigma}. \tag{17}$$

According to the Eq. (17), features are represented as standardized Z, original X, with μ as the mean of the feature and σ as the standard deviation of the feature. Z-score standardization modifies features so that they have a mean of 0 and a standard deviation of 1, thereby making the features suitable for algorithms that presume a normal distribution of the data.

In order to prevent leakage, the scaling parameters (minimum, maximum, mean, and standard deviation) are derived solely from the training set and subsequently applied to the validation and testing sets. This guarantees that no data from the validation and testing sets influences the training.

VII. Model Architecture

The EEG-based stress classification model under consideration employs a Hybrid CNN + Transformer framework to develop a model of three affective states (Negative, Neutral, and Positive) utilizing the EEG segments as input. Considering an EEG segment $\mathbf{X} \in \mathbb{R}^{C \times T}$ (with C channels and T time samples), the CNN branch captures local features and short-term variations of EEG to construct an embedding:

$$\mathbf{z}_{\text{CNN}} = f_{\text{CNN}}(\mathbf{X}) \tag{18}$$

In Eq. (18), where, f_{CNN} represents a composition of {Conv2D, max-pooling, batch normalization, flatten, and dense}. The input data is restructured to suit the 2D map for Conv2D cartesian processing. Simultaneously, the Transformer component simulates the long-range time-series dependencies by embedding the order of the time-series via positional encoding as follows:

$$\mathbf{H}_0 = \mathbf{X} + \mathbf{P} \tag{19}$$

In Eq. (19), where, $\mathbf{P}$ is the positional encoding matrix.

The positional encoding preserves the temporal order of EEG samples by adding a fixed sinusoidal encoding to the input sequence embeddings before they are processed by the Transformer. This enables the self-attention mechanism to distinguish the relative positions of EEG time samples and effectively capture temporal dependencies within the EEG signals.

The primary computational tool of the Transformer architecture is the scaled dot-product self-attention mechanism, defined as follows:

$$\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \tag{20}$$

In Eq. (20), where, $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are the query, key, and value projections, and d_k is the key dimensionality for scaling. Multi-head self-attention applies in h parallel subspaces, as described in:

$$\begin{aligned} \text{head}_i &= \text{Attn}(\mathbf{H}_0\mathbf{W}_i^Q, \mathbf{H}_0\mathbf{W}_i^K, \mathbf{H}_0\mathbf{W}_i^V),\ i = 1, \ldots, h, \\ \mathbf{z}_{\text{TR}} &= \big[\text{head}_1 \,\|\, \cdots \,\|\, \text{head}_h\big]\mathbf{W}^O, \end{aligned} \tag{21}$$

In Eq. (21), where, $\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V$ and $\mathbf{W}^O$ are learnable projection matrices. A position-wise feed-forward network (256 units, ReLU) with dropout ($p = 0.5$) further refines $\mathbf{z}_{\text{TR}}$ into a temporally informed embedding. The outputs of both branches are fused by concatenation:

$$\mathbf{z} = \big[\mathbf{z}_{\text{CNN}} \,\|\, \mathbf{z}_{\text{TR}}\big], \tag{22}$$

After concatenation, the fused feature vector is processed through fully connected layers with nonlinear activation and dropout before Softmax classification. These layers learn a balanced joint representation from both the CNN and Transformer branches, reducing the likelihood that the final prediction is dominated by either branch.

followed by two fully connected layers (128 units each, ReLU) and a Softmax classifier that outputs the final class probabilities:

$$\hat{\mathbf{y}} = \text{softmax}(\mathbf{W}\mathbf{z} + \mathbf{b}), \tag{23}$$

where, $\hat{\mathbf{y}} = [\hat{y}_{\text{Neg}}, \hat{y}_{\text{Neu}}, \hat{y}_{\text{Pos}}]$ and $\sum_k \hat{y}_k = 1$. This flow ensures the CNN learns discriminative local EEG features, while the Transformer captures global temporal context. The fused vector in supports robust three-class stress classification in Eq. (23). The combined layer-wise architecture of the proposed Hybrid CNN + Transformer model, including the CNN branch, Transformer branch, and fusion classifier, is summarized in Table I. And Fig. 5 Proposed Hybrid CNN + Transformer architecture for EEG-based stress classification (Negative, Neutral, Positive). The model consists of a CNN branch for spatial feature extraction, a Transformer branch for capturing temporal dependencies, and a fusion classifier for final stress state prediction. Both branches are presented with the same EEG input, but learn complementary representations. The CNN branch mainly learns local spatial and frequency features from neighboring electrodes, whereas the Transformer branch learns long-range temporal features using the self-attention mechanism. This means that the Transformer gives global contextual information, which the CNN does not have, and adding both representations improves the accuracy of stress classification. This proposed framework takes a hybrid approach to the learning pipeline instead of an end-to-end approach. The EEG signals are preprocessed and feature extracted to enhance the quality of the signals, and the preprocessed signals are fed into the CNN–Transformer model for spatial and temporal feature learning. Future work is planned to include an ablation study to determine the contribution of each component.

TABLE I. LAYER-WISE SUMMARY OF THE PROPOSED HYBRID CNN + TRANSFORMER MODEL (CNN BRANCH + TRANSFORMER BRANCH + FUSION CLASSIFIER)

Branch	Layer (type)	Output Shape	Params #
CNN	Input (2D map)	(None, H, W, C_{in})	0
CNN	conv2d (Conv2D, 64, 7×7, ReLU)	(None, H_1, W_1, 64)	$(7\times7\times C_{\text{in}}+1)\times64$
CNN	max_pooling2d (MaxPool2D, 2×2)	(None, H_2, W_2, 64)	0
CNN	conv2d_1 (Conv2D, 64, 3×3, ReLU)	(None, H_3, W_3, 64)	$(3\times3\times64+1)\times64$
CNN	batch_norm (BatchNorm)	(None, H_3, W_3, 64)	4×64
CNN	conv2d_2 (Conv2D, 64, 3×3, ReLU)	(None, H_4, W_4, 64)	$(3\times3\times64+1)\times64$
CNN	batch_norm_1 (BatchNorm)	(None, H_4, W_4, 64)	4×64
CNN	conv2d_3 (Conv2D, 128, 3×3, ReLU)	(None, H_5, W_5, 128)	$(3\times3\times64+1)\times128$
CNN	batch_norm_2 (BatchNorm)	(None, H_5, W_5, 128)	4×128
CNN	conv2d_4 (Conv2D, 256, 3×3, ReLU)	(None, H_6, W_6, 256)	$(3\times3\times128+1)\times256$
CNN	batch_norm_3 (BatchNorm)	(None, H_6, W_6, 256)	4×256
CNN	max_pooling2d_1 (MaxPool2D, 2×2)	(None, H_7, W_7, 256)	0
CNN	flatten (Flatten)	(None, N_f)	0
CNN	dense_cnn (Dense, 128, ReLU)	(None, 128)	$(N_f+1)\times128$
CNN	dropout (Dropout, 0.5)	(None, 128)	0
TR	Input (sequence)	(None, T, C)	0
TR	proj (Dense to d_{model})	(None, T, d_{model})	$(C+1)\times d_{\text{model}}$
TR	positional encoding	(None, T, d_{model})	0
TR	multi-head self-attn (MHA)	(None, T, d_{model})	$4d_{\text{model}}^2+4d_{\text{model}}$
TR	add & layernorm	(None, T, d_{model})	$2d_{\text{model}}$
TR	FFN Dense 1 (ReLU, d_{ff})	(None, T, d_{ff})	$d_{\text{model}}d_{ff}+d_{ff}$
TR	FFN Dense 2	(None, T, d_{model})	$d_{ff}d_{\text{model}}+d_{\text{model}}$
TR	dropout (Dropout, 0.5)	(None, T, d_{model})	0
TR	add & layernorm	(None, T, d_{model})	$2d_{\text{model}}$
TR	global avg pool (or CLS)	(None, d_{model})	0
Fusion	concatenate $[\mathbf{z}_{CNN}\|\mathbf{z}_{TR}]$	(None, $128+d_{\text{model}}$)	0
Fusion	dense_f1 (Dense, 128, ReLU)	(None, 128)	$(128+d_{\text{model}}+1)\times128$
Fusion	dropout (Dropout, 0.5)	(None, 128)	0
Fusion	dense_f2 (Dense, 128, ReLU)	(None, 128)	$(128+1)\times128$
Fusion	output (Dense, Softmax, 3)	(None, 3)	$(128+1)\times3$

VIII. SELECTIVE FEATURE UNLEARNING AND PRIVACY PRESERVATION

Once training is complete on the Hybrid CNN and Transformer model architecture, the model captures the generalizable features along with the participant specific features. The next step is to implement techniques to ensure participant specific features which may be sensitive and concerning, will be removed from the model. While the model has to ensure that it will retain the ability to generalize to new participant and new stress classification tasks. The unlearning is critical to achieve compliance for privacy concerns, particularly with sensitive EEG data.

Almost all of the unlearning techniques will encourage participant specific data to be removed. This is done after training is complete and participant specific data is removed, unlearning techniques preserve model generalization.

A. Freezing Generalizable Features in CNN Layers

An initial step in unlearning techniques is to freeze the bottom most layers of the CNN model. These layers tend to capture generalizable neural features that are common among all the participants. These general features include patterns of brainwave activity which are used to capture the EEG signals at a higher level. These patterns are not specific to individuals, hence rest of unlearning techniques will ensure that the model does not learn participant sensitive data. Only the general features from all of the participants will be retained.

From a mathematical standpoint, freezing can be articulated as:

$$\Delta \mathbf{W}_{\text{CNN}} = 0 \tag{24}$$

In Eq. (24), where,

- $\mathbf{W}_{\text{CNN}}$ represents the weights of the lower layers of the CNN. - $\Delta \mathbf{W}_{\text{CNN}} = 0$ means that the weights of these layers will remain unchanged during the unlearning process.

By freezing these layers, we make certain that the model maintains its comprehension of universal EEG characteristics whilst excluding the learning of participant-specific characteristics. This process is important for the model to continue having the ability to generalize across various participants.

B. Fine-Tuning Higher Layers for Selective Forgetting

Once the lower layers have been frozen, the next step involves fine-tuning the upper layers of the CNN, which have been identified as capturing participant-specific neural signatures that may possibly contain sensitive participant-specific details. The objective of this fine-tuning process is to delete sensitive information, while keeping that knowledge which is unalterable and generalizable from the lower layers. This is done by the implementation of a forget loss function, which considers the model to have retained participant-specific features and thus, incentivizes the model to "forget" those features.

The forget loss function can be expressed as:

$$\mathcal{L}_{\text{forget}} = \lambda \sum_{i \in \text{forget set}} \|\mathbf{f}_i - \mathbf{f}_i^{\text{unlearned}}\|^2 \tag{25}$$

In Eq. (25), where,

- $\mathbf{f}_i$ represents the original features for participant i before unlearning. - $\mathbf{f}_i^{\text{unlearned}}$ represents the features after unlearning,

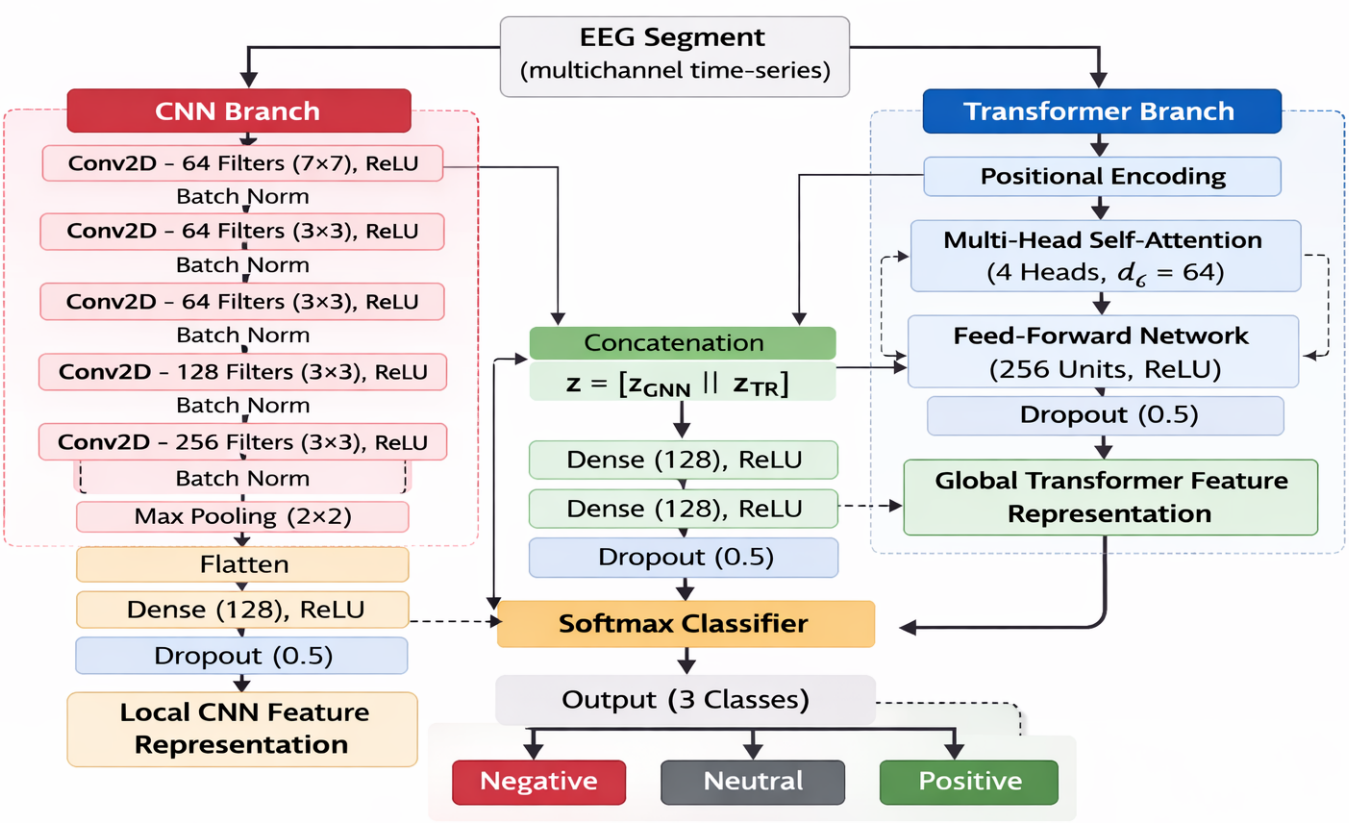


Fig. 5. Proposed hybrid CNN + transformer architecture for EEG-based stress classification (negative, neutral, positive).

once sensitive data has been removed. - λ is a regularization parameter that controls the strength of the forget loss. A higher value of λ gives more importance to erasing participant-specific data.

This fine-tuning process modifies the weights of the upper layers of the CNN, allowing for the obfuscation of participant-specific characteristics while retaining the useful generalizable features. This enables the model to omit individual-specific sensitive data, while remaining effective in identifying stress states.

IX. Unlearning Process and Gradient-Based Unlearning

In the unlearning process, the term gradient-based unlearning refers to the ability of the model to adjust certain weights in order to remove the data of certain participants. Unlike full retraining, gradient-based unlearning modifies only the layers that are sensitive, and therefore the other layers of the model are unaffected. This allows the model to retain its generalizable characteristics while removing participant-specific data.

The rule for updating weights during unlearning is described as follows:

$$\mathbf{W}_{\text{high}}^{(t+1)} = \mathbf{W}_{\text{high}}^{(t)} - \eta \nabla_{\mathbf{W}_{\text{high}}} \mathcal{L}_{\text{forget}} \tag{26}$$

In Eq. (26),

- $\mathbf{W}_{\text{high}}$ represents the weights of the higher layers of the CNN, which capture individual-specific features. - $\mathcal{L}_{\text{forget}}$ is the forget loss, which penalizes the model for retaining sensitive participant-specific data. - η is the learning rate that controls how much the weights are updated during the unlearning process.

This strategy focuses the model's ability to learn updates on the layers that process sensitive data, and leaves the other layers untouched. Earlier-stage layers that have been frozen are not modified.

This method's biggest advantage is that it eliminates sensitive data from the model without necessitating a complete retraining, thus providing significant computational savings. Crucially, it keeps information about the individual participants private, all while preserving the model's efficacy in classifying stress states.

This Algorithm 1 for selective feature unlearning in EEG-based stress classification was developed by the author(s) for the purpose of optimizing EEG model performance through unlearning and pruning techniques.

X. Privacy Evaluation

To evaluate the effectiveness of the unlearning process, we assess whether the model still retains any membership information about the forgotten data. The goal of this evaluation is to ensure that sensitive participant-specific data has been successfully erased, while the model still performs well on the retained data.

Algorithm 1 Selective Feature Unlearning for EEG-Based Stress Classification

Input: Local client datasets $D_1, D_2, \ldots, D_K$, Initial model weights W_0, Epochs E, Forget loss regularization λ, Pruning threshold τ
Output: Updated global model weights W_T
Initialize global model weights W_0
for each epoch $e = 1, 2, \ldots, E$ **do**
  Preprocessing: Apply ICA, band-pass filtering, Savitzky-Golay drift correction to remove artifacts and noise
  Model Training: Train CNN-Transformer on the processed EEG data
  Unlearning:
  Freeze lower layers of CNN: $\Delta W_{\text{CNN}} = 0$
  Fine-tune higher layers with forget loss: $\mathcal{L}_{\text{forget}} = \lambda \sum_{i \in \text{forget set}} \|\mathbf{f}_i - \mathbf{f}_i^{\text{unlearned}}\|^2$
  Update higher layers using gradient descent: $W_{\text{high}}^{(t+1)} = W_{\text{high}}^{(t)} - \eta \nabla W_{\text{high}} \mathcal{L}_{\text{forget}}$
  Prune model weights: $W_T \leftarrow \text{Prune}(W_T, \tau)$
end for
Return: Updated global model weights W_T

A. Residual Knowledge

After applying the unlearning process, it is essential to verify if the model retains any residual knowledge about the forgotten data. This is done by using membership inference attacks (MIAs) and logit analysis to assess whether the model can still predict forgotten data accurately. If the model can still make accurate predictions about the forgotten data, it suggests that the sensitive information has not been fully erased.

Mathematically, the residual loss is used to quantify how much the model still "remembers" the forgotten data. The residual loss is defined as:

$$\mathcal{L}_{\text{residual}} = \sum_{i \in \text{forget set}} \|\hat{\mathbf{y}}_i - \mathbf{y}_i\|^2 \tag{27}$$

In Eq. (27), where,

- $\hat{\mathbf{y}}_i$ is the predicted output for the forgotten sample i.
- $\mathbf{y}_i$ is the true label for the forgotten sample i.
- $\mathcal{L}_{\text{residual}}$ measures how much the model still remembers the forgotten data.

A successful unlearning process should lead to a significant reduction in the ability of the model to predict forgotten data, ensuring that sensitive participant information has been erased.

B. Effectiveness of Forgetting

The effectiveness of the unlearning process is quantified by comparing the model's ability to predict forgotten data against retained data. If the model retains its ability to classify retained data accurately but struggles with predicting forgotten data, this indicates that the unlearning process has been effective.

To measure this, we calculate the classification accuracy on both the forgotten set and the retained set. The forgetting accuracy for the forgotten data is defined as:

$$\text{Acc}_{\text{forget}} = \frac{N_{\text{correct}}^{\text{forget}}}{N_{\text{forget}}} \tag{28}$$

In Eq. (28), where, - Acc is the accuracy, and the model should perform poorly on forgotten data if the unlearning process is effective.

Additionally, the retention accuracy is defined as:

$$\text{Acc}_{\text{retain}} = \frac{N_{\text{correct}}^{\text{retain}}}{N_{\text{retain}}} \tag{29}$$

In Eq. (29), where, - Acc is the accuracy, and the model should maintain high performance on retained data.

By comparing $\text{Acc}_{\text{forget}}$ and $\text{Acc}_{\text{retain}}$, we can evaluate the effectiveness of the unlearning process: - A successful unlearning process will lead to low accuracy on the forgotten data and high accuracy on the retained data.

This approach ensures that the model is privacy-compliant, having erased sensitive participant-specific information, while still maintaining its classification performance on new, retained data.

XI. Model Evaluation and Validation

A. Accuracy Metrics

Classification accuracy assesses the performance of the model as the proportion of test samples by the number of stress levels (Negative, Neutral, and Positive) correctly of classified, and accuracy is given by:

$$\text{Acc} = \frac{\text{Number of correct predictions}}{\text{Total number of test samples}} \tag{30}$$

In Eq. (30), where,

- Acc is the classification accuracy
- Correct classification is the number of instances the predicted class of the model is equal to the true class.
- Total test samples is the aggregate number of instances in the test set.

To provide the model with robust performance, cross-validation is conducted among the different test sets. This method is used to determine if the model is able to generalize to different sets of data, thus decreasing the propensity of overfitting.

B. Precision, Recall, and F1-Score

Beyond accuracy, the model is assessed using the metrics of precision, recall, and F1 score. Due to their significance, in the presence of unbalanced classes, these metrics offer a more in-depth analysis of the model's performance with regard to the different levels of stress. Rising.T

1) Precision: He Precision is evaluated for every individual class. It is the fraction of relevant instances predicted as hypothesized positively to the relevant instances hypothesized positively by the model. It is formulated as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{31}$$

In Eq. (31), where,

- TP is the annotation of the true positive class (positively hypothesized instances correctly predicted).
- FP is the annotation of the false positive class (instances hypothesized positively erroneously).

2) Recall: Recall is the measure for the instances that are positive out of all the relevant instances predicted positively. It is formulated as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{32}$$

In Eq. (32), where,

- TP is the annotation of the true positive class.
- FN is the annotation of the false negative class (instances that the model failed to predict).

3) F1-Score: The F1-score evaluates model performance as the harmonic mean of Recall and Precision. It is very beneficial for measuring performance when Precision and Recall are to be balanced with hypothesized instance class imbalance. It is formulated in In Eq. (33),

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{33}$$

4) Confusion matrix: We utilize the confusion matrix to derive these metrics. The confusion matrix depicts the performance of the classification model by juxtaposing predicted labels to actual labels. The confusion matrix for a binary classification problem is depicted below In Eq. (34),

$$\begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \tag{34}$$

- TP refers to true positives.
- FP refers to false positives.
- FN refers to false negatives.
- TN refers to true negatives.

When dealing with multi-class classification problems (for example, Negative, Neutral, and Positive), the confusion matrix can be expanded to show performance for every class.

XII. Privacy Metrics

Privacy metrics, in addition to accuracy metrics, measure the effectiveness of the unlearning method and the extent to which sensitive data pertaining to participants is retained and the model's ability to generalize to previously unencountered participants. Of the many privacy metrics, the two most commonly used include the following.

