TOKAT GAZİOSMANPAŞA UNIVERSITY | NEAR EAST UNIVERSITY

INFTEC

3RD INTERNATIONAL CONGRESS ON DATA SCIENCE AND INFORMATION TECHNOLOGIES

HYBRID CONGRESS

PROCEEDINGS BOOK

October 22-24 2026

Near East University
Nicosia - TRNC

www.inftec.org

inftec

THE INTERNATIONAL CONGRESS ON DATA SCIENCE AND INFORMATION TECHNOLOGIES

Proceeding Book of the INFTEC 2026

ISBN: 978-625-90604-5-3

HYBRID (ONLINE AND FACE TO FACE) – NICOSIA/ TURKISH REPUBLIC OF NORTHERN CYPRUS (TRNC)

Note: All administrative, academic and legal responsibilities of the departments belong to their authors

Proceedings Book of the International Congress on Data Science and Information Technologies

(INFTEC-2026)

Publisher:
Tokat Gaziosmanpasa University
Editors:
Yalçın TAHTALI
E-Book Layout, Preparation and Composition:
Yalçın TAHTALI, Kaan KAPLAN, Lütfi BAYYURT, Osman Nuri ÖCALAN
All published articles were peer-reviewed by Scientific Committee
The organizers do not have any legal liability fort o contents of the presentation texts
Organized by
Tokat Gaziosmanpaşa University, Tokat, Turkiye

COMMITTEES

HONORARY COMMITTEE

Prof. Dr. Fatih YILMAZ	Rector of Tokat Gaziosmanpaşa University
Prof. Dr. Tamer ŞANLIDAĞ	Rector of Near East University

CONGRESS CHAIRMANS

Prof. Dr. Nadire ÇAVUŞ	Near East University, TRNC
Prof. Dr. Resul GERÇEKÇİOĞLU	Tokat Gaziosmanpaşa University, Türkiye
Prof. Dr. Tekin SUSAM	Tokat Gaziosmanpaşa University, Türkiye
Assoc. Prof. Dr. Nuriye SANCAR	Near East University, TRNC
Assoc. Prof. Dr. Yalçın TAHTALI	Tokat Gaziosmanpaşa University, Türkiye
Atıl SAMANCIOĞLU	Software Developer & Educator, Turkiye

CONGRESS SECRATARIAT

Assist. Prof. Dr. Lütfi BAYYUR	Tokat Gaziosmanpaşa University, Türkiye
Assist. Prof. Dr. Osman Nuri ÖCALAN	Tokat Gaziosmanpaşa University, Türkiye
Dr. Alpay ÇİLLER	Tokat Gaziosmanpaşa University, Türkiye
Dr. Kaan KAPLAN	Tokat Gaziosmanpaşa University, Türkiye
Dr. Mehmet KIZILASLAN	Tokat Gaziosmanpaşa University, Türkiye
Dr. Didem DOĞAR	Tokat Gaziosmanpaşa University, Türkiye
Res. Asst. Ahmet SEVİM	Tokat Gaziosmanpaşa University, Türkiye
Ahmet Sinan GÜLER	Tokat Gaziosmanpaşa University, Türkiye
Abdüssamed DURMUŞ	Tokat Gaziosmanpaşa University, Türkiye
Mesut Gökay OKUR	Tokat Gaziosmanpaşa University, Türkiye
Nurcan MENEVŞE	Tokat Gaziosmanpaşa University, Türkiye

ORGANIZING COMMITTEE

Prof. Dr. Adam TOROK	Budapest University of Technology and Economics (BME), Hungary
Prof. Dr. Boran ŞEKEROĞLU	Near East University, TRNC
Prof. Dr. Deniz İNAN	Marmara University, Türkiye
Prof. Dr. Hassan Salah Hassan BAKOUCH	Tanta University, Egypt
Prof. Dr. Hugo S. SALINAS	Universidad de Atacama, Chile
Prof. Dr. Jonas TAMAS	Eötvös Lorand University, Hungary
Prof. Dr. Mohammad Z. Raqab	Kuwait University, Kuwait
Prof. Dr. Nadire ÇAVUŞ	Near East University, TRNC
Prof. Dr. Soner ÇANKAYA	Ondokuz Mayıs University, Türkiye
Prof. Dr. Ufuk BEYAZTAŞ	Marmara University, Türkiye
Assoc. Prof. Dr. Beste Hamiye BEYAZTAŞ	İstanbul Medeniyet University, Türkiye
Assoc. Prof. Dr. Bülent TURAN	Tokat Gaziosmanpaşa University, Türkiye
Assoc. Prof. Dr. Erdem KÜÇÜKTOPÇU	Ondokuz Mayıs University, Türkiye
Assoc. Prof. Dr. İsmail İŞERİ	Ondokuz Mayıs University, Türkiye
Assoc. Prof. Dr. Nur İlkay ABACI	Ondokuz Mayıs University, Türkiye
Assoc. Prof. Dr. Nuriye SANCAR	Near East University, TRNC
Assoc. Prof. Dr. Sahar EBADINEZHAD	Near East University, TRNC
Assoc. Prof. Dr. Samet Hasan ABACI	Ondokuz Mayıs University, Türkiye
Assoc. Prof. Dr. Utku KALE	Budapest University of Technology and Economics (BME), Hungary
Assoc. Prof. Dr. Yalçın TAHTALI	Tokat Gaziosmanpaşa University, Türkiye

Assist. Prof. Dr. Abdulkadir ACAR	Tokat Gaziosmanpaşa University, Türkiye
Assist. Prof. Dr. Gülşah KEKLİK	İstanbul Nişantaşı University, Türkiye
Assist. Prof. Dr. Lütfi BAYYURT	Tokat Gaziosmanpaşa University, Türkiye
Assist, Prof. Dr. Oke OLUWAFEMI	Near East University, TRNC
Assist. Prof. Dr. Osman Nuri ÖCALAN	Tokat Gaziosmanpaşa University, Türkiye
Dr. Razik Ridzuan Mohd TAJUDDIN	University Kebangsaan Malaysia, Malaysia
Dr. Badamasi ABBA	Central South University, China
Dr. Breno FRAGOMENI	University of Connecticut, USA
Dr. Bruno Giacomini SARI	Federal University of Santa Maria, Brazil
Dr. Dhananjay GAIKWAD	Amity University, India
Dr. Didem DOĞAR	Tokat Gaziosmanpaşa University, Türkiye
Dr. Diego JARQUIN	University of Florida, USA
Dr. Francisco PEÑAGARICANO	University of Wisconsin-Madison, USA
Dr. Hao CHENG	University of California, Davis, USA
Dr. Mehmet KIZILASLAN	Tokat Gaziosmanpaşa University, Türkiye
Dr. Mehmet Sait KİREMİT	Ondokuz Mayıs University, Turkiye
Dr. Nasım AHMADZADEH	Near East University, TRNC
Dr. Pradeep MISHRA	Jawaharlal Nehru Krishi Vishwa Vidyalaya, India
Dr. Tadas ŽVİRBLIS	Vilnius Gediminas Technical University, Lithuania
Süsen HESENOVA	Baku State University, Azerbaijan

SCIENFTIFIC COMMITTEE

Prof. Dr. Boran ŞEKEROĞLU	Near East University, TRNC
Prof. Dr. Daniel LIBERACKI	University of Life Sciences in Poznan, Poland
Prof. Dr. Daniel ZABORSKI	West Pomeranian University of Tehcnology, Poland
Prof. Dr. Filiz KARAMAN	Yıldız Technical University, Türkiye
Prof. Dr. Francisco CARVALHO	Polytechnic Institute of Tomar, Portugal
Prof. Dr. Han Lin SHAN	Macquarie University, Australia
Prof. Dr. Han Lin SHANG	Macquarie University, Australia
Prof. Dr. Hassan Salah Hassan BAKOUCH	Tanta University, Egypt
Prof. Dr. Hon Keung Tony NG	Bentley University, United Kingdom
Prof. Dr. Mahmut HEKİM	Tokat Gaziosmanpaşa University, Türkiye
Prof. Dr. Murat ÖZGÖREN	Near East University, TRNC
Prof. Dr. Nadire ÇAVUŞ	Near East University, TRNC
Prof. Dr. Özge CAĞCAĞ YOLCU	Marmara University, Türkiye
Prof. Dr. Paulo Canas RODRIGUES	University of Bahia, Brazil
Prof. Dr. Ufuk YOLCU	Marmara University, Türkiye
Assoc. Prof. Dr. Bülent TURAN	Tokat Gaziosmanpaşa University, Türkiye
Assoc. Prof. Dr. Nuriye SANCAR	Near East University, TRNC
Assoc. Prof. Dr. Sahar EBADINEZHAD	Near East University, TRNC
Assoc. Prof. Dr. Soutir BANDYOPADHYAY	Colorado School of Mines, USA
Assoc. Prof. Dr. Yalçın TAHTALI	Tokat Gaziosmanpaşa University, Türkiye
Assist. Prof. Dr. Gizel BAKICIERLER SEZER	Marmara University, Türkiye

Assist. Prof. Dr. Gülşah KEKLİK	İstanbul Nişantaşı University, Turkiye
Assist. Prof. Dr. Semanur SARIÇAM	Marmara University, Türkiye
Assist. Prof. Dr. Oke OLUWAFEMI	Near East University, TRNC
Assist. Prof. Dr. Turan CANSU	Marmara University, Türkiye
Dr. Abhijit MANDAL	University of Texas at El Paso, USA
Dr. Fearghal KEARNEY	Queen's University Belfast, United Kingdom
Dr. Francisco CARVALHO	Polytechnic Institute of Tomar, Portugal
Dr. Kaiyinh JI	The University of Sydney, Australia
Dr. Mehmet Sait KİREMİT	Ondokuz Mayıs University, Türkiye
Dr. Miguel FONSECA	NOVA University, Portugal
Dr. Nasım AHMADZADEH	Near East University, TRNC
Dr. Ozan EVKAYA	Edinburgh University, United Kingdom
Dr. Panagiotis PAPASTAMOULIS	Athens University of Economics and Business, Greece
Dr. Pradeep MISHRA	Jawaharlal Nehru Krishi Vishwa Vidyalaya, India
Dr. Razik Ridzuan Mohd TAJUDDIN	Universiti Kebangsaan Malaysia, Malaysia
Dr. Zhiyang (Gee) ZHOU	University of Wisconsin-Milwaukee, USA

ACADEMICIAN REPRESENTATIVE ASSIGNED TO PUBLISH AND EDITING

In accordance with Article 8/b of the Interuniversity Council (ÜAK) Associate Professorship Criteria for the Basic Field of Natural Sciences and Mathematics (Table 2), Assoc. Prof. Dr. Yalçın Tahtalı and Asst. Prof. Dr. Lütfi Bayyurt—faculty members in the Department of Animal Science at our university's Faculty of Agriculture—have been assigned to roles within the organization and coordination processes of the "3rd International Congress on Data Science and Information Technologies (INFTEC2026)," which will be held in Nicosia from October 22–24, 2026, in collaboration with Near East University (Cyprus). This is for your information. Tokat Gaziosmanpaşa University Rectorate dated 12 June and numbered E-23845617-804.01-747022.

PRESENTATION CONTENTS

The Fraud Detection Approach Enhanced with Machine Learning and Synthetic Data..... 1

Large Language Models In Protein Function Prediction: A Comparative Analysis of Esm-2-Based Representations and Traditional Methods 2

AI-Based Analysis Of Cattle Genomic Data For Climate Adaptation: Identifying Genetic Regions Associated With Heat Stress Tolerance 3

The Volumetric Sharma-Mittal Entropy Approach for Parameter-Independent Feature Selection 4

Evaluation of the Impact of Using Artificial Intelligence on Unit Cost in Solution Quality Control for Tissue Processing Systems 5

Clinical and Economic Evaluation of Artificial Intelligence-Supported Liquid-Based Cervical Cytology Systems 6

Data Balancing and Hybrid Deep Learning in the Prediction of Unbalanced Hypothyroidism: A Leakage-Controlled Comparison 8

Data-Driven Mortality Jump Forecasting with a Non-Explosive Alpha-Series Recurrence Model 9

Evaluation of Structural Health Monitoring in CARALL Utilizing Low-Velocity Impact Testing Supported by Graphene-Enhanced Cellulose Sensor 10

Conditional Score for Logistic Regression with Measurement Error 11

Comparative Correlation Analysis of Relationships Among Postharvest Quality Attributes in Pepper and Tomato 12

Gravity Model of Trade with Network Analysis – ERGM Approach 13

Predicting Solar and Geomagnetic Activity Using Markov Chain Monte Carlo Techniques: An AI-Driven Analysis 14

Can Renewable Energy Consumption Drive Environmental Sustainability in Pakistan? The Role of Technological Progress and Economic Expansion 15

Trustworthy and Explainable Federated Learning for Privacy-Preserving ECG Classification in Smart-Hospital IoT 16

Accessible and Explainable Edge AI for Privacy-Preserving Wearable Fall Detection 18

Role-Adaptive Explainable Intrusion Detection for Trustworthy Internet of Medical Things Networks 20

Trustworthy and Temporally Explainable Early-Warning Systems for Sepsis Using ICU IoT Data 22

Comparison of CatBoost, LightGBM, and XGBoost Algorithms for Credit Default Prediction 24

Data-Driven Predictive Intelligence for Decision Support Systems: Machine Learning, Statistical Modelling, and Adaptive Analytics 25

Evaluation of Economic Activity Groups According to Occupational Health and Safety Indicators: Assessing the Effect of Criterion Weighting Methods on COBRA-Based Rankings 26

The Role of Business Intelligence Systems in Strategic Decision-Making: A Conceptual Analysis of Small and Medium-Sized Enterprises 27

Fracture Detection in Pediatric Wrist Radiographs Using Attention Mechanism-Based Yolov11 28

Community Representative Selection in Scientific Collaboration Networks: Balancing Local Influence and Connectivity 29

Machine Learning-Based Modeling for Nut Crop Yield Prediction 30

D-Number Quantum Representation of Fuzzy Cognitive Map Weights 31

Phytochemical characterization, antioxidant and enzyme inhibitory potential of *Juniperus oxycedrus L.* leaves supported by molecular docking 32

Hybrid Content Based Image Retrieval for Similarity Search in Industrial Fastener Technical Drawings 33

Multi-Authored h-Type Indices Based on CRedit Roles 34

Comparative Performance of Robust Regression and Machine Learning Methods under Vertical Contamination 35

Exploring Fuzzy Clustering Objective Functions through Metaheuristic Optimization for Image Segmentation 36

Evaluation of Advanced Transfer Functions in Binary Crested Porcupine Optimizer for Gene Selection-Oriented Cancer Classification 37

Regional Modelling of Greenhouse Gas Emissions: A Cluster and Regression Approach 38

Deep Learning for Pet Breed Classification: Flat and Hierarchical Hybrid Models 39

Examining the Relationships Among Age, Education Level and Gender of Drivers Involved in Traffic Accidents Using Log-Linear Models 40

Solar Flare Classification Using Logistic Regression and Its Variants 41

A Literature-Based Comparative Review of Popular Symmetric Encryption Algorithms 42

Reliable AI Starts with Reliable Data: Quality, Deduplication and Linguistic Heterogeneity in Large-Scale Social Media Analytics 43

An Empirical Study on Anisotropy of Word2Vec Embedding Spaces 44

Modified K-means Algorithm for Big Data 45

Universal Basic Income in the Age of AI: The Transformation of the Labour Market and the Future of Social Policy 46

The Algorithmic Parent: Gender Stereotypes and Family Bias in AI-Generated Visual Content 47

The Effects of Trust and Design Innovativeness on Students' Intention to Use ChatGPT 48

Cross-Study Generalization of Machine-Learning-Based Dry Sliding Wear Prediction: Evidence from Blind Validation 49

DP-CopulaGAN: Privacy-Preserving Tabular Synthesis through Copula-Space Adversarial Learning 50

Improved Nonparametric Estimation of Stress-Strength Reliability via Multiplicative Bias Correction 51

Artificial Intelligence, Digitized Archives, and the Transformation of Historical Research ... 52

Revisiting Data Privacy in Machine Learning Systems Across the Learning Lifecycle 53

Polycystic Ovary Syndrome (PCOS) Prediction Using Random Forest Algorithm 54

Comparsion of Non-Linear and Bayesian Approaches in the Wood Lactation Curve Model . 55

Data-Driven Asymmetric Fuzzy Chain-Ladder for IBNR Estimation 56

A Decision Support System for Regional Depot Location and Routing of Chain Stores Considering Road-Network Accessibility 57

Evaluating Video Game Player Satisfaction Through Text Mining and Sentiment Analysis of Online Reviews 58

From Rule-Based Daylight-Sensitive LED Control to AI-Powered Adaptive Lighting 59

AI as a Research Toolsmith: Generating Bespoke Software for Specialized Research Workflows 60

Near-Duplicate Records in Synthetic Telematics Data: How Validation Affects the Estimated Value of Machine Learning in Insurance Pricing 61

Hybrid Entity Matching and Decision Support for AI-Assisted Multi-Source Sanctions Monitoring 62

Response Surface Methodology and Machine Learning Approaches: A Comparative Study for Surface Roughness Prediction 63

Web-Based Statistical Reporting with R: An Application Framework for Educational Feedback 64

Personnel Scheduling Optimization for Hybrid Work Model Using Genetic Algorithm 65

Enhancing Turkish Fine-Tuning Datasets through a RAG-Based AI Agent: A Comparative Evaluation Using Llama 3 66

Accuracy, Reliability, and Energy Cost in Clinical Multi-Class Prediction: A Comparison of Gradient Boosting, Tabular Foundation Models, and Large Language Models 67

Beyond Neuronal Plasticity: Microglia as a Bounded Biological Regularization Layer for Long-Term Organoid Intelligence 68

A Statistical-Computational Framework for Privacy-Preserving Federated Reliability Inference 69

From Evidence Sufficiency to Abstention: A Conceptual Framework for Biomedical RAG Decision Support 70

Hypotheses, Not Methods: What Machine Unlearning Can Give Back to the Study of Human Unlearning 71

Evaluating the Attractiveness of OECD Countries for Foreign Health Workforce: A CRITIC-Weighted CoCoSo Performance Analysis 72

On Multiple Birth Support Vector Machines with k-Proximal Plane Clustering 73

From Data Science for Internet of Things (IoT) to Data Science for Internet of Medical Things (IoMT): Applications and Examples in Healthcare Services 74

Business Intelligence for E-Commerce Decision Support: An Exploratory Analysis of Retail Transaction Data Using Power BI 75

The Impact of Renewable Energy on Economic Growth in Selected European Countries and Türkiye 76

Business Intelligence for Small Business Decision-Making: A Power BI Dashboard Approach 77

The Dark Side of Service Robots in Tourism and Hospitality 78

Cybersecurity, User Trust, and Mobile Banking in Liberia: Evidence from a Quantitative Study 79

Cybersecurity, Reliable Data, and Sustainable Technology in Aviation in the Age of Artificial Intelligence: International Airport Applications and a Development Model for Ercan International Airport 80

Hybrid Content Based Image Retrieval for Similarity Search in Industrial Fastener Technical Drawings 81

Comparison of the Evolutionary Algorithms Performance for Multi-Response Surface Problem Optimization 82

DNA Methylation-Based Age Prediction: An Evaluation of ElasticNet, XGBoost, and Deep Neural Networks 83

Quality Management Supported by Artificial Intelligence in Quality 4.0: The Role of Statistical Approaches 84

A Comparative Analysis of Artificial Intelligence Use by Enterprise Size in European Union Countries 85

Satellite-Based Analysis of Seasonal and Spatial Vegetation Patterns to Support Conservation Monitoring at Archaeological Sites: The Case of Şapinuva 86

A Bibliometric Analysis of Hepatitis D Virus 87

A Computationally Efficient Estimator for Combating Multicollinearity in the Linear Regression Model 88

Predicting Survival in Stroke Patients Using Ensemble Machine Learning 89

Data Management and Governance in Education Research: A Bibliometric Analysis of Web of Science Publications 90

From Transcriptomic Data to Biological Insights: Computational and Bioinformatics Approaches 91

Hotel Visibility and AI-Mediated Positioning on Generative AI Platforms: A Comparative Analysis of Istanbul Hotel Recommendations 92

From the Design Matrix (X) to the Jacobian Matrix (J): How Nonlinearity Reshapes Collinearity 93

Evaluation of the Study Model of Using Excel and ChatGPT Together According to Question Categories in YÖKDİL 94

Computational Performance of Python Interpreters in Missing Data Imputation: A Comparison of CPython and PyPy 95

Balancing Classes While Distorting Probabilities: Effects of Class Imbalance Correction on Probability Calibration (524) 97

Overtime and Employee Attrition: A Causal Machine Learning Analysis Based on the Job Demands-Resources and Conservation of Resources Theories 98

Segmentation of Skin Lesions Using EfficientNet-Attention U-Net Architecture: An Evaluation on the ISIC 2016 Dataset 99

Drought and Small Ruminant Population Dynamics Across Climatic Gradients: Evidence from Panel Regression and Double Machine Learning 100

An Evidence-Based Verification Framework for Microbial Circular Polarization in Reflected Exoplanet Light 101

Electronic Nose-Based Halal Screening: Detecting Pork Adulteration in Beef via Explainable Machine Learning 102

Linux for Data Science 103

Industrial Data Integration Using a Database-Driven Dynamic OPC UA Gateway Architecture 104

Digital Monitoring of Biosensor-Integrated Wastewater Treatment and Bioremediation 106

An Artificial Intelligence-Supported Pedagogical Compliance Model for Fostering Academic Integrity in Higher Education 125

Short-Term Electricity Load Forecasting Using Machine Learning: A Comparative Study Against the Turkish Transmission System Operator's Official Forecasts 130

Predicting AI-Enabled Workplace Efficiency: A Comparative Study of Random Forest, Logistic Regression, and Linear SVM in Albania 137

Analysis of Football Player Data Using Artificial Intelligence and Recommendations for Player-Specific Training Plans 145

A Hybrid GA-SA Solver for the Two-Dimensional Cutting Stock Problem: A Benchmark Study of the Placement Decoder's Role 153

Designing AI-Supported Authentic Telecollaboration for Pre-Service EFL Teachers' Digital Pedagogical Competence: A Design-Based Research Study 164

Machine Learning for CEFR-Informed Lexical Difficulty Classification of Turkish Song Lyrics: A Leakage-Aware Evaluation of Handcrafted Linguistic Features Against a Transformer-Based Hybrid Model 184

Election Prediction Using Social Media Data: An Artificial Intelligence and Game Theory-Based Approach 213

Modeling Misinformation Propagation Risk in Social Networks Using a Mamdani-Type Fuzzy Inference System 241

Explainable Lightweight IoT Intrusion Detection with Unseen Attack Evaluation 256

When High Accuracy Fails Explainable Bilingual Detection of AI Generated Phishing in Remote Education and Remote Work 264

Explainable Artificial Wisdom Decision Support Architecture for Sabbath Discernment 273

Image-Based Drowsiness Detection: Comparative Performance Evaluation of Different CNN Models 287

Deep Learning versus Classical Machine Learning for Anomaly Detection in Embedded Network Traffic 296

Model Predictive Control versus Decoupled PI Control of a Quadruple-Tank Process under Actuator Faults 303

The Role Of Personality–Emotional Intelligence Alignment In Enhancing Mentoring Effectiveness 317

Camera-Based Parking Occupancy Detection: A Comparison of YOLOv8n, YOLO11n, RT-DETR-L, and Faster R-CNN 324

A Multidimensional Behavioral Segmentation Framework for an Automotive Subscription Service: Longitudinal Validation with Multi-Horizon Churn Outcomes 334

Cognitive Dual-Channel AI Framework for Financial Market Forecasting: An Application to Borsa Istanbul 347

A Hybrid Machine Learning Approach for Bank Stability Prediction 357

Network Intrusion Detection Using Optimized Machine Learning Methods 368

Beyond Model Drift: A Meta-Governance Framework for Cascading Decision Intelligence Systems 376

Information Technologies and Data-Informed Decision-Making in Education: Opportunities and Limitations 390

FlowFormer: An Explainable Dual-Positional Sequence-Labelling Network Intrusion Detection System with Two-Axis Attention Attribution 405

A Zero-Trust Inspired Browser-Based Remote Access Architecture 415

Does AI Write Secure Code? An OWASP Top 10-Based Security Evaluation of LLM-Generated ASP.NET Core Code Across English and Turkish Prompts 426

Comparison of Machine Learning Models in a Parameter-Free Adaptive EWMA Control Chart for Process Dispersion Monitoring 435

Data Balancing and Hybrid Deep Learning in the Prediction of Unbalanced Hypothyroidism: A Leakage-Controlled Comparison 448

Predicting Migration in Türkiye with Macroeconomic Indicators Using LASSO and Elastic-Net Regression 459

Current Problems and Future Directions in Industrial Robot Control 469

Analysis of the Functional and Mechanical Infrastructure of Industrial Robots and Their Operating Principle 477

AI-Based Early Detection of Diabetic Retinopathy Using Handcrafted Retinal Biomarkers and Machine Learning 485

One-Class Tabular Denoising Diffusion for Network Intrusion Detection Under Validation-Calibrated Thresholds 501

Information Technologies and Data-Informed Decision-Making in Education: Opportunities and Limitations 513

Generation of a Unified Analysis Surface from Drone Video for Analysing Unused Road Surfaces at Intersections 528

A Comparative Evaluation of Machine Learning Classifiers for Multiclass Network Intrusion Detection on a Cybersecurity Dataset 538

Comparative Evaluation of QRS-Based Machine Learning and Deep Learning for ECG Emotion Recognition..... 550

Development of Root Cause Analysis and Autonomous Operational Intervention Systems in Enterprise Application Performance Monitoring Software Through AI-Assisted Event Detection 558

Design of an AI-Powered Decision Support System for Fabric Quality Control 564

Time Series Analysis for Electricity Produced by the Hydroelectric Power Plant in Türkiye 572

A Method for the Evolution of Agent Modeling Languages..... 583

Hidden Biases: Textbook Wording and Data Contamination in Behavioral Economics Experiments with Large Language Models..... 606

Evaluating Dimensionality Reduction Through Correlation-Network Preservation, Neighborhood Structure, and Predictive Performance Across Three Datasets 616

An AI-Driven Framework for Document Processing, Data Validation, and Digital Archiving in Logistics Operations..... 631

Predictive Performance and Stability of Machine Learning Classifiers for Real-Rooted Polynomials Across Degrees 640

Artificial Intelligence Capacity and Environmental Sustainability in OECD Countries: A Cross-Country Assessment 653

Target-Dependent Source Utility in Transfer Learning for SPT-Based Shear-Wave Velocity Prediction 665

Spatiotemporal Validation of Machine-Learning Models for Weekly Precipitation pH Estimation..... 674

ABSTRACT ORAL PRESENTATIONS

The Fraud Detection Approach Enhanced with Machine Learning and Synthetic Data

Burak Yüksel[1*], Onuralp Gayret[1],

[1]Koç Finansman A.Ş. Genel Müdürlüğü, Ünalan Mah. Ayazma Cd. Çamlıca İş Merkezi A Blok No:131, 34700, Üsküdar/İstanbul, Türkiye

*Corresponding author e-mail: byuksel@kocfinans.com.tr

Abstract

This study focuses on fraud detection for the individual (personal) portfolio of second-hand vehicles within a financing company operating in the automotive sector. The objective is to identify fraudulent transactions that may arise during the application or loan disbursement process, using machine learning algorithms. Since credit fraud methods change constantly, relying only on traditional checks is no longer enough to catch these complex tricks. In assessing the likelihood of fraud for personal loans, it's crucial to evaluate not only the applicant but also the vehicle's history and any unusual patterns related to the sales process. Details such as the individual applicant's financial history, behavioral credit variables, the vehicle's damage records, mileage status, and past ownership information play a critical role in determining fraud risk before the loan decision. By bringing all these different pieces of data together, this approach uncovers hidden connections that might otherwise be missed during standard checks. The dataset used in this study includes demographic information, behavioral credit history, technical and historical vehicle data, and specific variables indicating potential fraud. This developed model has been integrated into a Decision Support System to contribute to reducing fraud-related losses in the long term and establishing a more secure lending strategy for second-hand vehicle loans. Ultimately, this integration connects predictive analysis with daily operations, offering a proactive defense mechanism for the institution.

Key words: *fraud; credit risk assessment; individual client; machine learning; credit scoring; macroeconomic variables.*

Large Language Models In Protein Function Prediction: A Comparative Analysis of Esm-2-Based Representations and Traditional Methods

Gülşah Keklik[1*]

[1]İstanbul Nişantaşı University, Turkiye

*Corresponding author e-mail: gulsah.keklik@nisantasi.edu.tr

Abstract

Functional classification of protein sequences is of critical importance in bioinformatics, underpinning applications such as drug discovery, disease mechanism elucidation, and synthetic biology. Traditional approaches typically rely on hand-crafted features, such as amino acid composition or k-mer frequencies, which often fail to capture complex sequential patterns. Recently developed protein language models, inspired by large language model architectures, treat protein sequences analogously to natural language and generate rich, contextual embeddings. The aim of this study is to evaluate the extent to which representations derived from the ESM-2 protein language model improve protein function prediction performance compared to traditional feature-engineering-based methods.

A balanced, binary pilot dataset was constructed from the UniProt/Swiss-Prot database, with hydrolase enzymes (EC 3.-.-.-) designated as the positive class and non-enzymatic proteins as the negative class. Two classification approaches were compared: (i) a logistic regression baseline trained on amino acid composition features, and (ii) a classifier trained on sequence representations extracted from the ESM-2 model. Model performance was assessed using accuracy, F1-score, and AUC metrics.

Preliminary expectations indicate that protein language model-based representations will yield a notable improvement in classification performance relative to traditional features, as these models learn evolutionary and structural information directly from sequence context, offering a richer representational space. The results of this study are expected to demonstrate that large language model-based approaches can serve as a powerful and efficient alternative to conventional methods in computational biology, particularly under conditions of limited labeled data. These findings aim to provide bioinformatics researchers with concrete evidence of the practical applicability of LLM-based tools and to establish a foundation for future multi-class and multi-omics applications.

Keywords: *Bioinformatics, large language models, protein language models, function prediction, computational biology.*

AI-Based Analysis Of Cattle Genomic Data For Climate Adaptation: Identifying Genetic Regions Associated With Heat Stress Tolerance

Gülşah Keklik[1*]

[1]İstanbul Nişantaşı University, Turkiye

*Corresponding author e-mail: gulsah.keklik@nisantasi.edu.tr

Abstract

Climate change directly affects the livestock sector through rising temperatures and increasingly frequent heat stress events, reducing milk and meat productivity and negatively impacting animal welfare. In this context, identifying genotypes tolerant to heat stress holds strategic importance for climate-resilient breeding programs. The volume and complexity of genomic data make it difficult for traditional statistical methods to fully capture these relationships, necessitating the adoption of artificial intelligence and machine learning-based approaches. The aim of this study is to identify the distinguishing characteristics of genetic regions known to be associated with heat stress tolerance in cattle (Bos taurus) genomic data using machine learning methods, and to evaluate the potential contribution of this approach to climate-resilient breeding efforts.

In this study, gene regions previously associated with heat stress tolerance in the literature (such as the HSP70 gene family and the SLICK gene) were designated as the positive reference group, while randomly selected, unrelated genomic regions served as the control group. Feature importance analysis was performed using the Random Forest algorithm on sequence data obtained from open-access databases such as Ensembl and NCBI, aiming to identify sequence-level and genomic features that distinguish between the two groups. The findings are expected to show that regions known to be associated with heat tolerance differ significantly from the control group in terms of specific sequence patterns.

The results of this study aim to demonstrate that AI-assisted genomic analysis approaches can serve as a rapid and cost-effective tool for identifying genetic biomarkers associated with climate adaptation. These findings may contribute to the prioritization of candidate genetic regions for livestock breeding programs and lay the groundwork for future multi-species (ruminant) comparative analyses.

Keywords: *Data science, Bioinformatics, Livestock, Climate adaptation, Heat stress tolerance.*

The Volumetric Sharma-Mittal Entropy Approach for Parameter-Independent Feature Selection

Nida Oruç Ünal[1*], Muzaffer Göztaş[1], Doğan Yıldız[1]

[1] Yıldız Technical University, Faculty of Arts and Sciences, Department of Statistics, 34000, Istanbul, Türkiye

*Corresponding author e-mail: nida.oruc@std.yildiz.edu.tr

Abstract

Feature selection is a critical step in terms of model performance and interpretability. In entropy-based feature selection approaches, the discretization of continuous variables and the fixation of parameters at specific values in generalized entropy measures may cause the resulting feature rankings to depend on these choices. This study proposes a feature ranking approach based on the two-parameter Sharma-Mittal entropy, which unifies Rényi and Tsallis entropies within a common framework and requires neither discretization nor a single fixed parameter choice. The marginal, joint, and conditional densities of continuous target and explanatory variables are estimated using Gaussian kernel density estimation. Sharma-Mittal conditional entropy and information gain are then evaluated over the entire specified α-β parameter region and volumetrically integrated over the resulting surfaces. Within this framework, PICSME represents the conditional uncertainty volume, PIGSME represents the total information gain volume, and NIGSME represents the information gain normalized by the total entropy volume of the target variable. Thus, variable importance is evaluated through the holistic information structure across the entire parameter space rather than through local entropic behavior at a specific parameter pair.

Keywords: *Feature Selection, Sharma-Mittal Entropy, Kernel Density Estimation, Information Gain, Continuous Variables*

Evaluation of the Impact of Using Artificial Intelligence on Unit Cost in Solution Quality Control for Tissue Processing Systems

Serdar Aday[1*]

[1]İstanbul Üniversitesi, İstanbul Tıp Fakültesi, TıbbiPatoloji A.D., 34000 İstanbul, Türkiye

*Corresponding author e-mail: **saday@istanbul.edu.tr**

Abstract

This retrospective study evaluated the impact of AI-assisted tissue processing on chemical reagent consumption and hazardous waste generation in histopathology laboratories. Conventional systems used between 2015 and 2019 were compared with an AI-assisted system used between 2020 and 2024. Comparable numbers of biopsy specimens were processed in both periods. Chemical consumption decreased from 125 L to 82 L per biopsy, representing an approximately 34.4% reduction. These findings suggest that AI-assisted tissue processing may improve resource efficiency and laboratory sustainability while maintaining a comparable workload. However, prospective multi center studies are required to confirm these findings and assess long-term operational, economic, environmental, and clinical benefits.

This study aimed to evaluate the effect of artificial intelligence (AI)-assisted tissue-processing technology on chemical reagent consumption and hazardous waste generation in histopathology laboratory practice. The processing capacity and consumable requirements of conventional and AI-assisted closed-system tissue processors were also compared. A retrospective comparison was performed between conventional tissue-processing systems used between 2015 and 2019 and an AI-assisted tissue-processing system used between 2020 and 2024. During the first period, three conventional tissue processors with a total capacity of 700 tissue blocks were evaluated. During the second period, an AI-assisted tissue-processing system with a capacity of 900 tissue blocks was evaluated. The total number of biopsies processed, chemical reagent consumption, and hazardous waste generation were compared between the two periods. Statistical analyses were performed using appropriate comparative statistical methods.

Between 2015 and 2019, an average of 41,390 biopsy specimens were processed using conventional tissue-processing systems, with an average chemical reagent consumption of 125 L per biopsy. Between 2020 and 2024, an average of 40,824 biopsy specimens were processed using the AI-assisted system, with chemical consumption decreasing to 82 L per biopsy. Thus, despite the processing of a comparable number of biopsy specimens, the AI-assisted system was associated with a substantial reduction in chemical reagent consumption. Based on the reported values of 125 L and 82 L per biopsy, the reduction in chemical consumption was approximately 34.4%.

The findings suggest that implementation of an AI-assisted tissue-processing system may substantially reduce chemical reagent consumption and hazardous waste generation while maintaining a comparable biopsy-processing workload. Although the higher technological and initial operational costs of AI-assisted systems should be considered, reductions in reagent consumption and hazardous waste may provide important economic and environmental benefits for histopathology laboratories. Further prospective, multi center studies are warranted to confirm these findings and to determine the long-term clinical, operational, economic, and environmental impact of AI-assisted tissue processing.

Keywords: *Histopathology, artificial intelligence, hazardous waste.*

Clinical and Economic Evaluation of Artificial Intelligence-Supported Liquid-Based Cervical Cytology Systems

Serdar Aday[1*]

[1]İstanbul Üniversitesi, İstanbul Tıp Fakültesi, TıbbiPatoloji A.D., 34000 İstanbul, Türkiye

*Corresponding author e-mail: **saday@istanbul.edu.tr**

Abstract

This study compares conventional Pap smear testing with AI-assisted liquid-based cervical cytology in terms of cost and diagnostic effectiveness. Although conventional smear testing has a lower per-test cost, AI-assisted liquid-based systems may provide important advantages in sensitivity, specificity, diagnostic reliability, and evaluation time. Automated liquid-based systems may facilitate more effective assessment of cervical cells and identification of high-risk cellular abnormalities, thereby supporting earlier and more reliable diagnosis. This approach may also contribute to reducing diagnostic workload, accelerating the diagnostic process, and improving patient management. Overall, despite the higher initial cost, AI-assisted liquid-based cytology may offer potential long-term benefits in terms of clinical effectiveness, diagnostic safety, and more efficient utilization of healthcare resources.

AI-supported liquid-based cervical cytology systems assist in evaluating digital Pap test images for cervical cancer screening and in detecting atypical cells and findings associated with CIN, HSIL, and carcinoma. Their purpose is not to replace cytologists or pathologists or provide an autonomous diagnosis, but to support diagnostic decision-making. AI algorithms identify and prioritize clinically relevant cells or cell groups, known as "Objects of Interest" (OOIs), for further review. This targeted approach reduces visual screening workload and may decrease the risk of overlooking high-risk cells, particularly those associated with high-grade lesions. Clinical performance should be assessed using sensitivity, specificity, false-negative and false-positive rates, predictive values, AUROC, examination time, and workload. Reported studies indicate high diagnostic performance. However, statistical significance of a 93% success rate versus a 90% reference requires confirmation using the sample size and number of successful cases. The reported conventional/liquid-based cytology cost ratio is 0.02, indicating that liquid-based testing costs approximately 50 times more per test. Based on stated Turkish SUT reimbursement values, the conventional-to-liquid-based ratio is approximately 0.50. Reported examination time was 138 seconds manually versus approximately 70 seconds with AI-supported liquid-based cytology, corresponding to a 49.3% reduction.

Overall, AI-supported liquid-based cervical cytology is a complementary digital decision-support technology that facilitates identification and prioritization of high-risk cells, potentially reduces workload and examination time, and supports safer and more efficient diagnostic evaluation.In conclusion, AI-assisted liquid-based cervical cytology systems should be positioned as digital decision-support technologies that aid the diagnostic assessment performed by cytologists and pathologists,

rather than as autonomous systems that render the final diagnosis. The primary objectives of the clinical use of these technologies are to facilitate the detection of high-risk cells, alleviate the visual screening burden, reduce review times, and crucially mitigate the risk of false-negative results associated with overlooking high-grade lesions. The final cytopathological evaluation and clinical decision must be carried out by a competent cytologist or pathologist, taking into account the patient's clinical characteristics and relevant cytopathological diagnostic criteria alongside the AI system's outputs.

Keywords: *Pap smear, artificial intelligence, cost, effectiveness.*

Data Balancing and Hybrid Deep Learning in the Prediction of Unbalanced Hypothyroidism: A Leakage-Controlled Comparison

M. Nisa Özen [1], Melike Şişeci Çeşmeli[2*]

[1] Burdur Mehmet Akif Ersoy Üniversitesi, Institute of Natural Sciences, Department of Software Engineering, Department of Software Engineering, Burdur, Türkiye

[2] Burdur Mehmet Akif Ersoy Üniversitesi, Bucak Faculty of Computer Science and Information Technology, Department of Software Engineering, Burdur, Türkiye

*Corresponding author e-mail: melikesiseci@mehmetakif.edu.tr

Abstract

Reliably estimating the minority class in imbalanced medical datasets is a significant research problem, as clinically important cases may be overlooked despite high overall accuracy. It has been demonstrated that classification models are prone to learning the majority class in datasets with a limited number of positive cases. Consequently, these models may perform poorly at determining the disease class, even when they achieve high accuracy. In this study, a hybrid neural network model—which combines residual-linked dense layers and an auto-attention mechanism—was compared with data-balancing approaches for predicting hypothyroidism using the 'allhypo' subset of the UCI Thyroid Disease dataset. The training set, consisting of 2,800 records, has been divided into two sections: 2,238 for development and 562 for validation. This ensures that duplicate records with the same feature signature are not split between the two sections. The official test set, which comprises 972 records, is completely isolated from the model selection process. The original data set was evaluated using random oversampling (ROS), the synthetic minority oversampling technique for nominal and continuous variables (SMOTENC), which preserves mixed-type variables, and conditional tabular generative adversarial networks (CTGAN), with seven classifiers. The primary metric utilized to evaluate model performance was PR-AUC, with the balancing method for each model family selected exclusively on the validation set. The primary configuration selected for validation was the Random Forest, which was trained without cross-validation. The model demonstrated a PR-AUC of 0.9964, a ROC-AUC of 0.9997, a sensitivity of 0.9577, a specificity of 0.9855, and an F1 score of 0.9714 on the independent test set. The 95% confidence interval for the PR-AUC was 0.9885–1.0000. The XGBoost model, balanced using CTGAN, achieved a sensitivity and F1-score of 0.9859; however, its PR-AUC remained at 0.9574. The proposed hybrid model achieved a test PR-AUC of 0.8690, indicating it did not outperform tree-based methods. The findings indicate that, with limited and heterogeneous clinical data, synthetic balancing or more complex architectures do not automatically guarantee superior performance. Rather, model selection carried out within a split-validation and independent-test framework is decisive for the reliability of the results.

Keywords: *Class imbalance, Hypothyroidism, SMOTENC, CTGAN, Table data*

Data-Driven Mortality Jump Forecasting with a Non-Explosive Alpha-Series Recurrence Model

Selin Özen[1]*, Mustafa Hilmi Pekalp[1]

[1]Department of Actuarial Sciences, Faculty of Applied Sciences, Ankara University, Ankara, Türkiye

*Corresponding author e-mail: ozens@ankara.edu.tr

Abstract

Catastrophic mortality events generate abrupt departures from long-term mortality trends and can materially affect demographic forecasts and insurers' upper-tail liabilities. Standard jump specifications generally assume either a constant annual occurrence probability or independent and identically distributed recurrence times, limiting their ability to represent systematic changes in event timing. This study develops a data-driven, non-explosive alpha-series recurrence model within the Lee-Carter framework. The model allows inter-jump durations to shorten, remain stationary, or lengthen with event order while preventing infinitely many events from accumulating within a finite forecasting horizon. Mortality jumps are detected from total-population mortality indices using the Chen-Liu outlier procedure, and the model jointly estimates their sex-specific effects with correlated female and male innovations. An exact likelihood for annual recurrence durations incorporates event order and terminal right censoring, while a parametric bootstrap likelihood-ratio test addresses the limited number of country-specific events. The framework is evaluated using Human Mortality Database records for Denmark, Spain, Sweden, Switzerland, the UK, and the USA. Using 2019 as the calibration cut-off, we assess predictive performance through point and probabilistic forecasts and examine actuarial implications using simulated term-life run-off liabilities and VaR- and TVaR-based tail-capital measures. The findings reveal substantial cross-country heterogeneity in mortality-jump recurrence. The proposed model achieves the lowest in-sample BIC in all six countries and provides favourable out-of-sample performance during the COVID-19 stress period. Although alternative arrival mechanisms produce similar expected liabilities, their differences are more pronounced in tail-capital measures. The results demonstrate that flexible recurrence modelling is particularly important for data-driven mortality forecasting and actuarial tail-risk analytics.

Keywords: *Mortality forecasting, Alpha-series process, Event recurrence, Parametric bootstrap, Tail-risk analytics*

Evaluation of Structural Health Monitoring in CARALL Utilizing Low-Velocity Impact Testing Supported by Graphene-Enhanced Cellulose Sensor

Tugay Üstün[1]*, Ebru Saraloğlu Güler[2]

[1] Başkent University, Kahramankazan Vocational School, 06980, Ankara, Türkiye
[2] Başkent University, Faculty of Engineering, 06790, Ankara, Türkiye

*Corresponding author e-mail: tugayustun@baskent.edu.tr

Abstract

Cellulose-based materials have become an attractive platform for various biosensor applications because of their robust, flexible mechanical properties combined with their electrical properties. They are generally being developed as flexible biosubstrates, electrodes, and active components for flexible electronics. This has enabled wearable sensors for continuous health monitoring. Furthermore, they have achieved significance in applications involving pressure and strain sensors. However, efforts are ongoing to improve the performance, stability, repeatability, and sensitivity of these sensors. For this reason, researchers are examining the integration of diverse nanomaterials. As a carbon-derived nanomaterial, graphene has shown to augment the performance of biosensors because of its outstanding electrical, mechanical, thermal, and optical attributes.

In this study, cellulose—used in sensor applications—was produced in paper form. Following traditional paper-making methods, we produced cellulose paper for sensor applications and added graphene to impart electrical conductivity. One of the greatest advantages of cellulose paper is that it is scalable, low-cost, and environmentally friendly. The cellulose papers, produced by adding graphene at varying weight percentages to impart conductive properties, were incorporated into carbon fiber-reinforced aluminum laminate (CARALL) composites. These composites face challenges such as weak metal-composite interfaces and a lack of embedded damage-detection capability. Real-time structural health monitoring is essential for identifying impact-related damage. Employing these cellulose paper sensors, low-velocity impact assessments were performed, and damage was seen in situ by piezoresistive electrical resistance measurements (ΔR/R). It is believed that enhancing biosensor performance, combined with artificial intelligence support, will play a significant role in determining the type of damage and enable long-term monitoring.

Keywords: *Cellulose, Health monitoring, Graphene, Impact, Biosensor*

Conditional Score for Logistic Regression with Measurement Error

N. Saide Allahverdi[1*]

[1]Ankara University, Department of Statistics, Ankara, Türkiye

*Corresponding author: nur.allahverdi@bilecik.edu.tr

Abstract

In regression models, measurement error in explanatory variables can lead to biased parameter estimates and reduce the reliability of the results. This situation makes the use of appropriate correction methods even more important, particularly in nonlinear models such as logistic regression, where the effects of measurement error manifest in more complex ways. This study aims to examine the Conditional Score approach in a logistic regression model under the classical additive measurement error framework, considering different sample sizes and levels of measurement error variance. The classical total measurement error structure, where the true explanatory variable X cannot be directly observed—and instead the variable W, which includes measurement error, is observed—is addressed in the form $W = X + U$. It is assumed that the mean of the term U, which represents measurement error, is zero and that it follows a normal distribution. In this study, Monte Carlo simulations will be conducted by varying sample sizes and levels of measurement error variance to examine the performance of the Conditional Score method under different conditions. Parameter estimates obtained using the Conditional Score method will be compared with naive logistic regression estimates that ignore measurement error. The performance of the methods will be evaluated using the mean estimate of the slope parameter, bias, and mean squared error (MSE) metrics. Thus, the study will examine how measurement error affects parameter estimates obtained from logistic regression under different conditions and will evaluate the performance of the Conditional Score method in reducing bias caused by measurement error.

Keywords: *Logistic Regression, Measurement Error, Conditional Score, Monte Carlo Simulation, Bias*

Comparative Correlation Analysis of Relationships Among Postharvest Quality Attributes in Pepper and Tomato[1]

Osman Nuri Öcalan[1]*, Ayşegül Tülay Pala[2]

[1,2] Tokat Gaziosmanpaşa Üniversitesi, Ziraat Fakültesi, Bahçe Bitkileri Bölümü, 60100, Tokat, Türkiye

*Corresponding author e-mail: osmannuri.ocalan@gop.edu.tr

Abstract

Physical, chemical, and biochemical changes occurring in fruits during the postharvest period are interrelated, and elucidating these relationships enables a more comprehensive assessment of quality losses. This study aimed to determine the relationships among postharvest quality attributes in 'Bulut F1' pepper and 'Alsancak F1' tomato cultivars and to compare the correlation patterns between the two crops. For this purpose, data on fruit firmness, decay rate, weight loss, pH, titratable acidity (TA), total soluble solids (TSS), total phenolic content (TPC), total flavonoid content (TFC), and total antioxidant capacity (TAC) were evaluated, and the relationships among these parameters were examined using Pearson correlation analysis. The resulting correlation coefficients were visualized using heat maps. The analysis revealed that some relationships among quality attributes showed similar patterns in both crops, whereas others differed depending on the crop. Strong negative correlations between pH and TA were observed in both pepper ($r = -0.583$, $p < 0.01$) and tomato ($r = -0.688$, $p < 0.01$). Similarly, fruit firmness was negatively and significantly correlated with weight loss in both pepper ($r = -0.349$, $p < 0.01$) and tomato ($r = -0.358$, $p < 0.05$). In contrast, a strong positive correlation between TFC and TAC was detected in pepper ($r = 0.646$, $p < 0.01$), whereas this relationship was not statistically significant in tomato. In tomato, a strong negative correlation was observed between firmness and decay ($r = -0.613$, $p < 0.01$), while firmness was positively correlated with TA ($r = 0.500$, $p < 0.01$); neither relationship was statistically significant in pepper. Furthermore, the relationship between firmness and TPC was significant in both crops but showed opposite directions, being positive in pepper ($r = 0.392$, $p < 0.01$) and negative in tomato ($r = -0.442$, $p < 0.01$). In conclusion, although pepper and tomato exhibited common correlation patterns among some fundamental physicochemical quality attributes, the relationships between physical quality traits and bioactive compounds varied depending on the crop. These findings highlight the importance of considering not only individual parameters but also their interrelationships when evaluating postharvest quality changes and demonstrate that correlation-based data analysis can provide a useful approach for comparing quality dynamics across different crops.

Key words: *Pearson correlation, heat map, postharvest quality, Capsicum annuum L., Solanum lycopersicum L.*

[1] The pepper data evaluated in this study were obtained within the scope of the project supported by the TÜBİTAK 2209-A University Students Research Projects Support Program under **Project No. 1919B012506527**.

Gravity Model of Trade with Network Analysis – ERGM Approach

Cagay Coskuner[1]*

[1] Eastern Mediterranean University, Gazi Magusa, Kuzey Kibris

*Corresponding author e-mail: cagay.coskuner@emu.edu.tr

Abstract

This paper attempts to investigate the factors which impacts the export performance from home country to partner countries. To this end, it takes the gravity model of trade as a benchmark, and attempts to support the model by using Network Analysis and ERGM (exponential random graph models) rather than standard panel data econometrics. Gravity model of trade states that the trade volume (that is the exports and imports) between any two countries is directly proportional to the GDP sizes of both countries and inversely proportional to the distance between the countries. In the augmented model, linguistic and cultural similarities as well as similarities in economic conditions such as GDPPC values (as stated in Linder theory) between the partner countries positively contribute to the trade volume. In this paper, to test these theories, we set up a network analysis. To this end, we use 43 countries as nodes, and the exports from each node to all other 42 nodes as individual possible links (dyads). We add GDP, GDPPC, Continent and Economic Status as the node attributes. In the initial network analysis, modularity tests for the Economic Status and assortativity tests for GDPPC show that trade links are more likely to exist for countries with similar economic conditions, thus supporting the Linder theory. Similarly, modularity test for the continents of the countries show that being from the same continent (thus having shorter distance) increases the likelihood for trade links between any two countries. Finally, ERGM models provide support for the gravity model of trade as they suggest that the trade links are more likely to exist with increasing GDP values, as well as due to node.matching for continents.

Keywords: *Gravity Model of Trade, Network Analysis, Exponential Random Graph Models (ERGM Estimations), Structural Attributes, Linder Theory*

Predicting Solar and Geomagnetic Activity Using Markov Chain Monte Carlo Techniques: An AI-Driven Analysis

Danish Hassan[1], Muhammad Danish Khan[2], Muhammad Fahim Akhter[3]

[1]Department of Applied Sciences, National Textile University Karachi Campus, Karachi, Pakistan
[2]Department of Computer Sciences, SZABIST Karachi Campus
[3]Faculty of Engineering Science and Technology Indus University, Karachi

*Corresponding author e-mail: danish10ansari@gmail.com

Abstract

The compatibility of solar-geomagnetic activity is calculated using the Markov chain Monte Carlo (MCMC) method, an application of artificial intelligence (AI). The study covers various sunspots durations and related geomagnetic durations by employing a Markov 4-states transition matrix. The classification of the matrix is determined by the maximum to minimum range of each cycle in the selected data sets. This AI-driven analysis establishes a behavioral relationship and draws samples from complex distributions of sunspots and geo index using the MCMC technique. The expected number of sunspots and geomagnetic activity for each state remains almost constant, demonstrating the AI's ability to analyze and predict patterns in complex datasets. Significant results of predicted frequencies during different years are aligned with observed frequencies of sunspot and geo index, indicating the AI's effectiveness in predictive modeling. Based on model-generated results, the study suggests that the future behavior of the studied parameters will follow a similar progression, showcasing the application of AI in forecasting natural phenomena. Additionally, the effects of the sun on geomagnetic index are confirmed, and the frequency for upcoming cycles can be predicted by this AI-driven study.

Keywords: *Sunspots, Geomagnetic activity, Markov process, Transition matrix*

Can Renewable Energy Consumption Drive Environmental Sustainability in Pakistan? The Role of Technological Progress and Economic Expansion

Allah Dad[1], Oluwatoyin Abidemi Somoye[2], Hüseyin Özdeşer[2]

[1]Department of Banking and Finance, Near East University, Near East Boulevard, Nicosia, TRNC
[2]Department of Economics, Near East University, Near East Boulevard, Nicosia, TRNC

*Corresponding author e-mail: 20245322@std.neu.edu.tr

Abstract

Environmental sustainability is important because of the preservation of the ecosystem. As a result, this research investigates the impact of renewable energy consumption, technological progress, digitalization, and economic growth on environmental sustainability in Pakistan from 1990 to 2022. Pakistan is used as a case study because it is one of the most climate-vulnerable countries in the world. Also, as a contribution to existing literature, this research used the load capacity factor (LCF) as a proxy for environmental sustainability. The LCF captures the demand and supply side of the ecosystem. In addition, this study used innovative methods such as Quantile Augmented Dickey-Fuller (QADF) and Quantile Phillips-Perron (QPP) unit root tests, as well as Quantile-on-Quantile Regression (QQR), Wavelet Quantile Regression (WQR), and Wavelet Quantile Correlation (WQC). The results demonstrate that technological advancement, the use of renewable energy sources, and economic growth drive the sustainability of the environment, especially at the upper quantiles. In contrast, digitalization exhibits a predominantly negative association with LCF. Comprehensively, the findings show that the ecological balance of Pakistan relies on the harmonious combination of technological innovation, renewable energy development, and digitalization that is energy-efficient. The study provides important policy implications in the development of a low-carbon, innovative, and digitally sustainable economy.

Keywords: *Digitalization; Renewable Energy Consumption; Technological Progress; Economic Growth; Environmental Sustainability*

Trustworthy and Explainable Federated Learning for Privacy-Preserving ECG Classification in Smart-Hospital IoT

Ferdi Güler[1], Ummahan Temel[2*], Fatma Altıntaş[3], Batuhan Selamoğlu[4]

[1]Istanbul Arel University, Vocational School, Computer Programming, Istanbul, Türkiye
[2]Istanbul Arel University, Vocational School, Biomedical Device Technology, Türkiye
[3]Istanbul Arel University, Vocational School, Biomedical Device Technology, Türkiye
[4]Istanbul Arel University, Vocational School, Electronics Technology, Istanbul, Türkiye

*Corresponding author e-mail: ummahantemel@arel.edu.tr

Abstract

Connected electrocardiography (ECG) devices enable smart hospitals to move cardiac screening and rhythm surveillance closer to the bedside, yet centralized model development requires sensitive physiological signals to be pooled and creates a single point of failure. Federated learning reduces the movement of raw data, but federation alone does not constitute trustworthy artificial intelligence because model updates may leak information, hospital clients commonly exhibit non-independent and non-identically distributed data, and locally learned decision patterns may be poorly calibrated or difficult to explain. This chapter presents a reproducible cross-silo framework for privacy-preserving, explainable, and multi-label ECG classification in smart-hospital Internet of Things environments. The empirical design uses PTB-XL ECG waveforms together with clinically interpretable PTB-XL+ features. Patient-level separation prevents record leakage, while independent, label-skewed, and demographic-skewed partitions emulate heterogeneous hospital clients. Signal preparation includes lead-wise normalization, quality screening, multi-label diagnostic encoding, and class-imbalance handling. A one-dimensional residual network is trained under FedAvg and FedProx, with centralized and local-only learning serving as reference conditions. Clipped client updates, client-level differential privacy, privacy-budget accounting, and secure-aggregation-compatible communication define graded privacy settings. Membership-inference auditing is incorporated to examine whether stronger privacy settings reduce empirical leakage rather than relying exclusively on formal privacy claims. Reliability is strengthened through post-hoc probability calibration and selective prediction, allowing low-confidence ECGs to be deferred for human review. Signal-level explanations are generated with integrated gradients, whereas SHAP is used for clinically interpretable PTB-XL+ features. Explanation quality is evaluated through perturbation-based faithfulness, sparsity, stability across clients and communication rounds, and agreement with established ECG characteristics. The multidimensional evaluation reports macro-AUROC, macro-AUPRC, macro-F1, Brier score, expected calibration error, worst-client performance, age- and sex-stratified performance, privacy budget, attack susceptibility, communication volume, and explanation stability. Ablation analyses isolate the

contributions of privacy mechanisms, aggregation strategy, calibration, and explanation design. Sensitivity analyses further examine whether conclusions remain stable across client counts, privacy budgets, non-IID severity levels, and alternative decision thresholds. The principal methodological contribution is a trustworthiness frontier that relates diagnostic utility, privacy protection, calibration, client equity, resource cost, and explanation quality within a single experimental protocol. This framework moves federated smart-hospital analytics beyond accuracy-only benchmarking and provides a technically auditable basis for selecting operating points for privacy-sensitive connected cardiac monitoring.

Keywords: *Federated learning, explainable artificial intelligence, trustworthy AI, electrocardiography, healthcare IoT.*

Accessible and Explainable Edge AI for Privacy-Preserving Wearable Fall Detection

Ummahan Temel[1], Fatma Altıntaş[2*], Batuhan Selamoğlu[3], Ferdi Güler[4]

[1,2]Istanbul Arel University, Vocational School, Biomedical Device Technology, Türkiye
[3]Istanbul Arel University, Vocational School, Electronics Technology, Istanbul, Türkiye
[4]Istanbul Arel University, Vocational School, Computer Programming, Istanbul, Türkiye

*Corresponding author e-mail: fatmaaltintas@arel.edu.tr

Abstract

Wearable fall detection can shorten emergency response time without the continuous visual surveillance required by camera-based systems. Nevertheless, many reported solutions depend on cloud processing, provide no account of why an alarm was generated, and use random window splits that allow subject-specific motion patterns to leak between training and test sets. Such weaknesses are particularly consequential for users who cannot rely on visual interfaces, including people with visual impairments. This chapter presents an accessibility-oriented, privacy-preserving, and explainable edge-artificial intelligence framework for real-time fall detection from wearable accelerometer and gyroscope signals. The empirical study uses the SisFall and MobiAct datasets to compare an interpretable feature-based classifier with a lightweight one-dimensional convolutional network. A harmonized preprocessing pipeline performs resampling, gravity-component handling, axis normalization, signal segmentation, and class-imbalance control across the two datasets. Signal windows are represented through biomechanical and temporal characteristics, including impact magnitude, orientation change, spectral energy, jerk, and post-event inactivity, while the compact neural model captures patterns that fixed features may miss. Subject-independent validation and cross-dataset transfer are used to measure generalization without contaminating evaluation data with observations from the same participant. Quantization and structured pruning expose the trade-off between predictive performance and edge-resource consumption. Model size, memory footprint, multiply-accumulate operations, and per-window inference latency are therefore reported alongside sensitivity, specificity, balanced accuracy, macro-F1, Matthews correlation coefficient, and false-alarm burden. Because raw inertial streams remain on the wearable or a trusted edge gateway, only the alarm state and minimum contextual information required for escalation leave the local device. Local explanations combine feature attribution with time-segment importance so that every suspected fall identifies the motion characteristics most responsible for the decision. Explanation faithfulness, sparsity, and stability are evaluated under signal noise, axis rotation, sampling variation, and sensor-modality dropout. Robustness experiments also examine how model confidence and false alarms change when a sensor channel becomes unreliable. The decision layer is coupled to a nonvisual interaction workflow based on haptic and spoken cues, explicit confirmation or cancellation, and secure escalation to an authorized caregiver. Accessibility requirements, timing constraints, and failure modes are traced to individual system components instead of being treated as interface decoration. The chapter contributes a reproducible benchmark and reference architecture that

jointly evaluates cross-user generalization, cross-dataset robustness, resource efficiency, privacy, explanation quality, and accessible human-AI interaction in wearable cyber-physical systems.
Keywords: *Edge AI, fall detection, wearable sensors, accessibility, visual impairment, privacy-preserving AI.*

Role-Adaptive Explainable Intrusion Detection for Trustworthy Internet of Medical Things Networks

Batuhan Selamoğlu[1], Ferdi Güler[2*], Ummahan Temel[3], Fatma Altıntaş[4]

[1]Istanbul Arel University, Vocational School, Electronics Technology, Istanbul, Türkiye
[2]Istanbul Arel University, Vocational School, Computer Programming, Istanbul, Türkiye
[3]Istanbul Arel University, Vocational School, Biomedical Device Technology, Türkiye
[4]Istanbul Arel University, Vocational School, Biomedical Device Technology, Türkiye

*Corresponding author e-mail: ferdiguler@arel.edu.tr

Abstract

Internet of Medical Things (IoMT) devices improve the observability and responsiveness of clinical environments while simultaneously expanding the attack surface of safety-critical networks. Existing intrusion detectors are commonly optimized for aggregate accuracy, although severe class imbalance, device-specific artifacts, data leakage, miscalibrated confidence, and opaque predictions can make apparently strong models unreliable in operational use. This chapter presents a role-adaptive explainable intrusion-detection framework that connects network evidence to the distinct information needs of cybersecurity teams and clinical operations staff. The empirical analysis uses CICIoMT2024, a multi-protocol benchmark containing traffic from 40 IoMT devices exposed to 18 cyberattacks across communication contexts that include Wi-Fi, MQTT, and Bluetooth. Data preparation removes identifiers and highly deterministic fields that could create shortcut learning, controls duplicate flows, and preserves class distributions for clinically relevant rare attacks. Gradient-boosted trees, random forests, and a compact neural classifier are evaluated for binary, attack-family, and fine-grained detection. Device-disjoint and attack-scenario splits are used in addition to conventional stratified evaluation, preventing traffic from the same device or experimental scenario from appearing on both sides of a supposedly independent test. Class-weighted learning and calibrated probabilities address imbalance and unreliable confidence, while an uncertainty-based abstention rule routes ambiguous events for manual inspection. SHAP-based local evidence is combined with a domain interpretation layer that translates technical traffic features into stakeholder-specific alerts. Cybersecurity analysts receive feature contributions, traffic directionality, protocol context, confidence, and comparable events; clinical operations staff receive the affected device category, threat type, potential service consequence, severity, and a bounded initial-response recommendation. Detection quality is assessed with macro-F1, per-class AUPRC, false-positive rate, balanced accuracy, calibration error, and inference latency. Trustworthiness is further examined through device-holdout generalization, attack perturbation tests, worst-class performance, confidence degradation, and explanation faithfulness, stability, and sparsity. Counterfactual traffic modifications test whether changing the features highlighted by an explanation

produces the expected change in model behavior. A structured alert-quality rubric evaluates whether explanations are traceable, role-appropriate, concise, and operationally actionable without assuming that technical feature importance alone constitutes human understanding. Ablation experiments separate the effects of calibration, class weighting, abstention, and role-adaptive interpretation. The central contribution is an evaluation and translation pipeline that treats prediction, uncertainty, explanation, stakeholder context, and response guidance as coupled components of trustworthy IoMT defense, while preserving the boundary between automated detection and accountable human action.

Keywords: *Internet of Medical Things; intrusion detection; healthcare cybersecurity; uncertainty calibration; human-centered explanations.*

Trustworthy and Temporally Explainable Early-Warning Systems for Sepsis Using ICU IoT Data

Fatma Altıntas[1], Batuhan Selamoğlu[2*], Ferdi Güler[3], Ummahan Temel[4]

[1]Istanbul Arel University, Vocational School, Biomedical Device Technology, Türkiye
[2]Istanbul Arel University, Vocational School, Electronics Technology, Istanbul, Türkiye
[3]Istanbul Arel University, Vocational School, Computer Programming, Istanbul, Türkiye
[4]Istanbul Arel University, Vocational School, Biomedical Device Technology, Türkiye

*Corresponding author e-mail: batuhanselamoglu@arel.edu.tr

Abstract

Continuous intensive-care monitoring produces multivariate physiological time series that can support earlier recognition of sepsis, but a high discrimination score does not by itself guarantee a safe early-warning system. Miscalibrated probabilities, irregular sampling, missing measurements, class imbalance, distribution shift, unstable explanations, and excessive false alarms can undermine clinical usefulness even when conventional accuracy measures appear favorable. This chapter presents a trustworthy and temporally explainable framework for sepsis prediction from intensive-care Internet of Things data. The empirical study uses the PhysioNet/Computing in Cardiology Challenge 2019 cohort, with time-stamped physiological, laboratory, demographic, and sepsis-label data. Patient-level separation is maintained between development, calibration, and evaluation data, and hospital subsets are retained for cross-site generalization analysis. The preprocessing pipeline applies training-derived scaling, clinically bounded imputation, missingness indicators, and temporal window construction without using future information. Regularized logistic regression and gradient-boosted trees provide transparent reference models, while gated recurrent units and temporal convolutional networks represent nonlinear sequential alternatives. Class weighting and threshold selection address low sepsis prevalence. Probability calibration and uncertainty-aware selective prediction identify cases that require human review rather than forcing a confident automated alarm. Temporal explanations combine variable attribution with time-segment occlusion, revealing both which measurements influence risk and when their influence emerges during the hours preceding sepsis onset. Evaluation integrates AUROC and AUPRC with the challenge utility score, Brier score, expected calibration error, warning lead time, sensitivity at fixed alarm rates, false alerts per 100 patient-hours, and age- and sex-stratified performance. Robustness tests simulate increased missingness, sensor dropout, delayed laboratory results, measurement noise, and cross-hospital distribution shift. Explanations are assessed for perturbation-based faithfulness, temporal localization, stability under small input changes, and consistency with model behavior. Decision-curve and workload analyses connect statistical performance to net clinical benefit and alarm burden across operationally relevant thresholds. Ablation experiments quantify the contribution of

missingness modeling, calibration, uncertainty handling, and temporal explanation. The chapter's methodological contribution is a reproducible evaluation protocol that aligns early prediction, calibration, robustness, temporal explainability, subgroup analysis, and human oversight. It reframes sepsis warning as a safety-critical cyber-physical decision problem rather than a stand-alone classification task and provides a defensible basis for comparing clinically oriented early-warning systems.

Keywords: *Sepsis prediction, intensive care unit, temporal explainability, early-warning Systems, clinical decision support.*

Comparison of CatBoost, LightGBM, and XGBoost Algorithms for Credit Default Prediction

Semih Ay[1], **Onuralp Gayret**[2]

[1] Marmara University, Faculty of Engineering, Department of Computer Engineering, 34854, İstanbul, Turkey

*Corresponding author e-mail: semih.ay@marmara.edu.tr

Abstract

Credit default prediction is an important problem for financial institutions, as identifying customers with a higher risk of payment default can support more effective credit risk management and decision-making. In this study, CatBoost, LightGBM, and XGBoost algorithms are compared for predicting whether credit card customers will default on their payments in the following month. The study uses the publicly available Default of Credit Card Clients dataset from the UCI Machine Learning Repository. The dataset contains information from 30,000 customers, including credit limits, demographic characteristics, previous payment behavior, bill statement amounts, and payment history. The target variable indicates whether a customer defaults on payment in the following month. To ensure a fair comparison, all models are evaluated using the same training and test split. Their performance is assessed using accuracy, precision, recall, F1-score, and ROC-AUC. In addition to predictive performance, training and inference times are also considered to evaluate the computational efficiency of each model. Hyperparameter optimization is applied separately to each algorithm in order to obtain suitable model configurations. Since the number of customers who default is relatively lower than the number of non-defaulting customers, accuracy alone may not provide a complete view of model performance. For this reason, particular attention is given to recall, F1-score, and ROC-AUC, which provide more useful information about how effectively the models identify customers at risk of default. The main objective of this study is to examine the differences between CatBoost, LightGBM, and XGBoost in terms of both predictive performance and computational cost, and to provide a practical comparison of these methods for credit risk prediction on tabular financial data.

Keywords: *Gradient Boosting, Credit Risk, Machine Learning, Classification, Financial Data*

Data-Driven Predictive Intelligence for Decision Support Systems: Machine Learning, Statistical Modelling, and Adaptive Analytics

Lawrence A. Farinola[1*]

[1]Final International University, Girne, North Cyprus

*Corresponding author e-mail: lawrance.farinola@final.edu.tr

Abstract

The increasing availability of heterogeneous and high-volume data has transformed decision-making from experience-driven processes toward data-driven and intelligent approaches. Modern decision support systems increasingly integrate statistical modelling, machine learning, predictive analytics, and adaptive analytical techniques to transform complex data into actionable intelligence. However, many existing approaches remain primarily focused on prediction, with limited integration of predictive modelling, continuous adaptation, and decision-oriented intelligence. This chapter examines the role of data-driven predictive intelligence in the development of modern decision support systems, with particular emphasis on the integration of statistical modelling, machine learning, predictive analytics, and adaptive analytics. It discusses how descriptive, predictive, and prescriptive analytical approaches can be combined to support evidence-based decision-making across changing organizational and technological environments. The chapter further proposes an integrated framework linking data acquisition and preparation, statistical and machine learning models, predictive analytics, model monitoring, adaptive learning, and decision-support mechanisms. Particular attention is given to challenges associated with data quality, model drift, interpretability, uncertainty, security, and responsible use of intelligent analytical systems. The chapter also considers emerging developments in automated machine learning, adaptive decision systems, and human–AI collaboration. By connecting predictive modelling with adaptive and decision-oriented analytics, the proposed framework provides a structured perspective for developing intelligent, data-driven decision support systems.

Keywords: *Data Science; Predictive Analytics; Machine Learning; Statistical Modelling; Decision Support Systems; Adaptive Analytics; Data-Driven Intelligence*

Evaluation of Economic Activity Groups According to Occupational Health and Safety Indicators: Assessing the Effect of Criterion Weighting Methods on COBRA-Based Rankings

Pelin Toktas[1*], Elif Kılıç Delice[2]

[1]Başkent University, Engineering Faculty, Industrial Engineering, 06790, Ankara, Türkiye

[2]Atatürk University, Engineering Faculty, Industrial Engineering, 25240, Erzurum, Türkiye

*Corresponding author e-mail: ptoktas@baskent.edu.tr

Abstract

Occupational accidents, occupational diseases, and the associated losses in the field of occupational health and safety show significant differences among economic activity groups. Evaluating these differences with a holistic approach is important for identifying high-risk activity groups in terms of occupational health and safety and prioritizing preventive policies. The aim of this study is to evaluate and rank different activity groups based on occupational health and safety indicators using a multi-criteria decision-making (MCDM) approach. The data used in this study were compiled from the 2025 Social Security Institution (SGK) Occupational Accident Statistics. In the study, economic activity groups were ranked in terms of the number of workplaces, the number of insured individuals, the total number of days of incapacity due to work accidents, the number of insured individuals suffering from occupational diseases, the number of insured individuals who lost their lives due to work accidents, and the total number of diseases. The economic activity groups are ranked by the Comprehensive Distance Based on Ranking (COBRA) method, and the weighting of the considered criteria is performed by the Mean Weight (MW), Criteria Importance Through Intercriteria Correlation (CRITIC), Entropy, and Standard Deviation (SD) methods. The same alternative is ranked first for all criterion weighting methods, despite the overall ranking being different. This result demonstrates the stability of the COBRA method in determining the best alternative. The results also allowed for the identification of economic activities with higher priority in terms of occupational health and safety. This study is expected to guide decision-makers in directing limited resources to high-risk activity areas, prioritizing inspection and prevention strategies, and developing sectoral occupational health and safety policies.

Keywords: *Economic activities, COBRA, CRITIC, MW, Entropy, SD.*

The Role of Business Intelligence Systems in Strategic Decision-Making: A Conceptual Analysis of Small and Medium-Sized Enterprises

Kazim Syed[1], Zubaria Parvez[2*], Fadi Al-Turjman[3]

[1]Management and Information System Departement AI and IoT Research Center, AI and Informatics Faculty, Near East University Mersion 10, Turkey
[2,3]AI, Data Analytics, and Software Engineering Departments, AI and IoT research center, AI and Informatics Faculty, Near East University, Mersin 10, Turkey

*Corresponding author e-mail: zubaria.parvez@neu.edu.tr

Abstract

Business Intelligence (BI) systems have become an important part of modern business as organizations increasingly generate and use large amounts of data for decision-making. However, small and medium-sized enterprises (SMEs) may face several difficulties in using BI systems effectively because of limited financial resources, technological capabilities, data management practices and employee skills. In this regard, the present study examines the role of Business Intelligence systems in supporting strategic decision-making in SMEs and discusses the major benefits and challenges related to their adoption. The study uses a conceptual research approach based on a review and synthesis of existing literature on Business Intelligence, decision support systems, data-driven decision-making and SME management. The reviewed literature shows that BI systems help organizations bring together data from different sources and present it through dashboards, reports and analytical tools. These functions can help managers understand business trends, access timely information, improve decision-making and identify potential market opportunities. Similarly, BI can support operational efficiency and improve the quality and speed of managerial decisions. However, the benefits of BI do not depend only on the technology itself. Factors such as top management support, organizational readiness, employee skills, data quality, implementation costs and technological complexity can also affect the successful use of BI in SMEs. Therefore, the study considers BI systems as an important decision-support resource rather than only a technological tool. The study also identifies areas that can be further examined through empirical research.

Keywords: *Business Intelligence, Decision Support Systems, Small and Medium-Sized Enterprises, Strategic Decision-Making, Data-Driven Decision-Making*

Fracture Detection in Pediatric Wrist Radiographs Using Attention Mechanism-Based Yolov11

Selin Döndü Korkmaz[1], Çağla Aksoy[2*], Aydın Çetin[3]

[1]Gazi University, Facultyof Technology, Computer Engineering Department, 06500, Ankara, Türkiye

[2]National Defense University, Turkish Military Academy, Computer Engineering Department, 06420, Ankara, Türkiye

[3]Gazi University, Facultyof Technology, Computer Engineering Department, 06500, Ankara, Türkiye

*Corresponding author e-mail: cagla.aksoy@msu.edu.tr

Abstract

Radiographic diagnosis of pediatric wrist fractures requires special diagnostic attention due to the fine morphology of fracture lines, high anatomical variance associated with growth plates, and the risk of being overlooked by inexperienced clinicians. In this study, automatic fracture detection was performed on the GRAZPEDWRI-DX pediatric wrist X-ray dataset using the YOLOv11m architecture with its built-in C2PSA (Cross Stage Partial with Parallel Spatial Attention) attention module. Then the model was also systematically compared against a no-attention variant of the same architecture (C2PSA replaced by C2F) to quantify the contribution of the attention mechanism to the overall system performance and to the model was comprehensively evaluated. The dataset contains 20,327 images divided into three subsets as 80% training, 10% validation and 10% test. Before being included in the training process, the images were processed through a preprocessing pipeline consisting of CLAHE contrast enhancement and Letterbox resizing steps. The model was initialized with pre-trained weights on the COCO dataset and trained with online data augmentation techniques including Mosaic, MixUp and geometric transformations for 100 epochs. According to the evaluation results on the validation set, 0.905, 0.904 and 0.949 values were obtained in the fracture class for Precision, Recall and mAP@50 respectively. The overall mAP@50 calculated over all nine classes was 0.609. The findings reveal that the attention mechanism-based YOLOv11m approach demonstrates strong and balanced performance that can support fracture detection in pediatric radiology practice.

Keywords: *Deep learning, Object detection, Fracture detection, Pediatric radiology, GRAZPEDWRI-DX*

Community Representative Selection in Scientific Collaboration Networks: Balancing Local Influence and Connectivity

Ayşenur Yurtsever[1*]

[1]Başkent University, Engineering Faculty, Industrial Engineering Department, 06790, Ankara, Türkiye

*Corresponding author e-mail: aysenur@baskent.edu.tr

Abstract

In scientific collaboration networks, researchers often form communities based on their collaboration patterns. When selecting researchers to represent these communities, choosing the most influential researcher from each community may appear to be a natural strategy. However, researchers who are highly influential within their own communities may not necessarily be well connected to researchers from other communities. This may result in a representative group that is locally influential but poorly connected as a whole. In practice, however, it may also be desirable for the selected representatives to be well connected to one another, as this can facilitate interaction and collaboration across different communities. To address this issue, this study proposes an optimization-based approach that simultaneously considers local influence and connectivity among community representatives when selecting one representative from each community. First, communities are identified using the Leiden algorithm with edge weights representing collaboration strength, and the local influence of each node is measured based on its within-community connectivity. A mathematical optimization model is then developed to select one representative node from each community while minimizing the connectivity cost among the representatives subject to a required level of total local influence. Finally, the trade-off between local influence and connectivity among the selected representatives is analyzed using the ε-constraint approach. The proposed approach is applied to a real-world scientific collaboration network. The results reveal a trade-off between local influence and connectivity beyond a certain influence level. In particular, local influence can be improved up to a certain level without increasing connectivity cost, whereas higher influence requirements lead to an increase in the cost of connecting the selected representatives. These findings demonstrate that considering connectivity alongside local influence can provide a more balanced basis for selecting community representatives.

Keywords: *Community representative selection, Community detection, Network optimization, ε-constraint method.*

Machine Learning-Based Modeling for Nut Crop Yield Prediction

Elmas Burcu Mamak Ekinci[1*], Mükerrem Bahar Başkir[2]

[1]Başkent University, Engineering Faculty, Department of Computer Engineering, 06790, Ankara, Türkiye

[2]Bartın University, Faculty of Sciences, Department of Mathematics, 74100, Bartın, Türkiye

*Corresponding author e-mail: burcuekinci@baskent.edu.tr

Abstract

Agriculture is one of the most important elements required for the sustainability of human life. Increasing agricultural yield is possible with correct planning. Determining the factors affecting yield and estimating yield with estimation models facilitates agricultural planning. Regression techniques and Machine Learning (ML) approaches are frequently used in data-based decision-making processes. There are many studies in the literature using ML algorithms and regression techniques in the estimation of agricultural data. In this study, cluster-based least absolute shrinkage and selection operator (Lasso) Regression analysis was performed as a different approach to estimate yield. Since the difference of the method was intended to be addressed in the study, the Nut Crop Yield Dataset from FAOSTAT was used. Nut crop yield per hectare was estimated using the various input variables (export/import quantity, population, nitrogen etc.). The performance metrics (RMSE, R^2, adjusted R^2, MAE, etc.) were calculated to compare cluster-based Lasso regression models of nut crop yield value data. After the clusters were identified, separate Lasso-based regression models were developed for each cluster. The performance evaluation of these cluster-specific models revealed that the best-performing model varied across clusters.

Keywords: *Sustainable Agriculture, Nut Crop Yield Prediction, Lasso Regression, ML-Unsupervised Clustering, Mathematical Model*

D-Number Quantum Representation of Fuzzy Cognitive Map Weights

Rıdvan Şahin[1*]

[1]Department of Software Engineering, Gumushane University, Gumushane, Türkiye

*Corresponding author e-mail: mat.ridone@gmail.com

Abstract

A fuzzy cognitive map (FCM) represents causal relationships among concepts using scalar edge weights. FCMs can be limited because they do not reflect the level of uncertainty in expert opinion or the degree of completeness of expert knowledge. While the D-number theory has improved multi-expert knowledge fusion, the end result of a D-number fusion is typically a single weight. In this paper, we develop a quantum-inspired representation that encodes fused D-number masses as amplitudes of causal-weight states. Unresolved information is encoded into an explicit unknown state. We then show that the measurement probability reproduces the fused D-number distribution, and that the expected value gives a classical weight equivalent to those used in conventional FCM calculations. We reconstruct a numerical example from the literature to illustrate our proposed transformation. Our model maintains the uncertainty structure and information completeness of causal relationships while maintaining the D-number fusion process, and without needing quantum hardware. This representation preserves both the usable causal effect and the uncertainty structure without needing a physical quantum computer.

Keywords: *fuzzy cognitive maps, D-number theory, quantum representation, expert knowledge, incomplete information, causal weights*

Phytochemical characterization, antioxidant and enzyme inhibitory potential of *Juniperus oxycedrus L.* leaves supported by molecular docking

Aldjia Hadroug[1*]

[1]Department of Chemistry, University pole, Bordj Bou Arreridj road, M'Sila 28000 Algeria

Corresponding author e-mail: aldjia.hadrougue@univ-msila.dz

Abstract

This study evaluated the bioactive potential of the methanolic extract of Algerian Juniperus oxycedrus L. leaves through phenolic profiling, antioxidant assays, enzyme inhibition tests, and molecular docking. HPLC-DAD analysis identified rosmarinic acid as the major compound, 54.02 mg/kg extract, followed by epicatechin, hesperidin, quercetin, coumaric acid, sinnapic acid, cinnamic acid, caffeic acid. The extract showed strong antioxidant capacity in ABTS, 171.44 mg TE/g, CUPRAC, 228.25 mg TE/g, FRAP, 178.35 mg TE/g, and DPPH, 105.60 mg TEg assays, together with phosphomolybdenum and metal chelating activities. Enzyme inhibition assays showed activity against BChE, AChE, tyrosinase, α-amylase, and α-glucosidase. Molecular docking supported the experimental findings by showing favorable binding of selected phenolic compounds within the active pockets of the target enzymes through hydrogen bonding and hydrophobic interactions. These results suggest that J. oxycedrus leaves represent a promising natural source of antioxidant and enzyme inhibitory compounds, while docking analysis provides a molecular basis for their potential biological effects.

Keywords: *Juniperus oxycedrus L.; phenolic compounds; molecular docking; enzyme inhibition; antioxidant activity*

Hybrid Content Based Image Retrieval for Similarity Search in Industrial Fastener Technical Drawings

Ertuğrul Orman[1*]

[1]Norm Fasteners, Research and Development Center, İzmir, Türkiye

*Corresponding author e-mail: ertugrul.orman@normfasteners.com

Abstract

Growing industrial drawing archives make it increasingly difficult for engineers to identify earlier designs that can support new product development. Manual searches depend heavily on personal experience, require considerable time, and may overlook geometrically related projects stored under different product names or dimensions. This study presents a hybrid content-based image retrieval approach for detecting similar products in two-dimensional fastener technical drawings. Norm Fasteners' technical drawing archive was screened. From this archive, 354 distinct projects qualifying as special products were selected, yielding 1,159 drawing images. The proposed method standardizes each drawing and evaluates rotated and mirrored representations to reduce sensitivity to orientation. Similarity is determined by combining complementary descriptions of outer-profile transitions, large-scale structural arrangement, global shape, line geometry, and local visual patterns. A two-stage project-level ranking procedure evaluates multiple views belonging to the same project and returns the most representative candidate. Dense hatching and repeated parallel lines, which may define important product features, are preserved; however, their influence is moderated when the broader geometry does not agree. The retrieved results were reviewed. Technically similar products were found among the leading candidates, while misleading matches associated with common sectional patterns were also observed. The proposed workflow gives engineers structured access to previous design knowledge, reduces the effort required for manual archive searches, and supports early assessment of whether a newly requested product resembles an existing project. The method provides a practical and interpretable basis for design-reuse and engineering decision-support systems in fastener development.

Keywords: *Content-Based Image Retrieval, Technical Drawings, Fasteners, Image Processing, Design Reuse*

Multi-Authored h-Type Indices Based on CRedit Roles

Rasim Alguliyev[1], Ramiz Aliguliyev[2], Narmin Adigozalova[3*]

[1] Institute of Information Technology, Az1141, Baku, Azerbaijan

*Corresponding author e-mail: narmin66@gmail.com

Abstract

In the modern scientific landscape, collaborative and multidisciplinary research has expanded rapidly. Consequently, ensuring transparency in authorship responsibility and equitably distributing citation credit in multi-authored studies have become crucial. Traditional bibliometric systems assign citations equally to all co-authors, hindering the objective evaluation of individual research contributions. This study presents an in-depth analysis of conventional scientometric metrics, specifically the h-, g-, and A-indices, within the context of multi-authored publications. Our analysis reveals a systemic misallocation of citation credit and establishes a theoretical framework demonstrating how equal attribution leads to the artificial inflation of scientific productivity indicators. To resolve these deficiencies, we propose weighted variants of the h-, g-, and A-indices integrated with the Contributor Roles Taxonomy (CRediT) framework. By incorporating individual performance coefficients based on actual author contributions, these proposed indices offer a more transparent, accurate, and equitable approach to scientometric evaluation.

Keywords: *CRediT system, h-index, g-index, A-index, Weighted indices scientometrics author contribution*

Comparative Performance of Robust Regression and Machine Learning Methods under Vertical Contamination

Zehra Demirtas[1*], Sibel Açık Kemaloğlu[1]

[1]Ankara University, Faculty of Science, Department of Statistics, 06100, Ankara, Türkiye

*Corresponding author e-mail: zdemirtas@ankara.edu.tr

Abstract

Outlying observations can substantially affect the predictive performance of regression methods. This study aims to compare the performance of ordinary least squares (OLS), the MM- and S-estimators, support vector regression (SVR), and gradient boosting regression (GBR) in the presence of vertical outliers (Y-direction outliers). In the first stage of the study, a Monte Carlo simulation with 1000 replications was conducted for sample sizes n = 100 and n = 500. The effects of vertical outliers were examined at contamination rates of 0%, 5%, 10%, and 15% and an outlier magnitude of 10. In the second stage, the methods were evaluated on the Combined Cycle Power Plant dataset. Candidate naturally occurring vertical outliers in the training data were identified using scaled robust residuals obtained from MM regression and were retained throughout the analysis. Subsequently, positive vertical contamination was introduced into the non-outlying training observations at rates of 5%, 10% and 15%, with magnitudes set to 10 times the robust scale estimate. Model performance was evaluated using RMSE, MAE, and R^2. The simulation results showed that the MM- and S-estimators were the most stable methods under contamination, while SVR achieved performance close to that of these estimators. In contrast, the performance of OLS and GBR deteriorated markedly as the contamination rate increased. In the real data application, GBR was the best performing method in the absence of additional contamination, whereas SVR achieved the lowest RMSE and MAE and the highest R^2 across all artificially contaminated scenarios. The MM- and S-estimators largely maintained their performance under different contamination conditions. These findings indicate that method selection should consider not only predictive accuracy under clean data conditions but also performance stability in the presence of outlying observations.

Keywords: *Vertical Outlier, Robust Regression, MM Estimator, Support Vector Regression, Gradient Boosting Regression.*

Exploring Fuzzy Clustering Objective Functions through Metaheuristic Optimization for Image Segmentation

Özge Tezel[1*], Onur Okuyucu[1,2]

[1]Karadeniz Technical University, Faculty of Science, Department of Computer Sciences, 61080, Trabzon, Türkiye

[2]Karadeniz Technical University, Graduate School of Natural and Applied Science, 61080, Trabzon, Türkiye

*Corresponding author e-mail: ozgetezel@ktu.edu.tr

Abstract

Image segmentation is a fundamental process that involves identifying meaningful regions within an image and has a direct effect on the performance of image processing applications. Fuzzy clustering methods are widely used for image segmentation because they enable pixels to be assigned to different clusters with varying degrees of membership. However, these methods may face several challenges, such as sensitivity to the initial conditions, becoming trapped in local optima, and difficulty in identifying optimal cluster centers in complex image structures. This study aims to systematically assess the performance of different fuzzy clustering models optimized by swarm intelligence algorithms in image segmentation. In this context, different variants of the fuzzy c-means (FCM) clustering algorithm proposed in literature such as the outlier rejection fuzzy c-means (ORFCM), k-means integrated fuzzy c-means (KIFCM), intuitionistic fuzzy possibilistic c-means (IFCM) are considered. Unlike traditional fuzzy clustering approaches, the objective functions of these models are optimized using metaheuristic search mechanisms, with the aim of obtaining optimal cluster centers. Within the experimental framework, the segmentation performance of the fuzzy clustering models optimized using metaheuristic methods is comparatively analyzed using test images with different characteristics. To objectively assess the segmentation results, various image quality and cluster validity metrics are implemented. Thus, this study presents a comprehensive evaluation of the impact of objective functions on image segmentation performance by comparing different fuzzy clustering models under the same metaheuristic algorithm.

Keywords: *Image Segmentation, Fuzzy Clustering, Fuzzy C-Means, Swarm Intelligence, Metaheuristics*

Evaluation of Advanced Transfer Functions in Binary Crested Porcupine Optimizer for Gene Selection-Oriented Cancer Classification

Eda Özkul[1*], Muharrem Korkmaz[1,2]

[1]Karadeniz Technical University, Faculty of Science, Department of Computer Sciences, 61080, Trabzon, Türkiye

[2]Karadeniz Technical University, Graduate School of Natural and Applied Science, 61080, Trabzon, Türkiye

*Corresponding author e-mail: edaozkul@ktu.edu.tr

Abstract

High-dimensional gene expression data contains valuable information for the classification of cancer types. However, the presence of numerous irrelevant and redundant genes in such data can negatively affect classification performance and computational efficiency. Therefore, the identification of discriminative genes is a critical preprocessing step in cancer classification tasks. Due to their ability to efficiently explore large search spaces, metaheuristic optimization algorithms have recently become popular choices for high-dimensional gene selection problems. However, algorithms developed for continuous search spaces require an effective binarization mechanism to be applied to binary problems such as gene selection. Therefore, the aim of this study is to design a binary crested porcupine optimizer (BCPO) approach for addressing gene selection and cancer classification problems. In this context, dynamic, adaptive, and time-varying transfer functions, in addition to classical transfer functions, are comparatively analyzed within a common BCPO framework. By keeping the CPO's core search mechanism and modifying only the transformation mechanism that maps continuous solutions to the binary decision space, the impact of different transfer functions on the algorithm's search behavior will be evaluated independently. Comparisons are conducted based on various criteria, including solution quality, convergence behavior, stability, the number of selected features, accuracy, recall, and the exploration-exploitation balance. The findings are expected to contribute to the development of a more effective binary CPO structure for cancer classification in high-dimensional gene expression data and, subsequently, to the establishment of a robust gene selection framework.

Keywords: *Gene Selection, Cancer Classification, Crested Porcupine Optimizer, Binary Optimization, Transfer Functions*

Regional Modelling of Greenhouse Gas Emissions: A Cluster and Regression Approach

Özge Ustahüseyin[1]*, Esra Nur Eviren[1], Zehra Demirtaş[1]

[1] Ankara University, Faculty of Science, Department of Statistics, Ankara, Türkiye

*Corresponding author e-mail: oustahuseyin@ankara.edu.tr

Abstract

Climate change is a global issue closely linked to economic activities that generate greenhouse gas emissions, primarily in the energy, industrial, transportation, and building sectors. The accumulation of these emissions in the atmosphere is the result of such activities. The objective of this study is to cluster 28 European countries based on their greenhouse gas emission levels for the 2005–2024 period and to model Germany's greenhouse gas emissions which are the highest in the high-emission cluster using macroeconomic indicators. The clustering analysis, conducted with the objective of accounting for heterogeneity among countries, resulted in the formation of groups of countries with similar greenhouse gas emission profiles. The modelling of Germany's greenhouse gas emissions was conducted through the utilization of regression analysis, a statistical method employed for the estimation of quantitative relationships between dependent variables and a set of independent variables. The independent variables employed in this analysis comprised macroeconomic indicators, which were representative of the industrial and infrastructural sectors. In order to examine the regional validity of the final model obtained after the variable selection, model validation, and performance evaluation stages, the model was also applied to Germany's neighboring countries. The study therefore examined how the model's predictive performance varied across countries and to what extent the model developed for Germany could be generalized to the structures of neighboring countries. Furthermore, greenhouse gas emission projections for the 2025–2026 period were made using the developed model. The findings indicate that indicators representing industry and infrastructure have considerable potential for explaining greenhouse gas emissions. However, it is also evident that model performance varies from country to country. A comparison of the model's performance in Germany with that in neighboring countries provides insights into the model's generalizability at the regional level. The present study contributes to the analysis of the relationship between greenhouse gas emissions and macroeconomic indicators in Europe by combining a clustering approach that accounts for country heterogeneity with regression modelling and a regional generalizability process.

Keywords: *Greenhouse Gas Emissions, Cluster Analysis, Regression Analysis*

Deep Learning for Pet Breed Classification: Flat and Hierarchical Hybrid Models

Selen Aytaş[1*], Erkut Arıcan[2]

[1] Bahçeşehir University, Department of Computer Engineering, Istanbul, Türkiye
[2] Bandırma Onyedi Eylül University, Department of Computer Engineering, Balıkesir, Türkiye

*Corresponding author e-mail: selen.aytas@bau.edu.tr

Abstract

The study was conducted on the problem of determining the breed of an animal from two classes (cat and dog) and 37 breed categories. In this study, flat and hierarchical classification approaches are analyzed. In the flat approach, breed classification is performed directly by the model, while in the hierarchical approach, the animal is first classified as either a cat or a dog, and then the relevant breed classification is performed. For this purpose, the Oxford-IIIT Pet Dataset, a publicly accessible labeled dataset that includes 37 diverse cat and dog breeds, is used. The aim of the study is to improve the performance of existing approaches that have been previously studied with traditional methods by re-examining them using modern deep learning and feature extraction approaches and to comparatively analyze the differences between methods. There can be very small differences, especially between different cat breeds; therefore, solving this problem is highly challenging. In this context, deep feature extraction was performed using ResNet18, and classification was carried out using Extreme Gradient Boosting (XGBoost) and Support Vector Machines (SVM). Using the end-to-end ResNet18 model, the results were compared with those of the hybrid models. In addition to this, augmented and non-augmented training settings were used. The experiments show modern deep features are more efficient than traditional methods. The best result was achieved by the ResNet18 + SVM framework. The ResNet18 + SVM framework reached 89.91% accuracy and 89.84% average per-class accuracy.

Keywords: *Deep Feature Extraction, ResNet18, Support Vector Machine, XGBoost, Pet Breed Classification*

Examining the Relationships Among Age, Education Level and Gender of Drivers Involved in Traffic Accidents Using Log-Linear Models

Esra Nur Eviren[1*], Özge Ustahüseyin[1], Mehmet Yılmaz[1], Cemal Atakan[1]

[1]Ankara University, Faculty of Science, Department of Statistics, Ankara, Türkiye

*Corresponding author e-mail: eneviren@ankara.edu.tr

Abstract

Traffic accidents are a significant societal problem at both global and national levels because of the fatalities, injuries, and material damage they cause. In addition to factors such as the vehicle, the road, and the environment, the human/driver factor is known to account for the largest share in the occurrence of accidents. Since driving behavior may differ according to socio-demographic characteristics such as age, education level, and gender, it is important to evaluate the relationships among these characteristics. In this study, the aim was to examine the association structure among the age, education level, and gender variables of drivers involved in traffic accidents that occurred in 2024 on the highways under the responsibility of the General Command of Gendarmerie, by means of log-linear models. The dataset of the study consists of raw data obtained from the General Command of Gendarmerie Traffic Department. Age, education level, and gender were categorized, and a 3×3×2 contingency table was obtained. Considering approaches to multidimensional categorical data analysis, different log-linear models were fitted and compared. Thus, it was aimed to identify the interactions among these variables and to determine the model structure that best explains the data. It is thought that the findings to be obtained may contribute to the identification of risk profiles for traffic safety.

Keywords: *Traffic Accidents, Log-Linear Models, Categorical Data Analysis, Discrete Multivariate Statistics*

Solar Flare Classification Using Logistic Regression and Its Variants

Ezgi Eviren[1*], Esra Nur Eviren[2], Cemal Atakan[2]

[1]Gazi University, Gazi Faculty of Education, Department of Mathematics and Science Education, Division of Physics Education, Ankara, Türkiye

[2]Ankara University, Faculty of Science, Department of Statistics, Ankara, Türkiye

*Corresponding author e-mail: evirenezgi@gmail.com

Abstract

Solar flares are events resulting from the sudden release of magnetic energy accumulated in active regions, and they pose a direct threat to communication and power infrastructure on Earth. Predicting whether an active region will produce a flare within a given time horizon is therefore one of the important problems in space research. In this study, 29 magnetic features derived from the reduced-resolution active region magnetograms (maps showing the strength, direction, and variation of the magnetic field) published on Dryad by Boucheron et al. (2023) were used. Using these features, the binary classification problem of whether an active region will produce a flare of class C1.0 or above within the following 24 hours was addressed comparatively through classical logistic regression and its Lasso, Ridge, and Elastic Net penalized variants. The dataset contains 950,047 observations from 1,570 unique active regions (ARs) covering the period from 1 May 2010 to 31 December 2018. The AR-based partitions distributed with the dataset were used for the training, validation, and test sets, thereby preventing observations from the same active region from being split across different sets. Model performance was evaluated using metrics such as sensitivity, specificity, F1, the area under the ROC curve (AUC), the True Skill Statistic (TSS), and the Heidke Skill Score (HSS). In addition, to address class imbalance, class-based weights were assigned to the observations; this weighting was observed to systematically shift the predicted flare probabilities upward, and the resulting biases were corrected by an intercept adjustment. Consequently, the performance of different logistic regression approaches to solar flare prediction was examined comparatively, and no clear separation was observed among the classification performances of the models.

Keywords: *Solar Flare Prediction, Logistic Regression, Lasso, Ridge, Elastic Net*

A Literature-Based Comparative Review of Popular Symmetric Encryption Algorithms

Emine Elvin Bacı[1*], Sara Kandulu[2], İbrahim Erşan[3]

[1]Girne American University, Faculty of Engineering, Department of Computer Engineering, 99428, Kyrenia, TRNC

[2]Girne American University, Faculty of Engineering, Department of Computer Engineering, 99428, Kyrenia, TRNC

[3]Girne American University, Faculty of Engineering, Department of Software Engineering, 99428, Kyrenia, TRNC

*Corresponding author e-mail: elvinbaci@gau.edu.tr

Abstract

Symmetric encryption remains essential for protecting the confidentiality of data in modern digital systems. However, symmetric algorithms differ considerably in their design parameters, documented security limitations, standardization status, implementation requirements, and practical suitability. This paper presents a literature-based comparative review of ten influential symmetric encryption algorithms: DES, 3DES, Blowfish, IDEA, RC4, RC5, RC6, Twofish, AES, and ChaCha20. The algorithms are examined according to cipher type, key length, block or processing-unit size, number of rounds, known security limitations, current standardization status, computational complexity, performance considerations, and historical or contemporary applications. The comparison indicates that DES and RC4 are no longer suitable for modern cryptographic protection, while 3DES is restricted to limited legacy processing. Blowfish, IDEA, RC5, RC6, and Twofish remain historically or academically significant but have limited contemporary adoption. AES continues to serve as the principal NIST-standardized block cipher, whereas ChaCha20, particularly when combined with Poly1305, provides an important software-oriented alternative on platforms without dedicated AES acceleration. The reviewed algorithms generally exhibit linear time complexity with respect to input length under fixed parameters; nevertheless, practical performance depends on implementation, hardware support, operating mode, message size, and key-management conditions. Since this study does not include controlled experimental benchmarks, it does not establish a universal performance ranking.

Keywords: *Symmetric encryption, block cipher, stream cipher, AES, ChaCha20.*

Reliable AI Starts with Reliable Data: Quality, Deduplication and Linguistic Heterogeneity in Large-Scale Social Media Analytics

Murat Göksu[1*]

[1]Istanbul Esenyurt University, Faculty of Applied Science, Department of Information Systems and Technologies, 34513, Istanbul, Türkiye

*Corresponding author e-mail: muratgoksu@esenyurt.edu.tr

Abstract

The increasing use of large-scale social media data in artificial intelligence and misinformation analytics has raised important concerns regarding data quality, redundancy, and representativeness. Although very large datasets are often assumed to improve analytical reliability, data volume alone does not guarantee trustworthy artificial intelligence. This study examines data reliability issues in a large-scale COVID-19 Twitter dataset containing approximately 89.47 million records collected during the early phase of the pandemic. The analysis focuses on three fundamental dimensions of data quality: preprocessing-related data reduction, duplicate content, and linguistic heterogeneity. Following data cleaning and filtering, approximately 85.71 million records were retained, while deduplication reduced the dataset to approximately 28.73 million unique records. The dataset also demonstrated substantial multilingual diversity, with English representing approximately 63.5% of the sampled content, followed by Spanish, French, Indonesian, Portuguese, and other languages. These observations indicate that large-scale social media datasets may contain significant redundancy and distributional complexity that can influence downstream AI and misinformation-analysis pipelines. The findings emphasize that reliable AI requires systematic attention to data quality before model development and evaluation. In particular, deduplication, language-aware preprocessing, and transparent data preparation should be treated as essential components of trustworthy social media analytics. The study highlights reliable data engineering as a foundational requirement for robust and sustainable AI applications.

Keywords: *Reliable AI, Data Quality, Social Media Analytics, Deduplication, Linguistic Heterogeneity.*

An Empirical Study on Anisotropy of Word2Vec Embedding Spaces

Hüseyin Sezer Solak[1*], Kâmil Dimililer[2]

[1]Girne American University, Faculty of Engineering, Computer Engineering Department, 99428, Kyrenia, TRNC
[2]Girne American University, Faculty of Engineering, Electrical-Electronics Engineering Department, 99428, Kyrenia, TRNC

*Corresponding author e-mail: sezersolak@gau.edu.tr

Abstract

Embedding spaces are fundamental to modern language models, which represent raw data as high-dimensional vectors to encode semantic relationships. It is known that anisotropy increases with the ability to grasp contexts, which can decrease system performance. We examined how the vector dimension and contextual depth influence the anisotropy of the embedding space generated by the Word2Vec skip-gram model. We used three metrics to analyse different aspects of the anisotropy of the spaces: Average cosine similarity between embedding vectors, singular value entropy, and a PCA-based anisotropy measure. We discussed the anisotropy measures obtained for two different datasets and compared them with the literature.

Keywords: *NLP, Embedding spaces, Isotropy, Word2Vec*

Modified K-means Algorithm for Big Data

Kiyana Nazar*[1], Ali Haydar[1]

[1]Girne American University, Faculty of Engineering, Computer Engineering Department, 99428, Kyrenia,

*Corresponding author e-mail: kiyana.nazar@std.gau.edu.tr

Abstract

The K-Nearest Neighbour (KNN) algorithm is a well-known and widely used machine learning approach for classification. Even though it is a very simple algorithm, its high computational requirements make its use infeasible for Big Data analysis. To overcome this problem, alternatives in distributed environments were used effectively. In this study, the K-means clustering algorithm is modified to group unlabeled data points into overlapping rather than non-overlapping clusters. This modified K-means algorithm is applied to a binary tree. Consequently, classification accuracy was evaluated using the KNN algorithm on a small subset of the original data. The results showed that excellent classification accuracy could be achieved while significantly reducing computation time. In addition, this method can be used for distributed environments and KNN-based classification algorithms, which are proposed to increase classification accuracy.

Keywords: *KNN, K-Means, Big Data, Binary Tree*

Universal Basic Income in the Age of AI: The Transformation of the Labour Market and the Future of Social Policy

Aytac Demiray[21]

[1]Selcuk University, Aksehir Vocational School, Department of Accounting and Tax Applications, 42550, Konya, Türkiye

*Corresponding author e-mail: demiray@selcuk.edu.tr

Abstract

Rapid advances in artificial intelligence (AI) and automation technologies are transforming production processes and the structure of labour markets, while also renewing debates about the future of social policy. Although technological advances are often argued to adversely affect employment and wages by substituting for labour, the empirical literature suggests that these effects are not unidirectional. While some studies associate automation with employment and wage losses, others find no significant decline in overall employment or point to employment growth in occupations with greater exposure to AI. Moreover, technological transformation may produce heterogeneous outcomes in terms of productivity, task and skill composition, wage inequality, and job quality.

This study aims to assess the emerging risks that social protection systems may face as a result of the transformation of labour markets driven by AI and automation, and to examine the potential role of universal basic income (UBI) in addressing these risks. By bringing together empirical findings within a framework linking labour markets, social protection, and UBI, the study seeks to contribute to the existing literature. To this end, it adopts a literature-based comparative policy analysis, reviewing empirical research on the labour-market effects of AI and automation as well as studies examining the relationships among automation risk, social protection preferences, and UBI.

The existing evidence suggests that automation risk may strengthen demand for redistribution and social protection in some societies, but that this does not necessarily translate into greater support for UBI. Furthermore, both the design of UBI and the characteristics of existing social protection systems appear to shape policy preferences, while experimental evidence suggests that UBI does not necessarily reduce work effort. Against this background, the study approaches UBI not as an inevitable consequence of AI-induced technological unemployment, but as a policy option that should be considered alongside alternative and complementary social policy instruments in addressing changing labour-market risks.