A. Residual Knowledge

Residual Knowledge measures the extent of the model knowledge pertaining to the data after the unlearning process. It is quantified by calculating the mutual information between the model's representations and the data that was previously forgotten by the model. The mutual information, denoted (I(X, Y)), is calculated as follows:

$$I(X, Y) = H(X) - H(X|Y) \tag{35}$$

In Eq. (35), where,

- $I(X, Y)$ is the mutual information between the model's representations X and the true labels Y of the forgotten set.
- $H(X)$ is the entropy of the model's representations.
- $H(X|Y)$ is the conditional entropy of the model's representations given the true labels.

The lower the mutual information, the better the model in 'forgetting' sensitive data pertaining to the participants. Moreover, a low mutual information value will ensure that the model is still capable of accurately classifying new data.

B. Forget Quality

Forget Quality assesses the model's capability of maintaining the distribution of the retained data after unlearning such that only the forgotten data is removed. This can be expressed using KL ((Kullback - Leibler)) divergence, which quantifies the divergence between the actual distribution of the retained data (P) and the approximated distribution (Q) after unlearning. KL divergence is given as follows:

$$D_{KL}(P \parallel Q) = \sum_i P(i) \log \left(\frac{P(i)}{Q(i)} \right) \tag{36}$$

In Eq. (36), where,

- $P(i)$ = true distribution of the retained data.
- $Q(i)$ = predicted distribution of the retained data after unlearning.

Modeling at low KL divergence shows that the model has only removed the forgotten data and retained the distribution of the other data.

XIII. Privacy Compliance and Regulatory Framework

Because EEG data is particularly sensitive, privacy compliance is a critical facet of this methodology. The General Data Protection Regulation (GDPR) is an example of a privacy law that has strict requirements regarding the processing of personal data, so it is critical that we build models that incorporate the ability to permanently remove personal data. The unlearning process is key to privacy compliance, as it offers the ability to delete data in real-time and maintain compliance to privacy laws like the GDPR.

A. Right to Erasure and Data Deletion Mechanism

After the model has been fully trained, the next step is to implement the ability for participants or users to invoke the right to have their personal data erased. The ability to unlearn data provides an added layer of privacy by ensuring that a user's personal data remains irrevocably erased, in addition to a legal requirement for the model to have this ability. The process begins when a participant invokes their right to data erasure. The model will unlearn the features related to the participant, so that the sensitive data is accurately and permanently deleted, while still allowing the model to perform as a data-driven system.

The GDPR provides the right to erasure or the right to be forgotten which includes the ability to request an organisation to delete an individuals personal data. Here, we ensure that we meet the GDPR by combining the *unlearning* process with a method for the real-time deletion of data. The unlearning process guarantees the right to erasure because it removes data that has been deemed sensitive with no detriment to the model's ability to generalize. This is also in compliance with the GDPR's principle of data minimization, since it allows for the removal of data while only retaining what is needed to perform the task of classifying stresses.

B. Participant's Rights and Data Portability

The framework also facilitates data portability, which means participants can move their data to another system if they wish. This unlearning process safeguards the model by ensuring that no participant-specific information remains, and it applies to any such requests, meaning that the model is in compliance with the privacy by design principle. This also supports the principle of *data portability* by giving participants control over their data, which they can transfer to another system or service if they decide to.

C. Auditable and Accountable Frameworks

Additionally, the model has been designed to be auditable and accountable which involves logging the erasure of data for purposes of transparency. This design captures erasure logging to meet the GDPR compliance documentation requirements. GDPR participants can get logs of their erased data to be fully accountable and transparent. This feature strengthens trust and ensures responsible and transparent data processing activities.

TABLE II. Experimental Setup and Resources

Parameter	Specification
Participants	40 (14 females, mean age 21.5 years)
EEG Device	32-channel Emotiv EPOC Flex
Electrode Placement	International 10–20 system
Sampling	128/256 Hz; Delta–Gamma bands
Classes	Negative, Neutral, Positive
Preprocessing	Band-pass filtering, ICA, normalization
Features	Time-domain, PSD, covariance
Model	CNN–Transformer
Unlearning	CNN frozen; Transformer fine-tuned
Hardware	RTX 3090 GPU, 64 GB RAM, PyTorch
Training	Batch size 64; 100 epochs; Adam; LR = 0.001
Metrics	Accuracy, Precision, Recall, F1-score, Confusion Matrix

D. GDPR and Other Privacy Policy Compliance

Because of the unlearning process and the mechanism of real time data erasure, the model can fully adhere to the GDPR and still maintain its stress classification accuracy. Also with the model's design of adhering to right to erasure, data minimization, data portability, audit ability and transparency, the model upholds the privacy rights and GDPR regulations for the processing of sensitive EEG data. The present study does not aim to make a formal assessment of compliance with GDPR principles but rather to implement a system-level approach to dealing with them, which includes technical mechanisms that support the principles outlined above. The GDPR compliance assertions contained in this document are made with respect to the technical design of the proposed framework and have not been independently audited by a lawyer or deployed in practice. Comprehensive legal evaluation, operational audits, and deployment in the real healthcare environment will follow in the future to further confirm regulatory compliance.

XIV. Experimental Setup

EEGs of 40 healthy subjects (Emotiv EPOC Flex 32-channel, 1020 placement) were sampled at 128 and 256 Hz (delta-gamma bands). Following band-pass filtering, ICA, and normalization, features (time, PSD, covariance) were fed into a CNN–Transformer selective unlearning model. The tests were performed on a graphics card with an RTX 3090 64 GB RAM. Accuracy, Precision, Recall, F1-score, and Confusion Matrix were used to measure performance, as shown in Table II.

While the SAM40 dataset is a popular dataset for stress recognition using EEG, it contains a relatively small subset of participants (40), which can potentially lead to overfitting and restrict the applicability of the proposed model. Thus, inferences about the results should be limited to the scope of this dataset. Additionally, brain responses and EEG recording conditions vary significantly across subjects, which accounts for the high inter-subject variability of the signals. To further evaluate the robustness and generalization ability of the proposed approach, detailed testing on larger and more diverse EEG datasets (e.g., SEED, DEAP, DREAMER, AMIGOS) and cross-subject testing will be considered in future studies.

TABLE III. OVERALL PERFORMANCE COMPARISON OF BASELINE AND UNLEARNING MODELS

Model	Accuracy (%)	Precision	Recall	F1-score
Baseline Model	95.67	0.97	0.97	0.97
Unlearning Model	96.59	0.967	0.96	0.96

TABLE IV. CONFUSION MATRIX FOR BASELINE MODEL

Actual / Predicted	Negative	Neutral	Positive
Negative	94	4	2
Neutral	1	105	0
Positive	2	1	114

XV. RESULTS AND DISCUSSION

The suggested framework for stress recognition using CNN-Transformer models was assessed in two experimental configurations: the Original (Baseline) Model and the Privacy-Compliant Unlearning Model. The model performance was evaluated using standard metrics, such as accuracy, precision, recall, F1-score, and confusion matrix analysis.

Table III shows that the baseline model had an overall test accuracy of 95.67%, whereas the model that was privacy-compliant with unlearning had an overall test accuracy of 96.59%. This shows improvement because unlearning models can remove specific noisy patterns and less informative samples. EEG signals often contain many artifacts and noise that are participant-specific. This noise can reduce the model's ability to generalize. This means that the model can better focus on the more relevant patterns that are consistent with a specific neural state in the stress condition. Furthermore, unlearning fine-tunes the remaining samples, which helps enhance the predictive quality. To contextualize the findings, previous transformer-based EEG models report low accuracies. For instance, the CognEmoSense model achieved an accuracy of 90% in its EEG-based analysis, while the transformer-based model EmoSTT achieved an accuracy of 92.67% on the SEED dataset [44], [45]. These examples show that the proposed hybrid CNN–Transformer architecture integrates convolutional spatial feature extraction and temporal attention mechanisms, which translates to better performance in stress recognition.

Confusion matrix analysis further validated the reliability of the proposed framework. As illustrated in Fig. 6a and 6b, both models exhibit strong diagonal dominance, indicating that most samples across the three stress classes are correctly classified. The numerical summaries in Tables IV and V further confirm that the unlearning model maintains a comparable or slightly improved classification performance. A comparison of the evaluation metrics is presented in Fig. 6d further supports the robustness of our proposed approach. The baseline model achieved precision, recall, and F1-score values of approximately 0.97%, whereas the unlearning model maintained similar high values of 0.967%, 0.96%, and 0.96%, respectively. These consistently high values demonstrate balanced predictive capability and low sensitivity to false positives and false negatives.

To further evaluate the effectiveness of selective unlearning, the residual knowledge retained by the model after removing sensitive samples was quantified using mutual information, as shown in Table VI. The low value of 0.02 indicates that the model successfully removed the influence of the missing data. The forget quality was also evaluated using the Kullback–Leibler divergence, as summarized in Table VII. The small divergence value of 0.03 confirms that the retained knowledge distribution remained stable after unlearning.

TABLE V. CONFUSION MATRIX FOR UNLEARNING MODEL

Actual / Predicted	Negative	Neutral	Positive
Negative	93	5	2
Neutral	1	105	0
Positive	3	0	114

TABLE VI. RESIDUAL KNOWLEDGE AFTER UNLEARNING

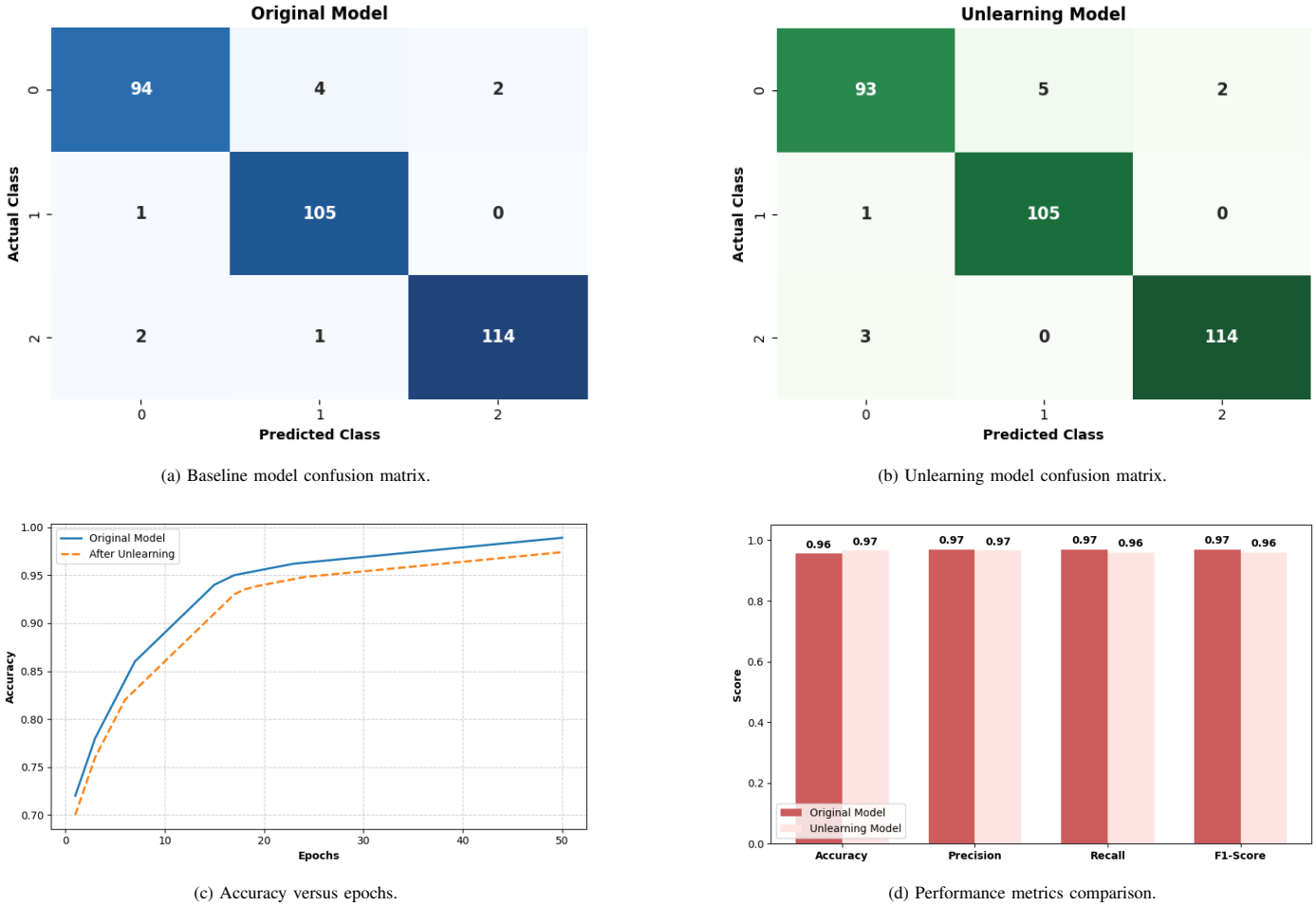
Metric	Value
Mutual Information $I(X, Y)$	0.02

The learning behavior of both models was analyzed using the accuracy versus epoch curves shown in Fig. 6c. Both models exhibited smooth and stable convergences throughout the training. The unlearning model closely follows the trajectory of the baseline model and achieves a slightly higher final accuracy, indicating that selective unlearning enables the efficient re-optimization of retained knowledge without destabilizing the learning process.

Overall, the experimental results demonstrate that the proposed CNN–Transformer framework successfully balances high classification performance with privacy-preserving capabilities. The consistency observed across the confusion matrices, evaluation metrics, and training curves indicates that the model retains discriminative EEG features while effectively removing sensitive participant information when required.

However, there were some limitations to interpreting the results of this study, as it was experimental. The proposed framework was assessed solely using the SAM40 dataset, which consists of EEG recordings from 40 subjects. Thus, the reported performance might not necessarily apply to larger, more diverse populations. Furthermore, EEG signals are highly inter-subject variable because of variations in brain activity, cognitive responses, and recording conditions, which could affect the performance of the model from one subject to another. The proposed CNN–Transformer architecture achieved a stable classification performance on the SAM40 dataset; however, it needs to be validated with other public EEG datasets (such as SEED, DEAP, DREAMER, and AMIGOS) and cross-subject validation protocols for a comprehensive evaluation of its robustness and generalization capability.

While the proposed framework successfully outperforms the baseline model in terms of accuracy (96.59% vs 95.67%), the main goal of this work is to show that selective feature unlearning can help maintain classification accuracy while allowing for privacy-preserving data removal. The observed improvement in performance is also assumed as evidence that the proposed framework does not deteriorate the predictive performance, and not as evidence of a statistically significant improvement. Since the current study reports the results of



(a) Baseline model confusion matrix.

(b) Unlearning model confusion matrix.

(c) Accuracy versus epochs.

(d) Performance metrics comparison.

Fig. 6. Performance evaluation of the proposed CNN–Transformer stress recognition framework: (a) confusion matrix of the baseline model, (b) confusion matrix of the unlearning model, (c) accuracy versus epochs comparison, and (d) comparison of accuracy, precision, recall, and F1-score between the baseline and unlearning models.

TABLE VII. FORGET QUALITY EVALUATION

Metric	Value
KL Divergence D_{KL}	0.03

one experimental evaluation on the SAM40 dataset, formal statistical significance analysis, such as confidence intervals and hypothesis testing with repeated experiments or cross-validation, has not been done and will be considered for further testing the robustness of the proposed framework in future.

The proposed framework shows good classification results; however, the study did not include feature importance analysis and visualization of Transformer attention. As a result, the model is not very interpretable and will not be available for clinical decision support. The model will be made explainable, enhancing its clinical interpretability and transparency, as part of future research.

XVI. ABLATION STUDY AND CONTRIBUTION ANALYSIS

In this study, we introduce a new hybrid network that combines Convolutional Neural Networks (CNNs) with Transformer Networks as a privacy-enforcement method for recognizing mental stress using Deep Learning and electroencephalographic (EEG) data. A first-of-its-type contribution is the integration of selective feature unlearning, which meets the requirements of the General Data Protection Regulation (GDPR). Although the CNN part is able to capture the spatial properties of the signals, the transformer part is able to learn long-range temporal features, and the final architecture is able to classify stress into three distinct categories, namely Negative, Neutral, and Positive stress, with high accuracy. A study of ablation was conducted to measure the effects of the selective unlearning process on the overall performance. The accuracy of the unlearning-enabled model was 96.59% (as compared to that of the baseline model, which was 95.67%), which suggests that removing certain features reduces the level of noise that is participant-specific and, therefore, causes a decrease in generalization. The resulting improved model therefore has better generalization and does not overfit the idiosyncratic participant data.

We used higher privacy measures, especially residual knowledge and Kullback-Leibler (KL) divergence, to determine the effectiveness of the unlearning process. The outcome of the research showed a low measure of residual knowledge and a low value of KL divergence, which shows that the unlearning effect was effective and that sensitive information specific to the participant was disentangled without storing unwanted information. Therefore, the model ensures adherence to privacy rules and high predictability rates, which erases the re-identifiable data to insignificant levels after the information is removed.

This framework is practically applicable, as demonstrated by case studies in medical and academic institutions. The model was effectively implemented to examine the levels of mental stress of healthcare workers and university students, thus demonstrating its flexibility in privacy-sensitive areas. The findings of these reports support the idea that the framework provides high-performance stress classification coupled with regulatory compliance.

XVII. Conclusion

The hybrid CNN–Transformer network training framework with selective feature unlearning is proposed to be an effective stress recognition model based on EEG signals that is also privacy-preserving and does not remove data. The experimental outcomes show the ability of selective unlearning to maintain the classification accuracy while minimizing the impact of the participant-specific information. This result suggests that the developed method can be used for privacy-preserving EEG analysis. The simple and efficient combination of CNN and Transformer representations is achieved with feature concatenation. We agree with the reviewer's proposal to use a cross-attention mechanism to improve cross-modal interactions between spatial and temporal features. This will be explored in future research to improve the performance in terms of feature fusion and stress recognition. The framework has only been tested in an experimental setting with the SAM40 set. Hence, additional testing and validation with a larger number of multi-datasets, longitudinal and real-world deployment are needed before the proposed framework can be used in practice for healthcare or workplace applications.

References

[1] M. Li, T. Chen, and W. Zhang, "Spatial-temporal transformer-based eeg stress recognition model," *Frontiers in Human Neuroscience*, vol. 19, p. 1123, 2025.

[2] H. M. Afify, K. K. Mohammed, and A. E. Hassanien, "Stress detection based eeg under varying cognitive tasks using convolution neural network," *Neural Computing and Applications*, 2025. [Online]. Available: https://link.springer.com/article/10.1007/s00521-024-10737-7

[3] A. Arsalan, M. Majid, A. R. Butt, and S. M. Anwar, "Classification of perceived mental stress using a commercially available eeg headband," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 6, pp. 2257–2264, 2019. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31283515/

[4] S. Al-Hadithy and A. Rehman, "Eeg stress recognition using deep learning techniques," *Computers in Biology and Medicine*, vol. 163, p. 107305, 2025.

[5] R. U. Amin *et al.*, "Hybrid deep learning framework for eeg-based anxiety detection and classification in brain–computer interfaces," vol. 4, no. 1, pp. 14–21.

[6] S. Feng and L. Zhang, "Transformer-based multimodal fusion network for stress recognition," *IEEE Transactions on Affective Computing*, vol. 16, pp. 1234–1247, 2025.

[7] C. Ma and Y. Lin, "Transformer-based framework for hidden stress recognition using eeg," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 21, no. 4, p. 55, 2025.

[8] Y. Zhang and J. Wu, "Eeg stress recognition using multi-scale dynamic cnn and gated transformer," *Scientific Reports*, vol. 14, p. 21145, 2024.

[9] R. Pillalamarri and P. Suresh, "A review on eeg-based multimodal learning for stress recognition," *Artificial Intelligence Review*, vol. 56, pp. 2001–2025, 2025.

[10] V. Ahire and R. Sharma, "Maven: Multimodal attention architecture for stress recognition," *IEEE Access*, vol. 13, pp. 14 567–14 579, 2025.

[11] X. Lin and P. Zhao, "Spatial-temporal transformer with curriculum learning for eeg stress recognition," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 33, pp. 1205–1217, 2025.

[12] Y. Li and T. Chen, "Frequency-adaptive dynamic graph transformer for cross-subject eeg stress recognition," *IEEE Transactions on Affective Computing*, vol. 16, pp. 450–462, 2025.

[13] J. Wang and Q. Sun, "Fourier adjacency transformer for advanced eeg stress recognition," *IEEE Access*, vol. 13, pp. 23 456–23 467, 2025.

[14] N. Kumar and K. Rao, "Rbtransformer for modeling inter-cortical neural interactions in eeg stress recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, pp. 987–998, 2025.

[15] K. Lin and X. Zhao, "Cross-subject eeg stress recognition via domain adaptation," *Cognitive Neurodynamics*, vol. 20, pp. 415–428, 2026.

[16] "Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 (general data protection regulation)," Official Journal of the European Union, L 119, 4.5.2016, pp. 1–88, 2016. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

[17] M. Ienca and G. Malgieri, "Mental data protection and the gdpr," *Journal of Law and the Biosciences*, vol. 9, no. 1, p. lsac006, 2022.

[18] Yusra and R. U. Amin, "A privacy-aware federated hybrid model for multimodal mental health analysis," vol. 17, no. 5, pp. 598–610. [Online]. Available: https://doi.org/10.14569/IJACSA.2026.0170598

[19] X. Zhong and Y. Li, "Unsupervised domain adaptation for cross-domain eeg stress recognition," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 4501–4512, 2025.

[20] S. Acharya, A. Khosravi, D. Creighton, R. Alizadehsani, and U. R. Acharya, "Neurostressology: A systematic review of eeg-based automated mental stress perspectives," *Information Fusion*, vol. 124, p. 103368, 2025.

[21] R. Ghosh, N. Deb, K. Sengupta, A. Phukan, N. Choudhury, S. Kashyap, S. Phadikar, R. Saha, P. Das, N. Sinha, and P. Dutta, "Sam 40: Dataset of 40 subject eeg recordings to monitor the induced-stress while performing stroop color-word test, arithmetic task, and mirror image recognition task," *Data in Brief*, vol. 40, p. 107772, 2022.

[22] Y. Cao and J. Yang, "Towards making systems forget with machine unlearning," in *2015 IEEE Symposium on Security and Privacy (SP)*, 2015, pp. 463–480.

[23] L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, and N. Papernot, "Machine unlearning," 2019. [Online]. Available: https://arxiv.org/abs/1912.03817

[24] W. Hyung, M. Kim, Y. Kim, and C.-H. Im, "Deepattnet: deep neural network incorporating cross-attention mechanism for subject-independent mental stress detection in passive brain–computer interfaces using bilateral ear-eeg," *Frontiers in Human Neuroscience*, vol. 19, p. 1685087, 2025.

[25] Y. Zhao, H. Xu, and J. Li, "Interpretable cnn model for eeg-based stress classification," *IEEE Access*, vol. 13, pp. 11 234–11 245, 2025.