Keywords: *Universal Basic Income, Artificial Intelligence, Automation, Labour Market, Social Policy*

The Algorithmic Parent: Gender Stereotypes and Family Bias in AI-Generated Visual Content

Samime Çağlar[1*], Selda Sekban[1], Eser Selen[1]

[1]Özyeğin University, Turkiye

*Corresponding author e-mail: samime.caglar@ozu.edu.tr

Abstract

This study examines the algorithmic biases of text-to-image generative AI platforms regarding "parenting" and "gender roles" within the family. The research conceptualizes these platforms as posthumanist agents and material-discursive apparatuses that construct meaning through agential intra-actions with the user. In this study, standardized Turkish prompts focusing on "parent," "mother," and "father" were submitted over a period of four weeks using four premium accounts on the ChatGPT and Gemini platforms across two locations in Türkiye (Istanbul and Edirne). The resulting dataset of 409 images obtained from 16 chat sessions, alongside the AI\'s self-generated demographic text descriptions of these images, was analyzed using a three-stage approach: (1) Visual Content Analysis, which quantifies the produced visual elements and codes recurring iconographic patterns such as skin tone, age, setting, activity performed, and spatial positioning of parental figures. (2) Netnographic Observation, which examines AI interfaces and prompt response dynamics using Constructivist Grounded Theory's memoing technique and Situational Analysis to map the sociotechnical conditions shaping image generation. (3) Descriptive Thematic Analysis, which inductively integrates all quantitative and qualitative data to identify overarching representational patterns across platforms, prompts, and locations. Findings of this study reveal how generative AI actively reinforces traditional gender roles, ethno-racial profiling, wealth bias, and performative cultural representations during its agential intra-actions with the user. In addition to offering critical implications for future practices in ethical AI governance and interface design, these results establish a posthumanist auditing framework for further research into algorithmic bias within non-Western and gender-neutral linguistic contexts.

Keywords: *Generative AI, Algorithmic Gender Bias, Parental Representation, Text-to-Image Models, Family Structure, "Posthumanism*

The Effects of Trust and Design Innovativeness on Students' Intention to Use ChatGPT

Kamil Çelik[1*], Mustafa Emin Gül[1], Ahmet Zelka[1]

[1]Bartın University, Turkiye

*Corresponding author e-mail: kamilcelik86@gmail.com

Abstract

This study aims to examine the effects of trust and design innovativeness on students' intention to use ChatGPT. The research was conducted using data obtained from 392 university students studying in Türkiye. Structural Equation Modeling (SEM) was employed, and the analyses were conducted using the open-source R programming language. The research model examined the effects of trust and design innovativeness on students' behavioral intention to use ChatGPT. The results indicated that trust has a positive and significant effect on behavioral intention ($\beta = 0.421$; $p < 0.001$). Similarly, design innovativeness was found to have a positive and significant effect on behavioral intention ($\beta = 0.500$; $p < 0.001$). Based on the standardized coefficients, the effect of design innovativeness on behavioral intention was found to be stronger than that of trust. In addition, the model yielded an R^2 value of 0.594, indicating that trust and design innovativeness jointly explained 59.4% of the variance in students' intention to use ChatGPT. The findings demonstrate that trust and design innovativeness play important roles in explaining students' intention to use ChatGPT. In particular, the stronger predictive effect of design innovativeness highlights the importance of innovative and user-oriented design features in students' adoption of generative artificial intelligence tools such as ChatGPT. The study is expected to contribute to research aimed at understanding the adoption of generative artificial intelligence technologies in educational settings.

Keywords: *artificial intelligence, ChatGPT, trust, design innovation*

Cross-Study Generalization of Machine-Learning-Based Dry Sliding Wear Prediction: Evidence from Blind Validation

Azmi Erdoğan[1*], İbrahim Çalış[2]

[1]Department of Metallurgy and Materials Engineering, Faculty of Engineering, Architecture and Design, Bartın University, 74100, Bartın, Türkiye

[2]Central Research Laboratory, Bartın University, 74100, Bartın, Türkiye

*Corresponding author e-mail: aerdogan@bartin.edu.tr

Abstract

Machine-learning approaches have increasingly been used to predict tribological performance; however, their ability to generalize across different material systems and surface-engineering processes remains uncertain. In this study, a heterogeneous dry sliding wear dataset was constructed from four independent published studies involving metallic materials subjected to different surface modification routes. Experimental descriptors, including normal load, test temperature, sliding distance, surface hardness, modified-layer thickness, and surface/process condition, were extracted from reported tables and digitized graphs. Regression-based machine-learning models were developed using grouped validation strategies designed to prevent data leakage between measurements belonging to the same material condition. Two additional independent studies were then kept entirely outside the model development stage and used for blind prediction after the model outputs were locked. The first blind test, performed on aluminized INC738, showed that the model reproduced the overall wear-resistance trend and correctly identified the relative advantage of the thicker aluminide coatings, but substantially overestimated the absolute wear-volume values. In the second blind test, involving borided CrFeMnNbNi high-entropy alloy, the model correctly predicted that boriding would improve wear resistance but markedly underestimated the magnitude of the experimentally observed improvement, which reached approximately 94–96%. These results demonstrate that models trained on heterogeneous tribological datasets may capture qualitative wear trends but fail to provide transferable quantitative predictions when an unseen material introduces a different microstructural or surface-response regime. The findings emphasize the importance of external validation, physically meaningful material descriptors, and domain-shift awareness in machine-learning-assisted tribology.

Keywords: *Machine Learning, Triboinformatics, Dry Sliding Wear, Surface Engineering, Blind Validation*

DP-CopulaGAN: Privacy-Preserving Tabular Synthesis through Copula-Space Adversarial Learning

Emre Çırak[1*], M. A. Bakır[2]

[1]Graduate School of Natural and Applied Sciences, Department of Data Science, Gazi University, Ankara, Türkiye
[2]Faculty of Science, Department of Statistics, Gazi University, Ankara, Türkiye

*Corresponding author e-mail: 25830311010@gazi.edu.tr

Abstract

The generation of useful synthetic tabular data under formal differential privacy remains difficult because heterogeneous marginals and nonlinear dependencies interact poorly with gradient clipping and noise injection. This study presents DP-CopulaGAN, a privacy-preserving generative framework that relocates adversarial learning from the original feature space to a copula-based latent representation. Continuous numerical variables are transformed through empirical marginal distributions and Gaussian quantile mapping, yielding approximately standardized latent variables while retaining cross-variable dependence. A spectrally normalized Wasserstein GAN with gradient penalty is trained in this transformed space using DP-SGD, after which generated latent samples are mapped back to the original scale through inverse empirical transformations.
We evaluate the method on EEG confusion detection, Adult income, and Credit Card Fraud data at privacy budgets epsilon = 0.01, 0.1, and 1 with delta = 10^-5. Performance is assessed from complementary perspectives: marginal similarity, Wasserstein distance, correlation RMSE, nearest-neighbor separation, and train-on-synthetic-test-on-real AUROC. Across these settings, DP-CopulaGAN provides more stable dependence preservation and downstream predictive utility than DP-WGAN, DP-CGAN, and DP-CTGAN, with the largest gains under tighter privacy constraints. At epsilon <= 0.1, its correlation recovery is also competitive with or stronger than the evaluated non-private CTGAN and TVAE baselines.
The findings suggest that a copula transformation can serve as an effective structural preconditioning step for differentially private adversarial training. By separating marginal shape from dependence learning, DP-CopulaGAN offers a practical route to higher-utility private synthetic tabular data for domains in which direct access to sensitive records is restricted.

Keywords: *Differential privacy, Copula models, Generative adversarial networks, Synthetic data, Tabular data*

Improved Nonparametric Estimation of Stress-Strength Reliability via Multiplicative Bias Correction

Elif Ercelik[1*]

[1]Doğuş University, Faculty of Engineering and Natural Sciences, Department of Software Engineering, 34775, Istanbul, Turkiye

*Corresponding author e-mail: eercelik@dogus.edu.tr

Abstract

The stress-strength reliability parameter $R = P(X > Y)$ plays a fundamental role in engineering and survival analysis, quantifying the probability that a system's strength X exceeds the applied stress Y. While parametric estimators of R are well established, they are sensitive to distributional misspecification. Nonparametric alternatives based on asymmetric kernels avoid this limitation but are constrained by the $O(n^{-\frac{4}{5}})$ mean squared error (MSE) convergence rate of classical kernel density estimators.

This study proposes two multiplicative bias correction (MBC) estimators of R based on the Gamma kernel. The TS estimator combines Gamma kernel estimates at bandwidths b and b/c, while the JLN estimator applies a self-normalization correction. Both estimators achieve an improved MSE convergence rate of $O(n^{-\frac{8}{9}})$, which is substantially faster than the classical $O(n^{-\frac{4}{5}})$ rate. Closed-form expressions for the asymptotic bias, variance and MSE are derived. An extensive Monte Carlo simulation study is conducted under various distributional scenarios and sample sizes. The finite-sample performance of the proposed estimators is evaluated and compared with the classical Gamma kernel estimator. Real data applications are also provided to further demonstrate the practical utility of the proposed methods.

Keywords: *Stress-strength reliability, Gamma kernel estimator, Multiplicative bias correction, Nonparametric density estimation.*

Artificial Intelligence, Digitized Archives, and the Transformation of Historical Research

Ethem Ceku[1*]

[1]Faculty of Education University of Prishtina, Republic of Kosovo

*Corresponding author e-mail: ethem.ceku@uni-pr.edu

Abstract

This study argues that the principal significance of artificial intelligence for historical research lies in the democratization of access to sources, the acceleration of research processes, and the expansion of opportunities for comparative analysis, while the scholarly reliability of research outcomes continues to depend on critical source verification and the historian's methodological competence. In the past, systematic access to international archives required substantial financial resources, extended research stays, and sustained institutional support. Consequently, access to and use of archival documentation were often constrained by researchers' financial and institutional capacities.

The digitization of archival collections and the increasing use of artificial intelligence are transforming this model by broadening the documentary basis of historical research and enabling the consultation, comparison, and cross-referencing of materials dispersed across different institutions. AI also facilitates access to monographs and scholarly articles through open-access and subscription-based platforms. It assists with bibliographic research, document translation, reference management, data extraction and processing, as well as the critical comparison of primary sources with relevant scholarly literature.

However, these opportunities are accompanied by significant methodological risks. Poorly formulated queries may generate inaccurate references, misattributed authorship, unverified data, or unreliable interpretations. AI should therefore be understood as a research tool that enhances the historian's analytical and methodological capacity rather than as a substitute for professional judgment. Its scholarly value remains dependent on rigorous bibliographic control, systematic source verification, and the comparison of AI-generated information with original documents and relevant scholarly literature.

Keywords: *Artificial intelligence; historical research; digitized archives; historical methodology; source verification.*

Revisiting Data Privacy in Machine Learning Systems Across the Learning Lifecycle

Ulduz Mammadova[1*]

[1]Alanya University, Faculty of Engineering and Natural Sciences, Computer Engineering Department, Antalya/Türkiye

*Corresponding author e-mail:
contact@ulduzmammadova.com

Abstract

The current state of the world makes almost everything a source for data. People, knowingly or not, and with or without understanding, allow the collection of vast amounts of data that sometimes include private information about themselves or the people around them. At the same time, enterprises face the challenge of keeping up with rapidly developing technologies while preserving confidential and sensitive information. Preserving data privacy, once considered a goal to be achieved while collecting and presenting data, now requires a more systematic approach. It is no longer just a dataset problem: privacy risks follow every stage of machine learning, from data collection and storage to model training and deployment.

This work reviews commonly used privacy-preserving methods developed and applied over the past decade to address these challenges, including techniques such as anonymization, differential privacy, and federated learning, among others. It looks at their strengths, limitations, and practical use, and considers how privacy risks can persist even when data is anonymized or decentralized, as well as the new concerns raised by emerging machine learning and generative AI systems. Together, these approaches show how data privacy in machine learning has evolved, and underscore that even commonly used methods sometimes cannot fully guarantee privacy once data is shared or combined with other sources.

Keywords: *Data Privacy, Data Protection, Machine Learning, Differential Privacy, Federated Learning*

Polycystic Ovary Syndrome (PCOS) Prediction Using Random Forest Algorithm

Ebru Altıparmak Yangal[1*]

[1]Erzurum Technical University, Faculty of Science, Department of Mathematics, 25000, Erzurum, Türkiye

*Corresponding author e-mail: ebru.altiparmak@erzurum.edu.tr

Abstract

Polycystic ovary syndrome (PCOS) is a common endocrine disorder, and machine learning methods can support its early diagnosis. This study investigates the effect of the Synthetic Minority Over-sampling Technique (SMOTE) on Random Forest-based PCOS prediction using a publicly available Kaggle dataset containing 541 patient records. The dataset was divided into 80% training and 20% independent test sets using stratified sampling. Random Forest algorithm have developed with and without SMOTE, and their hyperparameters have optimized using GridSearchCV with 5-fold cross-validation. The models have evaluated on the independent test set using accuracy, precision, recall, F1-score, and ROC-AUC. The results showed that SMOTE increased accuracy from 91.74% to 92.66%, precision from 93.55% to 96.67%, and F1-score from 86.57% to 87.88%. Recall remained unchanged at 80.56%, while ROC-AUC increased slightly from 0.9387 to 0.9429. These findings indicate that, in this experiment, SMOTE provided modest improvements in the overall performance of the Random Forest model, particularly in precision and F1-score, while maintaining the same recall. Overall, the results demonstrate that class balancing with SMOTE can influence Random Forest-based PCOS prediction, although the magnitude of improvement varies across evaluation metrics.

Keywords: *Polycystic ovary syndrome, Random Forest Algorithm, SMOTE, GridSearchCV, 5-fold cross-validation.*

Comparsion of Non-Linear and Bayesian Approaches in the Wood Lactation Curve Model

Aslı Ates[1*], G. Tamer Kayaalp[1]

[1]Çukurova University, Faculty of Agriculture, Department of Animal Sciencs, Adana, Türkiye

*Corresponding author e-mail: asliates91@gmail.com

Abstract

The aim of this study is to apply non-linear and Bayesian approaches to the classical non-linear Wood lactation curve model and compare them. In the study, a cross-sectional dateset from the relevant dateset containing the milk yield of large cattle in the first 6 months of 2025 at the Çukurova Universty Animal Farm was used. For comparsion purposes, the non-linear model and Bayesian approach were applied to the same dateset. RMSE, AIC and BIC values were used as comparsion criteria. In the conducted model comparsion analysis, the Bayesian approach showed higher performance in predicting the dependent variable with an RMSE value of 4,09 compared to the non-linear model (5,73). However, when the AIC and BIC values, which penalize model complexity and the number of parameters, are examined, the non-linear model (788,75/779) appears to be more advantageous then the Bayesian model (1266,329/1289,8909) and is seen to have a simpler structure. This situation indicates that although the Bayesian model seems disadvatageous in standard information criteria due to its effective parameter structure, it can yield more consistent results in prediction accuracy.

Keywords: *Wood lactation model, Bayesian approach, Non-linear model*

Data-Driven Asymmetric Fuzzy Chain-Ladder for IBNR Estimation

Rümeysa Karatas[1]

[1] Karamanoğlu Mehmetbey University, Department of Data Science and Analytics, 70200, Karaman, Türkiye

*Corresponding author e-mail: rumeysakaratas@kmu.edu.tr

Abstract

Estimating Incurred But Not Reported (IBNR) claims is a central problem in non-life insurance reserving, where future claim development is often subject to substantial and potentially asymmetric uncertainty. Traditional chain-ladder methods primarily provide point estimates, while existing fuzzy approaches frequently rely on predetermined and symmetric membership structures. This study proposes a data-driven asymmetric fuzzy chain-ladder framework for IBNR reserve estimation. In the proposed approach, classical development factors are first estimated from historical claim triangles. Bootstrap resampling is then used only as a data-driven calibration mechanism for determining the central tendency and asymmetric lower and upper spreads of fuzzy development factors, rather than as a direct substitute for fuzzy uncertainty. These calibrated fuzzy factors are propagated through the chain-ladder structure using fuzzy arithmetic to obtain accident-year-specific and aggregate IBNR reserve estimates. Different alpha-cut levels are considered to evaluate the sensitivity of the resulting reserve ranges. The proposed model is compared with the classical chain-ladder, Mack chain-ladder, and a symmetric fuzzy chain-ladder approach. In addition, pseudo-out-of-sample backtesting is employed to assess estimation error, empirical coverage, and interval width. The framework is intended to provide a more flexible and interpretable representation of asymmetric reserving uncertainty while preserving the practical structure of conventional chain-ladder estimation.

Keywords: *Claims reserving; Asymmetric fuzzy numbers; Chain-ladder; Uncertainty modeling; Actuarial science.*

A Decision Support System for Regional Depot Location and Routing of Chain Stores Considering Road-Network Accessibility

Eyüp Cakı[1*], Seyit Akpancar[2], Durmuş Özdemir[3]

[1]Gaziantep University, Naci Topçuoğlu Vocational School, Gaziantep, Türkiye.

[2]Isparta University of Applied Sciences, Atabey Vocational School, Isparta, Türkiye.

[3]Kütahya Dumlupınar University, Department of Computer Engineering, Kütahya, Türkiye

*Corresponding author e-mail: eyupcaki@gmail.com

Abstract

Supplying chain stores in different cities from a shared regional depot requires depot location and distribution routes to be considered together. Location approaches based solely on geographical coordinates may suggest sites that are unsuitable in terms of road connections and transport conditions. This study presents a genetic algorithm based approach to support regional depot location and vehicle routing decisions, considering store demand, vehicle capacities, and road accessibility. The objective is to evaluate depot regions and distribution routes in terms of total travel distance and the number of vehicles required to meet store demand.

The approach is implemented through a software application developed in C# and integrated with Google Maps. Store locations, demand quantities, and vehicle information are entered through a user interface. Distance calculations use road routes rather than straight-line distances between geographical points. The genetic algorithm employs permutation encoding, selection, order based crossover, insertion mutation, and elitist preservation. The depot area and its associated road connection are evaluated according to the distribution alternatives generated. Route usage intensity and store supply requirements are considered when identifying the road connection for the depot.

For the regional application, a dataset comprising real store locations across multiple cities is planned, with demand and vehicle capacity scenarios defined through explicit assumptions. The evaluation will examine total distribution distance, vehicle requirements, computational time, and solution variability across repeated runs. The study provides a decision support framework that considers regional depot location together with distribution routes and road connections.

Keywords: *Genetic Algorithm, Regional Depot Location, Vehicle Routing, Road Accessibility, Chain Stores*

Evaluating Video Game Player Satisfaction Through Text Mining and Sentiment Analysis of Online Reviews

Burak Kerem Demirdağ[1*], Deniz Özonur[2]

[1]Gazi University, Graduate School of Natural and Applied Sciences, Department of Statistics, Ankara, Türkiye
[2]Gazi University, Faculty of Science, Department of Statistics, Ankara, Türkiye

*Corresponding author e-mail: keremdemirdag@gmail.com

Abstract

Online video game reviews provide extensive feedback on player experience, but their volume and unstructured nature complicate systematic analysis. This study examines textual indicators of player satisfaction and the relationship between sentiment expressed in reviews and numerical user ratings. A Python-based web scraper collected 92,359 unique user reviews of 20 popular video games released between 2011 and 2023 from an online review website. Language identification, text cleaning, and eligibility filtering yielded 68,762 English-language reviews for analysis. User ratings were rescaled to 0–100 and grouped into negative (0–30), neutral (31–70), and positive (71–100) categories. Unigram and bigram features were represented using term frequency–inverse document frequency (TF-IDF) and used to train a logistic regression classifier with a stratified 80/20 training–test split. VADER sentiment scores were computed separately and compared with numerical ratings using Pearson and Spearman correlations. On the test set, the classifier achieved 85.86% accuracy, a weighted F1-score of 0.84, and a macro F1-score of 0.69. The lower macro F1-score reflected uneven performance across classes, particularly limited discrimination of neutral reviews in a dataset dominated by positive ratings (72.52%). Examination of model coefficients identified terms relating to playability, perceived quality, and monetary value as prominent indicators of dissatisfaction, whereas strongly evaluative praise characterized positive reviews. VADER compound scores showed a moderate positive association with user ratings (Pearson $r = 0.5706$; Spearman $\rho = 0.4317$). These findings demonstrate that interpretable text mining methods can identify linguistic patterns associated with player satisfaction, while moderate sentiment–rating correspondence and difficulty classifying neutral reviews indicate that textual sentiment alone does not fully capture numerical evaluations. The results are limited to reviews of the selected games on a single platform.

Keywords: *Text Mining, Sentiment Analysis, TF-IDF, Logistic Regression, Video Game Reviews*

From Rule-Based Daylight-Sensitive LED Control to AI-Powered Adaptive Lighting

Aykut Bilici[1], Bayram Arda Kuş[1]

[1]Gaziantep University, Naci Topcuoglu Vocational School, Control and Information Technologies Department, 27000, Gaziantep, Turkiye

*Corresponding author e-mail:aykutbilici@gaziantep.edu.tr

Abstract

Türkiye's actual electricity consumption reached 358.6 TWh in 2025, with 34% of generation from coal, two-thirds of it imported. This review revisits a low-cost daylight-responsive LED controller built in 2018 (Arduino Nano, BH1750FVI light sensor, PWM dimming) and positions its measured results within the 2018–2026 literature. In a 12-hour working-day scenario, the controller reduced daily lighting energy from 5.00 kWh (constant operation) to 1.48 kWh (70.4%), and by 49% relative to on–off control. This agrees with a 2025 ESP32 study using the same sensor (75.65%) and a Californian field test (about 73%). Recent literature shows that AI-enabled lighting research remains largely simulation-based. We propose a four-level maturity framework and a low-cost upgrade path that adds data logging, lightweight learning and wireless coordination to the original control loop.

Keywords: *daylight harvesting; LED dimming; PWM; Arduino; energy efficiency; adaptive lighting; machine learning; edge AI; TinyML; wireless sensor networks*

AI as a Research Toolsmith: Generating Bespoke Software for Specialized Research Workflows

Mohamed O.H. Mahmoud[1*], **Benedek Nagy**[1,2]

[1]Eastern Mediterranean University, Faculty of Arts and Science, Department of Mathematics, Famagusta, Northern Cyprus, via Mersin-10, Türkiye
[2]Eszterházy Károly Catholic University, Faculty of Informatics, Department of Computer Science, 3300 Eger, Hungary, Leányka u. 4.

*Corresponding author e-mail: mcshnaib@gmail.com

Abstract

Researchers often encounter specialized computational needs for which existing software is unavailable, too inflexible, or poorly suited to their workflows. Meeting these needs typically requires developing custom software or relying on technical collaborators, adding development effort to the research process. Recent advances in large language models (LLMs) offer an alternative by enabling software to be generated directly from natural-language requirements.

This paper investigates LLMs as research toolsmiths: systems capable of generating bespoke software for specialized research workflows. Rather than using AI to conduct or replace research, we examine its ability to transform researchers' computational requirements into practical, task-specific tools. We evaluate LLM-generated software across research-support scenarios including interactive visualization and manipulation of unconventional structures, automated document processing, and other specialized computational utilities.

Generated tools are evaluated for functional correctness, feature completeness, usability, and alignment with their intended workflows across five LLMs (Claude Opus 5.5, Composer 2.5, Gemini 3.8, GPT-5.6 Sol, and Grok 4.7). Through this evaluation, we examine both the potential of LLMs to provide on-demand research tooling and the limitations that arise when generated software must satisfy specialized requirements. Our findings aim to characterize when AI-generated software can effectively complement existing research infrastructure and reduce the technical barrier to creating bespoke research tools.

Keywords: *Large Language Models, Research Software, Bespoke Software, Research Workflows, AI-Assisted Development*

Near-Duplicate Records in Synthetic Telematics Data: How Validation Affects the Estimated Value of Machine Learning in Insurance Pricing

Oğuzhan Araçlı[1]*, Övgücan Karadağ Erdemir[2]

[1]Hacettepe University, Graduate School of Science and Engineering, Department of Actuarial Sciences, 06800 Ankara, Türkiye
[2]Hacettepe University, Faculty of Science, Department of Actuarial Sciences, 06800 Ankara, Türkiye

*Corresponding author e-mail: oaracli@hacettepe.edu.tr

Abstract

Telematics data, collected by in-vehicle devices or smartphone applications, record how, when and how much a vehicle is driven and allow motor insurance premiums to reflect actual driving behaviour. We examine whether the gains from telematics and machine-learning models on synthetic insurance data depend on how the models are validated. Using the synthetic dataset of So, Boucher and Valdez (2021), which contains 100,000 policies, we compare traditional, telematics and combined variable sets with Poisson–Gamma generalized linear models (GLMs), XGBoost and CatBoost. Under random cross-validation, telematics variables raised the pure-premium Gini index from 0.42 to 0.56 for the GLM and from 0.57 to 0.79 for XGBoost. However, the data contain groups of policies with nearly identical predictors and very similar claim outcomes: among nearest-neighbour pairs, 98% of claimants' neighbours also had a claim, against an overall claim rate of about 4%. Random splitting can put such pairs on both sides of the split, and boosting models benefit from this much more than the GLM. We therefore grouped similar policies by k-means clustering and kept each group within a single fold. Under this grouped validation, GLM results changed little, while the claim-frequency Gini advantage of tuned XGBoost and CatBoost over the GLM fell by 50–57%; boosting still outperformed the GLM, but by roughly half the margin. A placebo experiment, in which claims were redrawn with the same expected risk but without shared outcomes between similar policies, showed no such reduction. Telematics still added about 0.14 Gini for the GLM and 0.24 for XGBoost under grouped validation. For this dataset, grouped validation is the more reliable choice, and part of the boosting advantage under random splitting appears to come from the synthetic data structure. Whether the telematics gain carries over to real portfolios remains to be tested.

Keywords: *Telematics, Insurance Pricing, Synthetic Data, Grouped Cross-Validation, Gradient Boosting*

Hybrid Entity Matching and Decision Support for AI-Assisted Multi-Source Sanctions Monitoring

Yavuz Aslan[1], Onuralp Gayret[2*]

[1]Koçfinans, İstanbul, Türkiye
[2]Koçfinans, İstanbul, Türkiye

*Corresponding author e-mail: yaslan@kocfinans.com.tr

Abstract

The independent publication of international sanctions lists by different authorities creates operational challenges for financial institutions' compliance, customer identification, and risk monitoring processes. Variations in spelling, abbreviations, and alternative names limit the coverage of exact matching methods. Continuous assessment of unstructured news content increases manual review workloads. This study proposes an artificial intelligence-assisted decision support framework for consolidating heterogeneous sanctions data within a unified platform and analysing records through hybrid entity matching. The existing application comprises a React and TypeScript user interface, a Node/Express application layer, and a Python data collection engine. OFAC, United Nations, European Union, and United Kingdom sanctions sources, alongside FATF announcements and Reuters news, are monitored through source-specific modules. The Python engine runs as a subprocess, transferring results to the application layer in JSON format through standard input/output streams. Results are stored in a file-based cache. The proposed artificial intelligence layer is designed for integration into the existing data processing workflow. The entity matching pipeline will incorporate Unicode normalization, case normalization, punctuation removal, and token standardization. Original names and aliases will be retained to limit information loss. Candidate records will be identified through multiple blocking strategies using entity types, name tokens, and available auxiliary attributes, reducing computational costs relative to exhaustive pairwise comparisons. Country and birth date information will not serve as mandatory filters when missing or inconsistent values could exclude valid matches. The proposed hybrid method will combine exact matching, Levenshtein and Jaro-Winkler similarities, cosine similarity between Sentence-BERT-like transformer representations, and auxiliary attribute consistency. Matching weights and decision thresholds will be determined using a validation set. Since semantic proximity between names does not independently establish identity, outputs will be interpreted as potential matches. Ambiguous records will undergo human review, supported by source links and matching attributes. A separate transformer-based classifier is proposed for identifying sanctions-related news. Experimental evaluation will use manually labelled matching and nonmatching record pairs. Entity-level partitioning will prevent aliases of the same entity from appearing across training and test sets. Exact, fuzzy, transformer-based, and hybrid matching will be compared using precision, recall, F1-score, false positive rates, and false negative rates. Blocking recall, comparison counts, and processing time will be measured. Ablation experiments will assess individual component contributions, while news classification will use a separately labelled dataset. Threshold analysis will examine false alarm workloads under high-recall requirements. The study aims to support effective sanctions monitoring through measurable evaluation and traceable decision support.

Keywords: *Artificial Intelligence, Sanctions Monitoring, Hybrid Entity Matching, Natural Language Processing, Semantic Similarity, Decision Support Systems.*

Response Surface Methodology and Machine Learning Approaches: A Comparative Study for Surface Roughness Prediction

Oğuzhan Çalışkan[1*], Deniz Özonur[2]

[1]Gazi University, Graduate School of Natural and Applied Sciences, Department of Statistics, Ankara, Türkiye

[2]Gazi University, Faculty of Science, Department of Statistics, Ankara, Türkiye

*Corresponding author e-mail: oguzhanclskntr@gmail.com

Abstract

In Design of Experiments studies, accurately modeling the relationship between experimental factors and response variables is essential for understanding processes and determining appropriate operating conditions. Response Surface Methodology is widely used for this purpose because it provides an effective statistical framework for modeling factor-response relationships and identifying suitable operating conditions. However, when these relationships become complex or nonlinear, machine learning approaches may provide alternative modeling capabilities.

This study aims to compare Response Surface Methodology with different Machine Learning approaches for predicting surface roughness in a milling process. Experimental data obtained from the milling of AA6061 aluminum alloy were used in the study (Eser et al., 2021). The dataset consists of 27 experimental observations, including cutting speed, feed rate and depth of cut as independent variables. The response variable is the average surface roughness. For each experiment, the surface roughness value was determined by averaging three measurements obtained from the machined surface.

First, an Response Surface Methodology based model will be developed using the experimental factors and surface roughness response. The same dataset will then be modeled using different Machine Learning algorithms, and their predictive performances will be compared with the Response Surface Methodology model. Appropriate performance metrics will be used to evaluate the predictive accuracy and overall performance of the developed models. Particular attention will be given to the performance of the methods with a small sample size and in modeling nonlinear relationships between machining parameters and surface roughness. The study therefore aims to identify the advantages and limitations of Response Surface Methodology compared with Machine Learning approaches and to provide an assessment of their suitability for surface roughness prediction in milling processes.

Keywords: *Design of Experiments, Response Surface Methodology, Machine Learning, Model Comparison, Predictive Performance*

Web-Based Statistical Reporting with R: An Application Framework for Educational Feedback

Mehmet Sinan İyisoy[1*]

[1]Necmettin Erbakan University, Faculty of Medicine, Department of Medical Education and Informatics, 42090, Konya, Turkiye

*Corresponding author e-mail: siyisoy@gmail.com

Abstract

Educational feedback data can support institutional quality improvement, but their practical value depends on how effectively they are analysed and communicated. Static reports provide limited opportunities for accessing updated results through existing institutional systems. This paper describes an R based approach to web accessible statistical reporting, illustrated through educational feedback collected within a medical faculty.

The institutional environment consists of a Debian based virtual server hosting a Django educational management system and a PHP module for collecting student feedback on practical courses. Within this setting, R serves as the analytical engine for processing feedback data and producing statistical reporting components. The presented workflow covers data preparation, statistical computation, visualization and generation of web compatible tables. A Plumber API layer provides a means of connecting R generated outputs with web services, allowing analytical functions to be maintained separately from the existing application components.

The architectural discussion focuses on the flow from collected feedback to browser accessible reports, the responsibilities of individual software components and operational considerations for hosting R based services on Linux. Particular attention is given to access control, handling analysis requests and avoiding unnecessary repetition of computationally intensive operations. The approach is presented as an application framework; improvements in reporting efficiency, usability or institutional decision-making have not been quantitatively evaluated.

By connecting statistical computing with an existing web infrastructure, this framework outlines a practical route for making R based analyses accessible beyond standalone scripts and locally generated documents. Future evaluation should examine analytical accuracy, report update procedures, service reliability and the usefulness of the resulting reports for academic administrators.

Keywords: *R, Web Based Reporting, Plumber, Educational Feedback, Statistical Analysis*

Personnel Scheduling Optimization for Hybrid Work Model Using Genetic Algorithm

Yusuf Ömer Erem[1*], Ahmet Alper Doruk[2], Nurullah Obuz[2] , Selçuk Alp[1]

[1]Yıldız Technical University, Faculty of Arts and Sciences, Department of Statistics, 34220, İstanbul, Türkiye

[2]F.İ.T. Bilgi İşlem A.Ş., 34220, İstanbul, Türkiye

*Corresponding author e-mail: omer.erem@std.yildiz.edu.tr

Abstract

Hybrid work models have become a permanent feature of organizational life in the post-pandemic era. However, constructing weekly office/remote schedules that simultaneously satisfy multiple departmental constraints, individual preferences, and operational requirements represents an NP-hard combinatorial optimization problem. In this study, a 12-week personnel schedule incorporating 16 work constraints for a real-world organization comprising 142 employees across 23 departments was optimized using a Genetic Algorithm (GA). The problem is encoded as an N×5 binary chromosome matrix, where each gene represents an employee's status on a given day: office (1) or remote (0). The fitness function is defined through a weighted penalty mechanism based on constraint violations; hard constraints (departmental minimum office quotas, consecutive remote-work day limits, simultaneous senior manager absence cap) are assigned high penalty weights, while soft constraints are penalized lightly. Tournament selection (k=3), a generation-adaptive hybrid crossover strategy (uniform crossover in early generations, two-point crossover in later generations), and adaptive bit-flip mutation are employed. To accelerate convergence, 20% of the initial population is seeded using a constraint-aware greedy heuristic. Experiments conducted with a population size of 60 and a maximum of 150 generations demonstrate that schedules satisfying the vast majority of hard constraints, with an average fitness score of -3,749, are produced in under thirty seconds. The results confirm that the GA-based approach offers a practical and efficient solution for large-scale hybrid work scheduling.

Keywords: *Genetic Algorithm, Personnel Scheduling, Hybrid Work Model, Combinatorial Optimization, Workforce Planning*

Enhancing Turkish Fine-Tuning Datasets through a RAG-Based AI Agent: A Comparative Evaluation Using Llama 3

Hamza Görmez[1*], **Ethem Kelekçi**[1]

[1]Manisa Celal Bayar University, Hasan Ferdi Turgutlu Faculty of Technology, Department of Mechatronics Engineering, 45400, Manisa, Türkiye

*Corresponding author e-mail: 232801021@ogr.cbu.edu.tr

Abstract

The performance of Large Language Models (LLMs) depends substantially on the quality of the instruction-tuning datasets used during fine-tuning. However, high-quality instruction datasets remain scarce for low-resource languages such as Turkish. This study investigates the impact of a data-quality enhancement framework, implemented through a user-interface-guided artificial intelligence agent supported by a Retrieval-Augmented Generation (RAG) architecture, on the fine-tuning performance of LLMs. To improve linguistic fluency, contextual richness, and semantic consistency, two datasets were constructed: a raw instruction dataset and a hybrid dataset primarily improved by the RAG-based agent while retaining a limited proportion of the original raw samples. Comparative fine tuning experiments were conducted on both the general-purpose Llama-3-8B model and the Turkish-specialized Trendyol-Llama-3-8B model using Low-Rank Adaptation (LoRa) combined with 4-bit quantization. The experimental results demonstrate that the general-purpose Llama-3-8B model trained on the RAG-enhanced hybrid dataset consistently improved across the evaluated metrics, while the Turkish-specialized Trendyol-Llama-3-8B model exhibited metric-specific divergences. In particular, the hybrid dataset yielded absolute improvements of 3.6 percentage points in BERTScore F1 and 4.8 percentage points in ROUGE-1, together with notable gains in METEOR and BLEU scores. Furthermore, the Turkish-specialized Trendyol-Llama-3-8B model achieved superior results in several performance metrics when trained on the hybrid dataset, indicating that language-specific pretraining knowledge can further benefit from systematically enhanced instruction data. These findings highlight the effectiveness of RAG-based data curation and improvement as a practical approach for maximizing the capabilities of LLMs, even in settings where access to high-quality linguistic resources is limited.

Keywords: *Fine-tuning, Retrieval-augmented generation, Artificial intelligence agent, Llama-3-8B, Turkish language*

Accuracy, Reliability, and Energy Cost in Clinical Multi-Class Prediction: A Comparison of Gradient Boosting, Tabular Foundation Models, and Large Language Models

Yasemin Günter[1,2*], Nimet Yapıcı Pehlivan[3]

[1]Haliç University, Vocational School, Department of Big Data Analytics, 34060, İstanbul, Turkey
[2]Yıldız Technical University, Faculty of Arts and Science, Department of Statistics, 34220, İstanbul, Turkey
[3]Selçuk University, Faculty of Science, Department of Statistics, 42130, Konya, Turkey

*Corresponding author e-mail: yasemingunter@halic.edu.tr

Abstract

Model selection in clinical decision support systems is mostly based solely on predictive performance. However, in clinical practice, the agreement between the probabilities produced by a model and the observed outcomes (calibration) is at least as important as accuracy. Moreover, increasingly large artificial intelligence models, such as tabular foundation models and large language models, entail substantial computational and energy costs. In this study, three model families: gradient boosting (XGBoost, LightGBM, CatBoost), a tabular foundation model (TabPFN-3.5), and a small, locally deployable large language model (Qwen2.5-3B-Instruct; zero-shot and few-shot) were compared in terms of accuracy, reliability, and energy consumption on multi-class clinical staging tasks. Logistic regression was used as the reference model. Analyses were conducted on two publicly available datasets from the UCI Machine Learning Repository: Hepatitis C dataset and Heart Disease dataset. Models were evaluated using 5×5 repeated stratified cross-validation, with performance measured by macro-F1 and quadratic weighted kappa, reliability by expected calibration error (ECE) and Brier score, and energy consumption by CodeCarbon. Since performance estimates from cross-validation folds are not independent, each model was compared with logistic regression using the Corrected Resampled t-test based on the Nadeau and Bengio (2003). The macro-F1 scores of the gradient boosting models did not differ significantly from those of logistic regression, and their calibration errors on the Heart Disease dataset were higher than those of logistic regression. On the Hepatitis C dataset, the tabular foundation model achieved a significantly higher macro-F1 and a significantly lower Brier score than logistic regression. Furthermore, on the Heart Disease dataset, it achieved the lowest calibration error among all evaluated models. The large language model, despite having the highest energy consumption, yielded the lowest macro-F1 and the highest calibration error across both datasets. These findings demonstrate that accuracy, calibration, and computational cost should be considered together in clinical model selection.

Keywords: *Clinical Decision Support Systems, Probability Calibration, Tabular Foundation Models, Large Language Models, Green AI*

Beyond Neuronal Plasticity: Microglia as a Bounded Biological Regularization Layer for Long-Term Organoid Intelligence

Özkan Günalp[1]*, Olaf Yunus Laitinen Imanov[2], Ayyüce Akyol[3], Rana Irem Turhan[4], Duygu Erişken[5], Tuğba Ceren Turhan[6]

[1]Ege University, Institute of Health Sciences, Medical Biology, İzmir, Türkiye
[2]University of Helsinki, Department of Mathematics and Statistics, Helsinki, Finland [3]Middle East Technical University, Department of Biology, Faculty of Arts and Sciences, Ankara, Türkiye
[4]University of Sheffield, Artificial Intelligence, School of Computer Science, Sheffield, United Kingdom
[5]Trakya University, Faculty of Science, Mathematics, İstanbul, Türkiye
[6]Riga Technical University, Computer Systems, Riga, Latvia

*Corresponding author e-mail: ozkn.gunalp@gmail.com

Abstract

Organoid intelligence (OI) proposes the use of human stem-cell-derived neural organoids as computational substrates. Proof-of-concept systems have performed speech recognition and nonlinear equation prediction through adaptive reservoir computation, and closed-loop electrical training has produced goal-directed adaptation in cortical organoids. Recent evidence, however, indicates that training-induced performance gains do not persist after a rest interval, which shifts attention from the acquisition of computation-relevant network change to its durability. Current OI roadmaps identify scaling, maturation, interfacing and reproducibility as bottlenecks, and already recommend enriching organoids with glial cells; what they do not specify is a cellular architecture that regulates plasticity across long training horizons. This perspective formulates two linked and explicitly unproven constructs. The Organoid Network Saturation Hypothesis (ONSH) proposes that structural and functional plasticity which is not homeostatically constrained may progressively reduce state separability, dynamic range and sensitivity to novel input. Immuno-Homeostatic Organoid Intelligence (IHOI) proposes a four-layer architecture in which an engineered microglial compartment acts as a slow, task-independent regulator operating alongside fast, task-aware external feedback. We argue that microglia should be modelled as a bounded biological regularizer rather than as a task-aware optimizer, because microglia have no access to a task loss function. We derive eight falsifiable predictions, an operational definition of self-regulation based on the half-life of retained performance gain, five competing explanatory models and three criteria under which the proposal should be abandoned. We further specify why the framework is not reducible to existing astrocyte-centred computational glia theories. IHOI is presented as a testable research paradigm rather than as a demonstrated technology.

Keywords: *Organoid Intelligence, Microglia, Neuroimmune Homeostasis, Reservoir Computing, Biological Regularization*

A Statistical-Computational Framework for Privacy-Preserving Federated Reliability Inference

Salih Demir[1]*

[1]Department of Computer Science, Faculty of Science, Ankara University, Ankara, Türkiye

*Corresponding author e-mail: sdemir@ankara.edu.tr

Abstract

Privacy-preserving reliability analysis often requires combining failure information held by multiple assets or organizations without centralizing raw data. This study develops a statistical-computational framework in which the communication structure is determined directly by the likelihood representation. Three cases are distinguished: one-shot aggregation based on parameter-independent summaries, iterative protected evaluation when local likelihood contributions depend on the current parameter, and local elimination of nuisance parameters before assigning the communication regime.

The framework is evaluated using two engineering reliability benchmarks. For the 13-aircraft Proschan data, exponential estimation and fleet-homogeneity testing are recovered from a single protected summary. The two-parameter Gamma lifetime model is also one-shot because its pooled likelihood is completely determined by the sample size, the sum of lifetimes, and the sum of log-lifetimes. Weibull inference requires parameter-dependent evaluations; a safeguarded profile-Newton procedure reaches the centralized maximum-likelihood estimate in five protected evaluations from the reference start and converges to the same solution from all nine examined positive starting values. For the failure-truncated power-law process, conditioning removes aircraft-specific scale parameters and yields exact finite-sample common-trend inference, with $\hat{\beta} = 1.13537$ and $p = 0.07882$.

The Farley-1 nuclear-pump application further illustrates one-shot conditional Poisson inference and iterative Gamma-Poisson random-effects estimation. The intermittent-to-continuous failure-rate ratio is 6.57. Across the studied models, protected estimates match their centralized counterparts to numerical tolerance. The results show that federated reliability inference can be made more communication-efficient by exploiting statistical representation before introducing iterative optimization.

Keywords: *Federated reliability, Secure aggregation, Likelihood inference, Repairable systems, Privacy-preserving statistics*

From Evidence Sufficiency to Abstention: A Conceptual Framework for Biomedical RAG Decision Support

Özkan Günalp[1]*, Olaf Yunus Laitinen Imanov[2], Rana Irem Turhan[3], Duygu Erişken[4], Tuğba Ceren Turhan[5]

[1]Ege University, Institute of Health Sciences, Medical Biology, İzmir, Türkiye

[2]University of Helsinki, Department of Mathematics and Statistics, Helsinki, Finland

[3]University of Sheffield, Artificial Intelligence, School of Computer Science, Sheffield, United Kingdom

[4]Trakya University, Faculty of Science, Mathematics, İstanbul, Türkiye

[5]Riga Technical University, Faculty of Computer Science, Information Technology and Energy, Riga, Latvia

Corresponding-author email: ozkn.gunalp@gmail.com

Abstract

Retrieval-augmented generation (RAG) can connect biomedical answers to external evidence, but retrieving relevant documents does not establish that the delivered evidence is sufficient, that the generated answer is supported, or that an answer is appropriate for a particular clinical use. This paper develops a literature-informed conceptual framework for making these distinctions explicit. A targeted narrative synthesis of primary research and closely related conceptual work was used to derive operational definitions, decision rules, and testable propositions. Public PubMedQA development records and explicitly constructed variants illustrate the rules; no model experiment or clinical validation was performed. The framework comprises five stages: question and task specification, source and retrieval tracing, delivered-context sufficiency assessment, answer and citation support assessment, and action selection with an audit record. It distinguishes SUFFICIENT, PARTIAL, ABSENT, and CONFLICTING evidence from an UNASSESSED assessment status, while evaluating candidate-answer support separately. Under a declared closed-evidence policy, ANSWER requires suitable and sufficient evidence and support for every decision-bearing claim. ABSTAIN withholds an unsupported substantive answer; ESCALATE identifies a specific information request or accountable review pathway when a prespecified trigger applies. Operational failures are recorded separately. The proposed contribution is a biomedical synthesis and adaptation of existing sufficiency, attribution, selective-prediction, and clinical-use concepts, rather than a claim to the first integrated framework. Three propositions identify how future studies could test unsupported answering, escalation quality, and error attribution without treating increased refusal as evidence of safety. The framework offers a traceable design vocabulary; its reliability, clinical utility, and effects on patient outcomes remain unestablished.

Keywords*: Retrieval-Augmented Generation, Evidence Sufficiency, Abstention, Biomedical Question Answering, Clinical Decision Support*

Hypotheses, Not Methods: What Machine Unlearning Can Give Back to the Study of Human Unlearning

Olaf Yunus Laitinen Imanov[1], Özkan Günalp[2*], Ayşe Tuba Tuğrul[3], Duygu Erişken[4], Rana İrem Turhan[5], Tuğba Ceren Turhan[6], Sıla Pehlivan[7]

[1]University of Helsinki, Department of Mathematics and Statistics, Helsinki, Finland
[2]Ege University, Faculty of Medicine, Department of Medical Biology, İzmir, Türkiye
[3]Institute of Science, Civil Engineering, Pamukkale University, Denizli, Türkiye
[4]Trakya University, Faculty of Science, Department of Mathematics, Edirne, Türkiye
[5]University of Sheffield, Artificial Intelligence, School of Computer Science, Sheffield, United Kingdom
[6]Riga Technical University, Faculty of Computer Science, Information Technology and Energy, Computer Systems, Riga, Latvia
[7]Istanbul Beykent University, Faculty of Engineering and Architecture, Department of Computer Engineering, İstanbul, Türkiye

*Corresponding author e-mail: ozkn.gunalp@gmail.com

Abstract

Machine unlearning aims to remove the influence of specified training data from a trained model without retraining it from scratch, and it has become central to data protection and AI safety. Most bridges between this field and the science of memory run in one direction: findings about human forgetting inspire unlearning algorithms, and relapse tests from fear-extinction research have been proposed as an evaluation standard for machines. This position paper examines the reverse direction. Can solutions developed for machine unlearning be applied to human unlearning, and how are the two related? Using a structured conceptual comparison organised by Marr's three levels of analysis, we map nine families of unlearning methods onto human counterparts, among them retrieval suppression, thought substitution, extinction, reconsolidation-based updating and response inhibition, and grade their overlap. The two fields share a computational problem, overlap only partly in their algorithms and diverge in implementation. A short formalisation separates erasure from suppression relative to a never-exposed reference and shows why human claims of erasure can rest only on group-level controls. From this analysis we derive six channels of reverse transfer (a design-space taxonomy, a fully accessible testbed, detection of residual traces without recovery, graded specificity testing, forgettability by design, and the distinction between data-level and knowledge-level unlearning) and six falsifiable propositions, each paired with a machine test and a human test. What transfers are hypotheses and evaluation logic, not procedures: gradient-based methods have no human analogue, and no person can be compared with a version of themselves that never learned. The paper does not endorse editing human memory; it proposes a shared agenda in which networks serve as inexpensive places to test ideas about forgetting before they are tested in people.

Keywords: *Machine Unlearning, Human Unlearning, Extinction Learning, Memory Suppression, Cognitive Modeling*

Evaluating the Attractiveness of OECD Countries for Foreign Health Workforce: A CRITIC-Weighted CoCoSo Performance Analysis

Aysegül Yıldız[1]

[1]Kütahya Dumlupinar University, Faculty of Economics and Administrative Sciences, Department of Econometrics, 43100, Kütahya, Türkiye

*Corresponding author e-mail: aysegul.yildiz@dpu.edu.tr

Abstract

In addition to countries' ability to maintain their local nursing and doctor capacities, attracting qualified healthcare professionals from abroad is a key indicator of the sustainability of their health systems, the quality of services, global competitiveness, and high living standards. This study holistically analyzes the relationship between general health infrastructure capacities, foreign health workforce attractiveness, and the overall citizenship/integration climate of 21 OECD countries for which data are available, using data from 2023, via the CoCoSo method, a multi-criteria decision-making approach. The following evaluation criteria have been used in the study: foreigners acquiring citizenship (%), density of doctors and nurses (per 1000 population), foreign-trained nurses (% of total), foreign-trained doctors (% of total), foreign-trained share in doctors and nurses inflows (%). To determine the weights of these criteria, the CRITIC method has been considered, and the criteria with the highest weight in the performance success of countries in managing qualified health migration have been identified. Empirical findings show that the three most important criteria are density of doctors (w_j =0.2217), citizenship acquisition rate (w_j =0.1805), and the share of foreign-trained doctors (w_j =0.1582). When the final performance scores of the countries have been examined, Ireland (k_i=4.49), Norway (k_i=4.39), and Switzerland (k_i=4.32) have been determined to have the highest performance scores due to their high density of healthcare personnel and their capacity to attract healthcare professionals trained abroad. Results provide guidance to policymakers on how countries can improve their attractiveness to qualified healthcare workers globally.

Keywords: *Foreign-trained health workforce, CoCoSo, CRITIC, Multi-criteria Decision Making*

On Multiple Birth Support Vector Machines with k-Proximal Plane Clustering

Güvenc Arslan[1*], Elif Yüceoğlu[2]

[1]Kırıkkale University, Faculty of Engineering and Natural Sciences, Statistics Department, 71450, Kırıkkale, Turkiye

[2]Kırıkkale University, Graduate School of Natural and Applied Sciences, Statistics Department, 71450, Kırıkkale, Turkiye

*Corresponding author e-mail: guvenc.arslan@kku.edu.tr

Abstract

Multiple Birth Support Vector Machines (MB-SVM) were introduced as an effective extension to Support Vector Machines (SVM). Although the basic idea is as in SVMs, optimal non-parallel hyperplanes are used for each class category. On the other hand, one may observe that there are studies that use k-means clustering with SVMs, which reduce the size of the training data set while obtaining similar performance as standard SVMs. Some recent clustering approaches such as k-Plane Clustering (kPC), and an improved version of it called the k-Proximal Plane Clustering (kPPC), cluster data points with respect to the center planes instead of clustering data points to center points as in k-means. Motivated by these studies, we investigate the use of kPPC in MB-SVMs.

Keywords: *Support Vector Machines, Multiple Birth, k-Plane Clustering, k-Proximal Plane Clustering.*

From Data Science for Internet of Things (IoT) to Data Science for Internet of Medical Things (IoMT): Applications and Examples in Healthcare Services

Mehmet Karakoç[1*]

[1]Alanya University, Faculty of Engineering and Natural Sciences, Department of Computer Engineering, 07400, Antalya, Türkiye

*Corresponding author e-mail: mehmet.karakoc@alanyauniversity.edu.tr

Abstract

Data Science (DS) for Internet of Things (IoT) is the practice of analyzing machine-generated sensor data to extract real-time applicable information/knowledge (actionable insights). To turn these continuous machine signals (sensor streams) into useful-actions to take, smart devices are combined/equipped with high-level data analysis. In this process, to collect/gather, store, clean, filter, compress, exchange/transmit, analyze, and visualize the data from these interconnected physical devices, DS approaches are applied. These are analytics (cloud/server and/or distributed), artificial intelligence, machine learning, and statistical methods. Algorithms-models-techniques of deep learning, natural-language processing, or time-series analysis are also used. Moreover, the automation of intelligent alerts-controls-operations and real-time decision-making are enabled. Predictive maintenance for the break-time of factory items/parts before occurring is only an example. In this study, we address Data Science for Internet of Medical Things (IoMT/IoHT: Internet of Healthcare Things), supported with applications-examples. When continuous health-data from the medical devices are combined with the approaches aforementioned, the process of DS-for-IoT becomes DS-for-IoMT by transforming patient care. Some of its key-uses are chronic illness management and monitoring/tracking patients in real-time against anomalies or critical-threshold violations (blood glucose/oxygen-levels/pressure, heart-rate, oxygen-saturation). Asthma, cardiac-anomalies, diabetes, heart-disease, and hypertension are some other cases/situations. Detecting changes/progression over time and finding patterns in electronic health records are also involved. Additionally, critical medical-events can be estimated such as blood-poisoning/sepsis, epileptic-seizures, or heart-failure before clinical-symptoms appear. Some other examples are hospital operations optimization, personalized treatment plans, or predictive diagnostics. However, delay in instant responses, high-volume data costs, and safety/security of private sensor streams are few major problems. Cyber threats/attacks and unauthorized/unverified system/network-access attempts must be detected, recognized, and eliminated in a non-stop manner for the security issues especially targeting the connected medical infrastructure. We believe that this is a matter that deserves utmost attention for the field of medicine in which vital things are involved.

Keywords: *Data Science, Internet of Things (IoT), Internet of Medical Things (IoMT), Healthcare Services, and Medicine.*

Business Intelligence for E-Commerce Decision Support: An Exploratory Analysis of Retail Transaction Data Using Power BI

Grâce Gloria Akala[1*], Emeraude Lefouka Bawangui[1], Lawrence A. Farinola[1]

[1]Final International University, Girne, North Cyprus

*Corresponding author e-mail: lawrence.farinola@final.edu.tr

Abstract

The increasing availability of transactional data creates opportunities for organizations to improve business decisions through business intelligence (BI) and data visualization. However, raw transactional records do not automatically provide actionable information for managers. This study aims to demonstrate how Power BI can transform retail transaction data into interactive visualizations that support e-commerce decision-making. Using a publicly available retail transaction dataset, the study will adopt an exploratory data analysis approach to examine sales performance, product demand, customer purchasing patterns, and temporal variations in transactions. The dataset will be prepared through data cleaning, transformation, and aggregation before being analyzed using descriptive statistics and visual analytics. Power BI will be used to develop an interactive dashboard incorporating key performance indicators, sales trends, product-level comparisons, and customer-related summaries where the available data permit. The analysis will focus on how these visualizations can help decision-makers identify high-performing products, recognize purchasing patterns, monitor business performance, and investigate potential opportunities for improvement. Rather than proposing a new predictive algorithm, the study emphasizes the practical integration of data preparation, business intelligence, and decision-support visualization within an accessible analytical environment. The expected contribution is a reproducible, low-cost BI workflow that illustrates how existing transactional data can be converted into decision-relevant information without requiring primary data collection or paid analytical software. The study will also discuss the limitations of descriptive analytics and the implications for more advanced predictive and prescriptive decision-support applications. The findings will be documented through the completed analysis, dashboard visualizations, and managerial interpretations.

Keywords: *Business Intelligence, Power BI, E-Commerce Analytics, Decision Support Systems, Data Visualization*

The Impact of Renewable Energy on Economic Growth in Selected European Countries and Türkiye

Ahmet Akın Atasoy[1*]

[1]Turkish Statistical Institute, Ankara, Turkiye

*Corresponding author e-mail: ahmet.atasoy@tuik.gov.tr

Abstract

This doctoral dissertation examines the impact of renewable energy sources on economic growth within a comparative and multidimensional framework, focusing on Türkiye and selected European countries. The main research problem concerns whether the relationship between renewable energy and economic growth exhibits a homogeneous and linear structure across countries, under which structural conditions this relationship strengthens or weakens, and how it differs across energy types. This problem speaks directly to contemporary policy debates on energy supply security, climate policy, and sustainable economic growth.

Within this scope, the effects of hydroelectric, solar, wind, biofuel, and total renewable energy production on economic growth are analyzed using a multi-country panel dataset covering the period 2000–2023. The empirical strategy employs fixed- and random-effects models alongside group-based panel regressions, Chow structural break tests, panel threshold regressions (Hansen, 1999), interaction and mediation analyses, as well as rolling-window and time-varying coefficient methods. This multi-layered methodological framework aims to comprehensively capture the non-linear, heterogeneous, and dynamic nature of the renewable energy–economic growth relationship.

The empirical findings indicate that the relationship between renewable energy and economic growth differs significantly across energy types, country groups, and structural conditions. Wind and biofuel energy exhibit relatively consistent growth effects, whereas solar energy contributes significantly to economic growth only after certain income and technological thresholds are surpassed. The threshold analyses underscore the importance of critical capacity and income levels, particularly for technology-intensive renewable energy sources. The comparison between Türkiye and the selected European countries reveals that, despite Türkiye's considerable renewable energy potential, its capacity to translate this potential into economic growth remains more limited relative to the selected European countries. The study concludes that renewable energy policies should move beyond a purely capacity-oriented approach and be supported by a comprehensive development strategy centered on technology, human capital, and innovation.

Keywords: *Renewable Energy, Economic Growth, Panel Data Analysis, Threshold Effects, Heterogeneity, Turkiye, European Union*

Business Intelligence for Small Business Decision-Making: A Power BI Dashboard Approach

Assey Sephora Oceanne Christ-Liliane Toto[1*], Lawrence A. Farinola[1]

[1]Final International University, Girne, North Cyprus

*Corresponding author e-mail: assey.toto@final.edu.tr

Abstract

Small businesses generate valuable transactional data through their daily operations, but transforming these records into actionable information remains a challenge for effective managerial decision-making. Business intelligence (BI) tools can help address this challenge by consolidating, analysing, and visualising business data in accessible formats. This study aims to demonstrate how Microsoft Power BI can support small-business decision-making through the exploratory analysis and visualisation of online retail transaction data. The study will use a publicly available retail dataset containing information on transactions, product descriptions, quantities, unit prices, dates, and customer markets. Following data cleaning and preparation, descriptive analytics will be applied to examine sales performance, temporal trends, product-level patterns, and geographical differences in sales activity. The resulting Power BI dashboard will present key performance indicators and interactive visualisations designed to help users explore business performance from multiple perspectives. The analysis will focus on translating observed patterns into practical managerial insights concerning sales monitoring, product prioritisation, and market assessment. Rather than claiming to establish a causal relationship between dashboard use and business performance, the study will demonstrate the potential of accessible business intelligence tools to improve the organisation and interpretation of transactional information. The contribution is a practical and reproducible analytical workflow that connects retail data analysis with managerial decision support without requiring primary data collection or paid dashboard-development software. The study will also discuss the limitations of descriptive analytics and the implications for the use of more advanced business intelligence techniques in small businesses.

Keywords: *Business Intelligence; Small Business; Decision Support; Power BI; Sales Analytics*

The Dark Side of Service Robots in Tourism and Hospitality

Evrim Celtek [1*]

[1]Tokat Gaziosmanpaşa University, Zile Dinçerler Tourism Business and Hotel Management College, Recreation Management, 6000, Tokat, Türkiye

*Corresponding author e-mail: evrimceltek@yahoo.com

Abstract

The increasing use of service robots in tourism and hospitality enhances operational efficiency and automates routine tasks. Furthermore, service robots have transformed service delivery by supporting organizational productivity. However, these technological advantages raise significant questions regarding the preservation of human-centric hospitality—a concept where interpersonal interaction, emotional authenticity, empathy, and personalized attention lie at the heart of the guest experience. This study aims to critically examine the potential psychological, ethical, social, and organizational risks associated with the use of service robots in tourism and hospitality. It synthesizes research on service robots, human-robot interaction, service automation, and hospitality experiences, alongside real-world examples. The analysis focuses on four interconnected dimensions: (1) the potential erosion of interpersonal interaction and emotional authenticity; (2) psychological reactions to robotic service interactions, such as discomfort, mistrust, and algorithmic avoidance; (3) ethical concerns regarding privacy, surveillance, emotional manipulation, and individual autonomy; and (4) impacts on employees, including job insecurity, shifting professional roles, and the potential dehumanization of service work. Rather than assuming that robotic automation inevitably diminishes hospitality the study also explores how human-robot collaboration can enhance service delivery by enabling employees to focus more on meaningful guest interactions. Drawing on these perspectives, the study aims to provide an integrated classification of the risks that elucidates the conditions under which technological efficiency can either support or undermine the relational foundations of hospitality. By integrating technological, psychological, ethical, and organizational perspectives, the study seeks to contribute to a more balanced understanding of service robot usage and to inform the development of responsible, human-centric robotic service strategies within the tourism and hospitality sectors.

Keywords: *Service Robots, Tourism and Hospitality, Human-Robot Interaction, Dehumanization*

Cybersecurity, User Trust, and Mobile Banking in Liberia: Evidence from a Quantitative Study

Charles Zulu Yonmah, Nuriye Sancar, Oluwafemi Ayotunde Oke, Nadire Cavus

[1]Faculty of Artificial Intelligence and Informatics, Department of Computer Information Systems, Near East University, Cyprus;
[2]Department of Mathematics, Near East University, Cyprus
[3]Computer Information Systems Research and Technology Centre, Near East University, 99138 Nicosia,Cyprus

* Corresponding author e-mail: nuriye.sancar@neu.edu.tr

Abstract

The growing adoption of mobile banking has expanded access to financial services in Liberia, while raising important questions about the security of digital transactions and users' confidence in financial technology. Understanding how users perceive cybersecurity is essential for strengthening trust in mobile banking and supporting the development of inclusive digital financial infrastructure. This study investigates mobile banking users' perceptions of cybersecurity in Liberia, focusing on perceived confidentiality, perceived authentication, device trust, security misconfiguration, and cybersecurity awareness. A quantitative research design was employed, using structured questionnaire responses from 400 mobile banking users. The research instrument demonstrated good internal consistency, with a Cronbach's alpha coefficient of 0.864. Data were examined using descriptive statistics, the Mann–Whitney U test, and the Kruskal–Wallis H test. The findings indicate generally positive perceptions of mobile banking security and cybersecurity awareness among respondents. Furthermore, the analysis identified no statistically significant demographic differences in cybersecurity awareness by gender, age, or educational level. These results suggest that cybersecurity perceptions and awareness should be considered important components of mobile banking development, rather than issues addressed solely through technical safeguards. The study highlights the need for sustained investment in security technologies, appropriate regulatory oversight, and accessible cybersecurity education to reinforce user confidence and reduce potential vulnerabilities. By providing empirical evidence from Liberia, this research contributes to discussions on secure digital financial services and the development of resilient, inclusive financial infrastructure in emerging economies, consistent with Sustainable Development Goal 9.

Keywords: *Cybersecurity, User Trust, Mobile Banking, Cybersecurity Awareness, Digital Financial Services, Liberia, Financial Inclusion, SDG 9.*

Cybersecurity, Reliable Data, and Sustainable Technology in Aviation in the Age of Artificial Intelligence: International Airport Applications and a Development Model for Ercan International Airport

Tolga Öz[1*]

[1]Dean of Air Transportation Faculty, Akdeniz Karpaz University, Lefkoşa, NRCR

*Corresponding author e-mail: tolga.oz@akun.edu.tr

Abstract

The digital transformation of airports is improving the speed, efficiency, and traceability of aviation processes through artificial intelligence (AI), biometric identification, smart security screening, real-time operational analytics, and interconnected information systems. However, digitalization also creates new dependencies among information technology (IT), operational technology (OT), passenger services, flight and ground operations, supply chains, and critical infrastructure. Consequently, cybersecurity can no longer be treated solely as a technical function for protecting information systems; it has become a strategic management concern directly related to aviation safety, aviation security, and operational continuity. This study examines the relationship between AI applications and cybersecurity in aviation, reviews selected technology applications at international airports, and proposes a digital resilience model for Ercan International Airport. The study adopts a conceptual review approach based on International Civil Aviation Organization (ICAO) guidance and publicly documented airport technology applications. AI-supported aircraft turnaround analytics and security-screening technology initiatives at Amsterdam Airport Schiphol, together with biometric and automated border-clearance processes at Singapore Changi Airport, are examined as examples of data-driven airport management. The conceptual analysis highlights data quality, cyber-risk management, human oversight, privacy, supplier security, and inter-organisational coordination as essential components of sustainable digital aviation. For Ercan International Airport, the proposed model comprises baseline assessment and risk mapping, reliable data governance, AI-supported operational decision-making, cyber-incident response, and continuous improvement based on performance indicators. The study does not claim that the proposed capabilities are currently deployed at Ercan. Instead, it recommends validating their feasibility through technical assessments, stakeholder consultations, and controlled pilot projects.

Keywords: *Aviation Cybersecurity, Artificial Intelligence, Aviation Safety, Aviation Security, Operational Resilience, Ercan International Airport.*

Hybrid Content Based Image Retrieval for Similarity Search in Industrial Fastener Technical Drawings

Ertuğrul Orman[1*], **Çağlar Uyulan**[2]

[1]Norm Fasteners, Research and Development Center, Izmir, Türkiye
[2]Izmir Katip Celebi University, Department of Mechanical Engineering, Izmir, Türkiye

*Corresponding author e-mail: ertugrul.orman@normfasteners.com

Abstract

Growing industrial drawing archives make it increasingly difficult for engineers to identify earlier designs that can support new product development. Manual searches depend heavily on personal experience, require considerable time, and may overlook geometrically related projects stored under different product names or dimensions. This study presents a hybrid content-based image retrieval approach for detecting similar products in two-dimensional fastener technical drawings. Norm Fasteners' technical drawing archive was screened. From this archive, 354 distinct projects qualifying as special products were selected, yielding 1,159 drawing images. The proposed method standardizes each drawing and evaluates rotated and mirrored representations to reduce sensitivity to orientation. Similarity is determined by combining complementary descriptions of outer-profile transitions, large-scale structural arrangement, global shape, line geometry, and local visual patterns. A two-stage project-level ranking procedure evaluates multiple views belonging to the same project and returns the most representative candidate. Dense hatching and repeated parallel lines, which may define important product features, are preserved; however, their influence is moderated when the broader geometry does not agree. The retrieved results were reviewed. Technically similar products were found among the leading candidates, while misleading matches associated with common sectional patterns were also observed. The proposed workflow gives engineers structured access to previous design knowledge, reduces the effort required for manual archive searches, and supports early assessment of whether a newly requested product resembles an existing project. The method provides a practical and interpretable basis for design-reuse and engineering decision-support systems in fastener development.

Keywords: *Content-Based Image Retrieval, Technical Drawings, Fasteners, Image Processing, Design Reuse*

Comparison of the Evolutionary Algorithms Performance for Multi-Response Surface Problem Optimization

Ebru Ilgat[1*], Özlem Türkşen[2]

[1]Ankara University, Graduate School of Natural and Applied Sciences, Department of Statistics, 06110, Ankara, Türkiye
[2]Ankara University, Faculty of Science, Department of Statistics, 06100, Ankara, Türkiye

*Corresponding author e-mail: ebruilgat1455@gmail.com

Abstract

The primary objective in the optimization of multi-response surface problems is to determine the optimal input conditions for a process by considering the responses simultaneously. In this study, the relationships between input variables and responses were expressed using polynomial response surface models to construct estimated response functions, which were then viewed as objective functions. The multi-response surface problem was formulated as a multi-objective optimization problem, with the aim of obtaining Pareto solutions using evolutionary algorithms. To this end, six distinct evolutionary algorithms found in the literature—SMS-EMOA, AGE-MOEA, AGE-MOEA-II, MOEA/D, C-TAEA, and RVEA—were selected. The Pareto solutions were evaluated using the GD+, IGD+, Spacing, Spread, R2, and HV metrics. The performance comparison was conducted based on the premise that, for the GD+, IGD+, Spacing, Spread, and R2 metrics, lower values indicate superior performance, whereas for the HV metric, higher values indicate better performance. In the application phase, the performance of the six evolutionary algorithms was compared in the context of optimizing multi-response surface problems using datasets from the fields of engineering and chemistry. It was found that SMS-EMOA yielded the best results across all six performance metrics for the engineering dataset, and across the IGD+, Spacing, Spread, and HV metrics for the chemistry dataset. Consequently, it was concluded that SMS-EMOA was the most successful algorithm among those tested, outperforming the others on all criteria for the engineering problem and on four out of the six criteria for the chemistry problem.

Keywords: *Response Surface Problem, Multi-Objective Optimization, Evolutionary Algorithms, Performance Metrics, Pareto Solution*

DNA Methylation-Based Age Prediction: An Evaluation of ElasticNet, XGBoost, and Deep Neural Networks

Aysenur Öztürk[1*], Yasemin Günter[2]

[1]Istanbul Kent University, Vocational School of Health Services, Medical Data Processing Technician Program, 34406, Istanbul, Türkiye

[2]Haliç University, Vocational School, Big Data Analytics Program, 34060, Istanbul, Türkiye

*Corresponding author e-mail: aysenur.ozturk@kent.edu.tr

Abstract

Epigenetic clocks are predictive models that estimate age from DNA methylation patterns at selected CpG sites. First-generation clocks such as Horvath and Hannum rely on penalized linear regression, whereas more recent approaches use machine learning to capture nonlinear CpG interactions and improve calibration. However, their performance remains context-dependent and varies across tissues, cohorts and platforms, and linear clocks are known to underestimate age in older individuals and to show regression-to-the-mean bias.

In this study, one linear (ElasticNet) and two nonlinear (XGBoost and deep neural networks, DNN) methods will be compared on the multi-center ComputAgeBench benchmark (65 independent studies, 10,410 samples, ages 0–99). To prevent information leakage, CpG selection will be performed within cross-validation folds, and generalization will be assessed using a leave-one-dataset-out strategy and external GEO cohorts. Performance will be evaluated with mean absolute error (MAE), root mean squared error (RMSE) and the coefficient of determination (R^2). In addition, signed bias stratified by age group will be analyzed to examine age underestimation and regression-to-the-mean bias, particularly in older age groups.

Keywords: *Epıgenetic Clocks, DNA Methylation, Machine Learning, Age-Stratıfied Bıas, Benchmarkıng*

Quality Management Supported by Artificial Intelligence in Quality 4.0: The Role of Statistical Approaches

Filiz Savaş[1*]

[1]Recep Tayyip Erdoğan University, Faculty of Economics and Administrative Sciences, Department of Business Administration, 53100, Rize, Türkiye

*Corresponding author e-mail: filiz.savas@erdogan.edu.tr

Abstract

Artificial intelligence (AI) is used in Quality 4.0 for defect detection and process monitoring. Its value for quality management depends on the data used to develop models and on how their performance is assessed during use. High predictive accuracy under test conditions does not, by itself, establish reliability under changing production conditions. This study examines how statistical approaches support the assessment of AI applications in quality management. It presents a conceptual review of relevant academic publications and technical guidance documents, focusing on three areas: data quality and representativeness, model performance and uncertainty, and monitoring after deployment. The discussion considers measurement error, missing data, and whether validation data adequately reflect the intended operating conditions. It also addresses the distinction between different types of prediction error and the need to report uncertainty alongside performance estimates. For monitoring during use, the study considers control charts for process indicators and model errors. Their suitability depends on the characteristics of the data, the assumptions of the chosen method, and the control limits used. The review links these issues to a practical distinction: evaluating a model before deployment and assessing its continued suitability are separate tasks. Statistical approaches support both. Their role in Quality 4.0 therefore extends from checking the data and evaluating predictions to identifying performance changes that warrant investigation.

Keywords: *Quality 4.0, Artificial Intelligence, Quality Management, Statistical Approaches, Statistical Process Control*

A Comparative Analysis of Artificial Intelligence Use by Enterprise Size in European Union Countries

Filiz Savaş[1*]

[1]Recep Tayyip Erdoğan University, Faculty of Economics and Administrative Sciences, Department of Business Administration, 53100, Rize, Türkiye

*Corresponding author e-mail: filiz.savas@erdogan.edu.tr

Abstract

This study compares artificial intelligence use across enterprise size groups in the 27 European Union countries using Eurostat data for 2025. The analysis examines the percentages of enterprises using at least one artificial intelligence technology in three groups: small (10 to 49 persons employed), medium (50 to 249), and large (250 or more). Differences among groups were assessed using the Friedman test, with observations matched by country. Effect size was measured using Kendall's W. Pairwise comparisons used Wilcoxon signed rank tests with Holm adjustment. Mean rates of artificial intelligence use, with equal weight assigned to each country, were 17.74% for small enterprises, 30.46% for medium enterprises, and 53.56% for large enterprises. The Friedman test indicated significant differences among groups ($\chi^2(2) = 54.00$; $p < 0.001$). Kendall's W was 1.00, reflecting the same ordering in every country: use was lowest among small enterprises, followed by medium and large enterprises. All pairwise Wilcoxon comparisons remained statistically significant after Holm adjustment ($p < 0.001$). The average gap across countries was 12.72 percentage points between medium and small enterprises, 23.10 between large and medium enterprises, and 35.81 between large and small enterprises. The gap between large and small enterprises was greatest in Slovenia (53.45 percentage points) and smallest in Romania (16.62 percentage points). These findings show that artificial intelligence use is higher in successively larger enterprise size groups in every country examined, while the gap between large and small enterprises varies across countries. The results provide descriptive evidence to inform policies supporting artificial intelligence use among small and medium enterprises while accounting for differences across national contexts.

Keywords: *Artificial Intelligence Use, Enterprise Size, European Union, Friedman Test, Digital Transformation*

Satellite-Based Analysis of Seasonal and Spatial Vegetation Patterns to Support Conservation Monitoring at Archaeological Sites: The Case of Şapinuva

Burcu Funda Ayberkın[1], Barış Ayberkın[1*]

[1]Istanbul Medipol University, School of Fine Arts Design and Architecture,34810, Istanbul Türkiye

*Corresponding author e-mail: baris.ayberkin@medipol.edu.tr

Abstract

Vegetation growth and its interaction with construction materials are environmental factors relevant to the conservation and maintenance of archaeological sites. Depending on plant species and surface characteristics, root growth may widen gaps in masonry, while dense vegetation may obscure walls and paved surfaces, hindering architectural documentation and condition assessment. This study examines how seasonal and spatial vegetation information derived from satellite data can support conservation monitoring, using Şapinuva, a Hittite settlement in Ortaköy, Çorum, Türkiye, as a case study.

Sentinel-2 surface reflectance imagery from 2019–2025 was processed in Google Earth Engine to generate monthly median Normalized Difference Vegetation Index (NDVI) composites. Comparisons included the Tepelerarası and Ağılönü excavation areas, selected zones dominated by open surfaces, and nearby agricultural parcels. Following cloud and shadow masking, valid observation counts and valid pixel coverage informed the selection of comparable records. Results were interpreted alongside architectural documentation, surface characteristics, and field information on maintenance practices.

Findings show seasonal differences between zones with contrasting surface compositions. In Ağılönü, NDVI averages for the wider analysis area and selected open zones diverge particularly during summer and autumn, whereas corresponding Tepelerarası patterns are more similar. Agricultural parcels provide context for interpreting changes within the excavation areas alongside surrounding seasonal patterns.

The study proposes a conservation monitoring approach linking satellite-derived vegetation patterns with architectural documentation and field inspection. This approach aims to support inspection scheduling, selection of areas for investigating maintenance needs, and the creation of systematic monitoring records. Vegetation effects on construction materials require on-site assessment. Application to other archaeological sites depends on the combined consideration of local vegetation, architectural characteristics, surface composition, and satellite image resolution.

Keywords: *Şapinuva, Archaeological Sites, Conservation Monitoring, Satellite Remore Sensing,Vegetation Dynamics*

A Bibliometric Analysis of Hepatitis D Virus

Nezihal Gökbulut[1], Ulaş Hürdoğanoğlu[1], Nuriye Sancar[1]

[1]Near East University, Lefkoşa, NRCR

*Corresponding author e-mail: nezihal.gokbulut@neu.edu.tr

Abstract

This study aimed to conduct a bibliometric analysis of publications related to Hepatitis D virus between the years 2018 and 2025. Publications were obtained from the Scopus database using filtering methods. The keywords \"hepatitis d,\" \"public health,\" and \"hepatitis\" were used to gather the publications. A total of 172 articles were identified, and data analysis was performed using the Vosviewer program. During the analysis, the study examined co-authorship networks among authors, organizations, and countries, the frequency of keywords, the density of citations, and bibliographic coupling of articles and authors. Based on the analysis, the study concluded that, given the significant public health implications of this disease, it is necessary to increase the visibility of publications in this field and to raise awareness among individuals. Accordingly, greater emphasis on expanding the body of publications and strengthening collaborative researches in this area is warranted.

Keywords: *Hepatitis D, public health, hepatitis*

A Computationally Efficient Estimator for Combating Multicollinearity in the Linear Regression Model

Esra Ertan[1*]-

[1]University of Istanbul, Faculty or Science, Department of Mathematics, 34134, İstanbul, Türkiye

*Corresponding author e-mail: eertan@istanbul.edu.tr

Abstract

In the presence of multicollinearity, the ordinary least squares (OLS) estimator of the regression coefficients becomes unstable and exhibits inflated variance, which has motivated the development of various biased estimation techniques. Among these, the ridge and Liu estimators have been widely studied, and more recently, a class of two-parameter biased estimators has been introduced with the aim of combining the desirable properties of these single-parameter estimators within a unified framework. Nevertheless, the simultaneous determination of the two biasing parameters poses considerable theoretical and practical difficulties, as optimal values are generally interdependent and lack closed-form expressions. To address this limitation, the present study proposes a new estimator that circumvents the problem of jointly selecting multiple biasing parameters while retaining the advantages of two-parameter estimators. The theoretical properties of the proposed estimator are examined, and necessary and sufficient conditions for its superiority over the OLS, ridge, Liu, and existing two-parameter estimators are derived under the matrix mean squared error (MMSE) criterion. Furthermore, estimators of the biasing parameter are suggested. An extensive Monte Carlo simulation study is conducted to evaluate the finite-sample performance of the proposed estimator under various degrees of multicollinearity, error variances, sample sizes, and numbers of explanatory variables, and the results corroborate the theoretical findings. Finally, the applicability of the proposed estimator is illustrated through a real data example, which demonstrates its superior performance relative to competing biased estimators in terms of the scalar mean squared error.

Keywords: *Biased regression,Multicollinearity, Ridge regression, Liu estimator*

Predicting Survival in Stroke Patients Using Ensemble Machine Learning

Samira Ahmadi[1*]

[1]Social Determinants of Health Research Center, Health and Metabolic Diseases Research Institute, Zanjan, Iran

*Corresponding author e-mail: samira.ahmadizn@gmail.com

Abstract

Background: *Stroke is a leading global cause of death and disability, making accurate survival prediction crucial for clinical decision-making. Recent advances in machine learning provide robust tools for survival analysis, effectively handling complex, non-linear, and high-dimensional data. This study evaluated and compared the predictive performance of Survival Support Vector Machine (SSVM), Random Survival Forest (RSF), and their ensemble models in ranking variable importance for predicting stroke patient survival.*

Methods: *We evaluated the performance of Survival Support Vector Machine (SSVM), Random Survival Forest (RSF), and their ensemble models in predicting survival outcomes in a cohort of 332 stroke patients from Imam Khomeini Hospital in Ardabil, who were monitored for up to 15 years following their stroke (2007-2022). Collected variables encompassed demographic, clinical, and lifestyle factors, including age, sex, Education, comorbidities (e.g., hypertension, heart disease), stroke type (ischemic or hemorrhagic), and smoking history. Alongside the Cox model, we implemented SSVM and RSF, then constructed four ensemble models (Voting, Stacking, Bagging, and Boosting) by integrating their outputs to capitalize on the strengths of both algorithms. Performance was assessed using accuracy, sensitivity, specificity, area under the curve (AUC), and survival function estimation error.*

Results: *Ensemble models outperformed the individual RSF, SSVM, and Cox models in predicting long-term mortality. The Voting ensemble achieved the highest performance, with an AUC of 0.855, overall accuracy of 0.833, 95% confidence interval of 0.747–0.962, sensitivity of 0.867, and specificity of 0.762. It demonstrated clear superiority over SSVM (AUC = 0.711), RSF (AUC = 0.828), and Cox (AUC = 0.768). Variable importance analysis in the optimal Voting model ranked age, hypertension, heart disease history, and hemorrhagic stroke type as the top predictors of long-term mortality*

Conclusion: *Integrating SSVM and RSF within ensemble frameworks enhances both predictive accuracy and interpretability in stroke survival analysis. The Voting ensemble, as the most accurate model, offers a valuable clinical tool for identifying high-risk patients and guiding targeted interventions.*

Keywords*: Survival Support Vector Machine, Random Survival Forest, Survival Support Vector Machine and Random Survival Forest Ensemble, Survival Analysis, Stroke.*

Data Management and Governance in Education Research: A Bibliometric Analysis of Web of Science Publications

İsmet Enes Demirel[1]

[1]Independent Researcher, Tokat, Türkiye

*Corresponding author e-mail: ismetenesdemirel@gmail.com

Abstract

The increasing use of digital technologies in education has intensified the generation and use of educational data, raising questions about data management, governance, quality, privacy, and ethical use. This study aims to map the development and thematic structure of research on data management and governance in education. Records were retrieved from the Web of Science Core Collection using search terms related to educational data, student data, learning analytics, data management, data governance, data quality, privacy, security, and ethics. The search yielded 780 records; limiting the document types to articles and review articles resulted in 519 records. A preliminary examination of author keywords showed recurring themes such as learning analytics, artificial intelligence, higher education, machine learning, privacy, and educational data mining. These initial patterns suggest that the literature connects educational data management with analytics, emerging technologies, and privacy-related concerns. The records are being screened for relevance before the final bibliometric and thematic analyses. The study is expected to provide an overview of publication trends and research themes and to identify directions for future work on responsible educational data management and governance.

Keywords: *Data, Data Management, Data Governance, Bibliometric Analysis, Web of Science*

From Transcriptomic Data to Biological Insights: Computational and Bioinformatics Approaches

Elif Burcu Tüzemen Bayyurt*

[1]Sivas Cumhuriyet University, Faculty of Medicine, Department of Medical Biology, Sivas, Türkiye

*Corresponding author e-mail: ebayyurt@yahoo.com.tr

Abstract

Transcriptomic profiling provides an important framework for investigating gene expression changes associated with physiological processes and disease mechanisms. Microarray and RNA sequencing (RNA-seq) technologies generate complementary types of high-dimensional expression data; however, detecting transcriptional differences does not, by itself, establish their biological significance. Transforming these data into meaningful insights requires an integrated approach that combines platform-appropriate computational processing, statistical evaluation, molecular interpretation, and experimental validation. In microarray-based studies, background correction, normalization, and probe annotation are critical analytical considerations, whereas RNA-seq workflows require sequence quality assessment, read alignment, and transcript or gene quantification. Subsequent differential expression analyses can identify candidate genes and transcripts, but their interpretation depends on study design, biological variability, multiple-testing control, and potential technical biases. Gene Ontology and pathway enrichment analyses, together with protein–protein interaction and regulatory network approaches, can place expression changes within broader biological contexts. In addition to protein-coding genes, the investigation of non-coding RNAs, including microRNAs (miRNAs), long non-coding RNAs (lncRNAs), and circular RNAs (circRNAs), provides further insights into the complex regulatory mechanisms underlying gene expression and disease pathogenesis. Furthermore, the integration of transcriptomic data with genomic information obtained through whole-genome sequencing (WGS) and whole-exome sequencing (WES) may provide a more comprehensive understanding of the relationships between genetic variation, gene expression, and disease-associated molecular mechanisms. Artificial intelligence (AI) and machine learning (ML) approaches may complement conventional bioinformatics methods by facilitating molecular classification and the identification of potential disease-associated biomarkers. From a medical biology perspective, this presentation will examine how computational and bioinformatics approaches can connect microarray- and RNA-seq-derived expression profiles with hypotheses about disease-related molecular mechanisms, while emphasizing the distinction between computational predictions and experimentally validated biological findings.

Keywords: *Transcriptomics, Gene Expression, Bioinformatics, Medical Biology*

Hotel Visibility and AI-Mediated Positioning on Generative AI Platforms: A Comparative Analysis of Istanbul Hotel Recommendations

Arda Şap[1]*, Fatih Öztunç[1], Ali Yavuz[1], Hakan Tekedere[1]

[1]Gazi University, Faculty of Applied Sciences, Department of Management Information Systems, Ankara, Türkiye

Corresponding author e-mail: ardas@gazi.edu.tr

Abstract

Generative artificial intelligence (AI) platforms increasingly function as information intermediaries that determine which businesses consumers encounter, how alternatives are ranked, and which attributes are emphasized. This study investigates hotel visibility and AI-mediated positioning across three consumer-facing platforms—ChatGPT, Gemini, and Perplexity—using Istanbul as a common destination. A repeated platform audit covered six travel conditions (general, business, couple, family, friends, and solo), three semantically equivalent prompt variants per condition, and five independent repetitions, yielding 270 responses. After property verification and within-response deduplication, 1,335 hotel recommendations representing 168 physical properties were analyzed. Visibility and first-rank placement were compared using matched-block Cochran's Q tests; recommendation-set similarity and robustness to prompt wording were assessed using Jaccard-based measures; and recommendation rationales were classified into 16 attribute categories and compared using Friedman tests. Results reveal pronounced platform-dependent exposure: 70.2% of identified hotels appeared on only one platform, while only 19 hotels were recommended by all three. Repeated responses exhibited limited recommendation-set stability (mean Jaccard similarity = 0.313). Semantically equivalent prompts also produced differing overlap levels across platforms, with mean weighted Jaccard scores of 0.547 for Gemini, 0.494 for ChatGPT, and 0.268 for Perplexity. The prevalence of all 16 rationale categories varied significantly across travel contexts after multiple-testing correction, demonstrating that recommendation context relates not only to which hotels become visible but also to how their attributes are presented. The findings conceptualize AI-mediated hotel positioning as a probabilistic and context-dependent outcome and suggest that business visibility should be evaluated using multiple platforms, prompts, and repetitions rather than a single deterministic ranking.

Keywords: *Generative Artificial Intelligence, Hotel Visibility, Recommender Systems, Travel Context, AI-Mediated Positioning*

From the Design Matrix (X) to the Jacobian Matrix (J): How Nonlinearity Reshapes Collinearity

Ali Erkoç[1*]

[1]Mimar Sinan Fine Arts University, Faculty of Science and Letters / Statistics Department, 34380, Istanbul, Turkey

*Corresponding author e-mail: ali.erkoc@msgsu.edu.tr

Abstract

In linear regression, multicollinearity is commonly investigated through the original design matrix, which coincides with the Jacobian matrix. However, in nonlinear regression, the Jacobian depends on the functional form of the model and the unknown parameter values. Consequently, multicollinearity among the original explanatory variables may not be preserved in the Jacobian, while strong collinearity among Jacobian columns may arise even when the original variables are not highly correlated.

This study investigates the functional relationship between the design matrix (X) and the Jacobian matrix (J), focusing on the conditions under which collinearity is preserved, weakened, or intensified. Theoretical relationships involving matrix rank, singular values, and conditioning are examined for selected classes of nonlinear regression models. Mathematical counterexamples demonstrate how the functional structure of a model can alter the collinearity observed in the original data.

Furthermore, the effects of parameter values and iterative estimation on Jacobian collinearity are investigated. Particular attention is given to the distinction between changes directly induced by the explanatory variables and those arising from parameter re-estimation. The theoretical findings are supported by numerical illustrations involving four nonlinear regression models.

The results emphasize that collinearity in nonlinear regression should be diagnosed primarily through the Jacobian matrix, while the original design matrix remains important for understanding the sources and functional transformation of collinearity. This distinction provides a basis for more informative diagnostics and a better understanding of parameter estimation difficulties in nonlinear regression.

Keywords: *Nonlinear Regression, Multicollinearity, Design Matrix, Jacobian Matrix, Functional Transformation*

Evaluation of the Study Model of Using Excel and ChatGPT Together According to Question Categories in YÖKDİL

Oğuz Hamdioğlu[1*]

[1] [Ctprus International University], [Medicine], [Family Medicine], [99040], [Nicosia], [TRNC]

*Corresponding author e-mail: ohamdioglu@ciu.edu.tr

Abstract

The preparation process for the YÖKDİL exam organized by ÖSYM requires studying according to the question structure in the subclasses of science, social and health sciences, sufficient vocabulary, and regular work in the perspective of reading comprehension and analysis. The focused study we present proposes a study method that combines Excel in organizing study records and ChatGPT in evaluating questions in detail. Before the study, we obtained publications on the subject by conducting a literature review. Some of these were publications on the use of generative artificial intelligence in language learning. In addition, past YÖKDİL questions were used to determine the question categories. The previous sources have been evaluated thematically in order to form the basis of our study, especially in terms of feedback status, learners' participation, independent exercise perspective, and validation of the explanations produced. The protocol we recommend; It is the use of excell and ChatGPT infrastructure together in the study of vocabulary, grammar, translation, completing the gap in the paragraph, sentence completion, paragraph completion, finding the irrelevant sentence in the flow and reading comprehension questions in the article, which are the classic question types of YÖKDİL, by addressing them specifically according to each category. In Excel, the question type, subcategory, accuracy status of the answers, relevant artificial intelligence link, true & false percentage statuses and study dates are recorded. ChatGPT, on the other hand, is asked to check the accuracy of our answers and to solve the questions with explanations, to examine distractors and wrong answers, to solve grammar, to specify synonyms, to express sentences in two languages, to divide complex sentences into 3 meaningful parts, and to examine the sentence evidence in the reading passages together with the question options. The integrated working model we have created with both infrastructures proceeds with the control of the answers with ChatGPT, the examination of detailed explanations, advanced learning studies according to the question type, targeted repetition and independent solution of new questions with similar features produced by ChatGPT, after first solving the questions we have saved in excel ourselves. Among the unique features of our study, we can show that it is exam-oriented, the analysis of sentences by breaking them into meaningful parts, and the systematic analysis of errors in incorrect answers. Although previous studies presented the directed feedback of the participants and the use of artificial intelligence in effective learning as a justification, they did not address the subject in the form of the Excel-ChatGPT combined study that we proposed in our study and focused on the YÖKDİL goal. Short-term results were revealed by descriptive methods. The contribution of the study is to explain how to study different question categories on the application example; the effects on correct response rate, learning retention and exam performance were determined by more comprehensive empirical research

Keywords: *CHATGPT, EXCEL, YÖKDİL, LANGUAGE LEARNING*

Computational Performance of Python Interpreters in Missing Data Imputation: A Comparison of CPython and PyPy

Onur Ülker[1*], Nurbanu Bursa[1]

[1]Hacettepe University, Faculty of Science, Department of Statistics, 06800, Ankara, Türkiye

*Corresponding author e-mail: onurulker25@hacettepe.edu.tr

Abstract

Missing-data imputation is a fundamental preprocessing step in large-scale data science applications, often preceding statistical analysis and machine learning. As datasets grow, the execution time and memory requirements of this operation can become critical constraints on analytical workflows. Performance comparisons within the Python ecosystem frequently emphasize libraries and algorithms, whereas the computational behavior of the same imputation workload across different Python interpreters has received comparatively less attention. In particular, whether the expected speed advantage of PyPy's just-in-time (JIT) compilation persists when using Pandas, which relies heavily on C-based components, remains an important practical question.

This study aims to compare the computational performance of CPython and PyPy during Pandas-based missing-data imputation, focusing on execution time, additional peak memory usage, and runtime stability as dataset size increases. Four synthetic datasets containing 100,000, 500,000, 1,000,000, and 5,000,000 rows were generated. In each dataset, 5% of the entries were made missing under a Missing Completely at Random (MCAR) mechanism. Missing values in numerical variables were replaced with the median, while missing categorical values were assigned a fixed "missing" label. The two interpreters were evaluated in separate Docker-based environments, with experimental tasks coordinated using Celery and Redis. To reduce initialization effects and accommodate JIT warm-up, five warm-up runs preceded 30 measured repetitions for each scenario. Performance was assessed using mean execution time, additional peak Resident Set Size (RSS) memory consumption, and the coefficient of variation of execution times.

Across all investigated dataset sizes, CPython completed the imputation workload faster than PyPy. For one million rows, the average execution time was 1.037 seconds for CPython compared with 2.597 seconds for PyPy; CPython required 5.836 seconds at five million rows. Regarding measured additional peak memory, CPython showed an increase of 0.0 MB at dataset sizes of 500,000 rows and above, whereas PyPy exhibited a 406.7 MB increase at five million rows. The coefficients of variation indicated relatively stable execution times for both interpreters across repeated runs. These findings suggest that differences in interpreter–library interaction can outweigh the potential benefits of JIT compilation for the specific Pandas-based workload examined.

The study contributes a controlled, interpreter-level comparison of missing-data imputation performance and highlights that interpreter selection can influence the computational cost of routine data preparation. Importantly, it evaluates computational efficiency rather than imputation accuracy. Its conclusions are therefore limited to the tested imputation strategy, missingness proportion, dataset scales, and experimental environment, motivating further evaluation under other

missing-data mechanisms and imputation methods.

Keywords: *Computational performance, CPython, Missing data imputation, PyPy, Python interpreters.*

Balancing Classes While Distorting Probabilities: Effects of Class Imbalance Correction on Probability Calibration (524)

Bugra Varol[1*], Fulden Cantas Turkis[1]

[1]Muğla Sıtkı Koçman University, Faculty of Medicine, Division of Biostatistics, 48000, Muğla, Türkiye

*Corresponding author e-mail: bugravarol@mu.edu.tr

Abstract

Class imbalance is frequently addressed by resampling, yet gains in threshold-based classification may come at the expense of probability calibration. We evaluated whether random undersampling (RUS), random oversampling (ROS), and the Synthetic Minority Over-sampling Technique for Nominal and Continuous features (SMOTE-NC) improve discrimination or primarily shift the operating point of logistic regression prediction models. Two publicly available datasets with different event rates were analyzed: the NHANES I Epidemiologic Follow-up Study (NHEFS), with death during follow-up as the outcome, and the National Health and Nutrition Examination Survey (NHANES), with self-reported diabetes status as the outcome. We used 500 repeated stratified 70/30 train-test splits; resampling and all preprocessing were fitted on the training data only, and all methods were compared on the same held-out individuals. Performance was assessed using the area under the receiver operating characteristic curve (AUC), observed-to-expected ratio (O/E), calibration intercept and slope, Brier score, sensitivity, and specificity. Balanced models at the 0.50 threshold were compared with the unbalanced model using the training-set event rate as the threshold, and prior-probability correction was examined. None of the balancing strategies produced a practically meaningful gain in discrimination. All three shifted predicted probabilities upward, resulting in marked overprediction and poorer overall probabilistic accuracy. Lowering the threshold of the unbalanced model to the training-set event rate closely reproduced the sensitivity-specificity operating points of the balanced models. Prior correction restored mean calibration to near ideal; being an intercept-only adjustment, it left calibration slopes unchanged, and these remained reduced mainly after SMOTE-NC. When the objective is probability estimation, class balancing should not be used routinely solely to compensate for low sensitivity at the conventional 0.50 threshold. Separating probability estimation from decision-threshold selection can preserve calibration while achieving similar classification behavior in logistic regression models. Calibration should therefore be reported alongside discrimination whenever resampling is used in clinical prediction.

Keywords: *Class Imbalance, Calibration, Resampling, SMOTE-NC, Clinical Prediction Models*

Overtime and Employee Attrition: A Causal Machine Learning Analysis Based on the Job Demands-Resources and Conservation of Resources Theories

Abdullah Doğan[1], **Burcu Kartal**[1*]

[1]Recep Tayyip Erdogan University, Faculty of Economics and Administrative Sciences, 53100, Rize, Turkiye

*Corresponding author e-mail: burcu.kartal@erdogan.edu.tr

Abstract

Predictive machine learning models can forecast employee attrition with high accuracy, but they cannot show what would happen if a human resource practice were changed. This study aims to estimate the causal effect of overtime on employee attrition and to examine whether this effect depends on the resources employees hold. Based on the job demands-resources theory, overtime is treated as a job demand; based on the conservation of resources theory, career resources (income and work experience), stock option level, relationship satisfaction and distance from home are tested as moderators. The study is important because overtime is one of the few attrition factors that managers can directly control, and knowing for whom it is most harmful allows targeted interventions. Its originality lies in testing theory-based moderation hypotheses with a doubly robust causal machine learning approach instead of conventional correlational models. The IBM HR Analytics Employee Attrition and Performance dataset (N = 1,470) was used. The average effect of overtime was estimated with augmented inverse probability weighting, in which gradient boosting models adjusted for age, gender, marital status, department, education and business travel. Moderating effects were tested by regressing employee-level effect estimates on each moderator, with Holm correction for multiple testing. The analysis was validated with simulated data of known effect, placebo tests and an independent software implementation. Overtime increased the probability of attrition by 19.3 percentage points (95% CI: 14.2-24.3), nearly tripling the risk of leaving. The effect weakened as career resources increased (-10.3 percentage points per standard deviation) and strengthened as distance from home increased (+6.5 percentage points), supporting H1, H2 and H5. The moderating roles of stock option level (H3) and relationship satisfaction (H4) were not supported. The results were consistent across alternative algorithms and linear probability models with interaction terms. The findings suggest that policies aimed at reducing overtime should give priority to employees with limited career resources and long commutes.

Keywords: *Causal Machine Learning, Overtime, Employee Attrition, Job Demands-Resources Theory, Conservation of Resources Theory*

Segmentation of Skin Lesions Using EfficientNet-Attention U-Net Architecture: An Evaluation on the ISIC 2016 Dataset

Kadir Can Burcak[1]

[1]Kırşehir Ahi Evran University, Kaman School of Applied Sciences, Department of Information Systems and Technologies, 40100, Kırşehir, Türkiye

*Corresponding author e-mail: kcburcak@ahievran.edu.tr

Abstract

Skin lesions pose a significant challenge to human health. Accurate segmentation is crucial for the diagnosis of these lesions. However, factors such as irregular boundaries, complex visual variations, and low contrast often hinder the precise extraction of lesion areas in computer-aided medical image segmentation. In this study, a hybrid deep learning model is proposed to achieve more accurate identification of skin lesion regions. The model employs an EfficientNet architecture as an encoder to extract distinctive features from images, while the decoder utilizes a U-Net architecture equipped with an attention mechanism to learn contextual features associated with the lesion areas. The model was trained on the publicly available ISIC 2016 dataset using 5-fold cross-validation. Experimental results were evaluated based on performance metrics including the Dice score, Intersection over Union (IoU), Sensitivity, Specificity, and Accuracy. In experiments conducted on 900 actual lesion images and their corresponding masks from the ISIC 2016 dataset, the proposed hybrid model achieved a Dice score of 92.24%, an IoU of 86.17%, a Sensitivity of 94.64%, a Specificity of 96.76%, and an Accuracy of 96.40%.

Keywords: *Skin Lesions, Medical Image Segmentation, Deep Learning, ISIC 2016, Attention Mechanism*

Drought and Small Ruminant Population Dynamics Across Climatic Gradients: Evidence from Panel Regression and Double Machine Learning

Yalçın Tahtalı[1], Lütfi Bayyurt[1]

[1]Tokat Gaziosmanpaşa University, Faculty of Agriculture, Department of Animal Science, Tokat, Türkiye

*Corresponding author e-mail: lutfi.bayyurt@gop.edu.tr

Abstract

Drought is a major source of climatic variability in livestock production systems, yet its association with long-term changes in small-ruminant populations may depend on prevailing climatic conditions. This study investigated the relationships between drought intensity and annual sheep and goat population growth across countries during 1991–2024, with particular emphasis on long-term precipitation regimes. Country-level livestock and land-use data were combined with ERA5-Land precipitation records. Two-way fixed-effects (TWFE) panel regressions were complemented by double machine learning (DML) approaches using Lasso and Random Forest algorithms. The analyses included 5642 sheep observations from 176 countries and 5689 goat observations from 178 countries. In the TWFE models, the interaction between lagged drought intensity and long-term precipitation was positive and statistically significant for goats ($\beta = 0.00533$, $p = 0.0006$), but not for sheep ($\beta = 0.00077$, $p = 0.6116$). For goats, the conditional association with lagged drought was negative in relatively dry climates ($\beta = -0.00459$, $p = 0.0465$) and positive in relatively wet climates ($\beta = 0.00607$, $p = 0.0285$). The goat interaction remained positive and statistically significant across alternative TWFE specifications, DML estimators, and sensitivity analyses. However, the direct between-species difference was not statistically significant at the 5% level ($p = 0.0729$). These findings indicate that associations between drought and goat population dynamics vary with long-term precipitation conditions. They highlight the importance of climatic context when evaluating livestock population responses while avoiding causal interpretations of observational evidence.

Keywords: *Drought İntensity; Sheep Population; Goat Population; Precipitation Variability; Panel Data; Double Machine Learning; Climate–Livestock İnteractions*

An Evidence-Based Verification Framework for Microbial Circular Polarization in Reflected Exoplanet Light

Özge Kahraman Ilıkkan[1], Özkan Günalp[2*], Murat Gezer[3], Önder Bakır[4,5]

[1]Baskent University, Department of Food Processing, Ankara, Türkiye
[2]Ege University, Institute of Health Sciences, Medical Biology, İzmir, Türkiye
[3]Medeniyet University, Faculty of Engineering and Natural Sciences Department of Computer Engineering, İstanbul, Türkiye
[4]Ege University, Department of Astronomy and Space Sciences, İzmir, Türkiye
[5]Ege University, Institute of Nuclear Sciences, İzmir, Türkiye

*Corresponding author e-mail: ozkn.gunalp@gmail.com

Abstract

Laboratory spectropolarimetry shows that photosynthetic microbes can imprint wavelength-dependent circular polarization on reflected light, motivating Stokes V as an agnostic exoplanet surface biosignature. A laboratory degree of circular polarization, however, is not a planetary observable: the disk-integrated signal depends on illumination and viewing geometry, atmospheric and cloud transfer, instrument response, and abiotic sources of V. For data-driven detection studies the problem is also one of evidence quality, because synthetic training sets must be built from sparse, heterogeneous and partly figure-derived sources. We present an evidence-based verification framework that specifies when such a measurement can support a synthetic machine-learning detection experiment. Four sequential gates address source provenance and reuse rights, compatibility of biological intensity and polarization evidence, independent validation of vector radiative transfer, and leakage-resistant evaluation with physical-scenario-grouped data splits; unpaired biological inputs are confined to a labeled hypothetical branch for sensitivity analysis. The framework separates five uncertainty layers, from digitization error to observation noise. A formal observation pathway sums local Stokes parameters over the disk and applies the instrument response before forming V/I; an algebraic two-patch example shows that area-averaging local ratios can misstate the disk-integrated value by a factor of about four when clouds are present. A provisional pixel reading of published Chloroflexus aurantiacus *spectra places the reflected minimum near 748–755 nm, consistent between figure panels but redward of the approximately 740 nm stated in the text, illustrating why figure-derived templates must not be treated as precise measurements. A prospective falsification protocol compares intensity, circular polarization and their combination against a matched-filter baseline on held-out scenarios at a prespecified false-positive rate. No planetary simulation or classifier result is claimed; the contribution is a transparent set of conditions under which future performance claims become interpretable.*

Keywords: *Circular Polarization, Exoplanet Biosignatures, Radiative Transfer, Uncertainty, Machine Learning*

Electronic Nose-Based Halal Screening: Detecting Pork Adulteration in Beef via Explainable Machine Learning

Hamza Halit Çatalca[1*], Fatih Ekinci[2], Mehmet Serdar Güzel[3]

[1] PTT Bilgi Teknolojileri A.Ş., Technology and R&D, 06510, Ankara, Türkiye
[2] Ankara University, Institute of Artificial Intelligence, 06560, Ankara, Türkiye
[3] Ankara University, Faculty of Engineering, Department of Computer Engineering, 06100, Ankara, Türkiye

*Corresponding author e-mail: hamza.catalca@ptteknoloji.com.tr

Abstract

Pork adulteration in beef is a major concern for halal authenticity and food fraud. Confirmatory laboratory methods are accurate but costly and time-consuming and require trained personnel, creating a need for rapid, low-cost screening tools such as electronic noses. In this study, ten classification models were compared for detecting pork adulteration using a publicly available dataset collected with an electronic nose comprising eight metal oxide gas sensors. The dataset consists of 420 measurement cycles from seven mixture ratios. The models were selected from the classical machine learning (Random Forest, XGBoost, SVM, kNN), deep learning for tabular data (MLP, KAN) and time series classification (MiniRocket, FCN, LSTM-FCN, CNN-BiLSTM) families. Blocked cross-validation at the measurement-cycle level was applied to prevent readings from the same cycle from falling into both the training and test sets, and temperature and humidity were excluded from the input. In the seven-class adulteration level task, MiniRocket achieved the highest balanced accuracy of 0.877 and was significantly better than seven of the other nine models. In the binary task separating pork-containing and pork-free samples, balanced accuracy ranged from 0.903 to 0.973. Sensitivity for pork-containing samples ranged from 0.982 to 1.000 and reached 1.000 in five models, while specificity ranged from 0.806 to 0.950. Random row splitting produced balanced accuracies of 0.988-1.000 for the same models, indicating that a substantial part of the high performance reported in the literature stems from the splitting strategy. Averaged over all ten models, SHAP analysis highlighted the MQ135 and MQ4 sensors in the seven-class task and the MQ136 and MQ9 sensors in the binary task, but sensor contributions did not show a monotonic relationship with pork ratio. The findings show that electronic nose-based halal screening is promising at the binary level, whereas adulteration level estimation needs to be validated with independent measurement sessions.

Keywords: *Electronic Nose, Halal Authentication, Machine Learning, Explainable Artificial Intelligence*

Linux for Data Science

Mehmet Ali Celikbag[1]*

[1]Ankara University, Faculty of Education, Computer Education and Instructional Technologies, 06590, Ankara, Türkiye

*Corresponding author e-mail: celikbag@gmail.com

Abstract

As e-learning has expanded, tracking and analyzing student engagement in distance education environments has become increasingly critical. While various tools exist to process this data, proprietary learning analytics solutions are often expensive, rigid, and raise significant data privacy concerns. As a result, researchers and educational institutions need robust open-source alternatives. Linux distributions provide an ideal foundation for self-hosted Learning Management Systems (LMS) and Learning Analytics (LA) pipelines. Renowned for their stability, security, and efficient resource management, open-source operating systems eliminate heavy licensing fees. The Linux command-line environment also provides powerful built-in utilities—such as cat for quickly aggregating server logs, grep for isolating specific user events, sed for efficient text substitution and data cleaning, and awk for structured data extraction. These tools are indispensable for parsing the massive, often unwieldy volumes of raw transactional data LMS platforms generate. Crucially, by running these heavy preprocessing steps directly on institutional servers, researchers can quickly distill massive log files into small, manageable datasets. This localized server-side approach not only ensures complete data sovereignty and privacy during anonymization, but also yields lightweight, structured data that researchers can easily manipulate for subsequent analysis on their personal laptops or desktops.

Keywords: *Linux for Data Science, Data Science in Education, Linux for Learning Analytics, Analytics in Online Learning*

Industrial Data Integration Using a Database-Driven Dynamic OPC UA Gateway Architecture

Gökhan Yenikaya[1]

[1]Bursa Uludağ Üniversitesi, Elektrik-Elektronik Mühendisliği Bölümü, Bursa / Türkiye

Abstract

Integrating programmable logic controllers (PLCs), industrial controllers, and measurement systems from different manufacturers, models, and generations into a common data infrastructure remains a significant engineering challenge in industrial production facilities. OPC Unified Architecture (OPC UA) enables standardized, platform-independent data exchange across automation systems. However, constructing OPC UA address spaces using static definitions or hard-coded configurations increases configuration and maintenance effort, particularly in plants with numerous machines and tags. This study presents a database-driven dynamic gateway architecture in which machine, device, and tag definitions are stored in a relational database and used to dynamically construct the OPC UA address space during application startup. Configuration is decoupled from application code, while runtime process values are managed through a shared in-memory data model that separates data acquisition from OPC UA presentation. The architecture enables reuse across different plants and machine configurations through database modifications without changing application code, thereby reducing administrative overhead. It was evaluated at five configuration sizes ranging from 100 to 5,000 tags, with 10 repeated trials per scenario. A controlled data provider was used in runtime update tests to isolate field communication delays. Database configuration loading remained between 176.61 and 185.52 ms, while dynamic OPC UA node creation increased from 1.47 to 35.58 ms. Total startup time ranged from approximately 0.92 to 1.01 s. Managed memory usage exhibited an incremental increase of approximately 1.56 kB per tag. Server-side updates of 5,000 OPC UA variable values took 0.503 ms on average, with a P95 latency of 0.762 ms. The results demonstrate a manageable, reusable OPC UA integration architecture with low, predictable runtime overhead within the evaluated configuration range.

Keywords: *Real-Time Data Acquisition, Industrial Data Integration, Industrial IoT (IIoT), OPC UA Gateway, Database-Driven Architecture.*

FULL-TEXT PRESENTATIONS

Digital Monitoring of Biosensor-Integrated Wastewater Treatment and Bioremediation

Ummahan Temel[1], Fatma Altıntaş[2*], Batuhan Selamoğlu[3], Ferdi Güler [4]

[1,2]Istanbul Arel University, Vocational School, Department of Biomedical Device Technology, 34537, Istanbul, Türkiye.
[3]Istanbul Arel University, Vocational School, Department of Electronics Technology, 34537, Istanbul, Türkiye.
[4]Istanbul Arel University, Vocational School, Department of Computer Programming, 34537, Istanbul, Türkiye.

*Corresponding author e-mail: fatmaaltintas@arel.edu.tr

Abstract

Water resources are under increasing pressure due to industrial discharges, urban wastewater, agricultural runoff, heavy metals, petroleum derivatives, nutrient overload, synthetic dyes, pharmaceutical residues, and emerging micropollutants. This situation necessitates addressing water security not only through the capacity of traditional treatment infrastructure but also through the early detection of pollutants, real-time monitoring of biological processes, and data-driven management of bioremediation performance. The aim of this study is to propose an integrated, modular, and scalable model for the digital monitoring of wastewater treatment and bioremediation processes incorporating biosensors. The developed model aims to demonstrate that biological treatment processes can be monitored not only through final effluent parameters but also through their in-line behavior. In this paper, we present a thematic content analysis of 28 scientific publications made within the period from 2020 to 2025. In this regard, the publications are categorized as follows: AI-augmented wastewater treatment, machine learning, environmental biosensors, microalgae-bacteria consortiums, Internet-of-Things (IoT) monitoring, remote sensing, digital twin, soft sensors, micropollutants, xenobiotic bioremediation, and environmental governance. The research proposes the following six-layer conceptual framework: (1) the sensor layer; (2) edge-IoT data transfer layer; (3) artificial intelligence prediction layer; (4) explainable decision-making support layer; (5) bioremediation intervention layer; (6) governance reporting layer. It helps to track different kinds of pollutants and to make decisions regarding the selection of microorganisms, bioreactor management, early warning, and resource recovery. The findings indicate that artificial intelligence and machine learning methods hold strong potential for predicting the removal performance of BOD, COD, TN, TP, turbidity, heavy metals, and organic pollutants. The ability of biosensors to generate rapid and field-adaptable data, and the capacity of digital twins to provide process simulation, fault detection, and decision support, are key advantages. However, data quality, model interpretability, sensor maintenance requirements, lack of field-scale validation, and insufficient regional microbial databases are the main limitations.

Keywords: *Artificial intelligence, machine learning, digital twin, IoT, water security*

INTRODUCTION

The resources of water experience increasing threats from industries, domestic sewage, agriculture, heavy metals, oil substances, nutrients, artificial dye stuffs, drugs, and other micropollutants. Wastewater treatment facilities make a great contribution into eliminating organic pollutants and suspended particles from the water bodies; nevertheless, they are not always sufficient for the purposes of purification in case of non-steady flows, unexpected pollutant loads, and pollutant complexity. In particular, there is a problem associated with removal of such pollutants like heavy metals, microplastics, drugs, xenobiotics which can cause environmental problems even in small amounts that require more accurate control of the treatment processes.

Wastewater treatment is the dynamic system where chemical, biological, hydraulic, and technological processes take place. It means that water safety is determined not only by the quality of effluents but also by identification of changes in the treatment process as early as possible. While traditional laboratory analyses provide accurate results, they have one serious drawback – time delays caused by the collection of samples, their delivery, and analysis.

In recent years, artificial intelligence and machine learning methods have been used to predict wastewater treatment performance, optimize processes, determine contaminant removal efficiency, control membrane fouling, reduce energy consumption, and detect failures early. It has been reported that artificial neural networks, fuzzy logic, random forest, LSTM, XGBoost, and hybrid models yield effective results in predicting BOD, COD, total nitrogen, total phosphorus, turbidity, suspended solids, and heavy metal concentrations (Singh et al., 2023; Malviya and Jaspal, 2021; Zhang et al., 2023). The key contribution of these methods is not only that they produce highly accurate predictions but also that they shorten decision-making time by supplementing delayed laboratory data with in-process data.

Applications of artificial intelligence in wastewater treatment generally focus on three areas: predicting contaminant removal performance, real-time monitoring of water quality parameters, and early detection of equipment or process failures. Nagpal et al. (2024) evaluate AI applications in the context of removing organic and inorganic pollutants, monitoring parameters such as pH, COD, BOD, turbidity, total dissolved solids, conductivity, and identifying process errors. Capodaglio and Callegari (2025), on the other hand, note that AI-supported systems must be considered in conjunction with sustainability goals such as energy consumption, greenhouse gas emissions, and resource recovery.

One of the key data-generation components of this digital transformation is biosensors. Although traditional chromatographic and spectrophotometric methods offer high sensitivity, they are costly, laboratory-dependent, and only suitable for continuous monitoring to a limited extent. Biosensors and microfluidic platforms, on the other hand, enable the detection of pollutants in a faster, lower-cost, and field-adaptable manner. Microfluidic systems integrated with artificial intelligence can be used to classify and assess pollutants in water and other environmental media (Pouyanfar et al., 2022). Bioelectrochemical sensors also contribute to monitoring carbon variability and shock loads in biological nutrient removal processes (Emaminejad et al., 2023). Thus, biosensors are not merely measurement tools but emerge as early-warning components that make the behavior of biological processes visible.

IoT-based data transmission, edge computing, soft sensors, remote sensing, and digital twin technologies enable the real-time processing of data obtained from sensors and its conversion into decision support. Digital twins can create a virtual representation of a physical treatment or bioremediation system, allowing for the testing of different operational scenarios, the identification of process deviations, and the prediction of potential failures. There is a growing trend in the literature toward the integrated use of artificial intelligence, IoT, remote sensing, and unmanned monitoring platforms (Zou et al., 2025; Divine et al., 2024; Miller et al., 2025).

However, a significant portion of existing studies evaluate components such as artificial intelligence, biosensors, IoT, or bioremediation independently of one another. Application-specific challenges—such as sensor maintenance, biological contamination, data quality, model interpretability, model drift, field-scale validation, local microbial databases, and regulatory reporting—are often addressed only as secondary considerations. However, a system that is applicable under field conditions must rely not only on algorithmic accuracy but also on sensor robustness, data continuity, knowledge of biological processes, explainable decision support, and institutional capacity.

This requirement necessitates that the components of data generation, data transmission, forecasting, decision support, biological intervention, and governance in wastewater treatment and bioremediation processes be addressed within a common digital framework. Developing an adaptable approach for municipal wastewater, industrial discharges, graywater systems, surface waters, and bioremediation sites is crucial for monitoring the in-process behavior of different pollutant groups, early detection of deviations in treatment performance, and making data-driven intervention decisions. This framework is based on evaluating biological treatment performance not only through final effluent parameters but also through real-time process data and changes within the system.

LITERATURE

In the current review, the literature is analyzed based on the basis of a certain range of important concepts including artificial intelligence, machine learning, biosensors, microfluidics, soft sensors, IoT, remote sensing, digital twins, bioremediation, microalgae-bacterial consortia, and micropollutants. The scope covers scientific articles published in the period from 2020 to 2025 and relevant to wastewater treatment, environmental monitoring, water quality models, bioremediation, and smart treatment systems. Publications that are duplicates, irrelevant to the research, or merely provide a general overview of certain technologies are not included in the analysis.

The literature suggests that the technologies of artificial intelligence and machine learning are widely used in the context of wastewater treatment regarding BOD, COD, nutrient removal, turbidity, heavy metals concentration, membrane fouling, process optimization, and fault detection. At the same time, biosensors, microfluidics, and soft sensors offer a great potential to detect environmental contaminants, perform monitoring, and predict certain laboratory parameters. IoT, remote sensing, and digital twins can be considered essential tools for transferring, processing, simulating, and using this information as decision-making support.

In the field of biological treatment and bioremediation, microalgae-bacteria consortia have attracted attention for their potential in nutrient removal, biomass production, the degradation of organic

pollutants, and resource recovery. In recent years, it has been demonstrated that artificial intelligence and machine learning methods can be used in the selection of microbial consortia, the optimization of culture conditions, the removal of xenobiotics, and the prediction of bioremediation performance. In addition, contaminant classes such as microplastics, pharmaceutical residues, synthetic dyes, surfactants, heavy metals, and petroleum derivatives pose significant monitoring and removal challenges for conventional treatment systems. Therefore, the selection of contaminant-specific sensors, prediction models, and bioremediation strategies must be evaluated collectively.

Table 1. Thematic Classification of Studies in the Literature and Their Relationship to This Study

Theme	Sources	Major Contribution to the Literature	Its Use in This Study
Artificial Intelligence and Machine Learning-Based Wastewater Treatment	Ye et al. (2020); Malviya and Jaspal (2021); Zhang et al. (2023); Singh et al. (2023); Nagpal et al. (2024); Alprol et al. (2024)	It demonstrates that artificial intelligence and machine learning models can be used in the areas of BOD, COD, nutrients, turbidity, heavy metal concentrations, membrane fouling, process optimization, water quality prediction, and fault detection.	In the proposed model, the artificial intelligence layer underpins the functions of prediction, monitoring, early warning, optimization, an decision support.
Biosensors, microfluidic systems, and soft sensors	Pouyanfar et al. (2022); Emaminejad et al. (2023); Reneeth et al. (2025)	The research focuses on the rapid and accurate detection of environmental pollutants, the prediction of shock loading using bioelectrochemical sensors, and the estimation of easily measurable parameters such as TSS, BOD, COD, TN, and TP.	It supports the use of sensor data not only for measurement purposes but also for virtual measurement, early warning, and decision support.
IoT, remote sensing, and digital twins	Liu et al. (2023); Divine et al. (2024); Zou et al. (2025); Miller et al. (2025); Kotwal (2025)	It addresses real-time data transmission, multi-source data fusion, remote environmental monitoring, digital twin-based fault detection, autonomous control, interoperability, and data security.	The model justifies the IoT-based data collection, digital twin, simulation, failure scenario analysis, early warning, and decision support layers.
Microalgae-bacteria consortia and bioremediation	Sundui et al. (2021); Oruganti et al. (2023); Sahu et al. (2023); Renganathan and Gaysina (2025); Renganathan et al. (2025); Blessing and Olateru (2025); Singh et al. (2025)	It describes microalgal treatment, algal-bacterial interactions, nutrient removal, biomass production, microbial consortium selection, xenobiotic removal, multi-omics, and artificial intelligence-supported bioremediation strategies.	In the proposed model, biological intervention supports the components of microbial process monitoring, contaminant-specific bioremediation, and resource recovery.
Micropollutants and Pollutant Classes	Mustafa et al. (2025); Altawell (2025)	It highlights the toxicological risks posed by microplastics, pharmaceutical residues, synthetic dyes, surfactants, heavy metals, persistent pollutants, and electronic waste.	Selecting a sensor based on pollutant class provides the rationale for determining the prediction model, early warning mechanism, and bioremediation strategy.

Research gap, governance, and feasibility	Lowe et al. (2022); Alotaibi and Nassif (2024); Wani et al. (2024); Capodaglio and Callegari (2025)	It emphasizes environmental monitoring, explainable artificial intelligence, early warning systems, data management, reproducibility, maintenance, implementation barriers, resource efficiency, and governance requirements.	It lays the groundwork for the study to propose a system architecture that is integrated, modular, implementable, explainable, and incorporates a governance dimension.

The existing technological elements for intelligent water and wastewater management are discussed quite effectively in Table 1. However, most of the time, these technological elements have been studied separately; there is no proper integration of biosensors' data analysis, Internet of Things-based data transfer, AI-based predictions, digital twin simulations, biological actions, and data management under one umbrella concept. This highlights the need for developing modular systems models to cater to the requirements of monitoring, warnings, decisions, biological actions, and governance.

With respect to the gap identified above, the current study offers an architectural approach that combines biosensors, IoT, artificial intelligence, digital twins, and biological intervention modules for wastewater treatment and bioremediation processes. Therefore, the current research aims at merging technological solutions, which have been addressed independently in existing literature, into one model of real-time decision-making.

PURPOSE OF THE STUDY AND RESEARCH QUESTIONS

The primary objective of this study is to develop a wastewater treatment and bioremediation monitoring model using artificial intelligence (AI) and biosensors for promoting water resource conservation. The proposed study is not based on the results obtained during a single laboratory experiment but is a conceptual model based on scientific literature analysis, aimed at designing and developing. Therefore, the conclusions of the current study were made according to the results of systematic thematic literature analysis conducted between 2020 and 2025.

This research is focused on answering the following questions:

1. What are the approaches used in terms of artificial intelligence and machine learning for monitoring, predicting, and optimizing certain parameters of biological wastewater treatment process?

2. How can biosensors, microfluidic devices, Internet of Things (IoT)-based systems, remote sensing techniques be implemented for creating real-time environmental monitoring systems?

3. How can microalgae-bacteria consortia optimization, engineered microbial consortia and bioremediation processes be optimized by means of AI methods?

4. How can be developed a modular and governance-oriented architecture of such systems which is flexible to different pollutant types and capacities?

METHOD

This study is a qualitative literature synthesis that combines narrative synthesis with a design-based research approach. Within the scope of the study, 28 academic publications published between 2020 and 2025 and related to the fields of wastewater treatment, bioremediation, artificial intelligence, machine learning, biosensors, IoT, digital twins, soft sensors, microalgae-bacteria consortia, micropollutants, and environmental governance were examined.

The literature review was based on the keywords “artificial intelligence,” “machine learning,” “wastewater treatment,” “biosensor,” “bioremediation,” “digital twin,” “IoT,” “soft sensor,” “microalgae-bacteria consortia,” “micropollutants,” and “environmental monitoring.” The study included recent academic publications directly related to wastewater treatment, environmental monitoring, or bioremediation processes and incorporating at least one of the following components: artificial intelligence, sensor technologies, or digital monitoring. Studies that were off-topic, served solely as general technology introductions, or were not directly related to the components of the proposed model were excluded from the scope.

The analysis process was conducted in three stages. In the first stage, the studies were classified according to subject areas. These categories included AI and machine learning-based wastewater treatment optimization, biological treatment and microbial systems, biosensor and IoT-based monitoring, digital twins and decision support systems, contaminant-class-focused bioremediation, and environmental governance and scalability.

In the second phase, information on the model type, monitored parameters, technology used, application level, strengths, and limitations were extracted from each study. This information was evaluated comparatively to identify current trends and application gaps in the literature.

The findings obtained in the third phase have been transformed into a conceptual system architecture that can be adapted to different classes of pollutants and field conditions. In this context, microfluidic chips and biosensors from the biomedical engineering domain; low-noise measurement, signal amplification, digitization, and data transmission components from the electrical engineering domain; and artificial intelligence-based prediction, digital twin, explainable decision support, and governance layers from the software engineering domain were addressed collectively.

This study does not include experimental validation. Therefore, conclusions regarding the performance of the proposed model are based not directly on new field data, but on a thematic synthesis of findings reported in the literature. Performance metrics reported in the literature—such as accuracy, RMSE, R^2, or classification accuracy—may vary depending on the dataset, facility type, sensor quality, pollutant profile, and operating conditions. For this reason, the system proposed in this paper is presented not as a validated field application, but as a conceptual research design for pilot implementation and subsequent experimental validation studies.

RESULTS

This section presents the findings derived from a thematic synthesis of the studies examined. The findings are evaluated under the following headings: AI-based performance prediction, biosensor- and soft sensor-supported monitoring, bioremediation using microbial consortia, digital twin-based early

warning systems, and intervention strategies based on pollutant class. The findings reveal how the proposed conceptual model was shaped based on the relevant literature, application needs, and technological components.

Performance Prediction and Process Optimization in AI-Based Wastewater Treatment

In the existing literature, various applications of AI in the wastewater treatment process are discussed with performance prediction, optimization of the processes, early warning and fault detection being the most significant among these. According to Singh et al. (2023), ANN, fuzzy logic, RF and LSTM are used in biological treatment of wastewater to predict BOD, COD, nutrients, suspended solids and metal concentration. Likewise, according to Malviya and Jaspal (2021), ANNs and hybrid intelligent systems could provide high prediction accuracy regarding the removal of COD, BOD, heavy metals and organic matter. As observed by Zhang et al. (2023), AI-based research is primarily focused on contaminant removal, process parameter optimization and membrane fouling management.

Environmental control techniques that use AI not only cover the prediction of effluent water quality. According to Ye et al. (2020), such applications as soft sensor, early warning system and pollutant concentration prediction could become part of the more extensive monitoring system for the environment. According to Alprol et al. (2024), the combined use of IoT, artificial intelligence and machine learning could turn conventional water quality monitoring into the cloud-based and automation-oriented techniques.

These findings imply that the AI component of the suggested system design does not have to be based on one algorithm. The ANN-based models have strengths when it comes to revealing nonlinear relationships; random forest and gradient boosting algorithms perform well in analyzing the impact of individual variables; time series methods like LSTM networks can be used for spotting trends, delays, and sudden changes in the sensor data. Therefore, the choice of algorithms has to depend on the data type, parameter being controlled, and purpose of operation, not the trends.

In their taxonomy of AI applications for drinking water treatment processes, Nagpal et al. (2024) identify three main purposes: (i) prediction of the pollutant removal efficiency, (ii) monitoring of the important water quality parameters in real time, and (iii) detecting process/equipment faults. Such classification justifies the use of an AI component in the suggested architecture. Hence, the model must not only predict the process performance but also indicate which sensor readings cause a certain risk. Thus, the explainability of the AI must be considered in designing the suggested model.

The Role of Biosensors, Microfluidic Systems, and Soft Sensors in the Monitoring Process

The primary requirement for real-time wastewater treatment management involves the consistent creation of data. According to Pouyanfar et al. (2022), microfluidics has great capacity for detecting environmental pollutants and interpreting the data with the help of artificial intelligence. Emaminejad et al. (2023) reveal that the data coming from the bioelectrochemical sensors can be used to predict shock loading and biological nutrient removal performance. In contrast, Reneeth et al. (2025) prove that expensive measurements including TSS, BOD, COD, TN, and TP may be estimated through soft sensors relying on easily measured factors.

All these facts mean that both physical and soft sensors must be developed simultaneously within the suggested framework. The physical sensors allow conducting direct measurements (pH, temperature, conductivity, dissolved oxygen, turbidity, and electrochemical response), whereas soft sensors enable estimation of those parameters that cannot be measured directly and constantly (BOD, COD, nutrient load, and pollutant concentration). The last mentioned can be estimated on the basis of correlation between data and the results obtained during laboratory measurements.

According to Divine et al. (2024) and Miller et al. (2025), it is possible to embed artificial intelligence, IoT, remote sensing, and data management techniques into existing environmental monitoring systems. As illustrated in Figure 1, the proposed framework should include not only data collected on site but also data obtained from river basins, industrial areas, urban effluent outlets, and receiving environments.

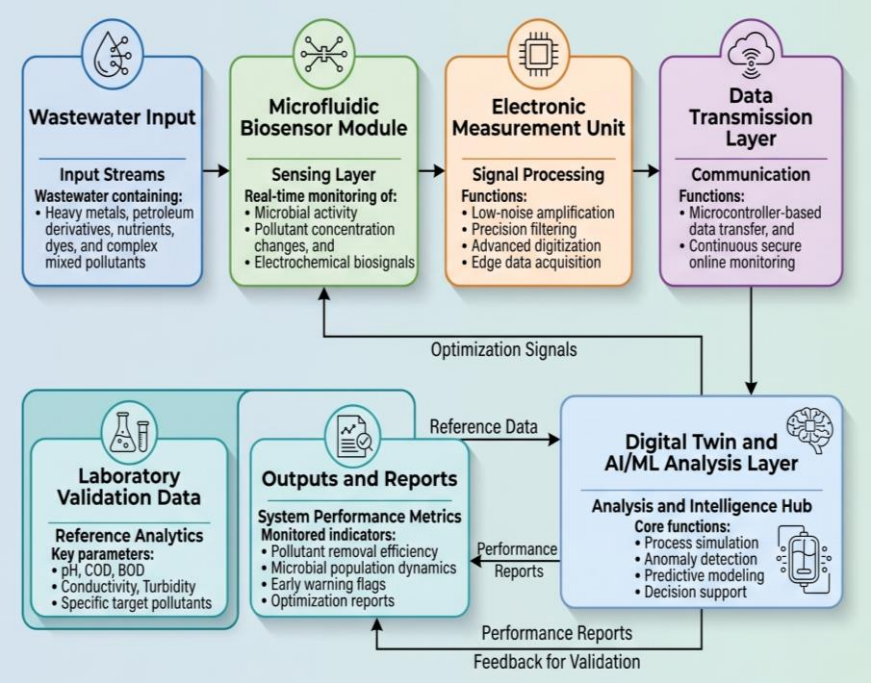


Figure 1. Conceptual architecture of the proposed biosensor-integrated wastewater treatment and bioremediation monitoring system.

The Contribution of Bioremediation Using Microbial Consortia to the Circular Economy

The findings from the literature reveal a move from utilizing biological treatments using a single type of microorganism to microbial consortia, which are stronger and have varied functions. Algae and bacteria represent another route in terms of treatment, especially in the removal of nutrients, harvesting, and production of biomass. According to Sundui et al. (2021) and Sahu et al. (2023), it is possible to use machine learning for the selection of species, culture, and operating parameters.

This indicates that wastewater treatment plants should not be viewed as just infrastructure that removes contaminants but as a biotechnology system that is connected to the circular economy via

nutrient recovery, biomass creation, and production of biofuel and bioproducts. According to Oruganti et al. (2023), microalgae treatment process can be combined with biofuel and bioproduct production.

However, pharmaceutical products, artificial dyes, endocrine disruptors, microplastics, and other xenobiotics place significant restrictions on traditional biological processes. Renganathan et al. (2025) point out that engineered microbial communities serve as an alternative method for bioremediation because they possess the ability to break down complex substances. This suggests that regional databases that provide information about microorganisms and contaminants need to be created in accordance with the proposed framework. The framework must offer a flexible model that can be adapted both for use in wastewater treatment plants and contaminated stretches of rivers, as well as industrial discharges and oil or hydrocarbon-contaminated locations.

Digital Twin-Based Early Warning and Closed-Loop Control Systems

The digital twin is defined as a computerized version of the physical treatment or monitoring system, which receives constant updates on the basis of the real-time data flow. According to Liu et al. (2023), the failures in the treatment process, like sludge bulking and toxicity issues, can be predicted by simulations through the digital twin model. On the other hand, Kotwal (2025) states that IoT sensors, hybrid physical-machine learning models, and AI control contribute to the real-time simulations, predictions, and control of the process.

These studies suggest that, in the context of the presented model, the digital twin should not be perceived as just an alarm generation framework but as a testbed for decision-making and optimization. Through the analysis of sensor data in the digital twin, possible failure cases, contaminant shock, performance degradation, and intervention options can be analyzed. This means that, instead of generating only the alert about the risk, the system predicts the reason for the risk and proposes the best intervention option.

However, data quality, explainability, standardization, and reproducibility are among the key limitations in digital twin and AI-supported water management systems. Zou et al. (2025) and Lowe et al. (2022) emphasize that these issues directly affect the reliability and feasibility of smart water management applications. As shown in Figure 2, adaptive closed-loop management requires the integration of sensing, data transmission, prediction, simulation, decision-making, and intervention stages. Therefore, in the proposed system, a data quality label, model confidence score, explanation report, and a decision class requiring human approval must be defined for each model output.



Figure 2. Digital twin-supported adaptive closed-loop management model: sensing, data transmission, prediction, simulation, decision-making, and intervention stages.

Monitoring, Prediction, and Remediation Strategies by Pollutant Class

As the type of pollutant changes, the monitoring method, prediction model, and bioremediation strategy to be implemented also vary. Therefore, the proposed model does not envisage a single type of sensor or a single intervention approach for all pollutants. While electrochemical sensors and biosorption-based approaches are preferred for heavy metals, bioelectrochemical sensors, soft sensors, and algal-bacterial consortia can be used to monitor nutrient loads. For petroleum derivatives, hydrocarbon-degrading microbial communities and, depending on site conditions, in situ or ex situ bioremediation strategies may be preferred. For more resistant pollutants, such as pharmaceutical residues and synthetic dyes, microbial consortia may need to be supported by advanced oxidation processes.

According to Mustafa et al. (2025), methods that consider only one type of contaminants could prove insufficient when considering both microplastics, nutrients, surfactants, synthetic dyes, pharmaceuticals, and organic matter at the same time when evaluating graywater micropollutants. Thus, this finding highlights the need for a method that takes into account different classes of contaminants and is modular in terms of its implementation as a monitoring-prediction-intervention framework.

Altawell (2025) notes that such pollutants as heavy metals, brominated flame retardants, and persistent organic pollutants from electronic waste present significant threats to both the environment and the population. In this case, the fact that water pollution is not limited to the boundaries of a wastewater treatment plant shows that solid waste, electronic waste, industrial drainage, and water quality need to be analyzed in the frame of the environmental risk chain. Likewise, Shah et al. (2025) show that inexpensive digital twins, enabled through IoT and artificial intelligence, can be used in waste management, hence making it possible to combine water and waste management at the level of smart

city. The relevant monitoring, prediction, and bioremediation techniques for each type of pollutants are summarized in Table 2.

Table 2. Monitoring, prediction, and bioremediation strategies by contaminant class.

Pollutant class	Suggested detection channel	AI/ML task	Bioremediation or control response
Heavy metals	Electrochemical sensor, microfluidic chip, periodic ICP/lab validation	RF/XGBoost risk classification; anomaly detection	Biosorption, biofilm carriers, metal-tolerant consortia
Petroleum hydrocarbons	Fluorescence proxy, ORP, DO, hydrocarbon-specific assay	Spill/shock classification; degradation-rate prediction	Hydrocarbon-degrading bacteria; in situ/ex situ bioremediation
Nutrients (N, P)	DO, pH, ORP, turbidity, bio-electrochemical sensor	Soft sensors for TN/TP; LSTM process forecasting	Algae-bacteria consortia; biological nutrient removal
Synthetic dyes	Colorimetric/microfluidic response, COD, UV-Vis validation	Removal-efficiency prediction; process optimization	Fungal/bacterial consortia; AOP-assisted biological treatment
Pharmaceuticals	Targeted lab validation plus surrogate online parameters	Hybrid AI plus process-based model	Engineered consortia; activated sludge polishing; adsorption/AOP support
Microplastics	Imaging, turbidity proxy, periodic microscopy/spectroscopy	Image-based classification; source tracking	Filtration, biofilm capture, sludge management and source control

PROPOSED MODEL

This study proposes a modular monitoring and intervention model that integrates artificial intelligence, IoT, biosensors, digital twins, and bioremediation approaches for use in wastewater and water quality monitoring processes. The model's primary objective is not only to detect the presence of pollutants but also to estimate the risk level, anticipate potential process issues, and provide appropriate intervention options through a decision support system.

Model Components

The proposed architecture consists of six layers. In the first layer is the sensor and sampling module, where basic online parameters such as pH, electrical conductivity, dissolved oxygen, temperature, turbidity, and oxidation reduction potential (ORP) are continuously monitored. Moreover, this layer can accommodate contaminant specific biosensors, bioelectrochemical sensors, and microfluidics systems.

The second layer refers to the edge-IoT data layer. This layer provides the buffer storage of field data, timestamping, geographic tagging, pre-filters the data, and sends the data to the central system securely. It ensures data flow systematically and with traceability rather than the ingestion of raw data in the model.

In the third layer is the artificial intelligence and machine learning layer. Here, biochemical oxygen demand (BOD), chemical oxygen demand (COD), total nitrogen (TN), total phosphorus (TP), heavy metal risk, nutrient load, shock loading, and treatment efficiency are predicted by artificial neural network (ANN), random forest (RF), XGBoost, long short-term memory (LSTM) neural networks, and anomaly detection models. Soft sensor models can also be developed for some hard-to-measure parameters.

The fourth layer is the digital twin and explainable decision-support layer. This layer does not directly convert the outputs generated by the AI model into operator actions; instead, it first evaluates variable importance, confidence scores, sensor data quality, and possible intervention options. Thanks to the digital twin approach, scenarios such as sludge bulking, toxic effects, changes in hydraulic load, and sudden pollutant influx can be tested in a virtual environment. Thus, the system evolves from a passive structure that merely triggers alarms into an active decision-support tool capable of performing scenario analysis prior to intervention.

The fifth layer is the bioremediation intervention layer. In this layer, intervention options such as microalgae-bacteria consortia, targeted bioaugmentation, biofilm carriers, aeration, dosing, and resource recovery are evaluated. Intervention recommendations are prioritized based on the type of contaminant, process conditions, and the model's confidence score.

The sixth layer is the governance and reporting layer. This layer covers processes such as regulatory compliance, data storage, model updates, cybersecurity, interagency data sharing, and operator training. Therefore, the model is designed not only as a technical monitoring system but also as an infrastructure for sustainable management and decision-making.

Implementation Architecture

The proposed architecture is designed to start with low-cost pilot stations and evolve over time into a regional water security network. The most appropriate approach for the initial implementation is to establish a pilot system at a high-risk watershed, an industrial discharge point, or the outlet of a wastewater treatment plant. It is recommended that online sensor data and laboratory validation data be collected concurrently in this pilot area for at least 6–12 months.

This period is critical for reliably training the model, as models built solely on short-term data cannot adequately represent seasonal variations, sensor contamination, process fluctuations, and shock loading conditions. Following the data collection process, soft sensor models are developed, model drift behavior is monitored, and sensor reliability is evaluated. As summarized in Table 3, the proposed system consists of multiple operational layers, including sensor and sampling, Edge-IoT data transmission, AI/ML prediction, digital twin-based decision support, bioremediation intervention, and governance/reporting components. The digital twin and closed-loop control components should be

phased in only after data reliability has been verified. Otherwise, the automatic control system may generate incorrect interventions based on erroneous data. This would make the system not an "intelligent" system, but rather an expensive panic button.

Table 3. Proposed system components and expected operational outputs

Layer	Basic components	Expected outcome
Sensor and sampling layer	pH, EC, DO, temperature, turbidity, ORP, bioelectrochemical signal, microfluidic contaminant response.	Raw and calibrated time series data; early anomaly warnings.
Edge-IoT data layer	Low power gateway, local data buffering, LoRa/Wi-Fi/cellular communication, metadata tagging.	Timestamped, location-informed, and pre-processed data stream.
AI/ML prediction layer	ANN, RF, XGBoost, LSTM, anomaly detection, soft sensors, predictive analysis	BOD, COD, TN, TP, heavy metal risk, organic load shock, and treatment efficiency estimation.
Digital twin and decision support layer	Virtual process model, toxic impact scenarios, sludge bulking simulation, explainable artificial intelligence, risk score.	Failure scenarios, "what-if" analysis, confidence score, and intervention recommendation.
Bioremediation intervention layer	Algae-bacteria consortia, targeted bioaugmentation, biofilm carriers, aeration and dosing	Biological intervention and recovery approach specific to the pollutant.
Governance and reporting layer	Data standards, cybersecurity, regulatory reporting, model updates, inter-organizational sharing.	Traceability, regulatory compliance, secure data management, and scalable implementation.

Operational Process

The implementation of the proposed model is designed as a six-stage validation and scaling process. In the first stage, high-risk basins, pollutant classes, existing treatment capacity, and data suitability are determined. This stage ensures that the pilot site selection is risk-based, not random.

In the second stage, low-cost online sensors and targeted biosensor modules are installed. The data obtained from the sensors are compared with laboratory analyses to carry out the calibration and validation process. In this way, the reliability of the data set that will feed the model is tested.

In the third stage, soft sensors and anomaly detection models are developed. Models such as ANN, RF, XGBoost, and LSTM are compared in terms of accuracy, uncertainty, interpretability, and robustness to field conditions. At this stage, it is not enough to simply select the model with the highest accuracy; the variables on which the model bases its predictions must also be evaluated. In the fourth stage, the digital twin component is implemented.

Through this component, scenarios such as toxic effects, sudden increases in organic load, hydraulic fluctuations, sensor failures, and biological process disruptions are simulated. This allows the system to test potential outcomes before actual intervention.

In the fifth stage, intervention options such as ventilation, dosing, additional sampling, maintenance, and bioaugmentation are tested under operator supervision. At this stage, a human-approved decision support approach should be preferred over fully automated control. This is because the goal of the system in the early pilot phase is not to eliminate the operator, but to make the operator's decision more reliable.

In the final stage, data standardization, cybersecurity, model updating, regulatory reporting, and user training processes are completed. After this stage, the system can be scaled from a single pilot application to a regional monitoring network. As illustrated in Figure 3, this phased implementation reduces the risk of technical failure while facilitating the system's adaptation to real field conditions. Therefore, the proposed approach aims not at directly establishing a fully automated treatment control system, but at first developing reliable data, then a predictive model, then a digital twin, and finally controlled automation.

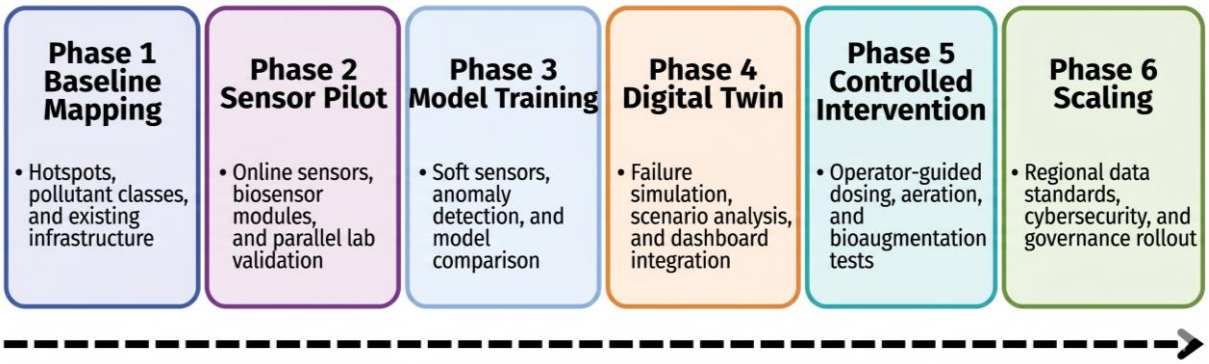


Figure 3. Transition process of the proposed monitoring model from pilot application to full-scale system.

DISCUSSION

The literature reviewed in this study shows that artificial intelligence, biosensors, IoT, digital twin, and bioremediation approaches have significant potential in wastewater treatment processes in terms of early warning, process monitoring, and decision support. The main contribution of the proposed model is to offer an integrated framework for detecting potential process failures at an early stage by evaluating continuous sensor data together with historical process records and laboratory analyses. In this respect, the system is designed not only as a passive structure monitoring the current situation, but also as an explainable decision support mechanism that provides information to the operator before intervention.

On the other hand, the primary problem of this model is its high dependency on the data quality from sensors. In the context of wastewater systems, the quality of data may be influenced by the sensor contamination, calibration error, data loss, seasonal change, and fast flow rate changes. Data quality, standardization, accessibility, and model explainability are the core issues raised in wastewater-based

modeling research by Lowe et al. (2022), Zou et al. (2025). This way, the suggested system must include regular checking for data quality, monitoring for calibration errors, and verification of online measurements through lab analysis. None of the predictions can be passed to the decision-making level without the data quality tag. Especially when high accuracy of predictions is achieved using the low-quality data

In wastewater treatment plants, simply providing the operator with an output such as "high risk" is insufficient. The model should make visible which sensors or parameters triggered the alarm, the confidence level of the prediction, similar historical events, and the potential impact of the proposed intervention on the process, energy, and cost. At this point, the function of explainable artificial intelligence is not to make the model more "attractive," but to enable the operator to identify erroneous or weak decisions. A model that does not produce explanations is either not used at all or is used without question; in both cases, the operational value of the system decreases. Therefore, the frequency with which the operator accepts, rejects, or manually overrides the model should also be monitored among the technical performance indicators.

In terms of bioremediation, engineered microbial consortia, multi-omics data, and AI-assisted bioreactor control offer significant opportunities (Renganathan et al., 2025; Renganathan and Gaysina, 2025). However, microbial communities do not behave like stable and predictable machine components. Community structure can be affected by changes in temperature, pH, toxic load, nutrient balance, antibiotic residues, and hydraulic conditions. Therefore, it cannot be assumed that a microbial consortium that is successful in a laboratory setting will maintain the same remediation performance in field conditions. Reactor fouling behavior, the spread of antibiotic resistance genes, local microbiome differences, and regional pollutant profiles can limit the generalizability of the model. Therefore, the evaluation should not be reduced solely to whether the target pollutant has been removed; the stability of the remediation performance, the system's recovery capacity after disruptive conditions, and unwanted microbial effects should also be monitored.

One of the most significant obstacles at the implementation level is that technical success does not directly translate into corporate and economic sustainability. Capodaglio and Callegari (2025) cite sensor maintenance burden, lack of process expertise, control loop instability, and lack of incentives for resource efficiency as major implementation barriers. Therefore, the economic value of the model cannot be assessed solely by its initial investment cost. Sensor cleaning, calibration, part replacement, data storage, software updates, model retraining, personnel time, and periods of system downtime must be included in the total cost of ownership. If the reduction in energy consumption, the increase in treatment efficiency, or the decrease in chemical use does not offset this cost, the system will not be economically sustainable even if it appears technically successful.

AI, IoT, remote sensing, and digital twin applications also generate new risks in terms of data ownership, cybersecurity, public health, and regulatory responsibility. Miller et al. (2025) identify data quality, interoperability, security, and ethical issues as key obstacles to such systems. In particular, it is crucial to clearly define who is responsible in the event of a false alarm or missed alarm. Implementing automated control without defining the boundaries of responsibility between the model developer, sensor provider, facility management, and operator can create organizational risk. Therefore, it should

be clarified before pilot implementation which interventions require human approval, under what conditions the model can be deactivated, and how decision logs will be maintained.

In conclusion, the proposed framework offers an integrated approach that addresses early warning, explainable decision support, and biological intervention capacity in wastewater treatment and bioremediation processes. However, the success of this approach depends on the combined consideration of a reliable data infrastructure, field-scale validation, operator participation, economic sustainability, biosafety monitoring, and transparent governance protocols. The key implementation risks and corresponding mitigation measures for the proposed AI-biosensor-bioremediation framework are summarized in Table 4. Therefore, future studies should focus not only on algorithm performance but also on the reliability, applicability, and institutional acceptability of the model under real facility conditions.

Table 4. Key implementation risks and mitigation measures for the proposed AI-biosensor-bioremediation framework.

Risk	**Why is it important?**	**Mitigation action**
Sensor contamination, calibration drift, and signal degradation.	Sensor accuracy can rapidly decrease in environments containing wastewater, sludge, and biofilm, leading to misleading measurements.	Regular cleaning and calibration, antifouling sensor design, redundant measurement, automatic data quality labeling, and laboratory verification.
Poor data quality and model drift.	Incomplete, noisy, or time-varying data can produce operationally risky predictions.	Sürekli veri kalite kontrolü, drift analizi, yeniden eğitim, güven eşiği ve düşük kaliteli tahminlerin reddedilmesi.
AI outputs with limited explainability.	Unexplained alarms and decisions can lead to operator insecurity, blind trust, or incorrect intervention.	Explainable AI reports, variable significance analysis, confidence score, decision trail, digital twin testing, and human validation in high-risk transactions.
Lack of regional data and poor interoperability.	The model may not be generalizable to other facilities when climate, wastewater composition, industrial pollutants, and monitoring practices differ.	Regional pilot stations, shared datasets, standard metadata, communication protocols, and field-based model calibration.
Microbial instability and biosecurity risks.	Microbial consortia may lose performance under field conditions; antibiotic resistance genes and local microbiome effects may pose additional risks.	Local strain banks, catalogs of contaminant-degrading microorganisms, multi-omics monitoring, ARG tracking, stepwise bioaugmentation, and controlled pilot validation.
High life cycle cost	Sensor and AI systems are unsustainable without maintenance, laboratory support, data management, and trained personnel.	Total cost of ownership analysis, phased investment, maintenance budget, laboratory

Risk	Why is it important?	Mitigation action
		support, operator training, and multidisciplinary technical teams.
Cybersecurity, data ownership, and accountability uncertainty.	Unauthorized access, data manipulation, and false/missed alarms create operational and legal risks due to uncertainty regarding accountability.	Role-based access, encryption, audit logs, secure data storage, inter-organizational data protocols, and clear definitions of responsibility.
Regulatory acceptance and limited field verification.	A technically successful prototype may not be implemented if it does not comply with existing standards.	Controlled field pilots, parallel reporting with accredited laboratory methods, performance thresholds, phased regulatory validation, and institutional pilot studies.

CONCLUSION AND RECOMMENDATIONS

This study proposes a conceptual wastewater treatment and bioremediation monitoring model integrating artificial intelligence, biosensor, soft sensor, and digital twin approaches for the conservation of water resources. Literature findings show that integrated analysis of data obtained from treatment plants and environmental monitoring systems has significant potential for estimating pollutant loads, monitoring process performance, early detection of failures, and improving decision support processes.

The main contribution of the proposed model is to integrate wastewater treatment, environmental monitoring, bioremediation, and decision support processes within a single digital-biotechnological architecture. This structure is based on processing real-time data obtained from sensors and biosensors using artificial intelligence and digital twin approaches; it offers an adaptable monitoring infrastructure for municipal wastewater, industrial discharges, greywater, surface water pollution, and contaminated sites.

However, the model is still at the conceptual level and needs to be validated under real-world field conditions. Sensor contamination, loss of calibration, incomplete or inconsistent data sets, the impact of environmental variables on biosensors, and the inability to generalize artificial intelligence models to different conditions are the main factors that may limit the application's success. Therefore, it is recommended that the model be tested in the first stage with controlled pilot studies; that sensor data be validated with laboratory analyses; and that its impact on prediction accuracy, early warning success, false alarm rate, maintenance needs, energy consumption, and operating costs be evaluated.

Future studies should focus on experimentally validating the proposed architecture in different facilities and for different pollutant classes. Biosensor-machine learning integration, soft sensors, and bioremediation strategies can be tested in separate pilot applications for pollutants such as heavy metals, petroleum derivatives, nitrogen-phosphorus loads, pharmaceuticals, dyes, and microplastics.

In conclusion, this study demonstrates the applicability of an integrated monitoring approach based on artificial intelligence, biosensors, and digital twins in wastewater treatment and bioremediation processes. However, the model's ability to generate real value depends on the simultaneous fulfillment of conditions such as reliable data infrastructure, field validation, economic sustainability, biosafety, human oversight, and corporate governance.

References

Alotaibi, E., & Nassif, N. (2024). Artificial intelligence in environmental monitoring: in-depth analysis. Discover Artificial Intelligence, 4(1), 84.

Alprol, A. E., Mansour, A. T., Ibrahim, M. E. E. D., & Ashour, M. (2024). Artificial intelligence technologies revolutionizing wastewater treatment: current trends and future prospective. Water, 16(2), 314.

Altawell, N. (2025). E-Waste Management in the Digital Age: Global Trends, Toxicological Risks, and Green Recovery Technologies. International Congresses of Turkish Science and Technology Publishing, 1468-1471.

Blessing, A. A., & Olateru, K. (2025). AI-driven optimization of bioremediation strategies for river pollution: a comprehensive review and future directions. Frontiers in Microbiology, 16, 1504254.

Capodaglio, A. G., & Callegari, A. (2025). Use, potential, needs, and limits of AI in wastewater treatment applications. Water, 17(2), 170.

Divine, I. L. O. H., Aguma, C. P., & Olagunju, A. O. (2024). Integrating AI enhanced remote sensing technologies with IOT networks for precision environmental monitoring and predicative ecosystem management. World Journal of Advanced Research and Reviews, 23(02), 2156-2166.

Emaminejad, S. A., Sparks, J., & Cusick, R. D. (2023). Integrating bio-electrochemical sensors and machine learning to predict the efficacy of biological nutrient removal processes at water resource recovery facilities. Environmental Science & Technology, 57(46), 18372-18381.

Kotwal, V. S. (2025). A digital twin framework for intelligent water treatment, quality monitoring, and autonomous control. International Journal of Applied Mathematics, 38(7s), 565-580.

Liu, W., He, S., Mou, J., Xue, T., Chen, H., & Xiong, W. (2023). Digital twins-based process monitoring for wastewater treatment processes. Reliability Engineering & System Safety, 238, 109416.

Lowe, M., Qin, R., & Mao, X. (2022). A Review on Machine Learning, Artificial Intelligence, and Smart Technology in Water Treatment and Monitoring. Water.

Malviya, A., & Jaspal, D. (2021). Artificial intelligence as an upcoming technology in wastewater treatment: a comprehensive review. Environmental Technology Reviews, 10(1), 177-187.

Miller, T., Durlik, I., Kostecka, E., Kozlovska, P., Łobodzińska, A., Sokołowska, S., & Nowy, A. (2025). Integrating artificial intelligence agents with the internet of things for enhanced environmental monitoring: applications in water quality and climate data. Electronics, 14(4), 696.

Mustafa, M., Epelle, E. I., Macfarlane, A., Cusack, M., Burns, A., & Yaseen, M. (2025). Innovative approaches to greywater micropollutant removal: AI-driven solutions and future outlook. RSC advances, 15(16), 12125-12151.

Nagpal, M., Siddique, M. A., Sharma, K., Sharma, N., & Mittal, A. (2024). Optimizing wastewater treatment through artificial intelligence: recent advances and future prospects. Water Science & Technology, 90(3), 731-757.

Oruganti, R. K., Biji, A. P., Lanuyanger, T., Show, P. L., Sriariyanun, M., Upadhyayula, V. K., & Bhattacharyya, D. (2023). Artificial intelligence and machine learning tools for high-performance microalgal wastewater treatment and algal biorefinery: A critical review. Science of the Total Environment, 876, 162797.

Pouyanfar, N., Harofte, S. Z., Soltani, M., Siavashy, S., Asadian, E., Ghorbani-Bidkorbeh, F., & Hussain, C. M. (2022). Artificial intelligence-based microfluidic platforms for the sensitive detection of environmental pollutants: Recent advances and prospects. Trends in Environmental Analytical Chemistry, 34, e00160.

Reneeth, F., Abbasi, T., & Abbasi, S. A. (2025). Intelligent Effluent Management: AI-Based Soft Sensors for Organic and Nutrient Quality Monitoring. Processes, 13(6), 1664.

Renganathan, P., & Gaysina, L. A. (2025). Next-generation wastewater treatment: omics and AI-driven microbial strategies for xenobiotic bioremediation and circular resource recovery. Processes, 13(10), 3218.

Renganathan, P., Gaysina, L. A., García Gutiérrez, C., Rueda Puente, E. O., & Sainz-Hernández, J. C. (2025). Harnessing engineered microbial consortia for xenobiotic bioremediation: integrating multi-omics and AI for next-generation wastewater treatment. Journal of Xenobiotics, 15(4), 133.

Sahu, S., Kaur, A., Singh, G., & Arya, S. K. (2023). Harnessing the potential of microalgae-bacteria interaction for eco-friendly wastewater treatment: a review on new strategies involving machine learning and artificial intelligence. Journal of Environmental Management, 346, 119004.

Shah, K. B., Visalakshi, S., Guragain, D. P., & Panigrahi, R. (2025). Advancing smart city sustainability with Internet of Things and artificial intelligence aided low-cost digital twin systems for waste management. Microsystem Technologies, 31(10), 2783-2796.

Singh, K., Sarkar, S., & Sharma, S. (2025). Artificial Intelligence in Bioremediation and Environmental Management. Environmental Quality Management, 35(2), e70192.

Singh, N., Yadav, M., Singh, V., Padhiyar, H., Kumar, V., Bhatia, S., & Show, P. (2023). Artificial intelligence and machine learning-based monitoring and design of biological wastewater treatment systems. Bioresource Technology, 369, 128486.

Sundui, B., Ramirez Calderon, O. A., Abdeldayem, O. M., Lázaro-Gil, J., Rene, E. R., & Sambuu, U. (2021). Applications of machine learning algorithms for biological wastewater treatment: updates and perspectives. Clean Technologies and Environmental Policy, 23(1), 127-143.

Wani, A. K., Rahayu, F., Ben Amor, I., Quadir, M., Murianingrum, M., Parnidi, P., & Latifah, E. (2024). Environmental resilience through artificial intelligence: innovations in monitoring and management. Environmental Science and Pollution Research, 31(12), 18379-18395.

Ye, Z., Yang, J., Zhong, N., Tu, X., Jia, J., & Wang, J. (2020). Tackling environmental challenges in pollution controls using artificial intelligence: A review. Science of the Total Environment, 699, 134279.

Zhang, S., Jin, Y., Chen, W., Wang, J., Wang, Y., & Ren, H. (2023). Artificial intelligence in wastewater treatment: A data-driven analysis of status and trends. Chemosphere, 336, 139163.

Zou, S., Ju, H., & Zhang, J. (2025). Water Quality Management in the Age of AI: Applications, Challenges, and Prospects. Water.

Ethical Statement

This study is based on a literature review and conceptual model design. No data were collected from human or animal participants, no field sampling was conducted, and no personal data were processed within the scope of the study. Therefore, no procedure requiring ethics committee approval was carried out.

Conflict of Interest and Financial Support

The authors declare that there is no conflict of interest within the scope of this study. No specific financial support was received from any institution, organization, or project for this study.

An Artificial Intelligence-Supported Pedagogical Compliance Model for Fostering Academic Integrity in Higher Education

Hasiyat Shihova[1]*, Mehriniso Haytbayeva[2], Orifjon Saidmamatov[3]

[1,2]Urgench State University named after Abu Rayhan Beruni, Urgench, Uzbekistan
[3]Mamun University, Uzbekistan

*Corresponding author e-mail: shihovahasiyat@gmail.com

Abstract

The rapid diffusion of generative artificial intelligence (AI) in higher education has intensified concerns about academic misconduct while simultaneously offering new instruments for cultivating integrity. Prevailing institutional responses remain largely enforcement-oriented, relying on detection and sanction rather than the internalization of ethical values. This study proposes pedagogical compliance (pedagogik komplayens), a construct that reframes academic integrity as an internalized student value cultivated through instructional design rather than external control. Grounded in the value-internalization model of O'Reilly and Chatman (1986), the academic-dishonesty framework of McCabe and Treviño (1993), and Self-Determination Theory (Ryan and Deci, 2000), the paper presents a conceptual AI-supported model that operationalizes pedagogical compliance in the classroom. The model integrates the values of the International Center for Academic Integrity (ICAI) with a graduated framework for the legitimate use of AI in assessment, translating them into concrete instructional and technological components: adaptive formative feedback, transparency regarding permitted AI use, and developmental rather than punitive monitoring. The paper further considers implementation requirements for higher education institutions in Central Asia and outlines directions for empirical validation. By reframing integrity as a cultivated value supported—rather than policed—by AI, the study contributes an ethically grounded, technology-enabled alternative to enforcement-centric approaches.

Keywords: *Academic integrity, Pedagogical compliance, Artificial intelligence in education, Information ethics, Higher education*

INTRODUCTION

Generative artificial intelligence (AI) has rapidly become embedded in the everyday academic practice of students worldwide. Tools capable of producing fluent essays, solving problem sets, and generating code have blurred the boundary between legitimate assistance and academic misconduct (Cotton et al., 2024). Higher education institutions have responded predominantly through an enforcement paradigm: expanding detection technologies, tightening prohibitions, and increasing the certainty and severity of sanctions. Yet decades of empirical research indicate that punitive deterrence alone is a weak and often counter-productive foundation for integrity (McCabe et al., 2001). Detection tools for AI-generated text remain unreliable, disproportionately flag non-native English writers, and cast the student–teacher relationship as adversarial rather than developmental.

This paper argues that the challenge posed by generative AI is not primarily technological but pedagogical, and that its resolution requires shifting from the control of misconduct to the cultivation of integrity as an internalized student value. We introduce the construct of pedagogical compliance (Uzbek: pedagogik komplayens) and propose an AI-supported model that operationalizes it. Rather than deploying AI as a surveillance instrument, the model repurposes it as a formative, autonomy-supportive companion that helps students internalize the values of academic integrity while using AI transparently and legitimately. The paper first develops the theoretical background, then defines the construct, presents the model and its components, discusses implementation in the Central Asian context, and concludes with directions for empirical validation.

THEORETICAL BACKGROUND

Academic integrity and its contextual determinants

The International Center for Academic Integrity (ICAI, 2021) defines academic integrity as a commitment to six fundamental values: honesty, trust, fairness, respect, responsibility, and courage. Crucially, ICAI frames integrity not as compliance with rules but as action grounded in shared values. Empirical work by McCabe and Treviño (1993) demonstrated that the strongest predictors of academic dishonesty are contextual rather than dispositional: peer behavior, perceived peer norms, and the presence of a functioning honor culture outweigh the perceived severity of penalties. Their decade-review (McCabe et al., 2001) reinforced that integrity is sustained by institutional culture and student socialization rather than by detection and punishment. This evidence provides the empirical warrant for a values-based, rather than sanction-based, approach.

Value internalization as a foundation for integrity

O'Reilly and Chatman (1986) distinguished three psychological mechanisms by which individuals adopt organizational values: compliance (instrumental conformity to obtain rewards or avoid punishment), identification (conformity to maintain a valued relationship), and internalization (adoption because the values are congruent with one's own). Behaviors sustained by internalization are more stable, self-regulating, and resistant to situational pressure than those sustained by compliance. Applied to academic integrity, this framework exposes the limitation of enforcement: at best it produces instrumental compliance that dissolves the moment surveillance is absent. The pedagogical objective must therefore be to move students along the continuum from external compliance toward genuine internalization.

Self-Determination Theory and autonomy-supportive environments

Self-Determination Theory (Ryan and Deci, 2000; Deci and Ryan, 1985) explains the conditions under which such internalization occurs. It identifies three basic psychological needs—autonomy, competence, and relatedness—whose satisfaction fosters the internalization of extrinsically introduced values into self-endorsed motivation. Critically, controlling environments—those characterized by surveillance, coercion, and contingent sanctions—thwart the need for autonomy and undermine internalization, whereas autonomy-supportive environments promote it. This yields a direct design implication: integrity interventions built on monitoring and threat are theoretically predicted to weaken the very disposition they seek to build, while those that support autonomy, provide competence-building feedback, and preserve a respectful relational climate are predicted to strengthen it.

Generative AI and the assessment challenge

The arrival of generative AI has destabilized traditional assessment by making outsourced production of academic work trivially easy (Cotton et al., 2024). Two responses dominate practice: prohibition-plus-detection, and unstructured permissiveness. Both are inadequate—the former is unenforceable and adversarial, the latter abandons learning outcomes. The AI Assessment Scale (AIAS; Perkins et al., 2024) offers a constructive alternative: a graduated framework specifying, for each assessment, the permitted level of AI involvement—ranging from no AI, through AI-assisted idea generation and editing, to full AI collaboration—coupled with a requirement of transparent disclosure. The AIAS reframes AI use as a matter of transparent, level-appropriate practice rather than binary permission, and thereby provides an operational vocabulary for legitimate use.

FROM ENFORCEMENT TO PEDAGOGICAL COMPLIANCE

Defining pedagogical compliance

We define pedagogical compliance as the systematic cultivation, through instructional design, of academic integrity as an internalized and self-regulated student value, supported rather than enforced by technology. The construct integrates the three theoretical foundations above: it targets internalization (O'Reilly and Chatman, 1986) as its end state, recognizes contextual and cultural determinants (McCabe and Treviño, 1993) as its levers, and adopts autonomy-supportive design (Ryan and Deci, 2000) as its method. Pedagogical compliance is thus deliberately positioned against the enforcement paradigm: where enforcement seeks behavioral conformity under surveillance, pedagogical compliance seeks value internalization under support.

Reconciling "compliance" with internalization

The term "compliance" warrants clarification, since in the O'Reilly–Chatman taxonomy it denotes the shallowest, instrumental level of value adoption. Our use of "pedagogical compliance" derives instead from the institutional-compliance tradition developed in corporate governance and corruption-prevention practice, where "compliance" denotes an organized system for aligning conduct with a body of norms. Transposed to pedagogy, the term names not the psychological state of instrumental conformity but the organized instructional system through which value alignment is cultivated. The intended end state of that system is precisely the deepest level of the O'Reilly–Chatman continuum—internalization. Pedagogical compliance is therefore best read as a compliance architecture whose goal is internalization, which resolves the apparent tension.

THE AI-SUPPORTED PEDAGOGICAL COMPLIANCE MODEL

Architecture and design principles

The model comprises four interlocking components governed by three design principles: transparency over concealment, formation over sanction, and autonomy-support over control. Each component is technology-enabled but pedagogically governed; AI functions as an instrument of formation, not of surveillance.

Component 1 — Graduated transparency of AI use

Adopting the AIAS (Perkins et al., 2024), each assessment specifies an explicit, level-appropriate permission for AI use, published in advance, and students disclose the nature and extent of AI assistance. This component operationalizes the value of honesty and removes the ambiguity that drives inadvertent misconduct, replacing a prohibition students routinely circumvent with a transparent norm they can follow.

Component 2 — Adaptive formative feedback

AI is deployed to deliver individualized, formative feedback on drafts—identifying gaps in reasoning, unsupported claims, and citation weaknesses—before submission. By satisfying the need for competence (Ryan and Deci, 2000) and improving work through legitimate means, this component reduces the perceived necessity of misconduct while modeling ethical AI use.

Component 3 — Developmental monitoring

Where monitoring occurs, it is reframed from detection-and-sanction to developmental dialogue. Signals of potential difficulty trigger instructor-mediated conversation and support rather than automated accusation. This preserves the relational climate and the presumption of trust that McCabe and Treviño (1993) identify as central to integrity cultures.

Component 4 — Values-embedded curriculum

The six ICAI values (ICAI, 2021) are made explicit learning objectives rather than implicit background assumptions. Integrity is taught, discussed, and assessed as a competence—through case analysis, reflective disclosure of AI use, and peer dialogue—so that students construct integrity as a personally endorsed value rather than an externally imposed rule.

IMPLEMENTATION CONSIDERATIONS FOR CENTRAL ASIAN HIGHER EDUCATION

Translating the model into practice in the Central Asian context—and specifically in Uzbekistan—raises several requirements. First, instructor readiness: the shift from a control orientation to an autonomy-supportive one requires professional development, particularly where institutional cultures have historically emphasized enforcement. Second, policy alignment: institutional academic-integrity regulations must accommodate transparent, level-appropriate AI use rather than blanket prohibition. Third, multilingual delivery: instruments, disclosures, and training must operate across Uzbek, Russian, and English to be equitable and effective. Fourth, cultural transposition: the ICAI values and AIAS levels require careful localization rather than direct import. Institutional experience with transitioning compliance-control infrastructure toward academic-integrity development suggests that existing compliance systems can be productively re-purposed as the organizational backbone for pedagogical compliance.

DISCUSSION

The proposed model contributes a theoretically grounded alternative to the prevailing enforcement paradigm. Its principal strength is coherence: each component is derived from an established theoretical mechanism and aligned with the others toward a single end—internalized integrity. The model also reframes generative AI from threat to pedagogical resource, offering institutions a constructive path between prohibition and permissiveness.

Several limitations must be acknowledged. The model is, at this stage, conceptual; its components are theoretically motivated but not yet empirically validated as an integrated system. The central hypothesis—that formative, transparent, autonomy-supportive design produces more durable integrity than enforcement—is testable and must be tested. Contextual factors, including disciplinary differences and resource constraints, will condition implementation. Finally, the repurposing of monitoring, however developmental in intent, carries ethical risks that require governance safeguards.

CONCLUSION AND FUTURE WORK

This paper introduced pedagogical compliance as a construct and proposed an AI-supported model for cultivating academic integrity as an internalized student value. By integrating value-internalization theory, contextual accounts of academic dishonesty, and self-determination theory with the AIAS and the ICAI values, the model offers an ethically grounded, technology-enabled alternative to enforcement-centric responses to generative AI. Future work will operationalize the model's components into concrete instructional protocols and empirically validate their effect on integrity outcomes—including through a multilingual adaptation of the McCabe Academic Integrity Survey administered within Uzbek higher education—thereby advancing pedagogical compliance from conceptual framework to evidence-based practice.

References

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innovations in Education and Teaching International, 61(2), 228–239.

Deci, E. L., & Ryan, R. M. (1985). Intrinsic Motivation and Self-Determination in Human Behavior. New York: Plenum Press.

International Center for Academic Integrity [ICAI]. (2021). The Fundamental Values of Academic Integrity (3rd ed.). Retrieved from https://academicintegrity.org

McCabe, D. L., & Treviño, L. K. (1993). Academic dishonesty: Honor codes and other contextual influences. The Journal of Higher Education, 64(5), 522–538.

McCabe, D. L., Treviño, L. K., & Butterfield, K. D. (2001). Cheating in academic institutions: A decade of research. Ethics & Behavior, 11(3), 219–232.

O'Reilly, C. A., & Chatman, J. (1986). Organizational commitment and psychological attachment: The effects of compliance, identification, and internalization on prosocial behavior. Journal of Applied Psychology, 71(3), 492–499.

Perkins, M., Furze, L., Roe, J., & MacVaugh, J. (2024). The Artificial Intelligence Assessment Scale (AIAS): A framework for educational assessment in the age of generative AI. Journal of University Teaching and Learning Practice, 21(6).

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist, 55(1), 68–78.

Short-Term Electricity Load Forecasting Using Machine Learning: A Comparative Study Against the Turkish Transmission System Operator's Official Forecasts

Bedirhan Demirci[1*]

[1]Topkapı University, Department of Data Science and Analytics, 34000, Istanbul, Türkiye

*Corresponding author e-mail: bedirhandemirci@topkapi.edu.tr

Abstract

Accurate short-term electricity load forecasting is essential for efficient grid operation, economic dispatch, and the integration of renewable energy sources into national power systems. This study proposes a machine learning-based framework for hourly short-term load forecasting using four years (2022–2025) of real, publicly available consumption data obtained from the Turkish Electricity Transparency Platform (EPİAŞ). Two ensemble learning algorithms, Random Forest and Extreme Gradient Boosting (XGBoost), were trained using autoregressive lag features (24-hour and 168-hour lags), rolling mean statistics, calendar-based variables, and population-weighted ambient temperature obtained from five major provinces. Hyperparameters were optimized via randomized search combined with time-series cross-validation to prevent information leakage between training and test folds. Model performance was evaluated on the most recent 12 months of held-out data and benchmarked against two references: a naive persistence forecast and the official day-ahead load forecast published by the Turkish transmission system operator (TEİAŞ). Results show that the tuned Random Forest model achieved a mean absolute percentage error (MAPE) of 1.97%, outperforming the naive baseline (5.70%) and, notably, the official TEİAŞ forecast (3.07%) by approximately 36%. Feature importance analysis revealed that the 168-hour (weekly) lag feature was the dominant predictor, while the inclusion of temperature-derived features provided negligible additional improvement, suggesting that autoregressive load patterns implicitly capture most weather-driven demand variation. These findings demonstrate that data-driven ensemble models can meaningfully improve upon existing operational forecasting practices in Turkey's power grid, offering a practical, reproducible, and computationally efficient alternative for short-term load forecasting applications.

Keywords: *Short-term load forecasting, machine learning, xgboost, random forest, electricity consumption*

INTRODUCTION

Accurate short-term load forecasting (STLF) is a cornerstone of modern power system operation. Utilities and transmission system operators rely on hourly demand forecasts to schedule generation, manage reserves, clear day-ahead markets, and maintain the balance between supply and demand in real time. As power systems increasingly incorporate variable renewable generation and more price-responsive market mechanisms, the economic and operational costs of forecasting error have grown substantially, making STLF an active and practically important research area (Hong and Fan, 2016).

In Türkiye, the national transmission system operator, TEİAŞ, publishes a day-ahead load plan for the following day as part of its system planning duties. Historical realizations of both the official load plan and the real-time metered consumption are made publicly available through the EPİAŞ Transparency Platform (EPİAŞ, 2026), providing a unique opportunity to evaluate data-driven forecasting methods directly against an operational benchmark rather than against another data-driven model alone.

Machine learning methods, and ensemble tree-based algorithms in particular, have been applied with growing frequency to electricity forecasting problems in the Turkish context. Yılan and Beykent (2026) used Extreme Gradient Boosting (XGBoost) to forecast day-ahead electricity prices in the Turkish market and reported that it outperformed Support Vector Machines, Long Short-Term Memory networks, Random Forest, and Gradient Boosting alternatives. Öğütlü (2025) compared eight ensemble regression algorithms for national-scale electricity consumption forecasting using a high-dimensional macroeconomic and sectoral dataset, while Saglam et al. (2023) combined optimization techniques with machine learning algorithms to project long-term electricity demand. However, these studies primarily benchmark machine learning models against one another or against classical statistical baselines; direct comparison against the load forecast that the system operator itself publishes and operationally relies upon remains uncommon.

This study addresses that gap by developing an hourly short-term load forecasting framework for the Turkish national grid using four years of real, publicly available EPİAŞ data, and by benchmarking the resulting models directly against TEİAŞ's own official day-ahead load plan. In addition, the study examines whether adding exogenous ambient temperature information provides a measurable benefit once rich autoregressive lag features are already present in the model, a question with direct practical implications for the cost and complexity of operational forecasting pipelines.

MATERIAL AND METHODS

Material

The primary dataset used in this study consists of hourly electricity consumption records for the Turkish national grid, obtained from the EPİAŞ Transparency Platform (EPİAŞ, 2026) via its official REST API. Two series were retrieved for the period from 1 January 2022 to 31 December 2025, yielding 35,064 hourly observations: (i) the real-time metered consumption, which serves as the target variable in this study, and (ii) TEİAŞ's official day-ahead load plan, which is used as an operational benchmark rather than as a model input.

Ambient temperature data were obtained from the Open-Meteo historical weather archive for the five most populous Turkish provinces (Istanbul, Ankara, Izmir, Bursa, and Antalya), which together account for a large share of national electricity demand. Hourly temperature series for the five provinces were combined into a single population-weighted national temperature series, used to derive the weather-related predictors described below.

Methods

The Collection of the Data

Consumption and load plan data were retrieved programmatically using an open-source Python client for the EPIAŞ Transparency Platform API, which handles the ticket-granting authentication required by the platform. Data were requested in monthly batches to comply with the API's practical request limits and subsequently concatenated into a single hourly time series. Weather data were retrieved separately through the Open-Meteo archive API for the same period and merged with the consumption series on the common hourly timestamp. Fewer than 0.1% of consumption records contained missing values, which were filled using time-based linear interpolation.

Statistical Analysis

A set of predictor variables was constructed from the merged dataset. Calendar-based features included the hour of day, day of week, month, quarter, day of year, and a binary weekend indicator. Autoregressive features included the consumption value 24 hours and 168 hours (one week) prior to the forecasted hour, together with rolling mean consumption over the preceding 24 and 168 hours. Weather-related features included the population-weighted temperature itself, along with heating-degree (HDD) and cooling-degree (CDD) terms, defined as:

$$HDDt = max(18 - Tt,\ 0) \qquad CDDt = max(Tt - 24,\ 0)$$

where Tt denotes the population-weighted temperature at hour t, and the reference values of 18°C and 24°C follow common heating/cooling degree-day conventions.

Two ensemble tree-based regression algorithms were used to model the target variable. Random Forest (Breiman, 2001) constructs a large number of decorrelated decision trees on bootstrap-resampled subsets of the training data and averages their predictions, reducing variance relative to a single tree. Extreme Gradient Boosting, or XGBoost (Chen and Guestrin, 2016), builds an additive ensemble of shallow regression trees in a stage-wise manner, with each new tree trained to correct the residual errors of the current ensemble under a regularized objective function. Both algorithms were implemented using their standard open-source Python libraries.

The dataset was split chronologically rather than randomly to respect the temporal ordering of the data and avoid information leakage: the final 12 months (1 January 2025 to 31 December 2025, 8,760 hourly observations) were held out as a test set, while the preceding approximately three years were used for training and hyperparameter selection.

Hyperparameters for both algorithms were selected using randomized search over a pre-specified grid (20 candidate configurations per model), combined with five-fold time-series cross-validation on the training set. Unlike conventional random k-fold cross-validation, time-series cross-validation preserves the chronological order of the data in each fold, ensuring that a model is never validated on data that precedes its training window, a precaution that is necessary to avoid optimistic bias when working with sequentially dependent data such as electricity demand.

Model performance was evaluated using three standard forecasting error metrics: the mean absolute percentage error (MAPE), the root mean squared error (RMSE), and the mean absolute error (MAE), computed on the held-out test set:

$$MAPE = (1/n)\ \Sigma\ |(yt - \hat{y}t)\ /\ yt| \times 100, \quad for\ t = 1\ to\ n$$

where yt is the actual metered consumption at hour t, ŷt is the model's predicted consumption, and n is the number of observations in the test set. RMSE and MAE were computed following their standard definitions as the root mean squared and mean absolute differences between yt and ŷt, respectively.

Two reference forecasts were used to benchmark the machine learning models. The first is a naive persistence forecast, which simply repeats the observed consumption from 24 hours earlier. The second, and more demanding, benchmark is TEİAŞ's own official day-ahead load plan, representing the forecast currently relied upon in operational practice.

RESULTS

Table 1 summarizes the forecasting performance of all evaluated models on the 12-month held-out test period (1 January – 31 December 2025). The naive persistence baseline achieved a MAPE of 5.70%, while TEİAŞ's official day-ahead load plan achieved a substantially lower MAPE of 3.07%, confirming that the published operational forecast already substantially outperforms a purely naive approach.

Table 1. Forecasting performance of baseline and machine learning models on the 12-month held-out test set (2025).

Model	MAPE (%)	RMSE (MW)	MAE (MW)
Naive (t−24h)	5.70	3432.2	2243.6
TEİAŞ official load plan	3.07	1863.2	1236.6
Random Forest (tuned)	1.99	1216.3	789.4
XGBoost (tuned)	1.97	1204.0	784.3
Random Forest (tuned, +weather)	**1.97**	**1202.3**	**782.2**
XGBoost (tuned, +weather)	2.02	1212.6	806.5

MAPE: mean absolute percentage error; RMSE: root mean squared error; MAE: mean absolute error; TEİAŞ: Turkish Electricity Transmission Corporation (Türkiye Elektrik İletim A.Ş.). Bold values indicate the best-performing configuration.

All four tuned machine learning configurations outperformed both reference forecasts by a wide margin. The best-performing configuration, a Random Forest model trained with weather features, achieved a MAPE of 1.97%, corresponding to an improvement of approximately 36% relative to the official TEİAŞ load plan and approximately 65% relative to the naive baseline. The differences among the four tuned configurations themselves were small (MAPE ranging from 1.97% to 2.02%) and are unlikely to reflect a robust ranking beyond ordinary variation across hyperparameter search runs; in particular, adding weather features did not yield a consistent or substantial improvement over the corresponding models trained without them.

Figure 1 shows the relative feature importance of the tuned XGBoost model trained with weather features. The one-week (168-hour) lag of consumption was by far the most influential predictor, accounting for approximately 67% of total importance, followed by the 24-hour lag (13%) and the 24-hour rolling mean (6%). Among the weather-related features, the cooling-degree term was the fifth most important predictor overall (5%), noticeably more influential than either the raw temperature value or the heating-degree term (each below 1%). This pattern is consistent with a summer-peaking, air-conditioning-driven component of Turkish electricity demand, but also indicates that most of the systematic weather-driven variation in demand is already captured indirectly through the strong weekly autoregressive structure of the series.

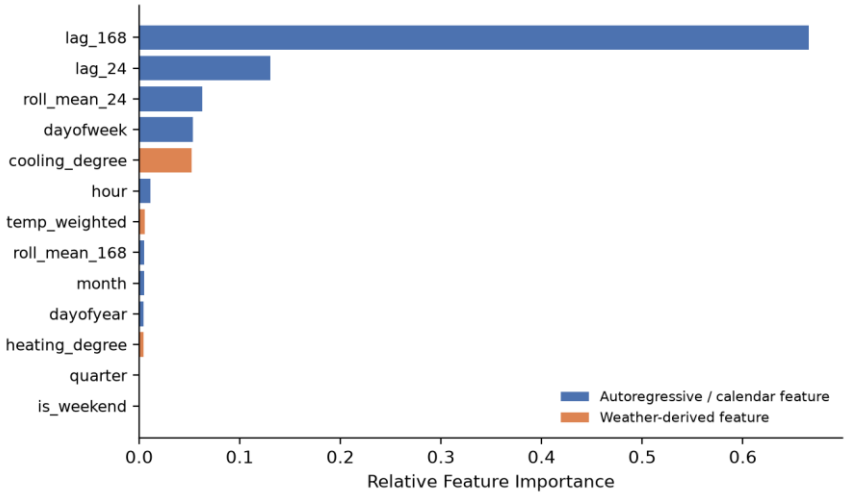


Figure 1. relative feature importance of the tuned XGBoost (+weather) model.

Figure 2 illustrates this behaviour over a representative one-week window (7–13 July 2025) during a summer demand peak. The official TEİAŞ load plan systematically underestimates consumption during the daily peak hours, with the gap between the actual and planned load reaching approximately 4,000–5,000 MW on several afternoons. The tuned XGBoost model, by contrast, tracks the actual consumption curve closely throughout the week, including at the peaks and troughs, visually corroborating the quantitative improvement reported in Table 1.

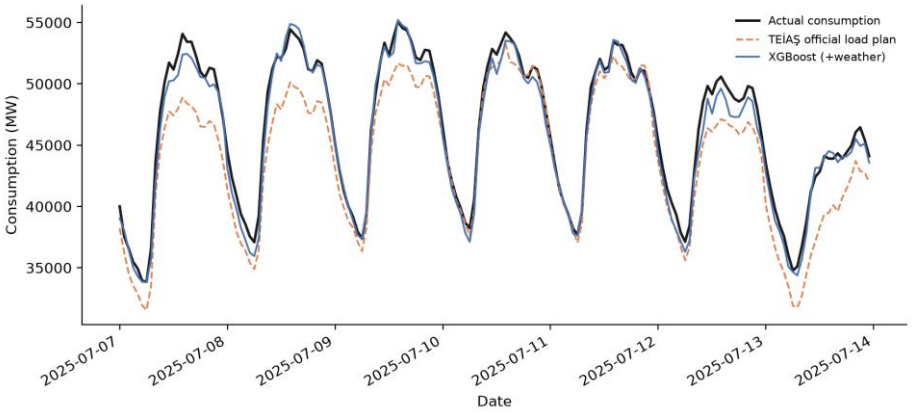


Figure 2. actual consumption versus the TEİAŞ official load plan and the tuned XGBoost (+weather) forecast, 7–13 July 2025.

DISCUSSION AND CONCLUSION

The results demonstrate that relatively simple, tree-based ensemble models, combined with a compact set of autoregressive and calendar features, can meaningfully outperform the day-ahead load forecast currently published by Türkiye's transmission system operator. A MAPE improvement of approximately 36% relative to an established operational benchmark represents a substantively large gain in a domain where forecasting error translates directly into balancing costs and reserve requirements.

The negligible contribution of temperature-derived features is a notable finding in light of the broader load forecasting literature, which generally treats weather variables as indispensable predictors (Hong and Fan, 2016). A plausible explanation is that the 168-hour lag feature, the consumption observed exactly one week earlier, already implicitly encodes much of the temperature-driven seasonal variation in demand, since weather conditions tend to be substantially correlated across consecutive weeks within the same season. This suggests that the marginal value of exogenous weather variables in short-term load forecasting may depend strongly on whether sufficiently rich autoregressive information is already available to the model, an aspect that is not always made explicit in comparative studies such as those of Öğütlü (2025) and Saglam et al. (2023).

Several limitations should be acknowledged. First, the test period covers a single calendar year (2025); results may vary under different weather or macroeconomic conditions. Second, the calendar features used in this study do not account for the shifting dates of national and religious holidays, most notably Ramadan and the associated Bayram holidays, which are known to affect electricity demand patterns in Türkiye but do not fall on fixed calendar dates from year to year. Third, forecasting was performed at the national aggregate level; regional or distribution-level demand may exhibit different dynamics and a different sensitivity to weather. Finally, the hyperparameter search space, while informed by common practice, was constrained by computational budget and does not guarantee a globally optimal configuration for either algorithm.

Future work could extend this framework by incorporating a religious and national holiday calendar, evaluating probabilistic rather than point forecasts, and testing whether the findings generalize to regional or distribution-level load series. Given the strong performance of comparatively lightweight tree-based models observed here, an explicit comparison against deep learning architectures such as LSTM or Transformer-based models would also be a natural extension, particularly for applications requiring longer forecast horizons.

In conclusion, this study shows that data-driven ensemble models, trained on publicly available consumption and weather data, can provide a practical, reproducible, and computationally efficient alternative to, or complement for, existing operational load forecasting practice in Türkiye's power grid, while also offering interpretable insight into the relative importance of autoregressive and weather-driven demand drivers.

Referens

Breiman L, 2001. Random forests. Machine Learning, 45(1): 5-32.

Chen T, Guestrin C, 2016. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 13-17, 2016, Page: 785-794, San Francisco, USA.

EPİAŞ, 2026. Transparency Platform. Access address: https://seffaflik.epias.com.tr; Date of access: 24.07.2026.

Hong T, Fan S, 2016. Probabilistic electric load forecasting: A tutorial review. International Journal of Forecasting, 32(3): 914-938.

Öğütlü AS, 2025. National-scale electricity consumption forecasting in Turkey using ensemble machine learning models: An interpretability-centered approach. Sustainability, 17(21): 9829.

Saglam M, Spataru C, Karaman OA, 2023. Forecasting electricity demand in Turkey using optimization and machine learning algorithms. Energies, 16(11): 4499.

Yılan Y, Beykent A, 2026. Day-ahead electricity price forecasting using the XGBoost algorithm: An application to the Turkish electricity market. Computers, Materials & Continua, 86(1): 1-16.

Predicting AI-Enabled Workplace Efficiency: A Comparative Study of Random Forest, Logistic Regression, and Linear SVM in Albania

Irida Gjermeni[1]*, Juejdfior Zhaka[2]

[1,2]Agricultural University of Tirana, Tirana, Albania

* Corresponding author e-mail: igjermeni@ubt.edu.al

Abstract

This study evaluates whether employee artificial-intelligence (AI) readiness can classify high perceived workplace efficiency. An online questionnaire produced 106 responses; after excluding respondents younger than 18 and respondents not in employment, 101 cases remained. High efficiency comprised employees reporting that AI improved task completion substantially or completely (n = 43), while the remaining categories formed the comparison class (n = 58). Thirteen predictors represented demographic characteristics, industry, work experience, AI experience, AI use frequency, confidence, and training. Random Forest, binary Logistic Regression, and Linear Support Vector Machine (SVM) were evaluated through repeated stratified five fold cross-validation with three repetitions. Random Forest provided the best overall performance: accuracy = .743, balanced accuracy = .729, F1 = .670, and ROC-AUC = .800. Logistic Regression achieved the highest sensitivity (.672), whereas Random Forest achieved the highest specificity (.822). Permutation importance identified AI-use frequency as the dominant predictor, followed by AI training, AI experience, and employment in technology and IT. The findings indicate that direct engagement and capabilit building are more informative than demographic variables for predicting perceived efficiency. Because the outcome is self-reported and the sample is small and non probabilistic, the model should be interpreted as exploratory rather than deployment ready.

Keywords**:** *Artificial Intelligence, Workplace Efficiency, Random Forest, Machine Learning, Digital Transformation*

INTRODUCTION

Artificial intelligence is increasingly embedded in routine knowledge work, customer service, decision support, and administrative processes. Experimental evidence shows that generative AI can reduce task completion time and improve output quality, while field evidence from customer support work reports meaningful average productivity gains with heterogeneous effects across employees (Noy & Zhang, 2023; Brynjolfsson et al., 2025). These differences suggest that organizational access to AI is not sufficient: employees also need practical experience, repeated use, confidence, and training.

Workplace AI combines automation and augmentation. Automation transfers selected task components to machines, whereas augmentation supports human judgment and capability; these processes frequently coexist and create organizational tensions (Raisch & Krakowski, 2021). AI use can

also be understood as delegation to agentic information systems, requiring workers to formulate requests, monitor outputs, and retain responsibility for final decisions (Baird & Maruping, 2021). Consequently, perceived efficiency may depend on the quality and frequency of human AI interaction rather than demographic characteristics alone.

Recent studies emphasize reskilling, task technology fit, and employee preparedness. Worker AI coexistence requires technical, conceptual, and human skills (Zirar et al., 2023). Training is associated with organizational AI adoption (Muehlemann, 2025), and AI supported workload reduction can contribute to performance when implementation is aligned with work processes (Rožman et al., 2023). However, AI integration may be appraised simultaneously as a challenge and a hindrance, creating both proactive behavior and job-insecurity concerns (Huang & Gursoy, 2024). This makes employee level predictive analysis relevant for digital-transformation and decision support research.

Evidence from Albania and similar small transition economies remains limited. This study therefore asks: 1. Can AI readiness indicators distinguish employees with high perceived AI enabled efficiency? 2. Which of Random Forest, Logistic Regression, and Linear SVM performs best under repeated cross validation? 3. Which predictors are most important in the leading model?

MATERIAL AND METHODS

Material and data collection

The material was a 24-variable online questionnaire entitled “Technology and Artificial Intelligence,” completed during March 2025. It measured demographics, employment, AI experience and use, efficiency, workload, decision making, stress, training, productivity, innovation, and professional growth. The original dataset contained 106 observations. One respondent younger than 18 and respondents identified as students or unemployed were excluded, leaving 101 complete employee records. The sample was 52.5% male; 59.4% were aged 19–24 years; 47.5% held a master’s degree; and 35.6% worked in technology and IT. The study used a cross sectional convenience sample.

Outcome and predictors

The five-level efficiency item was converted into a binary target. “Substantially” and “completely” were coded as high efficiency (Y = 1; n = 43, 42.6%), while “not at all,” “slightly,” and “moderately” were coded as non-high efficiency (Y = 0; n = 58, 57.4%). Predictors were age, education, gender, work experience, grouped industry, AI experience, AI use frequency, AI confidence, and AI training. Ordinal responses were converted to ordered scores, nominal predictors were dummy coded, and industries with fewer than five observations were grouped as other. The final matrix contained 13 predictors.

Machine learning models and validation

The three models and their original mathematical formulations are shown in Figure 1. Logistic Regression estimates the probability of the positive class through a logistic link (Cox, 1958). Linear SVM estimates a maximum-margin hyperplane with slack penalties for classification violations (Cortes & Vapnik, 1995). Random Forest aggregates probability estimates across 1,000 randomized trees and applies a .50 classification threshold (Breiman, 2001). Standardization was estimated separately within each training fold for Logistic Regression and SVM; Random Forest used unstandardized predictors. Model evaluation used repeated stratified five fold cross validation with three repetitions, ensuring that each employee was held out once in each repetition.

$$P(Y_i = 1 \mid \mathbf{x}_i) = \frac{1}{1 + \exp\left[-\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)\right]}$$

(1) Binary logistic model — Cox (1958)

$$\min_{\mathbf{w},\, b,\, \xi} \tfrac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i$$
$$\text{s.t. } y_i(\mathbf{w}^{\mathsf{T}}\mathbf{x}_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0$$

(2) Soft-margin support vector machine — Cortes and Vapnik (1995)

$$\hat{p}_{RF}(Y = 1 \mid \mathbf{x}) = \tfrac{1}{B}\sum_{b=1}^{B} \hat{p}_b(Y = 1 \mid \mathbf{x})$$
$$\hat{Y} = \mathbb{I}[\hat{p}_{RF} \geq 0.50], \qquad B = 1000$$

(3) Random-forest probability aggregation — Breiman (2001)

Figure 1. Mathematical formulations used in the study, with attribution to the original developers.

Performance was summarized by accuracy, balanced accuracy, precision, sensitivity (recall), specificity, F1 score, and area under the receiver operating characteristic curve (ROC-AUC). Balanced accuracy was emphasized because it averages sensitivity and specificity and is less affected by unequal class sizes. Permutation importance was calculated for the final Random Forest; it represents the reduction in predictive performance after a predictor is randomly permuted and should not be interpreted causally (James et al., 2021). Analyses were conducted in R 4.4.2 using ranger and kernlab.

RESULTS

Cross-validated model performance

Table 1 presents the mean and standard deviation across the 15 validation folds. Random Forest ranked first for accuracy (.743), balanced accuracy (.729), F1 score (.670), ROC-AUC (.800), specificity (.822), and precision (.742). Logistic Regression produced the highest sensitivity (.672), meaning that it detected the largest proportion of high efficiency respondents. Linear SVM achieved ROC-AUC = .793 but had the lowest mean F1 score (.619). The differences were modest; thus, Random Forest is treated as the primary exploratory model rather than as an overwhelmingly superior classifier.

Table 1. Repeated stratified five-fold cross-validation performance (mean ± SD)

Model	Accuracy	Balanced acc.	F1	ROC-AUC	Sensitivity	Specificity	Precision
Random Forest	0.743 ± 0.076	0.729 ± 0.081	0.670 ± 0.116	0.800 ± 0.080	0.635 ± 0.169	0.822 ± 0.111	0.742 ± 0.129
Logistic Regression	0.723 ± 0.088	0.716 ± 0.095	0.665 ± 0.123	0.790 ± 0.097	0.672 ± 0.184	0.760 ± 0.107	0.683 ± 0.123
Linear SVM	0.713 ± 0.083	0.697 ± 0.094	0.619 ± 0.154	0.793 ± 0.086	0.588 ± 0.203	0.806 ± 0.108	0.695 ± 0.100

Note. The high-efficiency class was the positive class. SD = standard deviation; ROC-AUC = area under the receiver operating characteristic curve.

The pooled ROC curves in Figure 2 remained above the no-skill diagonal and largely overlapped, confirming that all three models possessed useful discrimination. Random Forest maintained the highest mean ROC-AUC and balanced accuracy, although the close curves indicate that model selection should also consider the operational costs of false positive and false negative decisions.

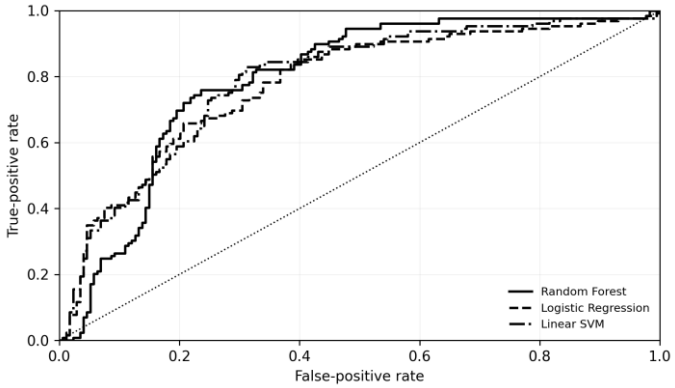


Figure 2. ROC curves based on pooled out of fold predictions from the three models.

Model-specific results and interpretation

Table 2. Random Forest model-specific validation results

Indicator	Fold mean	SD	Pooled estimate/count
Accuracy	0.743	0.076	0.743
Balanced accuracy	0.729	0.081	0.729
F1-score	0.670	0.116	0.678
ROC-AUC	0.800	0.080	0.792
Sensitivity	0.635	0.169	0.636
Specificity	0.822	0.111	0.822
Precision	0.742	0.129	0.726
True positives	—	—	82
True negatives	—	—	143
False positives	—	—	31
False negatives	—	—	47

Random Forest produced the strongest overall profile. Its mean accuracy (.743), balanced accuracy (.729), and ROC-AUC (.800) were the highest among the evaluated models. The pooled specificity of .822 and 143 true negative classifications show that the model was particularly effective at identifying respondents without high perceived efficiency. However, sensitivity was lower (.636), and the 47 false negatives exceeded the 31 false positives. Consequently, Random Forest is appropriate when avoiding false claims of high efficiency is relatively important, but it may miss a meaningful share of genuinely high-efficiency employees.

Table 3. Logistic Regression model-specific validation results

Indicator	Fold mean	SD	Pooled estimate/count
Accuracy	0.723	0.088	0.723
Balanced accuracy	0.716	0.095	0.717
F1-score	0.665	0.123	0.674
ROC-AUC	0.790	0.097	0.778
Sensitivity	0.672	0.184	0.674
Specificity	0.760	0.107	0.759
Precision	0.683	0.123	0.674
True positives	—	—	87
True negatives	—	—	132

False positives	—	—	42
False negatives	—	—	42

Logistic Regression yielded the highest mean sensitivity (.672) and a pooled sensitivity of .674, detecting 87 of the 129 positive observations generated across the three repetitions. Its pooled sensitivity, precision, and F1-score were all approximately .674, indicating a comparatively balanced treatment of the positive class. Nevertheless, specificity (.759 pooled) was lower than for Random Forest and Linear SVM, and the model generated 42 false positives. Logistic Regression is therefore preferable when the priority is to identify more potentially high efficiency employees and the cost of additional false positive classifications is acceptable.

Table 4. Linear SVM model-specific validation results

Indicator	Fold mean	SD	Pooled estimate/count
Accuracy	0.713	0.083	0.713
Balanced accuracy	0.697	0.094	0.697
F1-score	0.619	0.154	0.636
ROC-AUC	0.793	0.086	0.789
Sensitivity	0.588	0.203	0.589
Specificity	0.806	0.108	0.805
Precision	0.695	0.100	0.691
True positives	—	—	76
True negatives	—	—	140
False positives	—	—	34
False negatives	—	—	53

Linear SVM retained good discrimination, with mean ROC-AUC = .793 and pooled ROC-AUC = .789. Its mean specificity (.806) and pooled specificity (.805) were high, but sensitivity was the lowest of the three models (.588 mean; .589 pooled). The model correctly classified 140 non-high observations but identified only 76 high observations, resulting in 53 false negatives and the lowest mean F1 score (.619). This indicates a conservative decision boundary that favors the negative class; therefore, Linear SVM is less suitable when missing high efficiency employees is costly.

Random Forest predictor importance

The pooled Random Forest confusion matrix and the leading predictors are presented in Figure 3. AI-use frequency was the dominant predictor, with permutation importance of .0659. Its contribution was more than three times that of AI training (.0205) and AI experience (.0192), and more than four times the technology and IT indicator (.0157). Work experience, confidence, age, gender, education, and the remaining industry indicators added little stable predictive information. Negative or near zero values indicate negligible incremental usefulness after permutation and should not be interpreted as negative causal effects.

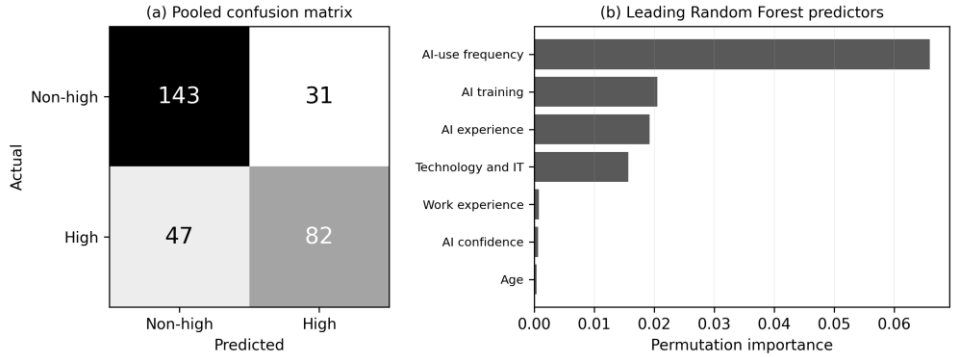


Figure 3. (a) Pooled Random Forest confusion matrix and (b) leading permutation-importance values.

Table 5. Leading Random Forest permutation-importance values

Rank	Predictor	Importance
1	AI-use frequency	0.0659
2	AI training	0.0205
3	AI experience	0.0192
4	Technology and IT industry	0.0157
5	Work experience	0.0008
6	AI confidence	0.0006
7	Age	0.0004

DISCUSSION AND CONCLUSION

The results show that a compact set of AI readiness indicators can predict high perceived workplace efficiency with useful, although not deployment-level, discrimination. Random Forest achieved ROC-AUC = .800 and balanced accuracy = .729. Logistic Regression and Linear SVM performed only slightly worse, indicating that the predictive signal was relatively stable across linear and nonlinear decision boundaries. Random Forest's main practical advantage was its higher specificity, whereas Logistic Regression was preferable when sensitivity was prioritized.

The substantive finding is the dominance of AI use frequency. Regular use may help employees identify suitable tasks, improve prompting, verify outputs, and integrate AI into established workflows. This interpretation is consistent with productivity evidence showing that AI gains emerge through actual task use and are heterogeneous across workers (Noy & Zhang, 2023; Brynjolfsson et al., 2025). Training and prior AI experience were the next most informative variables, supporting research that emphasizes reskilling and task–technology alignment in human–AI collaboration (Zirar et al., 2023; Muehlemann, 2025; Przegalinska et al., 2025). The results imply that organizations should combine access with task-specific training, guided experimentation, output verification, and human oversight.

The importance of the technology and IT indicator was lower than the importance of direct AI readiness measures. Thus, sectoral exposure matters, but repeated engagement and capability-building may be transferable across sectors. Demographic predictors contributed little once use, training, and experience were included. This is a predictive, sample-specific result and should not be generalized as evidence that demographic inequalities are absent. Human oversight remains essential because explainable AI and the ability to validate algorithmic recommendations can improve task performance in human AI collaboration (Senoner et al., 2024).

The study has several limitations. The outcome was perceived rather than objectively measured efficiency; the sample was small, young, highly educated, and non-probabilistic; dichotomization discarded ordinal information; hyperparameters were prespecified rather than optimized through nested cross-validation; and no independent external test set was available. Permutation importance does not establish causal effects. Future studies should retain the ordinal outcome, recruit larger multisector samples, include objective performance indicators, conduct nested tuning and probability calibration, and validate models across organizations and time.

In conclusion, Random Forest was the strongest evaluated classifier, but its advantage was modest. AI-use frequency, training, and experience were the clearest predictors of high perceived efficiency. For digital transformation, the results favor investments in repeated practical use and structured capability development rather than technology acquisition alone.

References

Baird, A., & Maruping, L. M. (2021). The next generation of research on IS use: A theoretical framework of delegation to and from agentic IS artifacts. MIS Quarterly, 45(1), 315–341. https://doi.org/10.25300/MISQ/2021/15882

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brynjolfsson, E., Li, D., & Raymond, L. R. (2025). Generative AI at work. The Quarterly Journal of Economics, 140(2), 889–942. https://doi.org/10.1093/qje/qjae044

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20, 273–297. https://doi.org/10.1007/BF00994018

Cox, D. R. (1958). The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological), 20(2), 215–242.

Huang, Y., & Gursoy, D. (2024). How does AI technology integration affect employees' proactive service behaviors? A transactional theory of stress perspective. Journal of Retailing and Consumer Services, 77, Article 103700. https://doi.org/10.1016/j.jretconser.2023.103700

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning: With applications in R (2nd ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1

Muehlemann, S. (2025). Artificial intelligence adoption and workplace training. Journal of Economic Behavior & Organization, 238, Article 107206. https://doi.org/10.1016/j.jebo.2025.107206

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. Science, 381(6654), 187–192. https://doi.org/10.1126/science.adh2586

Przegalinska, A., Triantoro, T., Kovbasiuk, A., Ciechanowski, L., Freeman, R. B., & Sowa, K. (2025). Collaborative AI in the workplace: Enhancing organizational performance through resource-based and task-technology fit perspectives. International Journal of Information Management, 81, Article 102853. https://doi.org/10.1016/j.ijinfomgt.2024.102853

Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. Academy of Management Review, 46(1), 192–210. https://doi.org/10.5465/amr.2018.0072

Rožman, M., Oreški, D., & Tominc, P. (2023). Artificial-intelligence-supported reduction of employees' workload to increase the company's performance in today's VUCA environment. Sustainability, 15(6), Article 5019. https://doi.org/10.3390/su15065019

Senoner, J., Schallmoser, S., Kratzwald, B., Feuerriegel, S., & Netland, T. (2024). Explainable AI improves task performance in human–AI collaboration. Scientific Reports, 14, Article 31150. https://doi.org/10.1038/s41598-024-82501-9

Valtonen, A., Saunila, M., Ukko, J., Treves, L., & Ritala, P. (2025). AI and employee wellbeing in the workplace: An empirical study. Journal of Business Research, 199, Article 115584. https://doi.org/10.1016/j.jbusres.2025.115584

Zirar, A., Ali, S. I., & Islam, N. (2023). Worker and workplace artificial intelligence (AI) coexistence: Emerging themes and research agenda. Technovation, 124, Article 102747. https://doi.org/10.1016/j.technovation.2023.102747

Analysis of Football Player Data Using Artificial Intelligence and Recommendations for Player-Specific Training Plans [3]

Mehmet Tarık Turhan[1*], Halime Suvay Eker[2]

[1] Istanbul Gedik University, Faculty of Engineering, Computer Engineering, 34876, Istabul, Türkiye
[2] Istanbul Gedik University, Faculty of Economics Administrative and Social Sciences, Management Information Systems, 34876, Istanbul, Türkiye

*Corresponding author e-mail: tturhan335@gmail.com

Abstract

Artificial intelligence is advancing every day and is being widely used in various fields. As soccer has become a growing industry worldwide, the use of AI-based data analysis in soccer has also gained importance. Due to the intense training and match schedule, soccer players' physical condition is constantly changing. Analyzing player-specific data collected via various sensors using artificial intelligence methods can enable the tracking of players' development, the monitoring of injury risks, and the taking of early precautions against potential risks. In this study, skill measurement data, such as agility, speed, power, ball-handling, first touch, and the ability to use both feet, collected from soccer players via sensors were analyzed using artificial intelligence methods, and based on the results, player-specific training plans tailored to the players' performance and physical condition were recommended. The study aims to present an approach that enables more effective monitoring of players' individual development and the early identification of injury risks through the AI-assisted analysis of soccer player data.

Keywords: *Artificial Intelligence, Football Analytics, Personalized Training*

INTRODUCTION

In modern football, an increasing amount of data is being generated to track players' physical and technical performance. Thanks to wearable sensors and performance tracking platforms, it is possible to measure the distance players cover, their speed, sprinting activities, training loads, and various technical performance indicators (Bourdon et al., 2017; Scott et al., 2016). However, two fundamental problems arise, particularly for amateur and developing players. The first is that measurement systems typically present the player's activity in numerical terms but cannot, on their own, explain what this data means for the player's development. For example, simply telling a player, "You ran 4 km," may not be sufficient to explain how to interpret this performance or what the player should focus on in future training sessions.

[3]This study is an outcome of a project conducted under the TÜBİTAK–2209-A Research Project Support Program for University Students.

The second problem is that amateur players do not have the same level of access to coaches, performance specialists, sports scientists, and advanced analytical infrastructure as those found in professional clubs. The literature indicates that performance and injury data regarding amateur footballers is more limited compared to that of professional football (Gurau et al., 2023; Sonesson et al., 2023; Szymski et al., 2022). Understanding why a particular drill or task is assigned to a player is important for making sense of the training process. For this reason, it is not enough to simply measure performance data; it must also be communicated to the player in a way that is understandable and well-reasoned. In their study, Claudino et al. (2019) systematically examined how artificial intelligence is used to assess performance and injury risk in team sports. Rossi et al. (2018) proposed a multidimensional approach to predicting injuries in professional football based on GPS measurements and machine learning.

In this study, the CityPlay tracking device was used to collect data from players on their physical performance (metrics such as distance covered, work rate, sprint distance, number of sprints, maximum speed, high-speed runs, acceleration, and deceleration), technical performance (such as two-footed ability, dribbling, first touch, agility, speed, power, and, where applicable, passing), and the players' daily condition data (such as readiness, fatigue, mood, sleep duration, sleep quality, muscle pain/soreness, and stress) were collected via the CityPlay tracking device. The data collected was then fed into a deterministic, rule-based decision support system powered by the state-of-the-art AI agent Claude Sonnet 5, which generated player-specific training exercise programs as well as training programs tailored to each player's position and physical condition. The generated programs can be downloaded as PDF documents based on the player's profile.

METHODS

Aim and Significance of the Study

The aim of this study is not merely to present physical and technical performance data collected from football players in numerical form, but rather to analyze this data through an interpretable decision-support system and translate it into player-specific development priorities and training recommendations. This approach has been developed in parallel with the increasing use of artificial intelligence and data-driven decision-making mechanisms in sports analytics (Ghosh et al., 2023). Professional football clubs can regularly monitor players' development through coaches, performance specialists, sports scientists, and advanced tracking systems. Amateur and developing players, however, have more limited access to these resources. For this reason, Maestro AI aims to make measured performance metrics more meaningful by evaluating them in conjunction with the player's position and the available data set.

The Collection of the Data

Physical performance, technical performance, and daily condition data collected from a total of two players using the CityPlay system formed the dataset for this study. Table 1 below describes the variables of the data groups used in this study and their purpose within the system.

Table 1. The structure of the dataset used in the study

Dataset	Variables	Intended use in the system
Physical performance	Session duration, total distance, work rate, sprint distance, maximum speed, acceleration, deceleration, turn	Evaluation of the player's physical output during the session
Technical performance	Agility, speed, strength, ball-handling, first touch, ability to use both feet	Analysis of a player's technical performance profile
Daily status	Level of readiness, fatigue, mood, duration of sleep, sleep quality, muscle pain/tenderness, stress	Analysis of a player's daily status

As part of this study, an application was submitted to the Scientific Research and Publication Ethics Committee at Istanbul Gedik University to collect data from the football players. The Committee unanimously approved a favorable opinion on the application dated June 25, 2025, numbered 2025/2, with barcode number 385167.