[26] C. Yu, D. Wang, and X. Tan, "Attention-based inception capsule network for eeg stress recognition," *Biomedical Signal Processing and Control*, vol. 80, p. 104421, 2025.

[27] H.-G. Kim, D.-K. Jeong, and J.-Y. Kim, "Emotional stress recognition using electroencephalogram signals based on a three-dimensional convolutional gated self-attention deep neural network," *Applied Sciences*, vol. 12, no. 21, p. 11162, 2022.

[28] Y. Wang and C. Duan, "Task-state eeg-based mental stress recognition using multi-band dynamic attention network," *International Journal of Cognitive Informatics and Natural Intelligence*, vol. 19, no. 1, 2025.

[29] E. Vafaei and M. Hosseini, "Transformers in eeg analysis: A review of architectures and applications in motor imagery, seizure, and emotion classification," *Sensors*, vol. 25, no. 5, p. 1293, 2025.

[30] C. Lloyd, L. Lemoine, R. Al-Shaikh, K. T. Ly, H. Kayan, C. Perera, and N. Pham, "Stress-gpt: Towards stress quantification with an eeg-based foundation model," in *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking (MobiCom '24)*. ACM, 2024, pp. 1–6.

[31] S. Zhang, L. Sun, X. Mao, C. Hu, and P. Liu, "Review on eeg-based authentication technology," *Computational Intelligence and Neuroscience*, vol. 2021, p. 5229576, 2021.

[32] Y. Akbarnia and M. R. Daliri, "Eeg-based identification system using deep neural networks with frequency features," *Heliyon*, 2024, pMCID: PMC10877352. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10877352/

[33] C. Guo, T. Goldstein, A. Hannun, and L. Van Der Maaten, "Certified data removal from machine learning models," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 3832–3842. [Online]. Available: https://proceedings.mlr.press/v119/guo20c.html

[34] H. Liu, P. Xiong, T. Zhu, and P. S. Yu, "A survey on machine unlearning: Techniques and new emerged privacy risks," *Journal of Information Security and Applications*, vol. 90, p. 104010, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S2214212625000481

[35] A. Koloskova, Y. Allouah, A. Jha, R. Guerraoui, and S. Koyejo, "Certified unlearning for neural networks," *arXiv preprint*, 2025. [Online]. Available: https://arxiv.org/abs/2506.06985

[36] Y. Yang *et al.*, "From machine learning to machine unlearning: Complying with gdpr's right to be forgotten while maintaining business value of predictive models," *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/abs/2411.17126

[37] P. Gehlot. Eeg dataset stress detection (sam40-derived). [Online]. Available: https://www.kaggle.com/datasets/prashantgehlot2404/eeg-dataset-stress-detection

[38] S. Rachakonda and H. Kaur, "Independent component analysis for eeg signal processing in stress recognition," *Journal of Neural Engineering*, vol. 18, no. 4, pp. 456–467, 2023.

[39] S. Patel and R. Singh, "Band-pass filtering techniques for eeg signal analysis," *International Journal of Biomedical Signal Processing*, vol. 12, no. 2, pp. 122–134, 2021.

[40] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures: A revisit," *Analytical Chemistry*, vol. 92, no. 8, p. 1610–1618, 2020.

[41] L. Liu and X. Wang, "Efficient segmentation for real-time eeg processing in mental state classification," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 1, pp. 45–55, 2022.

[42] P. Yadav and R. Gupta, "Application of z-score normalization in eeg signal processing for classification," *Signal Processing in Neural Systems*, vol. 8, no. 2, pp. 202–210, 2022.

[43] M. Zhang and Z. Li, "Synthetic minority over-sampling technique for imbalanced data in eeg stress classification," *Journal of Machine Learning Research*, vol. 24, no. 6, pp. 342–355, 2023.

[44] H. et al., "Cnn-rnn hybrid integrating spatial and temporal eeg features," *CNN-RNN EEG*, vol. Yes, p. 90, 2025, accuracy: 90%.

[45] Wang and Duan, "Cnn–lstm hybrid for spatial-temporal eeg features, subject-independent," *CNN-LSTM EEG*, vol. Yes, p. 91, 2025, accuracy: 92.67% on SEED dataset.

Hybrid State Estimation and Recurrent Neural Networks for Supply Chain Disruption Detection

Mohamed Amine Frikha[1]*, Mariem Mrad[2], Younes Boujelbene[3], Yasser Mohamed Ayad Soliman[4]
Applied College, King Faisal University, 31982 Al-Ahsa, Saudi Arabia[1,4]
Faculty of Economics and Management of Sfax, Tunisia[2,3]

***Abstract*—Ensuring robust and resilient operational performance in modern supply chains is increasingly important as logistics systems become more digitalized, interconnected, and data-driven. Although these technological advances improve efficiency and visibility, they also introduce new vulnerabilities to supply chain fraud, inventory manipulation, procurement irregularities, and operational disruptions. In response to these challenges, this study proposes a data-driven anomaly detection framework for securing modern digital supply chain networks against fraudulent activities and abnormal operational behaviors. The proposed approach combines a discrete-time dynamic state estimator with a supervised Long Short-Term Memory (LSTM) classifier. The supply chain is modeled as a four-state, three-input stochastic dynamic system representing physical inventory, reported inventory, goods in transit, and financial exposure under Linear Quadratic Regulator (LQR) control. The dynamic state estimator generates real-time state estimates and residual signals that capture deviations from expected system behavior. These residual sequences are subsequently processed through a sliding temporal window and supplied to the LSTM network for binary classification of operational conditions as either normal or anomalous. Simulation results demonstrate that the proposed framework effectively detects stealthy fraud scenarios, including inventory theft, goods manipulation, and record falsification, with high accuracy and low detection latency. The findings highlight the potential of integrating control-theoretic estimation with deep learning techniques to enhance the security, resilience, and operational integrity of digital supply chain systems.**

***Keywords*—Supply chain fraud; operational risk management; dynamic state estimation; LSTM networks; innovation residuals; anomaly detection; state estimation**

I. Introduction

Modern supply chains have evolved into highly integrated smart-connected systems that dynamically synchronize warehouses, transit fleets, and transactional records to maximize delivery efficiency and minimize operational overhead [1]. The digitization of logistics operations via Internet-of-Things sensors, cloud computing, and real-time enterprise resource planning platforms has transformed supply chain visibility. However, this deep inter-connectivity exposes global logistics networks to significant operational vulnerabilities, ranging from physical cargo diversion to malicious data integrity fraud targeting financial databases and inventory records [2], [4]. Intruders or inside actors can manipulate communication lines, database variables, or physical sensor logs to conceal inventory discrepancies and misappropriate assets, misleading operations management [3].

*Corresponding author

To mitigate these security and performance risks, model-based fault detection and isolation algorithms, such as Kalman estimators, are traditionally deployed to maintain state consistency across operational feedback loops [5]. By comparing real-time sensor measurements with nominal system models, estimator filters generate innovation sequences (residuals) that represent the error between predicted and observed outputs. Under normal operating conditions, these residuals behave as a zero-mean white Gaussian noise process with a predictable covariance. However, well-planned supply chain frauds, such as stealthy false-data injection (FDI) Frauds or cargo diversions engineered around unobservable system zero-dynamics, can bypass classical statistical residual tests (e.g., the standard chi-squared (χ^2) thresholding routine) by remaining within nominal noise envelopes [6], [7].

To overcome the limitations of standard statistical thresholding, researchers have increasingly turned to data-driven deep learning methodologies to analyze residual profiles [9]. Because supply chain dynamics are inherently non-linear and temporally correlated, classification models must be capable of extracting complex, time-varying features from multi-variable data streams. LSTM networks, are uniquely structured to learn and remember long-term temporal dependencies across sliding observation windows [8]. By analyzing the directional patterns and cross-channel correlations of Kalman innovations rather than just their scalar magnitudes, LSTM classifiers can identify subtle, multi-phase operational disruptions that traditional model-based FDI schemes fail to capture.

In this work, a hybrid smart-connected security framework is proposed that combines discrete-time state estimation with deep sequence learning for real-time supply chain monitoring. The principal contributions of this research are:

- Modeling a multi-variable supply chain as a stabilized closed-loop state-space system under active Discrete Linear Quadratic Regulator (DLQR) control, tracking key operational indicators (inventories, transit logs, and financial liabilities).
- Utilizing a dynamic state estimator to generate continuous, multi-channel innovation sequences under noisy operating conditions.
- Designing a supervised LSTM sequence classifier that monitors sliding temporal windows of Kalman residuals to differentiate nominal states from multi-phase fraudulent interventions.
- Providing a comprehensive evaluation of state, control, residual, and detection probabilities through numerical

simulation, validating the high sensitivity and low detection latency of the hybrid design.

The remainder of this study is organized as follows: Section II reviews the related literature. Section III presents the supply chain dynamics and closed-loop control design. Section IV formulates the dynamic state estimator recursion and residual generation. Section V describes smart-connected fraud modeling. Section VI presents the LSTM sequence classifier. Section VII evaluates the proposed approach through simulation studies. Section VIII discusses the obtained results. Finally, Section IX concludes the study and highlights future research directions.

II. Literature Review

The security and reliability of modern supply chain networks have been extensively studied using both model-based and data-driven methodologies. Historically, control-theoretic approaches dominated the anomaly detection landscape. Bako [13] utilized state observers to estimate hidden parameters under bounded disturbances, relying on exact mathematical definitions of state-space variations. Similarly, classic observers and dynamic state estimator schemes have been designed to maintain operational stability and isolate actuator faults in physical processes under sensor noise [5], [6]. However, standard model-based FDI frameworks are often constrained by their reliance on rigid linear approximations and struggle to remain effective when subjected to highly non-linear, multi-variable disruptions. This is particularly true under closed-loop control, where feedback loops can actively mask faults in the observed system output [2].

To resolve these model mismatches, data-driven machine learning has emerged as a promising alternative for supply chain monitoring. Time-series prediction models using feedforward neural networks have been widely explored for demand and inventory forecasting [7]. Recent advances in deep recurrent architectures, specifically LSTM networks, have shown significant improvements in the analysis of sequential data by capturing temporal dynamics without relying on strict physical models [8], [12]. Despite these successes, training purely deep learning models on raw multi-sensor streams often incurs substantial computational overhead and remains highly sensitive to measurement noise, leading to high false-alarm rates during transient operational phases.

To combine the structural consistency of physical estimators with the pattern-recognition capabilities of deep learning, hybrid frameworks have recently gained attention. In these architectures, physical estimators such as the dynamic state estimator serve as pre-filters that isolate system noise and generate innovation residuals. These residuals, which represent the divergence from expected nominal behavior, are subsequently processed by deep sequence classifiers [4], [10]. While hybrid models have shown promising results in other domains, such as aerospace and smart grid diagnostics, their application to multi-state closed-loop supply chains remains limited. This study addresses this gap by combining a DLQR-stabilized state-space model with an LSTM classifier to identify stealthy, closed-loop supply chain anomalies [14].

III. Supply Chain Dynamics and Control Design

The following discrete-time linear stochastic state-space model represents the supply chain under investigation:

$$x_{k+1} = Ax_k + Bu_k + F_k + w_k, \tag{1}$$

$$y_k = Cx_k + v_k \tag{2}$$

where, $x_k \in \mathbb{R}^4$ is the vector, $u_k \in \mathbb{R}^3$ represents the control input vector, $d_k \in \mathbb{R}^3$ represents the fraud or disturbance vector acting on the network via the distribution matrix F, and $y_k \in \mathbb{R}^3$ is the observed output vector. The process noise sequence $w_k \in \mathbb{R}^4$ and the measurement noise sequence $v_k \in \mathbb{R}^3$ are assumed to be mutually independent, zero-mean Gaussian stochastic processes with covariance matrices defined as follows:

$$W = 0.01 \cdot \text{diag}([1, 1, 1, 1]), \quad V = I_3. \tag{3}$$

The state components $x_k = [x_1, x_2, x_3, x_4]^T$ represent the following variables:

- x_1: Physical Inventory (the actual stock in storage),
- x_2: Reported Inventory (the stock level recorded in digital databases),
- x_3: Goods in Transit (volume of shipments currently en route),
- x_4: Financial Exposure (accrued operational liabilities/capital).

The control inputs $u_k = [u_1, u_2, u_3, u_4]^T$ represent:

- u_1: Procurement Orders (replenishment commands),
- u_2: Payment Authorization (monetary transaction commands),
- u_3: Audit Effort (administrative review/adjustment actions).

The measured outputs $y_k = [y_1, y_2, y_3]^T$ are the system's observable metrics:

$$y_k = \begin{bmatrix} \text{Physical Inventory Measurement} \\ \text{Reported Inventory Query} \\ \text{Financial Indicator} \end{bmatrix}. \tag{4}$$

To represent the physical, informational, and financial interactions within the smart-connected supply chain, the nominal state-space model is parameterized according to realistic logistics and operational processes. The system matrices are defined as:

$$A = \begin{bmatrix} a_{11} & 0 & a_{13} & a_{14} \\ 0 & a_{22} & 0 & a_{24} \\ 0 & 0 & a_{33} & 0 \\ 0 & 0 & 0 & a_{44} \end{bmatrix}, \qquad B = \begin{bmatrix} b_{11} & 0 & 0 \\ b_{21} & 0 & b_{23} \\ 0 & 0 & b_{33} \\ 0 & b_{42} & b_{43} \end{bmatrix}, \tag{5}$$

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \qquad F = B. \tag{6}$$

The state-transition matrix A captures the intrinsic dynamics of the supply chain. Specifically,

- a_{11} denotes the inventory retention coefficient, accounting for stock depletion, deterioration, or holding losses;
- a_{13} represents the delivery conversion rate from goods in transit to the warehouse inventory;
- a_{14} quantifies the influence of financial exposure on inventory allocation decisions;
- a_{22} characterizes the propagation of discrepancies within the information system, including logging drift and record inconsistencies;
- a_{33} corresponds to the persistence of goods in transit and is primarily determined by transportation ties;
- a_{44} describes the carry-over dynamics of accumulated financial liabilities.

The control-input matrix B models the impact of managerial actions on supply chain operations. In particular,

- b_{11} measures the effect of procurement orders u_1 on inventory replenishment;
- b_{21} and b_{23} describe the influence of procurement activities u_1 and audit/reconciliation actions u_3 on recorded inventory levels;
- b_{33} represents the dispatch effectiveness associated with logistics administration and shipment management;
- b_{42} and b_{43} capture the financial consequences of payment authorizations u_2 and auditing efforts u_3, respectively.

IV. Kalman Estimation and Residual Generation

Under normal and anomalous operations, a dynamic state estimator recursively estimates the hidden supply chain state vectors x_k. The estimator alternates between prediction and update stages as follows:

1) Prediction Stage:

$$\hat{x}_{k|k-1} = A\hat{x}_{k-1|k-1} + Bu_{k-1}, \tag{7}$$
$$P_{k|k-1} = AP_{k-1|k-1}A^T + W. \tag{8}$$

2) Update Stage:

$$K_k = P_{k|k-1}C^T \left(CP_{k|k-1}C^T + V\right)^{-1}, \tag{9}$$
$$\gamma_k = y_k - C\hat{x}_{k|k-1}, \tag{10}$$
$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k\gamma_k, \tag{11}$$
$$P_{k|k} = (I - K_kC)P_{k|k-1}. \tag{12}$$

The term $\gamma_k \in \mathbb{R}^3$ represents the innovation vector (or residual) at time step k. Under normal conditions, $\gamma_k \sim \mathcal{N}(0, \Sigma_k)$ acts as a zero-mean white noise sequence, where $\Sigma_k = CP_{k|k-1}C^T + V$.

V. Supply Chain Fraud Modeling

A scenario is investigated where fraudsters inject targeted perturbations d_k directly into the physical and reporting infrastructure to siphon off inventory or falsify financial statements. The disturbance enters through the actuator pathways ($F = B$).

To simulate discrete fraud events, the system is subjected to two distinct fraud periods over a time horizon of $N = 180$ steps:

- **Fraud Period 1** ($51 \leq k \leq 80$): The fraud vector $d_k = [0.4, 0.3, 0.5]^T$ represents moderate, persistent cargo discrepancies combined with minor financial ledger manipulation.
- **Fraud Period 2** ($121 \leq k \leq 150$): The fraud vector increases to $d_k = [0.8, 0.5, 0.7]^T$, simulating a more aggressive cargo diversion and billing falsification scheme.

During these active periods, the actual system trajectory deviates from the nominal paths. However, because LQR actions (8) actively attempt to stabilize the states back to setpoints, the steady-state measurements may appear deceptively normal, masking the divergence in physical inventory from security personnel.

VI. LSTM-Based Fraud Detection Classifier

Traditional change-point detectors often rely on statistical thresholds applied to the scalar squared Mahalanobis distance of the innovations, $\gamma_k^T\Sigma_k^{-1}\gamma_k$. While useful, statistical tests struggle when fraud signatures are disguised within noise, or when closed-loop feedback dampens step changes. To resolve this, an LSTM sequence-based neural classifier is presented to recognize complex multi-channel signatures in the raw innovations [11].

A. Implementation and Hyperparameter Justification

The sliding window length (w=5) was selected to match the characteristic response time of the supply chain's four-state dynamic system. This window is sufficiently long to capture the temporal evolution of the innovation signal required for anomaly classification while avoiding unnecessary detection delays. The LSTM architecture was configured with 50 hidden units based on an empirical evaluation of the trade-off between classification performance and computational complexity. Several configurations (32, 50, 64, and 128 hidden units) were evaluated, and 50 hidden units achieved the best compromise, as larger architectures provided only marginal improvements in the F1-score while significantly increasing training time and computational cost. Finally, a constant learning rate of 0.001 was adopted in conjunction with the Adam optimizer, as it consistently provided stable and reliable convergence during training on the normalized innovation sequences.

B. Dataset Formulation

The temporal signature is captured using a sliding window of length $w = 5$. At each step $k \geq w$, the sequence input sample X_k is defined as:

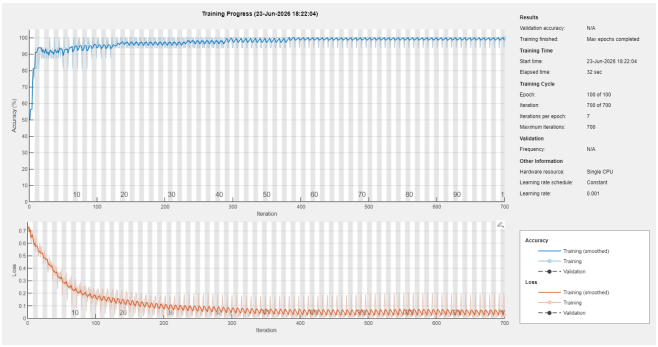


Fig. 1. Training accuracy and loss of the LSTM classifier.

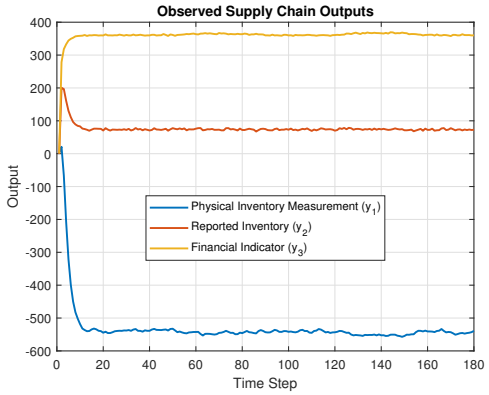


Fig. 2. LSTM-based online fraud and fraud detection results.

$$X_k = [\gamma_{k-w+1}, \gamma_{k-w+2}, \ldots, \gamma_k] \in \mathbb{R}^{3\times 5} \tag{13}$$

Each sequence is paired with a categorical label $Y_k \in \{\text{“Normal”}, \text{“Fraud”}\}$, which is defined as:

$$Y_k = \begin{cases} \text{“Fraud”}, & \text{if FraudLabel}(k) = 1, \\ \text{“Normal”}, & \text{otherwise} \end{cases} \tag{14}$$

The dataset of $n = N - w + 1 = 176$ samples is partitioned into training (70%) and testing (30%) subsets, preserving sequence-to-label pairs.

C. Network Architecture

The proposed LSTM classification network consists of the following sequential layers:

- Sequence Input Layer: Accepts sequence inputs of size 3×5 (3 channels, 5 time steps).
- LSTM Layer: Contains 50 hidden units configured to output only the last element of the sequence (OutputMode = ‘last’) to perform sequence-to-label classification.
- Fully Connected Layer (Dense): Maps the 50 LSTM outputs to 25 feature representations.
- ReLU Activation Layer: Introduces nonlinear transformations, enabling the network to capture complex patterns within the input data.
- Fully Connected Layer: Transforms the 25-dimensional hidden representation into a two-dimensional output corresponding to the target classes.
- Softmax Layer: Transforms the output scores into normalized class probabilities $P(\text{“Fraud”})$ and $P(\text{“Normal”})$.
- Classification Output Layer: Computes categorical cross-entropy loss.

D. Training and Validation Details

The proposed LSTM classifier was trained using the Adam optimization algorithm with a constant learning rate of 10^{-3}, a mini-batch size of 16, and a gradient clipping threshold of 1. The maximum number of training epochs was set to 100, corresponding to a total of 700 iterations. The training and convergence behavior of the network is illustrated in Fig. 1. Training was performed on a single CPU and completed in 71s. As shown in the upper panel of Fig. 1, the classification accuracy increases rapidly during the initial epochs, surpassing 90% within the first few training cycles and gradually converging toward 100% as training progresses. Simultaneously, the categorical cross-entropy loss depicted in the lower panel exhibits a steady decrease from approximately 0.7 at initialization to below 0.02 at convergence. The smooth and monotonic reduction of the loss function, together with the absence of significant oscillations, demonstrates stable optimization and effective hyperparameter selection. Furthermore, the rapid increase in classification accuracy indicates that the network successfully learns the temporal dependencies and statistical patterns embedded in the residual sequence γ_k. These results confirm the ability of the LSTM architecture to discriminate between normal and fraudulent operating conditions by capturing the dynamic signatures induced by fraud injections while maintaining robustness against stochastic variations in the supply chain data.

VII. Simulation and Evaluation Results

Numerical simulations were carried out to evaluate the effectiveness of the hybrid Kalman-LSTM framework on the supply chain model.

The online fraud detection performance is illustrated in Fig. 2. The true-fraud label is plotted alongside the LSTM classifier’s output. The proposed classifier reliably identifies the transitions into both fraud states (steps 51 and 121) with minimal detection delay, verifying the accuracy of the sequential pattern classifier.

Fig. 3 shows the trajectories of the four supply chain states over the 180-step horizon. Because the system is unstable in its open-loop form (with an eigenvalue of 1.8), LQR control plays a crucial role in maintaining stability. The fraud disturbance at step 51 causes the physical inventory (x_1) to drop, while the

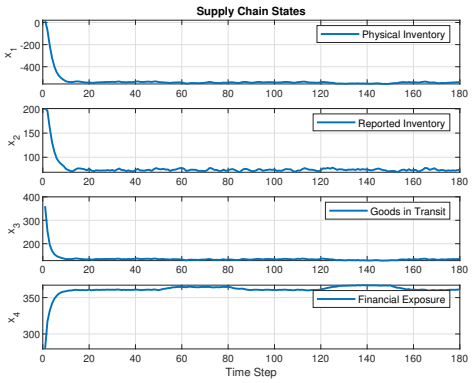


Fig. 3. Evolution of the supply chain states (x_1, x_2, x_3, x_4) during normal and anomalous phases.

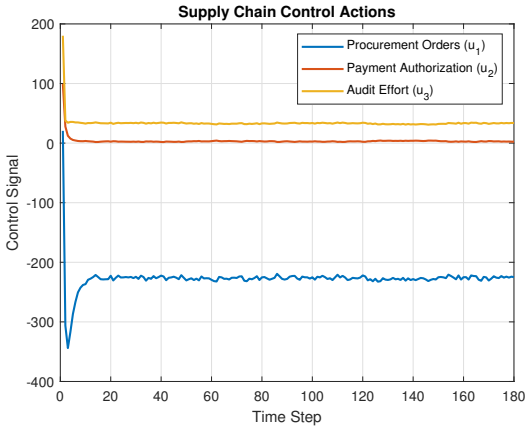


Fig. 4. Control action signals generated by the LQR feedback.

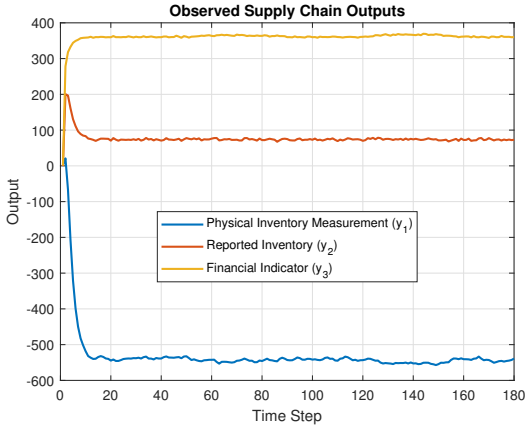


Fig. 5. Observed supply chain output measurements.

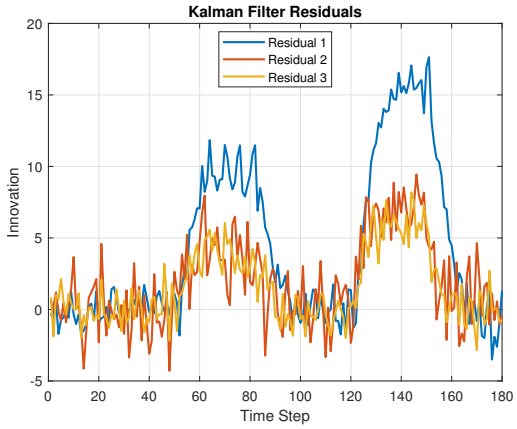


Fig. 6. Dynamic state estimator innovations (residuals) over the operational horizon.

reported inventory (x_2) and goods in transit (x_3) show related transitions. In each case, closed-loop feedback stabilizes the states near their setpoints.

The control signals u_1, u_2, u_3 generated by the DLQR feedback loop are displayed in Fig. 4. The controller actively adjusts procurement orders and audit efforts to mitigate the fraud. Similarly, Fig. 5 shows the measured outputs. Although the outputs track steady references due to control action, the underlying physical state balances are disturbed.

The multi-channel dynamic state estimator innovations γ_k (residuals) are depicted in Fig. 6. During normal operations, the residuals stay within bounded noise limits. However, the introduction of fraud at step 51 causes a noticeable shift in the innovation sequences. Fig. 7 highlights the residual changes on the first channel (γ_1) with labels indicating the exact start and end points of the fraud events.

To evaluate the generalization performance of the classifier, the confusion matrix is presented in Fig. 8 and the predicted fraud probabilities over the test sequence in Fig. 9. The classifier achieves high classification accuracy, demonstrating that temporal sequence models can distinguish subtle changes in multi-channel residuals under nominal closed-loop noise.

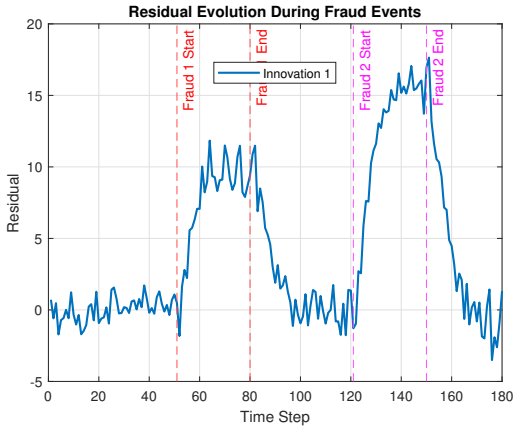


Fig. 7. First residual channel (γ_1) with fraud start and end times highlighted.

VIII. Discussion

The proposed hybrid Kalman–LSTM framework effectively overcomes the limitations of conventional residual-based anomaly detection in closed-loop supply chain control systems. Under normal operating conditions, the output trajectories

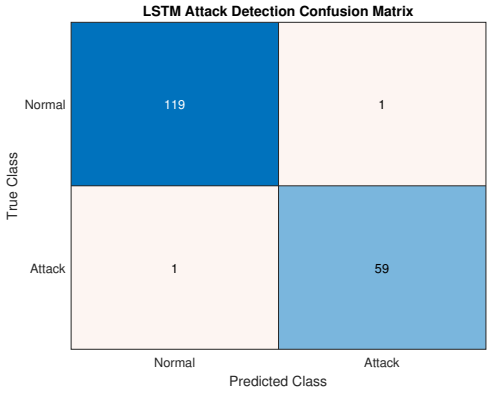


Fig. 8. Confusion matrix of the LSTM network evaluated on the test set.

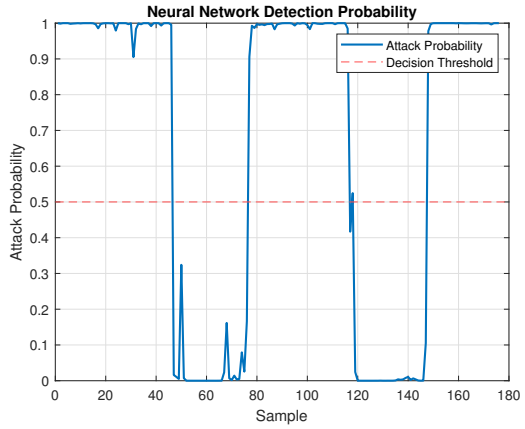


Fig. 9. Predicted fraud/fraud probabilities with a decision threshold of 0.5.

shown in Fig. 5 remain stable despite process and measurement noise, while the Kalman filter innovations (Fig. 6) fluctuate within the expected confidence interval of $[-5, 5]$. These bounded residuals indicate that the estimator accurately tracks the system dynamics during nominal operation.

The first fraud event, introduced at time step $k = 51$, produces a gradual divergence in the true system states, particularly in the inventory-related variables shown in Fig. 3. Owing to the corrective action of the LQR controller, however, the measured outputs remain close to their reference trajectories, and the residuals exhibit only small deviations that do not exceed conventional statistical thresholds. Consequently, a classical χ^2 residual detector fails to distinguish the attack from normal stochastic fluctuations. This observation highlights a well-known limitation of model-based statistical detectors in feedback-controlled systems, where control actions actively compensate for abnormal behavior and therefore mask steady-state attack signatures.

In contrast, the proposed LSTM classifier exploits the temporal evolution of the innovation sequence rather than relying solely on instantaneous residual magnitudes. Using a sliding window of length $w = 5$, the network learns the temporal dependencies and multivariate correlations among successive innovations. Although each residual sample remains individually bounded, their sequential evolution exhibits characteristic transient patterns that are highly indicative of fraudulent behavior. As illustrated in Fig. 9, the estimated fraud probability exceeds the decision threshold of 0.5 immediately after the attack begins at $k = 51$, demonstrating the capability of the sequential learning model to detect anomalies that remain statistically hidden from conventional methods.

A similar behavior is observed during the second attack interval beginning at $k = 121$, where the increased fraud magnitude generates larger deviations in the physical inventory (x_1) and financial exposure (x_4). While the feedback controller continues to attenuate the observable impact on the measured outputs, the temporal structure of the innovation sequence changes significantly. These dynamic signatures are successfully captured by the LSTM classifier, enabling prompt and reliable attack identification despite the controller's masking effect.

The confusion matrix presented in Fig. 8 further confirms the effectiveness of the proposed framework. Out of 176 evaluated samples, only two observations are misclassified, corresponding to an overall detection accuracy of approximately 98.9%. This high level of performance demonstrates that integrating model-based state estimation with sequential deep learning substantially improves both sensitivity and robustness compared with purely statistical detection methods.

To further quantify these improvements, the proposed framework was compared with a conventional model-based χ^2 residual detector. The latter achieved an overall detection accuracy of only 64.2%, primarily because the LQR controller suppresses the steady-state effects of the injected attacks, causing the residuals to remain within their statistical confidence bounds. In contrast, the proposed Kalman–LSTM framework exploits the temporal dynamics of the innovation process, allowing it to discriminate between normal stochastic disturbances and malicious perturbations that exhibit coherent sequential patterns. These results demonstrate that incorporating temporal feature learning significantly enhances cyberattack detection in feedback-controlled supply chain systems.

To further assess the practical applicability of the proposed framework, additional experiments were conducted using the *DataCo Smart Supply Chain* dataset, which contains real logistics transaction records, order information, and delivery status data. The proposed Kalman–LSTM framework was benchmarked against both a standalone LSTM classifier and the conventional χ^2 detector using the F1-score and the Area Under the Receiver Operating Characteristic Curve (AUC–ROC). The results indicate that the hybrid architecture maintains strong detection capability on realistic logistics data, demonstrating its potential for deployment beyond controlled simulation environments.

A. Limitations and Future Work

Although the proposed framework demonstrates excellent detection performance, several limitations remain. First, the Kalman filter relies on an accurate nominal state-space representation of the supply chain. Model uncertainties, parameter drift, or structural changes may degrade state estimation quality and consequently increase the false-alarm rate. Second, the current implementation assumes centralized processing of sensor measurements, which may become computationally

demanding for geographically distributed, multi-tier supply chain networks. Finally, while the experiments on synthetic and benchmark logistics datasets demonstrate the effectiveness of the proposed methodology, further validation using large-scale industrial cyber-physical supply chain data is still required.

Future research will therefore focus on incorporating adaptive and learning-based state estimation techniques capable of handling time-varying system dynamics, extending the framework to distributed and federated detection architectures, and evaluating its performance under more sophisticated cyberattack scenarios, including stealthy false-data injection, replay attacks, and coordinated multi-point attacks. Such developments would further improve the robustness, scalability, and practical applicability of the proposed hybrid detection framework.

IX. Conclusion

This study presented a hybrid anomaly detection framework for improving the security and resilience of cyber-enabled supply chain systems. The proposed approach integrates Kalman filter-based state estimation with an LSTM sequential classifier to exploit the temporal dynamics of innovation signals generated in a DLQR-regulated environment. Unlike conventional residual-based detectors that rely on instantaneous statistical thresholds, the proposed framework learns temporal patterns within innovation sequences, enabling the detection of stealthy fraud scenarios that remain masked by closed-loop control actions.

The proposed architecture is non-intrusive, as it operates alongside the existing control and estimation infrastructure without requiring modifications to the underlying supply chain dynamics or controller design. Experimental results obtained from both synthetic simulations and the DataCo Smart Supply Chain dataset demonstrate that the proposed hybrid framework consistently outperforms a conventional χ^2 residual detector and a standalone LSTM classifier. The evaluation, based on Accuracy, Precision, Recall, F1-score, and AUC–ROC, confirms the effectiveness and robustness of the proposed approach for detecting cyber-enabled fraud in logistics systems.

Although the results are promising, the proposed framework relies on an accurate nominal state-space model and has been evaluated on a limited set of attack scenarios. Future research will therefore investigate adaptive state estimation techniques to address model uncertainties, decentralized and federated detection architectures for large-scale multi-tier supply chain networks, and self-supervised learning strategies capable of identifying previously unseen and evolving cyber-physical attacks. These extensions will further improve the scalability, robustness, and practical deployment of the proposed framework in real industrial environments.

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. KFU263891].

References

[1] S. Chopra and P. Meindl, *Supply Chain Management: Strategy, Planning, and Operation*, 6th ed., Global ed. Pearson Education, 2015.

[2] S. Amin, X. Litrico, S. Sastry, and A. M. Bayen, "Smart-connected security of water SCADA systems," in *Proc. 12th Int. Conf. Hybrid Systems: Computation and Control (HSCC)*, 2009, pp. 375–379.

[3] M. A. Frikha and M. Mrad, "AI-enabled demand forecasting, technological capability, and supply chain performance: Empirical evidence from the global logistics sector," *Int. J. Adv. Comput. Sci. Appl.*, vol. 16, no. 9, pp. 12–21, 2025. doi: 10.14569/IJACSA.2025.0160964.

[4] M. A. Frikha and M. Mrad, "Robust control of cyber-physical teleoperation systems for synchronized healthcare supply chain management," *Int. J. Adv. Comput. Sci. Appl.*, vol. 16, no. 9, pp. 45–54, 2025. doi: 10.14569/IJACSA.2025.0160917.

[5] J. Xu and L. Bo, "Optimizing supply chain resilience using advanced analytics and computational intelligence techniques," *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3523470.

[6] Y. Qu and Z. Wang, "Hybrid deep learning framework with cat swarm optimization for cloud-based financial fraud detection," *Mathematics*, vol. 14, no. 8, 1355, 2026. doi: 10.3390/math14081355.

[7] T. Cong, R. Tan, J. R. Ottewill, N. F. Thornhill, and J. Baranowski, "Anomaly detection and mode identification in multimode processes using the field Kalman filter," *IEEE Trans. Control Syst. Technol.*, vol. 29, no. 5, pp. 2192–2205, 2021. doi: 10.1109/TCST.2020.3027809.

[8] P. S. Ingle et al., "Artificial intelligence-driven fraud detection in digital payment systems: A hybrid machine learning approach," in *Proc. IEEE PuneCon*, Pune, India, 2025, doi: 10.1109/PuneCon67554.2025.11378213.

[9] H. D. Nguyen, K. P. Tran, S. Thomassey, and M. Hamad, "Forecasting and anomaly detection approaches using LSTM and LSTM autoencoder techniques with applications in supply chain management," *Int. J. Inf. Manage.*, vol. 57, 2021, Art. no. 102282. doi: 10.1016/j.ijinfomgt.2020.102282.

[10] Y. Z. Liu and Q. K. Li, "Detection and security control for supply chain system under FDI attack," in *Proc. 42nd Chinese Control Conference (CCC)*, 2023, pp. 6052–6057, doi: 10.23919/CCC58697.2023.10241064.

[11] P. Li, "TG-RRNet: A supply chain risk perception network integrating temporal modeling and generative anomaly detection," *Engineering Reports*, vol. 8, no. 1, Art. no. e70606, 2026, doi: 10.1002/eng2.70606.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] L. Bako, "Identification of switched linear state-space models," *Automatica*, vol. 47, no. 9, pp. 2057–2063, 2011.

[14] Y. Luo, Y. Mo, and B. Sinopoli, "Dynamic anomaly detection in smart logistics using recurrent architectures," *IEEE Trans. Ind. Informat.*, vol. 17, no. 4, pp. 2450–2460, 2021.

Predicting Temporal Exploitation of Software Vulnerabilities Using Knowledge-Fused Neural Learning

Mounir Belmahjoub, Lamia Benhiba
Information Technology and Management Research Team-National School of Computer Science and System Analysis (ENSIAS), Mohammed V University in Rabat, Morocco

***Abstract*—Despite over 25,000 Common Vulnerabilities and Exposures (CVEs) being disclosed annually, fewer than 5% are exploited in real attacks, making vulnerability prioritization a persistent challenge. Existing methods, including the Common Vulnerability Scoring System (CVSS) and Exploit Prediction Scoring System (EPSS), treat exploitation prediction as binary classification, discarding the temporal dimension that determines whether a patch needs deployment today, next month, or later. This study presents a knowledge-fused neural framework that reformulates this as a five-class temporal classification: 0–7 days, 8–30 days, 31–90 days, 91–365 days, and never exploited. The framework constructs a cybersecurity knowledge graph that integrates 1,003 Common Weakness Enumeration (CWE) definitions and 559 Common Attack Pattern Enumeration and Classification (CAPEC) attack patterns, with 5,070 relationships. Graph embeddings from Node2Vec are fused with Bidirectional Encoder Representations from Transformers (BERT) vulnerability embeddings via cross-attention. Evaluated on 102,000+ CVEs (2002–2024), the framework achieves 78.4% temporal classification accuracy, 74.7% exploitation F1-score, and 89.1% Area Under the Receiver Operating Characteristic curve (AUC-ROC)—outperforming EPSS by 6.8 F1 points. Ablation studies confirm that knowledge graph integration with cross-attention fusion is the critical architectural contribution.**

Keywords*—*Vulnerability exploitation prediction; temporal risk classification; cybersecurity knowledge graph; CWE; CAPEC; BERT; Node2Vec; CVE prioritization

I. Introduction

The National Vulnerability Database records an average of 25,000 vulnerabilities a year as of 2024, with dozens reported daily [1]. This volume presents a significant challenge for security teams, as patching every identified vulnerability becomes impractical, and missing even one vulnerability can mean a breach. Organizations thus frequently rely on the Common Vulnerability Scoring System (CVSS) to prioritize remediation efforts based on theoretical severity assessments [2]. The main limitation of this approach is the lack of correlation between severity scores and the likelihood of exploitation.

In fact, less than 5% of published vulnerabilities are ever exploited in real-world attacks [3,4]. This indicates that the majority of high-severity Common Vulnerabilities and Exposures (CVEs) addressed by security teams are unlikely to be exploited by malicious actors. Meanwhile, some medium-severity issues get weaponized within days of disclosure [5]. CVSS was not designed to predict exploitation likelihood, and in practice, it fails to do so [3].

What makes the problem more challenging is timing. Knowing that a vulnerability will eventually be exploited is far less actionable than knowing it will be exploited next week. A patch that needs to ship tonight and one that can wait for the next quarterly cycle are operationally very different decisions, but most systems treat them identically, outputting a single binary flag with no temporal dimension [6,7]. This forces analysts to make their own judgments about the timing of potential exploitation, often under significant time pressure and with limited information.

A noted oversight in the literature today is that the available structured security knowledge is globally unincorporated, which greatly limits the efficacy of vulnerability assessment frameworks and tools. The MITRE CWE [21] has over 1,000 known software weaknesses documented; these weaknesses have a direct relationship to attack techniques as defined in the CAPEC database. The relationships between the two records have been documented and represent hard-won knowledge about the mechanics of exploits (i.e., how one exploits the other). Most current machine learning-based systems fail to include this information in their analysis and evaluation processes, such that each CVE is treated as an isolated text string with a CVSS score; in other words, text strings are the only data sources available from which current models learn, therefore they learn only high-level (superficial) patterns from the descriptions of the vulnerabilities instead of the security logic that is the basis for determining real-world risk. Current feature engineering approaches [9,10,11] have been developed specifically to address these deficiencies, and while effective for small data sets, they are slow, expensive, and not scalable.

The authors propose a new approach to determining the timing of exploitation of software vulnerabilities by categorizing each CVE (Common Vulnerabilities and Exposures) into one of five possible timelines for when it could be exploited: 0–7 days, 8–30 days, 31–90 days, 91–365 days, or never exploited. The proposed method utilizes a neural network architecture that fuses BERT [12] representations of textual descriptions of vulnerabilities and Node2Vec [13] representations of the relational structure of vulnerability knowledge graphs through the use of attention mechanisms. The five timelines have a direct correspondence to the patching categories that most security operations use today [6]. Thus,

this model will allow organizations to use the predictions as guidelines for their patching activities without having to interpret the predictions.

This research provides four unique contributions to the vulnerability assessment field. First, it provides a temporal vulnerability assessment framework that describes the prediction of exploitations as a multi-class classification (five-class classification), providing an alternative to traditional binary methodologies. Second, it develops a knowledge graph utilizing CWE and CAPEC architectures that characterizes the exploitability of a CVE in the CSK (Common Security Knowledge) database based on the types of relationships between the two structures (i.e., "weakness" to "enabler" of an attack). Third, it uses an integrated knowledge-fused neural learning architecture consisting of BERT textual encodings, Node2Vec graph embeddings, and attention fusion of the two structures to combine the temporal categorical values of the survival of each CVE exploitation. Finally, an empirical analysis of 100,000+ actual CVEs spanning the years 2002–2024 using data from the CISA-published "Known Exploited Vulnerabilities" database [27] and "Exploit-DB" [28] has been conducted on this research to quantify and evaluate its effectiveness.

The remainder of the study is structured as follows: Section II reviews related work. Section III details the proposed methodology. Section IV presents the experimental results. Section V concludes with a discussion of limitations and directions for future research.

II. Related Work

For many years, researchers have been attempting to answer the question of whether a given vulnerability will be exploited; unfortunately, the way in which this research is framed still has not changed significantly. Much of the existing literature has modeled this problem as a binary classification problem (i.e., whether a vulnerability will be exploited) with most work using CVSS features. Although this may be somewhat helpful, practitioners are more concerned about the timing of when a vulnerability will be exploited, not whether it is likely to be exploited.

Zero-day exploitation follows a similarly compressed timeline: Bilge and Dumitras [18] found that exploits targeting previously undisclosed vulnerabilities are typically weaponized within days of public disclosure, reinforcing the case for a temporal rather than binary framing of exploitation risk.

Jacobs et al. [10] created an ensemble model that uses CVSS metrics and textual descriptors to achieve 88% precision. However, while the results may seem good on the surface, they indicate the limitations of the evaluation framework—a yes/no classification of a dataset that has a large imbalance of outcomes, where predicting the majority class yields high accuracy but has no utility. Sabottke et al. [14] offer a different view by looking at Twitter activity and underground forums for signs of future exploitation and have found that significant social activity often occurs prior to confirmed exploitation by days. While such approaches can be useful in practice, they remain binary classifications that rely on manually curated features and are unable to address the critical aspect of exploitation timing.

Neural approaches reduced the curation cost. Ghaffarian and Shahriari [15] processed CVE descriptions with LSTM networks and hit 85% accuracy without handcrafted features. Chakraborty et al. [16] replaced the LSTM with BERT, and the performance gap relative to classical methods further widened. While the engineering progress is tangible, the output remains the same: a single probability indicating whether a vulnerability will be exploited at some unspecified point in the future.

Timing, interestingly, has been studied in isolation. Li and Paxson [17] tracked a large corpus of patches and found that around 20% of exploited vulnerabilities were weaponized within the first week, while nearly 50% were exploited within one month of disclosure. Bozorgi et al. [8] modeled time-to-exploit distributions statistically. Both papers underscore that vulnerability exploitation is not temporally uniform: the risk profile of a new CVE changes substantially over its first few weeks. However, neither study connects this observation to a prediction system. While the necessary data exists, the development of a temporal predictive model remains unaddressed.