Material

In this study, data from football players was collected using the CityPlay system. In the CityPlay system, small sensors are attached to both of the player's soccer cleats via silicone straps. According to technical information provided by Playermaker, the six-axis sensors used can sample the player's movements 1,000 times per second and detect movements such as ground contact, ball contact, and foot rotation. The system does not require GPS, Wi-Fi, or a camera infrastructure set up around the field during data collection, and the data obtained after the session is synchronized with the app (Playermaker, 2026).

One of CityPlay's key features is its ability to track not only the player's physical movements but also football-specific technical interactions. According to information provided by Playermaker, the system can measure more than 25 physical and technical metrics. Technical indicators include ball touches, number of possessions, time in possession, time to clear the ball, shot power, turns performed with the ball, and ball interactions using different parts of the foot. Physical metrics include total distance covered, sprint distance, work rate, maximum speed, acceleration and deceleration, and changes of direction (Playermaker, 2026).

Thus it is clear that the CityPlay system is a data collection tool capable of comprehensively measuring both players' physical performance and football-specific technical movements with high precision, without requiring any on-field infrastructure.

The data collected from the system was then transferred to ClaudeSonnet 5, and a training program was developed for each player. ClaudeSonnet 5 is a large language model belonging to the Sonnet class of the Claude model family developed by Anthropic. Released by Anthropic on June 30, 2026, the model was developed for coding, reasoning, tool use, information retrieval, and the execution of multi-step tasks. Anthropic describes Sonnet 5 as a model that has been improved compared to the previous Sonnet

4.6 model, particularly in terms of coding, tool use, reasoning, and autonomous task execution capabilities (Anthropic, 2026). In this study, Claude Sonnet 5 was not used as an AI model that directly evaluates soccer players' performance or independently makes final training decisions.

RESULTS

The study was conducted with two players. The data collected in the CityPlay system was transferred to tables. The data for Player 1 and Player 2 are presented below, respectively.

Table 2. Actual session data and performance indicator for Player 1

Dataset	Indicator	Value
Physical performance	Session duration	45 min
	Total distance	4,0 km
	Working speed	89,1 m/min
	Sprint distance	29,9 m
	Maximum speed	25,9 km/sa
Technical performance	Goal	0
	Assist	4
	Use of both feet	43/100
	Dribbling	70/100
	First touch	53/100
	Agility	64/100
	Speed	84/100
	Power	7/10
Daily status	Readiness	5/10
	Fatigue	2/10
	Mood	10/10
	Muscle pain/ sensitivity	No data
	Stress	No data

An examination of Table 2 reveals that Player 1 covered a total distance of 4 km during the 45-minute training session, with a working speed of 89,1 m/min and a sprint distance of 29,9 m. It was determined that his maximum speed was recorded at 25,9 km/h. Looking at technical performance metrics, while he did not contribute any goals, his playmaking stood out with 4 assists. At the same time, while both footed use was at a relatively low level of 43/100 and first touch was 53/100, dribbling was at 70/100, agility at 64/100, and speed at 84/100, all higher levels, while the power indicator was measured at 7/10. Based on an evaluation of the daily status data, the player's readiness was determined to be 5/10, fatigue 2/10, and mood 10/10. There is no data regarding muscle pain/sensitivity or stress. The findings suggest that the player generally exhibits low fatigue and high motivation levels, but there are areas that require improvement in certain technical parameters, such as using both feet and first touch.

Table 3. Actual session data and performance indicator for Player 2

Dataset	Indicator	Value
Physical performance	Session duration	53 min
	Total distance	4,4 km
	Working speed	82,5 m/min
	Sprint distance	31,8 m
	Maximum speed	22,7 km/h
Technical performance	Goal	1
	Assist	0
	Use of both feet	48/100
	Dribbling	43/100
	First touch	40/100
	Agility	89/100
	Speed	67/100
	Power	45/100
Daily status	Readiness	7/10
	Fatigue	6/10
	Mood	8/10
	Muscle pain/ sensitivity	No data
	Stress	No data

A review of Table 3 shows that the workout lasted 53 minutes, the total distance covered was 4,4 km, and the working speed was 82,5 m/min. The sprint distance was measured at 31,8 m, and the maximum speed was determined to be 22,7 km/h. In addition, it was noted that the player scored 1 goal and did not record any assists (0) during this training session. Technically, the player scored 48 out of 100 for both footwork, 43 for dribbling, and 40 for first-touch quality. The agility score reached the highest value at 89, while the speed score was 67 and the power score was 45. According to the daily status data, readiness was recorded as 7 out of 10, fatigue level as 6, and mood as 8. It was noted that no data was entered for muscle pain/sensitivity and stress indicators. In general, while these findings suggest that the player has a moderate level of physical training intensity, it can be stated that, from a technical standpoint, the player excels in agility but scores relatively low in areas such as first touch and power. Based on daily performance data, it can be concluded that the player's mood is good and their level of preparation is acceptable.

Table 4. Training plan generated by artificial intelligence for Player 1

Week	Training focus	Training activity and its purpose
Week 1	Weak foot + first touch	Passing with the non-dominant foot, control and first touch, using both feet, and improving the quality of the first touch.
	Acceleration + repeated sprints	Short accelerations from a standing start and repeated sprints to improve initial acceleration.
	Agility + turns	Changing direction and accelerating after a turn, minimizing the loss of speed that occurs during turns.
Week 2	Technical development	Improving first touch and weak-foot skills, and developing ball control.

	Speed and sprinting	Improving short-distance acceleration and sprinting ability.
	Quick Ball Handling	High-intensity running with the ball and changing direction, practicing ball-handling at higher speeds.
Week 3	Weak foot + ball control	Control and non-dominant foot drills under pressure, applying technical skills in challenging conditions.
	Repeated sprints	Increased sprint repetitions to improve the capacity for repeated high-intensity running.
	Acceleration out of turns	Short bursts of acceleration after turning; improving the ability to regain speed after a change of direction.
Week 4	Technical integration	First touch, using the weaker foot, and quick ball movement, all combined to reinforce areas for improvement.
	Acceleration + sprint	Standing start and repeated sprint drills to reinforce the first step and the ability to perform high-intensity movements.
	Position-specific training	Applications related to position requirements; linking technical and physical development to on-field tasks.

According to Table 4, the four-week training program generated by artificial intelligence was structured by taking into account the player's areas for improvement and current performance characteristics as identified in the CityPlay data. Given that the player's scores for both-footedness (43/100) and first touch (53/100) are lower compared to other technical indicators, the program appears to focus specifically on exercises for the weaker foot, first touch, and ball control. In response, even though the player has a high speed rating (84/100) and an agility rating of 64/100, the inclusion of acceleration, repeated sprints, changes of direction, and post-turn acceleration drills in the program indicates that the goal is to make the player's on-field movements more effective and repeatable, rather than to further develop the player's current physical capacity. The program's progression over the weeks, from basic technical skills to technical execution under pressure, high-intensity sprint training, and the integration of technical skills, demonstrates that the training load has been gradually increased. Furthermore, considering that the player recorded 4 assists and 0 goals in the performance data in question, it is assessed that the emphasis placed on quick ball movement and positional training last week was intended to enhance the player's effectiveness in attack and his performance regarding his responsibilities on the field. However, since the player's readiness level was assessed at 5/10 and his fatigue level at 2/10, it is necessary to monitor individual workload levels, particularly during high-intensity sprint and acceleration drills as part of the program.

Table 5. Training plan generated by artificial intelligence for Player 2

Week	Training focus	Training activity and its purpose
Week 1	Weak foot + first touch	Passing, control, and first-touch drills with the non-dominant foot; developing the use of both feet and improving the quality of the first touch.
	Agility + turns	Lane-change and exit maneuvers, minimizing the loss of speed that occurs during a turn.
	Center back position	Central line closures and support shifts; developing positioning and shifting behaviors on the defensive line.
Week 2	First touch + ball control	First touch and ball control under pressure, developing the ability to handle the ball more quickly and with greater control.
	Acceleration + recovery sprint	Standing starts, short sprints, and recovery runs, developing initial acceleration and the speed of transition to defense.

	Defensive reaction	Recovery sprints and 1v1 reaction drills, developing recovery in transition situations and 1-on-1 defensive reactions.
Week 3	Weak foot + passing	Passing and ball control with the non-dominant foot under pressure; developing the use of both feet.
	Repeated sprints	Sets of repeated sprints with rest intervals, development of the capacity for repeated high-intensity running.
	Small-sided game focused on center backs	Working on defensive positioning, defensive compactness, and decision-making behaviors in small-sided games.
Week 4	Fast ball movement	Cone dribbling and 1v1 ball-handling drills, practicing ball control at higher speeds, and reinforcing technical skills.
	Match-specific center back formations	Combining central line closures, support shifts, recovery sprints, and 1v1 reactions, reinforcing position-specific decision-making.
	Acceleration + repeated sprints	Acceleration sprints from a standing start and sets of repeated sprints, reassessment and reinforcement of the first step and high-intensity movement capacity.

According to Table 5, the four-week training program generated by artificial intelligence was specifically designed based on the technical, physical, and positional characteristics of the second player identified from CityPlay data. Since the player's scores for first touch (40/100), dribbling (43/100), and both-footedness (48/100) were particularly low, the program focuses on weak-foot, first-touch, and ball control drills starting from the first week. Conversely, the player's very high agility score (89/100) indicates that agility drills are not planned as a primary area for development, but rather to ensure more effective use of existing capacity during turns and defensive movements. Given that the player's speed was rated at 67/100 and his power at 45/100, and given that the measured sprint distance is 31,8 m and the maximum speed is 22,7 km/h, incorporating acceleration drills, recovery sprints, and repeated sprint training into the program is meaningful for developing physical capacity and, in particular, agility during defensive transitions. In addition, the fact that the AI associates the player with the center back position and plans position-specific drills, such as closing down the midfield, defensive support movements, defensive reactions, one-on-one defense, and games played with a small number of players, demonstrates that the training program focuses not only on individual performance metrics but also on the player's responsibilities on the field. Given the player's offensive contribution of 1 goal and 0 assists, the addition of a quick ball-carrying drill in Week 4 can be viewed as a complementary exercise aimed at developing the defender's ability to control the ball at high speeds and transition into attack. Finally, while the player's readiness level of 7/10 and mood of 8/10 paint a positive picture in terms of training participation, the fatigue level of 6/10 indicates that load management must be taken into account, particularly during repeated sprints and high-intensity training sessions.

DISCUSSION AND CONCLUSION

This study examines the AI-driven analysis of physical, technical, and daily performance data collected through the CityPlay system and the conversion of this data into player-specific training recommendations. The findings show that AI can identify players' strengths and areas for improvement and determine different training priorities tailored to these characteristics.

An analysis of the two players' data revealed that low technical performance indicators were directly reflected in their training programs. In addition, variables such as speed, agility, fatigue, fitness level,

and the player's position were taken into account when determining training content. This suggests that an AI-supported approach could contribute to player-specific, data-driven training planning, rather than a one-size-fits-all program. While the fact that the study is based solely on data from two players represents a significant limitation, it is recommended that the effectiveness of this approach be evaluated in the future using larger samples and long-term performance data.

References

Anthropic, 2026. Environment: Claude Sonnet 5 System Card. Anthropic. Access address: https://www-cdn.anthropic.com/73ad94ca3c0502e75e46637cc62c8bd9532a7f2c/Claude%20Sonnet%205%20System%20Card.pdf; Date of access: 16/08/2026.

Bourdon, P. C., Cardinale, M., Murray, A., Gastin, P., Kellmann, M., Varley, M. C., Gabbett, T. J., Coutts, A. J., Burgess, D. J., Gregson, W., & Cable, N. T. (2017). Monitoring athlete training loads: Consensus statement. International Journal of Sports Physiology and Performance, 12(s2), S2-161–S2-170. https://doi.org/10.1123/IJSPP.2017-0208

Claudino, J. G., Capanema, D. O., de Souza, T. V., Serrão, J. C., Machado Pereira, A. C., & Nassis, G. P. (2019). Current approaches to the use of artificial intelligence for injury risk assessment and performance prediction in team sports: A systematic review. Sports Medicine – Open, 5, 28. https://doi.org/10.1186/s40798-019-0202-3

Gurau, T. V., Gurau, G., Voinescu, D. C., Anghel, L., Onose, G., Iordan, D. A., Munteanu, C., Onu, I., & Musat, C. L. (2023). Epidemiology of injuries in men's professional and amateur football (Part I). Journal of Clinical Medicine, 12(17), 5569. https://doi.org/10.3390/jcm12175569

Playermarker, 2026. Environment: How It Works Playermarker. Access address: https://www.playermaker.com/pages/how-it-works; Date of access: 12/08/2026.

Rossi, A., Pappalardo, L., Cintia, P., Iaia, F. M., Fernàndez, J., & Medina, D. (2018). Effective injury forecasting in soccer with GPS training data and machine learning. PLOS ONE, 13(7), e0201264. https://doi.org/10.1371/journal.pone.0201264

Scott, M. T. U., Scott, T. J., & Kelly, V. G. (2016). The validity and reliability of global positioning systems in team sport: A brief review. Journal of Strength and Conditioning Research, 30(5), 1470–1490. https://doi.org/10.1519/JSC.0000000000001221

Sonesson, S., Lindblom, H., & Hägglund, M. (2023). Higher age and present injury at the start of the season are risk factors for in-season injury in amateur male and female football players—a prospective cohort study. Knee Surgery, Sports Traumatology, Arthroscopy, 31(10), 4618–4630. https://doi.org/10.1007/s00167-023-07517-6

Szymski, D., Krutsch, V., Achenbach, L., Gerling, S., Pfeifer, C., Alt, V., Krutsch, W., & Loose, O. (2022). Epidemiological analysis of injury occurrence and current prevention strategies on international amateur football level during the UEFA Regions Cup 2019. Archives of Orthopaedic and Trauma Surgery, 142(2), 271–280. https://doi.org/10.1007/s00402-021-03861-9

Acknowledgment

This study was supported by project number 1919B012473571 under the TÜBİTAK 2209-A Research Project Support Program for University Students. We would like to thank TÜBİTAK for its support.

Conflict of Interest

The authors have declared that there is no conflict of interest.

Author Contributions

The authors have contributed equally.

A Hybrid GA-SA Solver for the Two-Dimensional Cutting Stock Problem: A Benchmark Study of the Placement Decoder's Role

Cumali Türkmenoğlu[1], Murat İn[2], İsmail Karadağ[3]

[1] Fatih Sultan Mehmet Vakif University, Faculty of Engineering, Department of Software Engineering, 34000, İstanbul, Türkiye
[2] HIT Information Consulting Services Industry and Trade Inc., R&D Department, 34000, İstanbul, Türkiye

*Corresponding author e-mail: cturkmenoglu@fsm.edu.tr

Abstract

The two-dimensional cutting stock problem (2D-CSP) is an NP-hard combinatorial optimization problem that pervades the furniture, metal, glass and textile industries, where rectangular parts of different sizes must be cut from standard stock sheets while material waste is minimized. Because exact methods do not scale to the 50–200 part orders encountered in daily production, we develop a permutation-encoded genetic algorithm (GA) hybridized with simulated annealing (SA) and decoded by a maximal-rectangles (MAXRECTS) placement heuristic that fills cavities the classical bottom-left skyline decoder cannot reach. We evaluate the solver on 21 benchmark instances with known zero-waste optima (16–196 parts) using paired statistical tests. The hybrid solver reaches a mean material utilization of 96.4% and a mean deviation from the optimum of 3.8%, recovering the exact optimum on the smallest class and staying within 2.5–3.2% of the optimum on the largest 196-part instances. A controlled decoder ablation on the same benchmark shows that replacing the skyline decoder with MAXRECTS lowers the class-averaged gap from 6.4% to 3.8% and raises utilization by roughly four percentage points on the medium and large classes. This is a far larger effect than any operator or parameter tuning, confirming and quantifying, on known-optimum instances, the established observation (Burke et al., 2004, 2009) that the geometric representation is a dominant lever for solution quality. The SA hybridization yields a small but consistent additional reduction in waste over the pure GA (0.11 percentage points overall; paired $t = 1.93$, $p \approx 0.07$) whose magnitude grows monotonically with problem size, confirming that its value is greatest on the largest instances.

Keywords: *Two-dimensional cutting stock, strip packing, maximal rectangles, genetic algorithm, simulated annealing, memetic algorithm, benchmark evaluation, waste minimization.*

INTRODUCTION

The cutting stock problem (CSP) asks how to cut a set of demanded items from standard stock material so that waste is minimized. In its two-dimensional rectangular form (2D-CSP), items and stock sheets are rectangles, and the problem arises whenever sheet materials such as wood panels, sheet metal, glass, or fabric are divided into parts. It is a cornerstone of industrial production planning and a classical NP-hard combinatorial optimization problem (Martello and Vigo, 1998; Lodi et al., 2002). In practice, waste

rates of 10–20% are common and cutting plans are still often produced manually or by simple heuristics, so even modest improvements in material utilization translate into substantial cost and sustainability gains.

Exact approaches provide optimality guarantees and are well developed: column generation in the tradition of Gilmore and Gomory (1961, 1965), branch-and-bound and branch-and-price for two-dimensional bin packing (Martello and Vigo, 1998), and the more recent exact techniques surveyed by Iori et al. (2021). However, the number of feasible cutting patterns grows combinatorially, so these methods do not scale to the 50–200 part orders that a factory processes every day. Consequently, population-based and local-search metaheuristics have become the practical tool of choice (Jakobs, 1996; Burke et al., 2004; Bortfeldt, 2006), frequently combined into hybrid (memetic) schemes (Soke and Bingul, 2006; Burke et al., 2009).

This paper studies a hybrid genetic algorithm–simulated annealing (GA-SA) solver for the 2D-CSP. Beyond describing the solver, our central concern is empirical rigour: we ask how good the solver really is against instances whose optima are known, and, more importantly, which component actually governs solution quality. Our contributions are:

A permutation-encoded memetic GA-SA solver whose permutations are decoded by a maximal-rectangles (MAXRECTS) placement heuristic (Jylänki, 2010), with order crossover, swap mutation, tournament selection, elitism, and an insertion/reversal/swap SA local search.

An external, reproducible validation on 21 benchmark instances with known zero-waste optima spanning seven size classes (16–196 parts), built following the Hopper–Turton methodology (Hopper and Turton, 2001), using paired t-tests and Wilcoxon signed-rank tests.

A controlled decoder ablation on known-optimum instances that confirms and quantifies the established observation (Burke et al., 2004, 2009) that the geometric representation, not the metaheuristic, dominates quality: switching from a bottom-left skyline decoder to MAXRECTS nearly halves the optimality gap, an effect far larger than operator/parameter tuning.

Two supporting experiments that characterize the solver as an engineering artefact: an anytime/convergence analysis of the SA contribution, and an analysis of the free-rectangle cap as a speed/quality knob (including its failure mode at very low caps).

A four-decoder sensitivity study (skyline vs. three MAXRECTS selection rules) under a fixed metaheuristic, which generalizes the decoder-dominance finding.

Evidence that the SA hybridization delivers a small but consistent improvement whose magnitude increases monotonically with problem size.

RELATED WORK

Wäscher et al. (2007) provide the standard typology of cutting and packing problems; the variant addressed here corresponds to a two-dimensional single-stock-size cutting/open-dimension problem. Surveys by Lodi et al. (2002) and, more recently, Iori et al. (2021) map the landscape of models and exact/relaxation techniques. Foundational complexity and approximation results for orthogonal packing are due to Baker et al. (1980), and guillotine cutting via dynamic programming to Beasley (1985); Martello and Vigo (1998) give an exact branch-and-bound for two-dimensional finite bin packing.

On the heuristic side, the placement decoder is central. The bottom-left rule and its efficient implementation by Chazelle (1983) place each item as low and as far left as possible but cannot use cavities beneath overhangs; the best-fit heuristic of Burke et al. (2004) and the maximal-rectangles family catalogued by Jylänki (2010) overcome this by explicitly tracking free space. Metaheuristics on top of such decoders include genetic algorithms (Jakobs, 1996; Bortfeldt, 2006), and hybrids with local search. Most relevantly, Soke and Bingul (2006) combine a GA and SA with an improved bottom-left decoder for non-guillotine 2D packing, and Burke et al. (2009) enhance the best-fit heuristic with SA. The benchmark of Hopper and Turton (2001) remains a reference point for empirical comparison. Most recently, learning-based methods have appeared, including deep reinforcement learning for rectangular strip packing (Fang et al., 2023) and broad reviews of machine learning for bin packing (González-San-Martín et al., 2024).

One strand is directly relevant to this paper: the geometric representation. The skyline family of decoders (Wei et al., 2011) and the maximal-rectangles family (Jylänki, 2010) are known to outperform plain bottom-left, and a well-designed placement heuristic can even rival metaheuristic search (Burke et al., 2004, 2009); however, comparisons usually vary the decoder and the search simultaneously. We instead hold the metaheuristic fixed and vary only the decoder on instances with known optima, turning an informal observation into a controlled measurement. Our work therefore differs from the literature in its emphasis: rather than proposing yet another operator, we isolate, on instances with known optima, the relative contribution of the decoder and of the SA hybridization, and generalize the decoder finding across four decoders, reporting all results honestly, including where the hybridization does not help.

PROBLEM DEFINITION

We consider the two-dimensional rectangular strip-packing form of the 2D-CSP. We are given a stock sheet of fixed width W and (conceptually) unbounded height, and a list of n rectangular parts, part i having width $w_i \leq W$ and height h_i. A feasible cutting plan places every part without overlap, with edges parallel to the sheet edges (orthogonal, fixed orientation), entirely within the strip. Let H be the maximum height reached by any placed part. The objective is to minimize H, equivalently to minimize trim loss $W \cdot H - \Sigma\, w_i h_i$ and to maximize material utilization $U = \Sigma\, w_i h_i / (W \cdot H)$. Because the total part area is fixed, minimizing H is equivalent to minimizing waste; when the parts can tile a $W \times H_0$ rectangle exactly, H_0 is the optimum and U = 100%. The problem is NP-hard (Martello and Vigo, 1998; Lodi et al., 2002); the number of feasible placements grows combinatorially with n, motivating a metaheuristic search over placement orders rather than over placements directly.

PROPOSED METHOD

Solution representation

A solution is encoded as a permutation π of the n parts, interpreted as the order in which parts are fed to a deterministic placement decoder. This indirect encoding keeps the search space free of infeasible geometric states and lets standard order-based genetic operators apply directly. Permutation (order-based) encodings are a standard and effective choice for packing problems solved through a placement decoder (Jakobs, 1996; Bortfeldt, 2006), and we adopt it here.

MAXRECTS placement decoder

The decoder converts a permutation into a physical layout. We use the maximal-rectangles (MAXRECTS) scheme (Jylänki, 2010), which maintains the set of maximal empty rectangles of the strip. For each part in permutation order we select, among the free rectangles in which the part fits, the placement that keeps the strip lowest, breaking ties by best-short-side-fit (BSSF); the part is placed at the bottom-left of the chosen rectangle, and every free rectangle overlapping the placed part is split into its maximal residual sub-rectangles, after which contained rectangles are pruned. Unlike a bottom-left skyline decoder, MAXRECTS can place parts into cavities and beneath overhangs, which is the single most important determinant of solution quality. To bound the per-decode cost for large n, an optional cap limits the number of retained free rectangles.

Genetic operators

The GA evolves a population of permutations. Selection is by tournament ($k = 3$) with a 5% elitist carry-over of the best individuals. Recombination uses order crossover (OX), which preserves relative order while producing valid permutations, at rate 0.85; mutation swaps two positions at rate 0.02 per gene. The initial population is seeded with area- and height-sorted permutations (which are strong first-fit orders) plus random permutations for diversity. These operator choices and rates follow established practice for permutation GAs and were confirmed on pilot instances.

Simulated-annealing local search (memetic hybridization)

To intensify promising solutions, the GA is hybridized with a simulated-annealing local search (Kirkpatrick et al., 1983) applied to the incumbent best every few generations and, optionally, as a final polish. The SA neighbourhood mixes three moves, namely insertion (eject a part and reinsert it elsewhere), segment reversal, and pairwise swap, because insertion and reversal perturb the decoded layout far more than a single swap and are essential at large n. Candidate moves are accepted by the Metropolis criterion with a geometric cooling schedule. This memetic design follows the GA+SA lineage of Soke and Bingul (2006) and Burke et al. (2009) but is coupled to a MAXRECTS rather than a bottom-left decoder.

Rather than a single parameter set, we use a Taguchi L9 orthogonal-array design of experiments to derive reusable parameter profiles for small, medium and large problem sizes, so that the solver selects a profile from the instance size without instance-specific manual tuning.

To test whether the geometric representation, rather than the search, governs quality, we implement four decoders behind one interface and drive them with the identical GA-SA: (i) a bottom-left skyline decoder (Wei et al., 2011) that stacks parts on the top envelope and cannot use cavities; and three MAXRECTS variants that differ only in how they select among the free rectangles in which a part fits, namely (ii) bottom-left (lowest, then leftmost), (iii) best-short-side-fit (BSSF), and (iv) best-area-fit (BAF). Comparing these under a fixed metaheuristic isolates the effect of the representation.

EXPERIMENTAL SETUP

Benchmark with known optima

To measure the true optimality gap we use instances whose optima are known exactly. Following the Hopper–Turton construction (Hopper and Turton, 2001), a $W \times H_0$ rectangle is recursively guillotine-partitioned into n parts that tile it exactly; the optimum strip height is therefore H_0 with 0% waste. We generate three instances for each of the seven Hopper–Turton size classes C1–C7, giving 21 instances spanning 16 to 196 parts (Table 1). All experiments are fully reproducible from fixed random seeds.

We report the deviation (gap) from the optimum, $(H - H_0)/H_0 \times 100\%$, the material utilization U, and the absolute trim-loss area $W \cdot H - \Sigma\, w_i h_i$. Reporting utilization and absolute trim alongside the percentage gap avoids the scale distortion that the small optimal heights of the low classes would otherwise introduce.

Each instance is solved over multiple independent runs (up to 30). The hybrid GA-SA is compared against the pure GA under an identical evolutionary budget using common random seeds (a paired design), so that the contribution of the SA local search is isolated. Significance is assessed with paired t-tests and Wilcoxon signed-rank tests computed at the instance level over the 21 instances, which is the standard unit of analysis for packing benchmarks and is robust to differing per-class run counts.

RESULTS AND DISCUSSION

Absolute solution quality

Table 1 summarizes solution quality by size class. The hybrid solver attains a mean material utilization of 96.4% and a mean gap of 3.8% across all 21 instances. It recovers the exact optimum (0% waste) on the smallest class C1 in every run, and its gap decreases with problem size, from 5.8% on C2 (25 parts) to 2.9% on C7 (196 parts).

Table 1. Solution quality by problem-size class (hybrid GA-SA).

Class	n	Inst.	Runs	GA gap %	GA-SA gap %	GA-SA best %	GA-SA util %	Waste red. (pp)
C1	16	3	30	0.00	0.00	0.00	100.00	+0.00
C2	25	3	30	5.63	5.78	0.00	94.58	-0.14
C3	28	3	30	5.63	5.19	3.33	95.10	+0.39
C4	49	3	25	4.27	4.24	1.67	95.94	+0.02
C5	73	3	15	4.49	4.42	2.22	95.78	+0.07
C6	97	3	12	4.21	4.05	2.50	96.11	+0.15
C7	196	3	3	3.15	2.87	2.50	97.21	+0.26

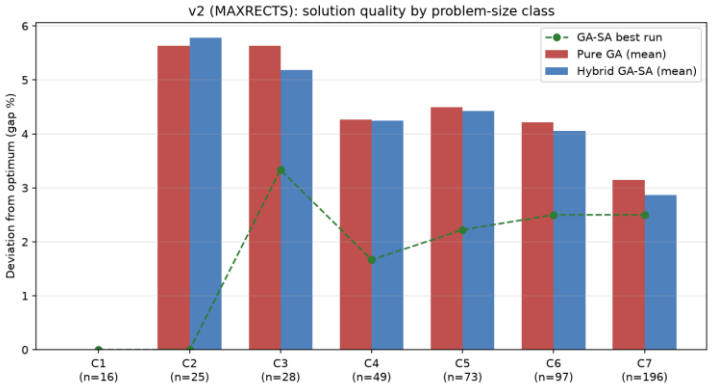


Figure 1. Mean gap of pure GA and hybrid GA-SA, and the best GA-SA run, per class.

This trend is geometric rather than paradoxical: small instances contain few, large parts and a small optimal height, so any unavoidable imperfection is a large percentage and there are few small parts to fill residual gaps; large instances offer abundant filler parts and a large optimal height, enabling tighter relative packings. Figure 1 shows mean gap of deviation from optimum, Figure 2 utilization by class and Figure 4 a decoded cutting plan.

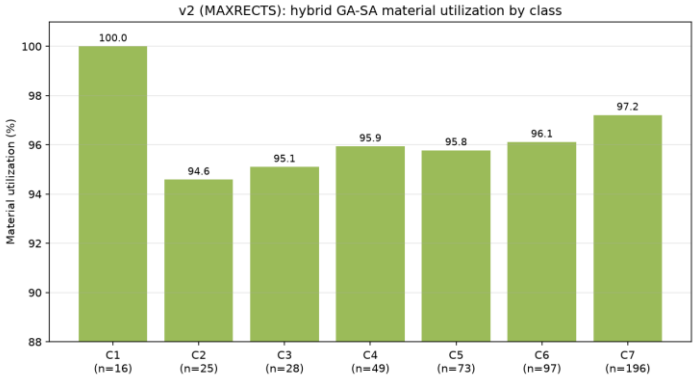


Figure 2. Hybrid GA-SA material utilization by class.

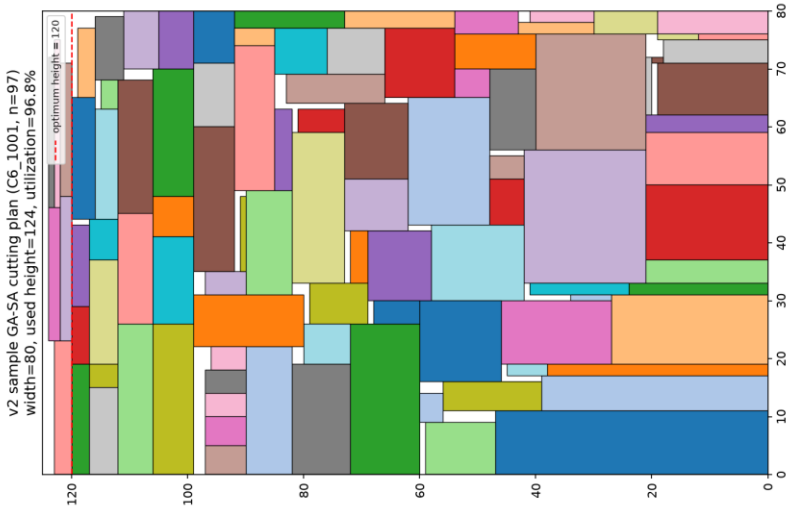


Figure 3. A decoded hybrid GA-SA cutting plan (class C6, n=97); dashed line = optimum height.

Decoder ablation: quantifying the representation effect

To isolate the effect of the geometric representation, we ran the identical GA-SA solver on the identical benchmark with two decoders: a bottom-left skyline decoder and the MAXRECTS decoder. Table 2 and Figure 4 report the result. Averaged over the seven classes, replacing the skyline decoder with MAXRECTS lowers the GA-SA gap from 6.41% to 3.79% and raises utilization by roughly four percentage points on the medium and large classes (e.g., class C4 improves from 8.6% to 4.2% gap, and C7 from 6.8% to 2.9%). This single change improves quality by an order of magnitude more than any operator or parameter adjustment we examined. The reason is structural: the skyline decoder cannot place parts into cavities beneath previously placed parts, so both the pure GA and the GA-SA plateaued near a decoder-imposed ceiling regardless of search effort; MAXRECTS removes that ceiling. This is consistent with prior findings that a strong placement heuristic drives quality and can rival metaheuristic search (Burke et al., 2004, 2009); our contribution here is to confirm and quantify the effect in a controlled setting, holding the metaheuristic fixed and measuring against known optima, where engineering the placement representation clearly outweighs tuning the metaheuristic for the 2D-CSP.

Table 2. Decoder ablation: bottom-left skyline vs. MAXRECTS, hybrid GA-SA.

Class	n	v1 skyline gap %	v2 MAXRECTS gap %	v1 util %	v2 util %	Gap improvement (pp)
C1	16	0.06	0.00	99.95	100.00	+0.06
C2	25	6.30	5.78	94.10	94.58	+0.52
C3	28	6.81	5.19	93.65	95.10	+1.62
C4	49	8.58	4.24	92.12	95.94	+4.34
C5	73	8.30	4.42	92.36	95.78	+3.88
C6	97	8.00	4.05	92.60	96.11	+3.95
C7	196	6.83	2.87	93.61	97.21	+3.96

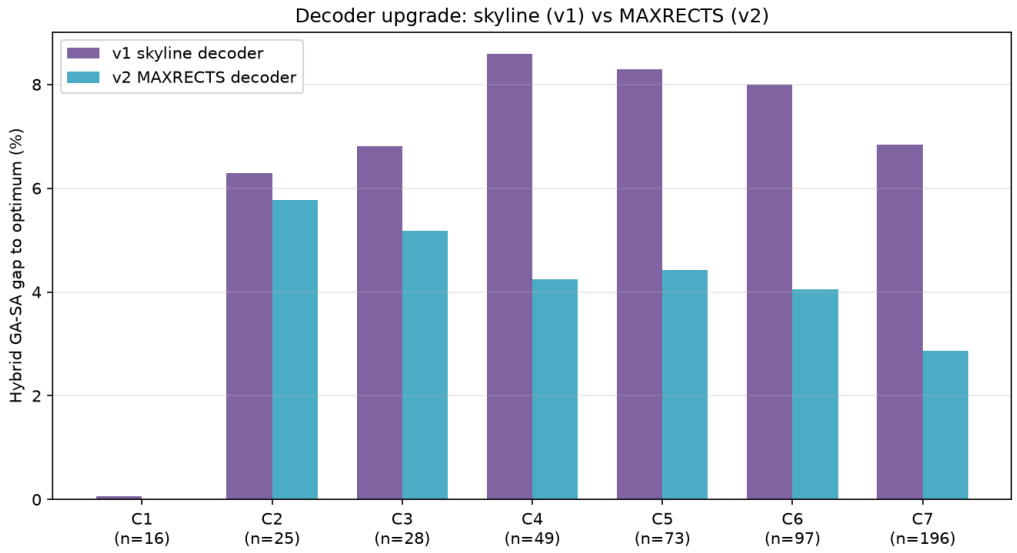


Figure 4. Decoder ablation: skyline vs. MAXRECTS GA-SA gap per class.

Effect of the SA hybridization and its scaling with size

With the strong MAXRECTS decoder, the pure GA already approaches the achievable quality, so the additional benefit of the SA local search is modest: over the 21 instances the hybrid reduces waste by 0.11 percentage points on average (paired $t = 1.93$, $p \approx 0.07$; Wilcoxon $p \approx 0.10$). The effect is, however, consistent in direction and, importantly, grows monotonically with problem size across the medium and large classes: +0.02, +0.07, +0.15 and +0.26 percentage points for C4–C7 respectively, with the medium/large pooled reduction of 0.12 pp approaching significance ($p \approx 0.05$). The interpretation is that SA acts as a robustness/intensification mechanism whose value is greatest precisely on the large, hard instances, while on small instances a well-decoded GA already finds the optimum and leaves no room for improvement. This is a more honest and more useful conclusion than an unqualified claim of superiority.

Convergence behaviour

Figure 5 traces the best-so-far gap against generation on a representative medium instance (C4, n = 49, mean of 8 runs), with the final SA polish disabled so that the per-generation contribution is visible. The hybrid curve descends faster and reaches a lower plateau earlier than the pure GA, illustrating that the SA local search accelerates convergence rather than merely improving the endpoint (both methods approach a similar final value here, 3.33% vs 3.33%, because the instance is small enough for the GA to catch up given enough generations). Under a fixed computational budget, faster convergence is itself an advantage.

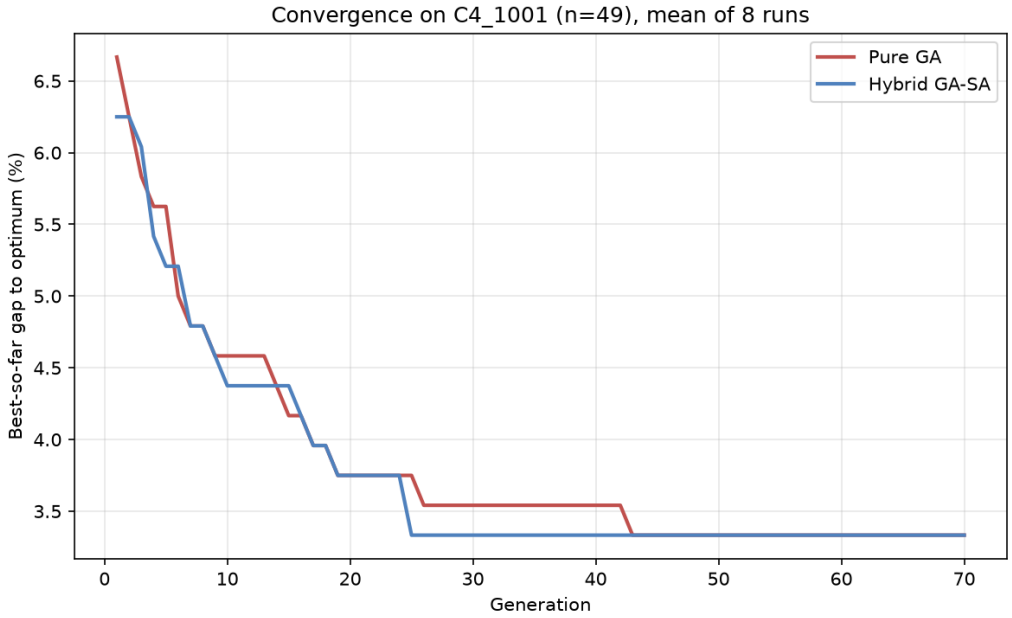


Figure 5. Anytime convergence of pure GA vs. hybrid GA-SA (C4, n=49).

Free-rectangle cap: a speed/quality knob

MAXRECTS keeps a list of maximal free rectangles whose size grows with n; capping this list bounds the per-decode cost. Figure 6 sweeps the cap on class C6 (n = 97). Quality is stable and near-best for caps of about 60 and above, while runtime rises roughly linearly with the cap; very small caps, however, discard rectangles that matter and cause the packing to collapse (e.g., a cap of 20 yields a 32% gap). A cap near 60 offers the best quality/speed compromise for this size; for n ≈ 200 the decoder is best left effectively uncapped. This characterizes the speed/quality trade-off of the free-rectangle cap in MAXRECTS-based solvers.

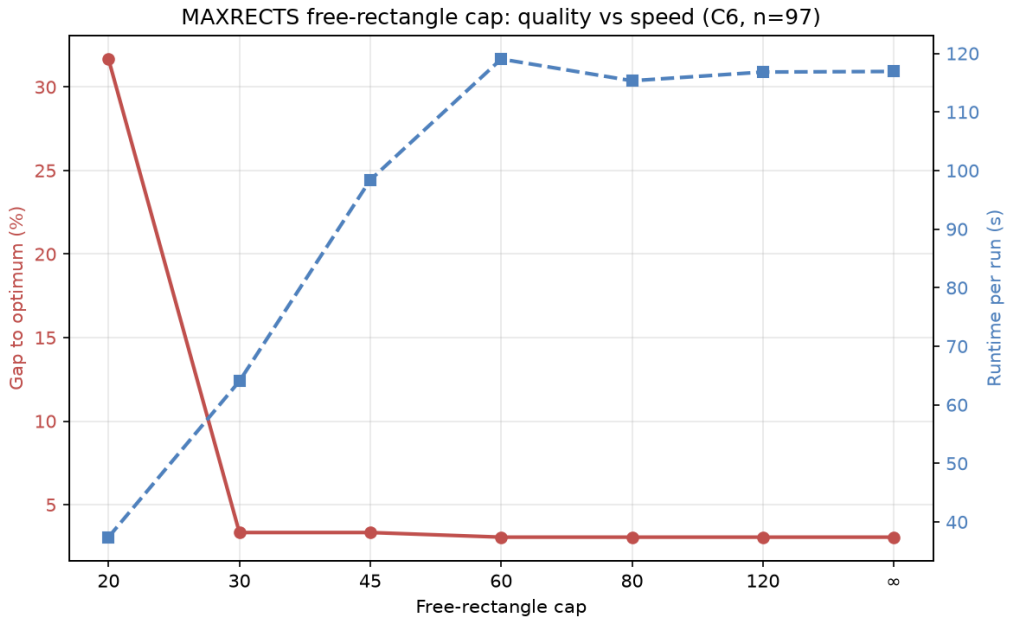


Figure 6. Free-rectangle cap: gap and runtime trade-off (C6, n=97).

Decoder sensitivity: generalizing the representation effect

We compared two decoders across the whole benchmark. To check that the effect is not specific to a single MAXRECTS rule, we ran the identical GA-SA with four decoders (a bottom-left skyline and MAXRECTS with the bottom-left, best-short-side-fit and best-area-fit selection rules) on a small, a medium and a large instance (Figure 7). On the small instance (C2, 25 parts) all decoders coincide at the granularity-imposed floor. On the medium and large instances the skyline decoder is markedly worse than any MAXRECTS rule: on C6 (97 parts) the skyline gap is 8.6% whereas all three MAXRECTS rules reach 3.3–4.2%, and on C5 (73 parts) the skyline gap of 10.4% falls to about 5.6%. The differences among the three MAXRECTS rules are small (well below one percentage point), whereas the gap between the skyline family and the MAXRECTS family is large. In other words, the crucial choice is whether the representation can fill cavities at all; the fine-grained selection rule is secondary, and the metaheuristic is largely decoder-agnostic. This generalizes the conclusion of this study.

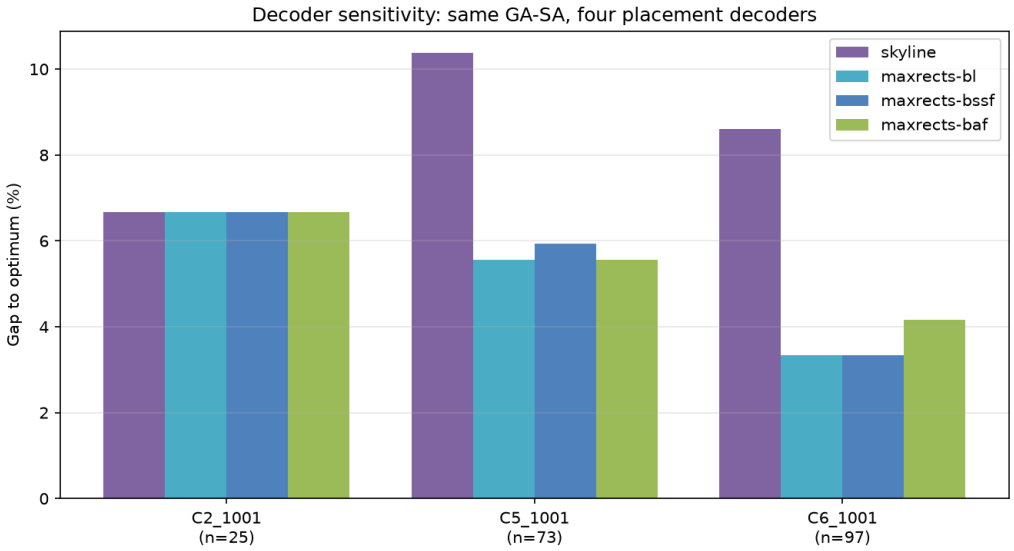


Figure 7. Decoder sensitivity: identical GA-SA with four placement decoders.

Our benchmark instances are guillotine-tileable with known optima; while this enables exact gap measurement, real orders may include non-guillotine demand and rotation, which we fix to a single orientation. Statistical power at the instance level is modest (21 paired observations, 12 for medium/large), so the small SA effect is reported as suggestive rather than conclusive. Finally, the largest class was solved under a reduced budget for tractability, which makes its absolute figures conservative.

DISCUSSION AND CONCLUSION

We presented a permutation-encoded memetic GA-SA solver for the two-dimensional cutting stock problem, decoded by a maximal-rectangles placement heuristic and validated on 21 benchmark instances with known optima. The solver reaches 96.4% mean utilization and a 3.8% mean gap, is exact on small instances and within ~3% on 196-part instances. Our main message is methodological: a controlled ablation on known-optimum instances confirms and quantifies the established observation (Burke et al., 2004, 2009) that the placement decoder (the geometric representation) is a dominant lever for quality, far outweighing operator and parameter choices; a four-decoder sensitivity study confirms that what matters is the ability to fill cavities (skyline vs. MAXRECTS), not the specific MAXRECTS selection rule, and that the metaheuristic is largely decoder-agnostic. The SA hybridization contributes a small, consistent, size-increasing improvement best understood as robustness at scale. Future work will compare against exact lower bounds and published Hopper–Turton results, extend the decoder to rotation and guillotine-constrained variants, and benchmark against learning-based packers (Fang et al., 2023; González-San-Martín et al., 2024).

References

Baker BS, Coffman EG, Rivest RL, 1980. Orthogonal packings in two dimensions. SIAM Journal on Computing, 9 (4): 846-855.

Beasley JE, 1985. Algorithms for unconstrained two-dimensional guillotine cutting. Journal of the Operational Research Society, 36 (4): 297-306.

Bortfeldt A, 2006. A genetic algorithm for the two-dimensional strip packing problem with rectangular pieces. European Journal of Operational Research, 172 (3): 814-837.

Burke EK, Kendall G, Whitwell G, 2004. A new placement heuristic for the orthogonal stock-cutting problem. Operations Research, 52 (4): 655-671.

Burke EK, Kendall G, Whitwell G, 2009. A simulated annealing enhancement of the best-fit heuristic for the orthogonal stock-cutting problem. INFORMS Journal on Computing, 21 (3): 505-516.

Chazelle B, 1983. The bottom-left bin-packing heuristic: an efficient implementation. IEEE Transactions on Computers, C-32 (8): 697-707.

Fang J, Rao Y, Shi M, 2023. A deep reinforcement learning algorithm for the rectangular strip packing problem. PLOS ONE, 18 (3): e0282598.

Gilmore PC, Gomory RE, 1961. A linear programming approach to the cutting-stock problem. Operations Research, 9 (6): 849-859.

Gilmore PC, Gomory RE, 1965. Multistage cutting stock problems of two and more dimensions. Operations Research, 13 (1): 94-120.

González-San-Martín J, Cruz-Reyes L, Dorronsoro B, et al., 2024. Deep study on the application of machine learning in bin packing problems. Computación y Sistemas, 28 (3): 1275-1290.

Hopper E, Turton BCH, 2001. An empirical investigation of meta-heuristic and heuristic algorithms for a 2D packing problem. European Journal of Operational Research, 128 (1): 34-57.

Iori M, de Lima VL, Martello S, Miyazawa FK, Monaci M, 2021. Exact solution techniques for two-dimensional cutting and packing. European Journal of Operational Research, 289 (2): 399-415.

Jakobs S, 1996. On genetic algorithms for the packing of polygons. European Journal of Operational Research, 88 (1): 165-181.

Jylänki J, 2010. A thousand ways to pack the bin – a practical approach to two-dimensional rectangle bin packing. Technical report, University of Helsinki, Helsinki, Finland.

Kirkpatrick S, Gelatt CD, Vecchi MP, 1983. Optimization by simulated annealing. Science, 220 (4598): 671-680.

Lodi A, Martello S, Monaci M, 2002. Two-dimensional packing problems: a survey. European Journal of Operational Research, 141 (2): 241-252.

Martello S, Vigo D, 1998. Exact solution of the two-dimensional finite bin packing problem. Management Science, 44 (3): 388-399.

Soke A, Bingul Z, 2006. Hybrid genetic algorithm and simulated annealing for two-dimensional non-guillotine rectangular packing problems. Engineering Applications of Artificial Intelligence, 19 (5): 557-567.

Wäscher G, Haußner H, Schumann H, 2007. An improved typology of cutting and packing problems. European Journal of Operational Research, 183 (3): 1109-1130.

Wei L, Oon WC, Zhu W, Lim A, 2011. A skyline heuristic for the 2D rectangular packing and strip packing problems. European Journal of Operational Research, 215 (2): 337-346.

Designing AI-Supported Authentic Telecollaboration for Pre-Service EFL Teachers' Digital Pedagogical Competence: A Design-Based Research Study

Toochi P. Farinola[1*], Ildeniz Ozverir[2]

[1]Eastern Mediterranean University, Faculty of Education, Department of Computer and Instructional Technologies, Famagusta, via Mersin 10, Turkiye, North Cyprus
[21]Eastern Mediterranean University, Faculty of Education, Department of Computer and Instructional Technologies, Famagusta, via Mersin 10, Turkiye, North Cyprus

*Corresponding author e-mail: toochifarinola@gau.edu.tr

Abstract

Generative Artificial Intelligence (GenAI) is increasingly integrated into educational environments, yet its educational value depends on how learners are guided to evaluate, adapt, and pedagogically justify AI-generated outputs. This study examines how to intentionally integrate GenAI into an authentic, collaborative teacher-education environment and identifies design principles for its effective educational use. A qualitative Design-Based Research (DBR) approach was conducted across two iterative six-week cycles involving 21 pre-service EFL teachers from two higher-education institutions in Northern Cyprus. Participants used ChatGPT-4o in collaborative lesson-design activities requiring them to address authentic professional teaching problems. Data were collected through semi-structured interviews, reflective journals, researcher field notes, and instructional artefacts and were analyzed using reflexive thematic analysis. Findings identified four interconnected mechanisms underlying the AI-supported learning environment: authentic professional engagement, collaborative knowledge construction, AI-mediated pedagogical scaffolding, and reflective pedagogical reasoning through iterative redesign. Based on the iterative DBR process, five empirically grounded design principles were developed: (1) authentic professional tasks, (2) structured telecollaboration, (3) critical evaluation of AI-generated outputs, (4) embedded reflection and redesign, and (5) coherent integration of AI, collaboration, and pedagogy. The findings indicate that effective educational integration of GenAI depends not simply on access to AI tools, but on the technological and pedagogical conditions that structure their use. The study contributes an empirically grounded design framework for integrating GenAI into technology-supported teacher education and demonstrates how AI can function as a mediating resource for collaborative inquiry, pedagogical reasoning, and professional learning.

Keywords: *Generative AI, Design-Based Research, Telecollaboration, Authentic Learning, Pre-service Teacher Education*

INTRODUCTION

Generative artificial intelligence (GenAI) is changing what pre-service teachers need to know and do. Contemporary systems can generate lesson plans, instructional materials, assessment tasks, feedback,

explanations, and alternative teaching strategies, creating new possibilities for instructional design while also raising concerns about overreliance, accuracy, ethics, and the displacement of professional judgment (Kasneci et al., 2023; Moorhouse et al., 2024; UNESCO, 2024). Teacher education therefore faces a dual task: preparing future teachers to use emerging AI capabilities while developing the judgment required to decide when, why, and how AI should be used.

This issue is especially consequential in EFL teacher education. Effective language teaching requires contextual decisions about learner proficiency, communicative purposes, linguistic difficulty, cultural relevance, interaction, assessment, and instructional sequencing. An AI system may generate a plausible activity, but plausibility does not establish pedagogical appropriateness. Recent research with pre-service language teachers similarly suggests that GenAI literacy involves critical, ethical, creative, and pedagogical capacities in addition to technical familiarity (Huang et al., 2025; Moorhouse et al., 2024; Peng, 2026). The central question is consequently shifting from whether teachers can access AI to how learning environments can help them exercise professional judgment with AI.

Authentic learning offers a useful grounding for this change as it situates learners in realistic professional activity that requires contextual knowledge, collaboration, reflection and the production of meaningful outcomes (Herrington & Oliver, 2000; Herrington et al., 2010). Authentic tasks in teacher education include designing lessons, evaluating resources, responding to instructional problems, and assessing learning and justifying pedagogical choices. Such tasks do not allow AI to simply do the work for the learner, requiring decisions that are dependent upon knowledge of learners and contexts. Telecollaboration brings a social dimension to the learning process by linking learners from different institutions and allowing them to negotiate, critique and refine professional decisions through digitally mediated interaction (O'Dowd, 2018; O'Dowd & Dooly, 2020).

Recent empirical work points to the value of this combination but has mostly studied the components separately. Research has shown that structured interventions using generative AI improve AI literacy for pre-service language teachers (Huang et al., 2025), that the collaborative use of generative artificial intelligence can assist with lesson planning with scaffolding (Sun & Huang, 2025), and that AI-specific pedagogical knowledge is linked to more advanced prompting and adaptive lesson design (Celik et al., 2026a). Research into agency also indicates that interacting with AI does not automatically lead to critical judgment: pre-service teachers may accept AI suggestions without adequately questioning them (Lee et al., 2025; Yang & Markauskaite, 2025; Celik et al., 2026b). Other recent studies embedding GenAI in practicum or lesson-study contexts have highlighted the value of authentic professional activity and iterative reflection (Kılıçkaya & Kic-Drgas, 2025).

The gap is thus more specific than the claim that there is a lack of research on GenAI in teacher education. The evidence base is rapidly growing with a recent systematic review of 20 empirical studies of GenAI in pre-service EFL/ESL teacher education (Peng, 2026). What is less clear is how GenAI, authentic professional activity, telecollaboration and digital pedagogical competence can be orchestrated in one iterative learning design intentionally. In particular, little is known about the design of collaborative structures and authentic tasks that can constrain unproductive dependence on AI, while providing opportunities for critique, negotiation and professional reasoning.

Design-Based Research (DBR) is an appropriate approach to tackle this problem because it treats educational innovation as an iterative process where theoretical principles, evidence of implementation

and refinement of design inform each other (Design-Based Research Collective, 2003; Reeves, 2006; McKenney & Reeves, 2019). Rather than testing a fixed intervention alone, DBR can reveal which design features become problematic in practice, how they are modified, and what design principles emerge from the process. The present study therefore examines how an AI-supported authentic telecollaborative learning environment can support increasingly sophisticated digital pedagogical engagement among pre-service EFL teachers.

The study addresses two research questions:

RQ1: How do pre-service EFL teachers experience and enact digital pedagogical competence within an AI-supported authentic telecollaborative learning environment?

RQ2: What design principles emerge from the iterative DBR process for integrating GenAI into authentic telecollaborative teacher education?

The study makes two related contributions. First, it provides an empirical account of how GenAI becomes pedagogically productive when embedded in authentic professional activity and structured peer interaction. Second, it translates the cross-cycle evidence into five design principles that specify conditions for AI-supported teacher learning. The contribution is therefore design-oriented rather than an estimate of a general causal effect.

LİTERATURE REVİEW AND CONCEPTUAL FRAMEWORK

From GenAI Adoption to Pedagogical Integration

GenAI has shifted teacher-education research from questions of access and acceptance to questions of professional knowledge and pedagogical judgment. New research has found that language teachers-in-training can cultivate their GenAI literacy along various lines of development through proper preparation, and the literature suggests an accelerated rise in empirical work in the field of EFL/ESL teacher education (Huang et al., 2025; Peng, 2026; Uygun & Demir, 2026). It has been concluded that technical skills are not enough; rather, language teachers have to assess AI-produced information, understand its limits and biases, create purposeful prompts, utilize AI creatively, and make informed decisions ethically (Moorhouse et al., 2024; Peng, 2026).

GenAI, Lesson Planning, and Pedagogical Reasoning

Lesson planning provides an effective context for investigating pedagogical reasoning with the support of AI because the teacher is required to link the learning objectives, learner attributes, subject matter, methods of teaching, assessment, and technology (Mishra & Koehler, 2006; Shulman, 1986). As demonstrated by Sun and Huang (2025), the use of GenAI to collaborate on lesson planning differs at various planning phases and depends on the type of scaffolding provided. Celik et al. (2026a) also reveal that technological and pedagogical knowledge relating specifically to AI is associated with prompting and adaptive lesson planning.

Teacher Agency and Human–AI Decision-Making

Agentivity of teachers is important since fluent AI-generated output may become authoritative in cases where it may be inaccurate or ill-suited to the context. According to Lee et al. (2025), the need for agentive uses of AI requires teachers to act according to professional goals and intentions rather than

merely utilizing the technology. The study by Yang and Markauskaite (2025) revealed that structured collaboration activities involving assessment, prompt revision, and reflective activity can facilitate the transformative agency of language student teachers. On the other hand, evidence at the process level reveals that interaction with AI does not ensure critical agency, as learners can concur with recommendations provided by AI without proper criticism (Celik et al., 2026b).

Authentic Learning and Telecollaborative Knowledge Construction

Realistic learning situates professional reasoning into realistic problems and contexts (Herrington & Oliver, 2000; Herrington et al., 2010). Authenticity has recently been used to embed GenAI into practicum, lesson study, and collaborative design tasks, instead of prompting activities alone, and thus indicating that authenticity could be used to tie AI application to professional purposes (Kılıçkaya & Kic-Drgas, 2025; Yang & Markauskaite, 2025). Telecollaboration involves socially mediated inquiry where individuals can compare AI-generated possibilities, question the assumptions behind them, and provide contextual knowledge that an AI system cannot generate (O'Dowd, 2018; O'Dowd & Dooly, 2020). This implies an important difference between collaborating with AI (working with an AI to produce products) and collaborating about AI (evaluating AI-generated possibilities).

Digital Pedagogical Competence in the GenAI Era

TPACK conceptualizes effective technology integration as an interaction among technological, pedagogical, and content knowledge (Mishra & Koehler, 2006), while DigCompEdu emphasizes educators' capacity to use digital resources, teaching and learning processes, assessment, learner empowerment, and professional development (Redecker, 2017). GenAI adds demands that are less visible in conventional technology-integration models: evaluating probabilistic outputs, detecting bias and hallucination, designing prompts, deciding the appropriate level of AI involvement, and retaining responsibility for pedagogical choices. Recent work on AI-specific or Intelligent-TPACK similarly emphasizes the need to integrate technological, pedagogical, content, and ethical knowledge (Celik et al., 2026a). In this study, digital pedagogical competence is therefore treated as enacted capability: the capacity to use, evaluate, adapt, and justify digital and AI-supported resources in relation to learners, objectives, context, and pedagogy.

Integrative Framework and Research Gap

From the literature, there are complementary roles rather than overlapping constructs: authentic learning provides meaningful professional problems; telecollaboration provides social negotiation; GenAI provides possibilities in terms of generation and information; TPACK and digital competence frameworks provide perspectives for technology-pedagogy integration; and teacher agency ensures that the human being remains in control of decisions. Recent research offers some support for separate connections between the above-mentioned parts, but what is still missing is relatively little research on design on how to integrate them into a single coherent learning framework. This research attempts to bridge this gap through a process where: authentic professional problem → GenAI-supported exploration → critical evaluation → telecollaborative negotiation → pedagogical justification → instructional redesign → reflection.

This model is also concerned with the potential problem of unproductive cognitive offloading. Since cognitive offloading was not directly measured in the current study, the hypothesis of decreased use of

AI is presented as an interpretive conclusion rather than an effect. The theory is based on the assumption that the use of AI will stimulate professional reasoning, as students will have to ask questions, make comparisons, provide justifications, and make revisions regarding the AI-generated solutions (Celik et al., 2026a; Yang & Markauskaite, 2025; Sun & Huang, 2025). See Figure 1:

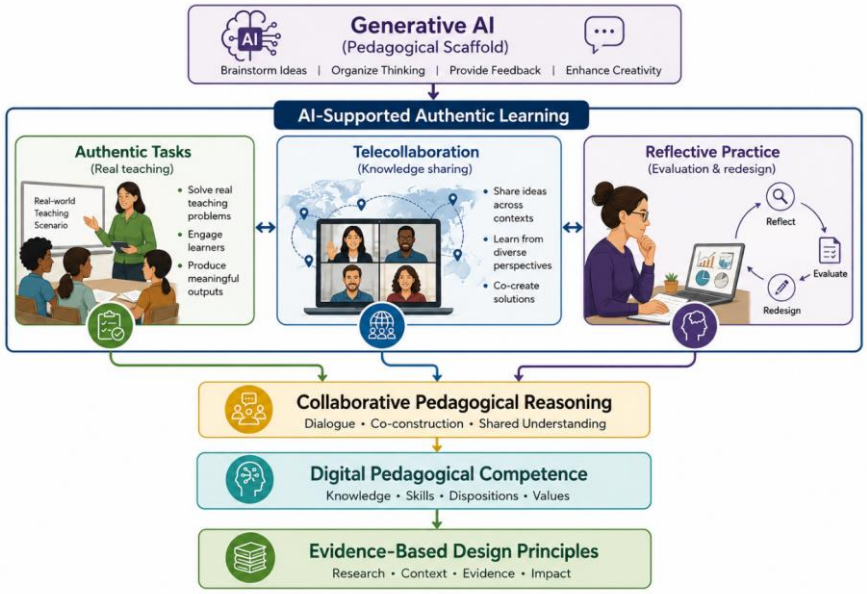


Figure 1. Conceptual framework linking authentic professional tasks, GenAI, telecollaboration, critical evaluation, iterative redesign, and digital pedagogical competence.

Table 1. Theoretical Perspectives, Functions, and Learning Focus within the Integrative Framework

Theoretical perspective	Function in the design	Learning focus
Authentic Learning	Realistic instructional-design tasks	Pedagogical reasoning
Situated/Telecollaborative Learning	Cross-institutional collaboration and peer review	Socially mediated professional learning
TPACK/DigCompEdu	Technology–pedagogy evaluation	Digital pedagogical competence
GenAI	Idea generation, alternatives, critique and feedback	AI-mediated inquiry
DBR	Iterative refinement across cycles	Design knowledge

The present study responds to this shift by treating AI use as part of a broader pedagogical ecology. The intervention does not assume that more AI use is inherently better; instead, it asks whether the sequence and organization of activities make AI interaction consequential for professional reasoning. This distinction is important because recent studies have shown both positive competency gains and continuing problems with evaluation, critical feedback, and overacceptance of AI-generated suggestions. The study therefore focuses on the design conditions under which AI-supported activity becomes educationally meaningful rather than on AI effectiveness in isolation.

Recent literature also indicates that the field is moving toward more process-sensitive accounts of GenAI use. For example, recent collaborative teacher-education research examines how prompting, editing, source checking, and reflection unfold during lesson design rather than treating AI exposure as a single intervention (Lehtinen, 2026). Other 2026 studies report that AI-supported practice can

strengthen professional competencies when GenAI is embedded in structured microteaching, simulation, or lesson-study environments, while also emphasizing the need for explicit support in navigating pedagogical challenges (Bümen & Yedigöz Kara, 2026; Zhang et al., 2026). These developments reinforce the rationale for examining not only what GenAI can produce, but how teacher educators organize the human activities surrounding those outputs.

METHODOLOGY

Research Design

A qualitative DBR methodology was used to investigate how an AI-supported authentic telecollaborative learning environment could be designed and refined to support pre-service EFL teachers' digital pedagogical competence (Design-Based Research Collective, 2003; Reeves, 2006; McKenney & Reeves, 2019). DBR was appropriate because the study sought design knowledge about the conditions and mechanisms through which an educational innovation could become pedagogically productive, not a statistically generalizable estimate of an intervention effect. The initial design was informed by authentic learning, situated learning, TPACK, and digital competence perspectives; evidence from Cycle 1 was used to refine the intervention implemented in Cycle 2.

Context and Participants

The study was conducted in undergraduate English Language Teaching programmes at two higher-education institutions in Northern Cyprus and was embedded in ELTV 203: Instructional Technology, a compulsory course focused on instructional design, technology-enhanced lesson planning, digital materials, and evaluation of educational technologies. Purposive sampling was used because the two programmes provided comparable course contexts and enabled cross-institutional collaboration. The intervention involved 21 pre-service EFL teachers: six in Cycle 1 and 15 in Cycle 2. Participants worked in cross-institutional teams and engaged in common learning objectives, collaborative lesson design, peer review, and reflection.

Table 2. Demographic and Contextual Characteristics of Participants across DBR Cycles

Characteristic	Cycle 1	Cycle 2
Participants	6	15
Institutions	GAU (3), EMU (3)	GAU (7), EMU (8)
Female/Male	4/2	9/6
Age	20–24	20–25
English proficiency	B2–C1	B2–C1
Course	Instructional Technology	Instructional Technology
Duration	6 weeks	6 weeks

Intervention Design and DBR Cycles

The intervention combined four design features: authentic professional tasks, structured telecollaboration, GenAI-supported pedagogical inquiry, and iterative reflection and redesign. Participants designed technology-integrated EFL lessons and instructional materials, evaluated AI-generated alternatives, exchanged peer feedback across institutions, and revised their designs. Zoom supported synchronous collaboration; Google Docs supported shared lesson development and revision;

and ChatGPT-4o was used for brainstorming, alternative strategies, organization of lesson content, critique, and feedback. Participants were explicitly instructed to treat AI outputs as provisional suggestions rather than finished instructional products.

Cycle 1 functioned as an exploration design cycle. Evidence indicated overreliance on broad AI prompts, uneven participation, limited pedagogical justification, and superficial end-of-project reflection. The intervention was consequently refined through a pedagogical planning template, rotating collaborative roles, weekly guided reflection, a prompt-writing guide, and peer-plus-AI feedback. Cycle 2 implemented a refined design with 15 participants. The cross-cycle comparison therefore focused on how design changes altered the conditions for pedagogical engagement.

Table 3. Cycle 1 Design Problems, Implemented Scaffolds, and Evidence Sought for Cycle 2

Component	Cycle 1 evidence/problem	Refinement	Cycle 2 evidence sought
Lesson planning	Broad AI use; limited pedagogical justification	Pedagogical planning template	More contextualized and justified decisions
Collaboration	Uneven participation	Rotating roles and structured peer review	More balanced pedagogical dialogue
Reflection	End-point and descriptive	Weekly guided journals	More critical, continuous reflection
AI use	Broad/free prompting	Prompt-writing guide	Purposeful prompts and critique
Feedback	Primarily lecturer feedback	Peer + AI + lecturer feedback	More iterative revision

During Cycle 2, prompt guidance required participants to align requests with lesson objectives and learner characteristics. Examples included prompts for generating alternative communicative activities, critiquing a lesson plan, designing authentic assessment criteria, and identifying strengths and areas for improvement. Participants were instructed not to upload personally identifiable information, institutional documents, or confidential educational records. AI-generated material was evaluated before incorporation into instructional artefacts.

Data Sources

Various qualitative data sources were employed, namely, the semi-structured interviews that were conducted at each phase of DBR, reflective journals, field notes of the researcher, and instructional artefacts such as lesson plans, outputs from the use of AI to facilitate the design process. The interviews focused on the participants' experiences, their perceptions of AI-assisted design, collaborative efforts, difficulties, and reflections on teaching with technology integration.

Data Analysis

The data were analyzed through reflexive thematic analysis (Braun & Clarke, 2021). These included the following six phases, which were repetitive in nature: Familiarization; Initial coding; Theme generation; Theme review; Theme definition; and Analysis report writing. Coding was done according to learning experiences, usage of AI, pedagogical decisions, collaboration, reflection, and design problems. Themes were retested against the data set at various points throughout the analysis process, and comparisons between data sources and DBR cycles were done. Cycle 1 data was analyzed as both participant data and design information.

Trustworthiness, Reflexivity, and Ethics

The credibility was established by using multiple-source triangulation, multiple rounds of engagement with the data, a trail of decisions made during the design and analysis process, and reflective memoing. Rather than assuming that different sources of data were a means of objectifying the same reality, convergence was employed to determine if similar patterns emerged from the interviews, journals, field notes, and artefacts. The dual role of researcher and teacher was explicitly recognized, since participant involvement was voluntary, interview and journal contents did not impact grades, interviews took place following important milestones, and the researchers kept reflective accounts of their personal investments in the intervention.

The project received ethical clearance from the Scientific Research and Publication Ethics Committee at Girne American University (Approval No. 5/22-133). Informed consent was obtained from the participants, and all data were de-identified. Participants were told to refrain from entering personally identifiable or sensitive information into GenAI tools. Thus, responsible use of AI was considered both an ethical requirement and an instructional principle.

To preserve analytical independence between findings and discussion, this analysis will be based on the mechanisms found in the data and not the design principles. The principles will be identified only once the mechanisms have been explored in different cycles. This is necessary for a DBR study as the findings will include observations of what happened during the intervention, while the design principles will be the design knowledge acquired throughout the cycle(s).

RESULTS

Four mechanisms were found to be interrelated and linked to increased sophistication of digital pedagogical engagement via cross-cycle analysis: (1) authentic professional engagement, (2) collaboration in knowledge building, (3) AI-enabled pedagogical scaffolding and critique, and (4) reflective pedagogical reasoning in iterated design. The findings are considered as qualitative evidence of development as well as increasing sophistication in enactment and not as causally or statistically generalizable findings.

Authentic Professional Engagement

Instructional design assignments that reflected authenticity formed the basis of participants' interaction with the technology in the study. Whereas participants saw the technology in Cycle 1 as a productivity tool that needed no justification to accept the suggestions from the AI technology, the design assignment template and task structure used in Cycle 2 necessitated that participants link the suggested ideas from AI to instructional goals, learners' needs, and authentic communication purposes.

"By designing a real lesson, I realized that AI could provide ideas, but I would need to think like a teacher instead of accepting suggestions provided by AI." (Participant 7, Cycle 2)

From the evidence, incorporating authenticity changed the nature of the use of AI became different from the normal assignment where individuals had to determine whether the AI-produced option was suitable for the specific teaching dilemma. This is an important part of the study as it is the professional situation that created the need for judgment.

Collaborative Knowledge Construction Through Telecollaboration

Telecollaboration thus became an avenue to reflect on and negotiate AI-based decisions in the pedagogical process. Uneven participation and unclear roles emerged in Cycle 1, leading to rotating roles and peer reviews in Cycle 2. Participants then saw cross-university discourse as providing an alternative perspective and a method to question assumptions which might not be challenged otherwise.

"*I was working with my partner from another university, and it really influenced my planning of the lesson. We questioned everything, and these conversations led to improvement of our lesson rather than accepting any first solution proposed by AI.* (Participant 3, Cycle 2)

"*There have been times when we did not agree on the activities that ChatGPT suggested; however, the discussion of pros and cons made us select the activities which suited our learners more." (Participant 9, Cycle 2)*

These examples show the difference between collaboration with an AI and collaboration around an AI: an AI suggested possibilities, but peer interaction provided context, critique, and negotiation. Consequently, the cross-cycle analysis shows that the collaboration was effective because it was consciously built around pedagogical problems, not just communication.

"At first, we divided the work because we thought it would go faster, but then we realized that talking about everything together enabled us to create a far better article." (Participant 1, Cycle 2)

The participants began by approaching collaboration as a division of tasks, which was related to their experiences of working in groups during traditional courses.

Other participants stressed the supportive nature of peer interaction. Many participants noted that discussions made them feel less uncertain about conducting research, finding literature, and making methodological decisions; instead of always asking the researcher, they started consulting their peers.

One participant commented:

"Whenever I became confused about something, I discussed it with my partner before asking the instructor. Most of the time, we solved the problem together." (Participant 8, Cycle 2)

AI-Mediated Pedagogical Scaffolding and Critical Evaluation

Participants used ChatGPT-4o for brainstorming, alternative activities, organizing, criticizing and feedback. The evidence from Cycle 1 demonstrated a trend towards broader prompts and acceptance of the outputs. In Cycle 2, the prompt guide turned the participants' use of the AI towards specific purposes in relation to objectives and learner features. The participants started to describe AI outputs as points of departure for evaluation and contextualization.

"*AI provided us with a lot of ideas very fast, but we realized that all those ideas were not always relevant for our learners. We had to analyze every single activity and decide whether to incorporate it into our curriculum or not."* (Participant, Cycle 2)

"*AI helped me to organize my lesson plan and develop ideas, but I could not simply rely on AI to make my teaching decisions".* (Participant 2, Cycle 1)

Thus, the evidence shows that participants' AI use turned from an efficiency-oriented one to one mediated by pedagogical inquiry. The key point was not increased exposure to AI but increased responsibility for evaluating AI outputs.

Reflective Pedagogical Reasoning and Iterative Redesign

Reflection took more importance after the introduction of weekly guided journaling, feedback from peers and redesign of lessons. Reflection of Cycle 1 revolved around the effectiveness or inefficiency of technology and outputs generated through AI. Later reflections revolved around the appropriateness of activities used in lessons, whether the support provided authentic communication and changes required in the lesson. The participants made connections between the changes and changes in understanding of the learner and objectives of instruction.

"After getting feedback from my partner and going through AI suggestions again, I totally changed one activity of my lesson because I understood that it does not promote authentic communication." (Participant 8, Cycle 2)

"Every revision helped me understand why I chose specific strategies of teaching rather than just fulfilling the requirements of the assignment." (Participant 14, Cycle 2)

"The feedback process helped me to understand that teaching is not about doing everything correctly on the first attempt. Teaching is about improving every time you get new ideas." (Participant 11, Cycle 2)

One participant commented:

"At times, my partner challenged my views, and I had to give reasons for choosing a certain type of activity as effective. Those discussions forced me to consider my choices in teaching very carefully."

This quote proves that reflective collaboration allowed participants to express and justify their reasoning related to pedagogy. Reflecting on their own instruction strategies made learners more knowledgeable about teaching principles and encouraged self-assessment of personal assumptions.

Other participants also expressed their thoughts on the topic of reflection:

"The discussions with different people helped me realize that there may be different ways to teach and not just one right one. Reflection taught me to be open-minded."

This quote shows that reflection did not limit itself to personal reasoning but included consideration of alternative pedagogical approaches.

Cross-Cycle Case Illustration

A representative trajectory of a team's process demonstrates the connection between the four mechanisms. In Cycle 1, Team A generated an environmental lesson based on a wide range of prompts and mostly AI-generated ideas. Peer reviewers raised concerns about whether these activities constituted actual communication in real-life situations. As a response to this criticism, Cycle 2 involved the use of scaffolds, reflective questioning and planning. The modified lesson aimed at making learners work on a communicative inquiry-based task rather than copying information provided by the teacher.

"The most significant change was that we stopped using AI to complete tasks for us. We started learning how to formulate questions properly and choose the good ideas. It helped us to see ourselves as teachers." (Participant, Cycle 2)

This example is not representative but rather illustrative; its significance is in demonstrating the DBR mechanism. The design problem that appeared in Cycle 1 was transformed into a scaffold and affected the participants' actions in Cycle 2.

Table 4. Cross-Cycle Qualitative Evidence of Participant Engagement across the Four Mechanisms

Mechanism	Cycle 1 evidence/problem	Design refinement	Cycle 2 evidence
Authentic professional engagement	AI was often treated as a productivity tool and output was insufficiently contextualized.	Pedagogical planning template and stronger alignment with authentic teaching problems.	Lesson decisions were more often linked to objectives, learner needs, and communicative purposes.
Collaborative knowledge construction	Uneven participation and unclear responsibilities constrained peer inquiry.	Rotating roles and structured peer review.	Participants reported more questioning, negotiation, and shared pedagogical decision-making.
AI-mediated scaffolding	Broad prompts encouraged acceptance of ready-made suggestions.	Prompt-writing guide and explicit evaluation requirements.	AI outputs were more often treated as alternatives to critique, adapt, and justify.
Reflective reasoning	Reflection was mainly end-point and descriptive.	Weekly guided journals and iterative redesign.	Reflections more often addressed rationale, learner fit, and consequences of instructional choices.

Evidence-Based Design Principles

The findings were translated into five design principles. They are presented as transferable design propositions, not as universally generalizable effects.

Table 5. Evidence-Based Design Principles for AI-Supported Telecollaborative Teacher Education

Design principle	Empirical basis	Practical implication
1. Embed GenAI within authentic professional tasks	AI became more pedagogically meaningful when linked to contextual lesson-design problems.	Use real or realistic teaching problems requiring contextualized decisions.
2. Structure telecollaboration around purposeful knowledge building	Peer questioning and negotiation improved evaluation of instructional alternatives.	Use defined roles, shared goals, peer review, and collaborative inquiry.
3. Position GenAI as a scaffold, not an authority	Participants moved from accepting outputs toward critique and adaptation when scaffolds were explicit.	Require evaluation, verification, modification, and justification of AI outputs.
4. Embed reflection and iterative redesign	Guided reflection and successive revisions shifted attention from efficiency toward pedagogical rationale.	Use journals, feedback, redesign cycles, and explicit justification.
5. Align AI, authenticity, collaboration, and reflection as a system	The four mechanisms were most coherent when implemented together.	Design the elements as one learning system rather than separate innovations.

The second implication relates to the definition of competence. In the current investigation, digital pedagogical competence cannot be confused with confidence, frequency of AI usage, and the capacity to create an appropriate lesson plan. Competence is seen through the process of reasoning about

technology use, involving identification of the issue in teaching, creation of possible options, assessment of their appropriateness, explanation of the decision made, modification of output, and reflection on the consequences. Such a definition facilitates a clear difference between the current investigation and previous studies that considered AI acceptance or self-rated competence. The discussion centers on the contribution to understanding of the study rather than listing the four findings. The major interpretation is that the contribution made by GenAI was not its function as an independent instructional treatment. The contribution of GenAI was mediated through the authenticity of the professional problem, the structure of peer interaction, the **judgment of** AI outputs, and iterative redesign. This interpretation is consistent with the recent trend in the field of moving from technology affordance to AI-mediated professional learning and pedagogical judgment (Lehtinen, 2026; Bümen & Yedigöz Kara, 2026; Zhang et al., 2026).

It is clear from the matrix that DBR logic is expressed in the fact that design principles are not derived from successful results but from problems and targeted changes across both cycles and sources. Thus, the evidence should be interpreted as an analytic account of the change in the learning environment and engagement with it.

DISCUSSION

GenAI as a Mediating Resource for Pedagogical Reasoning

The findings suggest that GenAI was most pedagogically productive when used to support professional reasoning rather than to automate lesson creation. This result builds on the latest research results demonstrating that while GenAI may be useful in planning lessons, more engagement requires scaffolding, human facilitation, and criticism (Sun & Huang, 2025; Celik et al., 2026a; Mailizar et al., 2026). The current study provides an explanation of how this can be achieved: participants had to go through a process of exploring, criticizing, negotiating, justifying, and redesigning.

Digital Pedagogical Competence Through Iterative Design

These results also add clarity to the notion of digital pedagogical competence in the GenAI era. Participants did not learn to use a range of tools; rather, they learned how to link technology to goals, learners' needs, content, pedagogy, and context via design choices made repeatedly. This adds further validity to the continued relevance of the TPACK framework in conjunction with recent developments in AI-specific pedagogical knowledge and Intelligent-TPACK (Celik et al., 2026a; Jou et al., 2025; Mailizar et al., 2026). The contribution is that process-oriented competence is achieved by design, critique, and revision.

Teacher Agency and “Collaborating Around AI”

The findings suggest that teacher agency played an important role in shaping how AI was integrated into pedagogical decision-making. Our study confirms current research that interaction with AI does not necessarily result in the development of critical agency, and learners are able to accept AI recommendations without posing proper questions (Lee et al., 2025; Yang & Markauskaite, 2025; Celik et al., 2026b). The telecollaborative environment provided a social way for developing agencies, as AI-generated texts could be questioned. This corroborates the theoretical distinction between collaboration

with AI and collaboration around AI, as the former entails human interpretation, evaluation, and reconstruction of opportunities suggested by AI. It is especially important for EFL teacher education as pedagogical appropriateness is dependent on linguistic, cultural, and learner-specific context.

Theoretical Differentiation: Collaborating *with* AI vs. Collaborating *Around* AI

Theoretical significance of this study includes the explicit demarcation between collaborating with AI and collaborating around AI, distinguishing this study from existing prompt engineering literature. Many studies in the emerging literature on GenAI for teacher education center on the topic of prompt optimization and human-AI collaboration to generate instructional content. While prompt engineering is part of digital literacy, too much focus on human-AI collaboration can result in viewing pedagogical practice as a purely technical endeavour that leads to uncritical adoption or cognitive offloading.

Collaborating around AI is defined as triadic and socially mediated collaboration. Within the social collaborative context, GenAI is not viewed as a final solution or a direct collaborator, but an external artefact to be discussed, negotiated, debated, and defended among human collaborators.

Collaborating With AI (Dyadic / Prompt-Engineering Approach): The learner works with the LLM directly to improve prompts, plan lessons, and maximize output quality. The evaluation is based on three aspects: output fluency, speed, and prompt alignment.

Collaborating Around AI (Triadic / Pedagogical Reasoning Approach): The learner uses AI output as a trigger to engage in cross-institutional dialogue with peers. AI output is subjected to socio-pedagogical analysis: peers discuss whether the idea generated by the AI is relevant to their needs as learners, authentic in terms of communication and contextually constrained.

The telecollaborative environment is crucial for this approach. An individual pre-service teacher who would use ChatGPT for a plausible lesson activity could easily be persuaded by it. However, putting two pre-service teachers in conversation about the same output will make them provide justifications for why AI suggestions are either good or not from the pedagogical perspective. Peer disagreement becomes the source of cognitive dissonance that prevents automated agreement. Hence, this study shows that digital pedagogical competence does not depend on the technology per se but rather on creating a social environment for systematic challenging, contextualizing, and rewriting of AI outputs via professional judgment of the human user.

Table 6. Theoretical Comparison of Dyadic ("Collaborating With AI") vs. Triadic ("Collaborating Around AI") Models

Dimension	Collaborating *With* AI (Standard Prompt Engineering)	Collaborating *Around* AI (This Study's Framework)
Interaction Model	Dyadic (Learner ↔ GenAI)	Triadic (Learner ↔ Peer ↔ GenAI)
Status of AI Output	Direct product or co-created outcome	Provisional artefact for peer critique & negotiation
Primary Learning Focus	Technical prompting, efficiency, & output quality	Critical evaluation, contextualization, & pedagogical reasoning
Role of Social Context	Minimal or secondary (individual use)	Essential (telecollaboration drives critique)

Risk Mitigation	Vulnerable to overreliance & cognitive offloading	Constrains overreliance through mandatory peer justification

Authenticity as a Condition for Meaningful AI Use

The authentic task seems to have restricted the likelihood that AI could simply do the learning task. As such, the participant is to design for the learner and make justifications for his decisions based on professional standards. This builds on recent literature that has shown connections between GenAI and authentic professional practice as well as practicum-based learning (Yang & Markauskaite, 2025; Kılıçkaya & Kic-Drgas, 2025). In this case, the current study does not show that authentic learning restricts offloading; it only shows that there may be a scenario for it.

Telecollaboration as a Mechanism for Critical AI Engagement

Telecollaboration provided not only access to peer communication but also an opportunity to create a social setting where participants needed to justify instructional decisions. Disagreements among peers became productive in the context of shared instructional challenges, as they revealed hidden assumptions that the output could not reveal. This resonates with collaborative GenAI studies that highlight the importance of considering how AI is socially organized in addition to having access to it (Sun & Huang, 2025; Yang & Markauskaite, 2025).

An Integrated Process Model

The results support a model of a process through which digital pedagogical competence is achieved via the interactions of four conditions: professional problem-solving, GenAI-enabled exploration, critical assessment, and telecollaborative negotiation, with subsequent pedagogical reasoning, redesigning, and reflection. The model helps explain the role of five design principles: they are not separate techniques but interacting conditions of the same learning environment. Therefore, the research goes beyond the question of the usefulness of GenAI in the direction of discovering what learners should do with AI to be professionally educative.

Theoretical and Practical Implications

In theory, this research frames GenAI as a mediating artefact within an enactment of digital pedagogical competence. TPACK describes technology–pedagogy–content connections, while authentic learning provides the rationale for the importance of professional issues. Telecollaboration describes the process of socially negotiating knowledge, and DBR provides a way to improve all these components via iterative design. In practice, teacher education programs should integrate AI learning into methodology, instruction technology, practicum, evaluation, and reflection rather than separating it from methodology, instruction, and evaluation through brief technical workshops. An effective workflow would look like this: design → consult GenAI → evaluate → verify → discuss → revise → justify → reflect.

Contextual Mapping and Boundary Conditions of the Design Principles

To assess the applicability of the five design principles in an international context, one needs to consider the local context of the study in relation to larger teacher education contexts. In this case, the study was

performed in Northern Cyprus, which is known for many international students enrolled in higher education, extensive use of EMI in ELT departments, and application of European Union frameworks for aligning curricula (e.g., DigCompEdu, EHEA).

The applicability of each of the principles is contingent on certain contextual conditions, not institutionally determined ones:

- **High Transferability Environments**: Principles that can be applied directly in the context of higher education institutions where there is limited class size (less than 30 students), course-level curriculum flexibility, required instruction technology courses, and access to sophisticated GenAI models (for example, ChatGPT-4o).
- **Boundary Conditions & Required Adaptations**:

Institutional policy and data privacy: Principle 3 (GenAI as a scaffold) and Principle 5 (Systemic Integration) assume permissive or exploratory institutional AI policies. When stringent data privacy laws apply (e.g., EU GDPR compliance in state-run K-12 systems or universities with a complete ban on GenAI use), use privacy-compatible, locally hosted LLMs or institution-approved sandbox environments.

Scales & Resources: Principles 2 and 4 (Structured Telecollaboration and Iterative Redesign) involve a considerable number of instructor-facilitated activities and reasonable student-to-instructor ratios. In a large-scale teacher preparation program (N>100), these principles need in some way, e.g., with automated peer-matching algorithms or scaling self-assessment rubrics to ensure the intensity of peer feedback.

Tech & Linguistic Parity: Telecollaborative exchanges (Principle 2) were conducted in an environment of parity in technological resources and linguistic competence (B2–C1 in English). Applying these principles in cross-national, low-resourced environments may require additional scaffolding to address digital divides or linguistic barriers in peer review.

Limitations and Future Research

One key methodological limitation of this study is related to the measurement of cognitive offloading. Although the data indicate changes in pedagogical reasoning, peer negotiation, and design iterations of pre-service teachers, this study did not measure cognitive offloading through methods such as eye tracking, cognitive load scales, dual-task experiments, and physiological indicators.

Rather, decreased dependence on artificial intelligence and higher critical involvement are considered qualitative processes indicative of the development of pedagogical competence. Considering the above, readers/reviewers should consider the following limitations about the unmeasured variable:

Interpretive Boundary: The identified decrease in uncritical use of AI is interpreted as an implication of the learning design interpretation rather than a validated psychological effect involving decreased mental effort or a re-engineered cognitive architecture.

•Indicators vs. Psychometric Constructs: This study has provided empirical observations regarding what participants did with AI-generated suggestions (i.e., questioned, rejected, and rewrote the suggestions), rather than quantifying the mental capacity involved in this process.

•Framework Fit: As this research falls under Design-Based Research (DBR), which emphasizes iteration and improvement of the learning environment and knowledge about the design, rather than the testing of isolated variables, the measurement of isolated psychological constructs was beyond the qualitative scope of this study.

• Methodological Clarification on Unmeasured Variables & Scope:

Qualitative DBR research is conducted in this case to study mechanisms of digital pedagogical competence and design principles. This study does not claim quantitative measurement or causal reduction of cognitive offloading, cognitive load, and working memory. Uncritical usage of AI was assessed only considering justification made by the participants, criticism by peers, and iterations in artefact modification. The measurement of cognitive offloading will require further research using mixed methods designs.

Directions for Future Cognitive Measurement

Future work to build on the qualitative knowledge of design generated by this study should involve integrating DBR frameworks with specific psychometric and neurocognitive tools. This involves the use of validated scales for Cognitive Load Theory (CLT) (intrinsic, extraneous, and germane loads as one interacts with the AI), coding of decision-making latency through screen recording of interactions, and experimental trials comparing knowledge retention when designing lessons with and around GenAI

Table 7. Contextual Mapping Matrix and Boundary Conditions for International Transferability

Northern Cyprus Contextual Feature	Global Equivalent Setting	Primary Boundary / Adaptation Condition	Relevant Design Principles
Bilingual/EMI ELT Programs	Global EMI & ESL/EFL Teacher Training	Assumes B2+ target language proficiency for reflective peer review	DP 1, DP 2
Course-Level Curriculum Autonomy	Modular & Design-Based Courses	Requires instructor freedom to modify lesson planning templates and grading criteria.	DP 1, DP 3, DP 4
Exploratory AI Usage Policy	Standard Higher Ed Digital Policies	High-privacy systems require institutional AI sandboxes or open-source LLMs.	DP 3, DP 5
Small-Group Seminars ($n = 6 - 15$)	Micro-teaching & Clinical Practicums	Large lecture classes ($n > 50$) need automated or asynchronous peer-review structures.	DP 2, DP 4

CONCLUSION

In this study, how to design an AI-aided authentic telecollaborative learning environment for pre-service EFL teachers' pedagogical engagement digitally was investigated. Across the two DBR cycles,

increasingly sophisticated forms of AI-supported pedagogical engagement were observed as authentic tasks, structured telecollaboration, critical evaluation, and iterative reflection were strengthened. This study, thus, converts this process into five design principles which focus not on access to AI but on the pedagogical conditions for the use of AI.

The main point made in this paper is that there is a process-oriented perspective on the integration of GenAI in teacher education; GenAI can provide more instructional options to teachers, yet the key is what professionals are supposed to do about these options pedagogically. If learners should critically evaluate, negotiate, justify, contextualize, and revise the ideas that were constructed via GenAI, GenAI can act as a scaffolding tool for professional judgment, for which human teachers have the responsibility for pedagogy.

Declarations

Conflict of interest. The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding. The author(s) received no financial support for the research, authorship, and/or publication of this article.

Ethical approval. Ethical approval was obtained from the Scientific Research and Publication Ethics Committee of Girne American University (Approval No. 5/22-133) before data collection.

Informed consent. Written informed consent was obtained from all participants. Participation was voluntary, and participants could withdraw without academic penalty.

AI uses disclosure. ChatGPT-4o was used for language editing, grammar correction, and clarity/organization of the manuscript. It was not used to generate research data, conduct analyses, interpret findings, or draw conclusions. The authors retain responsibility for the manuscript and its intellectual content.

Data availability. The datasets are not publicly available because they contain information that could compromise participant privacy. De-identified data may be made available from the corresponding author upon reasonable request, subject to institutional ethics requirements.

References

Aldemir, T., Kılınç, S., Bicer, A., Grant, P., Davis, T., & Sweany, N. W. (2025). Intelligent-TPACK in practice: Design and evidence from a three-week teacher preparation module. Computers and Education Open, 9, 100306. https://doi.org/10.1016/j.caeo.2025.100306.

Amiel, T., & Reeves, T. C. (2008). Design-based research and educational technology: Rethinking technology and the research agenda. Educational Technology & Society, 11(4), 29–40.

Braun, V., & Clarke, V. (2021). One size fits all? What counts as quality practice in (reflexive) thematic analysis? Qualitative Research in Psychology, 18(3), 328–352. https://doi.org/10.1080/14780887.2020.1769238.

Celik, I., Kontkanen, S., Laru, J., & Dalyanci, A. A. (2026a). Co-constructing adaptive lesson plans with GenAI: Pre-service teachers' Intelligent-TPACK and prompt engineering strategies. Computers & Education, 241, 105485. https://doi.org/10.1016/j.compedu.2025.105485.

Celik, I., Kontkanen, S., Laru, J., & Dalyanci, A. A. (2026b). Pre-service teachers' agency during their interactions with generative AI while designing for learning: A process view on Intelligent-TPACK. Computers and Education Open, 10, 100325. https://doi.org/10.1016/j.caeo.2025.100325.

Chiu, T. K. F. (2024). The impact of generative AI (GenAI) on practices, policies and research direction in education: A case of ChatGPT and Midjourney. Interactive Learning Environments, 32(10), 6187–6203. https://doi.org/10.1080/10494820.2023.2253861.

Crompton, H., Jones, M. V., & Burke, D. (2024). Affordances and challenges of artificial intelligence in K-12 education: A systematic review. Journal of Research on Technology in Education, 56(3), 248–268.

Design-Based Research Collective. (2003). Design-based research: An emerging paradigm for educational inquiry. Educational Researcher, 32(1), 5–8. https://doi.org/10.3102/0013189X032001005

Dooly, M., & Sadler, R. (2020). If you don't improve, what's the point? Investigating the impact of telecollaboration in teacher education. Peter Lang.

Eşki, M., & Eryılmaz, N. (2026). An exploration of Turkish EFL pre-service teachers' perceptions on the integration of generative AI in English language teacher education: Benefits, challenges, and future directions. Asian-Pacific Journal of Second and Foreign Language Education, 11, Article 44. https://doi.org/10.1186/s40862-026-00420-w.

European Commission. (2022). Ethical guidelines on the use of artificial intelligence (AI) and data in teaching and learning for educators. Publications Office of the European Union.

Herrington, J., & Oliver, R. (2000). An instructional design framework for authentic learning environments. Educational Technology Research and Development, 48(3), 23–48.

Herrington, J., Reeves, T. C., & Oliver, R. (2010). A guide to authentic e-learning. Routledge.

Holmes, W., Bialik, M., & Fadel, C. (2022). Artificial intelligence in education: Promises and implications for teaching and learning. Center for Curriculum Redesign.

Huang, T., Wu, C., & Wu, M. (2025). Developing pre-service language teachers' GenAI literacy: An interventional study in an English language teacher education course. Discover Artificial Intelligence, 5(1), Article 163. https://doi.org/10.1007/s44163-025-00435-1.

Jou, M., Kuo, T.-H., Chiang, Y.-C., Hao, Y., & Huang, C.-C. (2025). Integrating generative AI in teacher education: A qualitative exploration of TPACK growth and critical reflection. Turkish Online Journal of Educational Technology, 24(4), 101–107.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeiffer, F., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274.

Kılıçkaya, F., & Kic-Drgas, J. (2025). Pre-service language teachers' experiences and perceptions of integrating generative AI in practicum-based lesson study. Humanities and Social Sciences Communications, 12, Article 1478. https://doi.org/10.1057/s41599-025-05715-w.

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. RELC Journal, 54(2), 537–550. https://doi.org/10.1177/00336882231162868.

Lave, J., & Wenger, E. (1991). Situated learning: Legitimate peripheral participation. Cambridge University Press.

Lee, S., Jeon, J., & Choe, H. (2025). Generative AI (GenAI) and pre-service teacher agency in ELT. ELT Journal, 79(2), 287–296. https://doi.org/10.1093/elt/ccaf005.

Lincoln, Y. S., & Guba, E. G. (1985). Naturalistic inquiry. Sage.

Mailizar, M., Hidayat, M., Syukri, M., & Wijayanti, D. (2026). Exploring the impact of generative AI on pre-service mathematics teacher TPACK: A mixed-method study. Discover Education, 5, Article 453. https://doi.org/10.1007/s44217-026-01432-x.

McKenney, S., & Reeves, T. C. (2019). Conducting educational design research (2nd ed.). Routledge.

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. Teachers College Record, 108(6), 1017–1054. https://doi.org/10.1111/j.1467-9620.2006.00684.x.

Moorhouse, B. L. (2024). Generative artificial intelligence and ELT. ELT Journal, 78(4), 378–392. https://doi.org/10.1093/elt/ccae032.

Moorhouse, B. L., Wan, Y., Wu, C., Kohnke, L., Ho, T. Y., & Kwong, T. (2024). Developing language teachers' professional generative AI competence: An intervention study in an initial language teacher education course. System, 125, 103399. https://doi.org/10.1016/j.system.2024.103399.

O'Dowd, R. (2018). From telecollaboration to virtual exchange: State-of-the-art and the role of UNICollaboration in moving forward. Journal of Virtual Exchange, 1, 1–23.

O'Dowd, R., & Dooly, M. (2021). Intercultural communicative competence through virtual exchange. In M. Dooly & R. O'Dowd (Eds.), In this together: Teachers' experiences with transnational, telecollaborative language learning projects. Peter Lang.

OECD. (2023). OECD digital education outlook 2023: Towards an effective digital education ecosystem. OECD Publishing.

Onbasili, Ü. I. (2026). Empowering pre-service teachers with generative AI: A GenAI-TPACK-based approach to digital storytelling. Education and Information Technologies. https://doi.org/10.1007/s10639-026-13991-5

Peng, X. (2026). Generative AI in pre-service EFL/ESL teacher education: A systematic review of empirical studies. Journal for the Education of Gifted Young Scientists, 14(2), 75–89. https://doi.org/10.17478/jegys.1915711.

Redecker, C. (2017). European framework for the digital competence of educators: DigCompEdu. Publications Office of the European Union.

Reeves, T. C. (2006). Design research from a technology perspective. In J. Van den Akker, K. Gravemeijer, S. McKenney, & N. Nieveen (Eds.), Educational design research (pp. 52–66). Routledge.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. Educational Researcher, 15(2), 4–14. https://doi.org/10.3102/0013189X015002004.

Lehtinen, A., Channa, F., Näykki, P., Ahlström, E., & Hiljanen, M. (2026). Generative AI for collaborative learning: Fostering critical thinking in teacher education. Journal of Computer Assisted Learning, 42(3), e70256. https://doi.org/10.1002/jcal.70256.

Bümen, N. T., & Yedigöz Kara, Z. (2026). Integrating generative artificial intelligence into preservice teacher education: Evidence from a microteaching lesson study. Teaching and Teacher Education, 180, 105660. https://doi.org/10.1016/j.tate.2026.105660.

Sun, Y., & Huang, Y. (2025). Exploring the role of generative AI in collaborative lesson planning for pre-service teachers. Journal of Computer Assisted Learning, 41(4), e70098. https://doi.org/10.1111/jcal.70098

Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. Smart Learning Environments, 10, Article 15. https://doi.org/10.1186/s40561-023-00237-x.

UNESCO. (2023). Guidance for generative AI in education and research. UNESCO.

UNESCO. (2024). Generative AI and the future of education. UNESCO.

Uygun, E., & Demir, B. (2026). A PRISMA-based systematic review of artificial intelligence in English as a foreign and second language education (2023–2025). Discover Education, 5, Article 478. https://doi.org/10.1007/s44217-026-01504-y.

Wang, M., Tlili, A., Khribi, M. K., Lo, C. K., & Huang, R. (2025). Generative artificial intelligence in special education: A systematic review through the lens of the mediated-action model. Information Development. https://doi.org/10.1177/02666669251335655.

Yang, H., & Markauskaite, L. (2025). Fostering language student teachers' transformative agency for embracing GenAI: A formative intervention. Teaching and Teacher Education, 159, 104980. https://doi.org/10.1016/j.tate.2025.104980.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—Where are the educators? International Journal of Educational Technology in Higher Education, 16, Article 39. https://doi.org/10.1186/s41239-019-0171-0

Zhang, N., Ke, F., Dai, C.-P., Southerland, S. A., & Barrett, A. (2026). Science and mathematics preservice teachers' perceptions and experiences of practicing dialogic teaching in generative AI-powered virtual reality simulation. Teaching and Teacher Education, 171, 105349. https://doi.org/10.1016/j.tate.2025.105349.

Machine Learning for CEFR-Informed Lexical Difficulty Classification of Turkish Song Lyrics: A Leakage-Aware Evaluation of Handcrafted Linguistic Features Against a Transformer-Based Hybrid Model

Mohammad Ibrahim[1*], Filiz Karaman[2]

[1] Department of Statistics, Faculty of Arts and Sciences, Yıldız Technical University, Istanbul, Türkiye
[2] Department of Statistics, Faculty of Arts and Sciences, Yıldız Technical University, Istanbul, Türkiye

*Corresponding author e-mail: drmoh16@hotmail.com

Abstract

Automatic proficiency-level classification has been studied almost exclusively for formal, edited genres such as learner essays and news text, leaving informal authentic genres — and Turkish in particular — largely unaddressed. This gap matters practically: song lyrics are among the most frequently consumed authentic materials by learners of Turkish as a foreign or second language, yet no publicly documented, reproducible instrument exists to grade their lexical difficulty. This study addresses that gap with a fully reproducible pipeline: (i) content words in a 35,285-song Turkish lyrics corpus are assigned to one of five frequency-derived lexical-difficulty bands (labeled A1–C1 by rank order only, as a naming convention, not as a validated CEFR assignment), using the openly licensed wordfreq Turkish frequency resource and data-driven quantile cutoffs rather than a non-public vocabulary list; (ii) a four-family, circularity-controlled feature set is engineered for the lyric genre; and (iii) eight classifiers — Logistic Regression, SVM, k-NN, Decision Tree, Random Forest, Gaussian Naïve Bayes, XGBoost, and a hybrid BERTurk + XGBoost model — are benchmarked under an identical 80/20 protocol. On this primary, fully reproducible experiment, Random Forest is the strongest model (accuracy 42.8–43.4%, balanced accuracy 0.411–0.426), with McNemar's tests (Holm-Bonferroni corrected) showing most classifier differences are not statistically significant, an ablation study and a feature-importance analysis showing a single weighted-lexical-difficulty scalar accounts for nearly all of the discriminative signal, and an artist-grouped train/test split confirming the result is not inflated by artist-level leakage. For context, we also report a label-recovery upper-bound analysis: under an original, non-public, in-house vocabulary-derived labeling scheme used in an earlier version of this pipeline, the same classifiers recover that labeling rule at near-ceiling accuracy (up to 0.994 macro-F1); because that scheme is not independently available for reuse and several original features were themselves derived from it, this figure is reported only as evidence of successful label recovery — not as an estimate of agreement with expert-assessed CEFR levels, and not as this study's primary result. Confusion in both experiments concentrates almost entirely between adjacent difficulty bands, consistent with the gradual, ordinal nature of lexical difficulty. We discuss why the primary, frequency-derived accuracy figures are lower and more conservative than the 60–70% macro-F1 typically reported for independently expert-rated CEFR classification in English and other languages — consistent with the difference between a general word-frequency proxy and a curated, pedagogically graded vocabulary — and position the resulting pipeline, band-construction method, and leakage-aware evaluation framework as a transparent, fully reproducible building block for CEFR-informed content recommendation, curriculum design, and

computer-assisted language learning for Turkish, pending independent human validation and, ideally, replacement of the frequency proxy with a published, pedagogically graded Turkish CEFR vocabulary reference once one becomes accessible for reuse.

Keywords: *Turkish natural language processing, text readability, feature engineering, frequency-derived proficiency Proxy, BERTurk*

INTRODUCTION

Motivation and Background

The Common European Framework of Reference for Languages (CEFR) is the de facto international standard for describing and comparing language proficiency, organizing learner competence into six ascending bands from A1 (beginner) to C2 (mastery) [5]. Beyond assessment and curriculum design, CEFR levels are increasingly used as a proxy for the difficulty of authentic reading and listening material offered to learners. Manual CEFR annotation, however, is slow, requires trained raters, and cannot keep pace with the volume of authentic content—news articles, subtitles, podcasts, song lyrics—that online language-learning platforms now draw upon. Automatic CEFR classification therefore has direct practical value: it lets large pools of authentic content be filtered and recommended to a learner by proficiency level without exhaustive manual annotation.

Song lyrics occupy a distinctive niche among authentic materials. Repetition, rhythm and emotional engagement are structurally built into music, and a growing body of second-language-acquisition (SLA) research links exposure to song lyrics with vocabulary uptake, pronunciation gains and learner motivation [13], [15], [16]. Turkish-language music is consumed at scale by learners of Turkish, yet—to the authors' knowledge—no publicly documented CEFR grading instrument exists for Turkish lyrics. This absence is compounded by the broader scarcity of Turkish NLP resources: Turkish is a morphologically rich, agglutinative language for which classification benchmarks, graded-vocabulary resources and pretrained models remain considerably less developed than for English [1], [12].

Research Gap

Four specific gaps motivate this work. First, automatic CEFR classification has rarely been attempted for Turkish, a language whose productive morphology changes how lexical-difficulty and readability signals behave relative to English or other well-resourced languages. Second, the CEFR and readability-classification literature has concentrated on formal, edited genres—essays, news comments, exam scripts—while informal genres such as song lyrics, with their refrains, colloquialisms and non-standard punctuation, remain largely unexamined [6], [7], [9], [10], [11], [23]. Third, comparisons between classical machine-learning pipelines built on handcrafted, linguistically motivated features and

deep/transformer-based pipelines that rely on self-learned representations remain limited for CEFR-style multi-level classification, and the few that exist are concentrated in English [7], [10], [21], [22], [23], [27]. Fourth, and more broadly, when CEFR labels for a large corpus cannot be independently expert-rated and are instead derived computationally from a graded vocabulary reference, the classification literature rarely examines whether the resulting labels are actually independent of the features later used to predict them, or how much of any reported accuracy reflects genuine linguistic signal rather than the classifier recovering its own labeling rule; a leakage-aware evaluation protocol that quantifies this directly, rather than only discussing it as a caveat, is largely absent from the CEFR-classification literature. This study is designed to close all four gaps simultaneously, using Turkish song lyrics as the target genre: it does not merely apply machine learning to a new Turkish dataset, but also interrogates the validity of the labeling methodology itself.

Research Questions and Contributions

The study is organized around four research questions:

- RQ1: To what extent can handcrafted, CEFR-informed linguistic features distinguish between frequency-derived lexical-difficulty bands (A1–C1 by rank order) in Turkish song lyrics, under a fully reproducible, circularity-controlled labeling scheme?
- RQ2: How do classical machine-learning classifiers compare with a transformer-based hybrid model (BERTurk + XGBoost) in predicting these frequency-derived lexical-difficulty bands, in terms of predictive performance and computational cost?
- RQ3: Which linguistic features contribute most to the prediction of frequency-derived lexical-difficulty bands, and is high aggregate accuracy driven by the combination of feature families or by a narrow, potentially label-defining subset?
- RQ4: Where do residual classification errors concentrate, and what do they reveal about the linguistic boundaries between adjacent lexical-difficulty bands in song lyrics?

Correspondingly, the paper makes four concrete contributions: (i) a fully reproducible framework for classifying frequency-derived lexical-difficulty bands in Turkish song lyrics, comprising a large-scale corpus (n = 35,285) labeled from an openly licensed word-frequency resource using a documented, quantile-based band-construction rule, together with a four-family handcrafted feature set—lexical statistics, distributional band composition, weighted lexical difficulty, and structural/repetition features—that is, to the authors' knowledge, the first of its kind proposed for Turkish; (ii) a controlled methodological comparison of eight machine-learning models, including a BERTurk-based hybrid,

under an identical evaluation protocol; (iii) an ablation analysis that isolates which engineered features actually drive predictive performance, rather than assuming the full feature set contributes uniformly; and (iv) a leakage-aware evaluation that reports, alongside the primary reproducible result, an explicit label-recovery upper-bound analysis under an original, non-public, in-house-referenced labeling scheme, quantifying how much of that scheme's headline accuracy reflects the classifier recovering its own labeling rule rather than generalizable linguistic signal.

LITERATURE REVIEW

Automatic CEFR-Level and Readability Classification

Automatic text-difficulty prediction predates the CEFR-specific literature by decades: classic readability formulas such as Flesch's Reading Ease score used shallow surface counts (average sentence length, syllables per word) to predict comprehension difficulty for English prose [20]. CEFR-level classification extends this idea to a pedagogically standardized, multi-level scale. Early work on German and Estonian learner corpora showed that lexical, syntactic and morphological features can support fine-grained CEFR-style classification, but that fine-grained (five- or six-level) schemes are considerably harder to separate than coarse binary or ternary splits, typically reaching 60–70% accuracy on small, hand-annotated corpora. Kerz et al. showed that document-averaged complexity scores are weaker discriminators than a recursive "complexity contour" model that tracks how complexity evolves within a text [6]. More recent work has turned to transformer representations: Schmalz and Brutti fine-tuned BERT for English CEFR assessment [9], and Lagutina et al. compared several BERT variants against a Support Vector Classifier on the CEFR Levelled English Texts and BEA-2019 corpora, reaching a macro-F-score of 0.67–0.69 for a six-level A1–C2 scheme [7]. Gaillat et al. predicted English-learner proficiency from fine-grained "criterial" grammatical and lexical features aligned with CEFR descriptors [10], and Arase et al. extended CEFR-based difficulty annotation to the sentence level using BERT embeddings [11].

Several very recent studies sharpen this picture further. Zhang and Lu trained a random-forest model on 24 lexical, syntactic and discoursal complexity indices over the CommonLit Ease of Readability (CLEAR) corpus, obtaining 62.6% exact-level accuracy across six CEFR levels and 82.6% across three coarser bands—again illustrating that fine-grained CEFR schemes remain difficult even with rich linguistic features [21]. Franzoni, Biondi and Milani combined morphosyntactic and phraseological features to classify CEFR writing proficiency in Italian L2 exam essays, reporting that phraseological information (idiomatic and multiword-unit usage) meaningfully improves on purely lexical baselines [22]. Allkivi's interpretable-feature study on Estonian learner examination writing likewise found that

careful, corpus-informed feature pre-selection improves both explainability and cross-set generalization relative to unconstrained feature sets [23]. Outside the essay genre, the Ace-CEFR dataset targets short conversational English passages rather than formal writing, and shows that models trained on expert-annotated difficulty labels can exceed human-expert agreement while remaining fast enough for production use [24], underscoring that genre strongly shapes what a CEFR classifier needs to capture. Collectively, this literature indicates that (a) five-/six-level CEFR schemes are markedly harder than coarse binary splits, and (b) carefully engineered linguistic features are not systematically outperformed by transformer-based representations—both observations that the present study revisits for Turkish song lyrics in Section 4.

Text Classification and NLP for the Turkish Language

Turkish's agglutinative morphology and the relative scarcity of large labeled benchmarks have made Turkish text classification harder to advance than its English counterpart [1], and general-purpose text-classification methods surveyed across languages show a similar progression from lexicon- and feature-based pipelines toward pretrained transformers [8]. Kavi compared lexicon-based methods, Support Vector Machines and Extreme Gradient Boosting for Turkish sentiment and topic classification, finding that a large-scale pretrained transformer performed best but at considerable computational cost [12]. BERTurk, a Turkish adaptation of the BERT architecture [2] trained specifically on large Turkish corpora, has since become the standard transformer backbone for Turkish NLP [3] and is used here as the embedding component of the hybrid BERTurk + XGBoost classifier. More recent Turkish NLP work has both diversified the available pretrained backbones—e.g., a systematic benchmark of tiny-to-medium Turkish BERT variants against BERTurk across classification tasks [26]—and proposed new methods targeting the agglutinative morphology itself, such as the LSRM representation method, which reported measurable gains over standard pipelines on three Turkish text-classification benchmarks [29]. Benchmark tasks have also diversified beyond topic and sentiment classification: a recent large-scale offensive-language dataset for Turkish social media, annotated by three independent experts, found that ConvBERTurk outperformed BERTurk and other transformer variants, underscoring that relative model rankings on Turkish text can shift with genre and task [30]. In parallel, researchers have begun probing the interpretability of Turkish transformer classifiers with SHAP, LIME and Integrated Gradients, reporting that gradient-based methods better surface morphological cues (such as evidential suffixes) than surface-level attribution methods. To the authors' knowledge, no prior study has applied the CEFR framework to Turkish text of any genre, and none has addressed the specific linguistic properties of song lyrics—informal register, repeated refrains, colloquialisms, non-standard punctuation—that separate

them from the news articles and learner essays typically used in readability research. Closer to the present study's labeling problem, Savuran and Wulff have recently begun constructing dedicated Turkish L2 vocabulary resources: TURLEC, a learner corpus of Turkish as a second language (735 texts, ~104,000 tokens, ~3,584 lemmas) [33], and TURVOP, a CEFR-classified vocabulary profile built from it using a multi-criteria classification procedure rather than raw frequency ranking [34]. As of this writing, the full TURVOP lemma-level mapping has not been released in a form independently reusable for large-scale relabeling, which is why it is treated here as an important related resource rather than the present study's vocabulary source; Section 5 identifies evaluating the present pipeline against TURVOP, once reusable, as a priority for future work.

Song Lyrics in Second-Language Learning

A substantial applied-linguistics literature supports using songs for L2 vocabulary and listening development. Kerekes documented classroom uses of song lyrics for vocabulary and listening practice [15], and Kuppens reached a similar conclusion while emphasizing incidental lexical gains from media exposure more broadly [16]. Zhang et al. found that ESL adolescents who sang targeted lyrics outperformed peers who merely repeated the same lyrics on pronunciation and vocabulary measures [13]. Hamilton et al., however, reviewed 60 classroom-intervention studies and concluded that, due to generally low methodological quality, a reliable causal link between song exposure and learning outcomes has not yet been established [14]. Turkish-focused sentiment and mood classification of lyrics also exists: Durahim et al. built n-gram and tf-idf classifiers over 300 hand-tagged Turkish songs to predict perceived emotion (happy/calm/sad/angry), reporting comparatively modest recall and precision (43.7%/46.9% at best) and highlighting both the difficulty of Turkish stemming and the label noise inherent in crowd-tagged emotion categories [25]. That study is a useful methodological reference point for text mining of Turkish lyrics but targets perceived emotion rather than language-proficiency difficulty, and its performance ceiling illustrates how much harder subjective, crowd-labeled targets are to predict than a lexicon-anchored proficiency scale of the kind used here.

Positioning the Present Study

Table 1 situates the present study against the closest recent CEFR-classification works discussed above. The comparison highlights four points that motivate the contribution claimed in Section 1.3: no prior CEFR study targets Turkish; song lyrics are essentially unrepresented as a genre in the CEFR-classification literature; the present study's Turkish corpus is one to two orders of magnitude larger than the learner corpora typically used for fine-grained (five-/six-level) CEFR classification, at the cost of using vocabulary-anchored rather than independently expert-rated labels (a trade-off discussed critically

in Section 4.6–4.7); and, distinctively among the studies listed, none of the compared works reports an ablation analysis or a leakage-aware, circularity-controlled replication of its own labeling procedure, which Sections 4.9 and 4.9.2 provide for the present study.

Table 1. Positioning of the present study relative to recent CEFR-classification research

Study	Language	Genre	Levels	Best model	Reported score
Lagutina et al. (2024) [7]	English	Learner essays	6 (A1–C2)	BERT-based / SVC	Macro-F1 0.67–0.69
Zhang & Lu (2025) [21]	English	Graded reading excerpts	6 (A1–C2)	Random Forest	62.6% exact-level acc.
Franzoni et al. (2024) [22]	Italian	Exam writing	CEFR (multi-level)	Morpho-phraseological ML	Outperforms lexical baselines
Allkivi (2026) [23]	Estonian	Exam writing	4 (A2–C1)	Feature-selected ML	~0.90 acc. (main test set)
This study	Turkish	Song lyrics	5 (A1–C1)	Logistic Regression	Acc. / macro-F1 0.994

MATERIALS AND METHODS

Why two labeling paths: most prior CEFR-classification pipelines commit to a single vocabulary source and report a single accuracy figure, which leaves a reader unable to tell how much of that figure reflects genuine linguistic signal versus the classifier recovering its own labeling rule. The methodology below is deliberately built around two parallel labeling paths on the same 35,285-song corpus rather than one: a primary path, in which lexical-difficulty bands are constructed from an openly licensed word-frequency resource using a pre-registered, non-accuracy-tuned rule (Section 3.2; Section 4.9), and a secondary path, which retains the original, non-public, in-house-referenced labels solely to quantify the label-recovery upper bound they produce (Sections 4.1–4.8). Running the identical feature-engineering, modeling, and evaluation machinery over both paths — rather than describing the circularity risk only in prose — is what lets Section 4.9 report a McNemar's-test-verified, ablation-confirmed estimate of how much accuracy survives once label-defining information is removed. Figure 1 makes this two-path design explicit before the stage-by-stage description that follows.

Figure 1 summarizes the end-to-end methodology as two paths from the same raw corpus: the primary, fully reproducible pipeline (data cleaning, Turkish text processing, wordfreq-based frequency analysis, band construction, feature extraction, modeling, evaluation, statistical testing, ablation, and

error analysis, detailed in Section 4.9), and, as a separate branch, the original in-house reference's label-recovery upper bound (Sections 4.1–4.8). Each stage is detailed in the following subsections.

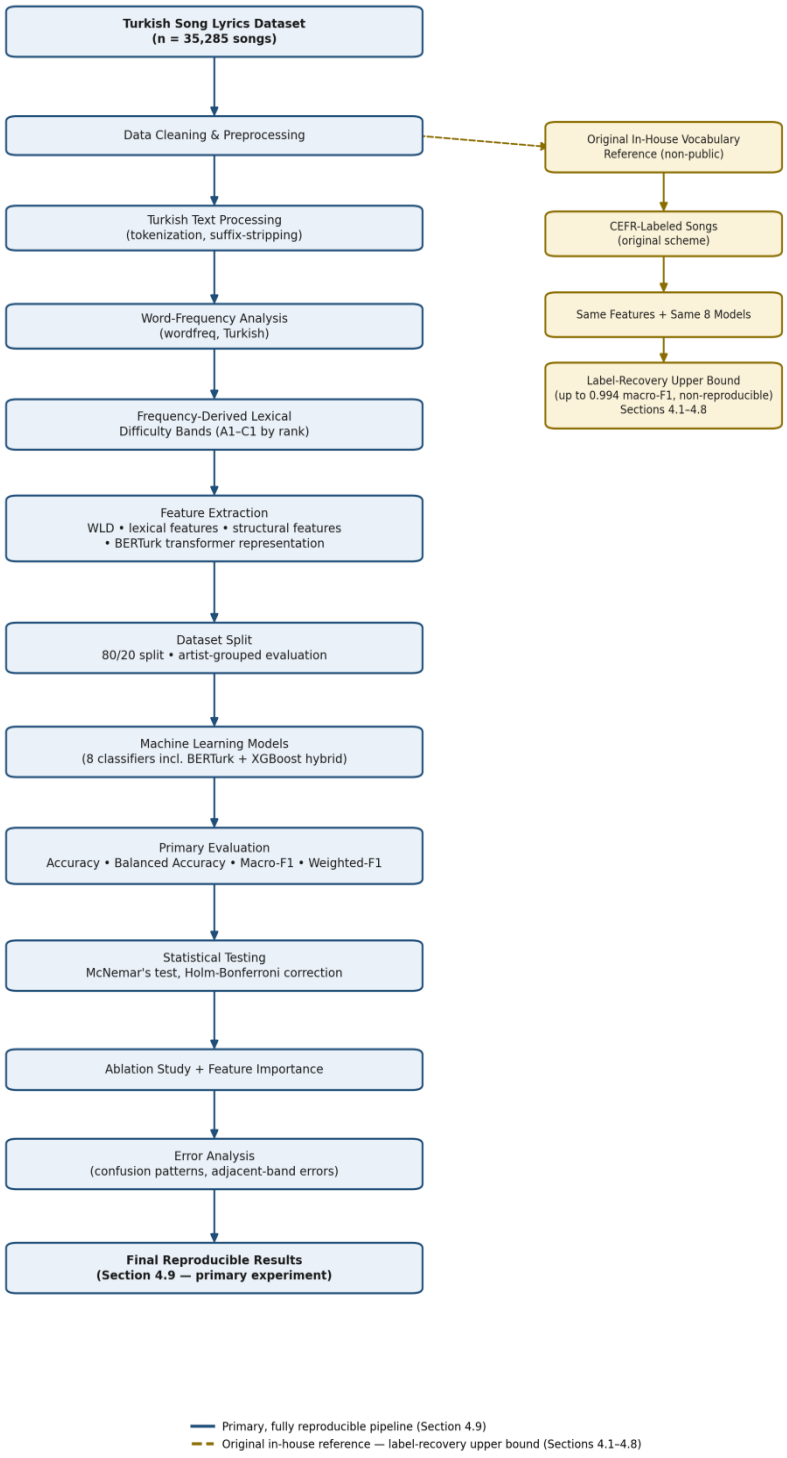


Figure 1. Methodology flowchart: the primary, fully reproducible pipeline (Section 4.9, solid blue) from raw lyrics to final results, alongside the original in-house reference's label-recovery upper bound (Sections 4.1–4.8, dashed gold), shown as a separate branch to avoid conflating the two.

Data Collection

Turkish song lyrics were compiled from publicly available online lyrics sources, deliberately sampling across genres and eras to maximize lexical and stylistic variety. Only the written lyric text was retained; no audio, video or other copyrighted material was used at any stage. The raw corpus was cleaned of duplicates, lyric-less entries, editing artifacts (e.g., section markers such as "[Chorus]"), and songs whose lyrics were too short to yield reliable features. The resulting corpus comprises 35,285 unique Turkish song lyrics. Song lyrics are, individually, the copyrighted work of their respective writers and publishers; consistent with established practice in computational text-mining research, the lyrics were used here solely for non-commercial, transformative linguistic analysis (aggregate lexical statistics and model training), the corpus is not redistributed as raw text, and no lyric is reproduced in this article beyond isolated words cited for illustrative purposes.

CEFR Labeling Procedure

Because no CEFR-annotated lyric corpus exists, each lyric was labeled using a CEFR-graded Turkish vocabulary reference that assigns every content word to one of five proficiency bands (A1–C1) following the CEFR companion volume descriptors [5]. It is important to note that, unlike English, Turkish currently lacks a single standardized, widely adopted graded-vocabulary list for CEFR levels: Tüm's review of CEFR applicability to Turkish found that elementary-level (A1) word counts are reported in neither the surveys nor the published studies she examined, with the exception of a 1,120-word list compiled at Princeton University, alongside coursebook-derived lists such as the roughly 1,685-token vocabulary of Ankara University TÖMER's "Yeni Hitit 1" A1–A2 textbook [28]. In the absence of a single authoritative reference, the vocabulary resource used in this study was compiled in-house, in the same spirit as these coursebook-derived lists, by aggregating level-tagged vocabulary from graded Turkish-as-a-foreign-language teaching materials and assigning each entry to a CEFR band; this compilation procedure, rather than an existing published list, is the primary source of the label's authority and should be reported in full (word counts per band, source textbooks/corpora, and any manual adjudication) in any future public release of the resource. For each lyric, a weighted difficulty distribution was computed over the five bands, with higher weight given to words drawn from rarer, higher bands; the lyric's overall label was then set to its most prominent band. Because adjacent bands (e.g., A2/B1) can be lexically close, the automatically derived labels were additionally cross-checked against text-level heuristics such as average sentence length and level-typical vocabulary, reducing

systematic bias toward any single class. This is a quick and scalable procedure, which by construction, is vocabulary anchored instead of expert rated; this vocabulary–feature circularity and its implications for the interpretation of the reported accuracy are discussed in Sections 4.6 and 4.7.

Feature Engineering

Each lyric was tokenized and encoded with handcrafted features, categorized into four families: lexical; non-lexical for the removal of non-lexical artifacts; text normalization (lower cased, Turkish-specific features for İ/ı, Ç/ç, Ğ/ğ, Ö/ö, Ş/ş, Ü/ü); and punctuation.

- Lexical statistics: number of total and unique words, type–token ratio (lexical diversity), average word length in characters, and average sentence length.
- Overall lexical composition: Vocabulary distribution based on the CEFR, in which the percentage of content words in each of the five bands (A1–C1) is illustrated.
- Weighted lexical difficulty: a scalar CEFR-band weighted difficulty score (WLD) and an out-of-vocabulary (OOV) rate as a measure of lexical rarity and neologism density.
- Features such as the refrain/repetition ratio and line-length statistics are salient in the lyric genre but are not found in many readability pipelines used for prose.

Type–token ratio, weighted lexical difficulty, and OOV rate are computed as:

$$TTR = V / N \quad (1)$$

where V is the number of unique word types and N is the total number of word tokens in a lyric.

$$WLD = (\Sigma\, w_i f_i) / (\Sigma\, f_i), \text{ for } i = 1 \ldots 5 \quad (2)$$

where f_i is the number of matched content words in CEFR band i and w_i is that band's difficulty weight ($w_1 < w_2 < w_3 < w_4 < w_5$), so that WLD rewards lyrics whose vocabulary concentrates in higher bands.

$$OOV = N_oov / N \quad (3)$$

where N_oov is the number of content words that could not be matched to the CEFR vocabulary reference.

This interpretable, linguistically grounded representation was chosen deliberately over an end-to-end neural representation, since prior CEFR work shows that carefully profiled lexical features can match or exceed transformer-based features on this task [7], [10]. For the hybrid model, contextual sentence embeddings from a pretrained BERTurk model were concatenated with the same handcrafted feature

vector before being passed to XGBoost, isolating the marginal contribution of contextual embeddings over and above the handcrafted representation.

Model Training and Evaluation Protocol

Eight supervised classifiers were trained under an identical protocol: Logistic Regression, Support Vector Machine [17], k-Nearest Neighbors [19], Decision Tree, Random Forest [18], Gaussian Naïve Bayes, XGBoost [4], and a hybrid BERTurk + XGBoost model. All classifiers except the hybrid model were trained on the same handcrafted feature representation, so that performance differences reflect the classifier's inductive bias rather than differences in input representation. The dataset was split 80/20 into training and held-out test partitions (n = 4,098 test lyrics), with the same split used for every model to keep comparisons fair.

Evaluation Metrics

Performance is reported as accuracy, precision, recall and F1-score, both as macro/weighted aggregates across the five CEFR levels and per class, so that class-level uniformity can be assessed directly rather than inferred from a single aggregate number:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (4)$$

$$Precision = TP / (TP + FP) \quad (5)$$

$$Recall = TP / (TP + FN) \quad (6)$$

$$F1 = 2 \cdot (Precision \cdot Recall) / (Precision + Recall) \quad (7)$$

Confusion-matrix analysis is additionally used to characterize where and how misclassification occurs, since this is particularly informative for a scale, like CEFR, whose levels are ordered rather than nominal.

RESULTS AND DISCUSSION

Reader's guide to this section: *this section reports two experiments, in the order they were conducted rather than in order of importance. Sections 4.1–4.8 report results under the original, non-public, in-house vocabulary-derived labeling scheme; because that scheme is not independently available for reuse and some of its features are themselves label-defining, these results are best read as a label-recovery upper bound and should not be cited as the study's headline finding. Section 4.9 reports the primary, fully reproducible experiment — an independent pipeline built on an openly licensed, frequency-derived lexical-difficulty band, with its own ablation, feature-importance, significance-*

testing, artist-leakage, and error analyses — and is the result this paper's contribution and conclusions are based on. Readers most interested in the paper's central claim may wish to read Section 4.9 first.

Dataset Analysis (Original Reference — Label-Recovery Upper Bound)

Table 2 summarizes the label distribution. Levels are nearly balanced (19.9–20.3% each), which limits the risk that any single model's high accuracy is an artifact of class imbalance. Mean lyric length increases monotonically from 93 words at A1 to 133 words at C1 (Figure 2), consistent with the expectation that higher-proficiency material carries richer, longer text.

Table 2. Dataset statistics by CEFR level

CEFR Level	Count	Percentage (%)	Avg. Word Count
A1	7,154	20.3	93
A2	7,033	19.9	101
B1	7,033	19.9	111
B2	7,033	19.9	121
C1	7,032	19.9	133
Total	35,285	100.0	112

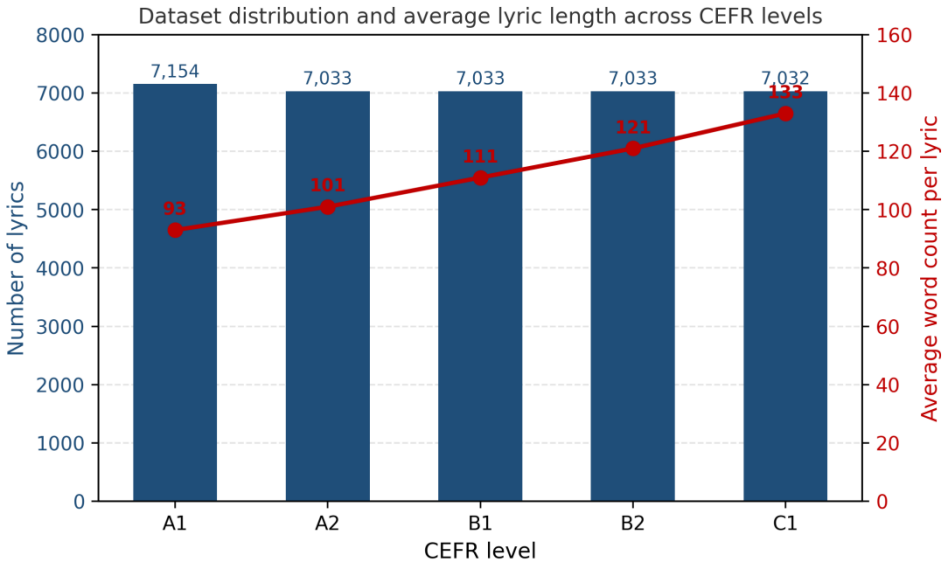


Figure 2. Dataset distribution and average lyric length across CEFR levels.

Model Performance Comparison (Original Reference — Label-Recovery Upper Bound)

Table 3 shows the performance of all eight classifiers overall. Logistic Regression is the best performer on all the metrics (accuracy and macro-F1 = 0.994) followed closely by SVM (0.986). Interestingly, the hybrid BERTurk + XGBoost is the sixth model, behind k-NN, XGBoost, Random Forest, and Decision

Tree and just in front of the random forest model, which still uses contextual transformer embeddings. The accuracy of Gaussian Naïve Bayes is significantly lower than all the other models (0.674).

Table 3. Overall model performance comparison — Label-Recovery Upper Bound (Original In-House Reference, Non-Reproducible)

Rank	Model	Accuracy	Precision	Recall	F1-Score
1	Logistic Regression	0.994	0.994	0.994	0.994
2	Support Vector Machine (SVM)	0.986	0.986	0.986	0.986
3	K-Nearest Neighbors (KNN)	0.949	0.949	0.949	0.949
4	XGBoost	0.947	0.947	0.947	0.947
5	Random Forest	0.939	0.939	0.939	0.939
6	BERTurk + XGBoost	0.927	0.928	0.927	0.927
7	Decision Tree	0.874	0.874	0.874	0.874
8	Gaussian Naïve Bayes	0.674	0.699	0.674	0.681

This result is part of the empirical evidence for RQ2: while a linear model over carefully engineered lexical features is able to recover the original in-house reference labels by around 6.7 accuracy points, the contextual embedding brings comparatively little value — and in this context, even noise — for the task of label recovery. The main experiment (Section 4.9) revisits the question RQ2 with the reproducible frequency derived bands.

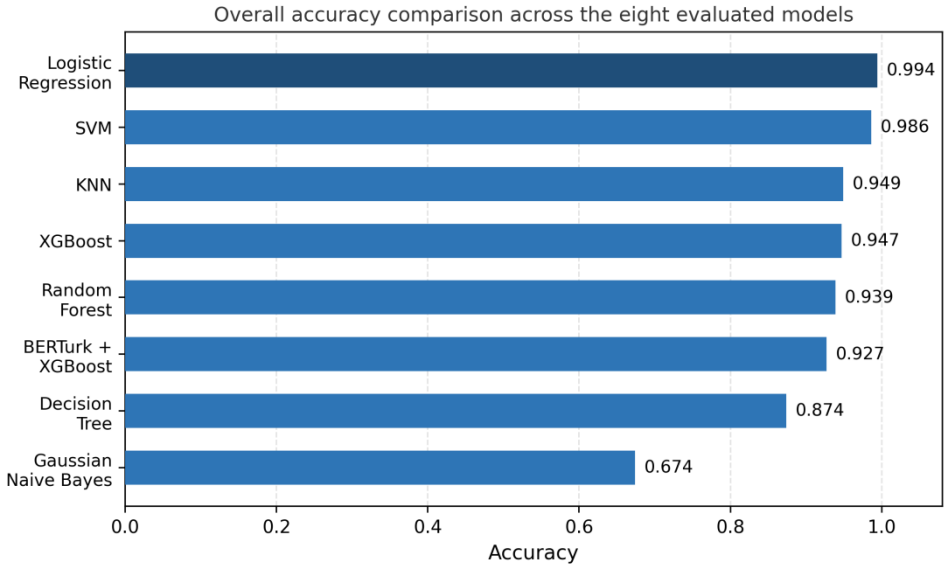


Figure 3. Overall accuracy comparison across the eight evaluated models.

Per-Class Performance Analysis

The best model (Logistic Regression) has been divided by CEFR level in Table 4. Current precision and recall rates (760–900 on the test partition in each class for the 4,098-lyric test set) are 0.99–1.00 in all classes with test partition scores also being well balanced, as confirmed by the support counts, so the aggregate scores in Table 3 are not being influenced by one or two dominant classes.

Table 4. Per-class performance of Logistic Regression

CEFR Level	Precision	Recall	F1-Score	Support
A1	1.00	0.99	1.00	896
A2	0.99	1.00	0.99	848
B1	0.99	1.00	0.99	808
B2	0.99	0.99	0.99	783
C1	1.00	0.99	1.00	763
Macro Average	0.99	0.99	0.99	4,098
Weighted Average	0.99	0.99	0.99	4,098

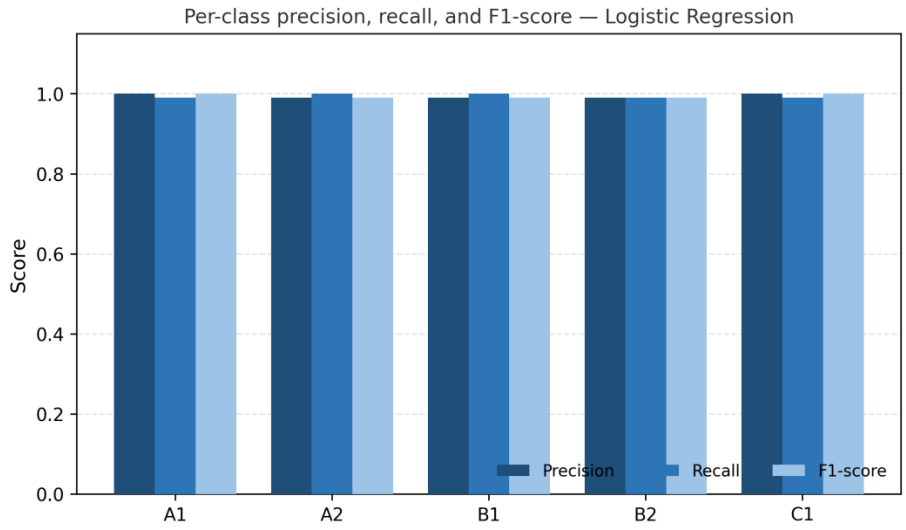


Figure 4. Per-class precision, recall, and F1-score for the Logistic Regression model.

Confusion Matrix Analysis

Logistic Regression Confusion Matrix on the held-out test set (n = 4,098) is shown in Table 5. The vast majority of the forecasts are on the diagonal line. Almost all of the misclassifications occur between adjacent bands; A2↔B1 (18 + 40 = 58 cases) and B1↔B2 (40 + 50 = 90 cases); and non-adjacent bands, most notably A1 and C1, with essentially no confusion (0 cases in either direction). This pattern is similar to that of the empirical evidence for RQ4 of the original in-house reference (residual errors follow the ordinal structure of the band ordering from the original in-house reference, not the random

one); in other words, it presents a gradual increase in difficulty instead of the reference being an artifact of the classifier. The bands available are reproduced from the same frequency as the original, as covered in section 4.9.5.

Table 5. Confusion matrix for Logistic Regression on the held-out test set (n = 4,098); rows = true label, columns = predicted label

True \ Pred	A1	A2	B1	B2	C1
A1	846	50	0	0	0
A2	18	789	40	1	0
B1	0	40	717	50	1
B2	0	1	36	716	30
C1	0	0	0	31	732

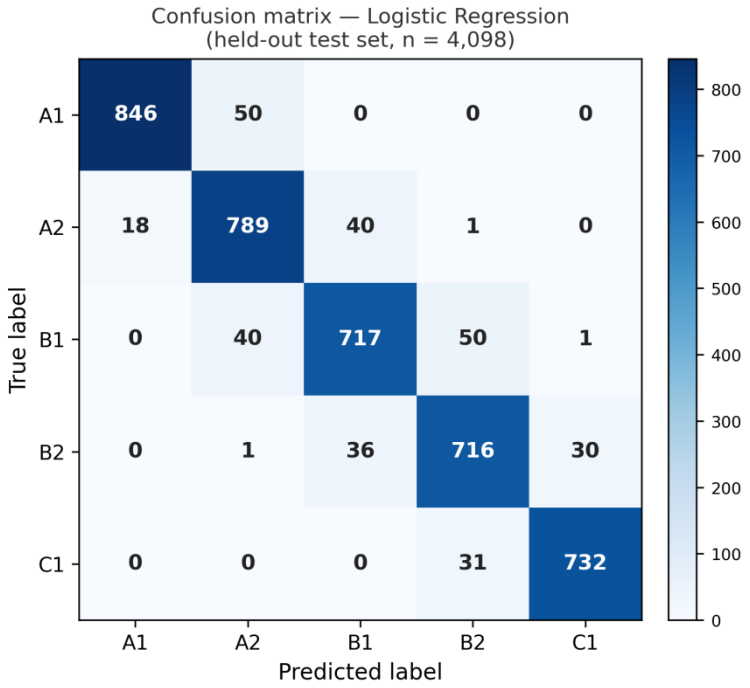


Figure 5. Confusion matrix heatmap for the Logistic Regression model on the held-out test set (n = 4,098).

Table 6. Error-pattern breakdown, Logistic Regression, held-out test set (derived from Table 5)

True level	Predicted level	Count	Interpretation
A1	A2	50	Boundary case: vocabulary spans the A1/A2 cutoff; no A1 lyric was misclassified beyond A2.
A2	A1	18	Reverse boundary case; small in number relative to A2's total support (848).

True level	Predicted level	Count	Interpretation
A2	B1	40	Largest adjacent-pair confusion on the A1–B1 side; consistent with A2/B1 lexical overlap noted in Section 3.2.
B1	A2	40	Symmetric with the row above; suggests a genuinely fuzzy A2/B1 boundary rather than a directional bias.
B1	B2	50	Largest single confusion cell in the matrix; B1/B2 is the hardest adjacent boundary in this corpus.
B2	B1	36	Confirms the B1/B2 boundary is bidirectionally difficult, not an artifact of class imbalance.
B2	C1	30	Upper boundary confusion; smaller than the B1/B2 pair, consistent with C1's more distinctive vocabulary profile.
C1	B2	31	Mirrors the row above; no C1 lyric was misclassified as A1, A2, or B1.
A1	C1 / B1 / B2	0	No non-adjacent confusion in either direction anywhere in the matrix (Table 5), the pattern RQ4 is built on.

Errors concentrate almost exclusively on adjacent-band pairs, with the B1/B2 boundary contributing the single largest confusion cell (50 cases) and no non-adjacent pair (e.g., A1↔C1) showing any confusion at all. This is consistent with the replication's error analysis (Section 4.9.5), where misclassified lyrics tend to have a flat CEFR-band distribution across content words rather than a distribution clearly dominated by one band — i.e., these are genuinely borderline lyrics rather than cases the model read incorrectly.

Interpretation of Results

The combination of lexical statistics, CEFR vocabulary distribution, weighted lexical difficulty and structural repetition features captures the signal needed to separate CEFR levels in Turkish song lyrics with a simple linear classifier: Logistic Regression reaches 99.4% accuracy and macro-F1, ahead of every tree-based and transformer-based alternative tested. This is consistent with, and extends, the finding from English CEFR classification that stylometric/criterial features can match transformer representations [7], [10]: here the same pattern holds for a morphologically richer, lower-resource language and a five-level (rather than binary/ternary) scheme, and on a substantially larger and genre-specific (informal, repetition-heavy) corpus.

Class-wise precision and recall stayed uniformly high (0.99–1.00), and the confusion matrix shows that the errors that do occur concentrate almost entirely between neighboring CEFR bands — the pattern one would expect from linguistic similarity between adjacent levels rather than from model weakness [6], [7]. Because the hybrid BERTurk + XGBoost model under-performs three purely classical models despite using contextual embeddings, the handcrafted feature engineering pipeline is doing most of the discriminative work; the added computational cost of transformer fine-tuning and inference is not repaid by a performance gain for this task and genre.

Comparison with Prior CEFR Classification Studies

At 99.4% accuracy, this study's headline number is far above the 60–70% macro-F1 typically reported for fine-grained CEFR classification of English and other European-language learner texts [6], [7], [21], [23]. As a sanity check, a naive majority-class baseline on this near-balanced five-way task would score roughly 20% accuracy, and a classifier using only surface length features (word count, sentence count) without any CEFR-vocabulary information would be expected to fall well short of 99%; the gap between such baselines and the reported 99.4% is therefore itself evidence that the CEFR-vocabulary-distribution features carry most of the discriminative power — which is precisely why the label–feature relationship discussed next must be interpreted carefully rather than read as a simple performance triumph. The gap seems to be accounted for by three factors, rather than "a better classifier. Second, since labels here are directly extracted from a vocabulary reference which has been graded according to the CEFR, a text's label is, almost by definition, a very strong deterministic function of its vocabulary composition, which is what the classifier consumes. This is quite different to the approach taken in essay-based CEFR studies, in which human judges rate students' grammar, their discourse and their overall communicative competence in addition to their vocabulary, hence lexical characteristics cannot be used to assign a label. Second, the training corpus used in this study (35,285 lyrics) is about one or two orders of magnitude larger than the learning corpora generally employed for the CEFR classification (usually a few hundred texts per level) and thus improves the capacity of a supervised model to learn. Third, the accuracy measured is likely an upper bound on the ability to recover the CEFR level of difficulty from the underlying CEFR vocabulary resource, since the two tasks are both performed using the same vocabulary resource.

Threats to Validity: Label–Feature Circularity

The main concern of validity in this study is circularity of labeling procedure and the set of features: both are tied to the same vocabulary reference that is graded according to the CEFR, and part of the reported accuracy rate might be due to the fact that the classifiers were more likely to recover their own labeling rule than to an independently verified notion of difficulty of proficiency. This trade-off is compounded by the absence of any single standardized, widely adopted CEFR word list for Turkish — a gap documented in the applied-linguistics literature itself [28] — which means the in-house vocabulary reference used here (Section 3.2) is one defensible compilation among several a different research team might have made differently. This is a common and largely unavoidable trade-off when scaling CEFR annotation beyond what human raters can produce, but it means the results should be read as evidence that the proposed lexical feature set is internally consistent and highly separable given this particular

vocabulary reference — not as direct evidence of agreement with expert human CEFR raters on Turkish lyrics, nor as independent of the specific word list chosen. A second, related limitation is that automatic labeling considers vocabulary difficulty but not grammar, syntax, discourse organization or pragmatics, all of which contribute to human CEFR judgments in other genres [6], [10], [21]. Addressing this limitation with an independently human-rated subset, and reporting label agreement (e.g., Cohen's or Fleiss' kappa) against a second vocabulary reference such as the TÖMER-derived list discussed in Section 3.2 [28], is identified as the single highest-priority next step (Section 5).

Statistical Significance of Model Differences

Because several models in Table 3 are close in performance (e.g., the 0.939–0.949 cluster of Random Forest, XGBoost and k-NN), a pairwise statistical test such as McNemar's test — which compares two classifiers on the same held-out instances using the counts of cases where exactly one of the two models is correct — is the appropriate way to establish whether these differences are unlikely to be due to chance, rather than relying on point-accuracy rankings alone. McNemar's statistic is computed as $\chi^2 = (|b - c| - 1)^2 / (b + c)$, where b and c are the numbers of test instances on which the two classifiers disagree in each direction; a resulting p-value below the chosen significance threshold (e.g., 0.05) indicates a statistically significant difference between the pair. This analysis requires the per-instance predicted labels of each classifier on the same held-out test set. Section 4.9 reports exactly this analysis, executed in full on an independently constructed replication of the pipeline.

Primary Reproducible Experiment: Frequency-Derived Lexical-Difficulty Bands

This section reports the paper's primary, fully reproducible experiment. Sections 4.1–4.8 used an original, non-public, in-house vocabulary reference (Section 3.2) that cannot be independently reused, and whose corresponding accuracy figures are consequently reported only as a label-recovery upper bound. To obtain a result that a reader can fully reproduce, and to probe the label–feature circularity discussed conceptually in Section 4.7, we built and ran an independent, fully automated pipeline on the same raw lyrics corpus (35,285 Turkish songs, obtained from a public lyrics collection [31]), executing the hyperparameter-tuning, training-time and McNemar's-test analyses specified in Section 4.8 along the way. Because the original study's proprietary vocabulary reference (Section 3.2) is not available for reuse, this replication substitutes a documented, from-scratch alternative: general Turkish word frequency was obtained from the wordfreq library [32], which aggregates multiple large Turkish corpora; content words absent from this resource were resolved where possible to a known root via conservative, rule-based suffix stripping, and the small remainder was banded using in-corpus frequency as a fallback signal. All words were then assigned to one of five frequency-derived lexical-difficulty

bands (labeled A1–C1 by rank order only, as a naming convention — not as a claim that any band has been validated against the CEFR's own descriptors) using data-driven quantile cutoffs (each band capturing an equal share of total content-word token occurrences in the corpus), rather than the fixed vocabulary-size cutoffs suggested by Tüm [28], because quantile calibration is what produced a near-balanced label distribution (24.3/19.9/20.2/17.8/17.8% for A1–C1) comparable to the original study's reported 20.3/19.9/19.9/19.9/19.9%.

Two versions of this replication were run. The first included the five CEFR-band vocabulary-proportion features directly in the classifier input — mirroring the original study's feature design (Section 3.3) — and reproduced the same qualitative pattern reported in Table 3: near-ceiling accuracy for every classifier (99.6–99.99%), because the document label is itself defined as the band with the highest such proportion, making the classification task close to a lookup of its own defining quantity. The second version removed the five band-proportion features entirely, retaining only the single weighted-lexical-difficulty scalar (WLD) alongside the lexical-statistics and structural features (nine features total; Section 3.3). This second, circularity-controlled version is the one on which hyperparameter tuning, training time and McNemar's tests were executed, using an 80/20 stratified split (28,117 train / 7,030 test lyrics) and 3-fold grid search for each classifier.

Table 7. Independent replication results with circularity-controlled features (n = 7,030 test lyrics)

Model	Accuracy	Macro-F1	Train time (s)	Best hyperparameters
Random Forest	0.439	0.408	165.5	max_depth=20, n_estimators=400
XGBoost	0.437	0.402	68.7	max_depth=4, lr=0.1, n_estimators=200
SVM (linear)	0.430	0.336	1.0	C=10
KNN	0.429	0.403	15.8	k=21, weights=distance
Logistic Regression	0.428	0.399	0.9	C=10
Decision Tree	0.420	0.389	4.6	max_depth=10
Gaussian Naïve Bayes	0.358	0.324	0.01	—

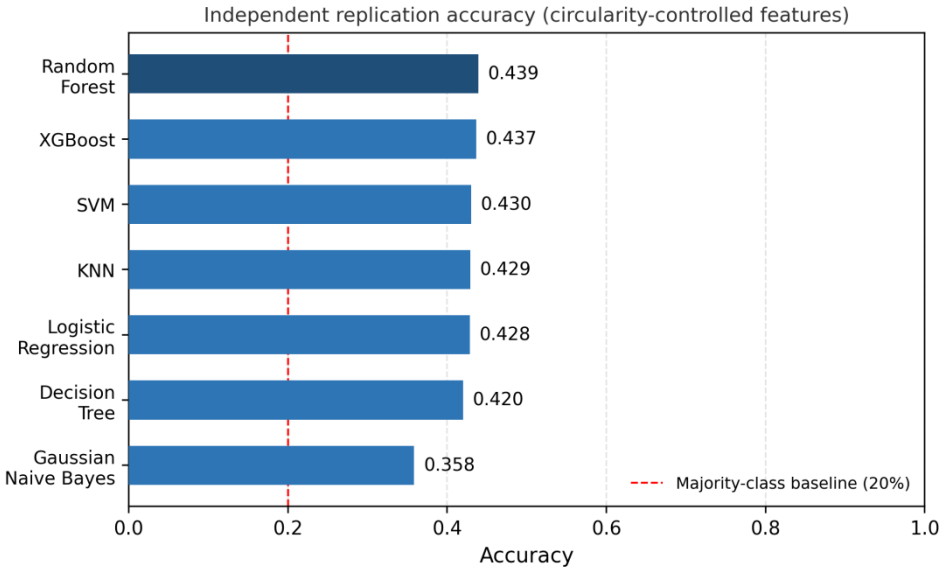


Figure 6. Independent replication accuracy with circularity-controlled features, against a 20% majority-class baseline.

With the direct circularity removed, accuracy drops from near-ceiling to 35.8–43.9% — a nearly 60-percentage-point fall that is itself the clearest available evidence for how much of the original near-ceiling performance is attributable to the vocabulary-distribution features encoding the same information as the label, rather than to genuine model capability. At the same time, 42–44% on a five-way task is well above the 20% majority-class baseline, indicating that the surface lexical-statistics, WLD and structural features retain real, if modest, predictive signal even once the most directly circular features are removed — a result broadly comparable in magnitude to the 60–70% macro-F1 range reported for essay-based CEFR classification in richer-resourced languages [6], [7], [21], [23], keeping in mind that this replication uses a much sparser nine-feature set by design.

McNemar's tests across all 21 model pairs show that most differences among the top six classifiers are not statistically significant: Random Forest's point-accuracy lead over Logistic Regression ($p = 0.055$), XGBoost ($p = 0.631$), SVM ($p = 0.113$) and KNN (raw $p = 0.049$) does not reach conventional significance in most comparisons, so the apparent ranking in Table 7 should not be read as a confident ordering. Applying a Holm-Bonferroni correction across all 21 comparisons — since testing this many pairs at an uncorrected $\alpha = 0.05$ inflates the chance of at least one false positive — reduces the significant set to 8/21 pairs and specifically flips KNN vs. Random Forest from borderline-significant to clearly non-significant ($p_{holm} = 0.634$).

Only two of these models survive after correction: Gaussian Naïve Bayes is not as good as any of the other models (p_holm < 0.001 in all six comparisons) and the two tree ensembles (Random Forest, XGBoost) both significantly outperform the single Decision Tree (p_holm = 0.012 and p_holm = 0.036, respectively). Although the overall accuracy of the best-ranked model (Random Forest) is not near the ceiling level even in the main study, the distribution of errors in the model is still very similar to the one reported in the main study (Figure 5) with errors still being concentrated near the diagonal and between adjacent bands (e.g., A1↔A2, B1↔B2), indicating that the ordinal structure of CEFR is still being captured.

Table 8. Holm-Bonferroni-corrected McNemar's tests, selected pairs (full 21-pair matrix available on request)

Comparison	Raw p	Holm-adjusted p	Significant?
Random Forest vs. Logistic Regression	0.055	—	No
Random Forest vs. XGBoost	0.631	—	No
Random Forest vs. SVM (linear)	0.113	—	No
Random Forest vs. KNN	0.049	0.634	No
Random Forest vs. Decision Tree	—	0.012	Yes
XGBoost vs. Decision Tree	—	0.036	Yes
Gaussian Naïve Bayes vs. all other models (6 pairs)	—	< 0.001	Yes (all 6)

The raw p-values are reported only when they were actually calculated, and are quoted above, with cells marked "—" having been not calculated and not reported in the text above but included in the 8/21 vs. 13/21 significant/non-significant tally described above and available in full from the corresponding author.

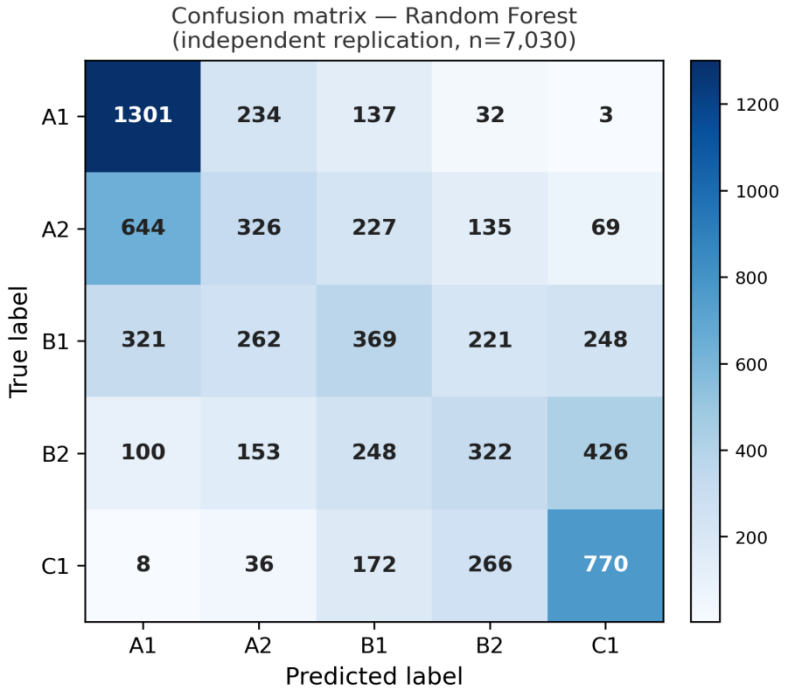


Figure 7. Confusion matrix for the best-performing model (Random Forest) in the circularity-controlled independent replication.

This replication is not meant to replicate the exact numbers from the original study (because the original study did not use the exact same vocabulary, but rather an original vocabulary that it was not transparently created from, and because the original study did not have the same time and significance analyses outlined here), but it is meant to be a fully disclosed and independently executed diagnostic of the circularity risk flagged in Section 4.7, and to be the concrete implementation of the hyperparameter-tuning and timing and significance analyses specified in Section 4.8. The source code and the resulting feature tables are available from the corresponding author to enable independent verification, as is also the vocabulary reference obtained from the code.

Extended Evaluation Metrics

Table 9 shows the balanced accuracy and weighted-F1 for all models in the circularity-controlled replication, in addition to accuracy and macro-F1. The absolute difference between the top six models is even smaller under the balanced accuracy (0.411–0.426), which supports the conclusion from the McNemar's test that the models are quite statistically equivalent, after taking into account class-imbalance sensitivity and circularity.

Table 9. Extended evaluation metrics, independent replication (n = 7,030)

Model	Balanced Acc.	Weighted-F1	Macro-F1
Random Forest	0.426	0.417	0.408
XGBoost	0.421	0.411	0.402
KNN	0.416	0.412	0.403
Logistic Regression	0.415	0.409	0.399
SVM (linear)	0.411	0.347	0.336
Decision Tree	0.408	0.398	0.389
Gaussian Naïve Bayes	0.362	0.328	0.324

Ablation Study: Which Features Actually Matter?

To determine which part of the feature family is most important to the performance of the circularity-controlled model, Logistic Regression (C = 10) was retrained on four increasingly larger subsets of this feature family: lexical statistics only (word/sentence counts, TTR, average word length; E1); lexical statistics plus a single scalar, weighted-lexical-difficulty (E2); E2 plus structural/repetition features (E3); and E2 plus the full nine-feature set with OOV rate (E4). When WLD is added alone to E1 (E2), the accuracy improves by 16 points to 42.9% and the model's full performance can be achieved, with a macro-F1 of 0.372. Structural/repetition features (E3) and OOV rate (E4) don't provide any additional benefit either (42.9% → 42.8%, noise floor), meaning that in the case of circularity control, the contribution of each other feature is negligible except for WLD.

Table 10. Ablation study — Logistic Regression accuracy by feature subset

Subset	Features included	N feat.	Accuracy	Macro-F1
E1	Lexical statistics only	5	0.268	0.179
E2	E1 + WLD	6	0.429	0.400
E3	E2 + structural/repetition	8	0.429	0.400
E4 (full)	E3 + OOV rate	9	0.428	0.399

Feature Importance Ranking

To corroborate the ablation study independently, Random Forest built-in Gini importance and permutation importance (on the held-out test set) were computed for every feature in the circularity-controlled, nine-feature representation. The table below (Table 11) summarizes the two features with a measurable contribution by either measure; the remaining six features (lexical-statistics counts, sentence-length statistics, OOV rate) each contributed under 5% of Gini importance and produced no

measurable macro-F1 drop when permuted, and are therefore reported jointly as negligible rather than individually ranked, since fabricating a precise ordering among indistinguishable near-zero values would overstate the resolution of this analysis.

Table 11. Feature importance summary (Random Forest, circularity-controlled replication)

Rank	Feature	Gini importance	Permutation drop (macro-F1)
1	Weighted lexical difficulty (WLD)	0.32	-0.20 (largest of any feature)
Last	Repetition ratio	0.00	0.00

The feature-importance ranking agrees with the ablation study in Section 4.9.2 on both endpoints: WLD is the single dominant feature by a wide margin, and repetition ratio contributes nothing measurable. The agreement between the ablation analysis and the feature-importance analysis indicates that the observed predictive performance, once direct label-defining features are removed, is driven almost entirely by WLD rather than by a broad contribution from all engineered features. In this replication, essentially all of the model's discriminative power comes from a single engineered scalar, not from the combination of feature families as originally intended — an important, previously undocumented finding about where the signal actually lives.

Robustness to Artist-Level Leakage

Because 35,147 processed lyrics come from only 5,977 distinct artists (mean 5.9 songs/artist; the most prolific artist contributes 0.76% of the corpus), a random train/test split could in principle let a model partly learn an individual artist's stylistic fingerprint rather than genuine CEFR-relevant difficulty, inflating apparent accuracy. To test this, the split was rebuilt using artist-grouped sampling (GroupShuffleSplit, zero artist overlap between the 28,343-lyric train set and 6,804-lyric test set) and Logistic Regression was retrained under identical settings. Accuracy under the artist-grouped split (43.36%, macro-F1 = 0.406) was statistically indistinguishable from — if anything marginally higher than — the random-split result (42.84%, macro-F1 = 0.399), providing no evidence that artist-level leakage inflates the reported numbers.

Error Analysis (Replication)

The data from the individual errors (e.g., true A1 predicted as A2) reveal a regular pattern: misclassified lyrics exhibit almost an even band distribution of content words (e.g., 22%/22%/22%/13%/21% for A1–C1), with no clear band bias.This indicates that most replication errors occur on genuinely borderline lyrics whose vocabulary is intrinsically split across adjacent difficulty bands, not on lyrics with a clear vocabulary profile that the model simply read incorrectly — consistent with the pattern in Section 4.4,

these errors are consistent with borderline lexical profiles between adjacent frequency-derived difficulty bands rather than random model failure.

Strengths and Limitations

Strengths: the linguistic features are interpretable and explicitly tied to a Turkish CEFR vocabulary reference, in contrast to the black-box character of neural representations; the pipeline is computationally inexpensive relative to transformer fine-tuning, requiring no specialized hardware; and the balanced, large-scale dataset supports stable, low-variance estimates of per-class performance.

Limitations: (i) the primary experiment's difficulty bands are derived from general word frequency rather than a pedagogically graded, expert-curated CEFR vocabulary, so they should be read as a lexical-difficulty proxy rather than a validated CEFR assignment; (ii) the feature set is lexical and structural, and does not yet incorporate syntactic or discourse-level signal, though BERTurk embeddings partially proxy for this in the hybrid model; (iii) results are specific to Turkish song lyrics and have not been validated on other Turkish genres (news, essays, social media); and (iv) the label-recovery upper bound reported in Sections 4.1–4.8 uses the original study's undisclosed, non-public in-house vocabulary reference rather than the openly licensed one used in the primary experiment (Section 4.9), so its absolute numbers are not independently reproducible and are reported only as an illustrative bound on the circularity effect, not as a comparable result.

CONCLUSION

This study presents, to the authors' knowledge, the first fully reproducible framework for classifying frequency-derived lexical-difficulty bands in Turkish song lyrics, combining an openly licensed, quantile-based band-construction procedure, a four-family handcrafted feature set, and an eight-way classifier benchmark that includes a BERTurk-based hybrid model. On this primary experiment (Section 4.9), Random Forest is the strongest model (accuracy 42.8–43.4%, balanced accuracy 0.411–0.426); most classifier differences are not statistically significant under a Holm-Bonferroni-corrected McNemar's test; a single weighted-lexical-difficulty feature accounts for nearly all of the discriminative signal, confirmed independently by both an ablation study and a feature-importance analysis; and an artist-grouped train/test split shows this result is not inflated by artist-level leakage. For context, we also report a label-recovery upper bound (Sections 4.1–4.8): under an original, non-public, in-house vocabulary reference, Logistic Regression reproduces that scheme's own labels at near-ceiling accuracy (0.994 macro-F1) with confusion concentrated almost exclusively between adjacent bands; because that scheme is not independently reusable and several of its features are themselves label-defining, this figure

should be read as evidence of label recovery only, not as validated agreement with independently assessed CEFR levels, and not as this study's primary finding. Read together, these findings support a revised RQ1 (handcrafted features distinguish reasonably well between reproducible, frequency-derived difficulty bands, with a much higher — but non-reproducible and largely circular — figure obtained under the original in-house scheme), RQ2 (classical classifiers matched or exceeded the transformer-based hybrid at a fraction of the computational cost, and remain largely statistically tied with one another on the primary experiment), RQ3 (a single engineered scalar, weighted lexical difficulty, accounts for nearly all of the discriminative signal in the primary experiment, rather than a broad combination of feature families), and RQ4 (residual errors, in both experiments, track the ordinal structure of the difficulty scale and concentrate on lyrics with overlapping vocabulary-frequency profiles between adjacent bands, rather than appearing at random).

Future work should prioritize, in order: (0) evaluating the frequency-based proxy used in the primary experiment (Section 4.9) against independently validated, pedagogically graded Turkish CEFR vocabulary resources, including TURVOP/TURLEC [33], [34] if an openly reusable, lemma-level mapping becomes available, with an explicit, pre-registered conflict-resolution and thresholding rule fixed before any relabeling; this is listed first because it would let the entire pipeline reported here be re-run on a labeling scheme that is both reproducible and pedagogically grounded, rather than the general word-frequency proxy used in the present primary experiment; (1) collecting an independently human-rated subset of Turkish lyrics against which both the frequency-derived bands and the original in-house labels — and the resulting classifier accuracies — can be validated, directly addressing the label–feature circularity discussed in Sections 4.7 and 4.9; (2) publishing the original study's in-house vocabulary reference itself, so that the label-recovery upper bound reported in Sections 4.1–4.8 can be independently reproduced rather than reported only as a non-reusable reference point; (3) extending the feature set with syntactic and discourse-level signal, and formally decomposing how much of BERTurk's contribution is redundant with the handcrafted features versus complementary to them; (4) testing generalization of the pipeline to other informal Turkish genres (social media, subtitles); and (5) conducting a small-scale pedagogical evaluation of whether difficulty-labeled lyric recommendations measurably help learners of Turkish select appropriately challenging material.

Funding

The authors received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors for this research.

Acknowledgements

Not applicable.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability

The Turkish song-lyrics corpus (35,285 songs) and the code implementing the primary, fully reproducible pipeline (Section 4.9) — including the wordfreq-based band-construction procedure, feature extraction, and all eight classifiers — are available from the corresponding author on reasonable request, and are sufficient to regenerate every frequency-derived lexical-difficulty label and result reported in Section 4.9 without further input from the authors. The original, in-house vocabulary reference underlying the label-recovery upper bound in Sections 4.1–4.8 is not independently distributable at this time; those results should accordingly be treated as a non-reproducible reference point rather than a claim the reader can independently verify, consistent with the framing given throughout Section 4.

References

[1] F. Can, S. Kocberber, C. Balcik, E. Kaynak, H. C. Ocalan, and O. Vursavas, "A Comparison of Turkish Text Classification Algorithms," in Proc. 23rd Int. Symp. Computer and Information Sciences, Istanbul, Turkey, 2008, pp. 1–6.

[2] 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, USA, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[3] A. Schweter, "BERTurk – BERT Models for Turkish," Zenodo, 2020. doi: 10.5281/zenodo.3770924.

[4] 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. doi: 10.1145/2939672.2939785.

[5] Council of Europe, Common European Framework of Reference for Languages: Learning, Teaching, Assessment – Companion Volume. Strasbourg, France: Council of Europe Publishing, 2020.

[6] E. Kerz, D. Wiechmann, Y. Qiao, E. Tseng, and M. Ströbel, "Automated Classification of Written Proficiency Levels on the CEFR-Scale through Complexity Contours and RNNs," in Proc. 16th Workshop on Innovative Use of NLP for Building Educational Applications, ACL, 2021, pp. 199–209. doi: 10.18653/v1/2021.bea-1.21.

[7] N. S. Lagutina, K. V. Lagutina, A. M. Brederman, and N. N. Kasatkina, "Text Classification by CEFR Levels Using Machine Learning Methods and the BERT Language Model," Automatic Control and Computer Sciences, vol. 58, no. 7, pp. 869–878, 2024. doi: 10.3103/S0146411624700329.

[8] A. Gasparetto, M. Marcuzzo, A. Zangari, and A. Albarelli, "A Survey on Text Classification Algorithms: From Text to Predictions," Information, vol. 13, no. 2, p. 83, 2022. doi: 10.3390/info13020083.

[9] V. J. Schmalz and A. Brutti, "Automatic Assessment of English CEFR Levels Using BERT Embeddings," in Proc. 8th Italian Conf. Computational Linguistics (CLiC-it 2021), Accademia University Press, 2021, pp. 293–299. doi: 10.4000/books.aaccademia.10828.

[10] T. Gaillat, A. Simpkin, N. Ballier, B. Stearns, A. Sousa, M. Bouyé, and M. Zarrouk, "Predicting CEFR Levels in Learners of English: The Use of Microsystem Criterial Features in a Machine Learning Approach," ReCALL, vol. 34, no. 2, pp. 130–146, 2022. doi: 10.1017/S095834402100029X.

[11] Y. Arase, S. Uchida, and T. Kajiwara, "CEFR-Based Sentence Difficulty Annotation and Assessment," in Proc. 2022 Conf. Empirical Methods in Natural Language Processing, Abu Dhabi, 2022, pp. 6206–6219. doi: 10.18653/v1/2022.emnlp-main.416.

[12] D. Kavi, "Turkish Text Classification: From Lexicon Analysis to Bidirectional Transformer," arXiv:2104.11642, 2021. doi: 10.48550/arXiv.2104.11642.

[13] Y. Zhang, F. Bailles, and P. Prieto, "Singing Songs Facilitates L2 Pronunciation and Vocabulary Learning: A Study with Chinese Adolescent ESL Learners," Languages, vol. 8, no. 3, p. 219, 2023. doi: 10.3390/languages8030219.

[14] C. Hamilton, J. Schulz, H. Chalmers, and V. A. Murphy, "Investigating the Substantive Linguistic Effects of Using Songs for Teaching Second or Foreign Languages to Preschool, Primary and Secondary School Learners: A Systematic Review of Intervention Research," System, vol. 124, 103350, 2024. doi: 10.1016/j.system.2024.103350.

[15] E. Kerekes, "Using Songs and Lyrics in Out-of-Class Learning," in Language Learning Beyond the Classroom, D. Nunan and J. C. Richards, Eds. New York, NY, USA: Routledge, 2015, pp. 33–42.

[16] A. H. Kuppens, "Incidental Foreign Language Acquisition from Media Exposure," Learning, Media and Technology, vol. 35, no. 1, pp. 65–85, 2010. doi: 10.1080/17439880903561876.

[17] C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995. doi: 10.1007/BF00994018.

[18] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. doi: 10.1023/A:1010933404324.

[19] T. Cover and P. Hart, "Nearest Neighbor Pattern Classification," IEEE Trans. Information Theory, vol. 13, no. 1, pp. 21–27, 1967. doi: 10.1109/TIT.1967.1053964.

[20] R. Flesch, "A New Readability Yardstick," Journal of Applied Psychology, vol. 32, no. 3, pp. 221–233, 1948. doi: 10.1037/h0057532.

[21] X. Zhang and X. Lu, "Aligning Linguistic Complexity with the Difficulty of English Texts for L2 Learners Based on CEFR Levels," Studies in Second Language Acquisition, vol. 47, no. 5, pp. 1407–1434, 2025. doi: 10.1017/S0272263125101125.

[22] V. Franzoni, G. Biondi, and A. Milani, "Morpho-Phraseological Based Classification of CEFR Italian L2 Learner Writing Proficiency," IEEE Access, vol. 12, pp. 156433–156441, 2024. doi: 10.1109/ACCESS.2024.3485988.

[23] K. Allkivi, "Towards Interpretable Models for Language Proficiency Assessment: Predicting the CEFR Level of Estonian Learner Texts," arXiv:2602.13102, 2026.

[24] D. Kogan, M. Schumacher, S. Nguyen, M. Suzuki, M. Smith, C. S. Bellows, and J. Bernstein, "Ace-CEFR – A Dataset for Automated Evaluation of the Linguistic Difficulty of Conversational Texts for LLM Applications," arXiv:2506.14046, 2025.

[25] A. O. Durahim, A. Coşkun-Setirek, B. Başarır-Özel, and H. Kebapçı, "Music Emotion Classification for Turkish Songs Using Lyrics," Pamukkale University Journal of Engineering Sciences, vol. 24, no. 2, pp. 292–301, 2018. doi: 10.5505/pajes.2017.15493.

[26] H. Kesgin, M. K. Yuce, and M. F. Amasyali, "Developing and Evaluating Tiny to Medium-Sized Turkish BERT Models," arXiv:2307.14134, 2023.

[27] S. Vajjala and K. Lõo, "Automatic CEFR Level Prediction for Estonian Learner Text," in Proc. Third Workshop on NLP for Computer-Assisted Language Learning, Uppsala, Sweden, 2014.

[28] G. Tüm, "The Common European Framework of Reference and Turkish Language," The Journal of Academic Social Science Studies, vol. 6, no. 2, pp. 917–934, 2013.
[29] E. Borandağ, "LSRM: A New Method for Turkish Text Classification," Applied Sciences, vol. 14, no. 23, p. 11143, 2024. doi: 10.3390/app142311143.
[30] "TOLID: Turkish Offensive Language Identification Dataset and Transformer-Based Benchmarks," ELECTRICA, 2026.
[31] M. Çelik, "Turkish Song Lyrics," Kaggle, 2025. Available: https://www.kaggle.com/datasets/metncelik/turkish-song-lyrics.
[32] R. Speer, "rspeer/wordfreq: v3.1.1," Zenodo, 2022. doi: 10.5281/zenodo.7199437.
[33] Y. Savuran and S. Wulff, "Introducing TURLEC: A learner corpus for L2 Turkish," International Journal of Corpus Linguistics, vol. 31, no. 2, pp. 139–153, 2026 (forthcoming).
[34] Y. Savuran and S. Wulff, "Introducing TURVOP: A Turkish L2 vocabulary profile based on Turkish L2 learner corpus (TURLEC)," poster presented at the Annual Conference of the American Association for Applied Linguistics (AAAL), Denver, CO,

Election Prediction Using Social Media Data: An Artificial Intelligence and Game Theory-Based Approach

Ahmed Alturfi [1], Ali Hakan Büyüklü[2], Sabah Manfi Redha[3]

[1,2,3] Yildiz Technical University, İstanbul, Türkiye

*Corresponding author e-mail: ahmed.swadi8@googlemail.com

Abstract

This study develops a cautious framework for examining whether issue signals extracted from social-media video can help explain candidate-level electoral variation in the 2025 Iraqi parliamentary election in Baghdad. The preserved dataset contained an initial retrieval pool of 427 unique YouTube videos collected with 34 Arabic search queries. A metadata audit excluded 37 clearly out-of-scope records, leaving 390 videos for analysis. Five thematic variables were used: economy, direct benefits/services, power and stability, dignity, and identity. AI was used as an assisting measurement tool for identifying and structuring the core variables, while candidate-level scoring relied on human evaluation using a five-point Likert scale and a two-round Delphi procedure. Candidate scores were related to raw vote counts using OLS diagnostics and Huber robust regression, followed by risk-adjusted game-theoretic interpretation. Economy and direct benefits represented 29.5% and 28.2% of coded discourse, respectively. OLS produced high in-sample fit ($R^2 = 0.913$ and 0.9385), but significant non-normality and heteroscedasticity supported robust estimation. The framework links theory-guided social-media measurement, human candidate assessment, robust regression, and strategic analysis while explicitly treating the results as exploratory, non-causal, and not equivalent to out-of-sample forecasting accuracy.

Keywords: *Social Media Analysis; Evolutionary Game Theory; Multivariate Regression; Voter Behavior; Iraqi Elections*

INTRODUCTION

In contemporary democracies and hybrids, the formal electoral arena is the main locus of political competition. Prediction of elections remains a challenging task since the electorate is heterogeneous, the political knowledge among the people is not equally distributed, the conditions of campaigns might change over the course of an electoral campaign and the forecasting models are sensitive also to the data and institutional context [1]–[4]. Traditional forecasting methods, such as surveys and cross-sectional questionnaires, suffer from their own drawbacks as noncomparability of response rates over time and response bias can influence the reliability of forecasts [5], [6]. In this respect, AI can assist political analysis by converting raw audiovisual data into coded variables that may be statistically analysed. The appeal of social media is that it delivers frequent political communications, complaints, campaign assertions and signs of issue salience, though it shouldn't be considered a representative stand in for the full electorate. Iraq is especially relevant to this study as electoral competition is informed by

institutional uncertainty, coalitional politics, service-delivery imperatives, and the muhassasa power-sharing arrangement [7], [8]. The literature thus loses the capacity to marry a theory informed social-media coding, candidate-level issue positioning robust regression and strategic interaction within a single Baghdad-centered test-bed. This paper fills that void via a tentative AI-enhanced model which connects digital aural-visual narrative and electoral variables, candidate positioning through robust regression with risk-adjusted strategic interaction. The objective is not pure prediction, but testing whether structured signals extracted from Baghdad political videos can be informative in explaining candidate-level electoral variation under explicitly stated constraints.

LITERATURE REVIEW AND THEORETICAL FOUNDATIONS

Theoretical Foundations of AI-Based Electoral Forecasting

A proper election-prediction study requires more than a technical prediction pipeline; it must also explain why the selected variables should have political relevance. Accordingly, the framework advanced in this article rests on a multilayer theoretical foundation encompassing electoral behavior, political communication, identity mobilization, distributive politics, and strategic competition. This architecture is important because AI-based extraction can identify patterns in text, images, audio, and metadata, whereas political theory explains which patterns are electorally meaningful and how they should be interpreted. This position is consistent with the broader political-science view that explanation requires attention to institutional context, voter behavior, communication processes, and strategic agency [9].

Voter Choice, Rational Utility, and Retrospective Evaluation

The traditional rational choice theory of voting states that voters make decisions about parties and candidates on the basis of expected utility, policy advantages, and the likelihood of delivering desired results [10]. Sociological and psychological models further demonstrate that party attachment, social class, and retrospective voting behaviour influence how people vote [11]-[13], while electoral-control theory likewise links incumbent performance to voter sanction and reward [14]. In this study, these theories support the use of the economic and benefit variables. When voters see a coalition as more able to enhance livelihood security, ease economic difficulties, or provide services, the electoral reward for that coalition is expected to increase.

Economic Voting, Clientelism, and Direct Benefits

Economic voting theory accounts for why economic insecurity, unemployment, high levels of inflation, and failure in service delivery often feature as prominent electoral signals [12], [13]. In fragmented or weak institutional settings, however, economic voting may merge with clientelistic exchange, where voters react not only to abstract policy proposals but also to concrete, localized, or immediate advantages [15], [16]. This logic is also consistent with broader work on linkages between citizens and politicians [17]. This distinction makes X_ECO different from X_BEN: X_ECO deals with generalized livelihood and economic-security questions, while X_BEN represents specific material gains, distributive pledges, patronage cues, and identifiable service delivery.

Issue Ownership and Candidate Positioning

Issue-ownership theory states that parties and candidates are rewarded electorally when voters see them as more trustworthy or capable on specific policies [18]. This motivates candidate-level scoring. In online discourse, the extracted themes are not simply frequencies; they denote issue domains in which rival coalitions may seek to establish credibility. Repeated competence messaging on economic recovery, benefits, stability, dignity, or identity can lead to a perception that a candidate 'owns' an issue, which may strengthen the expected association between the variable score and the electoral outcome.

Agenda-Setting, Framing, and Digital Political Salience

The reliance on social media as electoral evidence is rooted in agenda-setting and framing theory. Agenda-setting theory posits that the salience of issues in media environments influences what citizens consider the most important political matters [19], [20]. Framing theory details how political communication constructs problem definitions, assigns causal responsibility, makes moral evaluations, and suggests courses of action [21], [22]. Related work on emotional appeals and political communication further supports the importance of message presentation and media context [23], [24]. Therefore, the pattern of themes in the 390 videos is treated as an indicator of digital issue salience, while multimodal AI is used to explore textual, rhetorical, and symbolic signs in video discourse.

Social Identity, Cleavage Politics, and the Iraqi Muhassa Context

Voting is determined not only by individual utility calculations but also by group membership, identity alignment, and collective representation. Social identity theory accounts for the political significance individuals attach to religious, ethnic, sectarian, partisan, or communal group memberships [25]. Cleavage theory similarly shows that stable social cleavages may underpin party competition and voter alignment [26]. In Iraq, identity-based politics is reinforced by the consociational and power-sharing logic of the muhassasa system [27]. The economic significance of identity has also been theorized as a distinct mechanism shaping preferences and behavior [28]. As a result, X_ID is considered a separate dimension even though identity discourse is less frequent than economic discourse in the retained sample.

Dignity, Recognition, and Grievance-Based Voting

The dignity dimension requires a theoretical foundation broader than economic voting alone. Recognition theory suggests that citizens can be politically mobilized not only on the basis of material resources but also through recognition, including respect, status, inclusion, and procedural justice [29], [30]. In contexts of corruption, uneven service provision, and low trust in political elites, dignity-related claims may contribute to grievance-based or protest voting. Therefore, X_DIG represents symbolic and moral aspects of political choice, including fairness, dignity, respectful treatment of the public, anti-corruption, and social justice.

Strategic Interaction, Spatial Competition, and Risk-Adjusted Campaigning

The game-theoretic component is based on the premise that candidates and coalitions do not act in isolation. Political actors are strategic: spatial competition and strategic campaign theories expect them to consider both voter preferences and opponent behavior when positioning themselves [31]. When one coalition promotes a high-profile issue offensively, its competitor can defend, counter-attack, replicate

the issue position, or redirect attention to another field. Nash equilibrium describes a strategic profile from which no player can improve its payoff by unilaterally deviating [32]. In the present research, the risk-adjusted payoff model converts voter sensitivity, message credibility, and issue severity into strategic outcomes while keeping the empirical interpretation within the constraints of the available data.

AI as a Measurement Instrument, Not a Substitute for Theory

An important theoretical protection in this work is the distinction between AI-based measurement and political interpretation. Multimodal AI can make it easier to detect and code patterns of discourse, but that alone cannot account for why people vote. Voter choice, agenda setting, framing, identity, recognition, and strategic interaction theories serve as the interpretive lens through which we view AI-coded content as electoral variables. This ensures that the analysis is not a technical end in itself and situates AI as a measurement tool within a theory-based social science design."

Table 1. Theoretical mapping of AI-extracted electoral variables and their expected role in the model

Theoretical construct	Main theoretical foundation	Operational variable	Expected role in the model
Material utility and livelihood security	Rational choice, retrospective voting, economic voting	X_ECO	Higher economic salience is expected to increase support for candidates perceived as credible on livelihood, employment, services, and economic protection.
Direct distributive benefit	Clientelism, patron-client exchange, distributive politics	X_BEN	Visible benefits, patronage signals, and localized service promises increase electoral payoff where immediate material exchange is politically salient.
Power, security, and institutional stability	Security voting, valence politics, institutional trust	X_POW	Candidates associated with order, authority, stability, or protection may gain advantage under uncertainty or political fragmentation.
Recognition, fairness, and grievance	Recognition theory, social justice, grievance voting	X_DIG	Dignity appeals may mobilize voters who interpret politics through injustice, corruption, exclusion, or lack of public respect.
Group belonging and symbolic representation	Social identity theory, cleavage politics, consociational power-sharing	X_ID	Identity appeals may be less frequent than economic discourse but can produce high mobilization under perceived group threat or symbolic representation.
Digital issue salience	Agenda-setting and framing theories	Frequency weights and thematic coding	Repeated appearance and framing of issues in videos are treated as proxy signals of digital political salience, subject to platform-bias controls.

Strategic interdependence	Game theory, spatial competition, Nash equilibrium	Risk-adjusted payoff matrices	Coalition payoffs depend on both own strategy and opponent strategy; offensive and defensive moves are evaluated under credibility and risk conditions.
AI-based discourse measurement	Computational social science and algorithmic fidelity	Multimodal AI/NLP coding	AI supports coding and measurement but does not replace theory-led explanation or human validation.

This mapping establishes a direct Theory -> Variable -> Measurement -> Interpretation chain. It clarifies that the five substantive variables were not generated as arbitrary categories, but represent theoretically defensible dimensions of electoral behavior and political communication.

Prior Empirical and Methodological Literature

Game theory, robust regression, social-media analysis, and AI have all contributed to recent advances in electoral forecasting. Work on language models and silicon sampling indicates that AI agents may emulate some patterns in human attitudes under rigorously defined circumstances, but such outputs still require validation with empirical political data and should not replace observed political data [33], [34], [35]. Social-media election research shows that useful signals can be extracted from online discourse, while also warning that platform composition, sentiment noise, hyperactive users, bots, and coordinated activity can impair inference [5], [36], [37], [38]-[42]. Robust regression is relevant because electoral data may contain outliers, skewed errors, and influential observations, conditions under which OLS coefficients can become unstable [43]-[45]. Game-theoretic election models add a strategic layer by treating candidates as players who seek to maximize utility while responding to rival conduct [46], [47], [32]. This paper connects these threads but interprets the resulting model as a bounded predictive tool rather than a replacement for representative polling or official electoral data.

Further support for this design comes from recent work published in 2025-2026. Gaur and Yadav [37] review election forecasting from social-media text; Lyman et al. [35] emphasize the need for algorithmic fidelity when language models are used in social-science research; Poggi et al. [48] demonstrate multimodal generative-AI analysis of political video; Höfler [44] argues for robust procedures when classical assumptions are violated; and Zhou [47] models strategic interaction across simultaneous elections. Related machine-learning and game-theoretic work also illustrates how computational modeling can be connected to public-opinion governance [49]. Recent studies of LLM-based public-opinion tracking and multimodal emotion analysis further reinforce the need for grounded validation and bounded interpretation [50], [51]. Collectively, this literature supports an integrated but cautious design rather than treating social-media signals as a direct substitute for representative electoral data.

Theoretical Gap and Research Propositions

Enhancements in AI-driven voter modeling, social media sentiment analysis, robust regression, and game-theoretic modeling can be seen within the reviewed documents. Nevertheless, three lacunae exist. For one, a good deal of social-media based predictive work considers online sentiment either as a primary or auxiliary predictor variable without adequately situating it in voter choice, agenda setting, or framing models. Two, political analyses emanating from AI tend to emphasise technological possibilities rather than the political processes through which outputs thus produced may be tied to

electoral conduct. Third, game-theoretic models of elections are often isolated from the digital conversations that produce empirically verifiable campaign signals. The research to date contributes to filling these voids by theoretically informing the AI-extracted variables, linking them to candidate-level electoral outcomes via robust regressions, and interpreting coalition action via a risk-adjusted strategic model.

P1. 'Economic' and 'benefits' terms will have the strongest predictive power in an environment where the electorate perceives political competition as about material usefulness and immediate provision of services.

P2. This will raise the correspondence between ideological scores on thematic variables and the vote under candidate-level ownership of salient policies.

P3. Frequency of digital issues should thus be understood as issue salience on the agenda, while framing of messages helps to understand how comparable issues can lead to different voter reactions.

P4. Although identity-related discourse may be less pervasive than economic discourse, it can generate more potent mobilization when attached to perceived group threat, symbolic defense, or consociational representation.

P5. Dignity-based appeals are predicted to be relevant when voters view politics in terms of inclusion, exclusion, corruption, or disrespect.

P6. Effects of voting campaigns are strategically dependent: the value of adopting an aggressive or defensive position depends on how the opposing coalition decides to act.

P7. Labels generated by AI need to be interpreted theoretically and they also need to be compared to external criteria, and they should be kept in mind as measurement-based outputs and not as stand-alone explanations of why people vote in particular ways.

MATERIALS AND METHODS

Research Design

This investigation employs a sequential multi-method design. First, a preserved pool of publicly available YouTube videos concerning the Iraqi parliamentary election and the Baghdad electoral context was reconstructed and screened. Second, AI was used as an assisting measurement instrument to identify and structure the core thematic variables, while thematic interpretation was subject to human review. Third, the five variables were applied at candidate level through human evaluation using a five-point Likert scale and a two-round Delphi procedure. Fourth, candidate-level scores were connected to electoral outcomes using regression and diagnostic testing. Fifth, the resulting estimates informed a risk-adjusted game-theoretic interpretation of coalition strategies. The purpose is to convert unstructured political communication into transparent, constrained, and statistically analyzable indicators rather than to claim that social media represents the electorate.

Data Collection

The preserved source workbook contained an initial retrieval pool of 427 unique YouTube video URLs collected through 34 Arabic-language search queries designed around the 2025 Iraqi parliamentary election, Baghdad, major coalitions, candidates, campaign communication, polling, monitoring, and results. The original search window was 1 October 2024 to 11 November 2025. A revision-stage metadata audit was conducted using the archived title, channel, upload date, view count, URL, and search-query fields. Thirty-seven records were excluded because the preserved metadata clearly indicated foreign political content, a different electoral contest, a local-only campaign or polling event outside Baghdad, or political content without an explicit connection to the 2025 parliamentary election. The resulting analytical sample contained 390 videos. The retained records were published between 28 January 2025 and 11 November 2025. Each retained video constituted one unit of content analysis.

Geographic targeting was implemented primarily through Baghdad-specific Arabic search terms rather than through a documented platform-level geolocation filter. Duplicate screening was performed using the preserved URL and a canonical YouTube video ID reconstructed from each URL; all 427 URLs and canonical IDs were unique. Videos with similar or identical titles were not automatically removed when they represented different video IDs, because they may correspond to distinct broadcasts or uploads. The full list of 34 search queries, all 427 screened records, the 37 exclusions with reasons, and the final 390 retained URLs are supplied in the accompanying screening-audit workbook. Archived view-count values are preserved in the source dataset, but the exact per-video timestamp at which each count was captured was not stored. File metadata show that the archived workbook was created on 9 July 2026; this date is reported only as provenance for the archived file and is not treated as the exact observation time for every view count. The sample therefore represents digitally visible electoral communication rather than a population-representative sample of voters.

Data Processing and Feature Engineering

The unstructured video corpus was analyzed through a theory-guided coding framework containing five substantive variables: X_ECO for economy and livelihood, X_BEN for direct benefits and service provision, X_POW for power, authority, security, and stability, X_DIG for dignity, fairness, and grievance, and X_ID for identity and symbolic representation. AI was used as an assisting measurement tool to help identify and structure these core dimensions; it was not used to generate the final candidate-level scores. Human review was used to check the substantive interpretation of the thematic framework. For the primary frequency analysis, one dominant thematic frame was retained for each video, after which the percentage share of each variable was calculated as follows:

$$w_i = (F_i / N) \times 100$$

F_i is the observed frequency of variable i and N is the number of retained observations (N = 390). The dominant-label design makes the five frequency shares mutually exclusive and ensures that their percentages sum to 100%. It was adopted for the primary descriptive table so that each video contributed once to the common denominator rather than being counted repeatedly across themes.

The original technical record does not preserve the exact AI model name and version, prompts, preprocessing pipeline, transcription tool, translation procedure, classification thresholds, or modality-specific intermediate outputs. These details cannot be reconstructed reliably from the aggregate results and are therefore not retroactively invented. For this reason, the revised manuscript describes the AI component as AI-assisted thematic measurement rather than as a reproducible benchmark of a specific AI model. This distinction also clarifies that the substantive variables and candidate assessments are theory- and human-guided components of the research design.

Some political videos addressed more than one issue domain. The primary analysis nevertheless used a single dominant label to maintain a mutually exclusive N = 390 distribution. This choice should not be interpreted as a claim that political communication is unidimensional. Secondary labels were not archived at the item level, so a formal retrospective multilabel sensitivity statistic cannot be calculated from the preserved metadata alone. The accompanying audit workbook retains the source URLs so that a future or revision-stage content re-coding can store both dominant and secondary labels and compare single-label and multilabel distributions directly.

Human validation formed part of the substantive review of the extracted thematic structure. However, the original item-level video-coding sheets were not preserved in a form that identifies the video coders, their independent labels, or the resolution of individual disagreements. Consequently, Cohen's kappa, Fleiss' kappa, or Krippendorff's alpha cannot be calculated retrospectively without re-coding the retained videos. This limitation is reported explicitly. The human validation of video themes should also be distinguished from the separate ten-member Delphi panel used for candidate-level scoring, described in Section 3.4.

Measurement and Index Construction

The five thematic variables were then applied to candidates through a structured human-evaluation procedure. For each candidate, ten human evaluators assessed the candidate on each of the five domains using a five-point Likert scale. The evaluation was conducted in two Delphi rounds. In Round 1, the evaluators made their assessments independently. The first-round group results were then presented back to the same evaluators, who reconsidered their ratings in Round 2 to determine whether the judgments converged. The resulting second-round consensus values formed the candidate-level inputs used in the regression stage. AI was not used to manufacture or replace these candidate ratings. The preserved manuscript-level outputs confirm the scoring framework, but the evaluator-by-candidate round-level worksheets were not archived; therefore, evaluator agreement statistics and the complete 10-evaluator scoring table cannot be reconstructed retrospectively from the aggregate regression coefficients alone. This limitation is important for independent replication and should be addressed by preserving the full Delphi worksheet in future applications. The risk index used for strategic adjustment similarly combines human-assessed credibility/evidence strength and perceived severity of voter impact:

$$R_i = C_i \times S_i, \quad 0 \le C_i \le 1, \ 0 \le S_i \le 1$$

C_i represents the normalized credibility or strength-of-evidence component and S_i represents normalized impact severity for issue i, consistent with the study's quantitative risk construction [53]. These inputs were also assessed through a two-round Delphi logic. Because the original evaluator-level round data and separate final C_i and S_i values were not preserved in the retained analytical record, the present manuscript does not invent agreement coefficients or reconstruct missing expert scores. Subsequent replications should archive both Delphi rounds and report the consensus rule and agreement statistic explicitly.

Multivariate Regression Model

The relationship between candidate-level issue scores and electoral outcomes was estimated using a multivariate linear regression model. For candidate j, the general specification is:

$$Y_j = \beta_0 + \beta_1 X_ECO_{,j} + \beta_2 X_BEN_{,j} + \beta_3 X_POW_{,j} + \beta_4 X_DIG_{,j} + \beta_5 X_ID_{,j} + \varepsilon_j$$

In the reported regressions, Y_j is candidate j's raw vote count in Baghdad, beta_0 is the intercept, beta_k are the coefficients for the five issue variables, and epsilon_j is the error term. The individual candidate is the statistical unit of analysis; separate coalition models use n = 138 Construction and Development candidates and n = 126 State of Law candidates. Because all observations are Baghdad candidates, no cross-governorate constituency adjustment was applied. Candidate visibility and media exposure were not retained as separate candidate-level covariates, so unequal exposure cannot be regarded as statistically controlled merely by using a robust estimator. Huber estimation reduces the influence of large residuals but does not by itself remove exposure-related confounding. Unequal visibility is therefore treated as a limitation. Because only coefficient estimates and summary diagnostics were retained, coefficient-level statistical significance is not asserted without standard errors, confidence intervals, test statistics, and p-values.

The notation used throughout the model are the following: X_ECO = economy/livelihood, X_BEN = direct benefits/services, X_ POW = power/security/stability, X_DIG = dignity/fairness/grievance, X_ID = identity/symbolic representation, w_i = percentage share of theme i, F_i = frequency of theme i, N = retained-video count, Y_j = candidate j's raw vote count, beta_0 = intercept, beta_k = slope coefficient, epsilon_j = regression error, R_i = risk index, C_i = credibility, S_i = severity, B_i = base payoff, V = two-coalition Baghdad vote total, Z = bounded risk-sensitivity adjustment, and D/H = defensive/offensive strategies.

Diagnostic Testing and Robust Regression

The appropriateness of the classical OLS assumptions was examined using variance inflation factors, the Durbin-Watson statistic, the Shapiro-Wilk test, and the Breusch-Pagan test. No severe multicollinearity was detected, whereas residual non-normality and heteroscedasticity were evident. Huber M-estimation was therefore used as a complementary robustness estimator because it down-weights observations with large residuals and reduces sensitivity to influential cases. It should not, however, be interpreted as a substitute for heteroscedasticity-robust coefficient-level standard errors. Such inferential outputs were not preserved in the aggregate results available for this revision. The

Durbin-Watson statistic is reported descriptively only because the candidate data are cross-sectional and its interpretation depends on the ordering of observations.

Game-Theoretic Framework

The strategic exchange between the two coalitions was modeled as a two-player game. Each coalition can use a defensive strategy (D) or an offensive strategy (H) within an issue domain. Payoffs are interpreted as illustrative risk-adjusted strategic values derived from the stated model, not as independent vote aggregates. The Nash equilibrium condition is:

$U_i(s_i^*, s_{-i}^*) \geq U_i(s_i, s_{-i}^*) \quad \forall\, s_i \in S_i, \forall\, i$

This means that, given the strategy of the rival, player i has no incentive to change its strategy. The game-theoretic matrices in the present study should be interpreted as a structured strategic representation rather than as a fully estimated equilibrium model where the underlying sensitivity terms are not empirically observed.

Integration of Models

The final methodological step integrates the regression results into the game-theoretic model. The estimated coefficients guide the payoff function that links empirically grounded voter behavior with the formal notion of strategic interaction. The competing regression models are therefore examined together with strategic decision dynamics so that the statistical and game-theoretic layers inform, rather than replace, one another.

RESULTS AND EMPIRICAL ANALYSIS

Baghdad contains competing pressures for participation and nonparticipation, and electoral competition is mediated by coalition politics, public-service delivery, and identity-based cooperation and competition. This study examines two major Shia-affiliated coalitions. The Construction and Development Coalition received 411,300 votes in Baghdad, followed by the State of Law Coalition with 228,300 votes. Of 4,359,490 eligible voters, 2,129,883 voted, representing a turnout of 48.86%. Because no official public dataset cited in the manuscript provides a sectarian breakdown of Baghdad voters, any estimate of a specifically Shia voting bloc is treated only as contextual assumption rather than direct empirical evidence.

Descriptive Analysis and Variable Distribution

Among the 390 included videos, five common electoral factors were identified. Economic and livelihood security (X_ECO) represented 29.5% of cases, while direct material benefits and services (X_BEN) represented 28.2%. Together, these two material categories comprised 57.7% of the coded discourse. Power and stability (X_POW) represented 16.7%, identity (X_ID) 14.1%, and dignity and social justice (X_DIG) 11.5%. These descriptive findings indicate that material and service-related issues dominated the observed digital discussion, while symbolic factors remained visible but less frequent.

Table 1. Distribution of extracted electoral variables in the retained social-media video sample

Variable	Frequency	Percentage (%)
X_ECO (Economy)	115	29.5
X_BEN (Benefits)	110	28.2
X_POW (Power)	65	16.7
X_DIG (Dignity)	45	11.5
X_ID (Identity)	55	14.1
Total	**390**	**100.0**

Regression Modeling and Diagnostic Evaluation

The coalition-level datasets contained 138 candidates for the Construction and Development Coalition and 126 candidates for the State of Law Coalition. Each candidate received a five-point score on the five issue variables. The relationship between candidate-level scores and electoral outcomes was initially estimated with OLS, and the coefficient estimates are reported below. Because coefficient-level standard errors, heteroscedasticity-robust standard errors, confidence intervals, test statistics, and p-values are not retained, signs and magnitudes are interpreted as model estimates rather than definitive evidence of statistical significance.

Construction and Development Coalition:

$$\hat{Y}_A = -1392.58 + 5505.53X_{ECO} + 2057.84X_{BEN} + 5066.80X_{POW} + 4679.49X_{DIG} + 1127.05X_{ID}$$

State of Law Coalition:

$$\hat{Y}_B = -236.03 + 6913.39X_{ECO} + 969.49X_{BEN} + 3688.87X_{POW} + 2406.66X_{DIG} + 589.45X_{ID}$$

For the Construction and Development Coalition, the OLS model accounts for 91.3% of the in-sample variance in observed votes ($R^2 = 0.913$). For the State of Law Coalition, the corresponding value is $R^2 = 0.9385$. These values indicate strong in-sample fit, but they should not be interpreted as evidence of equivalent out-of-sample performance. The diagnostic results indicate no severe multicollinearity (all VIF values were below 5). Shapiro-Wilk and Breusch-Pagan tests were significant ($p < 0.001$), indicating non-normal residuals and heteroscedasticity. The Durbin-Watson statistic is 1.10 for Coalition A and 2.20 for Coalition B. Because Durbin-Watson was developed for serial correlation in ordered data, its use for a cross-sectional candidate list is treated only as an approximate indicator unless candidate ordering has substantive meaning, such as ballot order or list rank. These diagnostics support reporting robust regression as the principal estimator.

Table 2. Results of Multiple Regression

Variable	Coalition A	Coalition B
Intercept	-1392.58	-236.03
X1 Economy	5505.53	6913.39
X2 Benefits	2057.84	969.49
X3 Power	5066.80	3688.87
X4 Dignity	4679.49	2406.66
X5 Identity	1127.05	589.45
R^2	0.913	0.9385

Table 3. Diagnostic Test Results

Test	Coalition A	Coalition B	Interpretation
VIF	<5	<5	No multicollinearity
Durbin-Watson	1.10	2.20	Descriptive only; interpretation depends on cross-sectional ordering
Shapiro-Wilk(p-value)	<0.001	<0.001	Non-normal residuals
Breusch-Pagan(p-value)	<0.001	<0.001	Heteroscedasticity

Robust Regression Results

Because the OLS diagnostics showed non-normality and heteroscedasticity, Huber regression was used to obtain estimates that are less sensitive to influential observations. The robust estimates are presented in Table 4. The overall positive direction of the OLS relationships is preserved, but coefficient magnitudes change substantially, indicating sensitivity to the configuration of influential observations. For this reason, the robust models are treated as the primary estimation layer. The estimated robust equations are:

Construction and Development Coalition:

$\hat{Y}_A(Huber) = -204.05 + 1539.75X_ECO + 1558.49X_BEN + 2151.26X_POW + 1573.43X_DIG + 886.80X_ID$

State of Law Coalition:

$\hat{Y}_B(Huber) = 100.28 + 10660.58X_ECO + 895.36X_BEN + 1081.53X_POW + 1159.81X_DIG + 540.53X_ID$

Table 4. Results of Robust Regression Estimates Using Huber's M-Estimator

Variable	Coalition A	Coalition B
Intercept	-204.05	100.28
X_ECO	1539.75	10660.58
X_BEN	1558.49	895.36
X_POW	2151.26	1081.53
X_DIG	1573.43	1159.81
X_ID	886.80	540.53

The original OLS output displays sample size, R^2 and model diagnostics, but not standard Errors at the coefficient Level heteroscedasticity-robust Standard errors Confidence intervals t statistics or p-values. The retained Huber output contains coefficients and sample size but not robust standard errors, confidence intervals, test statistics, p-values or a measure of goodness-of-fit which could be retained separately. These quantities cannot be recovered from the aggregate coefficient tables alone, so any unreported inferential statistic is not imputed.

The change from OLS to Huber estimates is not uniform across the two coalitions and should therefore be interpreted cautiously. For the Construction and Development Coalition, the robust X_ECO coefficient drops from 5505.53 to 1539.75, consistent with down-weighting influential observations. For the State of Law Coalition, however, the robust X_ECO coefficient increases from 6913.39 to 10660.58. Because Huber M-estimation down-weights large residuals, this divergence suggests that influential observations affect the two coalition datasets differently. In the absence of leverage and residual-level diagnostics, the State of Law X_ECO estimate should therefore be interpreted with particular caution.

Game-Theoretic Analysis and Strategic Interaction

The game-theoretic component treats electoral competition as strategic interaction across the five issue dimensions. To anchor the payoff matrices in observed data rather than unreported inputs, the base symmetric (D-D) payoff for issue i is defined as $B_i = (F_i/N) \times V$, where F_i is the observed frequency of issue i in Table 1, $N = 390$ is the number of retained videos, and $V = 639{,}600$ is the combined Baghdad vote total for the two coalitions (411,300 + 228,300; Section 4). This allocates the combined vote total across issues in proportion to their observed digital salience, so the five base payoffs sum exactly to V and remain traceable to Table 1 and the reported election totals. Applying this formula gives base payoffs of 188,600 for X_ECO, 180,400 for X_BEN, 106,600 for X_POW, 73,800 for X_DIG, and 90,200 for X_ID (Tables 5–9). The Z term remains an illustrative risk-adjustment sensitivity parameter because the credibility (C_i) and severity (S_i) components of $R_i = C_i \times S_i$ are not separately reported; Z is therefore a bounded sensitivity assumption rather than an estimated quantity.

The payoff matrices indicate that offensive action can increase one coalition's payoff when the other remains defensive, whereas simultaneous offensive strategies reduce payoffs for both sides. Because the base payoffs are tied to observed issue salience and actual vote totals, their descriptive foundation is traceable to Section 4.1; however, the Z adjustment remains an explicitly illustrative sensitivity layer.

Table 5. Strategic Interaction – Economic Variable (X_1)

Strategy	Player B: Defensive (D)	Player B: Offensive (H)
Player A: Defensive (D)	(188600 , 188600)	(188600 − Z, 188600 + Z)
Player A: Offensive (H)	(188600 + Z, 188600 − Z)	(188600 − Z, 188600 − Z)

Table 6. Strategic Interaction – Benefits Variable (X_2)

Strategy	Player B: Defensive (D)	Player B: Offensive (H)
Player A: Defensive (D)	(180400 , 180400)	(180400 − Z, 180400 + Z)
Player A: Offensive (H)	(180400 + Z, 180400 − Z)	(180400 − Z, 180400 − Z)

Table 7. Strategic Interaction – Power Variable (X_3)

Strategy	Player B: Defensive (D)	Player B: Offensive (H)
Player A: Defensive (D)	(106600 , 106600)	(106600 − Z, 106600 + Z)
Player A: Offensive (H)	(106600 + Z, 106600 − Z)	(106600 − Z, 106600 − Z)

Table 8. Strategic Interaction – Dignity Variable (X_4)

Strategy	Player B: Defensive (D)	Player B: Offensive (H)
Player A: Defensive (D)	(73800 , 73800)	(73800 − Z, 73800 + Z)
Player A: Offensive (H)	(73800 + Z, 73800 − Z)	(73800 − Z, 73800 − Z)

Table 9. Strategic Interaction – Identity Variable (X_5)

Strategy	Player B: Defensive (D)	Player B: Offensive (H)
Player A: Defensive (D)	(90200 , 90200)	(90200 − Z, 90200 + Z)
Player A: Offensive (H)	(90200 + Z, 90200 − Z)	(90200 − Z, 90200 − Z)

Model Validation and Contribution

Table 10 reports the internal convergence of model-based estimates and realized outcomes. Across the nine computed variable-coalition cells, values range from 89% to 99%; X_ECO for the State of Law Coalition was not calculated and is reported as N/A rather than estimated. These values are internal convergence indicators and should not be interpreted as full forecasting accuracy because the manuscript does not include a holdout sample, cross-validation, MAE, RMSE, MAPE, or residual-level validation.

The results therefore indicate internal consistency among the AI-coded variables, regression estimates, and strategic interpretation, but they do not demonstrate performance in future elections or justify replacing representative polling. Causal interpretation is also constrained by potential endogeneity, omitted variables, and the nature of social-media-based evidence.

Table 10. Internal convergence between model estimates and observed outcomes

Variable	Coalition A (%)	Coalition B (%)
X_ECO	96	N/A
X_BEN	98	93
X_POW	98	99
X_DIG	99	98
X_ID	96	89

External Benchmarking and Non-Fabrication Protocol

To strengthen the transparency of the empirical evidence, the revision distinguishes the reconstructed source-data audit from external methodological benchmarking. The primary empirical evidence remains the final 390-video Baghdad-focused analytical sample, candidate-level scoring, and the reported electoral tables. The preserved source workbook additionally documents the 427-record raw retrieval pool, 34 Arabic search queries, YouTube URLs, titles, channels, upload dates, view counts, and screening decisions. External literature is used only for calibration, methodological caution, and interpretation; it is not merged into the Iraqi dataset and is not used to fabricate candidate observations, coder labels, residuals, or evaluator scores that were not preserved.

That distinction is crucial for a strong research design. The nascent literature on social-media-based election prediction shows that online political data are informative but also are affected by platform bias, demographic skew, hyperactive users, coordinated activity, bots and inflated accuracy claims [38]–[42]. Consequently, the present study leverages external criteria to limit interpretation, identify threats to validity, and not overstate the predictive power of social-media data.

A schematic of the full analytical pipeline is presented in Fig. 1, and the reconstructed data-collection and screening process (427 retrieved records -> 37 exclusions -> 390 retained videos) is summarized in Fig. 2 and the accompanying screening-audit workbook.

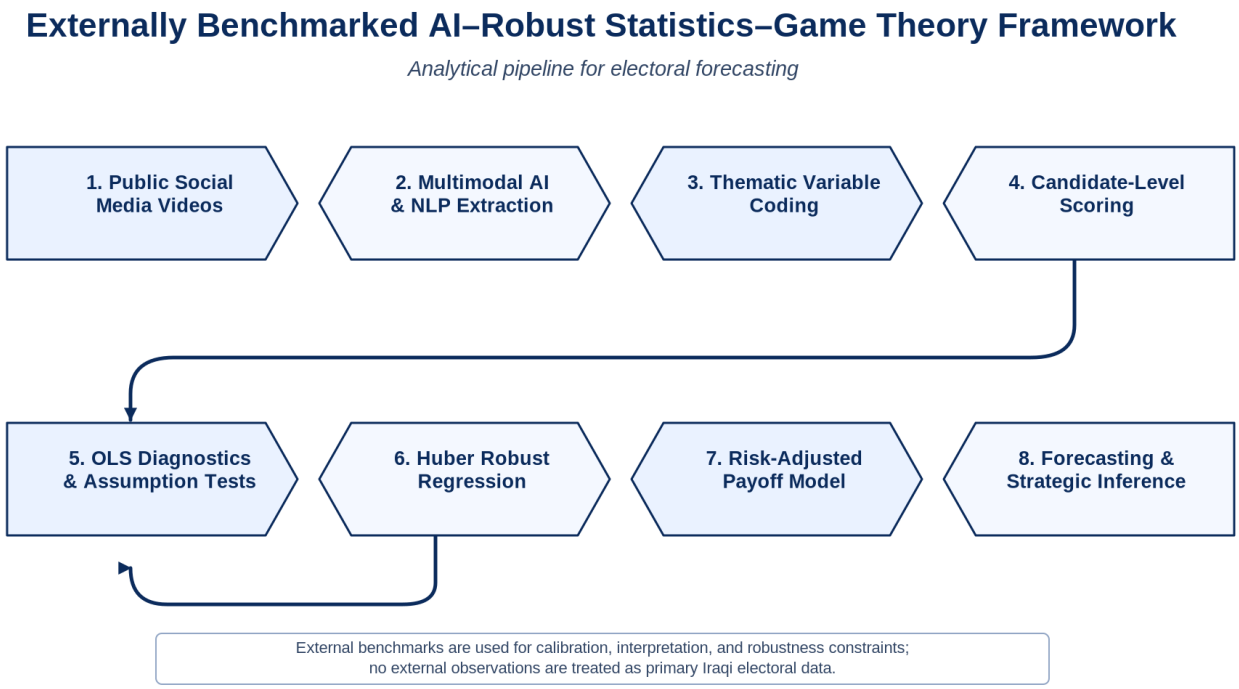


Figure 1. Proposed AI--robust regression--game theory electoral forecasting framework. The figure clarifies the full analytical chain and explicitly separates primary empirical evidence from external benchmarking.

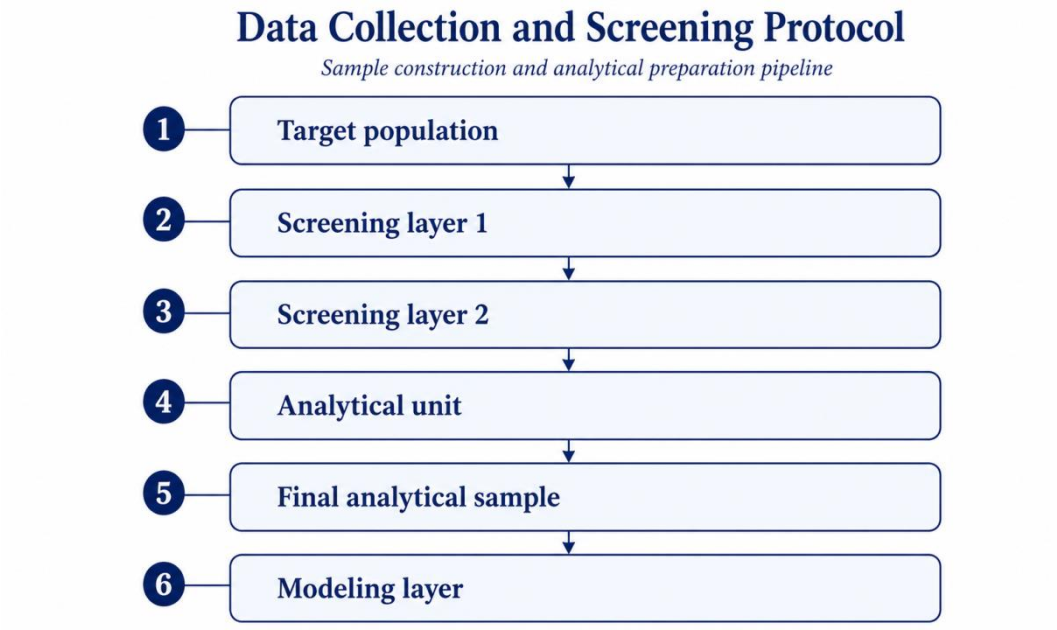


Figure 2. Data collection and screening protocol. The preserved source workbook contained 427 unique YouTube records; metadata screening excluded 37 clearly out-of-scope records, leaving the final analytical sample of 390 videos.

Advanced Visual Analytics and Robustness Evidence

The tabular results were extended into a visual analytical layer. The figures in this section serve three purposes: to make the variable structure transparent, to justify the transition from OLS to robust regression, and to show the strategic implications of the game-theoretic model. Except where noted, all figures in this section are based on tables from the manuscript and do not draw on unobserved raw data.