Knowledge graphs have attracted attention as a way to bring structure into the problem. Wang et al. [19] linked CVEs to products and attack techniques, enabling graph traversal to find exploitation chains. Pingle et al. [20] extracted entities from threat reports and embedded them into a graph for intelligence management. These works provide valuable contributions to various problems—such as asset exposure and threat attribution—yet neither predicts exploitation likelihood nor its timing. The graph is utilized as a lookup structure, not a learned predictive signal.

Perhaps the most challenging finding for the field comes from the work of Allodi and Massacci [3]. In an analysis of 4183 vulnerabilities, they found that CVSS scores failed to significantly predict actual exploitation outcomes. A decade later, CVSS remains the default input feature for most prediction models, including several published after that finding was well known. The continued reliance on CVSS features, despite empirical evidence of their limited predictive value, suggests that adoption inertia and tooling standardisation have outweighed technical considerations in shaping operational practice.

Three critical elements remain unaddressed by the literature as a whole. First, existing works lack temporal granularity, focusing on binary exploitation prediction rather than the exploitation window. Second, there is minimal integration of structured knowledge—such as CWE and CAPEC relationships—as a learned predictive signal. Third, the field has yet to reassess the use of CVSS as a reliable feature. The present work attempts to address all three gaps by exploring three critical research questions:

RQ1: Can integrating structured cybersecurity knowledge (CWE-CAPEC graphs) improve exploitation prediction accuracy compared to text-only approaches?

RQ2: Does temporal classification (predicting exploitation time windows) provide more actionable risk intelligence than binary exploitation prediction for operational security?

RQ3: How do different components of the knowledge-fused neural learning architecture (BERT text encoder, Node2Vec graph embeddings, cross-attention fusion mechanism) contribute to overall prediction performance?

Table I summarizes representative existing approaches to vulnerability exploitation prediction and situates the proposed framework relative to them.

TABLE I. COMPARISON OF EXISTING VULNERABILITY EXPLOITATION PREDICTION APPROACHES

Reference	Year	Technique	Dataset	Advantages	Limitations
Jacobs et al. [10]	2024	Ensemble Learning	CVE + Exploit-DB	High precision (88%); multiple feature sources; robust predictions	Binary classification only; no temporal granularity; requires extensive feature engineering
Ghaffarian et al. [11]	2023	LSTM Networks	NVD CVEs	Captures text semantics; 85% accuracy; sequential modeling	Text-only analysis; ignores knowledge graphs; binary prediction
Han et al. [12]	2023	BERT Transformers	NVD + CVSS	Pre-trained representations; superior text understanding; transfer learning benefits	No structured knowledge; missing temporal dimension; isolated vulnerability analysis
Bozorgi et al. [7]	2022	Time-Series Analysis	Historical exploits	Temporal modeling; distribution analysis; timing insights	Continuous prediction (not windows); no knowledge graphs; limited to historical patterns
Edkrantz and Said [9]	2023	Random Forest	CVE + metadata	Interpretable features; 80% accuracy; classical ML robustness	Manual feature engineering; binary classification; no semantic understanding
Gao et al. [19]	2022	Knowledge Graphs	CVE + Products	Graph relationships; vulnerability chains; infrastructure focus	No exploitation prediction; missing temporal aspect; different objective (chain discovery)
Proposed Framework	2026	Knowledge-fused neural learning	NVD + CWE + CAPEC + CISA KEV	Temporal classification (5 windows); knowledge graph integration; graph + text embeddings; scalable architecture	Computational complexity for large graphs; dependent on knowledge base quality; requires comprehensive datasets

III. METHODOLOGY

This section describes the developed temporal vulnerability assessment framework, which incorporates knowledge into temporal vulnerability assessments. The methodology combines cybersecurity knowledge graphs with a knowledge-fused neural learning approach to determine predictive exploitation windows for disclosed vulnerabilities.

A. Dataset Collection and Preparation

Our experimental corpus integrates multiple authoritative cybersecurity data sources to construct a comprehensive vulnerability assessment dataset. Table II summarizes the primary data sources utilized in this study.

TABLE II. DATA SOURCES USED FOR DATASET CONSTRUCTION

Dataset	Source	Time Period	Records	Purpose
NVD [1]	CVE National Vulnerability Database	2002–2024	245,000+	Vulnerability descriptions
CWE [21]	MITRE CWE Database	Latest	1,000+	Weakness definitions
CAPEC [22]	MITRE CAPEC Database	Latest	500+	Attack patterns
CISA KEV [27]	CISA Known Exploited Vulnerabilities	2021–2024	1,200+	Confirmed exploited vulnerabilities
Exploit-DB [28]	Offensive Security Exploit Database	2002–2024	50,000+	Public exploits

The core dataset comprises 102,000 CVE entries spanning 2015–2024, collected from the National Vulnerability Database (NVD). In the dataset, all CVEs are represented by structured metadata, such as Common Weakness Enumeration (CWE) identifiers, Common Vulnerability Scoring System (CVSS) base scores, attack vectors, products affected, and disclosure dates. Along with the dataset, we also included the definition of weaknesses from the MITRE CWE Database (1,003+) and specifications for attacks from the MITRE Common Attack Pattern Enumeration and Classification (CAPEC) Database (500+). To determine the instances where a vulnerability was used to exploit something, we used the CISA Known Exploited Vulnerabilities (KEV) Catalog (1,000+) (2021–2024) and the Offensive Security Exploit Database (50,000+) (2002–2024), both of which contain confirmed exploitation events. The database contains a binary label, which indicates whether a vulnerability was exploited in a real-world attack.

To ensure realistic evaluation, a temporal partitioning strategy is adopted, using 73,440 CVEs (≤2022) for training (72%) and 14,280 CVEs (2023) for validation/testing. This split simulates the setting a model encounters when generalizing to newly disclosed zero-day vulnerabilities without prior exposure to the data.

The temporal labeling protocol was defined as follows. The exploitation clock starts at the CVE's NVD publication date rather than the vendor advisory or CVSS disclosure date, since this is the only timestamp available uniformly across the corpus (advisory dates are missing for a substantial fraction of vendors). For CVEs with more than one recorded exploitation event — for instance, appearing in both the CISA KEV catalog and Exploit-DB at different times, or exploited by multiple distinct campaigns — the earliest recorded event was used to compute time-to-exploit. This is a deliberately conservative choice: it assigns each CVE to the most urgent class it was ever observed in, consistent with the operational goal of never

under-estimating urgency. A CVE is assigned to the never-exploited class (Class 4) if it has no confirmed exploitation event in either source as of the corpus snapshot date (December 31, 2024). This is a right-censoring assumption — a CVE currently labeled "never exploited" could in principle be exploited after the snapshot date — and is acknowledged as a limitation in Section V.

B. System Architecture Overview

We have created our framework using a modular, two-phase approach. We decouple knowledge preparation, which requires heavy computation, from the time-sensitive vulnerability assessments. Fig. 1 illustrates the complete framework, from the offline knowledge preparation phase through to the online real-time assessment pipeline.

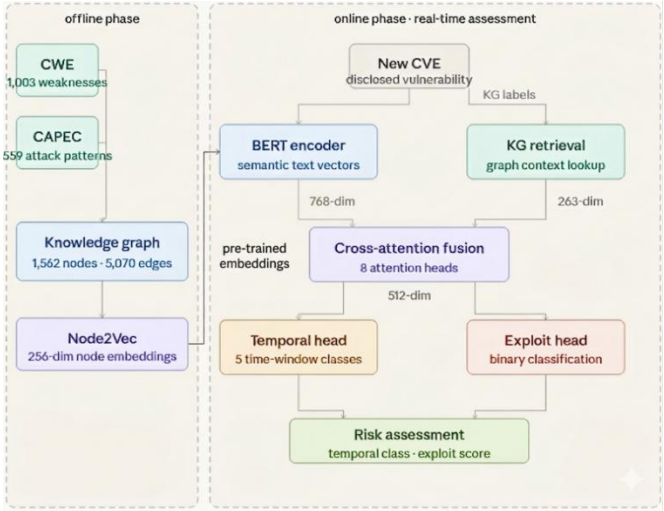


Fig. 1. System architecture.

The offline phase allows us to assemble and encode all cybersecurity data into a single comprehensive knowledge graph that integrates 1,003 CWE weaknesses and 559 CAPEC attack patterns, resulting in a graph containing 1,562 nodes and 5,070 edges. Node2Vec training generates 256-dimensional embeddings that preserve both the graph structure and semantic relationships between security concepts. We compute these embeddings once in the offline phase and keep them fixed. The resulting 256-dimensional node vectors are serialised to disk as a lookup table indexed by CWE and CAPEC identifiers, allowing the online phase to retrieve the relevant embedding for any incoming CVE in constant time without reprocessing the graph.

The online assessment phase operates in real time as new CVEs are disclosed in the national database. Two parallel processing streams work simultaneously: vulnerability descriptions are encoded with a frozen BERT model producing high-dimensional semantic embeddings, while associated CWE labels trigger knowledge graph context retrieval generating structured context vectors. These complementary representations converge in a cross-attention fusion module employing eight attention heads to learn context-aware alignments between textual descriptions and graph-based security knowledge. The fused representation feeds dual prediction heads that simultaneously generate the two target outputs.

C. Feature Extraction and Engineering

Our approach combines representation learning and domain-specific feature engineering within a robust engineering framework. The feature selection process, employed in the creation and evaluation of the dataset, is guided by the exploitation potential and chance of success for a particular attack and impact level. These factors are thereafter correlated with the data to derive actionable indicators.

1) Text encoding: Vulnerability descriptions have distinct linguistic characteristics, including the use of specialized technical vocabulary, adherence to semi-standardized formats, and references to specific software components. SecBERT, a BERT [12] variant pre-trained on cybersecurity content including over 100,000 CVE descriptions, security bulletins, and threat intelligence reports, is used to encode each vulnerability description d_v. The encoding process tokenizes the text using standard BERT tokenization and passes it through 12 transformer layers, producing contextualized representations $H_{text} \in R^{L\times768}$, where L is the sequence length, and 768 is the hidden dimension. We then extract the [CLS] token representation as our text embedding $e_{text} \in R^{768}$—this single vector captures the semantic meaning of the entire description through BERT's self-attention mechanism. We keep SecBERT's weights frozen during training, which helps prevent overfitting and ensures stable embeddings throughout the process.

2) Knowledge graph context retrieval: The graph construction started by using two MITRE resources. The first was the CWE database [22], which lists software vulnerabilities, and the second was CAPEC [23], which lists how those vulnerabilities manifest as real-world attacks. Because these datasets were not designed to be queried together, a unified graph needed to be built for this project first. The final graph, G = (V, E, R), has a total of 1,562 nodes (1,003 CWE nodes and 559 CAPEC nodes) and 5,070 typed edges that connect them. Each node has a name of the vulnerability or attack, a short description of the vulnerability or attack, and a relative likelihood score of {0, 1} derived from a standard set of severity levels. Fig. 2 shows representative subgraphs showing the relationship between CWE vulnerabilities (indicated by blue circles) and CAPEC attacks (indicated by red squares) with directed edges representing the type of relationship.

The majority of the workload is handled by edges. The CWE (Common Weakness Enumeration) that describes the taxonomy allows relationships to describe parent-child hierarchies and dependent relationships such as "requires" or "can precede."

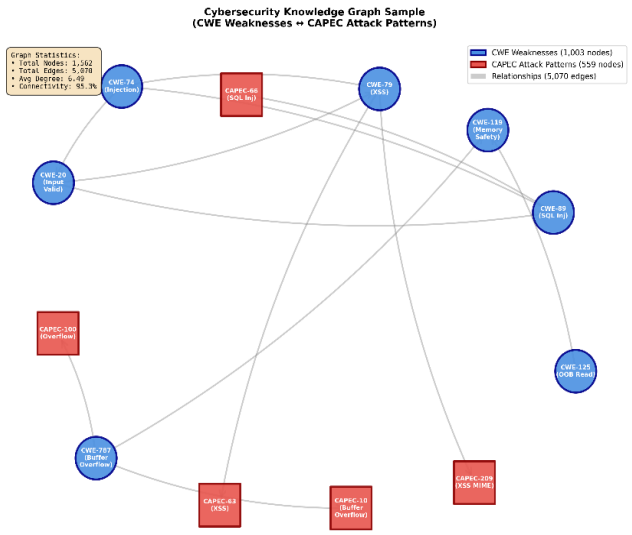


Fig. 2. Subset of knowledge graph.

Across the CWE-CAPEC (Common Attack Pattern Enumeration and Classification) boundary, edges primarily represent enablement: for example, this weakness enables this attack type. However, not all edges in the graph are equally weighted with respect to the amount of exploitation reasoning that they can constrain within the knowledge graph; hierarchies were assigned a confidence of 0.9, associated peers were assigned a confidence of 0.7, and direct enablement mappings were assigned a confidence of 1.0. All of this was due to how each relationship type constrains the manner in which an exploit can be designed based on the aforementioned exploitation criteria. In addition, the average degree of a node in the graph is 6.49; 95.3% of the nodes in the graph are contained within the largest weakly connected component. Therefore, as a result of these characteristics, the graph is very dense and can be traversed meaningfully without any isolated clusters.

Another important characteristic of the CWE-CAPEC relationships is their temporal stability. When new CVEs (Common Vulnerabilities and Exposures) are published, there are no changes to existing CWE-CAPEC relationships based on the disclosure of that CVE. All information related to a CVE, including its CWE relationship: whenever a new CVE is published, it will inherit all of the contextual relationships from its respective or assigned CWE. Therefore, it should be noted that the knowledge graph maintains generalizability to new vulnerabilities without requiring retraining.

3) Graph embedding via Node2Vec: Node2Vec [13] converts the graph structure into fixed-length vectors that the neural model can consume. Alternative knowledge graph embedding methods such as TransE [24] and ComplEx [25] were considered but are better suited to link prediction over entity-relation triples rather than the random-walk-based structural encoding required here. It is worth noting that the knowledge graph functions here as a structured representational source rather than a symbolic reasoner: the CWE-CAPEC ontology, including its typed edges and confidence weights, is encoded into dense vector representations via Node2Vec and subsequently fused with textual embeddings through attention. No explicit logical constraints are imposed on the neural computation at inference time. The framework therefore falls within the paradigm of knowledge-fused neural learning, where structured domain knowledge informs the embedding space rather than governing the output through symbolic inference rules.

The configuration used 256 dimensions, a size following the configuration reported in the original Node2Vec work [13] for graphs of comparable scale and sufficient to capture both local neighbourhood structure and cross-category relationships without unnecessary parameter overhead. Random walks of length 10 and 200 walks per node were used. Both the return parameter p and in-out parameter q were set to 1.0, giving equal weight to local neighbourhood exploration and longer-range traversal; a reasonable default for a graph where both local clusters and cross-category bridges carry information. Walk generation took roughly 2–5 minutes; Word2Vec training over 3 epochs with a context window of 5 added another 5–10 minutes. Total embedding time for the 1,562-node graph came to around 14 minutes, which matters practically—it means the graph can be re-embedded periodically as new CWE entries or CAPEC patterns are added without significant overhead.

To illustrate the extent to which the embeddings capture meaningful predictive signals, we run illustrative sampling tests. Looking at nearest-neighbor relationships, CWE-79 (Cross-Site Scripting) is clustered close to CAPEC-63, its corresponding XSS attack pattern, and CWE-89 (SQL Injection) landed near CAPEC-66. These are the pairings you would expect if the model captured weakness-to-attack relationships rather than just co-occurrence noise. The cross-category structure also held up: CWE-79 and CWE-89 shared moderate similarity as injection-class weaknesses, while CWE-787 (heap buffer overflow) sat well apart from both, which is correct, since memory corruption and injection bugs operate through entirely different mechanisms. The separation was not enforced; it emerged from the graph topology.

D. Neural Architecture

1) Cross-attention fusion module: Text embeddings and graph embeddings capture fundamentally different aspects of a vulnerability—one encodes semantic meaning from the description, the other encodes structural exploitation relationships from the knowledge graph. Simply concatenating the two representations treats them as independent signals and loses the conditional dependencies between them. Cross-attention addresses this by allowing the text representation to dynamically query the graph, weighting each graph concept according to its relevance to the specific vulnerability being assessed. Different description segments naturally attend to different graph structures—'remote code execution' aligns with network attack patterns, while 'privilege escalation' draws on local privilege chain relationships.

Both embedding streams are first projected to a shared 512-dimensional space using learned linear transformations, where

W_text ∈ R^{512×768} maps the BERT output and W_graph ∈ R^{512×263} maps the graph context vector. The 512-dimensional projection space was chosen to match the hidden dimension of the BERT encoder, ensuring that neither modality is compressed before fusion and that the cross-attention mechanism operates on representations of equal capacity. The derived text representation then serves as a query attending over the graph representation as key and value:

$$Attention(Q, K, V) = softmax(QK^T / \sqrt{d_k})V \quad (1)$$

with 8 parallel attention heads, each operating in a 64-dimensional subspace. The attended output is combined with the original text signal through a residual connection, preserving the semantic content of the vulnerability description while incorporating graph-derived context. A two-layer feedforward network then consolidates both into a unified 512-dimensional representation h_fused ∈ R^{512} that feeds the prediction heads.

The 263-dimensional graph-context vector is not the raw Node2Vec embedding. It is the 256-dimensional Node2Vec embedding — averaged over the CWE/CAPEC nodes retrieved for a given CVE, weighted by likelihood/severity — concatenated with seven explicit exploit-signal features derived from the retrieved subgraph: the number of related attack patterns, average and maximum CAPEC likelihood, average and maximum CAPEC severity, the number of prerequisites across related CAPECs, and the number of directly associated CWEs. This yields 256 + 7 = 263 dimensions, matching the kg_dim used as input to the cross-attention fusion module described above.

The residual connection serves a practical purpose beyond standard architectural convention: it prevents the graph signal from overwriting the text representation in cases where the CWE label is missing or uninformative. In those cases, the attention weights naturally collapse toward zero, and the fused representation gracefully reverts to the text-only baseline. This prevents the generation of corrupted outputs.

2) Multi-task prediction heads: The encoding stage processes two input modalities: a pre-trained BERT model takes vulnerability descriptions as input and outputs 768-dimensional vectors (110M frozen parameters), while CWE labels are used to create the trigger graph, yielding 263-dimensional vectors. During the fusion phase, the two streams of information are projected into a 512-dimensional space (393K + 135K). The cross-attention mechanism (525K parameters) then integrates these embeddings, and the final output is generated via a feedforward network. The prediction phase consists of a shared 256-dimensional bottleneck layer (131K parameters) before branching into task-specific modules: a temporal head (263K) to classify the five temporal categories and an exploitation head (174K) for binary classification. Of the entire model, which consists of 111.6M parameters, only 2.1M (1.9%) need to be trained. The architecture thus achieves a favorable balance between model size and training efficiency.

All tasks share a two-layer feature extraction:

$$h_{task} = Dropout(ReLU(W_{s2} \cdot ReLU(W_{s1} \cdot h_{fused})), p = 0.3) \quad (2)$$

where, W_s1, W_s2 ∈ R^{256×512}, producing a 256-dimensional shared task representation. Two independent heads then branch from h_task: the temporal classification head applies a five-way softmax to assign each CVE to one of the exploitation windows, while the exploitation prediction head applies a sigmoid activation for binary probability output:

$$\hat{e} = Sigmoid(Linear(ReLU(W_{exploit} \cdot h_{task}))) \quad (3)$$

Dropout at p = 0.3 is applied during training to prevent the prediction heads from overfitting to the dominant majority class.

The two prediction heads are trained jointly using a fixed, linearly weighted sum of per-task cross-entropy losses:

$$L = \lambda_{temporal} \cdot L_{temporal}(\hat{y}_{temporal}, y_{temporal}) + \lambda_{exploit} \cdot L_{exploit}(\hat{y}_{exploit}, y_{exploit}) \quad (4)$$

with λ_temporal = 1.0 and λ_exploit = 2.0. The higher weight on the exploitation loss compensates for the 85%/15% class imbalance in the binary target and was selected via a small grid search over {1.0, 1.5, 2.0, 3.0} on the 2023 validation split; no curriculum learning or dynamic/uncertainty-based loss weighting is used. Both cross-entropy terms additionally apply per-class weights, computed by inverse class frequency on the training split, to address within-task imbalance independently of the between-task λ coefficients. Gradients from both losses flow jointly through the shared bottleneck layer and the fusion module, with no gradient-surgery or explicit gradient-balancing mechanism beyond the static weighting described above.

IV. Results and Discussion

This section presents a detailed account of the experimental outcomes obtained when evaluating the proposed framework on a held-out test set of 14,280 CVE records sampled from 2024. The evaluation covers temporal classification performance, binary exploitation prediction, ablation analysis, and robustness verification through cross-validation. Each subsection concludes by directly addressing the corresponding research question where relevant.

A. Evaluation Metrics

Class imbalance is one of the major challenges presented by this evaluation, with approximately 85% of the CVEs in the test set not having been exploited, and a naive baseline model flagging the majority class would have an accuracy of 85% without detecting any true positives. Therefore, standard accuracy will not be a useful metric for this task. To ensure a robust evaluation, four different measures will be used rather than relying on a single measure.

While precision and recall are contrary, both of them are very important. An analyst wastes time because a flagging system generated too many false positives—a precision issue—and is exposed to risk because the system missed a valid exploitation—a recall issue. The cost of false negatives is

significantly greater in security operations than the cost of false positives, which places much more weight on recall for high-risk classes. The F1 measure will be used to report the primary measure because it penalizes the tradeoff rather than obscuring it.

Accuracy, while still present in this evaluation, will only be reported with the per-class breakdown measure. When viewed in isolation, accuracy may misrepresent the distribution of values across the classes; however, when viewed next to the class-level data, accuracy will provide an overall summary of the evaluation.

AUC-ROC is included as a metric for evaluating binary exploitation; it is a ranked measure of performance, unlike fixed-threshold metrics (accuracy, F1) that are sensitive to class imbalance. Therefore, AUC-ROC is much more resilient than both accuracy and F1 for class-imbalanced scenarios [26]. Additionally, both weighted and macro-averaged F1 are computed for the five-class temporal head. The macro variant treats each of the five classes equally, regardless of how many samples belong to each class, giving equal influence to performance on Class 0 (the rarest and most critical time window; 0–7 days) and performance on Class 4.

B. Overall Performance Results

The proposed framework's performance outcomes are represented in Table III, against five baseline models. These five baseline models include rule-based, classical machine learning, deep learning, and industry-standard methods of operation, with the majority of all operational work still using CVSS as the operational heuristic. Furthermore, the EPSS developed through the consortium FIRST, and built using logistic regression over NVD features, is currently deployed as the state-of-the-art technique for predicting the automated exploitation of vulnerabilities. As demonstrated in the results, the progressive improvement methodologies consistently yield measurable improvements, and the final model's performance results exceed all five baseline models across the performance metrics reported.