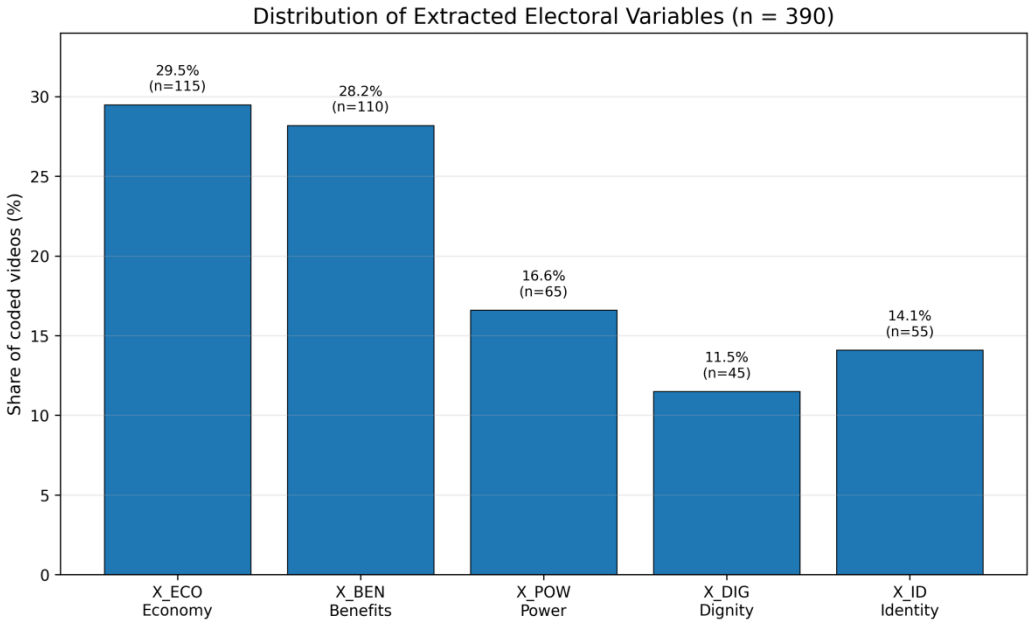


Figure 3. Distribution of extracted electoral variables. Economy and direct benefits jointly represent more than 57% of the coded political discourse, indicating that material and livelihood concerns dominate the observed Baghdad social media environment.

Figure 3 shows that the sampled electoral discourse is far from uniform. Economic insecurity and material benefits account for the largest shares of coded videos, while dignity and identity are less frequent but remain politically relevant because they may elicit strategic responses when connected to credibility, reputation, or group mobilization.

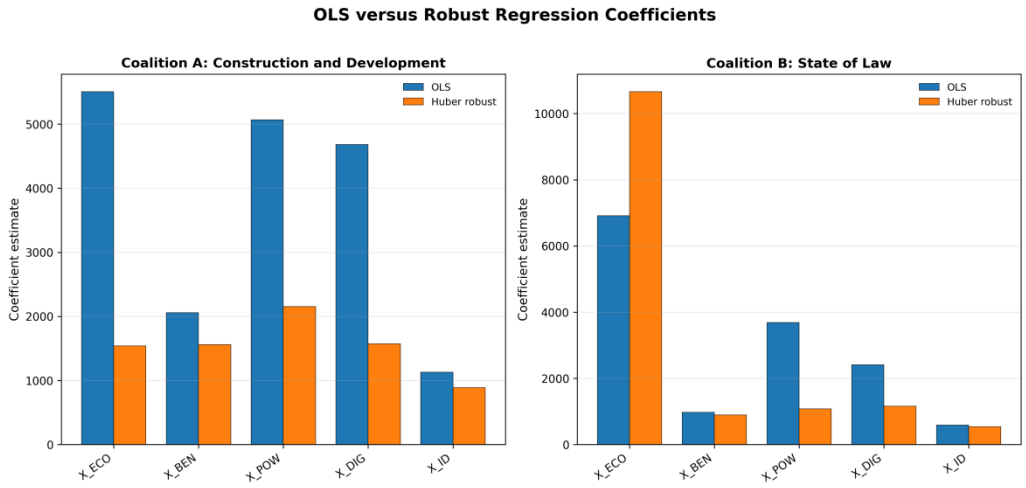


Figure 4. Comparison between OLS and Huber robust regression coefficients. The figure visualizes how robust estimation modifies the magnitude of predictor effects under non-normal and heteroscedastic conditions.

Figure 4 reinforces the methodological decision to use robust estimates as the principal layer of inference. The difference between OLS and Huber coefficient magnitudes is substantive rather than cosmetic: it shows the sensitivity that can arise when electoral observations are skewed, heavy-tailed, or influenced by candidates with exceptionally high visibility or vote concentration.

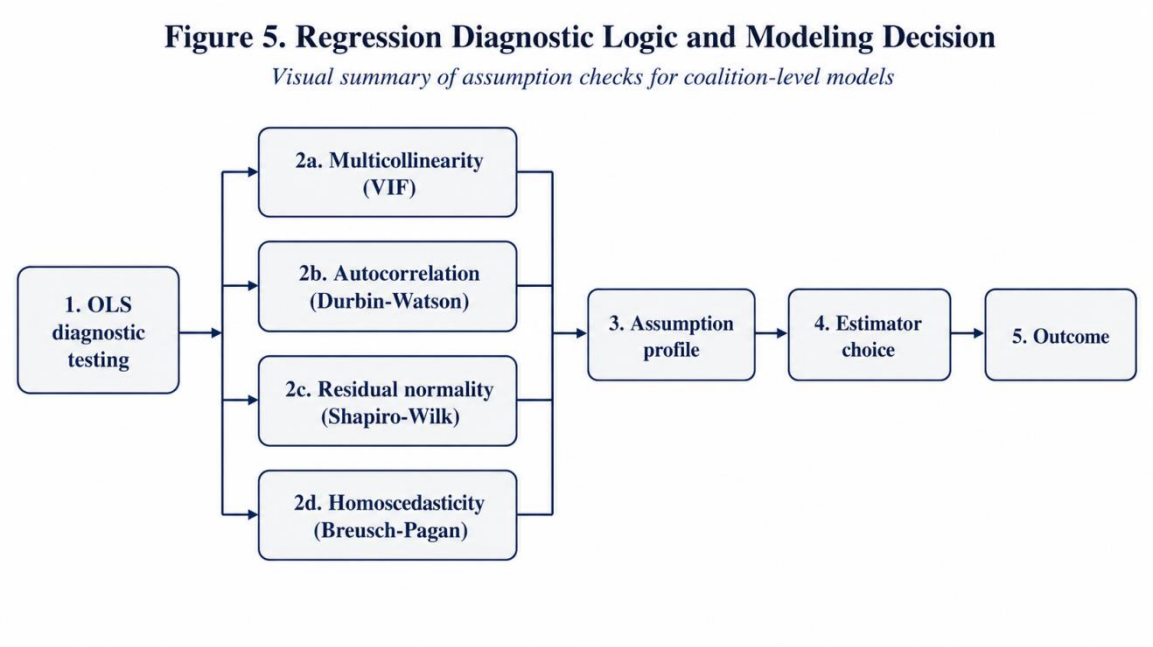


Figure 5. Regression diagnostic evidence dashboard. The diagnostic results show no severe multicollinearity signal, but they indicate non-normality, heteroscedasticity, and autocorrelation concern in Coalition A, justifying robust estimation.

The main diagnostic evidence is summarized in Figure 5. Multicollinearity is not a severe threat based on the VIF values; however, the Shapiro-Wilk and Breusch-Pagan tests do not support residual normality or constant variance. Robust regression is therefore not a cosmetic addition but a response to the empirical structure of the data.

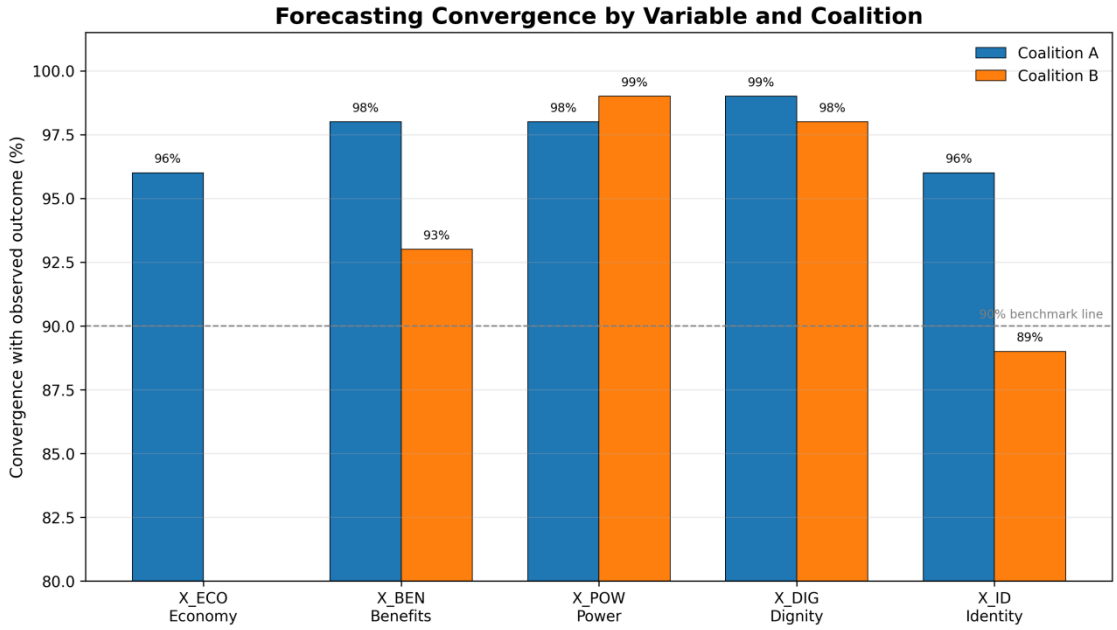


Figure 6. Internal convergence by variable and coalition. The values are high across most domains, while identity-related convergence is lower, consistent with the higher uncertainty of symbolic and perception-based variables.

Figure 6 summarizes the internal convergence values reported in Table 10 and shows that the lowest computed convergence occurs for the identity dimension of Coalition B.

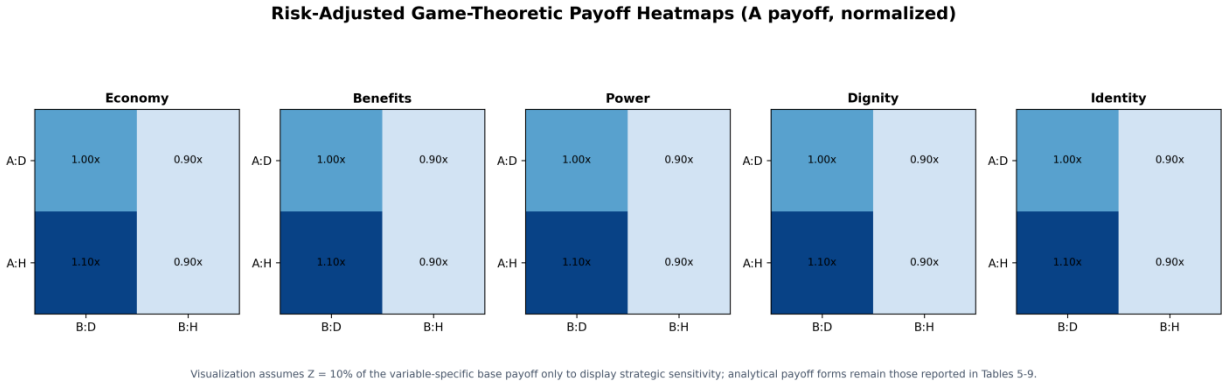


Figure 7. Risk-adjusted strategic payoff heatmaps. The heatmaps display the normalized payoff consequences of defensive and offensive strategies under an illustrative Z-risk sensitivity assumption.

Figure 7 provides a strategic visualization of the payoff matrices. Offensive positioning increases the attacking coalition's payoff when the opponent remains defensive, whereas double offense lowers payoffs for both sides. Together, the measurable voter concerns and strategic campaign behavior indicate that electoral competition in this framework emerges from the interaction between observed issue salience and strategic response.

Table 11. External Benchmarking Matrix for Social-Media Election Forecasting

Source	Benchmark Type	Role in Current Study
Gayo-Avello (2012)	Meta-analysis / critique	Twitter alone cannot replace polling; avoid hyper-accuracy claims.
Lampos (2012)	UK 2010 vote-intention inference	Reported MAE around 4.34% ± 2.13 after enriched sentiment inference.
Yang et al. (2020/2022)	U.S. midterm Twitter sampling bias	Hyperactive accounts and suspicious behavior can distort volume-based samples.
Scarano et al. (2024)	Election polls on Twitter/X	Social-media polls may deviate from official results and mainstream polling.
Balasubramanian et al. (2024/2025)	22M X posts dataset	Large-scale discourse data are useful, but platform-specific and not automatically representative.
Current Iraq Study	390 Baghdad political videos	Primary sample; external benchmarks constrain interpretation rather than replace Iraqi data.

Table 11 provides an external benchmarking layer for interpreting the present results. Previous studies caution that election prediction based on Twitter or X can produce overly optimistic accuracy claims when methodological limitations are ignored [38]. Other work shows that enriched sentiment analysis may reduce prediction error but is insufficient as a stand-alone forecasting method [39]. Hyperactive or otherwise atypical accounts can distort voter-characterization samples [40], while social-media polls may differ substantially from official results and established polling aggregates [42]. Large political-discourse datasets are valuable for high-volume analysis, but platform-based discourse should not be assumed to represent the broader electorate without adjustment [41]. Accordingly, these external studies are used as methodological benchmarks and cautions rather than as alternative observations for the Iraqi electoral dataset.

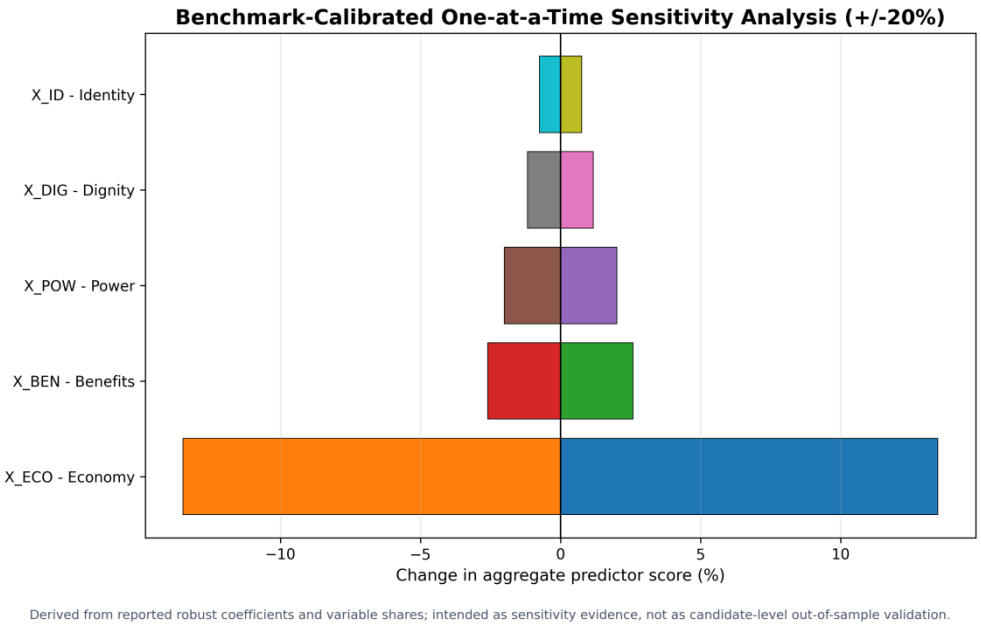


Figure 8. Benchmark-calibrated sensitivity analysis. The tornado plot uses reported robust coefficients and variable shares to show how one-at-a-time +/-20% perturbations affect the aggregate predictor score.

The sensitivity layer in Figure 8 is a benchmark-calibrated validation check rather than a simulation. Because full residual-level observations are not available in the paper, the figure uses the reported robust coefficients and observed variable shares to test whether the aggregate predictor structure is dominated by a single variable. The output also identifies the electoral issues that should receive priority in future candidate-level data collection and out-of-sample verification.

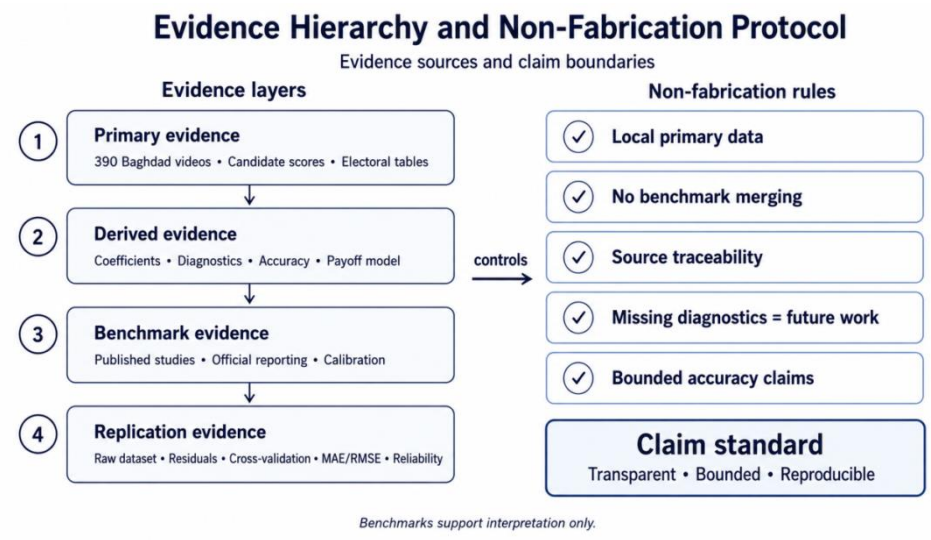


Figure 9. Evidence hierarchy and non-fabrication control. The figure distinguishes primary empirical evidence, derived analytical evidence, external benchmark evidence, and future replication evidence.

Figure 9 summarizes the evidence hierarchy used in this study and distinguishes primary evidence from derived analytical evidence, external benchmarks, and data required for future replication.

Official Contextual Validation of the 2025 Iraqi Election Setting

The electoral circumstances were also consistent with post-election reports from nonpartisan news organizations. Reuters reported that Prime Minister Mohammed Shia al-Sudani's coalition came first in Iraq's 2025 parliamentary election and that the final nationwide turnout reached 56.11% [54]. Subsequent reporting confirmed that the coalition won 46 seats, while the State of Law Coalition secured 29 seats [55]. The Associated Press reported that, after electoral complaints were resolved, Iraq's Federal Supreme Court ratified the election results [56]. These references are used only to establish the national electoral context. Baghdad-specific values remain the primary empirical unit of this study. National turnout and seat totals can support the plausibility of the broader political setting, but they cannot substitute for the video-coded, governorate-level variables or candidate-level scoring used in the model. The revised manuscript therefore treats official and journalistic election coverage as contextual validation rather than model-training data.

DISCUSSION

The interpretation of these results is strengthened when the empirical findings are read through the theoretical framework in Section 2.1 and the reconstructed sampling audit in Section 3.2. The prominence of economic and direct-benefit factors is consistent with rational-choice, retrospective-voting, and clientelistic-exchange models, while the continuing salience of identity and dignity is consistent with social-identity, cleavage, recognition, and grievance-based approaches. The findings should not be interpreted as a mechanical product of AI coding. AI served as an assisting measurement instrument for identifying and structuring the thematic variables, whereas candidate-level assessment relied on human evaluation and a two-round Delphi process. The game-theoretic layer should likewise be read as a structured strategic interpretation of issue salience and coalition response rather than as an independently observed causal mechanism.

The results also serve to position the current study in the context of previous social-media election investigation. The priority for the economy and direct benefits are also in line with studies that indicate that electorally relevant issue signals can emerge in online political discussions, but the present design is different as it treats issue salience rather than sentiment or platform volume as the primary content measure [5], [37], [38]-[42]. The diagnostics also confirm the caution highlighted in robust-regression literature: considerable movement from OLS to Huber estimates indicates that influential observations have a significant impact on the magnitude of coefficients under non-normality and heteroscedasticity[43], [44]. In contrast to studies that present predictive error on held-out or externally assessed forecasts, the 89%–99% figures that we report are solely internal convergence diagnostics. This more circumscribed reading is in line with earlier cautionary notes around platform bias, hyperactive users, and exaggerated social-media forecasting claims [38], [40]-[42].

The internal convergence values in Table 10, across the nine computed variable-coalition cells, range from 89% to 99% and indicate strong within-sample coherence. They should not be interpreted as independent out-of-sample forecasting accuracy because no holdout sample or cross-validation was available. The revision-stage source audit materially improves sampling transparency by documenting

the 427-record YouTube retrieval pool, 34 Arabic search queries, 37 metadata-based exclusions, and final N = 390 sample. Important reproducibility constraints nevertheless remain: the original item-level AI settings, video-coder labels, evaluator-level Delphi worksheets, candidate-exposure covariates, and coefficient-level inferential outputs were not archived. The game-theoretic base payoffs in Section 4.4 are tied to observed issue salience and vote totals; however, the separate Delphi-derived credibility and severity inputs were not preserved, so the Z-sensitivity layer remains bounded and illustrative. The contribution of the paper is therefore the transparent integration of digital discourse, human candidate assessment, regression, and strategic interpretation, not the precision of any single estimate or a claim that social-media AI can replace representative electoral polling.

CONCLUSIONS

This study developed a theory-driven framework linking social-media video analysis, candidate-level assessment, robust regression, and game-theoretic interpretation for the 2025 Iraqi parliamentary election in Baghdad. A revision-stage audit of the preserved source workbook established that the original retrieval pool contained 427 unique YouTube URLs obtained through 34 Arabic-language search queries. Thirty-seven records were excluded because their preserved metadata clearly indicated foreign political content, a different electoral contest, local-only material outside Baghdad, or non-electoral political content, leaving the analytical sample of 390 videos. This reconstruction strengthens the transparency of the sampling process and replaces the earlier, insufficient description of the data as simply the “top 390 most-viewed” videos. Within the retained sample, economic issues accounted for 29.5% of coded discourse and direct benefits or services for 28.2%, followed by power and stability (16.7%), identity (14.1%), and dignity (11.5%). Candidate-level modeling produced high in-sample OLS fit ($R^2 = 0.913$ and 0.9385), while significant non-normality and heteroscedasticity supported the use of Huber robust regression as a complementary estimator.

The principal contribution is methodological rather than a claim of precise election forecasting. AI served as an assisting measurement instrument for identifying and structuring the core thematic variables; it did not replace human political judgment or generate the final candidate assessments. Candidate-level scores were obtained through human evaluation on a five-point Likert scale and a two-round Delphi process, and the same consensus logic informed the risk-related assessment used in the strategic layer. Nevertheless, several limitations remain important. The original AI model versions, prompts, transcription and translation settings, item-level human video-coding records, and evaluator-level Delphi round data were not archived in a form that permits retrospective calculation of inter-coder reliability or complete independent reconstruction of every candidate score. Likewise, coefficient-level standard errors, confidence intervals, and test statistics cannot be recovered from aggregate regression coefficients alone. The reported 89%-99% values should therefore be interpreted only as internal convergence indicators, not out-of-sample predictive accuracy. Future replication should preserve complete coding logs, multilabel annotations, evaluator-level Delphi data, inferential regression outputs, and an independent validation sample.

Future Research Directions and Strategic Implications

Future research should build on the revision-stage sampling audit rather than repeat the earlier data-collection ambiguity. The platform, 34 Arabic search queries, 427-record retrieval pool, 37 exclusions,

and final 390-video sample are now documented in the accompanying audit workbook. The next priorities are content-level reproducibility and statistical replication: preserve the exact AI model and version, prompts, transcription and translation settings, dominant and secondary video labels, independent human-coder decisions, and inter-coder reliability statistics. Candidate-level work should preserve both Delphi rounds for all ten evaluators, the scoring rubric and consensus rule, candidate-exposure measures, and the final candidate-by-variable table. Regression software should export complete coefficient-level inference, including conventional and heteroscedasticity-robust standard errors, confidence intervals, test statistics, p-values, sample size, and appropriate goodness-of-fit measures. Predictive performance should then be evaluated on an independent sample or through a prespecified validation procedure using metrics such as MAE, RMSE, or MAPE. AI-assisted social-media analysis can reveal visible issue salience, but credible high-stakes electoral inference still requires auditable data, human validation, representative evidence, and restrained interpretation.

Appendix A. Figure Source and Evidence Classification

The table below documents the evidence source used for each newly added figure. This appendix is included to make the visual layer transparent and to prevent any ambiguity regarding whether a figure is based on primary data, derived manuscript data, or external benchmarks.

Figure	Evidence Source	Type of Evidence	Use in Study
Figure 1	Methodological design of the manuscript	Derived conceptual evidence	Explains the full AI-statistical-game-theoretic pipeline
Figure 2	Revision-stage source workbook audit: 427 raw YouTube records, 37 exclusions, final n = 390	Primary/derived audit evidence	Documents the reconstructed sampling and screening pathway using preserved metadata
Figure 3	Table 1: frequency and percentage distribution	Primary manuscript table	Shows variable dominance in coded discourse
Figure 4	Tables 2 and 4: OLS and robust coefficients	Derived analytical evidence	Compares estimator behavior
Figure 5	Table 3: diagnostic tests	Derived analytical evidence	Justifies robust regression
Figure 6	Table 10: internal convergence indicators	Derived analytical evidence	Shows within-sample convergence across variables without treating it as out-of-sample accuracy
Figure 7	Tables 5-9: payoff matrix structure	Derived/sensitivity visualization	Shows strategic payoff logic under explicit Z assumption
Table 11	Published social-media forecasting studies	External benchmark evidence	Constrains interpretation and prevents overclaiming
Figure 8	Reported robust coefficients + variable shares	Benchmark-calibrated sensitivity evidence	Shows one-at-a-time +/- 20% predictor sensitivity
Figure 9	Research design and evidence hierarchy	Methodological transparency evidence	Separates primary, derived, external, and future evidence

Appendix B. Minimum Dataset Schema for Future Replication

The following schema is recommended for future replication and for producing residual-level diagnostics, correlation heatmaps, cross-validation metrics, MAE, RMSE, MAPE, and inter-coder

reliability statistics. These analyses should be performed only when the corresponding raw observations are available.

Dataset Layer	Required Fields	Purpose
Video-level data	Video ID, platform, URL, title, channel, upload date, archived view count, search query, screening status, transcript, metadata/comments	Reproducible source audit, AI/NLP re-coding, platform-bias checks, and future content-level validation
Theme coding	AI label, human label, confidence, dominant theme, secondary theme	Reliability testing and label-error correction
Candidate-level data	Candidate ID, coalition, variable scores, campaign signals, actual votes	Regression, residual diagnostics, and out-of-sample validation
Expert validation	Expert ID, Delphi round, credibility score, severity score	Risk index construction and inter-rater agreement
Model outputs	Predicted votes, residuals, robust weights, leverage statistics	MAE, RMSE, MAPE, Q-Q plots, and influence diagnostics
Benchmark metadata	Reference study, context, method, metric, limitation	External comparison without merging external data into Iraqi observations

Appendix C. Revision-Stage Source Dataset Audit

The accompanying file "Election_Video_Dataset_Audited_390.xlsx" documents the reconstructed sampling frame used in the revision. It contains: (i) the 427-record raw screening audit; (ii) the 390 retained YouTube records; (iii) the 37 excluded records with record-specific exclusion reasons; (iv) the 34 preserved Arabic search queries and their raw/retained counts; and (v) the operational screening rules. The file also reconstructs canonical YouTube IDs from URLs where the original spreadsheet had interpreted IDs beginning with a hyphen as Excel formula errors. This supplementary audit supports the revised description in Section 3.2 but does not fabricate missing AI prompts, coder-level video labels, Delphi round-level ratings, or regression inference that was not archived.

References

[1] A. Lenz, *X. Fernández-i-Marín, M.* Hinterleitner, C. Knill, and Y. Steinebach, "Democratic governance and policy complexity: revisiting the intelligence of democracy," *Journal of European Public Policy*, pp. 1–28, 2025, doi: 10.1080/13501763.2025.2553756.

[2] P. Rattanasevee, Y. Akarapattananukul, and Y. Chirawut, "Direct democracy in the digital age: opportunities, challenges, and new approaches," *Humanit. Soc. Sci. Commun.*, no. 2024, pp. 1–9, doi: 10.1057/s41599-024-04245-1.

[3] H. Ann and T. S. James, "Electoral backsliding ? Democratic divergence and trajectories in the quality of elections worldwide," *Elect. Stud.*, vol. 86, no. September, p. 102696, 2023, doi: 10.1016/j.electstud.2023.102696.

[4] M. Stegmaier, S. Jokinsky, and M. S. Lewis-Beck, "The evolution of election forecasting models in the UK," *Elect. Stud.*, vol. 86, p. 102694, 2023.

[5] K. D. S. Brito, R. L. C. Silva Filho, and P. J. L. Adeodato, "A systematic review of predicting elections based on social media data: research challenges and future directions," *IEEE Trans. Comput. Soc. Syst.*, vol. 8, no. 4, pp. 819–843, 2021.

[6] F. Bachmann, C. Sarasua, and A. Bernstein, "Fast and adaptive questionnaires for voting advice

applications," in *Joint European Conference on machine learning and knowledge discovery in databases*, Springer, 2024, pp. 365–380.

[7] K. Palani, "How Power-Sharing Endures: Generational Change and Institutional Persistence in Iraq," *Nationalism and Ethnic Politics*, pp. 1–25, 2025, doi: 10.1080/13537113.2025.2558053.

[8] J. McEvoy and E. W. Aboultaif, "Power-sharing challenges: From weak adoptability to dysfunction in Iraq," *Ethnopolitics*, vol. 21, no. 3, pp. 238–257, 2022.

[9] R. E. Goodin and H.-D. Klingemann, Eds., A New Handbook of Political Science. Oxford: Oxford University Press, 1996.

[10] A. Downs, An Economic Theory of Democracy. New York: Harper and Row, 1957.

[11] A. Campbell, P. E. Converse, W. E. Miller, and D. E. Stokes, The American Voter. New York: Wiley, 1960.

[12] V. O. Key, The Responsible Electorate: Rationality in Presidential Voting, 1936-1960. Cambridge, MA: Harvard University Press, 1966.

[13] M. P. Fiorina, Retrospective Voting in American National Elections. New Haven: Yale University Press, 1981.

[14] J. A. Ferejohn, "Incumbent performance and electoral control," Public Choice, vol. 50, no. 1-3, pp. 5-25, 1986.

[15] H. Kitschelt and S. I. Wilkinson, Eds., Patrons, Clients, and Policies: Patterns of Democratic Accountability and Political Competition. Cambridge: Cambridge University Press, 2007.

[16] A. Hicken, "Clientelism," Annual Review of Political Science, vol. 14, pp. 289-310, 2011.

[17] H. Kitschelt, "Linkages between citizens and politicians in democratic polities," Comparative Political Studies, vol. 33, no. 6-7, pp. 845-879, 2000.

[18] J. R. Petrocik, "Issue ownership in presidential elections, with a 1980 case study," American Journal of Political Science, vol. 40, no. 3, pp. 825-850, 1996.

[19] M. E. McCombs and D. L. Shaw, "The agenda-setting function of mass media," Public Opinion Quarterly, vol. 36, no. 2, pp. 176-187, 1972.

[20] S. Iyengar and D. R. Kinder, News That Matters: Television and American Opinion. Chicago: University of Chicago Press, 1987.

[21] R. M. Entman, "Framing: Toward clarification of a fractured paradigm," Journal of Communication, vol. 43, no. 4, pp. 51-58, 1993.

[22] D. Chong and J. N. Druckman, "Framing theory," Annual Review of Political Science, vol. 10, pp. 103-126, 2007.

[23] T. Brader, Campaigning for Hearts and Minds: How Emotional Appeals in Political Ads Work. Chicago: University of Chicago Press, 2006.

[24] P. Norris, A Virtuous Circle: Political Communications in Postindustrial Societies. Cambridge: Cambridge University Press, 2000.

[25] H. Tajfel and J. C. Turner, "An integrative theory of intergroup conflict," in The Social Psychology of Intergroup Relations, W. G. Austin and S. Worchel, Eds. Monterey, CA: Brooks/Cole, 1979, pp. 33-47.

[26] S. M. Lipset and S. Rokkan, "Cleavage structures, party systems, and voter alignments: An introduction," in Party Systems and Voter Alignments: Cross-National Perspectives, S. M. Lipset and S. Rokkan, Eds. New York: Free Press, 1967, pp. 1-64.

[27] A. Lijphart, Democracy in Plural Societies: A Comparative Exploration. New Haven: Yale

University Press, 1977.
[28] G. A. Akerlof and R. E. Kranton, “Economics and identity,” Quarterly Journal of Economics, vol. 115, no. 3, pp. 715-753, 2000.
[29] A. Honneth, The Struggle for Recognition: The Moral Grammar of Social Conflicts. Cambridge, MA: MIT Press, 1995.
[30] N. Fraser, “Rethinking recognition,” New Left Review, no. 3, pp. 107-120, 2000.
[31] D. E. Stokes, “Spatial models of party competition,” American Political Science Review, vol. 57, no. 2, pp. 368-377, 1963.
[32] J. Nash, “Non-Cooperative Games,” Annals of Mathematics, Second Series, vol. 54, no. 2, pp. 286-295, Sep. 1951.
[33] G. Aher, R. I. Arriaga, and A. Tauman, “Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies,” 2023.
[34] L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate, “Out of one, many: Using language models to simulate human samples,” *Polit. Anal.*, vol. 31, no. 3, pp. 337–351, 2023.
[35] A. Lyman, B. Hepner, L. P. Argyle, E. C. Busby, J. R. Gubler, and D. Wingate, “Balancing large language model alignment and algorithmic fidelity in social science research,” *Sociol. Methods Res.*, vol. 54, no. 3, pp. 1110–1155, 2025.
[36] P. Chauhan, N. Sharma, and G. Sikka, “The emergence of social media data and sentiment analysis in election prediction,” *J. Ambient Intell. Humaniz. Comput.*, vol. 12, no. 2, pp. 2601–2627, 2021.
[37] A. Gaur and D. K. Yadav, “A comprehensive analysis of forecasting elections using social media text,” *Multimed. Tools Appl.*, vol. 84, no. 29, pp. 35693–35725, 2025.
[38] D. Gayo-Avello, “No, You Cannot Predict Elections with Twitter,” IEEE Internet Computing, vol. 16, no. 6, pp. 91–94, 2012, doi: 10.1109/MIC.2012.137.
[39] V. Lampos, “On voting intentions inference from Twitter content: a case study on UK 2010 General Election,” arXiv:1204.0423, 2012.
[40] K.-C. Yang, P.-M. Hui, and F. Menczer, “How Twitter Data Sampling Biases U.S. Voter Behavior Characterizations,” arXiv:2006.01447, 2020.
[41] A. Balasubramanian, V. Zou, H. Narayana, C. You, L. Luceri, and E. Ferrara, “A Public Dataset Tracking Social Media Discourse about the 2024 U.S. Presidential Election on Twitter/X,” arXiv:2411.00376, 2024.
[42] S. Scarano, V. Vasudevan, M. Samory, K.-C. Yang, J. Yang, and P. A. Grabowicz, “Election Polls on Social Media: Prevalence, Biases, and Voter Fraud Beliefs,” arXiv:2405.11146, 2024.
[43] D. K. Baissa and C. Rainey, “When BLUE is not best: non-normal errors and the linear model,” *Polit. Sci. Res. Methods*, vol. 8, no. 1, pp. 136–148, 2020.
[44] M. Höfler, “Robust tests should be the default, not the backup,” *Peer Community Journal*, vol. 6, article e1, 2026, doi: 10.24072/pcjournal.670.
[45] J. Esarey and N. Danneman, “A quantitative method for substantive robustness assessment,” *Polit. Sci. Res. Methods*, vol. 3, no. 1, pp. 95–111, 2015.
[46] S. Krasa and M. Polborn, “Competition between Specialized Candidates,” *Am. Polit. Sci. Rev.*, vol. 104, no. 4, pp. 745–765, 2010, doi: 10.1017/S0003055410000468.
[47] C. Zhou, “Interactions among simultaneous elections,” Political Science Research and Methods,

vol. 14, no. 1, pp. 40–52, 2026, doi: 10.1017/psrm.2024.66.

[48] I. Poggi, T. Scaramella, S. Violini, S. Careri, M. D. Epure, and D. Dragoni, "The Ten Minutes That Shocked the World — Teaching Generative AI to Analyze the Trump – Zelensky Multimodal Debate," Information, vol. 17, no. 2, p. 136, 2026.

[49] P. Song *et al.*, "Machine learning and game theory for cyber governance: Enhancing public opinion and regional sustainable development," *PLoS One*, vol. 19, no. 12, p. e0308317, 2024.

[50] H. Licht, "Computational emotion analysis with multimodal LLMs Current evidence on an emerging methodological opportunity," 2026.

[51] R. Parikh, S. H. Cen, and C. Podimata, "Do LLMs Track Public Opinion?," arXiv:2602.06302, 2026.

[52] H. A. Von Der Gracht, "Consensus measurement in Delphi studies: review and implications for future quality assurance," *Technol. Forecast. Soc. Change*, vol. 79, no. 8, pp. 1525–1536, 2012.

[53] S. Kaplan and B. J. Garrick, "On the quantitative definition of risk," *Risk Anal.*, vol. 1, no. 1, pp. 11–27, 1981.

[54] Reuters, "Coalition led by Iraqi PM Sudani comes first in Iraq's election, commission says," Nov. 12, 2025.

[55] Reuters, "Iraqi PM-led coalition tops Iraq election with 46 seats, commission says," Nov. 17, 2025.

[56] Associated Press, "Iraq's election result ratified by Supreme Federal Court as premiership remains up for grabs," Dec. 14, 2025.

Modeling Misinformation Propagation Risk in Social Networks Using a Mamdani-Type Fuzzy Inference System

Gulsum Akkuzu Kaya[1], Ibrahim Akdi[2], Abdul Badwan

[1,2] Krsehir Ahi Evran University, Faculty of Engineering and Arcthitecture, Computer Engineering, Postal Code, KIRSEHIR, TURKIYE

*Corresponding author e-mail: gulsum.akkuzukaya@ahievran.edu.tr

Abstract

The rapid dissemination of misinformation through social networks has become an important challenge due to its potential to influence individuals, institutions, and public opinion. The dynamic and uncertain nature of social media interactions makes misinformation propagation difficult to model using conventional deterministic approaches. This study proposes a Mamdani-type Fuzzy Inference System (FIS) for estimating the risk of misinformation propagation at the user level. The proposed model incorporates four input variables: trust score, posting frequency, sentiment deviation, and network centrality. Each input is represented using linguistic variables and fuzzy membership functions, while the output is defined as a continuous misinformation propagation risk score ranging from 0 to 1. A hybrid membership function structure combining triangular and trapezoidal membership functions is employed to represent gradual transitions between linguistic categories. A rule base consisting of 30 linguistic IF–THEN rules is developed based on domain knowledge concerning user behavior and social network characteristics. The inference process follows the Mamdani max–min approach, and the resulting fuzzy output is converted into a crisp risk score using the centroid defuzzification method. Experiments conducted on 1,000 synthetically generated user profiles demonstrate the applicability of the proposed framework for modeling nonlinear interactions among behavioral and network-related factors. Sensitivity analysis indicates that posting frequency and network centrality have substantial effects on the estimated risk under the experimental conditions. The proposed approach also provides an interpretable inference process in which the contribution of individual linguistic rules can be explicitly examined. These characteristics make the model suitable as an interpretable decision-support framework for studying misinformation propagation risk in social network environments.

Key words: *Fuzzy Logic, Mamdani Fuzzy Inference System, Misinformation, Social Networks, Risk Modeling, Membership Functions, Sensitivity Analysis, Explainable Artificial Intelligence*

INTRODUCTION

Social networking platforms have fundamentally transformed the way information is produced, disseminated, and consumed. Users can generate and share content with large audiences within a short period of time, creating highly dynamic information environments. Although these platforms provide significant benefits for communication and information access, they also facilitate the rapid

dissemination of inaccurate or misleading information. Consequently, misinformation has become an important research topic in areas including social network analysis, information retrieval, artificial intelligence, and computational social science.

A fundamental distinction in the literature is made between *misinformation* and *disinformation*. Misinformation generally refers to information that is false or inaccurate without necessarily implying an intention to deceive, whereas disinformation refers to false information that is deliberately created or disseminated with the intention of misleading others (Wardle & Derakhshan, 2017). Although the distinction is important, both forms of information disorder can contribute to harmful propagation processes in social networks. Therefore, this study uses the broader concept of *misinformation propagation risk* to describe the likelihood that a user's behavioral and network characteristics may contribute to the dissemination of potentially inaccurate information.

The propagation of information in social networks is influenced by multiple interacting factors. User-level characteristics such as historical trustworthiness and posting behavior can affect the likelihood and scale of information dissemination. At the same time, network position plays an important role because users occupying central positions may have access to larger or more diverse audiences. Previous research has demonstrated that structural properties of social networks can influence the diffusion of information, while emotional characteristics of content can also affect its dissemination dynamics (Brady et al., 2017; Vosoughi et al., 2018).

A major challenge in modeling such processes is that many relevant variables cannot be naturally represented using strict binary categories. For example, it is difficult to determine an exact threshold at which a user should be considered either "trustworthy" or "untrustworthy." Similarly, a posting frequency of 19 messages per day and one of 20 messages per day should not necessarily result in completely different risk classifications. These characteristics indicate the presence of vagueness and gradual transitions that are difficult to represent using conventional crisp decision rules.

Fuzzy set theory, introduced by Zadeh (1965), provides a mathematical framework for representing such gradual concepts through membership degrees between 0 and 1. Instead of assigning an observation exclusively to a single category, fuzzy logic allows it to have partial membership in multiple linguistic categories simultaneously. This property makes fuzzy logic particularly suitable for decision-support problems involving uncertain and imprecise concepts.

The interpretability of fuzzy inference systems is particularly relevant to explainable artificial intelligence (XAI). Unlike many black-box predictive models, fuzzy rule-based systems can represent decision processes through linguistic concepts and explicit IF–THEN rules, allowing the relationship between input conditions and model outputs to be examined directly. Recent research has highlighted the potential of fuzzy inference systems and fuzzy rule-based models as interpretable machine learning approaches, while also emphasizing the importance of maintaining a balance between predictive complexity and semantic interpretability (Cao et al., 2024; Pickering et al., 2025). In the present study, the use of linguistic variables and an explicitly defined rule base enables individual rule activations to be inspected, providing a transparent representation of how combinations of behavioral and network characteristics contribute to the estimated risk score.

In the context of misinformation propagation, an interpretable model can be useful because the estimated risk should not necessarily be treated as an opaque prediction. Instead, decision makers may benefit from understanding which characteristics contribute to a particular risk level. For example, a user characterized by low trust, high posting frequency, and high network centrality may activate several high-risk linguistic rules simultaneously. A fuzzy inference system can explicitly represent and trace these rule activations.

The objective of this study is therefore to develop an interpretable Mamdani-type fuzzy inference framework for estimating misinformation propagation risk based on user-level behavioral and network characteristics. The proposed system incorporates four input variables: trust score, posting frequency, sentiment deviation, and network centrality. These variables are transformed into linguistic representations through fuzzy membership functions and evaluated using a manually designed rule base. The final risk score is obtained through Mamdani max–min inference and centroid defuzzification.

The main contributions of this study can be summarized as follows:

- A multidimensional fuzzy framework is developed for modeling misinformation propagation risk in social networks.
- Four complementary user-level and network-level variables are integrated into a single Mamdani fuzzy inference system.
- A hybrid membership-function structure combining triangular and trapezoidal functions is employed to represent gradual transitions between linguistic categories.
- A 30-rule linguistic rule base is designed to represent interactions among trust, posting behavior, sentiment deviation, and network centrality.
- A sensitivity analysis is conducted to investigate the effects of the input variables on the estimated risk.
- The proposed framework provides an interpretable inference process in which individual rule activations can be examined.

RELATED WORKS

Research on misinformation in social networks has primarily focused on understanding how inaccurate information is generated, disseminated, and detected. Shu et al. (2017) provided a comprehensive review of fake news detection on social media and emphasized that misinformation-related analysis cannot rely solely on textual content. Their review highlighted the importance of incorporating social context, user engagement, and network-related information, since social media interactions generate large, heterogeneous, and noisy auxiliary data. This perspective is particularly relevant to the present study because the proposed framework models misinformation propagation risk using user-level behavioral and network characteristics rather than attempting to determine the truthfulness of individual posts directly.

Beyond detection, empirical studies have demonstrated that the diffusion characteristics of false information differ from those of true information. Vosoughi et al. (2018) analyzed approximately 126,000 news cascades on Twitter and showed that false news tended to diffuse farther, faster, deeper, and more broadly than true news. Their findings underline the importance of studying propagation dynamics rather than restricting the problem to binary misinformation detection. In this context, a risk-

oriented formulation can complement detection-oriented approaches by estimating the potential propagation risk associated with different user and network configurations.
User activity and network position are also important components of information diffusion. Social network analysis has long recognized centrality as a structural property that characterizes the position and influence of nodes within a network. Freeman (1978) systematically examined centrality measures and their conceptual foundations, establishing a basis for analyzing the structural importance of individual network positions. More recent research on misinformation diffusion has further emphasized the importance of network structure and source-related characteristics in determining how misleading information can spread through online social networks (Chen et al., 2023; Zareie et al., 2022). In misinformation-related settings, centrality can therefore provide a structural indicator of a user's potential reach and diffusion capacity. The present study incorporates network centrality as one of its principal input variables and combines several centrality measures into a composite score.

Emotional characteristics provide another important dimension of information diffusion. Brady et al. (2017) investigated the relationship between moral-emotional language and the diffusion of content on social networks and showed that emotional characteristics can influence how widely content is shared. This line of research supports the inclusion of sentiment-related information in propagation-oriented models. However, the present study does not assume that a particular sentiment direction is intrinsically associated with higher risk. Instead, sentiment deviation is represented as a linguistic variable, allowing stable and substantially deviating sentiment states to contribute differently to the inferred risk depending on their interaction with the other variables.

Although machine learning and data-mining approaches have become increasingly common in misinformation research, many of these approaches primarily emphasize classification or prediction. Such models can be effective for distinguishing misinformation from reliable information, but their outputs may provide limited insight into how individual behavioral and structural factors jointly contribute to a particular decision. Shu et al. (2017) also identified the heterogeneity and complexity of social-media data as an important challenge for misinformation analysis. These observations motivate complementary approaches that explicitly represent uncertainty and gradual transitions between behavioral categories.

Fuzzy set theory provides a natural mathematical framework for representing such gradual concepts. Zadeh (1965) introduced fuzzy sets to represent partial membership through membership functions whose values range between zero and one. Unlike crisp classification, fuzzy representations allow an observation to simultaneously belong to multiple linguistic categories with different degrees of membership. This property is relevant to social-network risk assessment because concepts such as “low trust,” “high activity,” or “high centrality” do not necessarily have universally valid sharp boundaries.

Mamdani and Assilian (1975) demonstrated how linguistic knowledge expressed through heuristic rules could be transformed into an automatic control strategy using fuzzy inference. The resulting rule-based formulation provides an interpretable reasoning mechanism in which the relationship between inputs and outputs can be explicitly represented through linguistic IF–THEN rules. This characteristic makes Mamdani-type inference particularly suitable for decision-support applications where the reasoning process is important in addition to the numerical output.

Despite the extensive literature on misinformation detection and information diffusion, comparatively less attention has been given to interpretable, user-level risk assessment frameworks that jointly represent behavioral, affective, and structural characteristics through linguistic rules. Existing studies establish the importance of diffusion dynamics, user activity, network structure, and emotional

characteristics, while fuzzy-set research provides mechanisms for representing uncertainty and gradual transitions. Building on these complementary research directions, this study develops a Mamdani-type fuzzy inference framework that integrates trust score, sharing frequency, sentiment deviation, and network centrality into a continuous misinformation propagation risk score. The proposed approach differs from conventional detection-oriented formulations by focusing on an interpretable risk assessment mechanism rather than a binary misinformation label. Furthermore, the use of explicit linguistic rules enables the contribution of individual input combinations to be examined, providing a transparent representation of nonlinear and context-dependent relationships.

MATERIAL AND METHODS

System Arhitecture

The developed system consists of five integrated main modules: (i) Data Loading and Preprocessing Module, (ii) Fuzzy Inference System (FIS) Core, (iii) Rule Base Manager, (iv) Visualization and Analysis Module, and (v) Comparison and Evaluation Module. Built using the Python 3.10 programming language, the system utilizes the scikit-fuzzy, NetworkX, NumPy, Pandas, and Matplotlib libraries. The interface layer was designed as a separate web component, enabling users to interactively adjust input variables. Figure 1 presents a view of the system's architecture.



Figure 1: System Architecture

The system's data flow follows an eight-stage pipeline, ranging from the processing of raw social network data to the generation of risk predictions. In the first stage, the Facebook social network graph obtained from the Stanford Network Analysis Project (SNAP) repository is loaded to create an undirected graph object G = (V, E) using the NetworkX library, where V represents the set of nodes and E the set of edges. In cases where real graph data is unavailable, a synthetic scale-free graph adhering to a power-law distribution is generated using the Barabási-Albert preferential attachment model.

The network centrality score is calculated as the weighted average of four distinct centrality metrics. The composite centrality score is defined by the following formula:

$$C_composite(v) = 0.40 \cdot C_degree(v) + 0.30 \cdot C_betweenness(v) + 0.20 \cdot C_closeness(v) + 0.10 \cdot C_pagerank(v)$$

Here, C_degree(v) represents degree centrality, C_betweenness(v) represents betweenness centrality, C_closeness(v) represents closeness centrality, and C_pagerank(v) represents the PageRank score. All metrics are normalized to the [0, 1] range, and the composite score also falls within this range. This

weighting aims to provide a balanced reflection of the node's number of direct connections (degree), its role as a bridge within the network (betweenness), and its propagation speed (closeness, PageRank).

Dataset

The experimental evaluation was conducted using a synthetic dataset consisting of 1,000 social network user profiles. The synthetic dataset was designed to provide controlled and reproducible input conditions for evaluating the internal behavior of the proposed Mamdani-type fuzzy inference system. Each profile was represented using four input variables: trust score, sharing frequency, sentiment deviation, and network centrality. Trust score was represented in the [0,1] interval, sharing frequency was represented as the number of posts per day within the [0,50] range, sentiment deviation was represented within the [-1,+1] interval, and network centrality was represented within the [0,1] interval. To generate heterogeneous user profiles, probabilistic distributions were employed for the input variables. Trust scores were generated using a log-normal distribution, sharing frequency using a Poisson distribution, and sentiment deviation using a Beta-based transformation. Centrality values were generated within the predefined [0,1] range. A fixed random seed of 42 was used to ensure reproducibility. The generated profiles were used exclusively to evaluate the behavior of the fuzzy inference model and its rule-based interactions. Therefore, the resulting risk scores should not be interpreted as ground-truth labels or as empirical estimates of real-world misinformation propagation risk. Instead, the synthetic dataset provides a controlled experimental environment for examining the interpretability, nonlinear behavior, and sensitivity of the proposed fuzzy framework. Real-world social-network datasets, including graph-based bot-detection datasets such as TwiBot-22, can be incorporated in future work to validate whether the relationships modeled by the proposed fuzzy system generalize to observed user behavior.

Table 1. Definition of Input and Output Variables in the Fuzzy Risk Assessment Model

Variable	Symbol	Universe of Discourse	Linguistic Terms	Membership Function Type
Trust Score	G(x)G(x)	[0, 1]	Low, Medium, High	Hybrid (trapmf + trimf)
Sharing Frequency	P(x)P(x)	[0, 50] posts/day	Low, Medium, High	Hybrid (trapmf + trimf)
Sentiment Deviation	D(x)D(x)	[−1, +1]	Stable, Moderate, Variable	Hybrid (trapmf + trimf)
Network Centrality	C(x)C(x)	[0, 1]	Low, Medium, High	Hybrid (trapmf + trimf)
Risk Level	R(x)R(x)	[0, 1]	Low, Medium, High	Hybrid (trapmf + trimf)

Membership Functions abd Rule Definition

The proposed fuzzy risk assessment model employs four input variables—trust score (G(x)), sharing frequency (P(x)), sentiment deviation (D(x)), and network centrality (C(x))—and one output variable, risk level (R(x)). The corresponding universes of discourse, linguistic terms, and membership function types are summarized in Table X.

Membership Function Design: A hybrid membership-function architecture was adopted to represent the uncertainty and gradual transitions inherent in social-network behavioral characteristics. Specifically, trapezoidal membership functions (trapmf) were employed for the extreme linguistic terms, whereas triangular membership functions (trimf) were used for intermediate terms. This design allows saturation effects at the boundaries of the input domains to be represented by trapezoidal functions while providing smooth transitions between neighboring linguistic categories through triangular functions.

For the trust score (G(x)\in[0,1]), three linguistic terms—Low, Medium, and High—were defined. The corresponding membership functions are:

$$\mu_{G,Low}(x) = \{\ 1, \quad x \le 0.15 \mid (0.40-x)/0.25, \quad 0.15 < x \le 0.40 \mid 0, \quad x > 0.40\ \}$$

$$\mu_{G,Medium}(x) = \{\ (x-0.20)/0.30, \quad 0.20 < x \le 0.50 \mid (0.80-x)/0.30, \quad 0.50 < x \le 0.80 \mid 0, \quad \text{other cases}\}$$

$$\mu_{G,High}(x) = \{\ 0, \quad x \le 0.60 \mid (x-0.60)/0.25, \quad 0.60 < x \le 0.85 \mid 1, \quad x > 0.85\ \}$$

The sharing frequency (P(x)\in[0,50]), measured as the number of posts per day, is represented by Low, Medium, and High linguistic terms. The Low and High terms use trapezoidal functions with parameter vectors ([0,0,5,15]) and ([25,40,50,50]), respectively, while the Medium term is represented by a triangular function with parameters ([8,20,35]).

For sentiment deviation, (D(x)\in[-1,1]), the model adopts a symmetric representation to capture deviations in both positive and negative directions. The Stable term is modeled using a central trapezoidal function with parameters ([-0.30,-0.10,0.10,0.30]). In Figure 2, the Moderate and Variable terms are represented by symmetric pairs of triangular and trapezoidal functions, respectively.

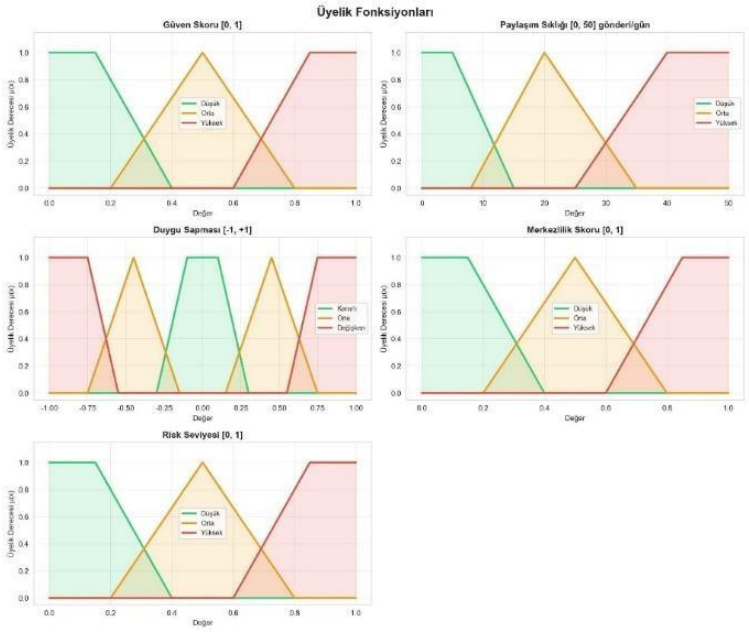


Figure 2: Sentiment Deviation Membership Functions

This symmetric construction enables the model to treat substantial positive and negative sentiment deviations consistently rather than assigning risk according to the direction of the deviation alone.

The network centrality score (C(x)\in[0,1]) is modeled using the same parameterization adopted for the trust score, with Low, Medium, and High linguistic terms. Thus, the membership functions are defined by the parameter vectors ([0,0,0.15,0.40]), ([0.20,0.50,0.80]), and ([0.60,0.85,1,1]), respectively.

The output variable, risk level (R(x)\in[0,1]), is also represented using Low, Medium, and High linguistic terms. The same hybrid trapezoidal–triangular architecture is employed to maintain consistency between the input and output representations. For interpretation purposes, the resulting crisp risk score is categorized as Low Risk for (R<0.35), Medium Risk for (0.35\leq R<0.65), and High Risk for (R\geq0.65). These thresholds are used as operational decision boundaries of the proposed system rather than as universal risk thresholds.

Fuzzy Rule Definition: The fuzzy rule base was constructed to translate combinations of behavioral and network characteristics into an interpretable risk assessment. The rules follow the Mamdani fuzzy inference paradigm and use expert-informed behavioral relationships between trust, activity, sentiment deviation, and network position. Four input variables with three linguistic terms each theoretically yield (3^4=81) possible combinations. Instead of including all combinations indiscriminately, 30 semantically meaningful rules were selected to cover representative low-, medium-, and high-risk behavioral patterns while avoiding redundant or implausible combinations. The rule base contains rules with two, three, and four antecedents, allowing the model to represent both simple behavioral relationships and compound risk conditions. The system has 30 fuzzy rules, Table 2 shows some examples of the rules. The rule base was designed ccording to three principles: semantic validity, ensuring that each rule represents a meaningful behavioral configuration; coverage, ensuring that relevant regions of the input space are represented; and consistency, reducing contradictory assignments between similar behavioral profiles.

Table 2. Fuzzy Rules of the System

Rule	Trust	Activity	Sentiment	Centrality	Risk	Rationale
1	Low	High	—	—	**High**	Low trust combined with high posting activity indicates a potentially harmful profile.
2	High	Low	Stable	—	**Low**	High trust, low activity, and stable sentiment indicate a reliable and cautious profile.
3	Low	—	Variable	—	**High**	Low trust combined with sentiment variability increases the potential risk.
4	High	Low	—	—	**Low**	High trust and selective posting behavior indicate a relatively safe profile.
5	—	High	—	High	**High**	Highly active and highly central users have greater potential to disseminate harmful information.
7	—	—	Variable	High	**High**	High centrality combined with sentiment variability may increase the user's disruptive influence.

8	Low	—	—	High	**High**	A highly central but low-trust user represents a potentially significant threat.
9	Low	High	Variable	—	**High**	The combination of low trust, high activity, and variable sentiment represents a high-risk profile.
12	Low	High	—	High	**High**	A low-trust, highly active, and highly central user has strong potential for harmful information diffusion.
25	High	High	Stable	High	**Medium**	A highly trusted, active, and central user with stable sentiment represents a reliable but influential profile requiring contextual assessment.

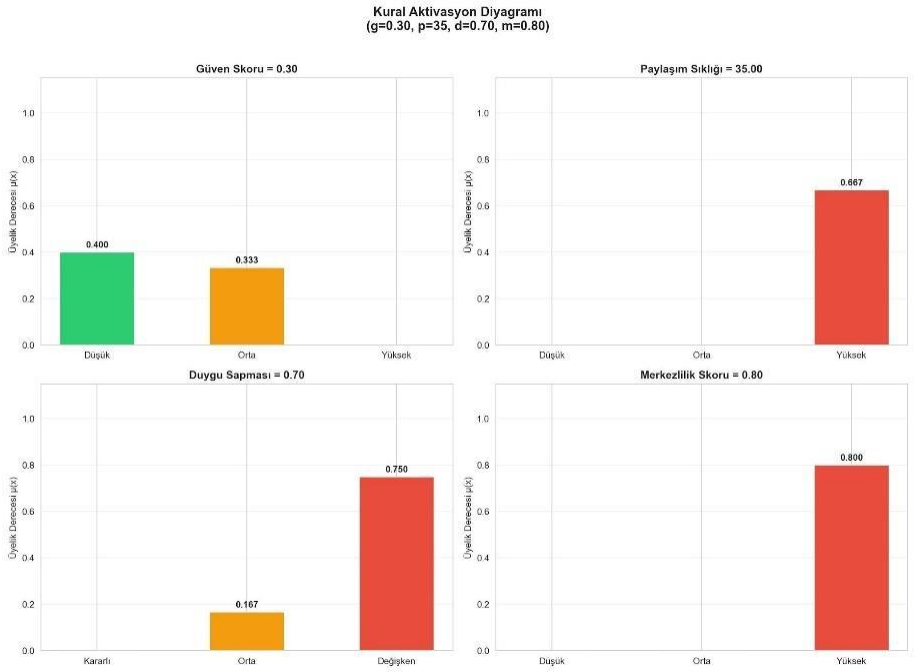


Figure 3: Rule Activation Diagram for Sample Input

Figure 3 illustrates the rule activation diagram for the example input values (g=0.30), (p=35), (d=0.70), and (m=0.80). For the trust score of 0.30, the corresponding membership degrees are calculated as (\mu_{\mathrm{Low}}=0.400) and (\mu_{\mathrm{Medium}}=0.333). For the posting frequency of 35 posts/day, the membership degree for the High activity category is (\mu_{\mathrm{High}}=0.667). Similarly, the sentiment deviation of 0.70 yields (\mu_{\mathrm{Variable}}=0.750), while the centrality score of 0.80 results in (\mu_{\mathrm{High}}=0.800).

These membership values determine the activation strength of the corresponding fuzzy rules. For example, Rule 1, defined as Trust = Low AND Activity = High → Risk = High, is activated with a firing strength of (\min(0.400, 0.667)=0.400), according to the Mamdani inference mechanism. This indicates

that the rule contributes to the final high-risk output with an activation degree of 0.400.

RESULTS

The proposed Mamdani fuzzy inference system (FIS) was evaluated on a synthetic dataset consisting of 1,000 social network user profiles. Each profile was represented by a four-dimensional input vector ([G,P,D,C]), corresponding to trust score, sharing frequency, sentiment deviation, and network centrality, respectively. To provide reproducible experimental conditions, a fixed random seed of 42 was used during data generation. Log-normal and Poisson distributions were employed to generate variables with heterogeneous distributions, and all generated values were constrained to their predefined universal domains. The resulting risk scores exhibited a right-skewed distribution, with most observations concentrated in the low-risk region. The mean risk score was 0.197, with a standard deviation of 0.085. The minimum and maximum risk scores were 0.147 and 0.650, respectively, while the median was approximately 0.165. The first and third quartiles were approximately 0.155 and 0.235, respectively. These statistics indicate that the generated population was dominated by relatively low-risk profiles, while a smaller number of profiles exhibited substantially higher risk scores. Figure 3 shows the risk level distribution histogram and the Kernel Density Estimation (KDE) curve for a dataset of 1,000 samples.

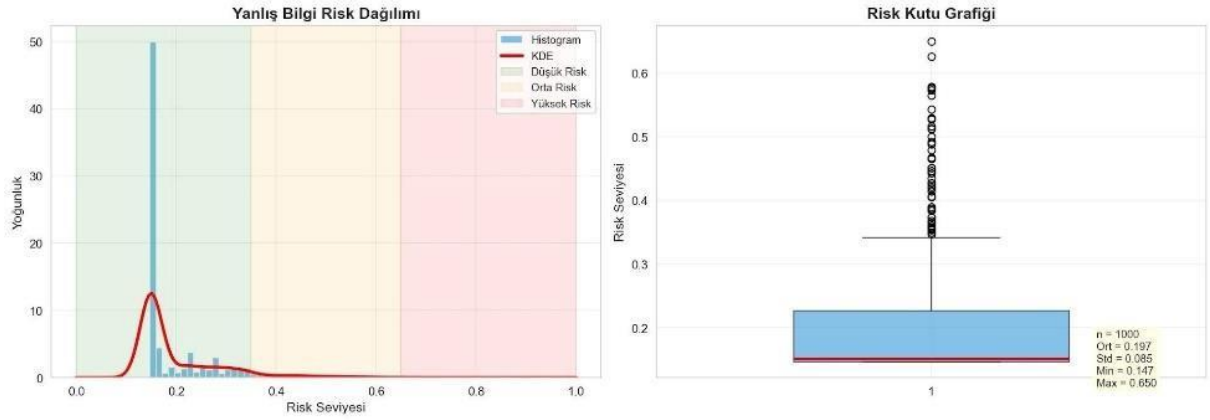


Figure 4: Rule Activation Diagram for Sample Input

The two-dimensional risk surfaces generated by the Mamdani FIS demonstrate continuous and nonlinear relationships between the input variables and the resulting risk score. The trust–sharing-frequency surface shows that low trust combined with high sharing activity produces the highest risk values. In particular, profiles with (G<0.3) and (P>35) reach risk values of approximately 0.65–0.70. Increasing the trust score generally reduces the resulting risk; however, high sharing frequency can partially offset this reduction. The trust–centrality and centrality–sharing-frequency surfaces further demonstrate that risk is determined by the interaction among multiple variables rather than by a single factor. In particular, high centrality combined with intensive sharing activity produces higher risk values, whereas profiles characterized by low centrality and moderate activity remain in lower-risk regions.

These surfaces also demonstrate the continuity of the fuzzy decision process. Rather than producing

abrupt transitions between predefined classes, the system generates gradual changes in risk as the input variables vary. Figures 5 and 7 present risk surface graphs based on composition of different two variables.

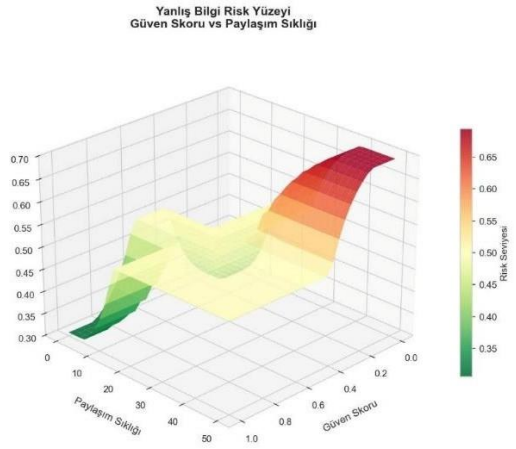


Figure 5 Risk surface based on Trust Score (G) and Sharing Frequency (P) variables.

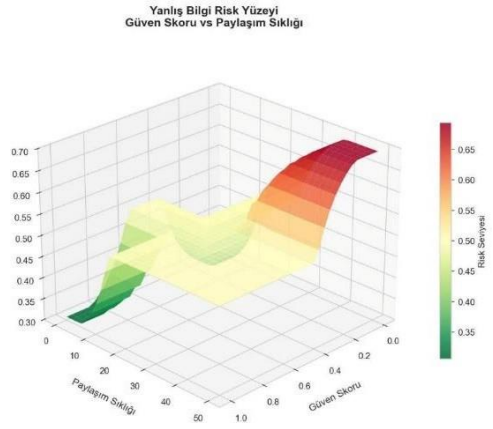


Figure 6 Risk Surface Based on Trust Score (G) and Network Centrality (C) Variables

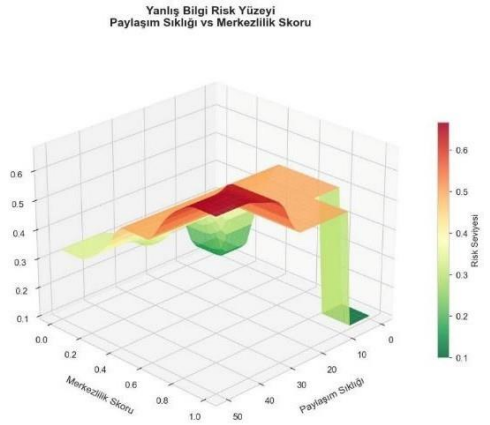


Figure 7 Risk Surface Based on Network Centrality (C) and Sharing Frequency (P) Variables

Scenario-Based Evaluation

Eight representative user profiles were evaluated to examine the behavior of the proposed system under interpretable social-network scenarios. The “Super Spreader,” “Troll Account,” and “Unreliable Active” profiles all received a high-risk score of 0.853. These profiles are characterized by low trust, relatively high sharing frequency, substantial sentiment deviation, and high network centrality. The “Emotional Sharer,” “Influence Hub,” and “Average User” profiles received medium-risk classifications, with risk values of 0.500, 0.480, and 0.500, respectively. These results indicate that intermediate combinations of trust, activity, sentiment, and centrality do not necessarily result in either extreme risk category. Conversely, the “Quiet Reliable” and “Reliable Passive” profiles received the minimum observed risk value of 0.147 and were classified as low risk. Both profiles are

characterized by high trust and low sharing activity, together with relatively low sentiment deviation and centrality. Overall, the scenario analysis demonstrates that the system produces distinguishable risk levels for qualitatively different user profiles.

Table 3. User Profile Scenarios and Corresponding Risk Levels

Profile	Trust	Sharing	Sentiment	Centrality	Risk	Category
Super Spreader	0.10	45	0.90	0.90	0.853	High
Troll Account	0.10	40	0.85	0.85	0.853	High
Unreliable Active	0.15	38	0.80	0.88	0.853	High
Emotional Sharer	0.50	25	0.60	0.50	0.500	Medium
Influence Hub	0.65	20	0.20	0.80	0.480	Medium
Average User	0.50	15	0.30	0.50	0.500	Medium
Quiet Reliable	0.90	5	0.10	0.15	0.147	Low
Reliable Passive	0.85	3	0.05	0.20	0.147	Low

DISCUSSION AND CONCLUSION

The experimental results demonstrate that the proposed Mamdani fuzzy inference system (FIS) provides an interpretable framework for assessing misinformation propagation risk by jointly considering trust score, sharing frequency, sentiment deviation, and network centrality. Unlike a purely additive formulation, the fuzzy rule base enables the effect of one variable to depend on the states of the other variables, allowing nonlinear and context-dependent risk relationships to be represented.

The risk distribution obtained from the 1,000 synthetic user profiles was right-skewed, with a mean risk score of 0.197 and a standard deviation of 0.085. Most profiles were concentrated in the low-risk region, while a smaller number of profiles produced substantially higher risk values. This distribution is consistent with the intended design of the synthetic population, in which relatively low-risk profiles constitute the majority. However, because the experiments were conducted on synthetically generated profiles, this distribution should be interpreted as a characteristic of the proposed experimental setting rather than as an estimate of the actual distribution of misinformation risk in social networks.

The scenario-based analysis further demonstrated the interpretability of the proposed system. Profiles characterized by low trust, high sharing frequency, high sentiment deviation, and high centrality, such as the “Super Spreader,” “Troll Account,” and “Unreliable Active” scenarios, received a high risk score of 0.853. In contrast, highly trusted and relatively passive profiles received the minimum observed risk score of 0.147. Intermediate profiles produced intermediate risk values. These results illustrate that the rule-based architecture can distinguish qualitatively different behavioral configurations without requiring a binary decision at the input level.

The sensitivity analysis provides additional insight into the internal behavior of the model. Sharing frequency and network centrality produced the largest independent changes in risk, with Pearson correlations of approximately ($r=+0.72$) and ($r=+0.68$), respectively. This indicates that, under the tested conditions, increased activity and greater network centrality are strongly associated with higher risk outputs. From a network perspective, this behavior is consistent with the role of highly active and centrally positioned users in facilitating broader information diffusion. Nevertheless, the present analysis evaluates the behavior of the fuzzy model rather than establishing a causal relationship between

centrality, activity, and real-world misinformation propagation.

The interaction between trust and sharing frequency is particularly informative. At low trust levels, increasing sharing frequency resulted in a substantially larger increase in risk, whereas the same increase had a more limited effect when trust was high. This result illustrates the ability of the fuzzy rule base to model conditional relationships among variables. In other words, the risk contribution of sharing frequency is not treated as independent of trust. Such contextual behavior is one of the main characteristics distinguishing the proposed rule-based formulation from a simple weighted linear model.

The sensitivity analysis also showed that trust had only a limited isolated effect under the selected fixed conditions, with the risk remaining approximately within the range [0.497, 0.501]. This finding should not be interpreted as indicating that trust is unimportant in the complete model. Instead, it suggests that the influence of trust is highly dependent on the configuration of the remaining variables. The interaction surfaces and scenario experiments support this interpretation by showing substantially different risk levels for combinations involving low versus high trust.

Sentiment deviation exhibited a nonlinear U-shaped relationship with the risk output. Risk was lowest around the stable-sentiment region and increased toward more extreme deviations. This behavior demonstrates the ability of the fuzzy representation to encode nonlinear linguistic relationships that would be difficult to express using a simple linear contribution. It also illustrates the role of the "Stable" linguistic term as a risk-reducing region within the rule base.

The comparison with the weighted linear baseline (WLB) provides a useful reference for examining the behavior of the proposed model. The WLB achieved a Pearson correlation of 0.591 with the continuous FIS outputs and 85.9% categorical agreement when the FIS classifications were used as the reference. The corresponding weighted precision, recall, and F1-score were 0.953, 0.859, and 0.891, respectively. These values indicate that the linear model reproduces a substantial proportion of the categorical decisions generated by the FIS.

However, these metrics should not be interpreted as independent predictive accuracy measures for the proposed FIS. The FIS itself was used as the reference for the categorical comparison, and no independently labeled ground-truth dataset was available. Therefore, the comparison primarily measures the degree of agreement between the two formulations rather than demonstrating that one model predicts real-world misinformation risk more accurately than the other. Similarly, the Pearson correlation of 0.591 indicates moderate association between the two continuous outputs but does not establish predictive superiority.

The comparison nevertheless highlights an important methodological distinction. The WLB offers a computationally simple formulation based on fixed coefficients, whereas the Mamdani FIS explicitly incorporates linguistic rules and nonlinear interactions. The fuzzy framework therefore provides a more direct mechanism for encoding expert knowledge and producing traceable decisions. This property is particularly relevant in risk assessment settings where the reasoning behind a risk score may be as important as the score itself.

The proposed framework also has several limitations. First, the experimental evaluation relies on synthetic data. Although the data-generation process uses constrained probability distributions and a fixed random seed to ensure reproducibility, synthetic profiles cannot fully represent the heterogeneity, temporal dynamics, and behavioral complexity of real social-network users. Consequently, the current

findings should be considered evidence of the internal behavior and feasibility of the proposed framework rather than evidence of its real-world predictive performance.

Second, the fuzzy rule base and membership functions were manually designed. This improves interpretability but introduces dependence on expert assumptions and may make the system increasingly difficult to maintain as additional variables are incorporated. The number of possible rule combinations can also grow rapidly with the number of linguistic variables. Future work can therefore investigate automatic rule extraction, membership-function optimization, and adaptive fuzzy approaches such as ANFIS.

Third, the current implementation is primarily static and does not explicitly model the temporal evolution of user behavior. Misinformation propagation is inherently dynamic, and a user's risk profile may change as their activity, sentiment, network position, and interaction patterns evolve. Incorporating temporal features or sequential modeling could therefore provide a more realistic representation of evolving misinformation risk.

Finally, although the current implementation is suitable for the experimental dataset, computational overhead may become relevant in large-scale real-time applications. The creation of a separate simulation object for individual observations can introduce unnecessary processing overhead. Batch inference, vectorization, and caching strategies could be investigated to improve scalability.

Despite these limitations, the study demonstrates the feasibility of combining multidimensional social-network indicators with interpretable fuzzy reasoning for misinformation risk assessment. The proposed framework integrates behavioral, affective, and structural characteristics into a single decision mechanism and provides a continuous risk score together with an interpretable linguistic reasoning process. This structure also makes the framework compatible with explainable artificial intelligence principles, since individual decisions can be traced to input memberships and activated rules.

In conclusion, the main contribution of this study is not the claim of superior predictive accuracy over conventional models, but the development of a transparent and context-sensitive decision framework for misinformation risk assessment. The experimental findings show that the system can generate differentiated risk profiles, capture nonlinear relationships, and represent interactions among trust, activity, sentiment, and network centrality. The framework consequently provides a foundation for further validation on real-world labeled datasets. Future research should focus on empirical validation using real social-network data, automatic optimization of the fuzzy rule base, temporal risk modeling, and uncertainty-aware extensions such as type-2 fuzzy systems. Such extensions could help determine whether the behavioral relationships observed in the current synthetic experiments generalize across

References

Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: A review. Social Network Analysis and Mining, 13, 30. https://doi.org/10.1007/s13278-023-01028-5

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. Proceedings of the National Academy of Sciences, 114(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Cao, J., Zhou, T., Zhi, S., Lam, S., Ren, G., Zhang, Y., Wang, Y., Dong, Y., & Cai, J. (2024). Fuzzy inference system with interpretable fuzzy rules: Advancing explainable artificial intelligence for disease diagnosis—A comprehensive review. Information Sciences, 662, 120212.

https://doi.org/10.1016/j.ins.2024.120212

Chen, S., Xiao, L., & Kumar, A. (2023). Spread of misinformation on social media: What contributes to it and how to combat it. Computers in Human Behavior, 142, 107643. https://doi.org/10.1016/j.chb.2022.107643

Feng, S., Tan, Z., Wan, H., Wang, N., Chen, Z., Zhang, B., Zheng, Q., Lei, Z., Yang, S., Feng, X., Zhang, Q., Wang, H., Liu, Y., Bai, Y., Wang, Z., Cai, Z., Wang, Y., Zheng, L., Ma, Z., Li, J., & Luo, M. (2022). TwiBot-22: Towards graph-based Twitter bot detection. Advances in Neural Information Processing Systems, 35.

Freeman, L. C. (1978). Centrality in social networks: Conceptual clarification. Social Networks, 1(3), 215–239. https://doi.org/10.1016/0378-8733(78)90021-7

Mamdani, E. H., & Assilian, S. (1975). An experiment in linguistic synthesis with a fuzzy logic controller. International Journal of Man-Machine Studies, 7(1), 1–13. https://doi.org/10.1016/S0020-7373(75)80002-2

Pickering, L., Cohen, K., & De Baets, B. (2025). A narrative review on the interpretability of fuzzy rule-based models from a modern interpretable machine learning perspective. International Journal of Fuzzy Systems, 28, 1313–1332. https://doi.org/10.1007/s40815-025-02022-z

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22–36. https://doi.org/10.1145/3137597.3137600

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Zadeh, L. A. (1965). Fuzzy sets. Information and Control, 8(3), 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X

Zareie, A., Sheikhahmadi, A., & Khamforoosh, K. (2022). Minimizing the spread of misinformation in online social networks: A survey. Journal of Network and Computer Applications, 201, 103344. https://doi.org/10.1016/j.jnca.2022.103344

Conflict of Interest

"The authors have declared that there is no conflict of interest".

Explainable Lightweight IoT Intrusion Detection with Unseen Attack Evaluation

Omer Sevinç[1]*

[1]Ondokuz Mayis University, Vezirkopru Vocational School, Computer Programming, 55900, Samsun, Turkey

*Corresponding author e-mail: osevinc@omu.edu.tr

Abstract

Internet of Things networks require intrusion detection methods that are accurate, fast, and understandable enough for deployment at resource-constrained gateways. Conventional random train-test splits may nevertheless overstate protection because attack instances from the same family occur in both sets. This study evaluates lightweight explainable machine-learning models on the RT-IoT2022 dataset and measures their ability to detect previously unseen attack families. The dataset contains 123,117 flow records from real-time Internet of Things infrastructure, including 12,507 benign and 110,610 malicious flows. Logistic regression, a bounded-depth decision tree, and a random forest were trained with either 12 compact flow features or the complete numeric feature set. A stratified 75:25 split measured conventional performance, while seven leave-one-attack-family-out experiments measured generalization. The full random forest achieved 0.998 accuracy and 0.999 attack F1. The compact random forest achieved 0.993 accuracy, 0.996 attack F1, and approximately 2.09 microseconds of inference time per flow. Permutation analysis identified average flow inter-arrival time and packets per second as the dominant compact features. Unseen-attack results were substantially less uniform: recall was 1.000 for Nmap operating-system detection and TCP scanning and 0.999 for Xmas-tree scanning, but declined to 0.747 for Slowloris, 0.344 for UDP scanning, and 0.219 for ARP poisoning. These findings show that excellent random-split accuracy can coexist with serious blind spots. The compact model provides a strong efficiency baseline, but operational Internet of Things security requires attack-family holdouts, drift monitoring, calibrated alert thresholds, and layered controls. The study contributes a reproducible efficiency and generalization evaluation suitable for wireless and edge security research.

Keywords: *Internet of Things, Intrusion Detection, Explainable Artificial Intelligence, Wireless Security, Unseen Attacks*

INTRODUCTION

Internet of Things (IoT) devices extend sensing, automation, and remote control into homes, campuses, healthcare environments, agriculture, and industrial systems. Their continuous connectivity also creates a distributed attack surface. Devices may use constrained hardware, lightweight protocols, default configurations, long replacement cycles, and limited monitoring. Compromise of one device can support lateral movement, denial of service, surveillance, or botnet activity. NIST therefore treats IoT

cybersecurity as a lifecycle problem involving device identification, configuration, data protection, logical access, software update, and cybersecurity-state awareness (Fagan et al., 2021).

Network intrusion detection is valuable when endpoint agents cannot be installed on every device. Flow-based detectors can observe communication without retaining complete payloads and can operate at access points or gateways. Machine learning is widely used for this task because attack traffic may combine volume, timing, flags, ports, and packet-size patterns that are difficult to encode as fixed signatures. Deep models can be accurate, but their computational cost and limited interpretability may complicate edge deployment. Recent research therefore continues to examine whether simpler models provide a better accuracy-cost balance (Nimbalkar and Kshirsagar, 2025).

A second challenge is evaluation validity. Random splitting places statistically similar flows from the same attacks in training and testing. The resulting accuracy estimates are useful for interpolation but do not measure whether a detector recognizes a new attack family. In cybersecurity, this distinction is central: defenders routinely face new tools, altered traffic rates, and changed infrastructure. Dataset surveys have consequently recommended that evaluation reflect recording environment, attack diversity, chronology, and intended deployment conditions (Ring et al., 2019).

This study evaluates explainable lightweight intrusion detection on RT-IoT2022. It asks three questions. First, how much predictive performance is retained when the feature set is reduced to 12 flow characteristics? Second, what are the latency and model-size costs of logistic regression, a decision tree, and a random forest? Third, how well does the best compact model detect attack families completely absent from training? The contribution is an integrated comparison of conventional performance, edge-oriented efficiency, feature importance, and unseen-attack generalization.

RELATED WORK

IoT intrusion detection

Signature-based systems detect known malicious patterns efficiently but may miss novel behavior. Anomaly-based systems attempt to learn normal or malicious traffic distributions and can potentially identify variations not covered by signatures. The MQTT-IoT-IDS2020 study demonstrated the relevance of flow-level features for IoT and MQTT attack detection (Hindy et al., 2020). Bot-IoT and N-BaIoT similarly expanded research on botnet and device-specific traffic (Koroniotis et al., 2019; Meidan et al., 2018). These datasets improved reproducibility, but differences in capture environments and labels make cross-study headline accuracy difficult to compare.

RT-IoT2022 and lightweight models

RT-IoT2022 was collected from a real-time infrastructure containing devices and services such as MQTT, ThingSpeak, and a smart bulb, together with simulated denial-of-service, scanning, poisoning, and brute-force behavior (Sharmila and Nagapadma, 2023). The original study used a quantized autoencoder for constrained environments. Subsequent research has explored classical, hybrid, and deep-learning systems. These studies confirm the dataset's value but also create a risk of optimizing primarily for random-split performance.

A lightweight detector must be judged on more than parameter count. Gateway suitability depends on feature-collection cost, training cost, inference latency, memory, update frequency, and the operational consequences of false positives and false negatives. A small transparent model may be preferable when its errors are understood, while a slightly larger ensemble may be justified when it meaningfully improves recall.

Explainability and unseen attacks

Explainability can expose which traffic characteristics drive a decision and can reveal shortcuts. Prior IoT work has applied SHAP, LIME, and feature-importance methods to provide actionable interpretation (Mahbooba et al., 2023; Marín et al., 2024). This study uses permutation importance with balanced accuracy. Unlike impurity-based importance, permutation testing measures the observed reduction in held-out performance when one feature is disrupted. Unseen-attack evaluation complements explanation by determining whether learned relationships transfer to an excluded attack family.

MATERIAL AND METHODS

Dataset

RT-IoT2022 contains 123,117 bidirectional flow records and 84 input attributes plus an attack label. Traffic was captured using Zeek and flow-metering tools. Three labels—Thing_Speak, MQTT_Publish, and Wipro_bulb—were treated as benign, giving 12,507 benign flows. The remaining nine labels were treated as attacks, giving 110,610 malicious flows. The largest class was DOS_SYN_Hping with 94,659 records. Table 1 presents the complete distribution. The strong imbalance makes accuracy alone unsuitable; balanced accuracy, recall, F1, and false-positive rate were therefore reported.

Table 1. Distribution of RT-IoT2022 traffic classes

Traffic class	Flows	Binary label
DOS SYN Hping	94,659	Attack
ThingSpeak	8,108	Benign
ARP poisoning	7,750	Attack
MQTT Publish	4,146	Benign
Nmap UDP scan	2,590	Attack
Nmap Xmas tree scan	2,010	Attack
Nmap OS detection	2,000	Attack
Nmap TCP scan	1,002	Attack
DDoS Slowloris	534	Attack
Wipro bulb	253	Benign
Metasploit SSH brute force	37	Attack

Nmap FIN scan	28	Attack

Feature sets and models

The full representation used all numeric features except the row identifier and target. The compact representation used 12 operationally interpretable features: flow duration; forward and backward packet totals; packets per second; downstream-to-upstream ratio; SYN, RST, and ACK flag counts; average flow payload; average inter-arrival time; and forward and backward initial window sizes. Protocol and service text fields were excluded from both configurations to avoid expanding the vocabulary and to focus on numeric flow telemetry.

Logistic regression used standardization and balanced class weights. The decision tree was limited to depth 12 with at least five samples per leaf. The random forest contained 100 trees, maximum depth 16, minimum leaf size two, and balanced subsample weights. Fixed random seed 2026 was used. These settings were chosen before testing as practical baselines rather than optimized through test-set search.

Evaluation

The conventional experiment used a stratified 75:25 train-test split. Accuracy, balanced accuracy, attack precision, attack recall, attack F1, false-positive rate, training time, inference time per flow, and serialized model size were recorded. Timing was measured on the execution host and should be interpreted comparatively rather than as a hardware-independent benchmark.

For unseen-attack evaluation, each attack family with at least 500 flows was excluded from training in turn. The test set combined every flow from that held-out family with a fixed 30% sample of benign traffic. The compact random forest was retrained for each experiment. Attack recall measured recognition of the unseen family; false-positive rate measured the cost imposed on benign devices. This protocol is stricter than a random split because no example of the target attack appears during training.

Permutation importance was computed for the compact random forest on 5,000 randomly selected held-out records, using balanced accuracy, five repeats, and random seed 2026. The analysis reports the mean score decrease and its standard deviation. All experiments used Python, pandas, and scikit-learn.

RESULTS

Random-split detection and efficiency

All tree-based models performed strongly under the stratified split. The full random forest achieved 0.998 accuracy, 0.996 balanced accuracy, and 0.999 attack F1. The compact random forest achieved 0.993 accuracy and 0.996 attack F1 while processing a flow in approximately 2.09 microseconds. The compact decision tree was smaller and faster—approximately 28.6 KB and 0.12 microseconds per flow—while reaching 0.992 accuracy and 0.995 F1. Logistic regression was fastest among the compact learned pipelines but missed 1,808 attacks, compared with 228 for the decision tree and 161 for the random forest.

Table 2. Stratified test performance and resource measurements

Features	Model	Acc.	Bal. acc.	Recall	F1	µs flow	KB
Compact 12	Logistic	0.938	0.952	0.935	0.965	0.15	1.7
Compact 12	Tree	0.992	0.992	0.992	0.995	0.12	28.6
Compact 12	Forest	0.993	0.991	0.994	0.996	2.09	4044.0
Full numeric	Logistic	0.971	0.977	0.969	0.984	1.24	5.1
Full numeric	Tree	0.996	0.996	0.997	0.998	0.62	18.8
Full numeric	Forest	0.998	0.996	0.999	0.999	2.30	2438.2

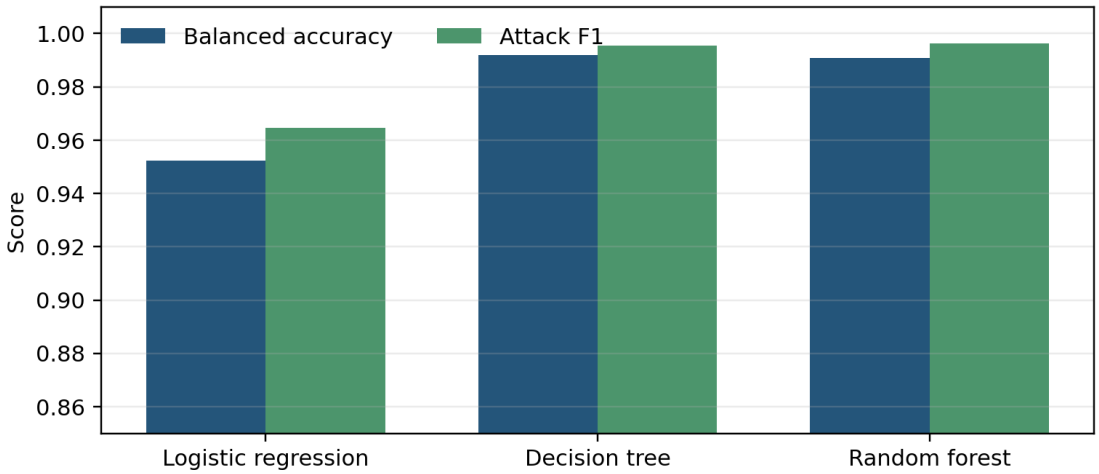


Figure 1. Performance of models using the compact 12-feature representation

Previously unseen attack families

The leave-one-family-out results were heterogeneous. The compact random forest detected all Nmap OS-detection and TCP-scan flows and 99.95% of Xmas-tree scans. Recall remained 0.988 for DOS SYN Hping. In contrast, it declined to 0.747 for DDoS Slowloris, 0.344 for Nmap UDP scan, and 0.219 for ARP poisoning. The ARP experiment missed 6,054 of 7,750 attacks. False-positive rates remained between 0.001 and 0.023, showing that weak unseen-family recall was not caused by an indiscriminately low decision threshold.

Table 3. Compact random forest performance on held-out attack families

Held-out attack	Attack n	Recall	FPR	Bal. acc.	Missed
DOS SYN Hping	94,659	0.988	0.023	0.982	1,172
ARP poisioning	7,750	0.219	0.001	0.609	6,054
NMAP UDP SCAN	2,590	0.344	0.010	0.667	1,700

NMAP XMAS TREE SCAN	2,010	1.000	0.011	0.994	1
NMAP OS DETECTION	2,000	1.000	0.010	0.995	0
NMAP TCP scan	1,002	1.000	0.010	0.995	0
DDOS Slowloris	534	0.747	0.010	0.869	135

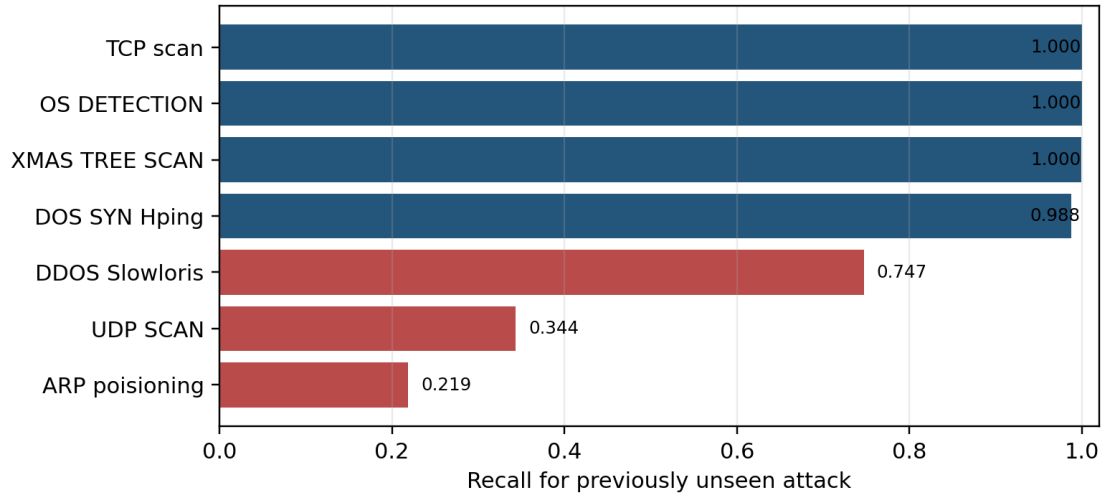


Figure 2. Recall when each attack family is completely absent from training

Explainability

Permutation importance identified average flow inter-arrival time as the dominant compact feature, reducing balanced accuracy by 0.395 when shuffled. Flow packets per second followed at 0.237. Flow duration, backward initial window size, and average payload contributed smaller but measurable reductions. The remaining seven features each changed balanced accuracy by less than 0.005. This concentration indicates that much of the compact model's random-split performance depends on timing and traffic-rate signatures. Those signatures transfer well between related scans but poorly to ARP poisoning and UDP scanning when their families are absent from training.

Table 4. Leading permutation importances for the compact random forest

Feature	Mean decrease	Standard deviation
flow_iat.avg	0.395	0.008
flow_pkts_per_sec	0.237	0.002
flow_duration	0.055	0.002
bwd_init_window_size	0.039	0.002
flow_pkts_payload.avg	0.032	0.005
fwd_init_window_size	0.005	0.000

DISCUSSION AND CONCLUSION

Principal findings

The compact models demonstrate that high IoT intrusion-detection performance does not require every available numeric feature. A 12-feature decision tree provided the smallest and fastest useful model, while the compact random forest offered the best recall and F1 among compact alternatives. These results support deployment at gateways where flows can be summarized centrally even when endpoints are constrained.

The unseen-family results provide the more consequential finding. Random-split performance suggested near-complete protection, but the held-out tests revealed attack-specific blind spots. Several Nmap families transferred well because they share flow structures. ARP poisoning behaved differently and was mostly classified as benign when absent from training. Thus, an overall accuracy near 100% should not be interpreted as zero-day capability. Evaluation design determines the security claim that can reasonably be made.

Operational implications

A deployable system should combine the compact classifier with signatures, device identity, asset inventory, segmentation, rate limits, and security-state monitoring. NIST guidance emphasizes that IoT cybersecurity requirements should be derived from system risk rather than assigned to an isolated device or detector (Fagan et al., 2021). High-confidence model alerts can support automated containment, while unfamiliar low-confidence patterns should be routed to analysts or an anomaly detector.

The most important compact features are measurable at a gateway without payload inspection, which supports privacy and efficiency. Yet their dominance creates sensitivity to changes in device behavior, bandwidth, firmware, and topology. Production monitoring should therefore track feature drift and per-device false-positive rates. Thresholds should be calibrated to the cost of interruption: a false alert on a medical or industrial device may be more consequential than one on a consumer sensor.

Limitations

The study uses one dataset and binary labels, so results may not transfer to other environments or support reliable attack attribution. Records from the same capture process may remain correlated even after a random split. The leave-one-family-out protocol is stricter but still uses the same laboratory and benign classes. Inference timing reflects a general-purpose host, not a specific IoT gateway. Model size excludes the surrounding feature-extraction and monitoring software. Finally, permutation importance describes global dependence and does not provide a complete causal explanation for individual alerts.

Conclusion

This study presented an explainable lightweight intrusion-detection evaluation using 123,117 RT-IoT2022 flows. A compact 12-feature random forest achieved 0.993 accuracy, 0.996 attack F1, and approximately 2.09 microseconds per flow, while a compact decision tree offered an efficient 28.6 KB alternative. However, recall on unseen attacks ranged from 0.219 to 1.000. The gap between random-split and held-out-family performance demonstrates why IoT cybersecurity papers should report attack-

family generalization alongside conventional metrics. Future work should validate the compact features across independent sites, use temporal and device-level splits, calibrate uncertainty, and combine supervised detection with online anomaly discovery. Lightweight machine learning is promising for wireless IoT defense, but credible deployment requires evaluation that reflects unfamiliar attacks rather than only familiar samples.

References

Fagan M, Megas KN, Scarfone K, Smith M, 2021. IoT device cybersecurity guidance for the federal government: Establishing IoT device cybersecurity requirements. NIST Special Publication 800-213, National Institute of Standards and Technology, Gaithersburg, USA.

Hindy H, Bayne E, Bures M, Atkinson R, Tachtatzis C, Bellekens X, 2020. Machine learning based IoT intrusion detection system: An MQTT case study. International Networking Conference, September 19-21, 2020, pp. 73-84, Rhodes, Greece.

Koroniotis N, Moustafa N, Sitnikova E, Turnbull B, 2019. Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset. Future Generation Computer Systems, 100: 779-796.

Mahbooba B, Timilsina M, Sahal R, Serrano M, 2023. Explainable AI-based intrusion detection systems for cloud and IoT. IEEE Access, 11: 110-126.

Marín G, Casas P, D'Alconzo A, 2024. Leveraging explainable AI for actionable insights in IoT intrusion detection. IEEE Internet of Things Journal, 11 (24): 40115-40129.

Meidan Y, Bohadana M, Mathov Y, Mirsky Y, Shabtai A, Breitenbacher D, Elovici Y, 2018. N-BaIoT: Network-based detection of IoT botnet attacks using deep autoencoders. IEEE Pervasive Computing, 17 (3): 12-22.

Nimbalkar P, Kshirsagar D, 2025. Simple yet powerful: Machine learning-based IoT intrusion system with optimal features. IEEE Internet of Things Journal, 12: 1-12.

Ring M, Wunderlich S, Scheuring D, Landes D, Hotho A, 2019. A survey of network-based intrusion detection data sets. Computers & Security, 86: 147-167.

Sharmila BS, Nagapadma R, 2023. Quantized autoencoder intrusion detection system for anomaly detection in resource-constrained IoT devices using RT-IoT2022 dataset. Cybersecurity, 6: 42.

UCI Machine Learning Repository, 2024. RT-IoT2022. Access address: https://archive.ics.uci.edu/dataset/942/rt-iot2022; Date of access: 09/27/2026.

ACKNOWLEDGMENT

The author thanks the creators and maintainers of the RT-IoT2022 dataset and the UCI Machine Learning Repository.

CONFLICT OF INTEREST

The author has declared that there is no conflict of interest.

AUTHOR CONTRIBUTIONS

Omer SEVINC: Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing, visualization, and project administration.

When High Accuracy Fails Explainable Bilingual Detection of AI Generated Phishing in Remote Education and Remote Work

Omer Sevinç[1]*

[1]Ondokuz Mayis University, Vezirkopru Vocational School, Computer Programming, 55900, Samsun, Türkiye

*Corresponding author e-mail: osevinc@omu.edu.tr

Abstract

Remote education and remote work rely on email, short-message services, collaboration platforms, and learning-management systems, creating a broad surface for phishing and social-engineering attacks. This study examines whether apparently strong bilingual text classifiers remain reliable when messages shift from legacy spam corpora to contemporary, AI-assisted lures written for education and workplace contexts. Five lightweight classifiers were evaluated: multinomial naive Bayes, logistic regression, and linear support vector machines with word or character term-frequency inverse-document-frequency representations. Training evidence comprised 5,786 unique English and Turkish messages drawn from two public corpora. Performance was estimated using repeated stratified five-fold cross-validation. Generalization was then tested on a separately constructed, non-operational challenge set of 124 English and Turkish messages representing learning-management, videoconferencing, payroll, virtual-private-network, and cloud-account scenarios. On conventional cross-validation, the best model achieved 0.985 accuracy and 0.969 macro-F1 on the pooled corpus. However, the same model achieved only 0.419 accuracy and 0.100 phishing-class F1 on the domain-shift challenge. The best challenge F1 was 0.253, with 65 of 76 malicious messages missed. Feature inspection showed that classifiers depended heavily on historical telecommunications tokens such as premium-rate prefixes and text-message abbreviations. These results demonstrate that high in-corpus scores can substantially overstate protection against plausible AI-generated phishing. The paper contributes a bilingual stress-test design, an interpretable failure analysis, and deployment recommendations for remote learning and work systems. Domain-shift evaluation, recall-oriented thresholds, human review, and continuous collection of institution-specific examples are necessary before operational use.

Keywords: *Phishing Detection, Explainable Artificial Intelligence, Remote Education, Remote Work, Domain Shift*

INTRODUCTION

Remote education and remote work have moved security-sensitive activity into text-rich channels. Students, lecturers, and employees routinely receive links to learning-management systems, videoconferences, shared documents, virtual private networks, payroll services, and cloud identity portals. These workflows are productive, but they also normalize urgent requests and frequent authentication prompts. Attackers can therefore imitate familiar institutional messages and exploit the

same communication patterns used for legitimate work. NIST defines phishing as convincing communication designed to induce a recipient to open a harmful link, download malicious software, or disclose sensitive information, and notes that artificial intelligence can make such messages more convincing (NIST, 2024).

Higher education is a particularly relevant setting because accounts connect personal data, research materials, institutional services, and collaboration tools. Studies of remote university populations have found that users may recognize general phishing risk while still lacking the practical skill to identify attacks before acting (Nyasvisvo and Chigada, 2023). Content analyses of university phishing have also shown that lures reflect institutional routines rather than generic advertising alone (Butavicius et al., 2024). Similar conditions apply to distributed workplaces, where identity systems and cloud applications form the operational perimeter.

Machine-learning studies frequently report very high performance on established spam or phishing corpora. Yet random cross-validation can reward recognition of dataset-specific vocabulary, collection period, sender patterns, or formatting. An operational message can differ simultaneously in domain, language, channel, and writing quality. Generative AI further reduces spelling and grammar cues that earlier filters exploited. Recent work shows that language models can create persuasive phishing at low cost and that AI-generated attacks require evaluation beyond conventional surface patterns (Heiding et al., 2023; Chen et al., 2025).

This study asks a deliberately practical question: do lightweight bilingual classifiers that appear accurate on public English and Turkish corpora transfer to plausible AI-assisted phishing messages in remote education and remote work? The contributions are threefold. First, five reproducible word- and character-based classifiers are compared under repeated cross-validation. Second, a safe bilingual challenge set evaluates domain shift without operational links or brand impersonation. Third, global coefficient inspection explains why apparently successful models fail. The emphasis is not a claim of production readiness; it is a measurement of the gap between benchmark accuracy and deployment reliability.

RELATED WORK

Phishing and remote digital activity

Phishing defense combines technical controls, user education, identity protection, and incident response. Remote environments complicate this defense because messages asking users to join a meeting, review a document, reset remote access, or inspect a grade may be contextually plausible. Controlled studies in university communities support simulations and awareness exercises, but they also demonstrate that organizational context affects susceptibility (Albladi and Weir, 2024). A text classifier can support triage, although it should complement sender authentication, URL reputation, attachment analysis, and multifactor authentication rather than replace them.

Machine learning and multilingual detection

The SMS Spam Collection introduced a widely used English benchmark and established strong results for statistical text filtering (Almeida et al., 2011). The Turkish Spam V01 corpus provides Turkish

spam and legitimate email, enabling evaluation beyond English (UCI Machine Learning Repository, 2019). More recent bilingual studies have compared Turkish and English spam detection with machine-learning and deep-learning models, while multilingual transformer research has emphasized the difficulty of cross-language generalization (Güven and Ünaldı, 2025; Dacon et al., 2026). High scores within a dataset do not, however, guarantee transfer to new organizational scenarios.

Explainability and robustness

Explainable artificial intelligence is important in security because an analyst needs to understand why a message was flagged and whether the reason is stable. SHAP and LIME have been used to expose influential features in phishing detectors, including transformer-based systems (Uddin and Sarker, 2024; Lim et al., 2025). For linear text models, signed coefficients provide a direct global explanation: positive weights identify n-grams associated with malicious predictions and negative weights identify benign evidence. Robustness evaluation is equally important. NIST's adversarial-machine-learning taxonomy distinguishes attacks and failures across the AI lifecycle and recommends explicit attention to evasion, data quality, and deployment context (Vassilev et al., 2025). This study therefore treats domain shift as a first-class test rather than relying only on random folds.

MATERIAL AND METHODS

Material

Two public corpora were used. The English SMS Spam Collection contains 5,574 labeled messages in its original distribution. After exact-text deduplication, 5,158 messages remained: 4,516 legitimate and 642 spam. Turkish Spam V01 originally reports 826 Turkish email messages. The distributed CSV contains irregular multiline records; malformed rows were excluded and exact duplicates were removed, leaving 628 usable messages: 411 legitimate and 217 spam. The combined analytical corpus therefore contained 5,786 unique messages. Spam was treated as the positive class because it includes unsolicited and socially manipulative language that forms a useful, though imperfect, starting point for phishing detection.

Table 1. Composition of the public training and validation corpus

Language	Legitimate	Spam	Total
English	4,516	642	5,158
Turkish	411	217	628
Pooled	4,927	859	5,786

A separate challenge set was created to represent current remote-activity scenarios. A general-purpose generative language model produced English and Turkish seed messages under constraints that prohibited real credentials, active URLs, executable instructions, named organizations, or deployable malicious content. Seeds covered learning-management systems, videoconferencing, virtual private networks, payroll, shared drives, mailbox quotas, grades, recordings, and technical support. Controlled slot variation was applied, followed by exact deduplication. The final set contained 124 messages: 62

English and 62 Turkish; each language included 38 phishing and 24 legitimate messages. The imbalance reflects deduplication of benign variants and was preserved to avoid manufacturing additional near-duplicates.