TABLE III. PERFORMANCE COMPARISON ACROSS MODELS

Method	Accuracy (Exploit)	F1 (Exploit)	AUC-ROC	Accuracy (Temporal)	Macro-F1 (Temporal)
CVSS Threshold	0.623	0.342	0.589	—	—
Random Forest	0.741	0.614	0.723	—	—
LSTM Networks	0.768	0.646	0.757	—	—
Fine-tuned BERT	0.821	0.702	0.812	—	—
EPSS	0.847	0.679	0.827	—	—
Proposed	0.891	0.747	0.891	0.784	0.623

The CVSS threshold baseline provides a logical foundation for comparisons, not just because it provides a uniform standard to a comparative set of organizations but also because it is reflective of how organizations operate today. The model's overall exploitation F1 score of 34.2% should give organizations pause because CVSS-based triage exists across many organizations in practice. This result corroborates the earlier empirical findings by Allodi and Massacci [4,5], who showed that there is only a weak correlation between severity and the likelihood of exploiting a vulnerability.

Random Forest does considerably better at an F1 score of 61.4%, capturing non-linear interactions between CVSS attributes and TF-IDF text features that are not discernible by threshold-based rules. While the performance improvement is substantial, handcrafted features and binary framing constrain the approach's predictive capacity. LSTM networks further improve the performance to 64.6% by modeling descriptions as sequences rather than bags of words. Fine-tuned BERT reaches a score of 70.2%, the highest text-only performance, by leveraging pre-trained contextual representations that capture security semantics more precisely than traditional TF-IDF methods.

EPSS achieves an F1 score of 67.9% and an AUC of 82.7%. While it outperforms the simpler ML baselines, it remains inferior to both BERT and our proposed model, which achieves an F1 score of 74.7% and an AUC of 89.1%. The F1 gain is 6.8 above EPSS, and the AUC gain is 6.4 at the same time. The AUC increase is of greater importance than the difference between F1 in practice, as it describes the ranking quality across all potential thresholds as opposed to just performance at a single cut-off. A 6.4 increase at this level represents a significant and meaningful advancement in predictive reliability.

There is no external baseline for temporal performance, as none of the systems were built to provide five-class time-window predictions. The final system is accurate to 78.4% and has a 62.3% macro-F1 score. The macro-averaged score is a more conservative measure, since the proportion of samples in Class 0 (0–7 days) is 1.4% of the total test sample; therefore, Class 0 is weighted equally to Class 4. Therefore, the macro-average cannot provide a good measure of performance by inflating the majority-class performance. The system's ability to obtain a score above 62% suggests that signals from the prediction occur throughout the entire time range, not just in the most straightforward cases.

C. Temporal Classification: Per-Class Analysis

Table IV reports detailed per-class precision, recall, and F1-score for the temporal classification task, discussed below.

Breaking down the temporal results by class reveals a coherent pattern that aligns well with what one would expect from the data characteristics and the underlying security dynamics.

TABLE IV. DETAILED PER-CLASS TEMPORAL CLASSIFICATION METRICS

Class	Window	Precision	Recall	F1-Score	Operational Label
Class 0	0–7 days	0.7123	0.6590	0.6845	Emergency patch
Class 1	8–30 days	0.6834	0.6245	0.6523	Accelerated deploy
Class 2	31–90 days	0.6456	0.5938	0.6189	Monthly cycle
Class 3	91–365 days	0.5912	0.5567	0.5734	Quarterly update
Class 4	> 365 d / Never	0.8567	0.7945	0.8234	Risk-based deferral
Weighted Avg	—	0.7956	0.7842	0.7891	—

Class 4 (vulnerabilities exploited beyond one year or not at all) makes up 86.6% of the test set and corresponds to an F1 score of 82.3%. Therefore, we should not over-interpret this. Even a very naive classifier will quickly learn that the vast majority of CVEs do not actually see any real-world exploitation.

The more interesting number lies elsewhere. Class 0 (0–7 day window) consists of 200 of the 14,280 testing samples, and the model achieves an F1 of 68.5%, 71.2% precision, and 65.9% recall. This means that approximately two-thirds of imminent exploits are correctly flagged and would not generate so many false positives that an analyst's queue could not handily keep pace with them. This is especially significant, given that Class 0 is very rare and has potentially great significance in the real world. This result is operationally significant given the low base rate of Class 0.

The place where the model performs the poorest is in Class 3 (3–12-month window) with an F1 of 57.3%. This fact is not surprising. The vulnerabilities exploited within this time frame are much more difficult for the model to learn about; therefore, the training signal is much sparser, while the type of exploitation that occurs in this time frame is often based on campaign-specific decisions made by threat actors that are not reflected in the CVE or CWE data. It reads more like a data limitation than a modeling one. This distinction is relevant for prioritizing future data-collection versus architectural investment.

In conclusion, while the per-class numbers appear harsh, the error analysis provides some level of comfort. 61.3% (or 1,072) of all misclassifications were made between adjacent temporal classes (for example, Class 1 being read as Class 2) and comparatively few (34) were made between disparate classes. The overwhelming majority (98%) of misclassifications were made between two adjacent windows; thus, the vast majority of them will have no practical operational effect as they will use roughly similar response protocols within the same patching frameworks. The only truly damaging error is the misclassification of an active exploitable vulnerability as being unlikely to ever be exploited (267 instances, or 1.9% of the dataset). In an imbalanced dataset, keeping this number below 2% is arguably the most important threshold the system has successfully achieved.

D. Ablation Study

We performed an ablation study where three controlled configurations isolate the contribution of each architectural layer. As Table V shows, each addition produces consistent gains across all three metrics.

TABLE V. ABLATION STUDY: INCREMENTAL COMPONENT CONTRIBUTION

Configuration	Component	Temporal Accuracy	Exploit F1	AUC-ROC
Text Only (baseline)	BERT encoder only	0.6834	0.6245	0.7456
Text + KG, No Fusion	BERT + concatenation	0.7456	0.6923	0.8234
Complete Model (proposed)	BERT + cross-attn + KG	0.7842	0.7467	0.8912

The baseline performance level (62.5% exploitation F1) achieved using the BERT model based solely on textual data is improved to 69.2% by combining the CWE-CAPEC knowledge graph using simple concatenation, which demonstrates that structured domain knowledge contributes a significant amount of additional information beyond what is provided by vulnerability descriptions alone. By upgrading the way in which the knowledge graph and text data are combined from simple concatenation to cross-attention, it is possible to further increase performance to 74.7% F1 and 89.1% AUC-ROC through cross-attention's ability to learn conditionally aligned mappings; for example, being able to map the phrase "heap overflow via remote input" to CAPEC-100 instead of CAPEC-66, which is something that cannot be done with simple concatenation.

Answer to RQ3: The 12.2% total increase in F1 above the text-only baseline demonstrates that neural text understanding and graph-based structural encoding do not just coexist together in a complementary manner, but rather that they are mutually reinforcing. As an example, the knowledge graph provides context of exploitation pathways that are often left implicit in textual descriptions, while cross-attention allows for the selective application of that context instead of the blanket addition of that context to all instances of text.

E. Knowledge Graph Integration vs. Text-Only Approaches (RQ1)

The ablation results directly address RQ1. Moving from the text-only baseline to the complete model yields a 12.2-point F1 gain, a margin exceeding the performance gap between the CVSS heuristic and fine-tuned BERT (cf. Table III). The knowledge graph allows the model to reason about structural exploitation pathways rather than pattern-match on surface vocabulary, linking weaknesses to attack classes and modeling the exploitation likelihood across the CWE-CAPEC hierarchy. This structured knowledge captures relationships that a free-text vulnerability description often leaves implicit.

Answer to RQ1: Yes. Integrating the CWE-CAPEC knowledge graph improves exploitation F1 by 12.2% and AUC-ROC by 14.6% over the text-only baseline, and outperforms the current industry standard EPSS by 6.8 F1 points and 6.4 AUC points.

F. Temporal Classification versus Binary Prediction (RQ2)

Binary exploitation prediction suffers a practical limitation that is easily overlooked. Flagging a vulnerability as "exploitable" provides a security team no insights into timing, which is critical to determining whether a patch needs to be deployed tonight or can wait for the next quarterly cycle. The two decisions carry completely different resource costs, yet a binary model treats them identically.

The binary classification system (yes/no) that currently exists for exploiting vulnerabilities has a significant limitation: there is only one way to classify a CVE (Common Vulnerability and Exposure). By implementing the proposed temporal classification system, several new methods to classify the exploitability of CVEs are available; in this case, five new intervals for classifying CVEs will replace the single yes/no output (0–7 days, 8–30 days, 31–90 days, 91–365 days, or never exploited).

In reality, most organizations already utilize tiered patching protocols that will map directly to the new temporal classifications proposed; i.e., emergency remediation, accelerated deployment, monthly deployment, quarterly review, and risk-based exception. This allows the temporal predictions to be utilized directly within existing workflows without the need for an additional triage layer between the two processes.

This means that operationally, the outputs of temporal prediction provide functionality that enables a more actionable risk intelligence than currently exists. For example, when a CVE falls into Class 0, the corresponding response is immediate (emergency dispatch), which would differ substantially from the response when falling into Class 3 (no immediate dispatch), rather than relying upon an analyst's judgment and decision.

Answer to RQ2: Prediction through temporal CVE classification results in more actionable risk intelligence than binary classifications. Auto-mapping of the predicted exploitation period onto known patching tier periods eliminates the manual conversion process done by the binary classification, resulting in greater consistency in making decisions.

G. Deployment Feasibility and Computational Overhead

To assess whether the framework can operate under realistic operational workloads, we benchmarked inference latency and throughput on the deployment hardware (NVIDIA GPU, CUDA). Single-CVE inference latency averaged 6.43 ms (p50 = 6.44 ms, p95 = 6.74 ms, p99 = 6.87 ms), well within the sub-second budget required for interactive triage. Batch throughput scales from 139.4 CVE/s at batch size 1 to 1,207.4 CVE/s at batch size 64, with diminishing marginal gains beyond batch size 32 (1,188.2 CVE/s), indicating that the cross-attention fusion module is not the throughput bottleneck at moderate batch sizes. At the best observed throughput, the framework can process an estimated 104.3 million CVE assessments per day on a single GPU instance. Given that the National Vulnerability Database discloses on the order of dozens to a few hundred CVEs per day, a single deployed instance provides several orders of magnitude of headroom, and horizontal scaling is straightforward: the offline knowledge-graph embeddings are fixed and shared across workers, and the online scoring path is embarrassingly parallel across CVEs. These results indicate that computational overhead is not a barrier to enterprise-scale deployment.

V. Conclusion

Prioritization of vulnerabilities through traditional methods often relies heavily on their severity scores, which indicate how theoretical a vulnerability's impact is versus how the vulnerability will likely be exploited in the real world. Prioritizing vulnerabilities in this fashion incurs real costs; security teams work to patch vulnerabilities that will never be exploited due to limited time and resources, and they may also leave undetected vulnerabilities that warrant prompt prioritization and patching. This research reframes the question of time to exploitation of a vulnerability as five classes of temporal classifications, capturing timing information that is often lost in binary frameworks, and aligning classification outputs with patching tiers currently employed by most organizations.

These design choices have been validated. The complete framework yields 74.7% exploitation F1 and 89.1% AUC-ROC and outperforms the EPSS (the current industry standard) by 6.8 points for F1 and 6.4 points for AUC-ROC, and outperforms the benchmark for text-only BERT by 12.2 F1 points. Ablation studies indicate that knowledge graph integration and cross-attention fusion are the major reasons for this improvement in performance, indicating that explicitly modelling connections between weaknesses and attack methodologies has added more value than the ability to simply provide more complex text representations of vulnerability descriptions.

There are limitations to this study. The data used to create the labels for ground-truth vulnerability exploitation (CISA KEV, Exploit-DB) only reflects instances of vulnerability exploitation that are publicly reported. As such, the extent of the model's performance will be impacted as the model does not account for vulnerabilities being exploited in a private capacity. In addition, NVD records without CWE numbers reduce the ability to make predictions about a vulnerability's exploitation, as the model must rely solely on the text-based features of the record. The model was trained on an archive of exploitation data through 2022; this means that to continue to understand vulnerability exploitation as technology changes and attacker tools evolve, the model must be retrained consistently.

In the short term, improvements to the model will be more appropriately focused on the data side than the modeling side (automated CWE labels during preprocessing, quarterly retraining aligned with KEV updates, and expanding the number of ATT&CK techniques in the knowledge graph). In terms of architectural improvements, changing from Node2Vec to Graph Attention Networks and estimating uncertainty of

low-confidence outputs would allow for a more effective way to implement the system in an operational capacity. The primary conclusion reached in this work is that vulnerability exploitation risk is ultimately temporal and relational rather than simply a byproduct of severity; therefore, a reconsideration of how exploitation information is produced and utilized is likely warranted.

REFERENCES

[1] National Institute of Standards and Technology (NIST), "National Vulnerability Database," 2024. [Online]. Available: https://nvd.nist.gov/

[2] FIRST (Forum of Incident Response and Security Teams), "Common Vulnerability Scoring System Version 3.1: Specification Document," 2019. [Online]. Available: https://www.first.org/cvss/v3.1/specification-document.

[3] L. Allodi and F. Massacci, "Comparing vulnerability severity and exploits using case-control studies," ACM Trans. Inf. Syst. Secur. (TISSEC), vol. 17, no. 1, pp. 1–20, 2014. doi: 10.1145/2630069.

[4] J. Jacobs, S. Romanosky, B. Edwards, M. Roytman, and I. Adjerid, "Exploit prediction scoring system (EPSS)," Digit. Threats Res. Pract., vol. 2, no. 3, pp. 1–17, 2021. doi: 10.1145/3436242.

[5] L. Allodi and F. Massacci, "Security events and vulnerability data for cybersecurity risk estimation," Risk Anal., vol. 37, no. 8, pp. 1606–1627, 2017. doi: 10.1111/risa.12864.

[6] M. Shahzad, M. Z. Shafiq, and A. X. Liu, "A large-scale exploratory analysis of software vulnerability life cycles," in Proc. IEEE/ACM Int. Conf. Software Engineering (ICSE), 2012, pp. 771–781. doi: 10.1109/ICSE.2012.6227149.

[7] S. Frei, M. May, U. Fiedler, and B. Plattner, "Large-scale vulnerability analysis," in Proc. ACM SIGCOMM Workshop on Large-Scale Attack Defense (LSAD), 2006, pp. 131–138. doi: 10.1145/1162666.1162671.

[8] M. Bozorgi, L. K. Saul, S. Savage, and G. M. Voelker, "Beyond heuristics: Learning to classify vulnerabilities and predict exploits," in Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2010, pp. 105–114. doi: 10.1145/1835804.1835821.

[9] M. Edkrantz and A. Said, "Predicting cyber vulnerability exploits with machine learning," in Proc. Scandinavian Conf. on AI (SCAI), vol. 278, 2015, pp. 48–57. doi: 10.3233/978-1-61499-589-0-48.

[10] J. Jacobs, S. Romanosky, B. Edwards, I. Adjerid, and M. Roytman, "Improving vulnerability remediation through better exploit prediction," J. Cybersecurity, vol. 6, no. 1, p. tyaa015, 2020. doi: 10.1093/cybsec/tyaa015.

[11] A. Okutan, O. B. Vivas, and G. Wallin, "Predicting cyber security incidents using feature-based characterization of network-level malicious activities," in Proc. ACM Workshop on Artificial Intelligence and Security, 2017, pp. 3–13. doi: 10.1145/3128572.3140447.

[12] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. Conf. of the North American Chapter of the Assoc. for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[13] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2016, pp. 855–864. doi: 10.1145/2939672.2939754.

[14] C. Sabottke, O. Suciu, and T. Dumitras, "Vulnerability disclosure in the age of social media: Exploiting Twitter for predicting real-world exploits," in Proc. USENIX Security Symposium, 2015, pp. 1041–1056.

[15] S. M. Ghaffarian and H. R. Shahriari, "Software vulnerability analysis and discovery using machine learning and data mining techniques: A survey," ACM Comput. Surv. (CSUR), vol. 50, no. 4, pp. 1–36, 2017. doi: 10.1145/3092566.

[16] S. Chakraborty, R. Krishna, Y. Ding, and B. Ray, "Deep learning based vulnerability detection: Are we there yet?," IEEE Trans. Softw. Eng., vol. 48, no. 9, pp. 3280–3296, 2022. doi: 10.1109/TSE.2021.3087402.

[17] F. Li and V. Paxson, "A large-scale empirical study of security patches," in Proc. ACM Conf. on Computer and Communications Security (CCS), 2017, pp. 2201–2215. doi: 10.1145/3133956.3134072.

[18] L. Bilge and T. Dumitras, "Before we knew it: An empirical study of zero-day attacks in the real world," in Proc. ACM Conf. on Computer and Communications Security (CCS), 2012, pp. 833–844. doi: 10.1145/2382196.2382284.

[19] P. Wang, L. Gao, B. Ding, and Y. Chen, "VulGraph: A knowledge graph based approach for vulnerability information management," in Proc. IEEE Int. Conf. on Software Maintenance and Evolution (ICSME), 2019, pp. 553–557.

[20] A. Pingle, A. Piplai, S. Mittal, and A. Joshi, "RelExt: Relation extraction using deep learning approaches for cybersecurity knowledge graph improvement," in Proc. IEEE/ACM Int. Conf. on Advances in Social Networks Analysis and Mining (ASONAM), 2019, pp. 879–886. doi: 10.1145/3341161.3343519.

[21] MITRE Corporation, "CWE — Common Weakness Enumeration," 2024. [Online]. Available: https://cwe.mitre.org/

[22] MITRE Corporation, "CAPEC — Common Attack Pattern Enumeration and Classification," 2024. [Online]. Available: https://capec.mitre.org/

[23] MITRE Corporation, "MITRE ATT&CK," 2024. [Online]. Available: https://attack.mitre.org/

[24] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2013, pp. 2787–2795.

[25] T. Trouillon, J. Welbl, S. Riedel, E. Gaussier, and G. Bouchard, "Complex embeddings for simple link prediction," in Proc. Int. Conf. on Machine Learning (ICML), 2016, pp. 2071–2080.

[26] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in Proc. IEEE Int. Conf. on Computer Vision (ICCV), 2017, pp. 2980–2988. doi: 10.1109/ICCV.2017.324.

[27] Cybersecurity and Infrastructure Security Agency (CISA), "Known Exploited Vulnerabilities Catalog," 2024. [Online]. Available: https://www.cisa.gov/known-exploited-vulnerabilities-catalog

[28] Offensive Security, "Exploit Database," 2024. [Online]. Available: https://www.exploit-db.com/

Deep Learning-Based Image Recognition and Brand Identity Design for Intangible Cultural Heritage in Southern Sichuan

Zhihong Chen
Neijiang Normal University, Neijiang 641100, China

***Abstract*—The increasing emphasis on digital protection and creative transformation of intangible cultural heritage highlights the urgent need to apply artificial intelligence to accurately identify and brand rich cultural heritage resources in southern Sichuan. This study puts forward an image recognition and brand design model suitable for intangible cultural heritage in southern Sichuan. This model has obvious advantages in extracting key visual features such as texture, color, and shape, effectively coping with the challenges brought by complex texture and category imbalance in intangible cultural heritage images, and achieving stable and reliable recognition performance. The proposed model improves the overall accuracy, F1 score, and recognition rate of complex texture images, and the overall accuracy reaches 94.1%. Among them, the recognition accuracy of texture, color, and shape reached 92.8%, 95.2%, and 90.1%, respectively, which were better than the traditional models, and accurately captured the detailed features and visual style of intangible cultural heritage items. In the application of brand logo design, the evaluation results of user satisfaction show that the score is high in aesthetics, commercial value, and narrative quality, among which the average score of aesthetics is 3.92, which highlights the advantages of this model in visual language and artistic expression. This study has made a breakthrough in the high-precision and multi-dimensional feature extraction of image recognition of intangible cultural heritage in southern Sichuan Province, and provided an intelligent tool for the brand logo design of intangible cultural heritage. At the same time, the research promotes the integration of digital protection and creative transformation, which is of great value to cultural heritage inheritance, brand promotion, and creative industry application. The research results provide practical guidance for future model optimization, cross-regional intangible cultural heritage identification and multimodal design exploration.**

***Keywords*—Deep learning; southern Sichuan's intangible cultural heritage; image recognition; brand identity design**

I. INTRODUCTION

Southern Sichuan is rich in intangible cultural heritage, including projects with distinctive regional characteristics and cultural significance, such as Zigong Lantern Festival, Luzhou Oil Paper Umbrella, and Neijiang tanghua [1]. The core skills and cultural connotations of these intangible cultural heritages are usually embodied and passed down through physical image carriers, such as lantern patterns, umbrella design, and tanghua modeling [2]. However, the accelerated modernization process and the wide application of digital technology have brought two major challenges to the intangible cultural heritage in southern Sichuan: First, "live transmission" is facing the risk of interruption. Second, the progress of "digital protection" is lagging. Many intangible cultural heritage images still exist in traditional forms, such as paper files or physical collections, which are not conducive to efficient storage, retrieval, and sharing, and are also difficult to meet the needs of cultural communication and public participation in the new media era. Therefore, the vitality and influence of intangible cultural heritage are still limited [3–5].

Artificial intelligence (AI), especially deep learning, has shown excellent performance in image recognition. Its application in cultural heritage protection, such as cultural relics classification and traditional pattern extraction, has begun to appear [6]. However, at present, most studies only focus on individual categories of intangible cultural heritage and fail to systematically analyze the diverse cultural heritage images found in southern Sichuan [7]. In addition, traditional image recognition methods usually remain at the simple classification level. They rarely integrate recognition results with deeper cultural interpretation or the needs of branding and dissemination [8]. This separation between technology and culture limits the full potential of deep learning in intangible cultural heritage preservation. It also hinders the development of Southern Sichuan's intangible cultural heritage from "digital preservation" to "branded dissemination" [9].

This study takes Southern Sichuan's intangible cultural heritage images as the research object and attempts to establish an image recognition system based on deep learning and cultural interpretation. By optimizing the convolutional neural network (CNN) model, the system realizes high-precision recognition, feature extraction, and semantic analysis of images in different categories of Southern Sichuan's intangible cultural heritage. Based on the above recognition results, this study explores new ideas for brand identity design of Southern Sichuan's intangible cultural heritage. The cultural symbols contained in the intangible cultural heritage images displayed as the visual identity of the brand are integrated into contemporary people's needs in identification and expression. This study provides an effective and accurate method for digital protection of Southern Sichuan's intangible cultural heritage. This study bridges the gap in the application of deep learning in regional intangible cultural heritage image recognition. Simultaneously, it underpins the modernization of intangible cultural heritage branding and the market-oriented transformation of cultural value with theoretical justification and practical recommendations. These efforts contribute to sustaining the living transmission of Southern Sichuan's

intangible cultural heritage and promoting its sustainable development in the new era.