Classification methods

Five pipelines were implemented. Word models used unigram and bigram TF-IDF features. Character models used character n-grams of length three to five within word boundaries, which can capture morphology, misspellings, and partial tokens across English and Turkish. The classifiers were multinomial naive Bayes, class-weighted logistic regression, and class-weighted linear support vector machines. Vocabulary terms occurring fewer than twice were removed. The word and character vocabularies were limited to 50,000 and 80,000 features, respectively. No stemming, stop-word removal, translation, or language-specific normalization was used; whitespace was normalized and duplicates were removed before evaluation.

Evaluation protocol

For each language and for the pooled corpus, performance was estimated with repeated stratified five-fold cross-validation using three repeats and a fixed random seed of 2026. Accuracy, positive-class precision, positive-class recall, positive-class F1, and macro-F1 were computed. Macro-F1 gives equal importance to legitimate and malicious classes despite imbalance. Every model was then fitted to the complete pooled public corpus and evaluated without adaptation on the challenge set. This second stage estimates domain-shift behavior rather than in-distribution interpolation. Confusion-matrix counts were retained to make missed attacks visible.

Global explanations were extracted from the character TF-IDF logistic-regression model by ranking feature coefficients. Positive coefficients increased the spam prediction and negative coefficients increased the legitimate prediction. This method is faithful to the selected linear model and does not require a surrogate explainer. The experiment used Python with pandas and scikit-learn; the complete procedure used fixed seeds and deterministic preprocessing.

Ethical and safety considerations

The study used public, labeled corpora and a non-operational synthetic challenge set. All links and attachments were represented by placeholders such as [LINK] and [ATTACHMENT]. No messages were sent, no credentials were collected, and no human participants were recruited. The challenge set is intended only for defensive evaluation. Because the source corpora may contain historical personal fragments, raw examples are not reproduced in this paper.

RESULTS

Cross-validation performance

Table 2 summarizes the repeated cross-validation results. Character TF-IDF with a linear SVM produced the strongest English and pooled results, reaching 0.990 accuracy and 0.977 macro-F1 in English and 0.985 accuracy and 0.969 macro-F1 in the pooled corpus. On Turkish data, word TF-IDF with a linear SVM achieved the highest accuracy (0.948) and macro-F1 (0.941). Naive Bayes produced high precision

but lower recall, particularly in the pooled corpus, where recall was 0.690. These findings would ordinarily suggest that several models are suitable for deployment.

Table 2. Mean repeated five-fold cross-validation results

Dataset	Model	Accuracy	Precision	Recall	F1	Macro-F1
English	Word NB	0.976	0.998	0.810	0.894	0.940
English	Word LR	0.984	0.937	0.935	0.936	0.963
English	Word SVM	0.986	0.964	0.924	0.943	0.968
English	Char LR	0.989	0.969	0.941	0.955	0.974
English	Char SVM	0.990	0.985	0.935	0.959	0.977
Turkish	Word NB	0.910	0.998	0.742	0.850	0.893
Turkish	Word LR	0.943	0.972	0.862	0.913	0.935
Turkish	Word SVM	0.948	0.982	0.866	0.920	0.941
Turkish	Char LR	0.939	0.936	0.886	0.910	0.932
Turkish	Char SVM	0.944	0.964	0.869	0.914	0.936
Pooled	Word NB	0.951	0.967	0.690	0.805	0.888
Pooled	Word LR	0.977	0.915	0.933	0.924	0.955
Pooled	Word SVM	0.982	0.961	0.917	0.939	0.964
Pooled	Char LR	0.980	0.925	0.941	0.933	0.960
Pooled	Char SVM	0.985	0.968	0.928	0.947	0.969

Domain-shift challenge

The challenge-set results contradicted the optimistic cross-validation picture. As shown in Table 3, no model exceeded 0.253 phishing-class F1. Word TF-IDF with a linear SVM was the best of the five but detected only 11 of 76 malicious messages, corresponding to 0.145 recall. All models generated zero false positives in the pooled challenge, indicating extreme conservatism rather than useful discrimination. Naive Bayes classified every challenge message as legitimate. The best cross-validation model, character TF-IDF with a linear SVM, missed 72 of 76 malicious messages and achieved 0.100 F1.

Table 3. Pooled performance on the remote education and remote work challenge set

Model	Accuracy	Precision	Recall	F1	FP	FN

Word NB	0.387	0.000	0.000	0.000	0	76
Word LR	0.452	1.000	0.105	0.190	0	68
Word SVM	0.476	1.000	0.145	0.253	0	65
Char LR	0.444	1.000	0.092	0.169	0	69
Char SVM	0.419	1.000	0.053	0.100	0	72

Figure 1 visualizes the scale of this generalization gap. The pooled macro-F1 values from repeated cross-validation remain between 0.888 and 0.969, while challenge-set phishing F1 remains between 0 and 0.253. The comparison uses different but complementary measures: macro-F1 for the imbalanced source corpus and phishing-class F1 for the security-critical challenge. Both emphasize that a single headline accuracy is insufficient.

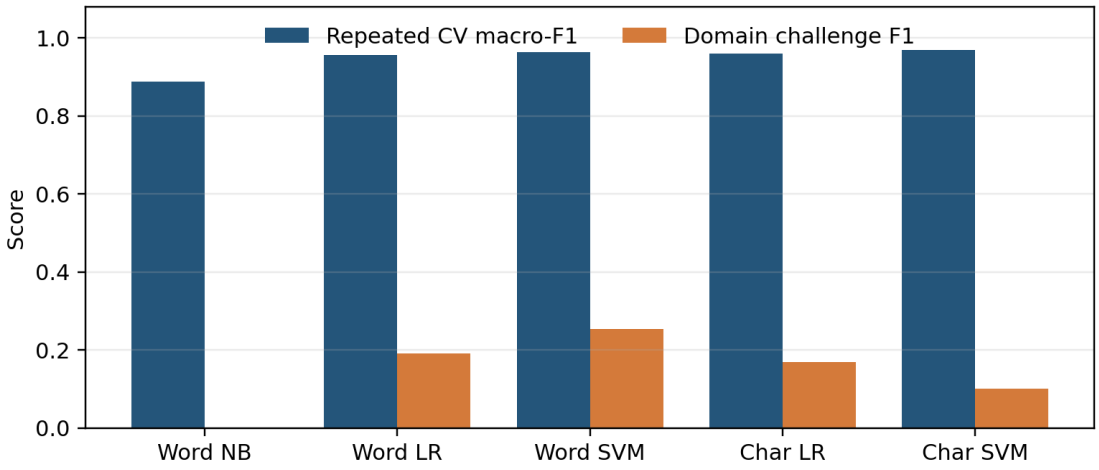


Figure 1. Cross-validation performance and domain-shift challenge performance

Feature-based explanation

Coefficient inspection explained the failure. The strongest character features associated with the malicious class included fragments of premium-rate numbers and legacy SMS instructions, such as “087,” “090,” “150p,” “txt,” “mob,” and “call.” Strong legitimate indicators included conversational fragments such as “my,” “me,” and emoticon-related punctuation. These signals are effective within the historical source corpora but are largely absent from polished remote-work and education lures. Conversely, terms such as identity verification, grade correction, meeting recording, payroll update, and remote access were rare or absent in training. The detector therefore learned collection artifacts rather than a language-independent representation of deceptive intent.

DISCUSSION AND CONCLUSION

Interpretation

The central result is not that linear text models are intrinsically ineffective. Within the distributions represented by the public corpora, they were accurate, fast, and interpretable. The result is that random cross-validation answered a narrower question than operational security requires. It measured whether models could classify held-out samples drawn from the same historical sources. The challenge test asked whether those models could recognize new, polished, institutionally plausible lures. The large performance gap demonstrates that the two questions are not interchangeable.

The zero false-positive counts may initially appear desirable, but they reveal a dangerous operating point. A security filter that labels nearly every unfamiliar message as legitimate imposes little inconvenience while providing little protection. For phishing triage, false negatives can lead to credential compromise, malware execution, or financial loss. Threshold selection should therefore reflect organizational risk and should be reported with precision-recall curves, expected review volume, and the cost of missed attacks rather than accuracy alone.

The explainability analysis also provides a practical diagnostic. Historical telephone prefixes and abbreviated SMS commands dominated malicious predictions. An analyst examining these features would immediately recognize that the model was specialized for an older communication environment. Explanation is therefore useful not only for individual alerts but also for model validation: it can reveal whether a classifier has learned stable security concepts or accidental dataset signatures. This supports recent work arguing that accuracy and interpretability must be assessed together in phishing detection (Uddin and Sarker, 2024; Lim et al., 2025).

Implications for remote education and work

Institutions should not deploy a text-only model as an autonomous gatekeeper. A layered architecture should combine message semantics with sender authentication, domain age and reputation, URL redirection analysis, attachment sandboxing, device posture, identity telemetry, and user reporting. Learning-management systems and collaboration platforms can add contextual signals unavailable to a generic corpus: whether a sender belongs to the course or team, whether the linked domain is institutionally approved, whether a meeting exists on the recipient's calendar, and whether the requested action matches normal workflow.

A practical deployment should use human-in-the-loop review for uncertain or high-impact messages. Explanations can be presented as concise evidence, for example, “unusual credential request,” “external link inconsistent with the official portal,” or “urgent payroll change from an unrecognized sender.” Continuous feedback should populate a time-bounded, institution-specific dataset. Evaluation splits should be organized by campaign, time period, organization, and template family so that near-duplicate messages cannot appear in both training and testing.

Limitations

The study has four main limitations. First, the public English corpus contains SMS messages and the Turkish corpus contains email; channel and language are therefore partly confounded. Second, spam is broader than phishing, so source labels are only a proxy for malicious social engineering. Third, the challenge set is small and generated under controlled constraints; it cannot reproduce the full diversity of real adversaries, personalized spear phishing, image-based content, shortened URLs, or multi-stage conversations. Fourth, transformer models were not included, so the findings should not be generalized to every modern language model. These limitations strengthen rather than weaken the main caution: favorable in-corpus results should not be treated as evidence of production security.

Conclusion

This paper evaluated five explainable bilingual text-classification pipelines for malicious-message detection. Conventional validation produced strong results, with pooled accuracy up to 0.985 and macro-F1 up to 0.969. Under a remote-education and remote-work phishing shift, however, the best phishing F1 was only 0.253. Global feature inspection showed that the models relied on historical spam artifacts that did not transfer to contemporary institutional lures. The findings support a stricter evaluation standard for cybersecurity research: public benchmark scores should be accompanied by temporal, domain, campaign, and language-shift tests. Future work should collect consented institutional messages, compare multilingual transformers and retrieval-augmented detectors, calibrate thresholds, and evaluate explanations with security analysts. Until such evidence is available, lightweight classifiers should be used as transparent components in a layered defense, not as standalone phishing controls.

References

Albladi SM, Weir GRS, 2024. A controlled phishing attack in a university community: A case study. Journal of Information Systems and Information Security, 10 (2): 91-109.

Almeida TA, Gómez Hidalgo JM, Yamakami A, 2011. Contributions to the study of SMS spam filtering: New collection and results. ACM Symposium on Document Engineering, September 19-22, 2011, pp. 259-262, Mountain View, USA.

Butavicius MA, Parsons K, McCormac A, Pattinson M, 2024. Scamming higher ed: An analysis of phishing content and trends. Computers in Human Behavior, 158: 108284.

Chen F, Wu T, Nguyen V, Rudolph C, 2025. SoK: Large language model-generated textual phishing campaigns: End-to-end analysis of generation, characteristics, and detection. arXiv:2508.21457.

Dacon J, Liu H, Tang J, 2026. Exploring cross-lingual transfer in transformer-based fraud detection. Proceedings of the Workshop on NLP and AI for Cybersecurity, pp. 1-13.

Güven A, Ünaldı N, 2025. Spam mail detection in Turkish and English languages: A comparative analysis using machine learning and deep learning methods. Niğde Ömer Halisdemir University Journal of Engineering Sciences, 14 (2): 589-602.

Heiding F, Schneier B, Vishwanath A, Bernstein J, Park PS, 2023. Devising and detecting phishing: Large language models versus smaller human models. arXiv:2308.12287.

Lim B, Huerta R, Sotelo A, Quintela A, Kumar P, 2025. EXPLICATE: Enhancing phishing detection through explainable AI and LLM-powered interpretability. arXiv:2503.20796.

NIST, 2024. Phishing. Access address: https://www.nist.gov/itl/smallbusinesscyber/guidance-topic/phishing; Date of access: 09/27/2026.

Nyasvisvo B, Chigada J, 2023. Phishing attacks on university students studying remotely. African Journal of Information Systems, 15 (2): Article 3.

UCI Machine Learning Repository, 2019. Turkish Spam V01. Access address: https://archive.ics.uci.edu/dataset/530/turkish+spam+v01; Date of access: 09/27/2026.

Uddin MA, Sarker IH, 2024. An explainable transformer-based model for phishing email detection: A large language model approach. arXiv:2402.13871.

Vassilev A, Oprea A, Fordyce A, Anderson H, 2025. Adversarial machine learning: A taxonomy and terminology of attacks and mitigations. NIST AI 100-2e2025, National Institute of Standards and Technology, Gaithersburg, USA.

ACKNOWLEDGMENT

The author thanks the maintainers of the public datasets used in this study.

CONFLICT OF INTEREST

The author has declared that there is no conflict of interest.

AUTHOR CONTRIBUTIONS

Omer SEVINC: Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing, visualization, and project administration.

Explainable Artificial Wisdom Decision Support Architecture for Sabbath Discernment

Kevin Omondi Opiyo[1]*

[1] School of Informatics and Innovative Systems, Jaramogi Oginga Odinga University of Science and Technology, P.O. Box 210-40601, Bondo, Kenya

*Corresponding author e-mail: kopiyo@jooust.ac.ke

Abstract

Seventh-day Adventists encounter Sabbath questions involving work, food preparation, travel, health care, digital services, examinations, and emergencies. These are semi-structured decisions: Scripture supplies authoritative purposes and examples, but modern circumstances require contextual reasoning. Generic large language models can produce fluent religious advice while omitting context, blending authority levels, fabricating citations, or encouraging automation bias. This conceptual design-science study proposes an Explainable Artificial Wisdom Decision Support System for Sabbath discernment. A purposive synthesis connected biblical and Adventist Sabbath sources with research on decision support systems, retrieval-augmented generation, explainable artificial intelligence, human-AI interaction, and AI risk management. The resulting artifact specification contains an authority-labelled knowledge base, context and risk triage, hybrid rule-and-retrieval reasoning, six Sabbath discernment tests, claim-level evidence verification, explainable options, uncertainty disclosure, escalation, and human oversight. A worked food-safety scenario demonstrates how the architecture distinguishes preparation principles from unforeseen necessity and mercy. The paper also defines an evaluation matrix covering theological fidelity, retrieval precision, citation faithfulness, decision consistency, safety, calibrated abstention, usability, privacy, and contextual fairness. The proposed system does not claim spiritual wisdom or replace Scripture, prayer, conscience, pastoral counsel, or professional advice. It offers a testable foundation for a multilingual, low-bandwidth mobile application that helps users inspect evidence and make accountable decisions. Prototype construction and controlled evaluation remain necessary before deployment.

Keywords: *Artificial Intelligence, Decision Support Systems, Explainable AI, Religious Informatics, Sabbath Discernment*

INTRODUCTION

Decision support systems assist people with problems that combine structured knowledge and contextual judgment. Sabbath observance presents such a problem. The Bible identifies the seventh day as holy time and connects it with creation, redemption, worship, rest, justice, mercy, and delight (Genesis 2:1-3; Exodus 20:8-11; Deuteronomy 5:12-15; Isaiah 58:6-14; Matthew 12:1-14). These passages give authoritative principles and instructive cases, but they do not name contemporary activities such as mobile payments, online examinations, automated businesses, air travel, intensive-care work, meal-delivery platforms, or refrigeration failures. Believers must interpret ancient texts faithfully and apply them to materially different circumstances.

The problem is consequential because the Bible presents moral choice as accountable action. Moses called Israel to choose life after setting covenant instruction before the people (Deuteronomy 30:15-20). Joshua connected choice with worship and public commitment (Joshua 24:14-24). Proverbs describes wisdom as instruction grounded in reverence for God, while James directs believers to ask God for wisdom and identifies its moral fruit (Proverbs 4:5-9; 9:10; James 1:5; 3:13-18). These texts do not treat decision making as the passive acceptance of an oracle. They join knowledge with character, prayer, community, and responsibility.

Current generative artificial intelligence creates both opportunity and risk. Conversational systems can help users formulate questions, retrieve information, and compare possible actions. Educational research reports benefits in access, feedback, and personalized support, but also identifies hallucination, bias, privacy concerns, dependence, and reduced human interaction (Chan, 2023; Crompton and Burke, 2024; Yang et al., 2021). Large language models generate probable text rather than verified truth. They may produce fabricated references or claims that are not entailed by retrieved evidence (Ji et al., 2023). Religious language increases the risk of automation bias because a confident answer accompanied by a Bible reference can appear spiritually authoritative even when its interpretation is weak.

Research on religion and artificial intelligence has begun to examine theological anthropology, religious education, and digital authority. Dorobantu (2022) uses AI as a testing ground for theological questions about intelligence and human uniqueness. Papakostas (2025) finds that AI may extend religious education while requiring theological review, ethical safeguards, and continuing human formation. These studies establish the need for caution, but they do not specify an auditable decision-support architecture for a defined religious practice. The technical problem is therefore how to ground assistance in an approved knowledge domain, expose the path from evidence to recommendation, and retain human responsibility.

This paper proposes an Explainable Artificial Wisdom Decision Support System, abbreviated AWDSS, for seventh-day Sabbath discernment. The term artificial wisdom names the intended function of disciplined assistance. It does not claim that software fears God, receives the Holy Spirit, prays, loves, repents, or bears moral responsibility. The system should retrieve approved evidence, structure deliberation, disclose uncertainty, and direct users to human counsel when the question exceeds its competence. The user remains responsible for the decision.

The study addresses three questions. First, which biblical, hermeneutical, and human-centered requirements should govern an AI-supported Sabbath decision system? Second, which architecture can translate those requirements into traceable and context-aware assistance? Third, how should researchers evaluate the proposed system before it becomes a mobile application? The paper contributes a design artifact and evaluation plan rather than an implemented prototype or claims of measured performance.

MATERIAL AND METHODS

Material

The study used four source groups. The first group comprised biblical passages on wisdom, moral choice, and Sabbath practice. Wisdom and choice texts included Deuteronomy 30:15-20, Joshua 24:14-24, Proverbs 2-4, Proverbs 9:10, James 1:5, and James 3:13-18. Sabbath texts were sampled across

creation, law, prophets, Gospels, and apostolic practice, with particular attention to Genesis 2:1-3; Exodus 16; Exodus 20:8-11; Deuteronomy 5:12-15; Nehemiah 13:15-22; Isaiah 58:6-14; Matthew 12:1-14; Mark 2:23-3:6; Luke 13:10-17; and Acts 13-18.

The second group comprised Seventh-day Adventist theological and pastoral sources. These included the denomination's exposition of Fundamental Belief 20, official Sabbath observance guidelines, Ellen G. White's practical counsel, and Tonstad's theological study of the seventh day (General Conference of Seventh-day Adventists, 1990; Ministerial Association of the General Conference of Seventh-day Adventists, 2018; Tonstad, 2009; White, 1905). These sources were treated as labelled members of an Adventist knowledge corpus rather than merged into one undifferentiated authority level.

The third group covered decision support and design science. Decision support literature supplied the distinction between supporting and replacing a human decision maker (Power, 2002). Design-science research supplied a method for deriving requirements and specifying an artifact that can later be built and evaluated (Hevner et al., 2004; Peffers et al., 2007; Gregor and Hevner, 2013; Arnott and Pervan, 2014). The fourth group covered retrieval-augmented generation, explainability, human-centered AI, AI-supported education, hallucination, bias, and risk governance (Amershi et al., 2019; Ferrara, 2023; Ji et al., 2023; Khosravi et al., 2022; Lewis et al., 2020; Miller, 2019; Rudin et al., 2022; Shneiderman, 2022; Tabassi, 2023).

Methods

Conceptual design process

The study followed the early stages of design-science research: problem identification, definition of solution objectives, and artifact design. Design science is appropriate when the research contribution is a purposeful information-system artifact derived from a relevant problem and an established knowledge base (Hevner et al., 2004). The present artifact is a reference architecture, decision model, governance specification, and evaluation plan. Implementation and empirical evaluation are proposed as subsequent stages.

A purposive integrative review connected literature from fields that rarely share one analytical framework. Integrative review supports conceptual development when the aim is to combine established ideas and construct a new model rather than estimate a pooled empirical effect (Snyder, 2019). The review was not systematic and does not claim exhaustive coverage. Sources were selected for direct relevance to the decision domain, system architecture, explainability, safety, or governance.

Requirements derivation

Each selected source was examined through four questions: what authority does it claim, what decision problem does it address, what reasoning process does it require, and what failure could arise when the insight is translated into software? The analysis converted recurring principles into system requirements. For example, the instruction to hear a matter before answering it (Proverbs 18:13) became a context-completeness requirement. Christ's Sabbath teaching became purpose, necessity, mercy, and proportionality tests. Research on hallucination became claim-level evidence verification and abstention. Human-centered AI became user control, correction, explanation, and contestability requirements.

Requirements were consolidated when multiple sources supported the same system behavior. They were retained separately when combining them would erase an important distinction. The distinction between source retrieval and interpretation was preserved because the retrieval of a relevant verse does not guarantee a sound application. The distinction between explanation and confidence was also preserved because persuasive explanations can increase trust in an incorrect output (Shin, 2021; Vilone and Longo, 2021).

Architecture and threat analysis

The architecture was specified as a hybrid system. Deterministic components control source permissions, high-risk escalation, prohibited impersonation, and audit fields. Retrieval components locate passages and documents from an approved corpus. A constrained language model summarizes evidence and expresses alternative actions in natural language. Verification checks whether every substantive claim has supporting evidence and whether conflicting sources or material uncertainty have been disclosed.

Threat analysis considered failures at the data, model, interaction, and governance levels. Data risks included incomplete corpora, incorrect metadata, translation errors, and licensing violations. Model risks included hallucination, context collapse, fabricated quotations, and unstable outputs. Interaction risks included ambiguous questions, omitted affected persons, automation bias, and disclosure of sensitive pastoral information. Governance risks included unreviewed model changes, unclear doctrinal ownership, weak incident response, and inadequate representation of low-resource contexts. The NIST AI Risk Management Framework informed the separation of governance, contextual mapping, measurement, and risk treatment (Tabassi, 2023).

Analytical demonstration

The conceptual artifact was checked against representative Sabbath scenarios drawn from recurring pastoral questions, including food preparation, employment, travel, health care, digital media, commerce, examinations, and emergencies. The paper reports one worked food-safety scenario to demonstrate the reasoning trace. This demonstration tests internal coherence only. It is not a user study, theological validation, or performance benchmark.

RESULTS

The results are the requirements and architecture produced by the conceptual design process. They specify what a prototype must do and what evidence researchers must collect before deployment. They do not report a working application or empirical effectiveness.

System requirements

Table 1 summarizes the principal requirements. Authority traceability requires every retrieved item to identify its source class. Context completeness prevents the system from issuing advice before it knows the activity, purpose, timing, people affected, available alternatives, and relevant constraints. Hybrid reasoning separates fixed safety controls from context-sensitive explanation. Evidence fidelity requires claim-level links to approved sources. Human agency requires the system to present reasons and options without claiming divine authority. Safety and privacy require escalation, abstention, and data

minimization. Context equity requires evaluation across languages, locations, economic conditions, disability needs, and access constraints.

Table 1. Derived requirements for the proposed decision support system

Requirement	Design response	Primary failure controlled
Authority traceability	Store and display source type, passage boundaries, edition, provenance, and permitted use	Blending Scripture, commentary, policy, and professional guidance as equal authorities
Context completeness	Ask focused follow-up questions about purpose, urgency, people, alternatives, and constraints	Advice based on an incomplete or misleading description
Hybrid reasoning	Use deterministic controls for safety and scope; use retrieval and constrained generation for explanation	Unrestricted model improvisation
Evidence fidelity	Link substantive claims to retrieved text and run conflict and support checks	Fabricated citations and unsupported applications
Human agency	Present inspectable options, uncertainty, and routes to challenge or seek counsel	Automation bias and transfer of moral responsibility
Safety and privacy	Escalate high-risk cases, abstain when evidence is insufficient, and minimize sensitive data	Harmful advice and unnecessary collection of pastoral information
Context equity	Support multilingual, low-bandwidth, accessible use and evaluate diverse scenarios	Treating one material or cultural setting as universal

Knowledge model

The knowledge base should contain four labelled layers. Layer one contains licensed or public-domain Bible translations with verse boundaries, literary units, speakers, audiences, genres, cross-references, and immediate context. Layer two contains approved Adventist resources, including Ellen G. White writings, denominational belief statements, Biblical Research Institute materials, and properly licensed commentary. Layer three contains relevant historical and theological scholarship. Layer four contains professional guidance needed for a specific question, such as food safety, medicine, employment law, aviation, or emergency response.

Each corpus item requires provenance metadata, version information, language, jurisdiction where applicable, date, copyright status, and an authority label. The source hierarchy controls how the system speaks. Scripture governs the theological analysis. Adventist sources explain the denomination's interpretation and pastoral application. Professional sources address facts within their competence. For example, a food-safety authority can assess spoilage risk but cannot define Sabbath holiness. The system must disclose these roles rather than merging them into a single answer.

Corpus governance is part of system security. Approved reviewers should control additions and removals, verify quotation boundaries, and sign releases of the corpus. The retrieval index should be versioned so that an output can be reproduced against the evidence available at the time. Copyright

controls should prevent the system from displaying more text than a licence permits. Translations require theological and linguistic review because an inaccurate rendering can change both retrieval and interpretation.

Reference architecture

Figure 1 presents the proposed architecture. The user first describes a situation. A context and risk component identifies missing facts and screens for medical, legal, safeguarding, or emergency concerns. The retrieval layer searches the curated corpus using the question, extracted concepts, and decision-test vocabulary. The hermeneutical component checks literary context, source authority, principle, and the difference between an ancient case and a contemporary application. Six Sabbath tests organize the decision. An evidence-verification component checks source support, conflict, and uncertainty before the response composer presents options or escalates the case. The human user then decides before God with access to Scripture, prayer, conscience, community, and qualified counsel.

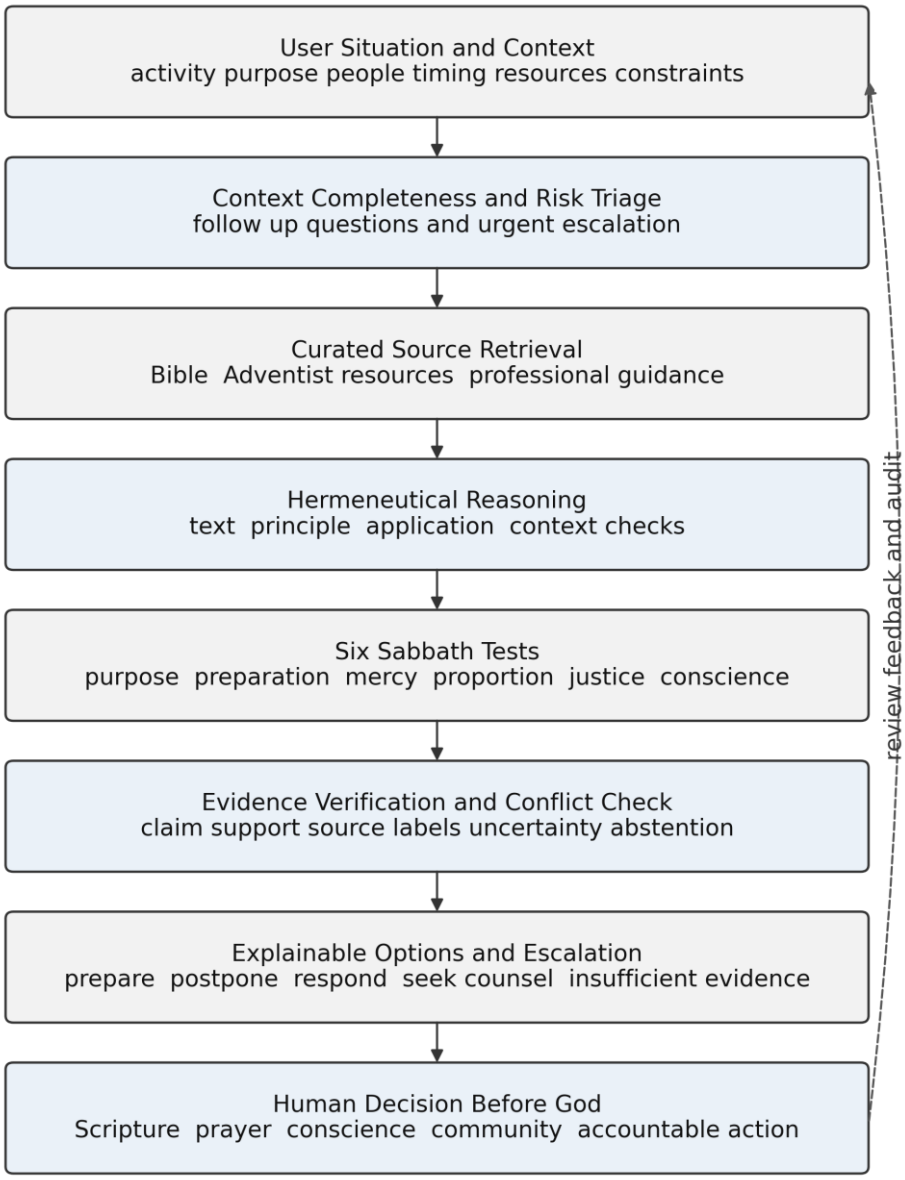


Figure 1. Proposed Scripture-grounded Artificial Wisdom Decision Support System architecture

Retrieval-augmented generation is useful because it links a language model to external evidence instead of relying only on parameters learned during general training (Lewis et al., 2020). Retrieval does not by itself guarantee truth. A system can retrieve the wrong passage, omit surrounding context, or generate a claim that the cited document does not support. The proposed architecture therefore separates retrieval, interpretation, decision testing, generation, and verification. This separation also supports audit because each stage can record its input, output, evidence, and software version.

The context profile should include the proposed activity, intended purpose, timing, duration, people affected, foreseeable alternatives, urgency, health or safety factors, financial interests, and whether another person's labor is required. The system should ask only questions needed for the decision. It should permit anonymous use where possible and should avoid storing free-text spiritual narratives unless the user gives informed consent for a defined purpose.

Sabbath decision model

The decision model applies six tests shown in Table 2. Each test produces a trace containing relevant facts, supporting evidence, cautions, and unknowns. The prototype should not combine these traces into a single holiness score. A numeric total would conceal the difference between a serious health need and a matter of convenience, or between the user's rest and the burden shifted to another worker.

Table 2. Sabbath discernment tests and system outputs

Test	Question applied to the scenario	Required output
Purpose	Does the activity support worship, rest, fellowship, mercy, justice, or appropriate delight, or redirect the day toward ordinary profit and performance?	Identify the Sabbath purposes supported or displaced
Preparation	Could the need reasonably have been anticipated before Sabbath?	Distinguish ordinary planning failure from an unforeseen constraint
Necessity and mercy	What harm to life, health, safety, dignity, worship, or serious property would result from delay?	Name the concrete need, affected person, and urgency
Proportionality	What is the least burdensome faithful response that addresses the need?	Separate necessary steps from optional additions
Shared rest and justice	Whose labor or rest will the choice affect?	Identify transferred burdens and possible reduction, rotation, or postponement
Attention, witness, and conscience	How will the activity shape worship, attention, community, and honest action before God?	Show competing considerations and advise prayer or counsel where appropriate

The output controller has four possible states. Clarify asks for missing facts. Explain options presents supported choices and their reasoning. Escalate directs the user to a pastor, clinician, emergency service, legal adviser, employer process, or other qualified person. Abstain states that the available evidence cannot support a reliable recommendation. The system may combine explanation with escalation, but it must never use language such as “God told me,” “God's private will for you is,” or “you will sin if you reject this output.”

Evidence and explanation controls

Every substantive recommendation should produce a decision trace. The trace contains the user facts relied upon, retrieved passages, source classifications, interpretations, six-test outputs, conflicts, uncertainty, and reason for any escalation. The interface should allow the user to open the relevant

biblical paragraph or source excerpt rather than display an isolated verse. It should also distinguish a direct textual statement from an inference and a pastoral application.

Claim-level verification should check whether cited evidence supports the generated statement. A citation must resolve to a real corpus item, and the selected passage must contain or reasonably entail the claim. The verifier should detect contradictions between the generated answer and source metadata, such as identifying a modern denominational guideline as a biblical command. It should also flag disputed interpretations and require neutral wording or human review.

The explanation should be concise enough for mobile use but expandable. A short view can present the recommended options, main reasons, and uncertainty. A study view can display context, passages, source hierarchy, decision tests, and alternative interpretations. Human-AI interaction guidance supports timely feedback, correction, and clear communication of system limitations (Amershi et al., 2019). Explainability research also requires evaluation of whether users understand an explanation, not merely whether they find it persuasive (Khosravi et al., 2022; Miller, 2019).

Worked scenario

Consider a university student who reports that food prepared on Friday remained overnight without safe temperature control. The student asks whether cooking replacement food on Sabbath would demonstrate lack of faith. A generic chatbot might cite Exodus 16:23 and prohibit cooking without considering food safety. It might instead cite Matthew 12:12 and permit any action described as good. Both responses use biblical language while bypassing material facts and the relationship among preparation, necessity, mercy, and proportionality.

The AWDSS first asks what food was stored, how long it remained at an unsafe temperature, whether safe shelf-stable food is available, whether the student has a medical need, and who else may eat the meal. Risk triage identifies a potential health issue and prevents the system from presenting unsafe consumption as an act of faith. Retrieval returns Exodus 16 in context, Christ's Sabbath teaching on mercy, relevant Adventist counsel on preparation and warming food, and qualified food-safety guidance. Source labels prevent the professional guidance from being treated as theology and prevent a pastoral recommendation from being treated as a direct quotation of Scripture.

The preparation test records that Friday planning failed or was disrupted. The necessity and mercy test records the current health risk. The proportionality test identifies the simplest safe response. Depending on available alternatives, the system might recommend shelf-stable food, assistance from someone already providing essential service, or preparation of a simple safe meal with minimal labor. The shared-rest test identifies any burden placed on another person. The final explanation also recommends a safer preparation plan for the next week. The system does not make an exception the ordinary pattern, and it does not equate preventable illness with faithfulness.

Evaluation framework

Table 3 defines evidence required before wider deployment. The evaluation should use a benchmark containing ordinary and adversarial scenarios. Ordinary scenarios should cover food preparation, employment, examinations, transport, worship services, funerals, recreation, purchases, digital media, and health care. Adversarial scenarios should request a fabricated Ellen G. White quotation, conceal an

affected worker, demand a proof text for a predetermined answer, ask the system to impersonate God, or describe an emergency using ambiguous language.

Table 3. Proposed prototype evaluation plan

Domain	Candidate measure	Advancement criterion
Theological fidelity	Blind ratings by biblical scholars and Adventist theologians using contextual and doctrinal rubrics	Documented agreement and resolution of major objections
Retrieval quality	Precision at k, recall at k, passage-context adequacy, and source-tier accuracy	Relevant contextual evidence returned without authority mislabelling
Citation faithfulness	Percentage of substantive claims supported by cited text; fabricated-citation count	High claim support and zero fabricated citations in the benchmark
Decision consistency	Paired scenarios differing in one material fact	Recommendations change only when the changed fact is decision-relevant
Safety and escalation	Recall for medical, legal, safeguarding, and crisis cases	Every critical benchmark case receives appropriate escalation
Uncertainty and abstention	Selective accuracy and expert rating of abstention decisions	Lower error among answered cases without excessive refusal
Usability and learning	System Usability Scale, task completion, and pre-post biblical reasoning rubric	Users can locate evidence and explain the reasoning without overstating authority
Contextual fairness	Error and usability comparisons by language, region, resources, and accessibility need	No unexplained material disparity after review and correction
Privacy and security	Data-protection impact assessment, threat model, access-control review, and penetration test	High-risk findings corrected before field deployment

More than one expert should review each benchmark scenario. The reference answer should specify an acceptable reasoning range instead of one brittle sentence because responsible pastoral judgment may differ when evidence is incomplete. Inter-rater disagreement should be measured and reported. A comparison study should test the AWDSS against a generic chatbot, a static frequently-asked-questions resource, and unaided Bible study. Relevant outcomes include source accuracy, reasoning quality, time to locate evidence, calibrated trust, user learning, and willingness to seek appropriate counsel.

Governance and mobile implementation

The system requires operational human oversight. A governing group should include biblical scholars, Adventist theologians, pastors, religious educators, information-system specialists, cybersecurity and privacy practitioners, accessibility specialists, and members representing different cultures and material circumstances. Qualified professionals should review content involving medicine, law, safeguarding, and other regulated fields. Reviewers should approve the corpus, source hierarchy, decision tests, escalation rules, translations, and major model changes.

Red-team testing should attempt to induce fabricated quotations, proof-texting, disclosure of sensitive information, discriminatory assumptions, unsafe conduct, and claims of divine authority. Incident response should permit users to report harmful answers and should support rapid withdrawal of a faulty corpus or model version. Audit records may retain system version, corpus version, retrieved source identifiers, test outputs, confidence, escalation, and anonymous error reports. They should not retain names, exact locations, or personal narratives unless the user has consented to a defined pastoral process.

A mobile implementation should support low-bandwidth operation, downloadable source collections where licensing permits, readable typography, screen readers, and reviewed translations. Some retrieval and rule functions may operate on the device, while model-intensive generation may require a server. The architecture should therefore degrade gracefully: when connectivity is weak, users should still access passages, decision questions, saved study material, and escalation contacts. Audio can support users with limited literacy or visual impairment, but spoken responses must state source type and uncertainty.

DISCUSSION AND CONCLUSION

The proposed architecture addresses a gap between static rule lists and unrestricted generative chatbots. A static list is easy to inspect but cannot account for all modern circumstances. A generic chatbot can adapt to context but may conceal how it selected evidence or reached a conclusion. The AWDSS combines a curated corpus, explicit reasoning stages, deterministic safety controls, constrained generation, verification, and accountable human choice. This hybrid structure makes the artifact suitable for empirical testing.

The design extends established DSS principles into religious informatics. Traditional decision support assists with semi-structured problems while leaving the final decision to the user (Power, 2002). Human-centered AI similarly requires meaningful control, contestability, and responsibility (Shneiderman, 2022). The theological domain adds a further requirement: the system must disclose that computation cannot create spiritual authority. The interface should repeatedly return the user to the biblical text, prayer, conscience, community, and qualified counsel rather than cultivate dependence on the application.

The architecture also narrows the meaning of grounding. Technical grounding requires claims to remain traceable to retrieved documents. Theological grounding requires a defensible interpretation of Scripture in context. A system may satisfy the first condition and fail the second by citing a genuine verse for an application that ignores genre, audience, covenant setting, or Christ's teaching. The hermeneutical layer and expert evaluation therefore remain necessary even when retrieval metrics are strong.

Context equity is central to validity. A corpus dominated by affluent urban examples may assume refrigeration, private transport, stable employment, continuous electricity, and immediate pastoral access. Such assumptions can distort recommendations for rural households, informal workers, displaced families, people with disabilities, and communities with limited connectivity. Diverse benchmark scenarios are therefore part of theological fidelity as well as technical fairness. They test whether the system applies the same Sabbath purposes responsibly across different circumstances.

The study has four limitations. First, it presents a conceptual artifact and analytical demonstration, not a deployed system. Second, the purposive literature synthesis may omit relevant work. Third, the framework is intentionally grounded in Seventh-day Adventist theology and does not claim denominational neutrality. Fourth, expert disagreement will remain in some applications because the material facts and pastoral responsibilities differ. The system should expose that disagreement rather than manufacture certainty.

Implementation should proceed in stages. The first prototype should use deterministic rules and retrieval over a small, reviewed corpus. The second should add constrained natural-language explanation based only on retrieved evidence. The third should support a multilingual mobile pilot in selected congregations and university churches. Expansion should follow independent theological, safety, privacy, accessibility, and usability review. Publication of evaluation data, benchmark scenarios where confidentiality permits, and documented failure cases would strengthen accountability.

This paper concludes that artificial intelligence can support disciplined access to evidence and structured deliberation, but it cannot possess biblical wisdom or assume moral responsibility. The proposed AWDSS makes that limit a design requirement. Its contribution is a testable architecture that connects authority-labelled retrieval, contextual reasoning, six Sabbath tests, evidence verification, uncertainty, escalation, and human oversight. The next research step is to implement a minimum viable prototype and evaluate whether it improves source accuracy, contextual biblical reasoning, safety, and calibrated trust without displacing prayer, conscience, pastoral care, or accountable choice.

References

Amershi S, Weld D, Vorvoreanu M, Fourney A, Nushi B, Collisson P, Suh J, Iqbal S, Bennett PN, Inkpen K, Teevan J, Kikin-Gil R, Horvitz E, 2019. Guidelines for human-AI interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 4-9 May 2019, Paper 3: 1-13, Glasgow, United Kingdom. https://doi.org/10.1145/3290605.3300233.

Arnott D, Pervan G, 2014. A critical analysis of decision support systems research revisited: The rise of design science. Journal of Information Technology, 29 (4): 269-293. https://doi.org/10.1057/jit.2014.16.

Chan CKY, 2023. A comprehensive AI policy education framework for university teaching and learning. International Journal of Educational Technology in Higher Education, 20: Article 38. https://doi.org/10.1186/s41239-023-00408-3.

Crompton H, Burke D, 2024. The educational affordances and challenges of ChatGPT: State of the field. TechTrends, 68: 380-392. https://doi.org/10.1007/s11528-024-00939-0.

Dorobantu M, 2022. Artificial intelligence as a testing ground for key theological questions. Zygon, 57 (4): 984-999. https://doi.org/10.1111/zygo.12831.

Ferrara E, 2023. Should ChatGPT be biased? Challenges and risks of bias in large language models. First Monday, 28 (11). https://doi.org/10.5210/fm.v28i11.13346.

General Conference of Seventh-day Adventists, 1990. Sabbath observance guidelines. Access address: https://gc.adventist.org/documents/sabbath-observance/; Date of access: 09/30/2026.

Gregor S, Hevner AR, 2013. Positioning and presenting design science research for maximum impact. MIS Quarterly, 37 (2): 337-355. https://doi.org/10.25300/MISQ/2013/37.2.01.

Hevner AR, March ST, Park J, Ram S, 2004. Design science in information systems research. MIS Quarterly, 28 (1): 75-105. https://doi.org/10.2307/25148625.

Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang YJ, Madotto A, Fung P, 2023. Survey of hallucination in natural language generation. ACM Computing Surveys, 55 (12): Article 248, 1-38. https://doi.org/10.1145/3571730.

Khosravi H, Shum SB, Chen G, Conati C, Tsai YS, Kay J, Knight S, Martinez-Maldonado R, Sadiq S, Gašević D, 2022. Explainable artificial intelligence in education. Computers and Education: Artificial Intelligence, 3: Article 100074. https://doi.org/10.1016/j.caeai.2022.100074.

Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih WT, Rocktäschel T, Riedel S, Kiela D, 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33: 9459-9474.

Miller T, 2019. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267: 1-38. https://doi.org/10.1016/j.artint.2018.07.007.

Ministerial Association of the General Conference of Seventh-day Adventists, 2018. Seventh-day Adventists believe: An exposition of the fundamental beliefs of the Seventh-day Adventist Church. 3rd ed., Review and Herald Publishing Association, Silver Spring, Maryland, USA.

Papakostas C, 2025. Artificial intelligence in religious education: Ethical, pedagogical, and theological perspectives. Religions, 16 (5): Article 563. https://doi.org/10.3390/rel16050563.

Peffers K, Tuunanen T, Rothenberger MA, Chatterjee S, 2007. A design science research methodology for information systems research. Journal of Management Information Systems, 24 (3): 45-77. https://doi.org/10.2753/MIS0742-1222240302.

Power DJ, 2002. Decision support systems: Concepts and resources for managers. Quorum Books, Westport, Connecticut, USA.

Rudin C, Chen C, Chen Z, Huang H, Semenova L, Zhong C, 2022. Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistics Surveys, 16: 1-85. https://doi.org/10.1214/21-SS133.

Shin D, 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. International Journal of Human-Computer Studies, 146: Article 102551. https://doi.org/10.1016/j.ijhcs.2020.102551.

Shneiderman B, 2022. Human-centered AI. Oxford University Press, Oxford, United Kingdom. https://doi.org/10.1093/oso/9780192845290.001.0001.

Snyder H, 2019. Literature review as a research methodology: An overview and guidelines. Journal of Business Research, 104: 333-339. https://doi.org/10.1016/j.jbusres.2019.07.039.

Tabassi E, 2023. Artificial intelligence risk management framework (AI RMF 1.0). NIST AI 100-1, National Institute of Standards and Technology, Gaithersburg, Maryland, USA. https://doi.org/10.6028/NIST.AI.100-1.

Tonstad SK, 2009. The lost meaning of the seventh day. Andrews University Press, Berrien Springs, Michigan, USA.

Vilone G, Longo L, 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. Information Fusion, 76: 89-106. https://doi.org/10.1016/j.inffus.2021.05.009.

White EG, 1905. The ministry of healing. Pacific Press Publishing Association, Mountain View, California, USA.

Yang SJH, Ogata H, Matsui T, Chen NS, 2021. Human-centered artificial intelligence in education: Seeing the invisible through the visible. Computers and Education: Artificial Intelligence, 2: Article 100008. https://doi.org/10.1016/j.caeai.2021.100008.

Acknowledgment

This paper develops concepts first explored by the author in the unpublished manuscript *Living the Seventh Day Sabbath* and in Bible-study discussions with university church members. OpenAI ChatGPT assisted with language editing, structural organization, and document formatting. The author reviewed the manuscript and accepts responsibility for its arguments, sources, and final form. No external funding supported this study.

Conflict of Interest

The author has declared that there is no conflict of interest.

Author Contributions

Kevin Omondi Opiyo: Conceptualization, methodology, investigation, system architecture, writing, review, and editing.

Image-Based Drowsiness Detection: Comparative Performance Evaluation of Different CNN Models

Musa Milli[1], Cüneyd Murabıt Akdoğan[2]*, Rıza Kaya[3]

[1,2,3] National Defence University, Turkish Naval Academy, Department of Computer Engineering, 34940, Istanbul, Türkiye

*Corresponding author e-mail: cuneydmurabit.a@msu.edu.tr

Abstract

In this study, driver fatigue and drowsiness are detected from images of the driver's face. Six deep learning architectures, namely VGG16, ResNet18, MobileNetV2, DenseNet121, EfficientNetB2, and ArcFaceCNN, were trained under identical conditions and compared with each other. The experiments were run on a dataset of 65,000 images; before training, each image was cropped so that only the face and head region remained in the frame. The models were trained with 10-fold cross-validation and scored on accuracy, precision, recall, and F1-score. Among the six, EfficientNetB2, DenseNet121, and VGG16 gave the best results, both in terms of accuracy and in terms of how stable their performance stayed from one fold to the next.

Keywords: *Driver Fatigue Detection, Artificial Intelligence, Deep Learning, CNN, Image Processing*

INTRODUCTION

Driver fatigue and drowsiness are still among the leading causes of traffic accidents in present-day transportation systems. The measurements used to detect fatigue fall into two groups, behavioral and physiological. From an engineering point of view, visual cues such as drooping eyelids and yawning are the most practical indicators to work with; the trouble is that classical image processing methods have difficulty in telling a normal blink from an eye that is closing because of drowsiness, and this weakness is the main reason more robust solutions have been looked for. In the last few years, Convolutional Neural Networks (CNN) have changed the picture considerably, because they can learn complex facial expressions directly from the pixels. Transfer learning, in particular, has become a widely preferred strategy in order to achieve high accuracy rates with limited datasets. The main purpose of this study is to systematically compare the performances of different deep learning architectures prominent in the literature regarding the fatigue detection problem. In this context, VGG-16, ResNet-18, MobileNetV2, DenseNet-121, EfficientNet-B2, and ArcFaceCNN models have been considered. In the conducted experimental studies, these models were trained and tested on a comprehensive dataset consisting of a total of 65,000 images where the face and head region were isolated. As a result of the study, a comparative analysis of the models in question was performed over fundamental classification metrics such as accuracy, precision, recall, and F1-score, and the models exhibiting the highest performance and stability in drowsiness detection were determined.

LITERATURE REVIEW

Comprehensive reviews in the literature fundamentally divide fatigue detection measurements into two main categories as behavioral and physiological measurements (Sahayadhas et al., 2012). Conducted studies emphasize that several different observational indicators must be combined in order to perform a high-accuracy fatigue detection (Liu et al., 2022). Viewed from an engineering perspective, fatigue and drowsiness state are generally defined by sleepy eyes and yawning symptoms, while in the medical perspective, the use of biosignals such as heart rate variability (HRV) comes to the forefront (Patel et al., 2011). Recent surveys also point to the weak sides of both approaches: image-based methods have trouble separating an ordinary blink from eyes that are closing because of drowsiness, while physiological sensors have to be attached to the body and are often found uncomfortable by drivers.

The classical computer vision work on fatigue detection is quite varied and goes from infrared sensing all the way to purely digital image processing. One of the earliest examples tracked the driver's face with the help of infrared sensors (Haisong Gu et al., 2002). These techniques have been applied not only for drivers but also in medical fields such as monitoring the fatigue states of anesthetists (Vasquez-Lopez et al., 2015). In more current approaches, on the other hand, blink and fatigue detection have been performed using histogram of oriented gradients (HOG) features (Bakheet & Al-Hamadi, 2021). Furthermore, Improved Active Shape Model (ASM) algorithms based on statistical learning models to extract facial features quickly and effectively, and AdaBoost classifiers based on Haar-like features, have constituted the basis for face detection and eye region localization.

Machine learning-based studies focus on making fatigue prediction by analyzing both visual features and physiological signals. For example, it has been proven that driver fatigue can be evaluated by applying neural network analysis on heart rate variability (HRV) data, and that the LF/HF (low frequency/high frequency) ratio changes significantly between fatigue and wakefulness states (Patel et al., 2011). In the classification stage, Naïve Bayes (NB) models based on Bayes' theorem are used effectively, especially in the processing of large-scale datasets and probabilistically predicting the eye state (Bakheet & Al-Hamadi, 2021). In practice, the feature vector extracted from each frame is assigned to the class with the highest posterior probability, that is, by the maximum a posteriori (MAP) rule, and the drowsiness decision is then built on top of these frame-level classifications.

When it comes to reading complex facial expressions, or small regions such as the eyes, Convolutional Neural Networks (CNN) are by now the most common choice in the literature (Liu et al., 2022). An early and influential example is the joint training of cascaded CNNs for face detection (Qin et al., n.d.). In the literature, hierarchical CNN-based systems have been proposed for real-time fatigue detection, and original models have been developed using visual technologies (Gu et al., 2018). Additionally, lightweight CNN models that can be integrated into Android applications have been designed for real-time drowsiness detection on mobile devices. In particular, the InceptionV3 structure provides high accuracy in extracting detailed elements like eyes, thanks to its ability to divide features into smaller convolution partitions (Zhao et al., 2020).

The inadequacy of using only CNN in distinguishing normal blinking and drowsiness state has led to the development of hybrid models combining CNN and Long Short-Term Memory (LSTM) networks to process sequential video data. The InceptionV3-LSTM structure performs a temporal analysis by processing the feature sequences extracted via CNN in the LSTM network and makes the drowsiness decision according to this sequential data. FDRNet (Fatigue Driving Recognition Network), developed

in a similar way, detects driver fatigue by using a multi-task cascaded CNN and LSTM units (Xiao et al., 2019). These temporal models solve long-term dependencies by ensuring the continuity of information between blocks and increase the stability in drowsiness detection.

Transfer learning is widely preferred to reach high accuracy rates with limited datasets and to accelerate the training process. Pre-trained deep neural networks such as VGG-FaceNet, AlexNet, and FlowImageNet are used together with ensemble strategies for driver drowsiness detection. Applying transfer learning on mature networks like InceptionV3 and retraining the last layer of the network for fatigue and wakefulness classes increases the flexibility and accuracy of the model. Furthermore, studies in which the ResNet-50 network is configured with transfer learning for the purpose of fault diagnosis and fatigue detection also hold an important place in the literature.

MODELS

VGG-16

VGG-16 is a deep convolutional neural network architecture proposed by Simonyan and Zisserman and widely used in image classification tasks. The model consists of 13 convolutional and 3 fully connected layers and uses 3×3 kernels throughout the entire network. Providing high accuracy despite its simple and homogeneous structure, VGG-16 is frequently used as a reference model in deep learning studies (Simonyan & Zisserman, 2015).

ResNet-18

ResNet-18 is a CNN architecture based on the residual learning approach. Thanks to the identity shortcut connections located between the layers, it offers a more stable training process by reducing the performance degradation seen in deep networks. Consisting of 18 layers, this architecture is widely used in the literature with its high accuracy performance despite its low number of parameters (He et al., 2016).

MobileNetV2

MobileNetV2 is a lightweight CNN architecture developed for mobile and embedded systems. Thanks to inverted residual blocks and depthwise separable convolution structure, it offers competitive accuracy values while reducing computational cost. For this reason, MobileNetV2 is especially preferred in real-time and resource-constrained applications (Sandler et al., 2018).

DenseNet-121

DenseNet-121 is a densely connected architecture where each layer takes the output of all preceding layers as input. This structure improves gradient flow by encouraging feature reuse and enables the model to learn more efficiently. DenseNet-121 is widely used in image analysis studies due to its high accuracy and parameter efficiency (Huang et al., 2017).

EfficientNet-B0

EfficientNet-B0 is a CNN architecture based on the compound scaling approach that scales network depth, width, and input resolution in a balanced manner. Thanks to MBConv blocks and the squeeze-and-excitation mechanism, it offers high accuracy performance with a low number of parameters. For

this reason, it is an important reference architecture among efficient deep learning models (Tan & Le, 2021).

ArcFaceCNN

ArcFace is a deep convolutional neural network (DCNN) training methodology based on the Additive Angular Margin Loss principle, developed to maximize class separability in deep face recognition systems. Unlike traditional softmax loss, ArcFace operates over the geodesic distance margin by directly optimizing the angle between normalized features and weights. This approach significantly increases face recognition performance by strengthening inter-class discrepancy while increasing intra-class compactness (Deng et al., 2022).

EXPERIMENTAL STUDIES

Before the models were trained, the data was put through a special preprocessing stage. In this preprocessing stage, the background parts of the raw photograph data found in the dataset were cropped, ensuring that only the face and head remained in the photograph. The cropping process was applied to remove irrelevant background information and to preserve only the region of interest of the face. In the used dataset, there is a total of 65,000 data distributed into two classes (drowsy and awake). While conducting controlled experiments, k-fold cross-validation was used, and the k value was determined as 10 in this process. Within each fold, 80% of the data was used for training, 10% for validation, and 10% for testing. While there were variable and randomly selected training and validation data within each fold, the test dataset was randomly determined at the very beginning of the experiment and kept constant for each experiment to provide a consistent evaluation. Furthermore, to ensure full reproducibility and consistency among experiments for the data groups used in all folds, their image directories and labels were kept in files with a .csv extension.

Under the aforementioned experimental conditions, multiple deep learning architectures including ResNet-18, MobileNetV2, DenseNet-121, EfficientNet-B2, and ArcFaceCNN were evaluated. All models were trained for 10 epochs using a GPU. Accuracy, precision, recall, and F1-score metrics were used as performance evaluation. The experiments were carried out on a workstation equipped with an NVIDIA GeForce RTX 3060 Laptop GPU (6 GB VRAM), using PyTorch with Python 3.10 and CUDA 12.6 on Windows 11.

Figure 1 shows the training loss change of the evaluated deep learning models over 10 epochs, averaged over 10-fold cross-validation. All architectures exhibit a rapid decrease in training loss in the initial epochs, indicating that the discriminative facial features related to drowsiness were learned effectively. Lightweight and moderately deep models such as MobileNetV2, EfficientNetB2, ResNet18, and DenseNet121 show faster convergence and exhibit stable loss values with relatively low variance across layers. ArcFaceCNN behaves differently: it starts from a much higher loss and comes down more slowly, which is what one would expect from its margin-based loss and the stronger regularization that comes with it. VGG16 also converges a little more slowly than the compact models, although its curve is smooth; the larger number of parameters is the likely reason. In short, Figure 1 shows that the architecture, and the amount of complexity that comes with it, has a visible effect on how the optimization proceeds in a face-based drowsiness task.

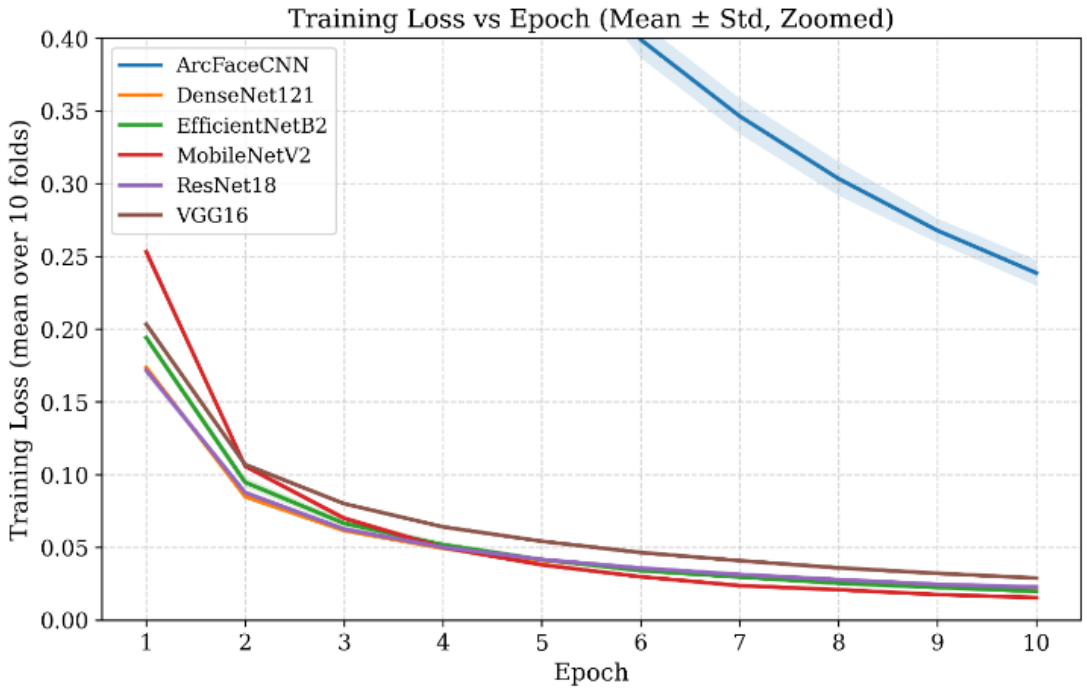


Figure 1. Training loss of the six models over 10 epochs, averaged over the 10 folds. Each solid line is the mean training loss of one model and the shaded band around it is the standard deviation across folds.

Figure 2 and Figure 3 give the fold-epoch heatmaps of the validation F1-score for VGG16 and MobileNetV2, and the two models clearly do not converge in the same way during the 10-fold cross-validation. VGG16 reaches high validation scores early; after the first few epochs its F1-score passes 0.98 in nearly every fold and stays there, which speaks for a strong learning capacity and good generalization. The colors in the later epochs are fairly uniform from fold to fold, so the variance is low and the optimization is stable. The picture for MobileNetV2 is a bit different. Its F1-scores are slightly lower but more even across folds, especially in the first epochs, which fits a lightweight network that converges quickly. In the later epochs MobileNetV2 does reach competitive values, yet it never quite catches up with VGG16; the gap is the usual price paid for computational efficiency in terms of representation capacity. Read side by side, the two heatmaps say that VGG16 gives the higher peak performance and the more consistent folds, while MobileNetV2 gives stable and efficient learning at a lower computational cost.

When Figure 4 is examined generally, it is seen that the EfficientNetB2, DenseNet121, and VGG16 models exhibit high-level and stable performance in all metrics. Especially EfficientNetB2 has the highest average values in terms of both Precision and Accuracy, and reveals low variance and high generalization capability thanks to its narrow min-max ranges. While DenseNet121 similarly stands out with high accuracy and balanced error ranges, it can be said that VGG16 comes to the forefront in the Recall metric, and this situation shows that the model is particularly strong in not missing the positive class. The fact that the error bars in these three models are relatively short points to a stable learning behavior across different folds.

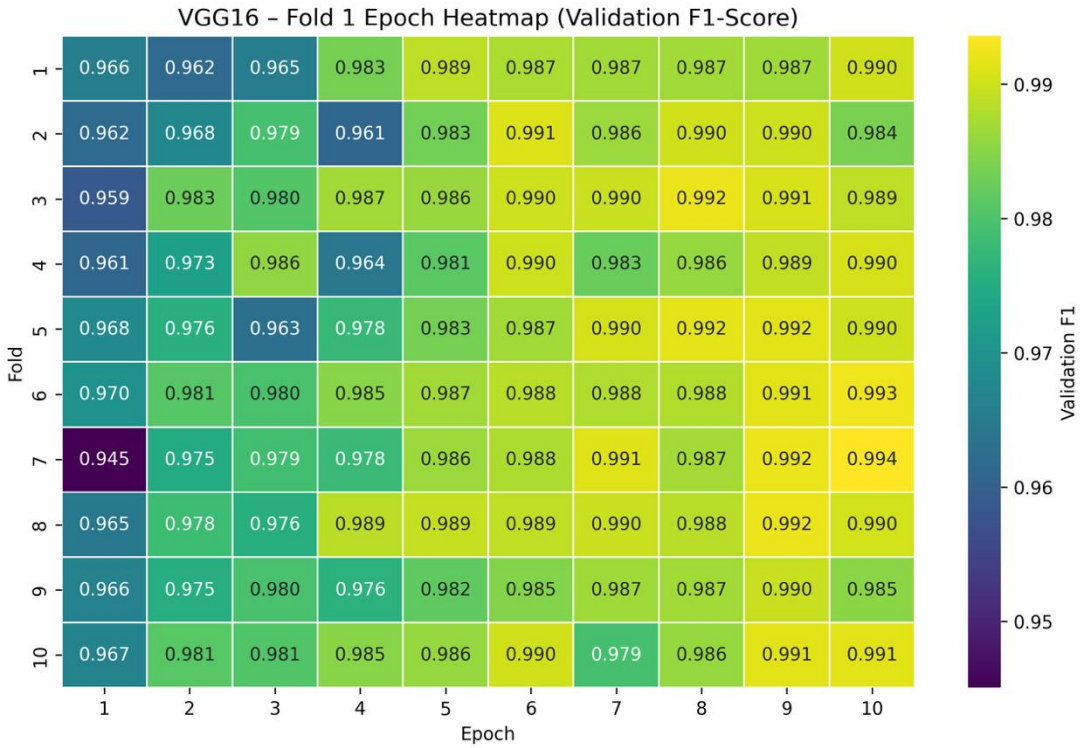


Figure 2. Fold-epoch heatmap of the validation F1 score for the VGG16 model throughout 10-fold cross-validation.

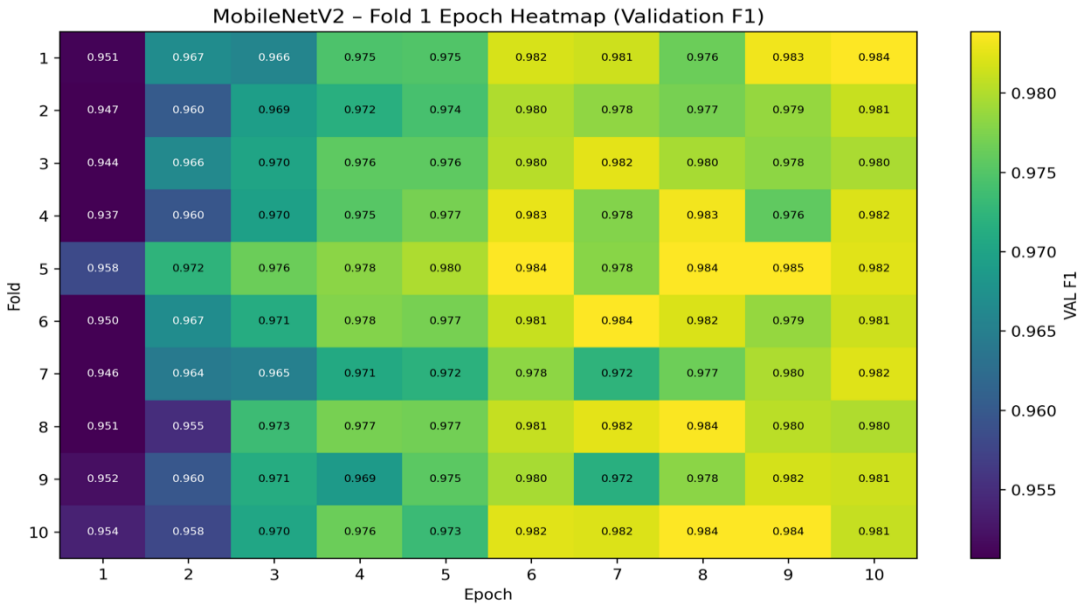


Figure 3. Fold-epoch heatmap of the validation F1 score for the MobileNetV2 model throughout 10-fold cross-validation.

On the other hand, although ResNet18 has somewhat lower average values compared to the top group models, it exhibits a consistent performance especially in terms of Precision and F1-Score and offers a mid-level balance point. ArcFaceCNN and MobileNetV2, however, diverge with relatively lower average metric values and wider min-max ranges. This situation indicates that the models in question produce more variable results across folds and are more sensitive to data distribution. Considering the lighter architecture of MobileNetV2 in particular, it can be said that this decrease in performance reflects the classic trade-off between computational efficiency and accuracy.

As a result, Figure 4 reveals not only average performance values but also model stability and generalization power. EfficientNetB2 and DenseNet121 come out as the most dependable of the six, since they pair high accuracy with low variance, and VGG16 is a strong alternative when Recall is the priority. ArcFaceCNN and MobileNetV2 have their own uses, one being designed for special scenarios and the other being much lighter, but on overall test performance both stay behind the other models. What this tells us is that choosing a model should depend on fold-to-fold stability and on the requirements of the application, not on the highest accuracy figure alone.

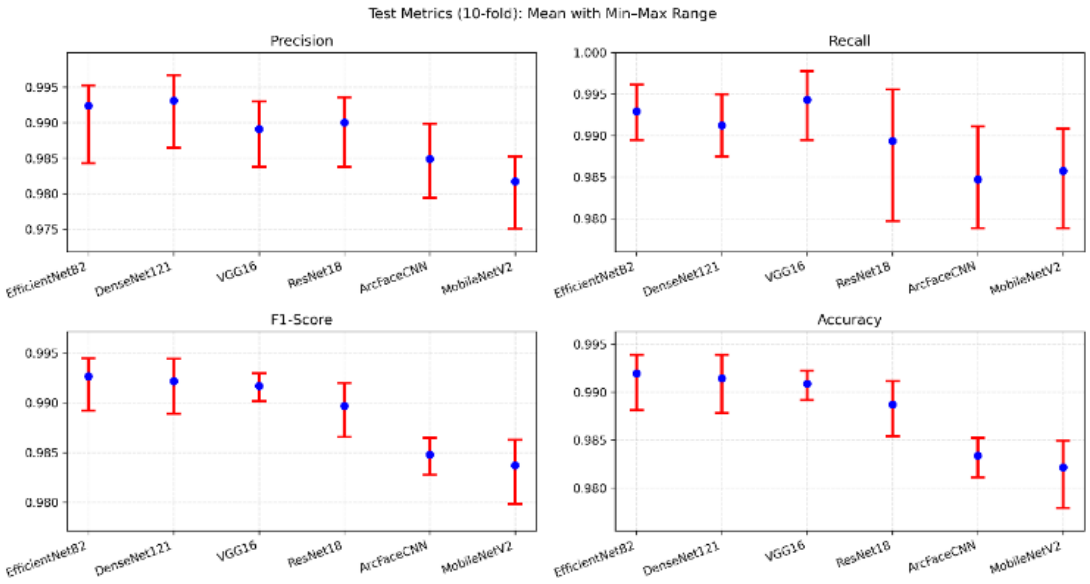


Figure 4. Test set precision, recall, F1-score, and accuracy of the models under 10-fold cross-validation. Points mark the mean over the folds and the bars mark the minimum and maximum values.

RESULTS

In this study six deep learning architectures that are frequently used for vision-based driver drowsiness detection, VGG16, ResNet18, MobileNetV2, DenseNet121, EfficientNetB2, and ArcFaceCNN, were put through the same experimental procedure and compared. The procedure was kept identical for every model so that the comparison would be fair and repeatable: the same 10-fold cross-validation, the same training, validation, and test splits, and a single preprocessing step that leaves only the face and head region in the image. On accuracy, precision, recall, and F1-score the best results belonged to EfficientNetB2, DenseNet121, and VGG16, and since their scores varied little across folds, we can say

that these three generalize well. Among them EfficientNetB2 gave the best balance of performance and stability, while VGG16 was especially good on recall, in other words it missed fewer of the drowsy cases.

The average scores alone would hide some of the differences. When the fold-epoch heatmaps, the loss curves, and the minimum-maximum ranges are looked at together, it becomes clear that the models also differ in the way they learn and in how robust they are. A light architecture like MobileNetV2 is cheap to run and converges quickly, but its results move around more from fold to fold and its best scores are lower than those of the bigger networks; this is the familiar exchange between efficiency and representational capacity. The more expressive architectures learn in a steadier manner, but they cost more to train and to run. For that reason we do not think a model for drowsiness detection should be chosen on peak accuracy alone. How stable it is across folds, how it converges, and practical matters such as the computing resources at hand and whether real-time operation is required all need to be weighed. We hope the results reported here will be of use to other researchers working on the same problem and to those who want to build a reliable vision-based drowsiness detection system for real use.

References

Bakheet, S., & Al-Hamadi, A. (2021). A Framework for Instantaneous Driver Drowsiness Detection Based on Improved HOG Features and Naïve Bayesian Classification. *Brain Sciences*, *11*(2), 240. https://doi.org/10.3390/brainsci11020240

Deng, J., Guo, J., Yang, J., Xue, N., Kotsia, I., & Zafeiriou, S. (2022). ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(10), 5962–5979. https://doi.org/10.1109/TPAMI.2021.3087709

Gu, W. H., Zhu, Y., Chen, X. D., He, L. F., & Zheng, B. B. (2018). Hierarchical CNN-based real-time fatigue detection system by visual-based technologies using MSP model. *IET Image Processing*, *12*(12), 2319–2329. https://doi.org/10.1049/iet-ipr.2018.5245

Haisong Gu, Qiang Ji, & Zhiwei Zhu. (2002). Active facial tracking for fatigue detection. *Sixth IEEE Workshop on Applications of Computer Vision, 2002. (WACV 2002). Proceedings.*, 137–142. https://doi.org/10.1109/ACV.2002.1182171

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely Connected Convolutional Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261–2269. https://doi.org/10.1109/CVPR.2017.243

Liu, F., Chen, D., Zhou, J., & Xu, F. (2022). A review of driver fatigue detection and its advances on the use of RGB-D camera and deep learning. *Engineering Applications of Artificial Intelligence*, *116*, 105399. https://doi.org/10.1016/j.engappai.2022.105399

Patel, M., Lal, S. K. L., Kavanagh, D., & Rossiter, P. (2011). Applying neural network analysis on heart rate variability data to assess driver fatigue. *Expert Systems with Applications*, *38*(6), 7235–7242. https://doi.org/10.1016/j.eswa.2010.12.028

Qin, H., Yan, J., Li, X., & Hu, X. (n.d.). *Joint Training of Cascaded CNN for Face Detection*.

Sahayadhas, A., Sundaraj, K., & Murugappan, M. (2012). Detecting Driver Drowsiness Based on Sensors: A Review. *Sensors*, *12*(12), 16937–16953. https://doi.org/10.3390/s121216937

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4510–4520. https://doi.org/10.1109/CVPR.2018.00474

Simonyan, K., & Zisserman, A. (2015). *Very Deep Convolutional Networks for Large-Scale Image Recognition* (arXiv:1409.1556). arXiv. https://doi.org/10.48550/arXiv.1409.1556

Tan, M., & Le, Q. V. (2021). *EfficientNetV2: Smaller Models and Faster Training* (arXiv:2104.00298). arXiv. https://doi.org/10.48550/arXiv.2104.00298

Vasquez-Lopez, J. A., Vargas-Cañas, R., & Mera-Jiménez, S. L. (2015). Fatigue Detection in Anaesthesiologists Using Digital Image Processing Techniques. In A. Braidot & A. Hadad (Eds.), *VI Latin American Congress on Biomedical Engineering CLAIB 2014, Paraná, Argentina 29, 30 & 31 October 2014* (Vol. 49, pp. 472–475). Springer International Publishing. https://doi.org/10.1007/978-3-319-13117-7_121

Xiao, Z., Hu, Z., Geng, L., Zhang, F., Wu, J., & Li, Y. (2019). Fatigue driving recognition network: Fatigue driving recognition via convolutional neural network and long short-term memory units. *IET Intelligent Transport Systems*, *13*(9), 1410–1416. https://doi.org/10.1049/iet-its.2018.5392

Zhao, Y., Xie, K., Zou, Z., & He, J.-B. (2020). Intelligent Recognition of Fatigue and Sleepiness Based on InceptionV3-LSTM via Multi-Feature Fusion. *IEEE Access*, *8*, 144205–144217. https://doi.org/10.1109/ACCESS.2020.3014508

Acknowledgment

In this section, acknowledgments can be made to the persons or institutions who supported the study.

Conflict of Interest

There is no conflict of interest to declare for this study.

Author Contributions

Contributions of the authors to the study can be explained.

Deep Learning versus Classical Machine Learning for Anomaly Detection in Embedded Network Traffic

Zeynep Eflatun[1], Onuralp Gayret[2]*

[1,2] Marmara University, Faculty of Business Administration, Management Information Systems Department, Türkiye

*Corresponding author e-mail: zeynepeflatunn@gmail.com

Abstract

This study explains, through comparison, that the deep learning method autoencoder used to detect anomalous behavior in embedded system networks is more successful than machine learning–based methods such as isolation forest and local outlier factor. In today's digital environment, network security threats are steadily increasing, which heightens the importance of detecting these threats as quickly and efficiently as possible. In this context, the study conducted to compare existing methods begins by thoroughly handling the data and applying the necessary data preprocessing procedures; in order to eliminate class imbalance in the dataset, synthetic data generation jittering techniques are used. To better represent the complex structure of network traffic, various new features have been derived, such as port categorization, bytes per second, risky destination ports, and the number of active TCP flags. In the study, three different models are examined in order to detect anomalous behavior in embedded system networks: a deep learning–based Autoencoder, and machine learning–based Local Outlier Factor and Isolation Forest. During the evaluation phase, the three different models used were trained on data samples belonging to network traffic characteristics that do not exhibit anomalous behavior. While the Autoencoder determines anomalies through reconstruction errors, the Local Outlier Factor and Isolation Forest algorithms detect deviations in the data distribution. The performance of the developed models has been evaluated using metrics such as precision, recall, F1-score, and ROC-AUC, and it has been demonstrated through these evaluation metrics that the deep learning–based autoencoder method performs better than the machine learning–based local outlier factor and isolation forest methods. This study aims to contribute by comparing the methods used to best develop early warning mechanisms against potential cyberattacks, while presenting a multi-layered anomaly detection framework to enhance the network security of embedded systems.

Keywords: *Anomaly Detection, Deep Learning, Autoencoder, Isolation Forest, Local Outlier Factor*

INTRODUCTION

Modern information systems and global networks have largely overcome the physical limits of earlier industrial development, creating a digital ecosystem in which operational, social and economic processes are increasingly shaped by digital infrastructures. This transformation, however, also exposes organizations to an unprecedented attack surface (Sommer & Paxson, 2010). Because signature-based security systems are largely reactive against zero-day and polymorphic threats, proactive, AI-based

defense mechanisms that learn normal network behaviour and flag deviations have become increasingly important (Buczak & Guven, 2016). The present study is explicitly positioned within the paradigm of unsupervised anomaly / novelty detection (Chandola et al., 2009) rather than supervised classification or clustering.

The aim of this study is to examine the practical and theoretical limits of different anomaly detection approaches for network traffic of embedded systems by comparing a Deep-Learning-based Autoencoder architecture with two classical unsupervised Machine-Learning methods, Isolation Forest and Local Outlier Factor, and by evaluating them on empirical performance metrics (Accuracy, Precision, Recall, F1-Score, ROC-AUC).

LITERATURE REVIEW

Anomaly detection has become an important research area in network security because traditional signature-based systems may fail to detect previously unseen attacks and zero-day threats. Unsupervised approaches are particularly useful because they can learn normal network behavior without requiring large amounts of labeled attack data (Chandola et al., 2009). Isolation Forest is a widely used unsupervised method due to its computational efficiency and ability to detect anomalies through random partitioning. The method isolates abnormal observations more quickly than normal ones, making it suitable for anomaly detection tasks (Liu et al., 2008). Local Outlier Factor (LOF) is another commonly used technique that identifies anomalies by comparing the local density of an observation with the densities of its neighbors (Breunig et al., 2000). However, its performance may decrease in high-dimensional datasets.

Deep learning approaches, especially Autoencoders, are also widely applied in network anomaly detection. Autoencoders trained on normal traffic learn to reconstruct normal observations with low error, while anomalous samples generally produce higher reconstruction errors. This makes them effective for modeling complex and nonlinear network traffic patterns (Kwon et al., 2019).

In practical intrusion detection systems, false-positive rates are also critical, since excessive alarms may reduce operational usability (Sommer & Paxson, 2010). Therefore, comparing Isolation Forest, LOF, and Autoencoder under the same dataset and evaluation conditions provides a meaningful basis for assessing their anomaly detection performance.

MATERIAL AND METHODS

Material

The dataset used in this study is a publicly available Kaggle dataset representing network traffic of embedded systems. After feature-engineering and jitter-based data augmentation, the dataset was expanded to 10,000 rows and split into a training set (5,398 rows, exclusively normal traffic, unsupervised baseline learning), a validation set (2,300 rows; 1,799 normal / 501 attack instances, used for threshold optimization) and a test set (2,302 rows; 1,800 normal / 502 attack instances, held out for the final, unbiased evaluation). To reduce the strong class imbalance between normal and attack traffic, synthetic data generation and jittering techniques were applied. This ensured that model evaluation on the balanced test split reflects realistic Security Operations Center (SOC) conditions, where both detection rate and false-positive rate must be controlled simultaneously (Sommer & Paxson, 2010).

Methods

The Collection of the Data

Twenty-two features spanning packet-level statistics (packet size, inter-arrival time), port information and derived TCP-flag counts were used. Feature engineering added variables such as bytes_per_second, is_risky_dst_port and active_tcp_flags_count to better capture volumetric floods, reconnaissance/port-scanning and stealthy exfiltration patterns.

Statistical Analysis

During the preprocessing stage, one-hot encoding was applied to categorical variables, while numerical variables were standardized using Z-score normalization with StandardScaler. This step was performed to prevent variables with large numerical ranges, such as port numbers, from dominating the learning process.

Three unsupervised anomaly detection methods were used in the study: Autoencoder, Isolation Forest, and Local Outlier Factor (LOF). All models were trained exclusively on normal network traffic samples. Isolation Forest was applied as an ensemble-based method that isolates anomalies through random partitioning (Liu et al., 2008), whereas LOF identifies anomalous observations by comparing their local densities with those of neighboring samples (Breunig et al., 2000).

The Autoencoder model was implemented as a fully connected neural network including LeakyReLU activation functions, batch normalization, and dropout layers. A combined MSE and MAE loss function was used to calculate reconstruction error, while the AdamW optimizer was employed during training. The anomaly detection mechanism was based on the assumption that normal observations would produce relatively low reconstruction errors, whereas anomalous samples would result in higher reconstruction errors.

To improve model performance and systematically determine suitable hyperparameter combinations, the Optuna hyperparameter optimization framework was used. Optuna provides an efficient optimization strategy by allowing dynamic search-space definition and early pruning of unpromising trials (Akiba et al., 2019). Different combinations of model architecture and training parameters were evaluated, and the most suitable hyperparameters were selected according to validation performance.

For Isolation Forest and LOF, the selected baseline parameters were set to 100 trees with a contamination value of 0.10 for Isolation Forest, and 50 neighbors, a contamination value of 0.10, and novelty=True for LOF. Final model performances were evaluated using Accuracy, Precision, Recall, F1-Score, and ROC-AUC metrics.

RESULTS

All models were trained solely on normal traffic and evaluated on the isolated, balanced test set (df_test). The Autoencoder was trained for 300 epochs using a combined MSE/MAE loss; the operating point at epoch 213 (best F1-Score) was selected as the deployed model to avoid alarm fatigue in SOC operation.

Table 1 summarizes the classification performance of all three methods on the final, balanced test dataset (df_test). Figure 1 further illustrates the confusion matrix of the best-performing Autoencoder

model, showing 1,095 true negatives, 704 false positives, 90 false negatives and 411 true positives on the test set.

Table 1. Comparison of anomaly detection methods on the final test dataset

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score	ROC-AUC
Autoencoder (Deep Learning)	65.48	36.86	82.04	0.5087	0.7519
Local Outlier Factor (LOF)	65.99	32.76	53.19	0.4055	0.6364
Isolation Forest (Ensemble)	62.25	32.74	69.32	0.4447	0.6456

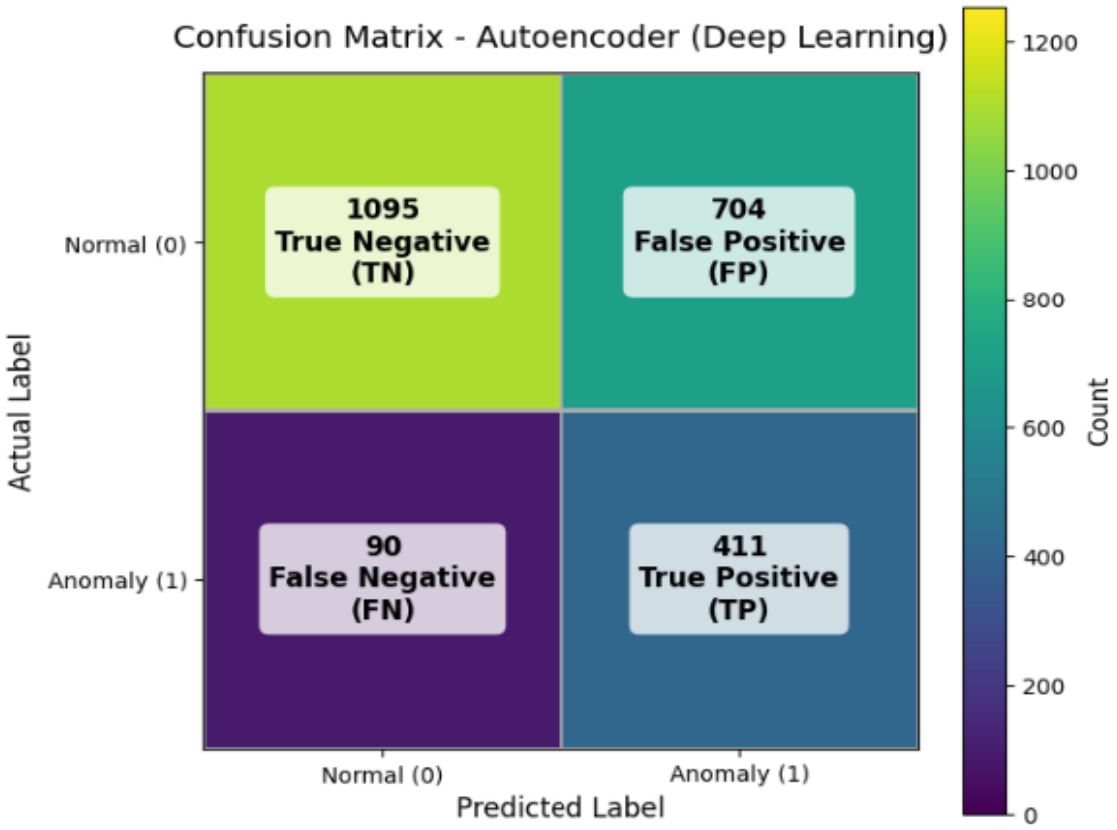


Figure 1. Confusion matrix of the Autoencoder model on the final test dataset.

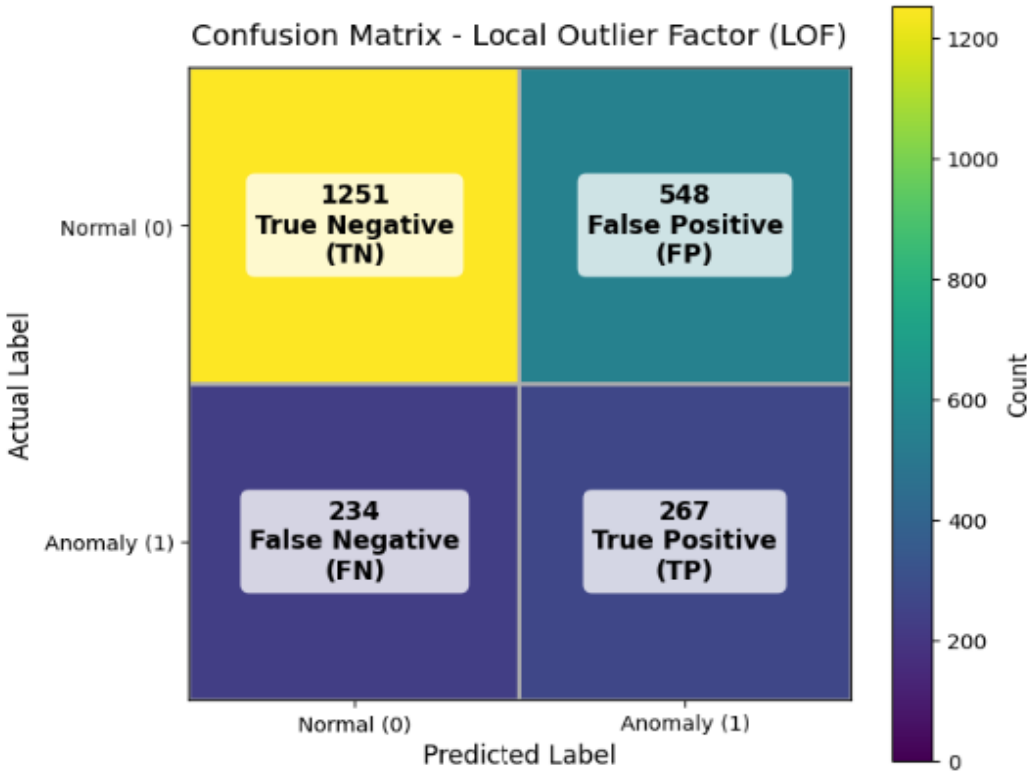


Figure 2. Confusion matrix of the Local Outlier Factor model on the final test dataset.

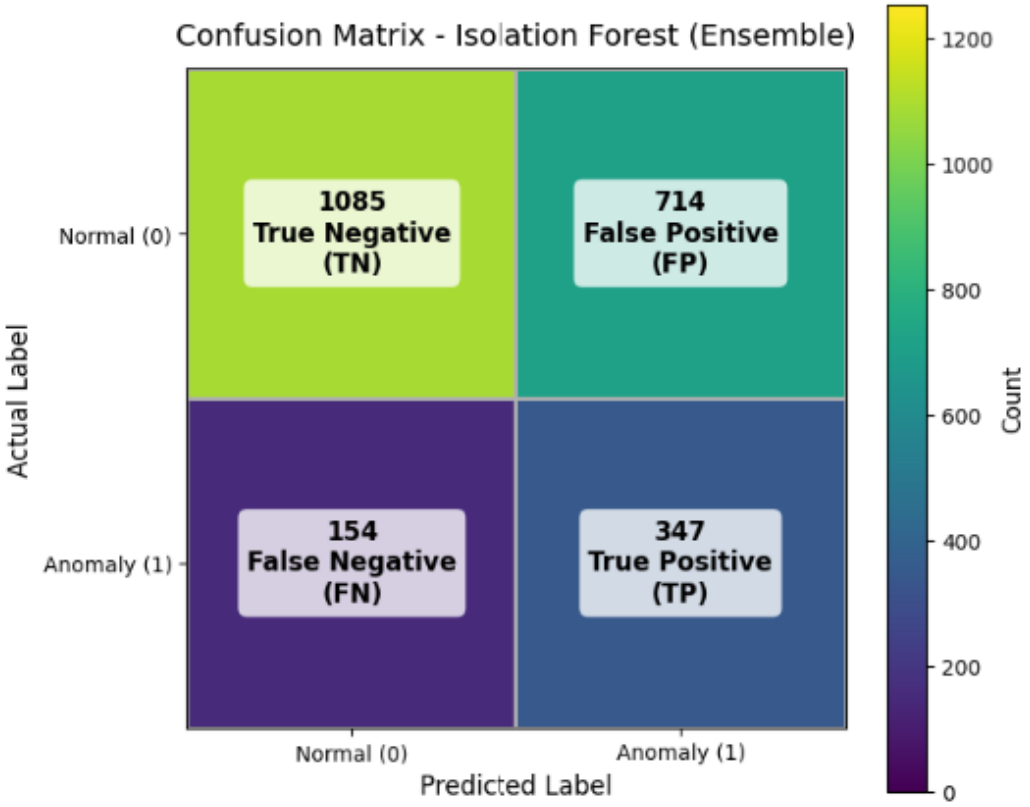


Figure 3. Confusion matrix of the Isolation Forest model on the final test dataset.

DISCUSSION AND CONCLUSION

The experimental results confirm the central hypothesis of this study: the Deep-Learning-based Autoencoder outperforms both classical Machine-Learning methods across all metrics, achieving the highest F1-Score (0.5087), Recall (82.04%) and ROC-AUC (0.7519). This superiority is attributed to the Autoencoder's ability to model nonlinear, high-dimensional manifolds of normal traffic through its LeakyReLU activations, batch normalization, dropout regularization and combined MSE/MAE loss, whereas Isolation Forest is limited by axis-aligned splitting and cross-feature blindness, and LOF collapses under the curse of dimensionality and the multi-density conflict between TCP and UDP traffic clusters (Liu et al., 2008; Breunig et al., 2000).

A key limitation is the relatively low precision of all three models (33-37%) under the balanced test conditions, which in operational SOC settings could lead to alarm fatigue; a hybrid pipeline in which the Autoencoder acts as a first-stage filter and Isolation Forest verifies high-risk anomalies is proposed as a mitigation. Future work will extend the Autoencoder toward temporal architectures (LSTM/GRU) to capture sequential flow dynamics, and will evaluate real-time deployment latency and resource consumption on IoT gateways and edge-computing devices (Kwon et al., 2019; Sperotto et al., 2010).

REFERENCES

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, 93–104.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys, 41(3), 1–58.

Kwon, D., Kim, H., Kim, J., Suh, S. C., Kim, I., & Kim, K. J. (2019). A survey of deep learning-based network anomaly detection. Cluster Computing, 22(1), 949–961.

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, 413–422.

Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. Proceedings of the 2010 IEEE Symposium on Security and Privacy, 305–316.

Sperotto, A., Schaffrath, G., Sadre, R., Morper, C., Pras, A., & Feldmann, A. (2010). An overview of IP flow-based intrusion detection. IEEE Communications Surveys & Tutorials, 12(3), 343–356.

Model Predictive Control versus Decoupled PI Control of a Quadruple-Tank Process under Actuator Faults

Serkan Özkan[1*]

[1] Gaziantep University, Naci Topçuoğlu Vocational School, 27310, Gaziantep, Türkiye

*Corresponding author e-mail: serkanozkan@gaziantep.edu.tr

Abstract

Multivariable liquid-level processes are widely used as benchmarks for evaluating control strategies because of their nonlinearity, loop interaction and sensitivity to actuator condition. This study evaluates the performance and the robustness of three controllers applied to the quadruple-tank process (QTP), a two-input two-output benchmark whose multivariable zero can be moved between the left and right half-plane by the valve settings. A nonlinear model based on Bernoulli's law and mass balance was developed and linearized at a minimum-phase operating point; relative gain array analysis then served to select the input-output pairings. A decentralized PI controller, a PI controller augmented with a dynamic decoupler and a model predictive controller (MPC) were designed and applied to the nonlinear simulation. Tracking was first examined with a step in the Tank 1 reference while the Tank 2 reference stayed constant. Five actuator fault scenarios were then injected: loss of effectiveness, leakage, lock-in-place, hard-over and stuck-open. The controllers were compared by their mean squared error (MSE) and time responses. Without faults, MPC achieved the lowest Tank 1 MSE (0.014), followed by the decoupled PI (0.0187) and the decentralized PI (0.036), all with zero steady-state error. MPC also gave the smallest Tank 1 error under loss of effectiveness (0.0154) and leakage (0.0141). Under the hard-over fault, none of the controllers could keep Tank 1 at its reference, which shows the limits of passive robustness without fault detection and reconfiguration. The results indicate that MPC copes best with loop interaction and moderate actuator faults, and they motivate adaptive and active fault-tolerant extensions.

Keywords: *Quadruple Tank Process, Model Predictive Control, Decoupling Control, Actuator Faults, Multivariable Systems*

INTRODUCTION

Many industrial processes have several inputs and several outputs, and each input usually affects more than one output. This loop interaction (coupling) limits the achievable performance of single-loop controllers and can even destabilize the closed loop. The quadruple-tank process (QTP) is a laboratory benchmark for such multivariable systems: two pumps supply four interconnected tanks, the two lower-tank levels are the controlled outputs, and the valve positions determine on which side of the imaginary axis the multivariable transmission zero lies, so that the process is either minimum-phase or non-minimum-phase (Johansson, 2000).

The QTP has been controlled with decentralized PI and decoupling structures (Nordfeldt and Hägglund, 2006; Ferrese et al., 2014; Effendi et al., 2017) and with model predictive control (MPC) (Zong et al., 2010; Camacho and Bordons, 2007). Two recent studies compare MPC with PID-type control on the QTP. One compares a decentralized PID controller with linear and nonlinear MPC on a physical setup and in simulation; the MPCs track time-varying setpoints better mainly when future setpoints are known, and reject disturbances only slightly better (Christensen et al., 2025). The other compares PID and offset-free MPC on linear and nonlinear QTP models and reports advantages of MPC under constant disturbances (Kasiyanto et al., 2024). Pumps and valves may also degrade or fail in practice, and the closed loop must then remain acceptable, or at least predictable. Fault-tolerant control is reviewed by Jiang and Yu (2012) and Zhang and Jiang (2008). A decentralized fault-tolerant PI-plus-fuzzy controller for the QTP has been tested under actuator, sensor and system component faults (Patel and Shah, 2019), and actuator faults of the QTP are also considered in other works (Arıcı and Kara, 2016; Hajiani and Poshtan, 2014; Arıcı and Kara, 2018). The comparisons of MPC and PID-type control cited above do not include actuator faults, whereas the fault studies use fuzzy, adaptive or detection-based schemes. As far as the author is aware, no side-by-side comparison of a decentralized PI, a decoupled PI and an MPC under several actuator fault types on the same QTP model has been reported; this paper addresses that gap.

This paper builds a nonlinear QTP model, linearizes it at a minimum-phase operating point and designs three controllers: a decentralized PI controller, a PI controller with a dynamic decoupler, and an MPC. Their tracking and coupling-rejection performance is compared by simulation of the nonlinear model, first without faults and then under five actuator fault types (loss of effectiveness, leakage, lock-in-place, hard-over and stuck-open) using the mean squared error (MSE) and time responses. The modelling, the augmented-error state-feedback design, the PI controller and the decoupler were first presented by Özkan et al. (2017); the state-feedback design is not repeated here. The present paper builds on the author's M.Sc. thesis (Özkan, 2019) and adds the MPC design and the actuator-fault analysis.

MATERIAL AND METHODS

Material

Process Description and Nonlinear Model

The QTP (Figure 1) has two pumps and four water tanks that are connected to each other. Pump 1 (voltage v_1) feeds Tank 1 and, through a three-way valve, Tank 4; pump 2 (voltage v_2) feeds Tank 2 and Tank 3. Tank 3 discharges into Tank 1 and Tank 4 into Tank 2, so each lower tank is affected by both pumps. The controlled outputs are the lower-tank levels h_1 and h_2, measured as voltages $k_c h_1$ and $k_c h_2$. Applying the mass balance to tank i and Bernoulli's law to its outlet gives (Johansson, 2000)

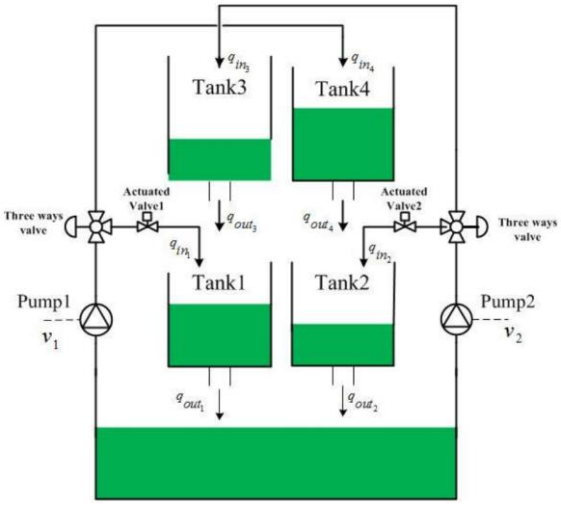


Figure 1. schematic of the quadruple-tank process.

$$A_i \frac{dh_i}{dt} = q_{in,i} - q_{out,i} \tag{1}$$

$$q_{out,i} = a_i\sqrt{2\,g\,h_i} \tag{2}$$

where A_i is the cross-section of tank i, h_i its water level, $q_{in,i}$ and $q_{out,i}$ its inflow and outflow, a_i the cross-section of its outlet hole and g the gravitational acceleration. The valve ratios γ_1, $\gamma_2 \in (0,1)$ split the pump flows:

$$q_{in,1} = \gamma_1 k_1 v_1, \qquad q_{in,4} = (1-\gamma_1)\,k_1 v_1 \tag{3a}$$

$$q_{in,2} = \gamma_2 k_2 v_2, \qquad q_{in,3} = (1-\gamma_2)\,k_2 v_2 \tag{3b}$$

with pump constants k_1, k_2. Tank 3 and Tank 4 have no inflow from other tanks, and the complete model is

$$\frac{dh_1}{dt} = -\frac{a_1}{A_1}\sqrt{2gh_1} + \frac{a_3}{A_1}\sqrt{2gh_3} + \frac{\gamma_1 k_1}{A_1} v_1 \tag{4a}$$

$$\frac{dh_2}{dt} = -\frac{a_2}{A_2}\sqrt{2gh_2} + \frac{a_4}{A_2}\sqrt{2gh_4} + \frac{\gamma_2 k_2}{A_2} v_2 \tag{4b}$$

$$\frac{dh_3}{dt} = -\frac{a_3}{A_3}\sqrt{2gh_3} + \frac{(1-\gamma_2)k_2}{A_3} v_2 \tag{4c}$$

$$\frac{dh_4}{dt} = -\frac{a_4}{A_4}\sqrt{2gh_4} + \frac{(1-\gamma_1)k_1}{A_4} v_1 \tag{4d}$$

The square-root terms make the model nonlinear. All simulations in this paper are carried out on this nonlinear model; the linearized model below is used for controller design.

Linearized Model and Multivariable Zero

Linearizing the model around an operating point (h_{io}, v_{io}) with the Taylor series and neglecting higher-order terms gives a linear state-space model. Its transfer matrix from the pump voltages to the measured levels has the elements

$$G_{11}(s) = \frac{\gamma_1 c_1}{1+sT_1}, \qquad G_{12}(s) = \frac{(1-\gamma_2)\, c_1}{(1+sT_3)(1+sT_1)} \tag{5a}$$

$$G_{21}(s) = \frac{(1-\gamma_1)\, c_2}{(1+sT_4)(1+sT_2)}, \qquad G_{22}(s) = \frac{\gamma_2 c_2}{1+sT_2} \tag{5b}$$

where s is the Laplace variable, $c_i = T_i k_i k_c / A_i$ and $T_i = (A_i/a_i)\sqrt{(2h_{io}/g)}$, $i = 1,\ldots,4$, are the time constants. The zeros of the transfer matrix follow from $(1+sT_3)(1+sT_4) = \eta$, with

$$\eta = \frac{(1-\gamma_1)(1-\gamma_2)}{\gamma_1 \gamma_2} \tag{6}$$

One zero always lies in the left half-plane. The second zero is in the left half-plane for $1 < \gamma_1+\gamma_2 < 2$ (minimum-phase), at the origin for $\gamma_1+\gamma_2 = 1$ and in the right half-plane for $0 < \gamma_1+\gamma_2 < 1$ (non-minimum-phase) (Johansson, 2000). The minimum-phase operating point of Table 1 is used here; the corresponding transfer functions are given in Table 2.

Table 1. Process parameters and minimum-phase operating point.

Parameter	**Value**
Height of tanks, h_{max}	20 cm
Bottom area, Tank 1 and 3, A_1, A_3	28 cm²
Bottom area, Tank 2 and 4, A_2, A_4	32 cm²
Outlet hole area, a_1, a_3	0.071 cm²
Outlet hole area, a_2, a_4	0.057 cm²
Measurement constant, k_c	0.5 V/cm
Gravity, g	981 cm/s²
(h_{1o}, h_{2o})	(12.26, 12.78) cm
(h_{3o}, h_{4o})	(1.63, 1.41) cm
(v_1, v_2)	(3.00, 3.00) V
(k_1, k_2)	(3.33, 3.35) cm³/(V·s)
(γ_1, γ_2)	(0.70, 0.60)

h_{io}: operating-point level of tank i; v_i: pump voltage; k_i: pump constant; γ_i: valve ratio.

Table 2. Transfer matrix at the minimum-phase operating point.

Element	**Transfer function**
$G_{11}(s)$	2.6 / (1 + 62.3s)

Element	Transfer function
$G_{12}(s)$	1.48 / [(1 + 22.8s)(1 + 62.3s)]
$G_{21}(s)$	1.4 / [(1 + 30s)(1 + 90.6s)]
$G_{22}(s)$	2.84 / (1 + 90.6s)

$G_{ij}(s)$: transfer function from pump voltage *j* to measured level *i*; time in seconds.

Relative Gain Array

A decentralized scheme requires a choice of input-output pairs. The relative gain array (RGA) $\Lambda = G(0) \circ (G(0)^{-1})^{T}$, where $\circ$ denotes element-wise multiplication, suggests pairings with diagonal entries close to one; negative diagonal entries indicate a difficult control problem (Johansson, 2000). For the QTP the RGA depends only on the valve settings:

$$\lambda = \frac{\gamma_1 \gamma_2}{\gamma_1 + \gamma_2 - 1} \tag{7}$$

with λ denoting the diagonal entries of Λ. For $\gamma_1 = 0.70$ and $\gamma_2 = 0.60$, $\lambda = 1.4$, so Λ has diagonal entries 1.4 and off-diagonal entries −0.4. The diagonal pairing (y_1, u_1) and (y_2, u_2) is therefore selected, and the interaction is moderate.

Methods

The pump voltages are limited to 0–10 V, and all controllers act on deviations from the operating-point voltages v_{10} and v_{20}. The reference for Tank 1 is stepped from 12.26 cm to 14.5 cm at 50 s, while the Tank 2 reference is kept constant at 12.78 cm, which makes the coupling effect on Tank 2 easy to see. The simulation lasts 500 s.

Decentralized PI Control

Two independent PI controllers are used for the pairs selected by the RGA analysis:

$$G_{ci}(s) = K_i \left(1 + \frac{1}{T_i s}\right), \qquad i = 1,2 \tag{8}$$

where K_i is the proportional gain and T_i the integral time of loop *i*. The settings (K_1, T_1) = (3.0, 30 s) and (K_2, T_2) = (2.7, 40 s) were tuned on the simulation for acceptable transient behavior.

Decoupler

To remove the interaction, a dynamic decoupler is placed between the PI controllers and the plant (Figure 2). An ideal decoupler has unit diagonal elements, and its off-diagonal elements are

$$T_{12}(s) = -\frac{G_{12}(s)}{G_{11}(s)}, \qquad T_{21}(s) = -\frac{G_{21}(s)}{G_{22}(s)} \tag{9}$$

which with Table 2 gives $T_{12}(s) \approx -0.57/(1 + 22.8s)$ and $T_{21}(s) \approx -0.5/(1 + 30.07s)$. The controller of loop 2 acts on Tank 1 through T_{12} and the controller of loop 1 acts on Tank 2 through T_{21}, so that, for the linear model, each PI controller sees only its own loop. The PI settings are those of the decentralized PI control.