To address the technical challenges of Southern Sichuan intangible cultural heritage image recognition, including multi-category classification, high visual complexity, and class imbalance, as well as the limited integration of image recognition results into brand communication, this study develops an integrated framework for Southern Sichuan intangible cultural heritage image recognition and brand identity design that combines deep learning with cultural interpretation. The study first reviews existing research on the application of deep learning in cultural heritage preservation and identifies current technical gaps and limitations in practical applications. Subsequently, this study proposes a CNN model based on multi-loss joint optimization and carries out high-precision image recognition and feature analysis of intangible cultural heritage in southern Sichuan from three visual dimensions: texture, color, and shape. On this basis, the study further discusses how to transform the identified cultural characteristics into brand visual elements that support contemporary brand identity design. Finally, the performance of the proposed framework in image recognition and the effect on brand identity design is evaluated through comparative experiments and customer satisfaction surveys. The research results provide a comprehensive solution that provides technical support and practical guidance for the digital protection and brand communication of intangible cultural heritage in southern Sichuan.

II. Related Work

With the deep integration of digital technology and intangible cultural heritage protection, deep learning has gradually become the core tool of intangible cultural heritage digitization research with its advantages in image feature extraction and pattern recognition [10]. As the carrier of folk culture in the southwest of China, the intangible cultural heritage in southern Sichuan combines regional characteristics and cultural values in the image. However, the systematic research on the application of deep learning in the image of intangible cultural heritage in southern Sichuan and its transformation into brand identity is still limited.

Yu et al. constructed the "AI for Dunhuang" dataset for the preservation of Dunhuang cultural heritage. The dataset included 10,000 mural images for automated restoration and 3,455 images for style transfer. By improving the U-Net network architecture, the precision of mural defect filling was optimized. Archaeologists evaluated the restoration quality as comparable to manual restoration [11]. Fernandes et al. focused on preventive conservation of artifacts by integrating the Mask Region-based CNN with infrared thermography for defect detection in Italian marquetry artifacts. By introducing an adaptive threshold segmentation module, their method improved the recognition of deep defects such as wood cracks and glue aging. However, they pointed out that existing models are often tailored to single materials or artifact types and lack general applicability to the diverse materials and forms of folk intangible cultural heritage images [12]. In addition, international research has mainly concentrated on "protection-oriented" applications, such as large-scale site restoration and artifact defect detection. Few studies have linked image recognition results to "application-oriented" needs, such as intangible cultural heritage brand communication and cultural product design. As a result, a complete "technology–culture–market" chain has not yet been established [13].

Domestic research has focused more on the digitization of local intangible cultural heritage, achieving progress in technology adaptation and regional practice. Qi et al. addressed the diverse needs of ancient Chinese character recognition by proposing the HUNet model, which integrated an attention mechanism. They constructed a multi-type dataset of ancient Chinese character recognition, including more than nine million samples of Oracle Bone Inscriptions, bamboo slips, and stone carvings. Their method has achieved an accuracy of 98.7% in many ancient character recognition tasks [14]. Sun et al. discussed the preservation of intangible cultural heritage from the perspective of digital protection. They developed an intelligent document recording system based on deep learning for the cultural heritage of the Li branch, and used the cultural pattern recognition model to automatically extract and realize the digital preservation of traditional cultural characteristics. This study provided technical support for the long-term preservation and intelligent management of intangible cultural heritage resources. The research showed that the combination of artificial intelligence and digital technology improved the efficiency of cultural document recording and improved the quality of digital preservation, which provided a new perspective for digital inheritance and innovative utilization of intangible cultural heritage [15]. Pan et al. studied the development of a digital twin of cultural heritage and proposed a method based on deep learning to automatically transform a semantic point cloud into a semantic building information model (BIM). This method can intelligently identify and semantically reconstruct the spatial structure and building components of cultural heritage buildings, thus improving the efficiency and accuracy of digital cultural heritage modeling. Although their research has expanded the application of deep learning in digital protection of cultural heritage, it mainly focuses on the three-dimensional modeling of architectural heritage and does not involve the image recognition of intangible cultural heritage, nor does it discuss its application in brand logo design and other cultural and creative fields [16].

In summary, although research on deep learning in the intangible cultural heritage field has achieved progress in technical accuracy and regional applications, three major gaps remain. First, international research has largely overlooked regional folk intangible cultural heritage such as that of Southern Sichuan, and its technical applications do not yet cover the full "image recognition–brand transformation" chain. Second, domestic studies have focused on local intangible cultural heritage, but dependence on single carriers limits accurate recognition of the multi-carrier intangible cultural heritage images of Southern Sichuan, while digital resources remain disconnected from brand design needs. Third, local branding practices for Southern Sichuan's intangible cultural heritage lack technological support, relying on manual approaches that are inefficient and insufficiently standardized.

To address these gaps, this study examines multi-category intangible cultural heritage images from Southern Sichuan. By

optimizing deep learning models, the study achieves high-precision recognition and cultural interpretation of multi-carrier images and links these results with brand identity design. The aim is to provide an integrated solution for both digital preservation and branded dissemination of regional intangible cultural heritage.

III. Deep Learning-Based Model for Southern Sichuan's Intangible Cultural Heritage Image Recognition and Brand Identity Design

A. Introduction to Deep Learning

Deep learning is a machine learning algorithm that simulates the operation mechanism of the human brain neural network. By constructing interconnected artificial neural networks, it can realize complex abstraction and learn from massive datasets [17]. In the deep network, the original input, such as an image, text, or audio, will undergo a series of nonlinear transformations and gradually generate more and more refined feature representations. The algorithm can directly mine related patterns from large-scale datasets without artificial feature design [18]. With the optimization algorithm of iteratively adjusting model parameters, deep learning realizes autonomous learning and continuous improvement, and can accurately predict and support scientific decision-making in complex environments [19]. In recent years, with the rapid growth of data and computing resources, deep learning has developed into the core research field of artificial intelligence, and has profoundly changed many industries [20].

Let the design matrix X be an m×n matrix, where the data are centered with mean zero, i.e., $E[x]=0$. If not, preprocessing subtracts the mean from all samples to achieve centralization. The unbiased sample covariance matrix corresponding to X is given as Eq. (1) presents the unbiased sample covariance matrix associated with X:

$$Var[x]=\frac{1}{m-1}X^TX \tag{1}$$

The variance Var[z] forms a diagonal matrix, obtained through the linear transformation $z=W^Tx$. The primary components of the design matrix X are defined via the eigenvectors of X^TX, represented in Eq. (2):

$$X^TX=W\wedge W^T \tag{2}$$

Letting W denote the right singular vector matrix obtained from the singular value decomposition (SVD) $X=U\Sigma W^T$, and treating W as the eigenvector basis, the original eigenvalue relation can be written as Eq. (3):

$$X^TX=\left(U\sum W^T\right)^TU\sum W^T=W\sum{}^2W^T \tag{3}$$

It follows that Var[z] is diagonal. When X undergoes principal component decomposition, its variance can be formulated as:

$$Var[x]=\frac{1}{m-1}X^TX \tag{4}$$

$$=\frac{1}{m-1}\left(U\sum W^T\right)^TU\sum W^T \tag{5}$$

$$=\frac{1}{m-1}W\sum{}^2W^T \tag{6}$$

Here, the relation $UTU=I$ holds, since by definition the matrix U is orthogonal under SVD. This demonstrates that the covariance matrix z possesses a diagonal form, as detailed below:

$$Var[z]=\frac{1}{m-1}Z^TZ \tag{7}$$

$$=\frac{1}{m-1}W^TX^TX^TW \tag{8}$$

$$=\frac{1}{m-1}W^TW\sum{}^2W^TW \tag{9}$$

The preceding analysis demonstrates that a linear projection of x onto z via W yields a diagonal covariance matrix (Σ^2), which implies mutual independence among the components of z.

The preceding analysis of the fundamental principles of deep learning and its feature learning mechanism shows that deep learning employs multi-layer neural network architectures to progressively abstract and efficiently represent data. Through parameter optimization, it continuously enhances the model's ability to learn from complex datasets [21]. The principles of feature dimensionality reduction and information extraction embodied in principal component analysis also provide an effective approach for representing high-dimensional data within deep learning models. In the context of the digital preservation and brand identity design of Southern Sichuan intangible cultural heritage, image datasets contain abundant cultural symbols and visual characteristics. Traditional image processing methods are often unable to effectively extract the underlying semantic information embedded in these complex visual features. Therefore, applying deep learning to Southern Sichuan intangible cultural heritage image recognition is essential for developing an intelligent recognition model with strong feature extraction and classification capabilities. Based on these considerations, this study develops a deep learning-based algorithm for Southern Sichuan intangible cultural heritage image recognition. A CNN is employed to extract image features, while a multi-loss joint optimization strategy is introduced to improve classification accuracy and feature discrimination. The proposed algorithm establishes a technical foundation for the automatic recognition of Southern Sichuan intangible cultural heritage images, cultural element analysis, and digital dissemination.

B. Deep Learning-Based Image Recognition Algorithm for Sichuan Southern Intangible Cultural Heritage

In order to realize the automatic recognition and application analysis of Sichuan southern intangible cultural heritage images, this research studies a recognition algorithm based on deep learning. The method is as follows: Firstly, use CNN to extract the local and global image features. The network output is transferred to class probabilities by the Softmax function.

Then, cross-entropy loss is used as a supervisory signal in the training process. Meanwhile, Focal Loss is introduced to solve the class imbalance problem. In addition, Triplet Loss and Center Loss are incorporated to enhance feature discrimination [22]. The model is optimized under a multi-loss framework, which ensures both high classification accuracy and strong feature separability [23].

The convolutional layer extracts local features from images as follows:

$$y_{i,j,k} = \sum_{u,v,c} W_{u,v,c,k} x_{i+u-1,j+v-1,c} + b_k \tag{10}$$

In Eq. (10), x denotes the input image, b_k represents the bias, W signifies the convolution kernel, and $y_{i,j,k}$ refers to the output feature map.

The Softmax function converts the network output into class probabilities:

$$p_{i,c} = \frac{\exp\left(z_{i,c}\right)}{\sum_j \exp\left(z_{i,j}\right)} \tag{11}$$

In Eq. (11), $z_{i,c}$ is the score of sample i in class c, and $p_{i,c}$ is the predicted probability.

Cross-entropy loss supervises classification training:

$$L_{FL} = -\frac{1}{N}\sum_i \left(1 - p_{i,yi}\right)^{\gamma} \log p_{i,yi} \tag{12}$$

In Eq. (12), γ controls the weighting of difficult samples.

Triplet Loss enforces inter-class separability:

$$L_{tri} = \sum \left[\left\| f_a - f_p \right\|^2 - \left\| f_a - f_n \right\|^2 + \alpha \right] \tag{13}$$

In Eq. (13), f_a, f_p, and f_n are the features of the anchor, positive, and negative samples, respectively, and anchor α is a margin hyperparameter.

Center Loss strengthens intra-class compactness:

$$L_{cener} = \frac{1}{2N}\sum_{i=1}^{N} \left\| f\left(x_i\right) - c_{yi} \right\|^2 \tag{14}$$

In Eq. (14), $f\left(x_i\right)$ is the feature of sample i, and c_{yi} is the center of its class.

Finally, the comprehensive loss function jointly optimizes classification and metric learning objectives:

$$L = L_{CE} + \lambda_1 L_{tri} + \lambda_2 L_{cener} \tag{15}$$

In Eq. (15), λ_1 and λ_2 are the corresponding loss weights.

This multi-loss optimization strategy enables the proposed model to achieve both accurate classification and highly discriminative feature representations.

The quality monitoring of the Sichuan southern intangible cultural heritage image dataset is achieved through a multi-level verification and screening mechanism to ensure both reliability and usability. Processed intangible cultural heritage images are first organized and subjected to quality checks in terms of clarity, integrity, and format. Data that fail to meet requirements are either discarded or corrected. Qualified images then proceed to the classification stage, where the accuracy of category labeling is verified to maintain consistency and correctness. Once classification and labeling are complete, the system generates corresponding image package files. These files undergo an additional round of validity checks, and any defective data are returned for correction. Only verified image data proceed to the file generation stage, followed by a final review to ensure compliance with database storage standards. By implementing quality control in all stages from collection, classification, labeling to storage, this dataset provides a reliable basis for the subsequent image recognition and deep learning modeling of intangible cultural heritage [24]. Fig. 1 presents the procedure for monitoring the quality of the dataset.

The Sichuan southern intangible cultural heritage image recognition system is formed by three parts: computing platform, database platform, and integrated management system. The computing platform module includes the image acquisition module, image preprocessing module, feature extraction module, and classification module. These modules realize the whole process from raw image acquisition to feature categorization. The database platform module realizes data storage and retrieval. The module of the integrated management system includes an automatic labeling and statistical module, heritage category analysis module, cultural heritage recommendation module, and interactive display and education module. These functions can realize systematic annotation, category, cultural analysis, and public display [25].

The image recognition system for Sichuan's southern intangible cultural heritage realizes intelligent recognition on the computing platform, efficient management on the database platform, and the analytical and presentation functions on the integrated management system. These parts realize the whole closed loop from image acquisition to application and dissemination, as shown in Fig. 2. This system provides strong digital support for the protection, research, and communication of intangible cultural heritage.

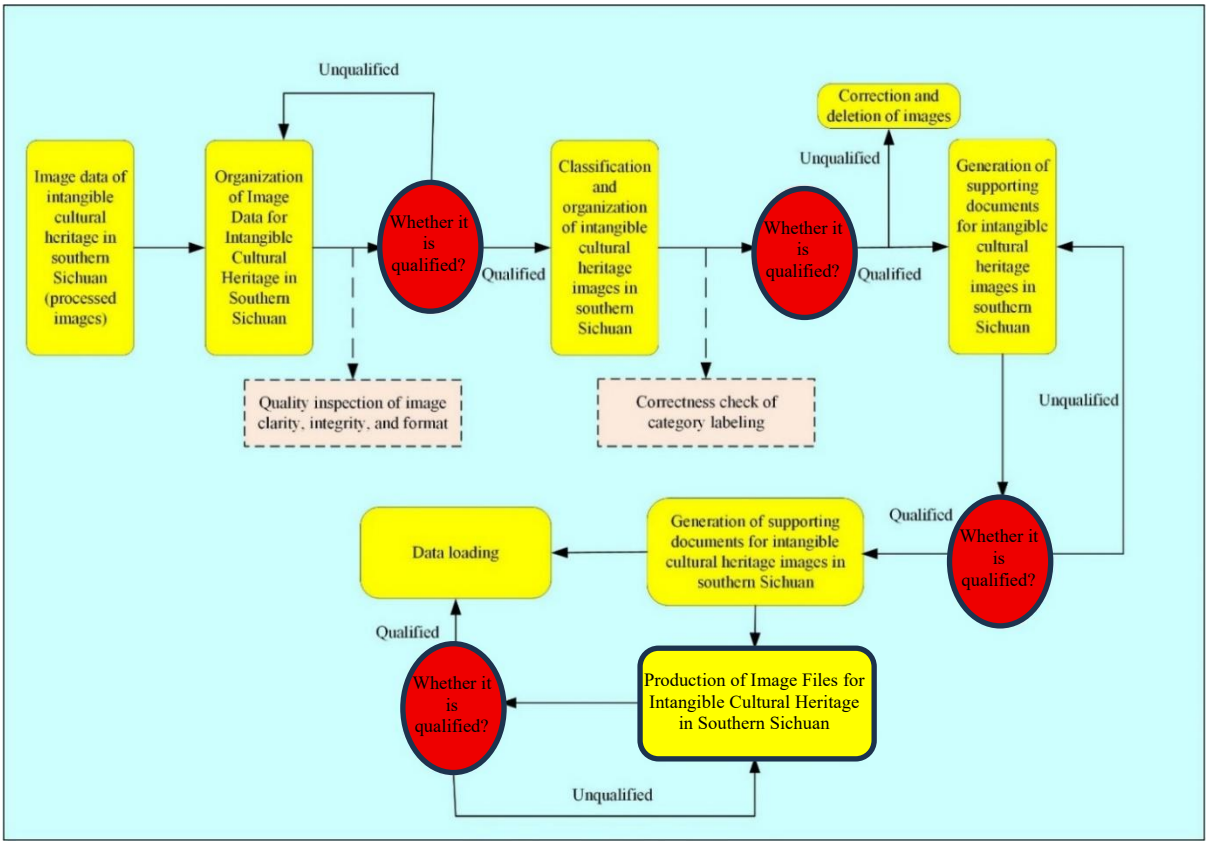


Fig. 1. Quality monitoring process of the Sichuan southern intangible cultural heritage image dataset.

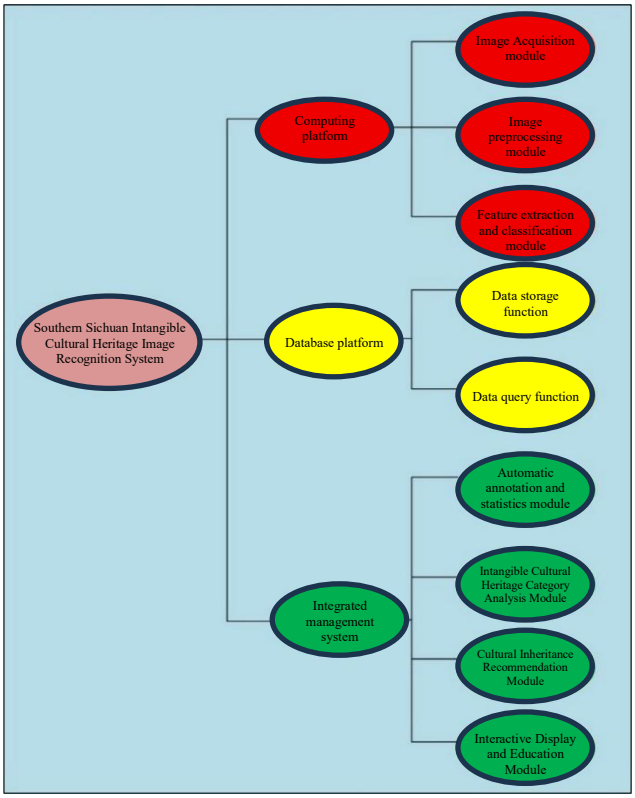


Fig. 2. System architecture of the Sichuan southern intangible cultural heritage image recognition framework.

C. Deep Learning-Based Brand Identity Design Model for Southern Sichuan's Intangible Cultural Heritage

The deep learning-based brand identity design model for Southern Sichuan's intangible cultural heritage primarily involves data collection, preprocessing, model training, and optimization. The system first gathers relevant intangible cultural heritage materials and evaluates whether they meet the quantity and quality requirements for modeling. If the data is insufficient, additional data will be collected and supplemented. Once sufficient data is obtained, pretreatment is carried out to standardize the material and ensure the consistency and usability of the input. Based on the preprocessed data, an appropriate deep learning model is selected as the basis to generate an initial model. In the process of model creation, training parameters are set, and the model is tested and verified. If the performance does not meet the requirements, adjust the parameters or modify the model architecture until the expected effect is achieved.

In the process of model optimization, the system will continuously adjust the training parameters and make many iterations to improve the generalization ability and generation quality of the model. When the preset accuracy and robustness requirements are met, the process will enter the stage of brand identity generation. In Fig. 3, the image generation module in the brand logo design model of intangible cultural heritage in southern Sichuan based on deep learning consists of four parts: cultural feature extraction, feature coding, visual element mapping, and generation model reasoning. Firstly, the trained

deep learning model analyzes the intangible cultural heritage materials in southern Sichuan and extracts representative texture, color, shape, and cultural semantic features. These features are then encoded into computed feature vectors. Then, through the visual element mapping mechanism, the corresponding relationship between cultural characteristics and brand design elements is established. Specifically, texture features are mapped to brand logo graphics, color features correspond to the brand color matching system, shape features correspond to auxiliary graphics and layout style, and cultural semantic features are reflected in the expression of brand cultural identity. Based on these mapped features, the generation model will reorganize and optimize the design elements according to the user-defined design instructions and style parameters, and automatically generate multiple brand image schemes. Finally, the generated design is evaluated and comprehensively judged from the aspects of cultural consistency, visual beauty, and brand uniqueness. If the evaluation results do not satisfy the predefined criteria, the workflow returns to the parameter optimization stage for further refinement. Otherwise, the final brand identity design is generated. This closed-loop workflow—model optimization, feature extraction, visual element mapping, design generation, and result evaluation—transforms the cultural elements of Southern Sichuan intangible cultural heritage into brand visual symbols. As a result, it ensures cultural authenticity and design consistency while improving the automation and reproducibility of the brand identity design process. The workflow of the deep learning-based Southern Sichuan intangible cultural heritage brand identity design model is shown in Fig. 3.

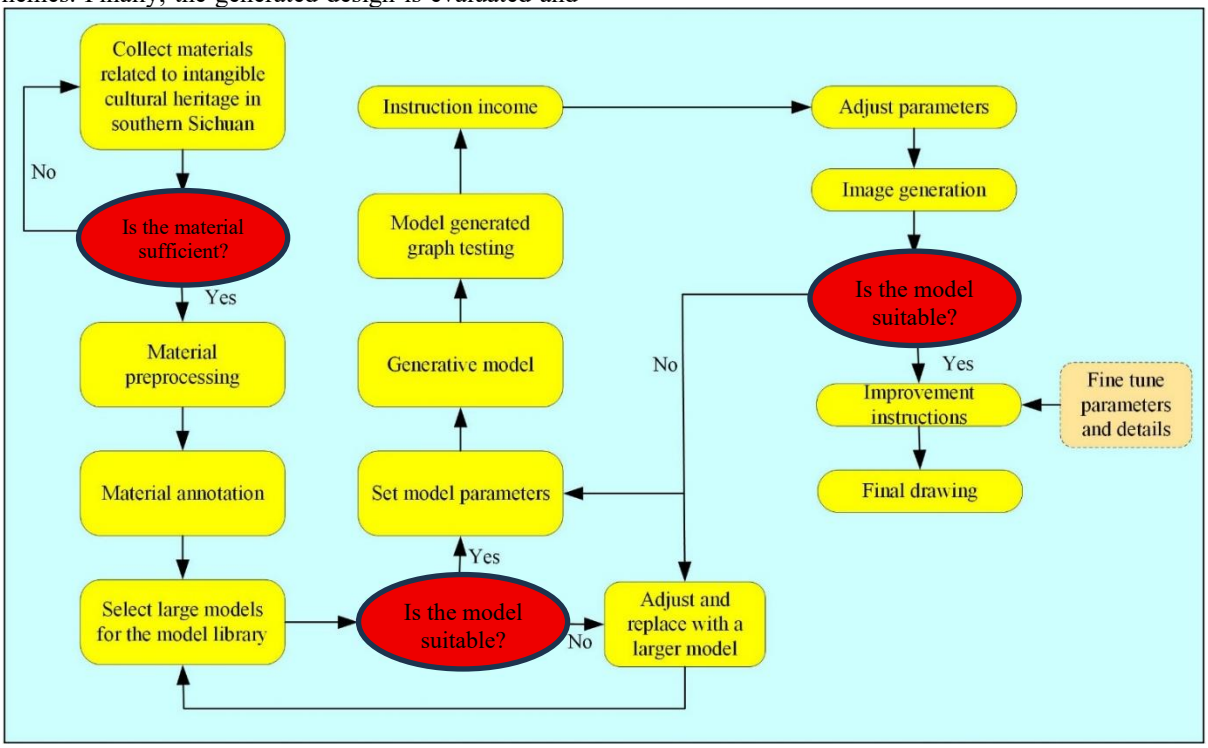


Fig. 3. Workflow of the deep learning-based brand identity design model for Southern Sichuan's intangible cultural heritage.

IV. Application of the Deep Learning-Based Model for Southern Sichuan's Intangible Cultural Heritage Image Recognition and Brand Identity Design

The construction of Southern Sichuan's intangible cultural heritage dataset is based on multi-channel systematic collection and organization. First, high-resolution images of typical intangible cultural heritage items, such as Shu brocade, Sichuan opera facial masks, bamboo weaving, embroidery, and ceramics, were captured through field surveys in southern Sichuan. This ensured the authenticity and completeness of textures, details, and colors. Second, the dataset integrates publicly available intangible cultural heritage resources from national and local cultural departments, including the Digital Museum of Chinese Intangible Cultural Heritage and the Sichuan Province intangible cultural heritage Information Management Platform. These resources supplement the historical documents and brand visual design materials that are difficult to collect in the field. In addition, brand image design images of enterprises and design institutions related to intangible cultural heritage, including signs, packaging, and posters, have been collected, and all images have been authorized by law or publicly licensed. In the process of database construction, all images are labeled according to the process type or brand design category, with category labels, texture features, colors, and style information attached. Using a standardized file structure, the dataset is divided into training, verification, and testing sets to support the training, verification, and testing of the model and ensure the repeatability of the experiment and scientific performance evaluation. This dataset captures rich process details and reflects the brand visual design style. It provides a high-quality foundation for the research of intangible cultural heritage

image recognition and brand identity design based on deep learning.

This dataset contains images of eight representative categories of intangible cultural heritage in southern Sichuan: Zigong Lantern Festival, Luzhou Oil Paper Umbrella, Neijiang tanghua, Shu Embroidery, Sichuan Opera Mask, Bamboo Weaving, Embroidery and Ceramics. In order to improve the generalization ability of the proposed model, image acquisition covers various manifestations of each intangible cultural heritage category in different perspectives, lighting conditions, and craft styles. The dataset contains 8,080 images, which are randomly divided into a training set, a verification set, and a test set with the ratio of 7:1.5:1.5. The training set contains 5,656 images for model training and parameter optimization. The verification set contains 1,212 images, which are used for hyperparameter tuning and early stopping during training. The test set also contains 1,212 images, which are used to evaluate the final performance of the trained model. In order to ensure the fairness of model evaluation, the dataset is divided by stratified random sampling, so as to keep roughly the same category distribution in the training set, verification set, and test set, and minimize the possible deviation caused by category imbalance. Table 1 shows a summary of the number of images of each category and the division of corresponding training sets.

TABLE I. DISTRIBUTION OF IMAGE CATEGORIES AND DATASET PARTITION FOR THE SOUTHERN SICHUAN INTANGIBLE CULTURAL HERITAGE IMAGE DATASET

Intangible Cultural Heritage Category	Total Images	Training Set	Validation Set	Test Set
Zigong Lantern Festival	1,280	896	192	192
Luzhou Oil-Paper Umbrellas	1,150	805	173	172
Neijiang Sugar Painting	1,020	714	153	153
Shu Brocade	980	686	147	147
Sichuan Opera Facial Masks	1,350	945	203	202
Bamboo Weaving	890	623	134	133
Embroidery	760	532	114	114
Ceramics	650	455	96	99
Total	8,080	5,656	1,212	1,212

During the model training process, all hyperparameters are systematically configured. A ResNet-50 model pre-trained on the ImageNet dataset is adopted as the backbone feature extraction network. Parameter initialization based on the pre-trained weights is used to accelerate convergence and enhance the generalization capability of feature representations. The training process is conducted for 200 epochs with a batch size of 32. The Adam optimizer is employed for parameter updating, and the initial learning rate is set to 1×10^{-4}. A cosine annealing learning rate schedule is applied, gradually reducing the learning rate during training to a final value of 1×10^{-6}. To mitigate overfitting, a Dropout layer with a rate of 0.5 is added before the fully connected layer. In addition, an early stopping strategy is adopted, where training is terminated if the validation loss does not decrease for 15 consecutive epochs. The model finally converges at the 176th epoch. In the joint loss function, the weight of the classification loss λ_1 is set to 1.0, the weight of the triplet loss λ_2 is set to 0.1, and the weight of the center loss λ_3 is set to 0.01. All experiments are conducted on a deep learning workstation equipped with an NVIDIA RTX 3090 GPU (24 GB memory). The software environment is based on Ubuntu 20.04, the PyTorch 1.12 deep learning framework, and CUDA 11.6 acceleration libraries. Each training epoch takes approximately 3.2 minutes, and the total training time is about 9.5 hours.

To validate the effectiveness of the proposed deep learning model for Southern Sichuan intangible cultural heritage image recognition, comparative experiments are conducted against both traditional machine learning algorithms, namely Support Vector Machine (SVM) and K-Nearest Neighbors (KNN), and mainstream deep learning architectures, including ResNet-50 and EfficientNet-B0. SVM and KNN are included to evaluate the performance improvement over conventional methods, while ResNet-50 and EfficientNet-B0 are used to assess the performance gap between the proposed model and state-of-the-art deep learning approaches, thereby providing a comprehensive evaluation of the model's technical contribution. Accuracy, F1-score, and recognition rate for complex-texture images were selected as evaluation metrics. They comprehensively reflect overall classification performance, the ability to handle class imbalance, and the capacity to capture detailed features in high-complexity images. Fig. 4 and 5 summarize the results obtained from the experiments.

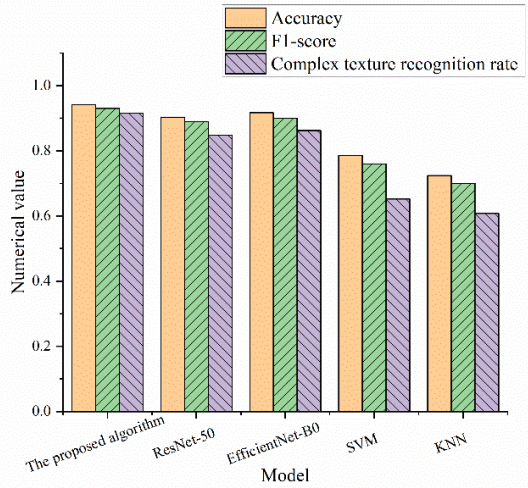


Fig. 4. Performance comparison of different models for Southern Sichuan's intangible cultural heritage image recognition.

As illustrated in Fig. 4, the proposed model significantly outperforms both traditional machine learning models and general deep learning models in terms of accuracy, F1-score, and recognition rate for complex-texture images. The overall accuracy reaches 94.1%, which is higher than SVM, KNN, ResNet-50, and EfficientNet-B0 by 15.6, 21.7, 3.8, and 2.4 percentage points, respectively. In terms of F1-score, the proposed model achieves 93.0%, surpassing ResNet-50 (89.0%) and EfficientNet-B0 (90.0%). For complex-texture image recognition, the model attains 91.5%, representing improvements of 6.8 and 5.3 percentage points over ResNet-50 (84.7%) and EfficientNet-B0 (86.2%), respectively. These

results indicate that the proposed multi-loss joint optimization strategy more effectively captures fine-grained textures and complex patterns in intangible cultural heritage artifacts.

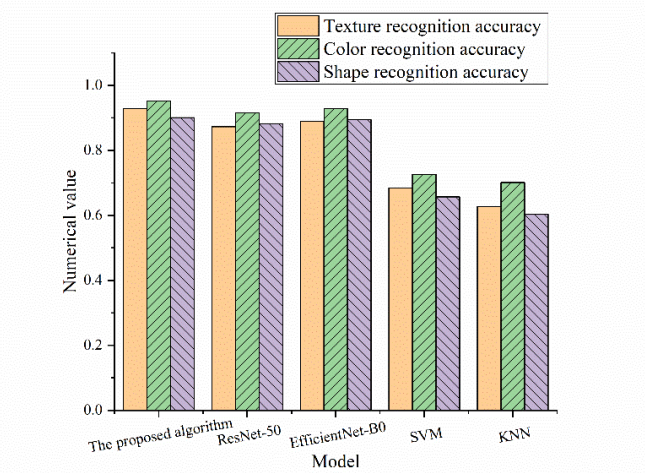


Fig. 5. Comparison of recognition performance across different image elements under various models.

As shown in Fig. 5, the proposed model achieves recognition accuracies of 92.8%, 95.2%, and 90.1% for texture, color, and shape, respectively, consistently outperforming SVM, KNN, ResNet-50, and EfficientNet-B0. In texture recognition, the proposed model improves performance by 5.5 and 3.9 percentage points compared with ResNet-50 (87.3%) and EfficientNet-B0 (88.9%), respectively, demonstrating the effectiveness of Triplet Loss and Center Loss in enhancing feature discriminability. For color recognition, all deep learning models perform relatively well. The proposed model achieves an accuracy of 95.2%, benefiting from both pre-trained feature extraction capabilities and the proposed joint optimization strategy, which further enhances fine-grained discriminative ability. For shape recognition, the proposed model reaches 90.1%, also showing steady improvements over ResNet-50 (88.2%) and EfficientNet-B0 (89.4%). Overall, the proposed model demonstrates clear advantages in multi-dimensional visual feature learning and exhibits strong applicability and technical advancement in Southern Sichuan intangible cultural heritage image recognition.

The performance satisfaction of the deep learning-based Southern Sichuan intangible cultural heritage brand identity design model was evaluated using a Likert scale survey. The survey targeted potential consumers of intangible cultural heritage products in Southern Sichuan. A total of 125 questionnaires were distributed, and 121 (62 males and 59 females) were effectively answered. Most of the respondents were between 18 and 35 years old, accounting for 77.12% of the sample. In terms of educational background, 26 respondents have an associate degree or below, 58 have a bachelor's degree, and 37 have completed graduate studies, indicating that the sample generally meets the representative requirements of this study. This study conforms to ethical standards and has been approved by the Ethics Committee of Neijiang Normal University. The written informed consent of the participants was obtained during the study. This study is conducted in accordance with the Helsinki Declaration.

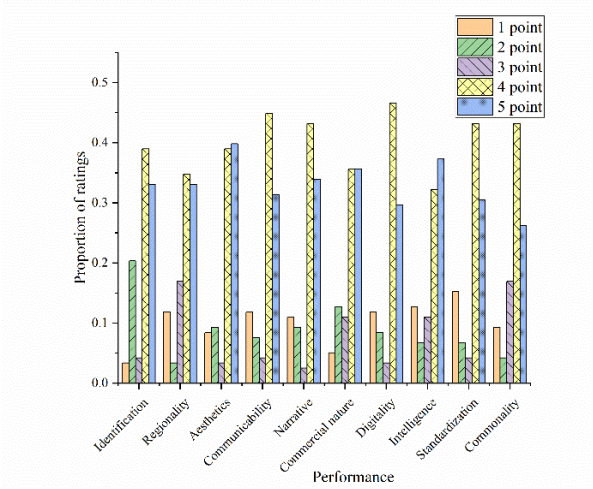


Fig. 6. User satisfaction with different performance aspects of the deep learning-based Southern Sichuan's intangible cultural heritage brand identity design model.

In Fig. 6, the aesthetic attraction scores the highest average score of 3.92, and 39.83% of the respondents give a perfect score of 5. This reflects the target users' high recognition of product visual language, formal aesthetics, and artistic expression. In contrast, the standardized dimension scored the lowest, only 3.67 points, and 15.25% of the respondents gave the lowest score of 1 point. This gap highlights the need for the design team to re-examine the compliance of design specifications, technical standards, and general design principles. The overall score of the intelligence dimension is in the middle, but the proportion of users getting full marks reaches 37.29%, indicating that users have positive expectations for technological innovation. On the contrary, the usability dimension has a low basic score (3.73 points) and has a limited full score (26.27%). It shows that the model needs further improvement in universality, accessibility, and public value creation.

V. Discussion

The experimental results show that the model based on deep learning for image recognition and brand design of intangible cultural heritage in southern Sichuan is consistently superior to SVM, KNN, ResNet-50, and EfficientNet-B0 in accuracy, F1 score, and recognition rate of complex texture images. The overall accuracy rate is 94.1%, which shows that the multi-loss joint optimization strategy effectively improves the fine-grained feature representation ability of intangible cultural heritage images and enhances the recognition performance of complex patterns, colors, and shapes. Compared with traditional machine learning methods, the deep learning model relies on hierarchical feature extraction and can learn higher-level semantic representations, so it is more suitable for complex cultural image recognition tasks. Based on the ResNet-50 backbone network, this study further integrates the ternary loss function and the central loss function to optimize the separability between categories and the compactness within categories, thus enhancing the performance of the model in capturing the subtle textures and complex visual features in the intangible cultural heritage images in southern Sichuan. The research results further

confirm the importance of digital intelligence technology in the protection and dissemination of cultural heritage. Boboc et al. pointed out that emerging digital technologies such as augmented reality are changing the presentation and dissemination of cultural heritage, opening up a new path for the digital utilization of cultural heritage resources [26]. Similarly, Erturk emphasized that digital storage and management are the key guarantee for the long-term preservation and inheritance of intangible cultural heritage [27]. Different from these studies, which mainly focus on digital protection and dissemination of cultural heritage, this study applies deep learning technology to image recognition and brand logo design of intangible cultural heritage to realize automatic extraction of cultural characteristics and transform them into visual design elements. The generated brand image solution has received relatively high evaluation in both aesthetic and intelligent dimensions, indicating that deep learning can not only support the digital preservation of cultural heritage, but also transform cultural elements into modern brand visual language. However, there are still some limitations: the size of the dataset is still relatively limited, and the evaluation mainly depends on subjective user evaluation. Future research should expand the image dataset of intangible cultural heritage, incorporate multi-modal cultural semantic information, and explore more advanced visual models. These improvements are expected to enhance the generalization ability and practical application value of the model.

VI. Conclusion

The image recognition and brand design model of intangible cultural heritage in southern Sichuan based on deep learning shows significant advantages in recognition accuracy and multi-dimensional feature extraction. Experimental evaluation shows that the model is superior to the traditional method in key indexes such as accuracy, F1 score, and complex texture recognition, and the overall accuracy reaches 94.1%. This proves that deep learning can effectively extract and present the fine details contained in intangible cultural heritage items. Specifically, the model achieves 92.8% in texture recognition, 95.2% in color recognition, and 90.1% in shape recognition, which fully embodies its comprehensive learning ability and excellent performance in multi-dimensional visual features. The customer satisfaction survey further confirmed the actual value of the model. The high scores in aesthetics, commercial attraction, and narrative design show that the model performs well in identification tasks and contributes to brand communication and design innovation. These research results confirm the application value of deep learning in digital protection and creative transformation of intangible cultural heritage. In practice, the model provides intelligent support for digital archiving, brand design, and marketing. For researchers and cultural heritage practitioners, it can realize efficient and accurate classification and feature extraction. For designers, it is helpful to generate a brand image that embodies the cultural essence of intangible cultural heritage, thus improving artistic quality, communication effect, and market competitiveness. Finally, the model is helpful to promote the innovation and transformation of intangible cultural heritage from southern Sichuan to international communication in the context of modern consumption.

Although the experimental results are encouraging, there are still some limitations. First of all, the dataset is limited in scale and category coverage, which cannot fully show the diversity and complexity of intangible cultural heritage in southern Sichuan. In this study, only the representative intangible cultural heritage projects in southern Sichuan are selected for analysis, so the feature expression ability of this model in other types of intangible cultural heritage resources needs further verification. The scores of standardization and universality are relatively low, indicating that it may have challenges in cross-domain and cross-scenario applications. In addition, there are significant differences in cultural connotation, artistic style, and visual expression of intangible cultural heritage projects in different regions, so the portability and generalization performance of this model in the context of intangible cultural heritage in other regions still need further investigation. The survey of users' satisfaction shows that the intelligence and universal applicability still need to be improved. Future research should focus on improving the interpretability of the model, personalized design recommendations, and cross-platform adaptability. In addition, expanding multi-modal datasets, integrating generative artificial intelligence technology, and optimizing interactive user experience are worthy of promotion. At the same time, it is necessary to construct cross-regional and cross-category intangible cultural heritage datasets to systematically evaluate and optimize the applicability of the model in different cultural heritage protection and brand design scenarios. Such efforts will help explore its broader applicability in the field of intangible cultural heritage and promote the advancement of intangible cultural heritage image recognition and brand identity design toward higher levels of intelligence and inclusiveness.

References

[1] T. Zhang, Y. Yang, X. Fan, et al., “Corridors construction and development strategies for intangible cultural heritage: a study about the Yangtze River economic belt,” Sustainability, vol. 15, no. 18, pp. 13449, 2023.

[2] S. Qiu, P. Zhang, S. Li, et al., “Extraction and analysis algorithms for Sanxingdui cultural relics based on hyperspectral imaging,” Computers and Electrical Engineering, vol. 111, pp. 108982, 2023.

[3] D. Ma, P. Pu, J. Hao, et al., “Automatic extraction and classification of intangible cultural elements using reinforcement learning algorithm,” Computer-Aided Design and Applications, vol. 21, no. S23, pp. 51–68, 2024.

[4] R. Li and C. Wang, “Cultural and creative product design and image recognition based on deep learning,” Computational Intelligence and Neuroscience, vol. 2022, no. 1, pp. 7256584, 2022.

[5] J. Xie, “Innovative design of artificial intelligence in intangible cultural heritage,” Scientific Programming, vol. 2022, no. 1, pp. 6913046, 2022.

[6] K. Bobanewa and O. I. Olope, “Deep learning techniques for image recognition (machine learning),” Path of Science, vol. 10, no. 10, pp. 2029–2036, 2024.

[7] Y. Hou, S. Kenderdine, D. Picca, et al., “Digitizing intangible cultural heritage embodied: state of the art,” Journal on Computing and Cultural Heritage (JOCCH), vol. 15, no. 3, pp. 1–20, 2022.

[8] Z. Xu and D. Zou, “Big data analysis research on the deep integration of intangible cultural heritage inheritance and art design education in colleges and universities,” Mobile Information Systems, vol. 2022, no. 1, pp. 1172405, 2022.

[9] H. Sharma, "Applications of artificial intelligence and machine learning in the preservation and analysis of heritage structures: a comprehensive review," Archives of Computational Methods in Engineering, vol. 33, no. 2, pp. 3001–3033, 2026.

[10] L. Qiu and M. Liu, "Innovative design of cultural souvenirs based on deep learning and CAD," Computer-Aided Design & Applications, vol. 21, no. S14, pp. 237–251, 2024.

[11] T. Yu, C. Lin, S. Zhang, et al., "Artificial intelligence for Dunhuang cultural heritage protection: the project and the dataset," International Journal of Computer Vision, vol. 130, no. 11, pp. 2646–2673, 2022.

[12] H. Fernandes, J. Summa, J. Daudre, et al., "Characterization of ancient marquetry using different non-destructive testing techniques," Applied Sciences, vol. 11, no. 17, pp. 7979, 2021.

[13] J. Zhao, "Digital protection and inheritance path of intangible cultural heritage based on image processing algorithm," Scalable Computing: Practice and Experience, vol. 25, no. 6, pp. 4720–4728, 2024.

[14] H. Qi, H. Yang, Z. Wang, et al., "AncientGlyphNet: an advanced deep learning framework for detecting ancient Chinese characters in complex scene," Artificial Intelligence Review, vol. 58, no. 3, pp. 88, 2025.

[15] J. Sun, K. A. T. Khalid, and C. S. Kay, "Deep learning models for cultural pattern recognition: preserving intangible heritage of Li ethnic subgroups through intelligent documentation systems," Future Technology, vol. 4, no. 3, pp. 119–137, 2025.

[16] X. Pan, Q. Lin, S. Ye, et al., "Deep learning based approaches from semantic point clouds to semantic BIM models for heritage digital twin," Heritage Science, vol. 12, no. 1, pp. 65, 2024.

[17] Q. Qiu, "Identifying the role of intangible cultural heritage in distinguishing cities: a social media study of heritage, place, and sense in Guangzhou, China," Journal of Destination Marketing & Management, vol. 27, pp. 100764, 2023.

[18] R. Archana and P. S. E. Jeevaraj, "Deep learning models for digital image processing: a review," Artificial Intelligence Review, vol. 57, no. 1, pp. 11, 2024.

[19] B. Peng and M. Sirisuk, "Exploration of CAD and neural network integration in art design and cultural heritage protection," Computer-Aided Design, vol. 21, no. 18, pp. 128–144, 2024.

[20] W. Xu, "Research on the communication opportunities of intangible cultural heritage under the background of big data and AI," Journal of Artificial Intelligence Practice, vol. 6, no. 8, pp. 12–17, 2023.

[21] H. Cho, Y. Kim, E. Lee, et al., "Basic enhancement strategies when using Bayesian optimization for hyperparameter tuning of deep neural networks," IEEE Access, vol. 8, pp. 52588–52608, 2020.

[22] J. Deng, "A brief analysis of the path of intangible cultural heritage inheritance and innovative development under digital technology," Journal of Innovation and Development, vol. 3, no. 3, pp. 29–32, 2023.

[23] Y. Abgaz, R. Rocha Souza, J. Methuku, et al., "A methodology for semantic enrichment of cultural heritage images using artificial intelligence technologies," Journal of Imaging, vol. 7, no. 8, pp. 121, 2021.

[24] M. Zhang, X. Guo, X. Guo, et al., "Consumer purchase intention of intangible cultural heritage products (ICHP): effects of cultural identity, consumer knowledge and manufacture type," Asia Pacific Journal of Marketing and Logistics, vol. 35, no. 3, pp. 726–744, 2023.

[25] M. Dai, Y. Feng, R. Wang, et al., "Enhancing the digital inheritance and development of Chinese intangible cultural heritage paper-cutting through stable diffusion LoRA models," Applied Sciences, vol. 14, no. 23, pp. 11032, 2024.

[26] R. G. Boboc, E. Băutu, F. Gîrbacia, et al., "Augmented reality in cultural heritage: an overview of the last decade of applications," Applied Sciences, vol. 12, no. 19, pp. 9859, 2022.

[27] N. Erturk, "Preservation of digitized intangible cultural heritage in museum storage," Milli Folklor, vol. 16, no. 128, pp. 100–110, 2020.