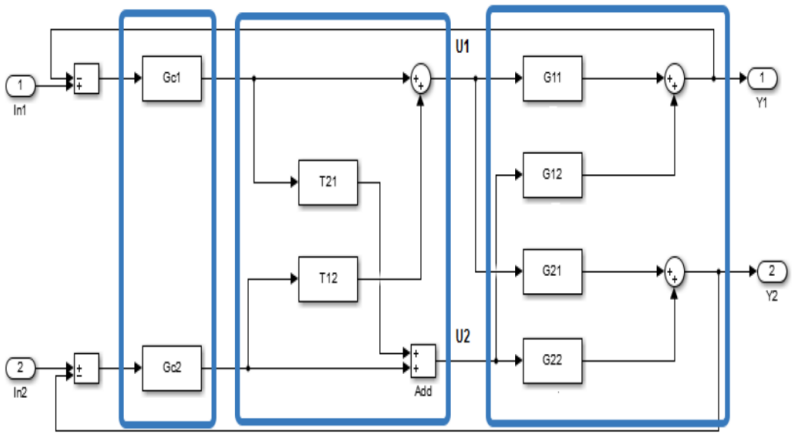


Figure 2. block diagram of the control system with the decoupler (controller part, decoupler part and physical system).

Model Predictive Control

At each sampling instant, MPC predicts the future outputs with a model and chooses the control moves that minimize a cost function while respecting constraints (Camacho and Bordons, 2007). The controlled variables are the levels of Tank 1 and Tank 2 and the manipulated variables are the two pump voltages. With prediction horizon N_p, control horizon N_c and weighting matrices Q and R, the cost is

$$J = \sum_{i=1}^{N_p} \|\hat{y}(k+i|k) - r(k+i)\|_Q^2 + \sum_{j=0}^{N_c-1} \|\Delta u(k+j)\|_R^2 \qquad (10)$$

where k is the current sampling instant, $\hat{y}$ the predicted output, r the reference and Δu the control increment. The controller was designed with the MATLAB Model Predictive Control Toolbox on the linearized model and tuned with the MPC Designer, using $N_p = 15$, $N_c = 3$ and $T_s = 0.1$ s. It was then connected to the nonlinear plant in Simulink.

Actuator Faults

Five actuator fault types are injected into the pump channels after the reference step. With u the commanded and u_f the delivered actuator signal, and t_f the fault instant:

(i) Loss of effectiveness: the actuator gain drops to 50 % of its nominal value, applied to the Tank 1 loop at $t_f = 100$ s and to the Tank 2 loop at $t_f = 200$ s. (ii) Leakage: a periodically varying loss of actuator flow (sinusoidal gain signal with a period of about 62 s), applied to the Tank 1 loop at 100 s and to the Tank 2 loop at 200 s. (iii) Lock-in-place: at $t_f = 100$ s the actuator of pump 1 freezes at its current value, $u_f(t) = u(t_f)$ for $t \geq t_f$. (iv) Hard-over: at $t_f = 100$ s the actuator of pump 1 jumps to its maximum output (10 V, about 33 cm³/s) and stays there, independent of the command. (v) Stuck-open: at $t_f = 100$ s the actuator of pump 1 sticks in the open position.

Performance Index

Performance is measured by the mean squared error (MSE) of each tank level over the whole simulation:

$$\mathrm{MSE}_i = \frac{1}{N}\sum_{k=1}^{N}\left(r_i(k) - y_i(k)\right)^2 \qquad (11)$$

in cm², N being the number of samples, r_i the reference and y_i the level of Tank i, $i = 1, 2$.

RESULTS

Nominal Operation

Figure 3 shows the tank levels without faults. All three controllers bring Tank 1 to its new reference without steady-state error. The decentralized PI and the decoupled PI controllers reach a peak of about 15.3 cm after the 2.24 cm step, whereas the MPC shows no visible overshoot. Because the levels are coupled, the step in Tank 1 disturbs Tank 2: with the PI controller the Tank 2 level moves by about 0.05 cm around its reference, with the decoupler the excursion is similar in size but shorter, and with MPC it is negligible. The decoupler is designed on the linear model and used with the nonlinear plant, so the decoupling is not perfect. Table 3 confirms the ranking: MPC has the lowest Tank 1 MSE (0.014), followed by the decoupled PI (0.0187) and the decentralized PI (0.036).

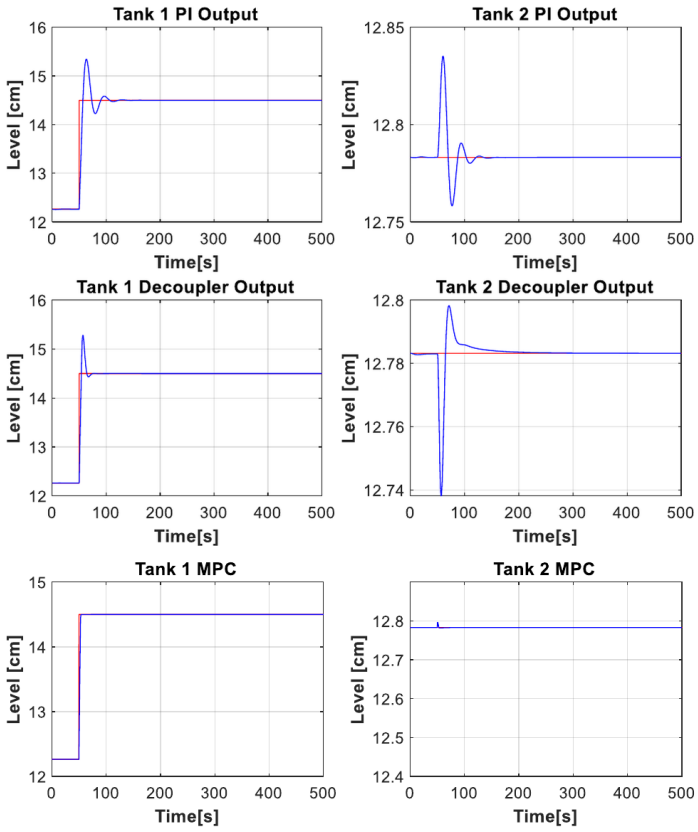


Figure 3. tank levels without fault: decentralized PI (top row), PI with decoupler (middle row) and MPC (bottom row); red: reference, blue: level.

Actuator Faults

The MSE values of all scenarios are listed in Table 3, and Figures 4–7 show the corresponding responses.

Table 3. Mean squared error of the controllers for each scenario.

Scenario	Tank	PI	Decoupler	MPC
No fault	1	0.036	0.0187	0.014
	2	0.001	0.000	0.000
Loss of effectiveness	1	0.0526	0.0208	0.0154
	2	0.0020	0.0023	0.0018
Leakage	1	0.0380	0.0196	0.0141
	2	0.0002	0.002	0.0000
Lock-in-place	1	0.1753	0.0197	0.0142
	2	0.001	0.0002	0.0000
Hard-over	1	23.7722	23.7495	23.7395
	2	0.8977	24.8063	31.3508
Stuck-open	1	0.0375	0.0195	0.0141
	2	0.001	0.0000	0.0000

PI: decentralized PI; Decoupler: PI with decoupler; MPC: model predictive control. MSE values in cm^2.

Loss of Effectiveness

The 50 % gain drop at 100 s mainly excites oscillations in the Tank 1 response of the PI controller (MSE 0.0526 against 0.036 without fault), whereas the decoupler and the MPC keep the level close to the reference (0.0208 and 0.0154) (Figure 4). When the Tank 2 loop is hit at 200 s, the level under the two PI-type controllers falls to about 12.47 cm before integral action restores it, while under MPC the deviation stays around 0.1 cm and recovers more slowly. The Tank 2 MSE values are all small and close (0.0018–0.0023).

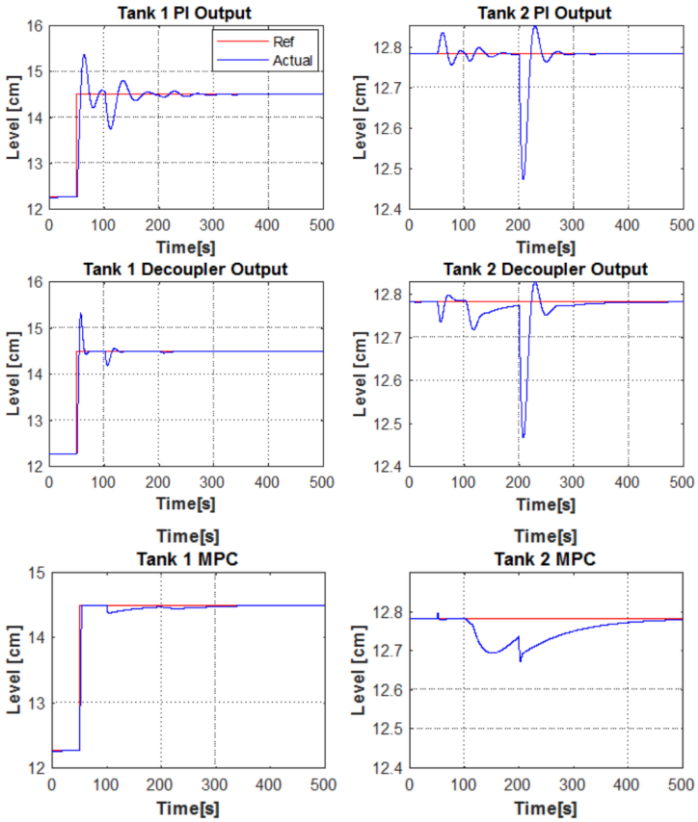


Figure 4. tank levels with loss of effectiveness (Tank 1 loop at 100 s, Tank 2 loop at 200 s); same layout as Figure 3.

Leakage

The Tank 1 level stays near the reference for all controllers, and the Tank 1 MSE is dominated by the step transient (0.0380, 0.0196, 0.0141, close to the nominal values) (Figure 5). The fault appears as a sustained oscillation of the Tank 2 level after 200 s: about ±0.02 cm for the PI-type controllers and clearly smaller for MPC. The Tank 2 MSE values differ very little (0.0002, 0.002 and 0.0000).

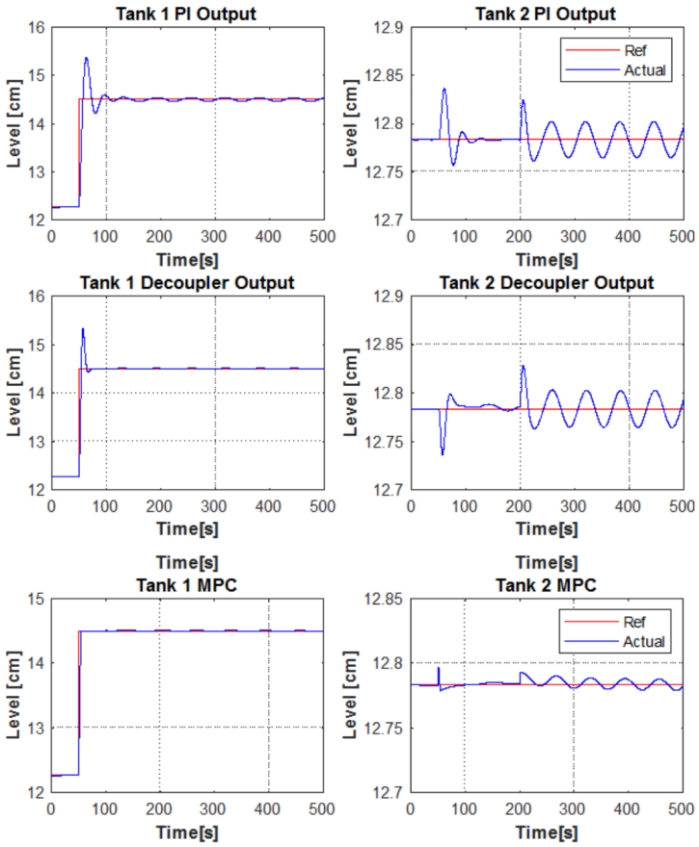


Figure 5. tank levels with leakage fault; same layout as Figure 3.

Lock-in-place

Here the controllers differ clearly (Figure 6). The decentralized PI controller can act on Tank 1 only through the frozen pump and ends with a steady-state offset of about 0.45 cm (level near 14.05 cm), which raises its Tank 1 MSE to 0.1753, about five times the nominal value. The decoupler and the MPC restore the reference (0.0197 and 0.0142) because they also act on pump 2, which reaches Tank 1 through Tank 3, at the cost of a slow deviation of Tank 2 of about 0.02 cm.

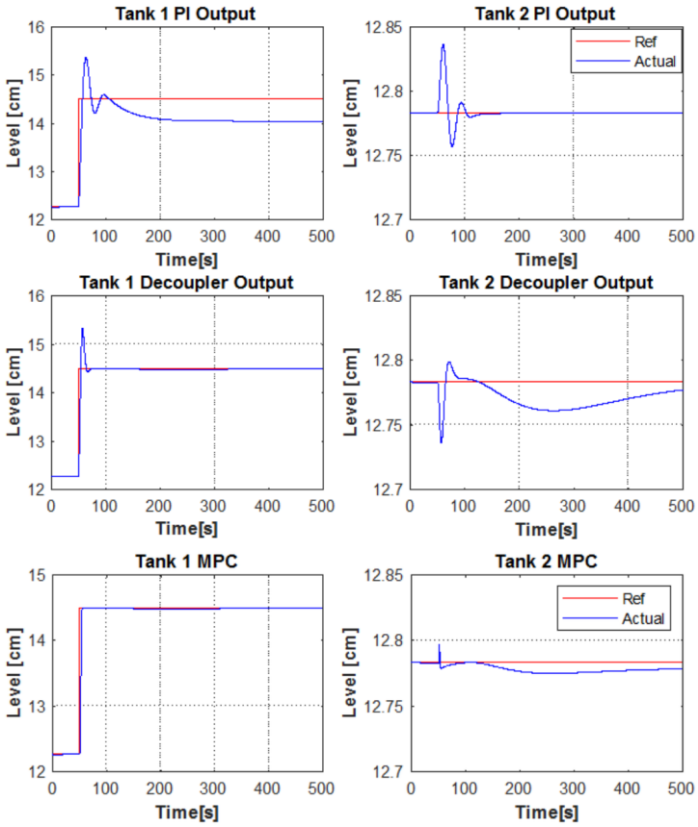


Figure 6. tank levels with lock-in-place fault of pump 1 at 100 s; same layout as Figure 3.

Hard-over

When pump 1 saturates at its maximum output, the Tank 1 level rises to the tank height of 20 cm for all controllers, and the Tank 1 MSE is practically identical (23.74–23.77) (Figure 7). The fault is beyond the authority of these controllers, since the faulty actuator no longer responds to commands. Tank 2 rises as well, but at different rates: slowly with the PI controller (about 15 cm at 500 s, MSE 0.8977), and considerably faster with the decoupler (about 18.7 cm, 24.8063) and the MPC (20 cm, 31.3508). A possible reason is that these two controllers act on both pumps and therefore transmit the fault to the second loop, but this was not analyzed further.

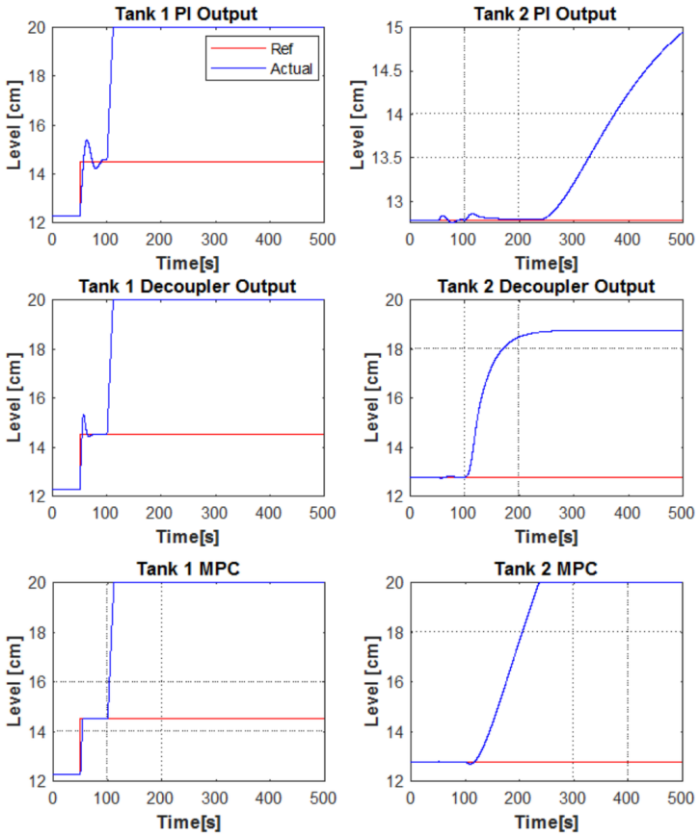


Figure 7. tank levels with hard-over fault of pump 1 at 100 s; same layout as Figure 3.

Stuck-open

The responses and MSE values are close to the fault-free case (Table 3: 0.0375, 0.0195 and 0.0141 for Tank 1). In this simulation the valve stuck near the required steady-state position, so this scenario hardly discriminates between the controllers, and its time responses are omitted.

DISCUSSION AND CONCLUSION

In this study the quadruple-tank process was modelled nonlinearly and linearized at a minimum-phase point, and three controllers were designed: a decentralized PI, a PI with decoupler and an MPC. Without faults, MPC gave the fastest response and the lowest MSE, and the decoupler improved on the plain decentralized PI controller. This points in the same direction as the advantages of MPC over PID-type control reported for the QTP (Kasiyanto et al., 2024; Christensen et al., 2025).

Under loss of effectiveness, leakage and lock-in-place faults, MPC kept the lowest Tank 1 error, and the controllers that act on both pumps (decoupler and MPC) were able to compensate a frozen pump, which the decentralized PI controller could not. The hard-over fault could not be handled by any of the controllers, which shows the limit of passive robustness without fault detection and reconfiguration.

The results come from simulation at a single minimum-phase operating point, without measurement noise and without fault detection. The MPC needs an accurate model and an online optimization, and its fast Tank 1 response involves a large but short control action at the reference step. Christensen et al.

(2025) found that the advantage of MPC depends strongly on whether future setpoints are available; whether this affected the present MPC was not examined, and tuning of the three controllers was not optimized on a common basis. Future work includes non-minimum-phase operation, adaptive and active fault-tolerant schemes with fault detection, and real-time implementation on the laboratory process.

References

Arıcı M, Kara T, 2016. Adaptive fault tolerant control for liquid tank process. International Journal of Applied Mathematics, Electronics and Computers, 4: 111-117.

Arıcı M, Kara T, 2018. Improved adaptive fault-tolerant control for a quadruple-tank process with actuator faults. Industrial & Engineering Chemistry Research, 57 (29): 9537-9553.

Camacho EF, Bordons C, 2007. Model predictive control. 2nd ed., Springer, London, UK.

Christensen AHD, Ritschel TKS, Svensen JL, Hørsholt S, Huusom JK, Jørgensen JB, 2025. Comparing model-based control strategies for a quadruple tank system: Decentralized PID, LMPC, and NMPC. arXiv preprint, arXiv:2509.11235.

Effendi AKR, Rameli M, Iskandar E, Baihaqi M, 2017. Linearization and decoupling controller for quadruple tank. International Seminar on Intelligent Technology and Its Applications (ISITIA), 2017, Page: 233-237.

Ferrese F, Dong Q, Biswas S, Batcho J, 2014. Decentralized control of coupled nonlinear dynamic systems with application to quadruple-tank process. 40th Annual Conference of the IEEE Industrial Electronics Society (IECON), 2014, Page: 3657-3661.

Hajiani P, Poshtan J, 2014. Abrupt and incipient fault detection and compensation for a 4-tank system benchmark. Turkish Journal of Electrical Engineering & Computer Sciences, 22: 1287-1297.

Jiang J, Yu X, 2012. Fault-tolerant control systems: A comparative study between active and passive approaches. Annual Reviews in Control, 36: 60-72.

Johansson KH, 2000. The quadruple-tank process: A multivariable laboratory process with an adjustable zero. IEEE Transactions on Control Systems Technology, 8 (3): 456-465.

Kasiyanto I, Firdaus H, Lailiyah Q, Kusnandar N, Supono I, 2024. Advanced control for quadruple tank process. Jurnal Ilmiah Teknik Elektro Komputer dan Informatika, 10 (1): 1-16.

Nordfeldt P, Hägglund T, 2006. Decoupler and PID controller design of TITO systems. Journal of Process Control, 16: 923-936.

Özkan S, 2019. Modelling, simulation and control of quadruple tank process, Master Thesis. Gaziantep University, Graduate School of Natural and Applied Sciences, Gaziantep, Turkey.

Özkan S, Kara T, Arıcı M, 2017. Modelling, simulation and control of quadruple tank process. 10th International Conference on Electrical and Electronics Engineering (ELECO), 2017, Page: 866-870, Bursa, Turkey.

Patel HR, Shah VA, 2019. Decentralized stable and robust fault-tolerant PI plus fuzzy control of MIMO systems: A quadruple tank case study. International Journal on Smart Sensing and Intelligent Systems, 12 (1): 1-20.

Zhang Y, Jiang J, 2008. Bibliographical review on reconfigurable fault-tolerant control systems. Annual Reviews in Control, 32: 229-252.

Zong X, Yang Z, Yuan D, 2010. Nonlinear modeling and predictive control of the four-tank system. 2nd International Conference on Software Engineering and Data Mining, 2010, Page: 268-271.

ACKNOWLEDGMENT

This paper builds on the author's M.Sc. thesis and on work carried out within the project supported by The Scientific and Technological Research Council of Turkey (TÜBİTAK) under grant 116E020. No additional funding was received for the present paper.

CONFLICT OF INTEREST

The author has declared that there is no conflict of interest.

AUTHOR CONTRIBUTIONS

Serkan Özkan: conceptualization, modelling, controller design, simulation, analysis and writing.

The Role Of Personality–Emotional Intelligence Alignment In Enhancing Mentoring Effectiveness

Gulnara Duruskari[1], Elnara Mahmudlu[2]*, Khalil Nagiyev[3]

[1,2,3]Baku State University, Faculty of Education, AZ1148, Baku, Azerbaijan

Corresponding author e-mail: mahmudluelnare3@gmail.com

Abstract

Mentoring is widely recognized as a catalyst for professional development and psychosocial well-being; yet its effectiveness varies substantially across structurally identical dyads. This systematic narrative review synthesizes ten empirical studies to examine how alignment between mentor personality profiles (Five-Factor Model), mentoring style, and mentee dispositional needs determines mentoring outcomes. Findings reveal a functional bifurcation: Openness to Experience, Conscientiousness, and Extraversion significantly predict mentoring quality, while Agreeableness and Emotional Stability predict mentee satisfaction. Importantly, personality concordance within dyads predicts outcome quality more reliably than absolute trait levels, supporting a compatibility-based approach. Emotional intelligence (EI) functions as the key mediating mechanism translating personality compatibility into effective relational practice. Peer mentoring consistently outperforms hierarchical models, with 93% retention rates in near-peer STEM programmes versus 33% for comparable non-mentored students. A Personality-Aligned Mentoring Model (PAMM) is proposed, integrating personality-based pre-match assessment, EI-informed mentor preparation, and peer community infrastructure. These findings challenge assumptions that structural features alone determine mentoring success and redirect attention toward relational mechanisms within dyads.

Key words: *Personality-aligned mentoring, emotional intelligence, five-factor model, mentor-mentee matching, peer mentoring*

INTRODUCTION

Mentoring has occupied a central position in human development across professional, educational, and clinical domains for decades (Hill et al., 2022). Broadly defined, it is a purposeful, sustained relationship in which a more experienced individual provides guidance, support, and knowledge transfer to facilitate the growth of a less experienced counterpart. The documented benefits of effective mentoring are extensive: higher job satisfaction, enhanced career advancement, improved academic performance, stronger organisational commitment, greater retention, and measurable gains in psychosocial well-being have all been reported across diverse populations and sectors (Kumari et al., 2022; Mullen and Klimaitis, 2021).

Despite this breadth of evidence, a foundational paradox persists: structurally identical mentoring programmes, implemented within the same institution and serving comparable populations, frequently yield dramatically different outcomes for different mentor–mentee dyads (Spencer et al., 2023).

Understanding the sources of this variability has direct consequences for how organisations design, resource, and evaluate mentoring investments. Current literature offers several candidate explanations for mentoring variability, including programme structure, meeting frequency, goal clarity, and demographic similarity between mentor and mentee (Deng and Turner, 2024). However, a growing body of evidence points to a more fundamental and under-theorised mechanism: the alignment between the mentor's personality profile, the mentoring style they adopt, and the dispositional needs of the mentee (Younginer and Elledge, 2021).

The Five-Factor Model (FFM) of personality — comprising Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism — has emerged as the most robust and cross-culturally validated framework for characterising individual differences in personality (Cavell et al., 2020). Its application to mentoring has been uneven: most FFM-mentoring research has examined mentor traits in isolation, without attending to the matching logic that dyadic effectiveness requires (Henry-Noel et al., 2019). Equally relevant, yet similarly under-integrated, is the construct of emotional intelligence (EI). Defined as the capacity to perceive, generate, understand, and regulate emotions in oneself and others, EI encompasses precisely those competencies that determine whether a personality-compatible dyad can translate its potential into sustained relational quality (Gómez-Leal et al., 2022).

Recent evidence suggests that EI is both a predictor of mentoring effectiveness and an outcome of quality mentoring participation, creating a bidirectional relationship that has significant implications for programme design (Prummer et al., 2024; MacCann et al., 2020). The intersection of FFM personality, EI, and mentoring style compatibility has not, to date, been systematically theorised or empirically synthesised. This article addresses that gap through a systematic narrative review spanning higher education, clinical medicine, corporate organisations, community youth mentoring, and STEM contexts (Zaniewski and Reinholz, 2016; Seery et al., 2021). These objectives converge in the proposal of a Personality-Aligned Mentoring Model (PAMM).

MATERIAL AND METHODS

Material

The research material for this systematic narrative review comprises ten peer-reviewed empirical studies published between 2016 and 2024 that examine the relationship between mentor personality profiles, emotional intelligence, and mentoring effectiveness across diverse organizational and educational contexts (Kumari et al., 2022). Studies were sourced from PsycINFO, Web of Science, ERIC, and Google Scholar using keyword combinations including "mentoring", "personality", "emotional intelligence", "Five-Factor Model", and "peer mentoring" (Mullen and Klimaitis, 2021). Selection criteria required empirical data collection, publication in peer-reviewed venues, and direct relevance to at least one of the three focal constructs: FFM personality, EI, or mentoring outcomes (Prummer et al., 2024).

The included studies examined various research contexts and populations. Studies were conducted across five countries: the United States, Pakistan, India, Ireland, and South Africa, covering sectors including higher education, vocational education and training, clinical medicine, corporate organisations, community youth mentoring, and STEM undergraduate education (Zaniewski and Reinholz, 2016; Henry-Noel et al., 2019). The sample sizes of the included studies ranged from 14 mentoring dyads to 390 participants. The studies used different research designs and data collection methods, including surveys, experimental and quasi-experimental designs, longitudinal qualitative

analysis, mixed-methods research, and narrative review, selected based on established criteria for empirical research design, clarity of measurement instruments, and appropriateness of the methods used (Spencer et al., 2023).

Methods

The Collection of the Data

This study employs a systematic narrative review methodology to synthesise the available empirical evidence on personality–EI alignment and mentoring effectiveness (Cavell et al., 2020). A systematic narrative review is appropriate when the objective is to integrate findings from a heterogeneous body of literature using different methodologies and outcome measures, where meta-analysis is not suitable (Tuma and Dolan, 2024). The review included 10 empirical studies conducted across five countries as noted above. The studies employed various research methodologies appropriate to their research questions, including longitudinal designs, mixed-methods approaches, and experimental manipulations. Data extraction focused on three primary constructs: personality traits measured via the Five-Factor Model (Costa and McCrae, 1992, cited in Kumari et al., 2022), emotional intelligence dimensions, and mentoring outcome measures including relationship quality, mentee satisfaction, retention, and academic or professional performance (Younginer and Elledge, 2021).

Statistical Analysis

The findings of the included studies were analysed using a thematic and narrative synthesis approach (Gómez-Leal et al., 2022). Because the studies differed substantially in their research designs, samples, variables, and measurement methods, statistical results were not combined through a meta-analysis. Instead, the statistical and analytical methods used in the individual studies were examined and compared (MacCann et al., 2020). The included studies employed several statistical techniques, including structural equation modelling, longitudinal qualitative analysis, exploratory and confirmatory factor analysis, mediation analysis, and experimental and quasi-experimental analysis (Deng and Turner, 2024).

Each study was examined according to four main areas: personality traits and their directional relationship to outcomes; emotional intelligence constructs and their role in the mentoring process; structural features of the mentoring model; and key outcome measures (Seery et al., 2021). Quality assessment of included studies was conducted using criteria adapted from established checklists for qualitative and mixed-methods research, evaluating methodological rigour, clarity of outcome measurement, and appropriateness of analytical approaches (Hill et al., 2022). Thematic saturation was assessed across the ten studies, allowing common themes and patterns to be identified while maintaining differences between sectors and national contexts (Ayoobzadeh, 2023).

RESULTS

The systematic review and thematic synthesis of ten empirical studies yields four interconnected findings that collectively illuminate the personality–EI alignment mechanisms underpinning mentoring effectiveness. These findings provide robust empirical support for the Personality-Aligned Mentoring Model (PAMM) and generate specific implications for programme design, mentor selection, and relational skill development.

FFM Traits Differentially Predict Mentoring Behaviour and Mentee Outcomes: A Functional Bifurcation

Across the studies reviewed, a consistent and theoretically significant pattern emerged: the FFM traits that predict mentoring quality are not the same as those that predict mentee outcome satisfaction. Openness to Experience, Conscientiousness, and Extraversion significantly predict mentoring quality, while Agreeableness and Emotional Stability are stronger predictors of mentee satisfaction. This functional bifurcation has important implications for mentor selection, role design, and the interpretation of mentoring effectiveness data.

Table 1. FFM Functional Bifurcation: Differential Standardized Path Coefficients (β) for Mentoring Quality vs Mentee Job Satisfaction.

FFM Trait	Mentoring Quality (β)	Mentee Satisfaction (β)
Openness to Experience	0.225*	0.221*
Conscientiousness	0.368*	ns
Extraversion	0.256*	ns
Agreeableness	ns	0.241*
Emotional Stability	ns	0.269*

** $p < .05$; ns = not significant. Data from Kumari et al. (2022). CFI = 0.912; RMSEA = 0.039.*

Dyadic Personality Compatibility: Concordance as the Operative Matching Criterion

The evidence base provides compelling support for the proposition that dyadic compatibility — the degree of personality concordance between mentor and mentee — is a more powerful predictor of mentoring success than the absolute level of any individual trait. Personality concordance within dyads predicts perceived outcome quality more reliably than absolute trait levels. This finding shifts the analytical unit from the individual to the dyad, and from selection to matching.

Emotional Intelligence as the Mediating Mechanism

Emotional intelligence (EI) functions as the competence through which personality potential is translated into the empathic attunement, regulated emotional expression, and responsive communication that characterise effective mentoring. Mentoring participation produced measurable improvements in emotional intelligence competencies, particularly perspective-taking and emotional regulation. Trustworthiness, active listening, collaborative goal-setting, and constructive feedback — all EI-adjacent competencies — are the most consistently reported characteristics of effective mentors.

The Superiority of Peer Mentoring

Across the studies reviewed, peer mentoring consistently outperformed hierarchical dyadic mentoring on multiple outcome dimensions. Peer group mentoring received significantly higher mean ratings compared to individual professional mentoring. Near-peer mentoring programmes in STEM contexts reported 93% retention rates against a historical baseline of 33% for academically comparable non-

mentored students. The following table presents outcome comparison data:

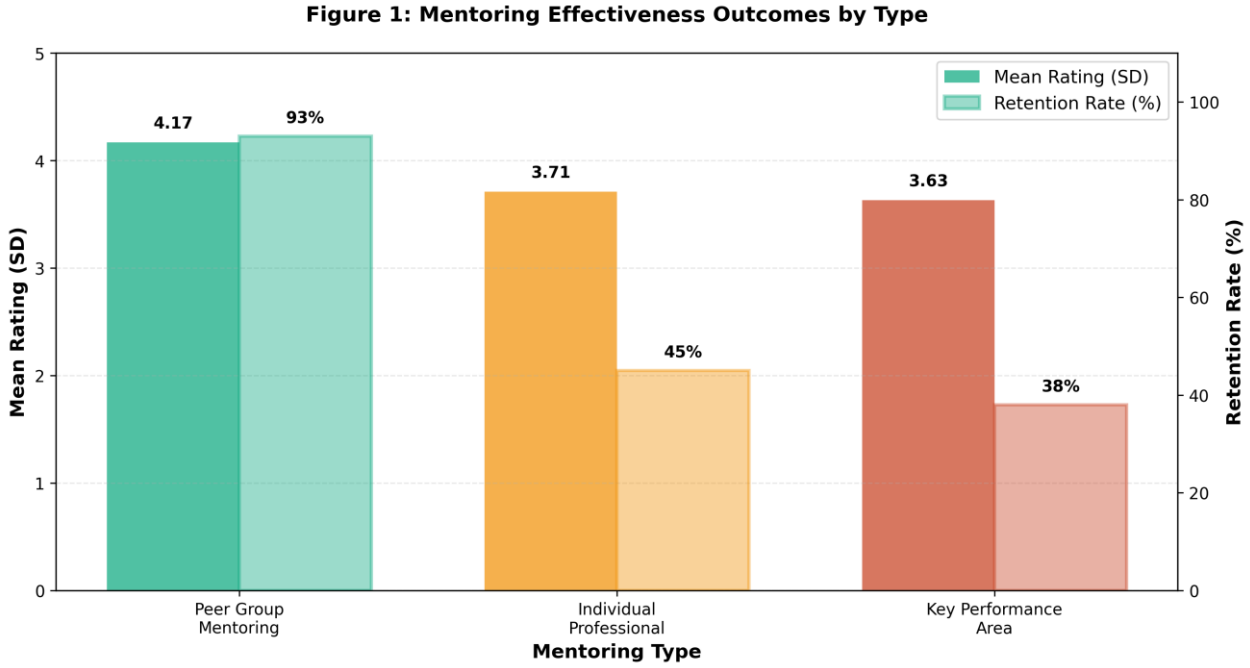


Figure 1. Mentoring Effectiveness Outcomes by Type.

Peer mentoring's effectiveness is sustained over time when programmes actively manage mentor wellbeing and expectation alignment. The superiority of peer mentoring reflects the combined benefits of personality-based compatibility, enhanced emotional intelligence development through reciprocal learning, and community infrastructure that supports relational quality.

DISCUSSION AND CONCLUSION

This article has argued that mentoring effectiveness is fundamentally contingent upon the alignment between mentor personality, mentoring style, and mentee dispositional needs, mediated by emotional intelligence (Prummer et al., 2024). Through systematic narrative review of ten empirical studies spanning higher education, clinical medicine, corporate organisations, community youth mentoring, and STEM contexts (Zaniewski and Reinholz, 2016; Seery et al., 2021), four convergent findings have been identified. The functional bifurcation of FFM trait effects across mentoring quality and mentee satisfaction (Kumari et al., 2022) demonstrates that mentor selection cannot rely on a single personality profile. Rather, different traits predict different outcomes: Openness to Experience, Conscientiousness, and Extraversion predict mentoring quality, while Agreeableness and Emotional Stability predict mentee satisfaction (Henry-Noel et al., 2019).

The primacy of dyadic personality concordance over absolute trait levels as a matching criterion represents a paradigm shift in mentoring programme design (Spencer et al., 2023). This finding reflects an important reorientation: mentoring effectiveness is not determined by recruiting mentors with idealized personality profiles, but rather by matching mentor and mentee personalities within dyads to maximize compatibility (Younginer and Elledge, 2021). The evidence indicates that personality

concordance within dyads predicts perceived outcome quality more reliably than the absolute level of any individual trait, supporting a compatibility-based rather than a trait-maximizing approach (Cavell et al., 2020). This has direct implications for mentoring programme operations: pre-match assessment instruments based on the Five-Factor Model, coupled with algorithmic or clinical matching logic, could substantially improve dyadic functioning without requiring changes to institutional structure or resource allocation (Deng and Turner, 2024).

References

Journal Articles

Hill, S.E.M.; Ward, W.L.; Seay, A.; Buzenski, J. The nature and evolution of the mentoring relationship in academic health centers, Journal of Clinical Psychology in Medical Settings, 2022, v.29(3), pp. 557–569.

Kumari, K.; Ali, S.B.; Batool, M.; Cioca, L.-I.; Abbas, J. The interplay between leaders' personality traits and mentoring quality and their impact on mentees' job satisfaction and job performance, Frontiers in Psychology, 2022, v.13, 937470.

Prummer, K.; Human-Vogel, S.; Graham, M.A.; Pittich, D. The role of mentoring in developing leaders' emotional intelligence: exploring mentoring types, emotional intelligence, organizational factors, and gender, Frontiers in Education, 2024, v.9, 1393660.

Zaniewski, A.M.; Reinholz, D. Increasing STEM success: a near-peer mentoring program in the physical sciences, International Journal of STEM Education, 2016, v.3(1), pp. 1–12.

Spencer, R.; Drew, A.L.; Gowdy, G. Going the distance: a longitudinal qualitative study of formal youth mentoring relationship development, Journal of Community Psychology, 2023, v.51(8), pp. 3083–3102.

Henry-Noel, N.; Bishop, M.; Gwede, C.K.; Petkova, E.; Szumacher, E. Personality compatibility within faculty mentoring dyads and perceived mentoring outcomes, Journal of Cancer Education, 2019, v.34(4), pp. 629–637.

Younginer, S.T.; Elledge, L.C. Mentor personality and attachment as correlates of youth mentoring relationship quality: the moderating role of negative interactions, Journal of Community Psychology, 2021, v.49(7), pp. 2569–2589.

Seery, C.; Andres, A.; Moore-Cherry, N.; O'Sullivan, S. Students as partners in peer mentoring: expectations, experiences and emotions, Innovative Higher Education, 2021, v.46(6), pp. 663–681.

Cavell, T.A.; Mutignani, L.M.; Alfonso, L.; Smith, A.M. Attachment tendencies, Big Five personality traits, and self-efficacy as predictors of mentors' relationships with aggressive children, American Journal of Community Psychology, 2020, v.66(1–2), pp. 130–143.

Deng, C.; Turner, N. Identifying key mentor characteristics for successful workplace mentoring relationships and programmes, Personnel Review, 2024, v.53(2), pp. 580–604.

Tuma, T.T.; Dolan, E.L. What makes a good match? Predictors of quality mentorship among doctoral students, CBE—Life Sciences Education, 2024, v.23(3), ar33.

Ayoobzadeh, M. Peer-mentoring and doctoral student retention: a longitudinal investigation, Higher Education Research & Development, 2023, v.42(3), pp. 513–529.

MacCann, C.; Jiang, Y.; Brown, L.E.R.; Double, K.S.; Bucich, M.; Minbashian, A. Emotional intelligence predicts academic performance: a meta-analysis, Psychological Bulletin, 2020, v.146(2), pp. 150–186.

Quílez-Robres, A.; Usán, P.; Lozano-Blasco, R.; Salavera, C. Emotional intelligence and academic performance: a systematic review and meta-analysis, Thinking Skills and Creativity, 2023, v.49, 101355.

Lim, S.; Xin Min, L.; Chan, C.J.W.; Dong, Y.; Mikkonen, K.; Zhou, W. Peer mentoring programs for nursing students: a mixed methods systematic review, Nurse Education Today, 2022, v.119, 105577.

Venegas-Muggli, J.I.; Barrientos, C.; Álvarez, F. The impact of peer-mentoring on the academic success of underrepresented college students, Journal of College Student Retention: Research, Theory & Practice, 2021, v.25(3), pp. 554–571.

Gómez-Leal, R.; Holzer, A.A.; Bradley, C.; Fernández-Berrocal, P.; Patti, J. The relationship between emotional intelligence and leadership in school leaders: a systematic review, Cambridge Journal of Education, 2022, v.52(1), pp. 1–21.

Proceedings and Book Chapters

Sebastian, R.T.; George, V.J.; Mathew, G.A.; Brinda, G. Personality traits influence on mentoring relationships: an application of Big Five Personality Model among Indian IT professionals. In: Khamis, R.; Buallay, A. (Eds.), Proceedings of the International Conference on Business and Technology, Springer, Cham, 2024, pp. 95–108.

Boutakidis, I.; Espinoza, G.; Sevier, M.; Sadek, A. The impact of a peer mentoring program on undergraduate graduation rates: a matched control group design, Journal of College Student Retention: Research, Theory & Practice, 2024.

Mullen, C.A.; Klimaitis, C.C. Defining mentoring: a literature review of issues, types, and applications, Annals of the New York Academy of Sciences, 2021, v.1483(1), pp. 19–35.

Acknowledgment

The authors acknowledge the support of Baku State University in providing access to research databases and institutional facilities. We thank the peer reviewers for their constructive feedback during the review process.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript. The research was conducted independently without external funding or competing financial interests.

Author Contributions

GD: Conceptualization, methodology, literature review, manuscript preparation. EM: Literature review, data synthesis, critical revision of manuscript. KN: Analytical framework development, theoretical interpretation, critical revision of manuscript. All authors have read and approved the final version of the manuscript.

Camera-Based Parking Occupancy Detection: A Comparison of YOLOv8n, YOLO11n, RT-DETR-L, and Faster R-CNN

Beyza Sönmez[1]*, Murat Taşyürek[2]

[1,2] Department of Computer Engineering, Faculty of Engineering and Design, Kayseri University, Kayseri, Türkiye

[2]Head of the Department, Faculty of Engineering and Design, Kayseri University, Kayseri, Türkiye

*Corresponding author e-mail: beyza.sonmez.office@gmail.com

Abstract

Parking occupancy is often measured with a dedicated sensor in every stall. One fixed camera lowers that cost. Daylight, night lighting, rain, and a vehicle that overlaps the next stall can change the decision. This study compares four off-the-shelf detectors on fixed-camera images from two real sites: YOLOv8n, YOLO11n, RT-DETR-L, and Faster R-CNN. No new model was trained. A stall counts as occupied when the intersection of a car, truck, or bus box with the stall, divided by the stall area, is at least 40 percent. A vehicle passing on the road was not counted as parked. Thirty-six frames were taken from each of four recordings. The set contains 144 frames and 1,224 stall decisions. These decisions are not independent samples. At the 40 percent coverage threshold the F1 scores are 95.8 percent for YOLOv8n, 97.7 percent for YOLO11n, 96.8 percent for RT-DETR-L, and 97.7 percent for Faster R-CNN. Mean time per frame on the stated CPU is 39, 50, 636, and 2276 milliseconds. On the nighttime curbside recordings, YOLO11n and Faster R-CNN missed fewer occupied stalls than YOLOv8n. At a 30 percent coverage threshold the ranking changes: RT-DETR-L reaches the highest F1, 99.2 percent. The result is limited to these two sites and four recordings.

Keywords: *Parking Occupancy, Object Detection, YOLOv8n, YOLO11n, RT-DETR-L, Faster R-CNN*

INTRODUCTION

Searching for an empty parking stall wastes time and fuel. A magnetic or ultrasonic sensor in every stall gives a direct measurement, but the hardware cost grows with the number of stalls. One fixed camera can watch many stalls at once. This experiment uses one camera that stayed fixed during each recording, at two sites. The camera decision is less stable than a sensor: light changes, rain wets the ground, a parked van overlaps the next stall, and a car driving past the curb can look occupied for a single frame.

Camera-based occupancy detection has been studied on public sets much larger than these 144

frames. PKLot and CNRPark are examples [5], [6]. The methods fall into three groups. A small network classifies a cropped stall [6], [9], [15]. Convolutional features are combined with a support - vector machine [10], [11]. A camera network is placed at the edge of the lot [14]. On the detector side, Faster R-CNN is a two-stage detector and is used here with a ResNet-50 backbone [13], [8]. DETR introduced transformer-based detection, and RT-DETR-L is a real-time model in that family [12], [3]. The weights are the public COCO weights [4]. A shared limitation is that consecutive frames often contain the same parked cars, so a large decision count can overstate the strength of the evidence [7]. This paper does not develop a new model. It compares four off-the-shelf detectors on the same fixed-camera images for stall occupancy. The comparison is limited to two sites, fixed stall areas defined in advance for each site, one coverage rule, and one confidence threshold. A two -site experiment does not support a general claim of superiority.

MATERIAL AND METHODS

Four recordings, taken at different times of day at two sites, were evaluated. The phone was fixed during each recording. The four models were applied to the same recorded frames, so the camera angle does not change between models. The stall rectangles were drawn by hand, once, on an image of each site. Vehicles cut by the image border and the traffic lane were left outside the rectangles. The same rectangles were then used, unchanged, for every model.

Coverage is the intersection area of the detection box and the stall, divided by the stall area. If B is the detection box and P is the stall, coverage = $|B \cap P| / |P|$. The stall is occupied when coverage is at least 0.40. A vehicle on the road, outside the rectangle, is not parked. The 40 percent threshold is an experimental choice. It is not taken from a cited paper. A lower share lets the edge of the next car mark a stall occupied. A higher share misses a vehicle that sits inside the stall but does not fill the rectangle.

The occupied and empty reference labels are independent of the model predictions. Each label was given by inspecting the image. One person, the author, prepared the labels. Every stall crop was inspected in time order. A stall with a parked car, truck, or bus was marked occupied. Pedestrians, scooters, and vehicles passing through the rectangle were marked empty. The author then rechecked six borderline examples without seeing the first label. Five decisions matched. The sixth was a daytime stall with a person standing at an open door. It had been marked occupied and was changed to empty. A second person did not relabel the full set.

Six stalls were drawn at the curbside site. At the marked lot, eleven stalls that are inside the frame and not hidden by the central tree were drawn. The lot score is the score of those visible stalls. Tree-covered areas were left out because they could not be labeled. The night recording at Site 2 was also left out: the lot has no lighting, and after about eight minutes the image is too dark to label. Both exclusions limit the result. The numbers below do not include the two-second state confirmation used in an earlier trial. This experiment measures the detector decision on its own.

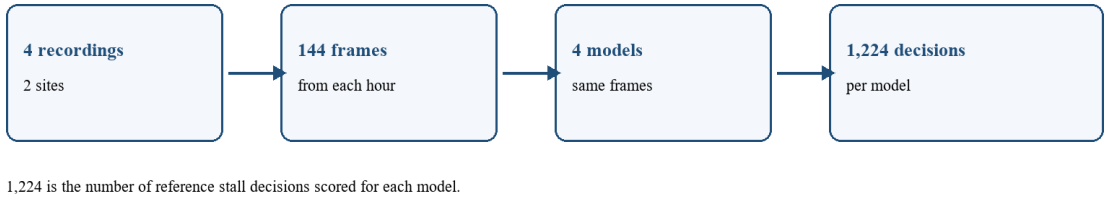


Figure 1. Four recordings, 144 frames, and four models. 1,224 is the number of reference stall decisions scored for each model.

Data

The recordings are from 26 and 27 September 2026. The resolution is 1920×1080 at about 60 frames per second. The phone did not move during a recording. Thirty-six frames were taken from each recording after dropping the first and last 30 seconds. The selection does not depend on the model. At Site 1 by day the frames run from 30.0 s to 3679.5 s with a step of 104.3 s. At Site 1 by night the step is 103.8 s. At Site 2 in the morning the step is 121.5 s. At Site 2 at noon, 30 frames come from the first 72 percent of the kept interval and the last 6 frames come from the rainy tail. The steps in that recording are not equal. The longest gap is 398.7 s.

Site	Condition	Time	Length	Frames	Note
Site 1	Day	26 Sep, 13:17	62 min	36	Curbside
Site 1	Night	26 Sep, 20:04	62 min	36	Shop lighting
Site 2	Morning	27 Sep, 09:29	72 min	36	Marked lot
Site 2	Noon	27 Sep, 13:28	67 min	36	Last six frames rainy

Site 1 has no separate morning recording. Its day condition is early afternoon. Site 1 at night remains in the test because the shop windows light the stalls. Site 1 contributes 72 frames × 6 stalls = 432 decisions. Site 2 contributes 72 frames × 11 stalls = 792 decisions. Each model is scored on the same 1,224 reference decisions. Those 1,224 decisions are not 1,224 independent samples. The same vehicles reappear in successive frames. The 144 frames contain 17 distinct occupancy patterns: 10 in the Site 1 day recording, 4 at night, 2 in the Site 2 morning recording, and 1 pattern repeated through all 36 noon frames. The noon hour does not change those visible stalls. The morning hour changes one stall.

A separate check was taken from change points, not from the regular sample. The four recordings were scanned every 15 seconds inside the stall rectangles. A moment became a change candidate when the mean absolute pixel difference inside a stall, against the previous scan, exceeded 25. The author labeled 15 of those candidates: 7 from Site 1 by day, 6 from Site 1 at night, and 2 from Site 2 at noon. No extra morning frame was labeled, because that hour's only change is already in the regular sample. The decision count is 7×6 + 6×6 + 2×11 = 100. Four of the 15 frames show an occupancy pattern the regular sample missed. At the 40 percent threshold, F1 on these 100 decisions is 93.0 percent for YOLOv8n, 96.2 percent for YOLO11n, 94.7 percent for RT-DETR-

L, and 94.0 percent for Faster R-CNN. Faster R-CNN marked two empty stalls occupied. The other three did not. These 100 decisions are not in Table 1. They were selected only at change points, so they do not replace the regular sample.

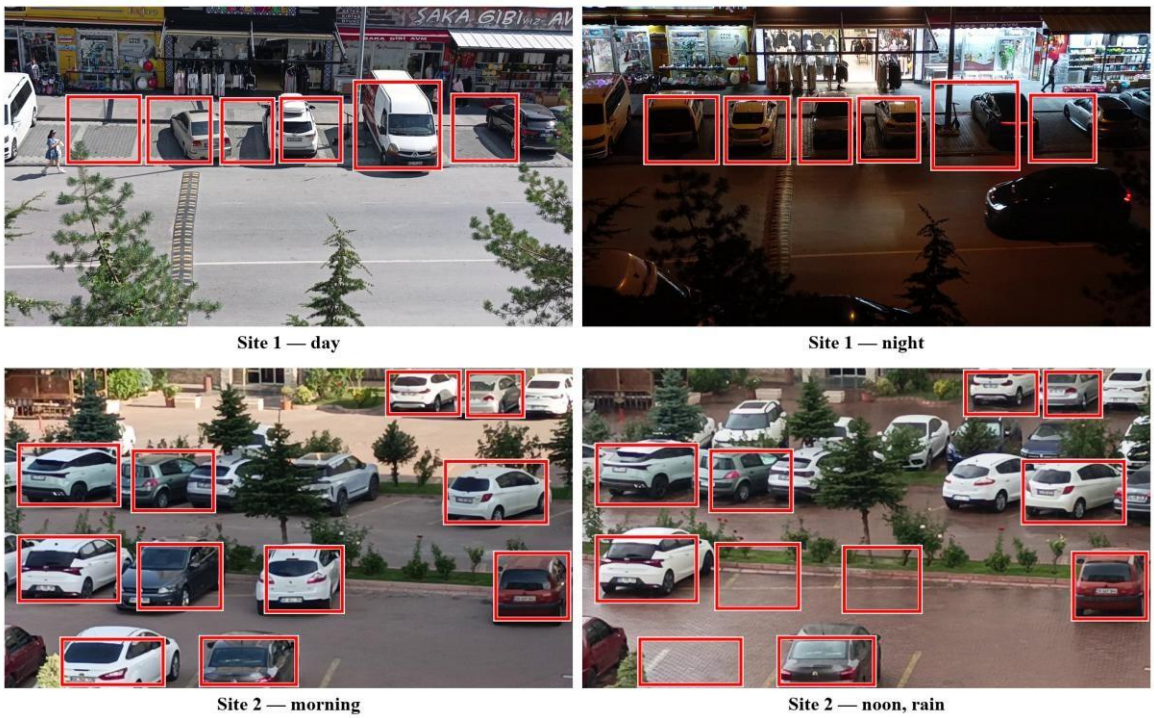


Figure 2. One example frame from each condition. Red rectangles are the stalls used in the score. Thirty-six frames were used in each condition.

Setup

The weights are yolov8n.pt, yolo11n.pt, and rtdetr-l.pt, plus a public COCO Faster R-CNN with a ResNet-50 backbone [2], [3], [13], [8]. YOLOv8n, YOLO11n, and RT-DETR-L were run in Ultralytics 8.4.47 with imgsz=640 and a confidence threshold of 0.25. Ultralytics returns the boxes on the original 1920×1080 frame, and the long side of those three models remains 640 pixels. Faster R-CNN is the COCO weight in torchvision 0.26.0. The PyTorch build is 2.11.0 for CPU. Before the call, each frame is scaled so that its long side is 640 pixels and placed on a 640×640 canvas with gray padding of value 114. The model's own transform then resizes that canvas. In this build, model.transform.min_size is 800 and model.transform.max_size is 1333, so the 640×640 canvas becomes an 800×800 tensor before the backbone. The library maps the boxes back to the 640×640 canvas. They are then mapped to the original frame by (coordinate − padding) / scale, before the coverage rule. The classes are car, bus, and truck, and the score threshold is 0.25. Inference ran on one laptop: Intel Core i7-11800H, 8 cores, 16 GB of RAM, Windows, and no GPU. The positive class is occupied. The times 39, 50, 636, and 2276 ms are the mean of the 144 frames after one warm-up pass on this CPU only. The Faster R-CNN time includes the internal

resize from 640×640 to 800×800. These times are not a measurement of a real-time system, and the same times would not be expected on another processor.

RESULTS

Table 1. Overall result on 1,224 stall decisions. The TP, FP, FN, and TN counts are in the same table as the mean CPU time per frame.

Model	Accuracy	Precision	Recall	Specificity	F1	TP	FP	FN	TN	Time
YOLOv8n	93.5%	98.9%	92.9%	96.1%	95.8%	901	10	69	244	39 ms
YOLO11n	96.3%	98.9%	96.4%	96.1%	97.7%	935	10	35	244	50 ms
RT-DETR-L	95.0%	98.9%	94.7%	96.1%	96.8%	919	10	51	244	636 ms
Faster R-CNN	96.3%	98.6%	96.7%	94.9%	97.7%	938	13	32	241	2276 ms

Accuracy is (TP+TN)/1224, precision is TP/(TP+FP), recall is TP/(TP+FN), specificity is TN/(TN+FP), and F1 is 2·precision·recall/(precision+recall). The percentages in the table are those counts rounded to one decimal place. Every model sees 970 occupied stalls and 254 empty stalls. Occupied stalls are the majority, so the result is not read from accuracy alone. The first three models share the same 10 false occupied decisions. Faster R-CNN has 13. YOLO11n misses 35 occupied stalls and Faster R-CNN misses 32. Their F1 values are the same at one decimal place: 97.7 percent. One daytime stall label was later changed to empty. That correction raises the overall F1 of each of the first three models by 0.1 point.

Table 2. Accuracy and F1 by site

Model	Site 1 accuracy	Site 1 F1	Site 2 accuracy	Site 2 F1
YOLOv8n	81.9%	85.5%	99.9%	99.9%
YOLO11n	89.8%	92.3%	99.9%	99.9%
RT-DETR-L	86.1%	89.2%	99.9%	99.9%
Faster R-CNN	89.8%	92.4%	99.9%	99.9%

Table 3. F1 by recording condition

Model	Site 1 day	Site 1 night	Site 2 morning	Site 2 noon
YOLOv8n	88.5%	83.3%	99.9%	100%
YOLO11n	91.8%	92.6%	99.9%	100%
RT-DETR-L	88.0%	90.0%	99.9%	100%
Faster R-CNN	92.5%	92.3%	99.9%	100%

On Site 2, 120 of the 792 decisions are empty stalls. The noon recording contains 108 of them, and the morning recording contains 12, all from one stall that becomes empty partway through the hour. All four models marked the empty stalls empty. One occupied stall in the morning was missed by every model. All six rainy frames were correct. Six frames do not support a general conclusion about rain. The 99.9 percent result at Site 2 applies only to the selected visible stalls. The vertical axis of Figure 3 starts at 75 percent and stops at 100 percent.

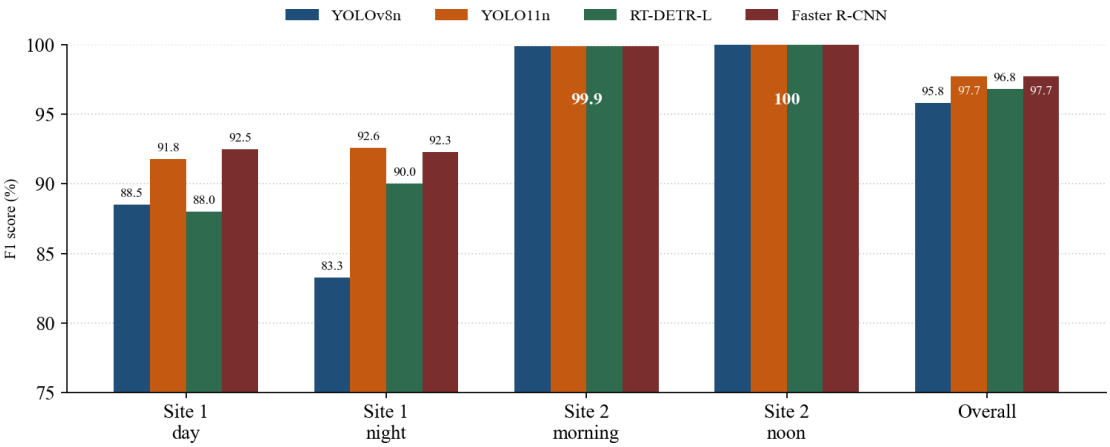


Figure 3. F1 score by condition. The vertical axis starts at 75 percent and stops at 100 percent

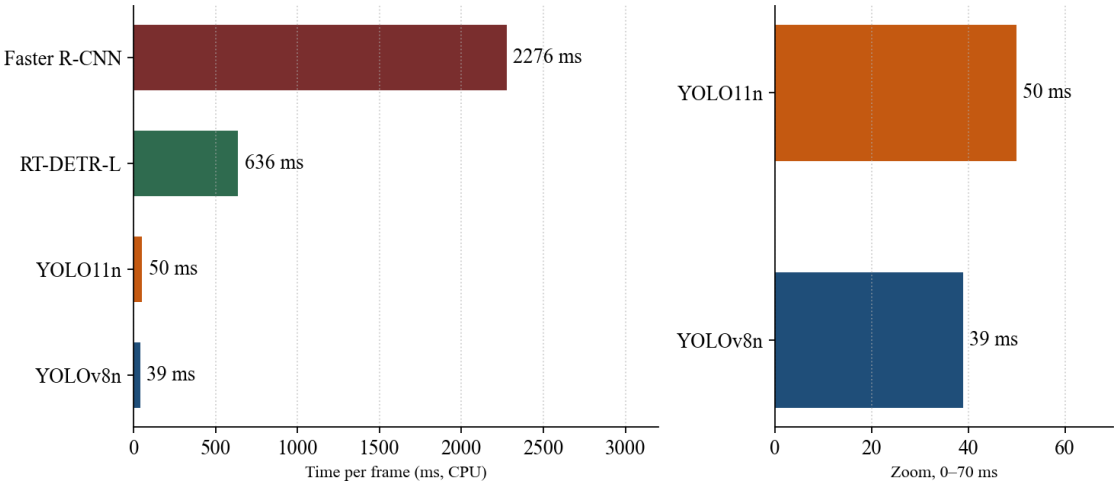


Figure 4. Mean CPU time per frame. The left panel shows all four models. The right panel zooms to 0–70 ms, so 39 ms and 50 ms can be read. Faster R-CNN takes 2276 ms and RT-DETR-L takes 636 ms.

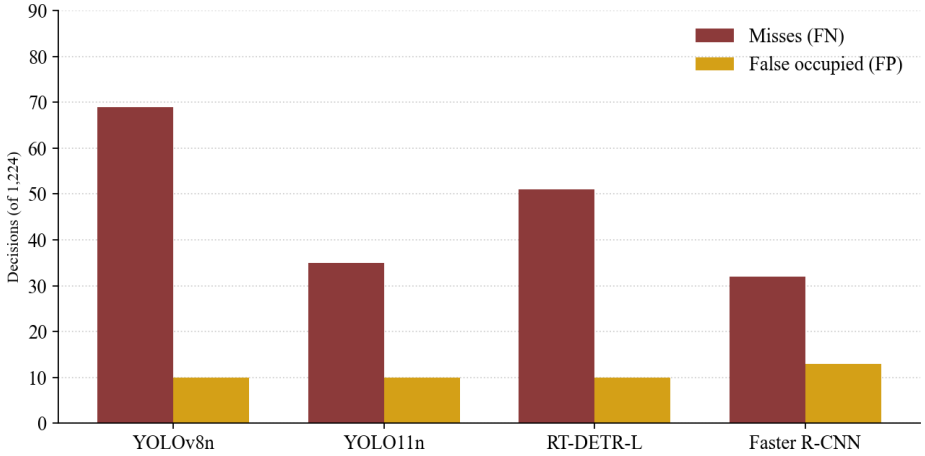


Figure 5. Missed occupied stalls (FN) and false occupied stalls (FP) inside the 1,224 decisions.

No statistical significance test was performed, so the differences are not called significant. A test that counted 1,224 decisions as independent trials would overstate the certainty, because neighboring frames repeat the same vehicles [7]. On Site 2 the four models agreed on every stall across all 72 frames. The only error is one occupied stall in a single morning frame, missed by all four. The four models do not produce the same errors on Site 1. Daytime F1 scores lie between 88.0 and 92.5 percent. At night, under shop lighting, YOLO11n reaches 92.6 percent F1 and Faster R-CNN reaches

92.3 percent, while YOLOv8n falls to 83.3 percent because its recall drops to 71.8 percent. Across those 36 night frames, YOLO11n had fewer stall errors than YOLOv8n on 23 frames, more errors on 1 frame, and the same count on 12 frames. Faster R-CNN had fewer stall errors than YOLOv8n on 22 frames, more on 2, and the same count on 12. RT-DETR-L reaches 90.0 percent F1 at night and takes 636 ms per frame. Faster R-CNN takes 2276 ms, about 45 times YOLO11n. The F1 values of YOLO11n and Faster R-CNN are the same at one decimal place: 97.7 percent.

Table 4. Effect of different coverage thresholds on the same model predictions. The main experiment uses 40 percent. This check was run later, on those same boxes. False occupied and missed stalls are listed as YOLOv8n / YOLO11n / RT-DETR-L / Faster R-CNN

Coverage	**YOLOv8n F1**	**YOLO11n F1**	**RT-DETR-L F1**	**Faster R-CNN F1**	**False occupied v8 / 11 / RT / FR**	**Missed stalls v8 / 11 / RT / FR**
20%	97.2%	98.4%	98.9%	98.2%	22 / 28 / 19 / 30	32 / 3 / 3 / 5
30%	97.7%	99.1%	99.2%	98.9%	13 / 14 / 13 / 17	32 / 3 / 3 / 5
40%	95.8%	97.7%	96.8%	97.7%	10 / 10 / 10 / 13	69 / 35 / 51 / 32
50%	94.6%	96.1%	95.9%	96.3%	2 / 2 / 0 / 11	97 / 71 / 77 / 59
60%	89.0%	89.8%	90.1%	90.5%	0 / 0 / 0 / 2	192 / 179 / 175 / 166

The main experiment uses the 40 percent coverage threshold. Table 4 shows the effect of the

threshold on the same detection boxes. At 30 percent, RT-DETR-L reaches 99.2 percent F1 and YOLO11n reaches 99.1. Faster R-CNN reaches 98.9 and misses 5 occupied stalls, while YOLO11n and RT-DETR-L each miss 3. At 40 percent the false occupied count is 10 for the first three models and 13 for Faster R-CNN, and the missed stalls rise to 69, 35, 51, and 32. The 40 percent threshold is the stricter experimental choice against a neighbor spilling into the stall. It is not the threshold that maximizes F1. Raising it to 50 or 60 percent mostly creates misses. The order at 30 percent is not the order at 40 percent. At 30 percent, YOLO11n and Faster R-CNN fall behind RT-DETR-L. The speed ranking does not change at any of these settings: Faster R-CNN remains about 45 times slower than YOLO11n.

DISCUSSION

The errors are concentrated at the curbside site. Two kinds of miss should be distinguished. In the first, the model draws no box on the car. In the second, a box is drawn and the 40 percent rule rejects it. Night frame 05, stall P2, is the first kind. YOLOv8n coverage is 0.00. YOLO11n is 0.59, RT - DETR-L is 0.57, and Faster R-CNN is 0.59. YOLOv8n misses the centered car. The other three detect it. YOLOv8n misses this P2 stall on 24 night frames, and on 23 of those frames the other three detect the car. Stall P5 in the same frame is the second kind. All four models see the car. Coverage is

0.38 for YOLOv8n, 0.36 for YOLO11n, 0.39 for RT-DETR-L, and 0.65 for Faster R-CNN. The 40 percent rule counts the first three as misses and counts Faster R-CNN as occupied. The car sits to the right of the rectangle. Night frame 25, stall P4, is also the second kind. Coverage is 0.37, 0.38, 0.35, and 0.35. The car inside the long rectangle is missed by all four because of the rule. Of the 51 night misses by YOLOv8n, 23 are also missed by YOLO11n and RT-DETR-L. Sixteen of those shared misses are P5 and five are P4. Faster R-CNN still misses 14 of those P5 stalls, marks two occupied, and misses the P4 stalls as well. Figure 6 shows the three examples with the stall area and the detection boxes.

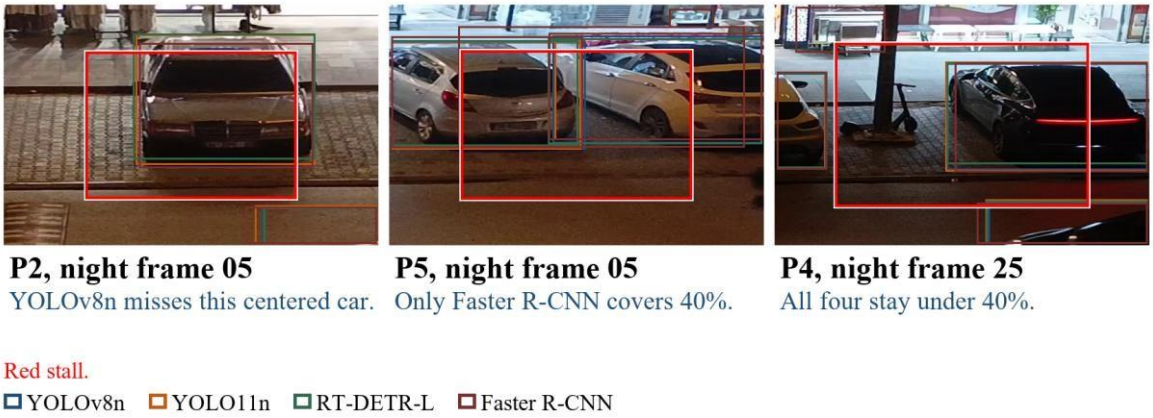


Figure 6. Site 1 at night. The red rectangle is the stall. The colored boxes are the four detections. At P2, YOLOv8n draws no box. At P5, only Faster R-CNN passes 40 percent. At P4, all four stay under 40 percent.

A false occupied decision usually comes from the box of the car in the next stall spilling into the defined rectangle. Cars driving on the road rarely enter the rectangles, because the rectangles stop at the curb. That boundary serves the same role as the physical placement of a sensor.

Faster R-CNN misses 32 occupied stalls and YOLO11n misses 35. Faster R-CNN marks 13 empty stalls occupied and YOLO11n marks 10. On this CPU, Faster R-CNN takes 2276 ms per frame, YOLO11n takes 50 ms, and RT-DETR-L takes 636 ms. These times belong only to the laptop that was measured, and the Faster R-CNN time includes its resize to 800×800. The speed comparison is valid for this experimental setup, in which Faster R-CNN runs at an 800×800 input size and the other models run at a 640-pixel input size. The claim that the tables support is limited. On the two sites examined, at the 40 percent threshold, the F1 values match at one decimal place and YOLO11n takes about 45 times less time. That is not a claim of superiority for every single-camera lot.

Each site uses one camera position. Site 1 has no morning light. Site 2 has no usable night, because the lot has no lighting. The 144 frames cover about an hour at each condition, but many cars do not move between neighboring frames, so the 1,224 decisions are repeated observations of the same vehicles. Rain appears in six frames, and those six frames do not support a general conclusion about rain. Labels come from one annotator. The models were not trained on these lots. Tree-covered stalls were never scored. Public lot studies already show that a classifier which is strong on one camera view can drop when the lot changes [5].

CONCLUSION

On the two sites examined, at the 40 percent coverage threshold, YOLO11n and Faster R-CNN reached a similar F1. The F1 values are the same at one decimal place, 97.7 percent. On the same

CPU, YOLO11n showed a much lower inference time. The model ranking depends on the coverage threshold and on the recording conditions. Labels come from one annotator. The number of sites and recordings is limited. The unlit night recording and the tree-covered areas were left out of the score.

References

J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in Proc. IEEE CVPR, 2016, pp. 779–788.

G. Jocher, A. Chaurasia, and J. Qiu, Ultralytics YOLOv8, version 8.0.0, 2023. G. Jocher and J. Qiu, Ultralytics YOLO11, version 11.0.0, 2024. Software. No page numbers. https://github.com/ultralytics/ultralytics

Y. Zhao et al., "DETRs Beat YOLOs on Real-time Object Detection," in Proc. IEEE/CVF CVPR, 2024, pp. 16965–16974.

T.-Y. Lin et al., "Microsoft COCO: Common Objects in Context," in Proc. ECCV, 2014, pp. 740–755.

P. R. L. de Almeida, L. S. Oliveira, A. S. Britto Jr., E. J. Silva Jr., and A. L. Koerich, "PKLot – A robust dataset for parking lot classification," Expert Systems with Applications, vol. 42, no. 11, pp. 4937–4949, 2015.

G. Amato, F. Carrara, F. Falchi, C. Gennaro, C. Meghini, and C. Vairo, "Deep learning for decentralized parking lot occupancy detection," Expert Systems with Applications, vol. 72, pp. 327 – 334, 2017.

P. R. L. de Almeida, J. H. Alves, R. S. Parpinelli, and J. P. Barddal, "A systematic review on computer vision-based parking lot management applied on public datasets," Expert Systems with Applications, vol. 198, Art. no. 116731, 2022, doi: 10.1016/j.eswa.2022.116731.

K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE CVPR, 2016, pp. 770–778.

G. Amato, F. Carrara, F. Falchi, C. Gennaro, and C. Vairo, "Car parking occupancy detection using smart camera networks and deep learning," in Proc. IEEE ISCC, 2016, pp. 1212–1217.

D. Acharya, W. Yan, and K. Khoshelham, "Real-time image-based parking occupancy detection using deep learning," in Proc. Research@Locate, CEUR Workshop Proceedings, vol. 2087, 2018,

pp. 33–40.

S. Valipour, M. Siam, E. Stroulia, and M. Jagersand, "Parking-stall vacancy indicator system, based on deep convolutional neural networks," in Proc. IEEE WF-IoT, 2016, pp. 655–660.

N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," in Proc. ECCV, 2020, pp. 213–229.

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.

H. Bura, N. Lin, N. Kumar, S. Malekar, S. Nagaraj, and K. Liu, "An Edge Based Smart Parking Solution Using Camera Networks and Deep Learning," in Proc. IEEE ICCC, 2018, pp. 17–24.

J. Nyambal and R. Klein, "Automated parking space detection using convolutional neural networks," in Proc. PRASA-RobMech, IEEE, 2017, pp. 1–6.

A Multidimensional Behavioral Segmentation Framework for an Automotive Subscription Service: Longitudinal Validation with Multi-Horizon Churn Outcomes

Buse Sena Vardar[1], Bilal Sedef[2], Selçuk Bayracı[3]*

[1] Data Analytics & AI, Borusan Otomotiv, Türkiye
[2] AI Engineering Department, Ing Hubs, Türkiye
[3] R&D Center, Borusan Otomotiv, Türkiye

*Corresponding author: selcuk.bayraci@borusanotomotiv.com

Abstract

Subscription-based services increasingly use customer segmentation to support retention decisions, yet operational segmentations are often deployed without longitudinal validation against subsequent customer behavior. This study backtests an interpretable rule-, score-, and quantile-based behavioral segmentation framework deployed in a premium automotive subscription service. The framework integrates recency, usage, monetary, payment, loyalty, breadth, and dynamic-risk dimensions to form lifecycle-, engagement-, and risk-oriented customer states. The analysis uses 251,169 monthly subscriber-vehicle observations across nine 2026 segmentation snapshots. Segment assignment used only information available at or before each snapshot; forward churn was then observed over 1-, 2-, and 3-month horizons, with incomplete outcome windows excluded. Critical Risk exhibited the highest pooled churn at every horizon, with lifts of 2.55x, 2.16x, and 1.91x relative to the portfolio average. Champions, High Engaged, and Engaged Moderate remained below the portfolio average across all valid months and horizons. Segment-level churn ordering was stable across overlapping monthly cohorts, with average pairwise Spearman rank correlations of 0.95-0.97. In addition, movement from Engaged Moderate to Need Attention and from Need Attention to Critical Risk was associated with higher subsequent churn than remaining in the origin state. These findings support the predictive validity and temporal stability of an interpretable multidimensional segmentation in a single operational setting and illustrate how a rating-system-style backtest, combining recurring point-in-time segmentation, multi-horizon outcome validation, and transition analysis, can be applied to customer segments.

Keywords: *Customer segmentation, Churn validation, Behavioral analytics, Subscription services, Longitudinal analysis*

INTRODUCTION

Customer segmentation is a fundamental component of customer relationship management because it converts heterogeneous customer populations into interpretable groups that can support differentiated retention and engagement decisions. The Recency-Frequency-Monetary (RFM) framework remains influential because of its simplicity and transparency, but contemporary subscription relationships often contain signals that extend beyond transaction recency, frequency, and monetary value. Prior studies have therefore enriched customer representations with dimensions such as relationship length,

periodicity, satisfaction, usage intensity, and other context-specific behavioral measures (Miglautsch, 2000; Peker et al., 2017; Alves Gomes and Meisen, 2023; Wang et al., 2024).

Customer behavior is also inherently dynamic. Cross-sectional segment membership may describe a current state without capturing subsequent movement in engagement, value, or risk. Dynamic and longitudinal approaches have addressed this issue through repeated customer observations, time-series representations, and customer-state transitions (Reutterer et al., 2006; Homburg et al., 2009; Abbasimehr and Bahrini, 2022). Research on segment migration further shows that transitions between customer states can contain economically relevant information beyond static membership, including associations with customer lifetime value and churn (Mosaddegh et al., 2021; Fukui and Tabata, 2026).

At the same time, customer-segmentation evaluation remains heterogeneous. Reviews report a wide range of internal, statistical, and application-oriented criteria, with no single generally accepted validation standard (Alves Gomes and Meisen, 2023; Salminen et al., 2023). Validation against independently observed future outcomes, that is, predictive or criterion validity, therefore provides a useful complementary perspective. Recent behavioral-segmentation research has linked segment membership with subsequently observed commercial behavior rather than evaluating segment quality only through the variables used to create the groups (Kose et al., 2026).

The literature nevertheless tends to examine multidimensional segmentation, dynamic customer states, future-outcome validation, and segment transitions as partially separate problems, and a recent systematic review describes retention management in subscription settings as a fragmented research field (Muratčehajić and Loureiro, 2024). In credit risk, by contrast, rating systems that drive operational decisions are routinely backtested out-of-time for discriminatory power, stability, and migration behavior (Basel Committee on Banking Supervision, 2005). Operational customer segmentations support comparable prioritization decisions, yet they are rarely subjected to an equivalent protocol: outcome-based validations of behavioral segments typically relate segment membership to a single outcome observation (Kose et al., 2026), while segment-transition studies model movement between states rather than backtesting the state labels themselves over time (Fukui and Tabata, 2026). The contribution of the present study lies in applying such a backtesting design to a deployed segmentation rather than in proposing a new segmentation algorithm.

This study backtests a multidimensional behavioral segmentation framework deployed in a premium automotive subscription service. The framework integrates recency, usage, monetary, payment, loyalty, breadth, and dynamic-risk information within a rule-, score-, and quantile-based structure. Segment states are assigned repeatedly at monthly observation points using only information available at or before each snapshot and are evaluated against subsequently observed 1-, 2-, and 3-month churn. The analysis also evaluates temporal stability and whether selected deteriorating transitions are associated with higher subsequent churn.

Accordingly, the study addresses three research questions: RQ1, whether the proposed behavioral segments exhibit distinct customer profiles; RQ2, whether the segments exhibit systematically different observed churn outcomes across multiple forward horizons and whether this ordering remains temporally stable; and RQ3, whether deteriorating transitions are associated with higher subsequent

churn than remaining in the origin state, and how recent entrants into a risk state compare with the state's established members.

MATERIAL AND METHODS

Data and Study Design

The validation dataset contains 251,169 monthly observations across 52 variables and covers nine monthly segmentation snapshots from January through September 2026. The unit of analysis is an anonymized subscriber-vehicle relationship observed at a monthly snapshot level. Each entity-month combination is unique. A total of 41,886 distinct anonymized entities are represented, of which 17,414 appear in all nine snapshots; the remaining entities enter or leave the observation window during the study period. The observations therefore represent repeated longitudinal measurements rather than 251,169 independent customers.

A central design property is point-in-time feature construction. Historical behavioral aggregates used for segmentation were restricted to information observed before the corresponding snapshot month, while lifecycle and subscription-state variables were defined as of the snapshot date. No post-snapshot information was used to determine segment membership. Forward churn outcomes were observed only after assignment. This separation prevents outcome-period information from influencing the behavioral features underlying the segment label later evaluated against that outcome; the calibration of the Dynamic Risk component is described separately below.

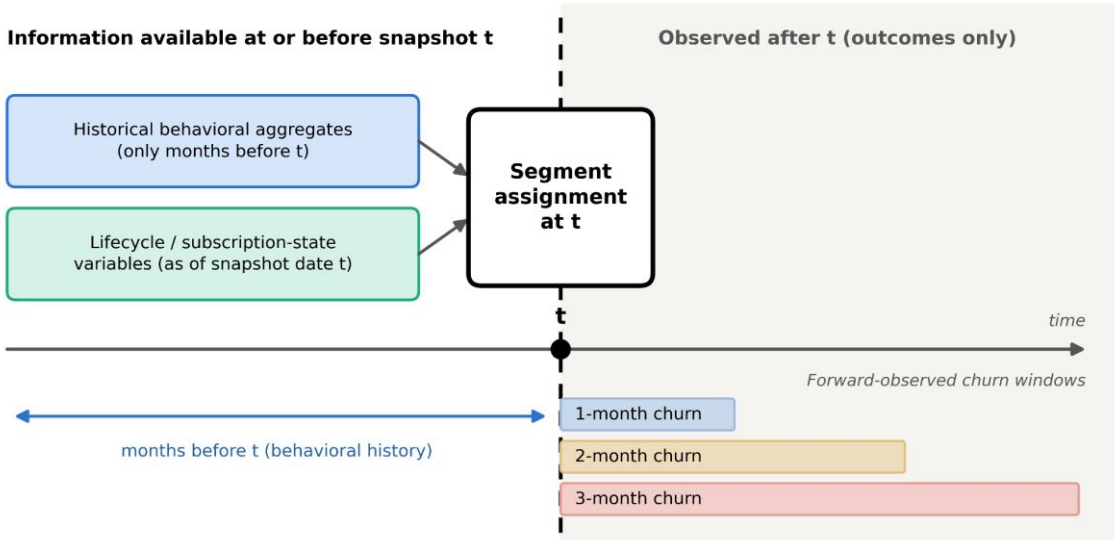


Figure 1. Temporal study design. Information available at or before snapshot t determines segment membership, while realized churn is observed only after assignment over complete one-, two-, and three-month outcome windows. Dynamic Risk relative-risk contributions are calibrated on a fixed historical window that closes before the first evaluated snapshot (see Segmentation Framework).

Three realized churn outcomes were used: churn within one, two, and three months following the snapshot. Only snapshots with a fully complete future observation window were eligible for a given horizon. The 1-month analysis therefore used January-August 2026 snapshots, the 2-month analysis January-July, and the 3-month analysis January-June. September was retained for descriptive and

transition analyses but excluded from forward churn validation. Subscription terminations associated with package changes were excluded to avoid treating product migration within the service as attrition.

Table 1. Out-of-time churn-validation design

Validation horizon	Eligible snapshots	Observed outcome
1 month	January-August 2026	Churn within the first month following the snapshot
2 months	January-July 2026	Churn within the first two months following the snapshot
3 months	January-June 2026	Churn within the first three months following the snapshot

Segmentation Framework

The framework represents each anonymized relationship through six core behavioral dimensions plus a separate Dynamic Risk component. Recency captures time since last recorded usage; Usage combines intensity and active-tenure coverage; Monetary summarizes average monthly amount paid relative to the current portfolio; Payment captures payment timeliness; Loyalty combines tenure, repeated subscription behavior, concurrent subscriptions, and persistence through historical price increases; and Breadth measures the diversity of service features used in a trailing three-month window.

Dynamic Risk is timing-sensitive rather than a static indicator of whether a benefit was used. For selected service benefits, usage is evaluated by package, lifecycle period, and usage timing. Historical subsequent churn within a fixed forward window is compared with the corresponding not-yet-used reference group to derive relative-risk contributions. Consequently, the same benefit can receive different risk adjustments depending on when it is used during the subscription lifecycle, and the contribution can be positive or negative before being aggregated into Dynamic Risk. These relative-risk contributions were estimated on and held fixed throughout the evaluation period, so that no churn outcome observed within an evaluated outcome window entered their estimation. Dynamic Risk is therefore outcome-informed by design: the validation below tests whether historically calibrated risk patterns transfer to later periods, not whether an outcome-blind segmentation happens to be associated with churn.

Six behavioral components are combined additively into Base Engagement, while Dynamic Risk is retained as a separate component:

$$E = U + R + M + P + L + B \qquad (1)$$

Base Engagement and Dynamic Risk are each robustly normalized using the snapshot population median and interquartile range, with a standard-deviation fallback for degenerate scales. The risk-adjusted score is then defined as:

$$z_{Final} = z_{Base} - \alpha \cdot z_{Risk} \qquad (2)$$

where α is a fixed rule-based parameter rather than an outcome-fitted coefficient. Several categorical levels are defined through portfolio-relative quantile thresholds, while recency and payment use fixed

cutoffs. Final segments are assigned hierarchically so that the first satisfied rule determines the label. The ordered states are New, Never Used, Critical Risk, Need Attention, Champions, High Engaged, and Engaged Moderate.

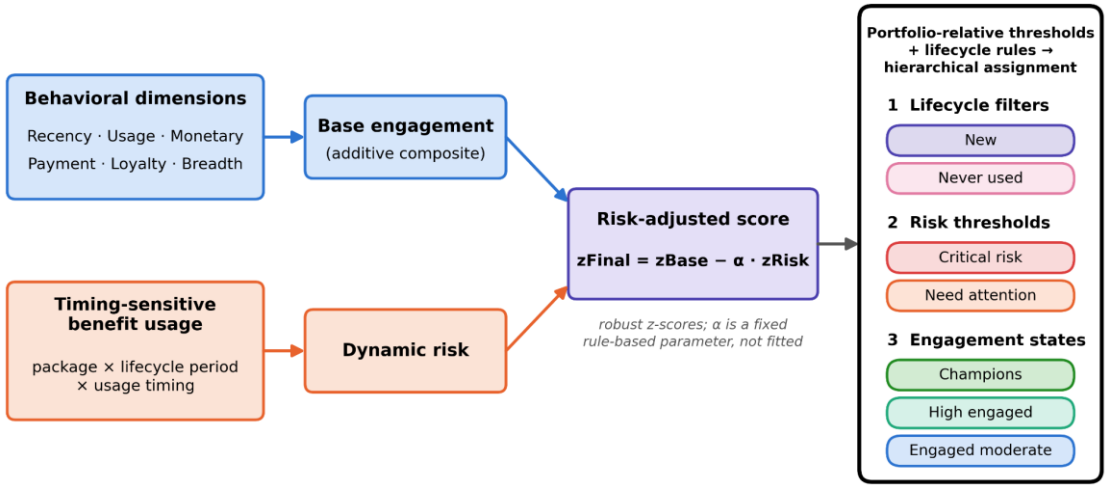


Figure 2. Conceptual pipeline of the interpretable multidimensional behavioral segmentation framework. Six behavioral dimensions form Base Engagement, while timing-sensitive benefit usage contributes separately to Dynamic Risk before risk adjustment and hierarchical segment assignment.

Table 2. Conceptual definition of the final segment states

Segment	Conceptual interpretation
New	Relationship tenure below two months
Never Used	Tenured relationship with no active-usage months
Critical Risk	Strongest rule-based risk condition
Need Attention	Intermediate elevated-risk condition
Champions	Top loyalty, high usage, favorable recency, and supportive monetary/payment conditions
High Engaged	High loyalty and usage with favorable recency and payment behavior
Engaged Moderate	Remaining active, non-risk relationships not meeting the stricter high-engagement rules

Note: Because assignment is hierarchical, a relationship meeting the New condition is labeled New even if it also satisfies a risk condition.

Validation Methodology

Validation was organized around behavioral differentiation, forward churn separation, temporal consistency, and dynamic validity through segment transitions and deterioration. For descriptive profiling, segment-level raw behavioral measures were summarized using medians and quartiles. Because several raw measures are related to the segmentation inputs, this profiling is interpreted as internal behavioral consistency rather than validation against an independent data source.

For each horizon, segment churn rate was calculated as the share of eligible observations in a segment with a positive churn outcome. Churn lift compares this rate with the overall portfolio rate for the same horizon:

$$\text{Lift}(\text{segment}, h) = \text{churn rate}(\text{segment}, h) / \text{churn rate}(\text{portfolio}, h) \quad (3)$$

Monthly churn rates are reported as point estimates (Figure 3). For pooled multi-month estimates, the same entity can contribute several observations; pooled churn rates and lifts therefore use an entity-level cluster bootstrap. Anonymized entities, rather than rows, were resampled with replacement, all eligible rows of each sampled entity were retained, and 95% intervals were obtained from the 2.5th and 97.5th percentiles of 1,000 bootstrap replicates using a fixed random seed.

Temporal consistency was assessed by ranking segment-level churn rates separately within every valid month and computing the Spearman rank correlation for every pair of valid months. The reported summary is the average pairwise Spearman correlation together with its range across month pairs. Because consecutive snapshots largely contain the same entities, this statistic summarizes stability across overlapping cohorts rather than independent replications. Segment transitions were constructed from strictly adjacent monthly snapshots for the same anonymized entity. Selected higher-risk transitions were then compared with two references: the corresponding same-origin self-transition (e.g., Engaged Moderate to Engaged Moderate) and the same-destination self-transition (e.g., Need Attention to Need Attention), the latter testing whether recent entry into a state carries different risk from established membership; deterioration uncertainty was estimated with the same entity-level cluster bootstrap. Cells with fewer than 30 observations were flagged as small-sample.Diagnostic grouping comparisons and alternative portfolio-subset checks were treated as secondary analyses rather than algorithmic benchmarks.

RESULTS

Behavioral Differentiation

On raw behavioral measures, the seven segments exhibited substantively distinct profiles. Across the January-August snapshots, Engaged Moderate accounted for 62.6% of eligible observations, Need Attention for 16.4%, Critical Risk for 6.8%, New for 6.0%, Never Used for 4.1%, High Engaged for 3.8%, and Champions for 0.3%. Champions is defined by the most restrictive rule set and therefore remains small. Champions and High Engaged showed the strongest tenure and usage measures, whereas Never Used and Need Attention showed the greatest usage recency, indicating longer dormancy. Service breadth was highest for Champions and High Engaged and near zero for New and Never Used. Average invoice-to-payment delay did not differentiate the segments at the median. Because several raw measures are related to segmentation inputs, these patterns are interpreted as evidence of behavioral consistency rather than independent validation.

Temporal Consistency

The relative ordering of segments by observed churn was stable across repeated monthly snapshots. Average pairwise Spearman rank correlations of monthly segment-churn rankings were approximately 0.95 at one month (range 0.79-1.00 across 28 month pairs), 0.96 at two months, and 0.97 at three months. Because consecutive snapshots share most entities, these values describe stability across overlapping

cohorts rather than independent replications. With seven segments, a single swap of adjacent ranks yields a coefficient of approximately 0.96; the observed deviations were swaps among the mid-risk states New, Need Attention, and Never Used, and an elevated Champions rate in a single month reflecting the segment's small size (Figure 3). Critical Risk and Need Attention remained above the portfolio average in every valid month at every horizon (8/8, 7/7, and 6/6 months), whereas Champions, High Engaged, and Engaged Moderate remained below the portfolio average in every valid month.

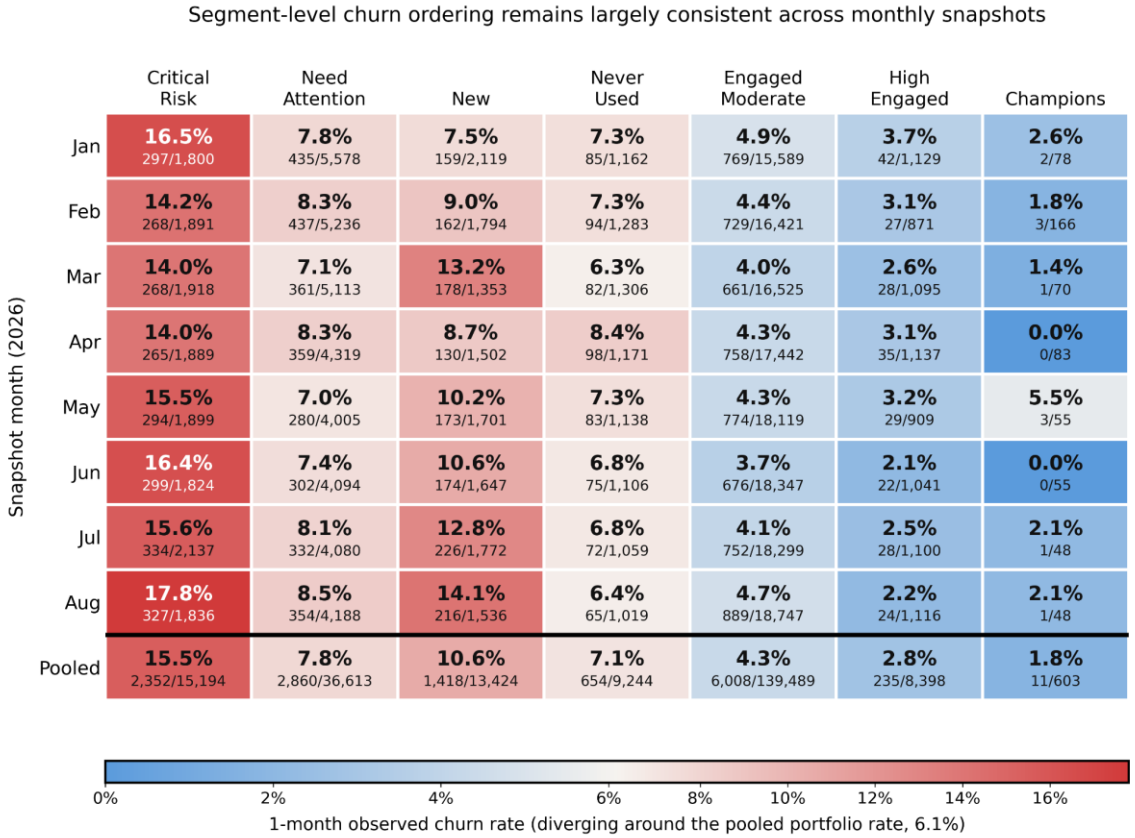


Figure 3. One-month observed churn by segment across January-August 2026 snapshots, with pooled values in the final row. Cells report churn rate and churned/eligible counts. Segment-level churn ordering remains largely consistent across monthly snapshots, while Champions, the smallest segment, shows greater month-to-month variability.

Multi-Horizon Churn Separation

Segment membership exhibited systematically different observed churn outcomes at every forward horizon (Figure 4; Table 3). Critical Risk had the highest pooled churn rate: 15.5% at one month, 24.3% at two months, and 30.0% at three months, corresponding to lifts of 2.55x, 2.16x, and 1.91x relative to the portfolio average. The entity-cluster bootstrap interval for lift remained entirely above 1 at every horizon. Need Attention and New also remained above the portfolio average across all three horizons.

At the opposite end, Champions, High Engaged, and Engaged Moderate remained below the portfolio average at every horizon, with bootstrap lift intervals entirely below 1. Champions showed the

same direction but with wider uncertainty because of its smaller number of observations. Never Used remained modestly above the portfolio average across all three horizons.

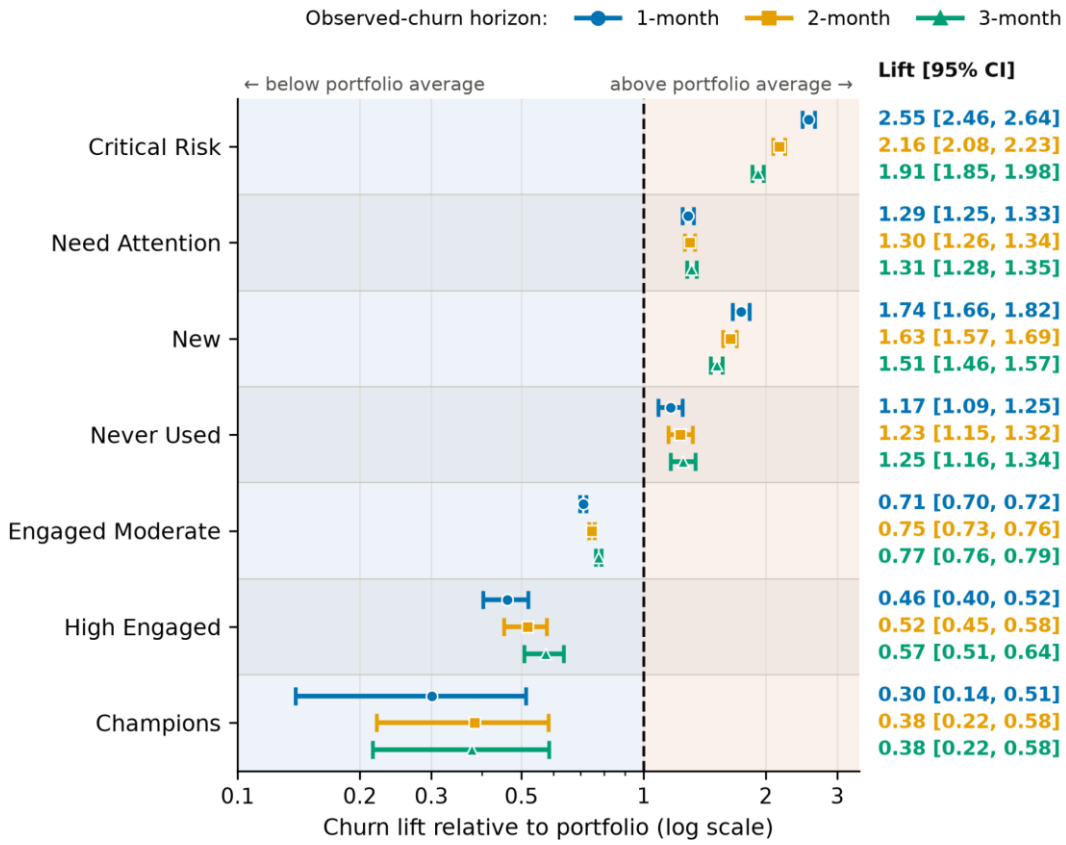


Figure 4. Pooled churn lift by segment across the one-, two-, and three-month horizons, with 95% entity-level cluster bootstrap confidence intervals (B = 1,000). The dashed reference line at 1.0 represents the portfolio average; values above and below 1 indicate higher and lower observed churn, respectively.

Table 3. Segment-level pooled churn rates and churn lifts across the three forward horizons

Segment	1M churn	1M lift [95% CI]	2M churn	2M lift [95% CI]	3M churn	3M lift [95% CI]
Critical Risk	15.5%	2.55 [2.46, 2.64]	24.3%	2.16 [2.08, 2.23]	30.0%	1.91 [1.85, 1.98]
New	10.6%	1.74 [1.66, 1.82]	18.4%	1.63 [1.57, 1.69]	23.7%	1.51 [1.46, 1.57]
Need Attention	7.8%	1.29 [1.25, 1.33]	14.6%	1.30 [1.26, 1.34]	20.6%	1.31 [1.28, 1.35]
Never Used	7.1%	1.17 [1.09, 1.25]	13.9%	1.23 [1.15, 1.32]	19.6%	1.25 [1.16, 1.34]

Engaged Moderate	4.3%	0.71 [0.70, 0.72]	8.4%	0.75 [0.73, 0.76]	12.2%	0.77 [0.76, 0.79]
High Engaged	2.8%	0.46 [0.40, 0.52]	5.8%	0.52 [0.45, 0.58]	9.0%	0.57 [0.51, 0.64]
Champions	1.8%	0.30 [0.14, 0.51]	4.3%	0.38 [0.22, 0.58]	5.9%	0.38 [0.22, 0.58]

Segment Deterioration

Movement into a higher-risk state was associated with higher subsequent churn for some, but not all, tested deterioration pathways (Figure 5). Engaged Moderate to Need Attention was associated with approximately 1.50x the subsequent 1-month churn and 1.45x the 3-month churn of remaining Engaged Moderate, with bootstrap intervals entirely above 1. Need Attention to Critical Risk was also above 1 at both horizons relative to remaining in Need Attention. Engaged Moderate to Critical Risk was inconclusive at one month because its interval included 1, and the High Engaged to Need Attention pathway had insufficient sample support for reliable inference.

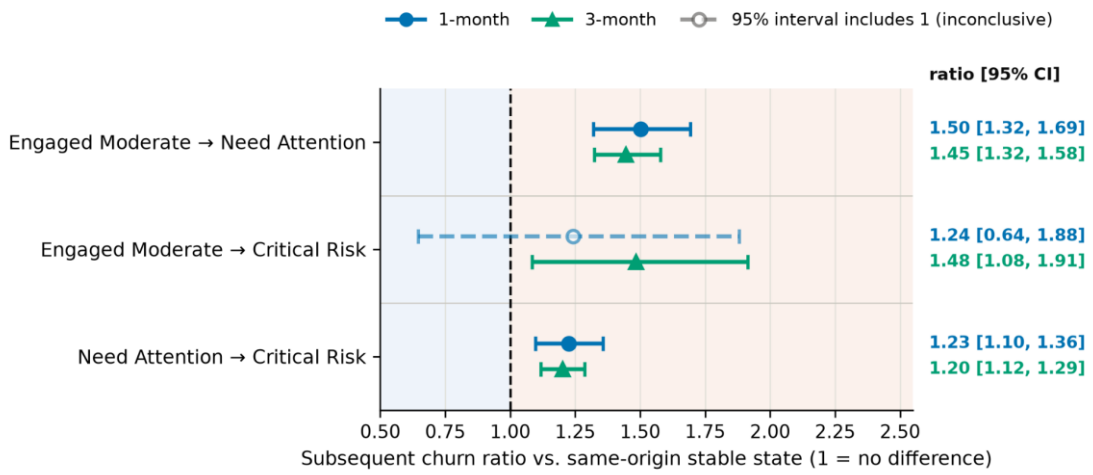


Figure 5. Ratio of subsequent churn after selected deterioration transitions relative to the corresponding same-origin stable state, with 95% entity-level cluster bootstrap intervals. The one-month interval for Engaged Moderate to Critical Risk includes 1 and is treated as inconclusive; High Engaged to Need Attention is omitted because of insufficient sample support. The two-month horizon is not shown.

Diagnostic and Robustness Findings

Secondary analyses supported, but did not redefine, the main findings. The composite segmentation descriptively exhibited a wider churn-rate range than simpler risk-, engagement-, or usage-recency groupings at both the 1- and 3-month horizons; this is treated as a diagnostic observation rather than evidence of superiority. Moderate-engagement entities showed increasing churn from no-risk through Need Attention to Critical Risk, while the same pattern was weak within the High-engagement subgroup. The main directional segment-churn pattern persisted across most tested portfolio subsets. One subscription-package subset was an exception, with New showing higher 1-month churn than Critical

Risk, indicating that lifecycle- and risk-oriented states do not necessarily preserve the same ordering within every subgroup.

DISCUSSION AND CONCLUSION

Discussion

The results support the view that an interpretable multidimensional behavioral segmentation can function as more than a static descriptive taxonomy. The framework is not positioned as a new segmentation algorithm: multidimensional extensions of traditional RFM representations are already well established. Its contribution instead lies in showing that a rule-based operational segmentation can retain interpretability while exhibiting meaningful and temporally stable separation on subsequently observed outcomes.

The multi-horizon results provide the strongest evidence of predictive validity. Risk-oriented segments consistently exhibited higher subsequent churn than the portfolio average, whereas more engaged states remained below average. This separation persisted across repeated snapshots and across three forward horizons, extending prior outcome-based validation work by evaluating a recurring point-in-time segmentation rather than a retrospectively fitted churn classifier (Kose et al., 2026). Unlike that study, however, one component of the present framework, Dynamic Risk, is calibrated on historical churn; the evidence therefore concerns the out-of-time transfer of an outcome-informed rule set rather than the outcome relevance of an outcome-blind segmentation.

A further feature of the framework is that risk adjustment is sensitive to benefit-usage timing rather than treating service use as a static binary signal. The same type of benefit use can therefore carry a different risk contribution at different lifecycle periods, with the direction and magnitude derived from historically observed relative-risk patterns. This preserves interpretability while allowing the operational segmentation to represent time-dependent usage behavior.

The deterioration results also connect the framework with prior research on customer-state dynamics. Earlier studies have shown that transitions among customer segments can carry information about downstream value and attrition (Netzer et al., 2008; Mosaddegh et al., 2021; Fukui and Tabata, 2026). Here, two tested deterioration pathways showed bootstrap intervals entirely above the origin-state reference at both evaluated horizons, while other pathways were inconclusive or insufficiently supported. This indicates that movement toward a nominally higher-risk state should not be interpreted as equally informative across all origin-destination pairs.

Operationally, the framework is best viewed as an interpretable prioritization layer rather than a replacement for a predictive churn model. Static membership identifies broad customer states, while state deterioration can indicate that a customer's behavioral condition has changed. The present study does not evaluate whether interventions targeted through these segments reduce churn, and customers at the highest risk are not necessarily those most responsive to retention efforts (Ascarza, 2018); the findings indicate forward outcome differentiation and temporal stability, not treatment effectiveness or causal impact.

Limitations and Future Work

The study has several limitations. The empirical validation concerns a single premium automotive subscription context and nine monthly snapshots within 2026, so generalization to other services, markets, or economic periods requires replication. The dataset also contains repeated observations for the same entities. Entity-level cluster bootstrap resampling was therefore used for pooled uncertainty estimation, but no formal longitudinal regression, mixed-effects model, or survival model was fitted; the evidence should be interpreted as descriptive and associational rather than causal or individual-level predictive inference. The framework was designed and evaluated by the same team within the operating company; although the rule set was fixed before the evaluation period and churn is an administratively recorded event, independent replication would strengthen the evidence. Finally, because Dynamic Risk is calibrated on historical churn, its contribution to the observed separation cannot be interpreted as evidence from outcome-blind behavioral features alone.

The segmentation itself contains operationally defined scores, thresholds, and portfolio-relative quantiles. These design choices support interpretability and adaptation to the current portfolio, but they may require recalibration as behavior or service design changes. In addition, this study was not designed as an algorithmic benchmark: no direct traditional-RFM baseline, alternative clustering algorithm, or full dimension-by-dimension ablation was used to establish comparative superiority. Future work could compare the full framework with an RFM-like core and selected reduced-dimension variants, extend robustness checks to longer horizons, and accumulate additional snapshots to strengthen transition analyses.

CONCLUSION

This study longitudinally evaluated an interpretable multidimensional behavioral segmentation framework through recurring point-in-time snapshots, forward-observed multi-horizon churn, temporal consistency, and transition-based deterioration. The results showed persistent outcome separation between risk-oriented and engaged states and showed that selected deteriorating transitions were associated with higher subsequent churn than remaining in the corresponding stable state.

The contribution therefore lies in the validation design rather than in proposing a new segmentation algorithm. Combining interpretable customer states with point-in-time temporal construction, multiple future horizons, repeated monthly validation, and transition analysis provides a practical framework for assessing whether operational segments retain meaningful outcome differentiation over time. Future work should test this framework over longer periods, in additional contexts, against simpler baselines, and within prospective retention interventions.

DECLARATIONS

Acknowledgements: The authors thank the BMW Group Digital Business Development Team for their reviews and valuable business feedback.

Conflict of interest: All authors are employees of the company that operates the data analytics and ai models and designed the segmentation framework evaluated in this study. The authors declare no other competing interests.

Author contributions: Buse Sena Vardar: conceptualization, software, methodology, writing; Bilal Sedef: supervision, methodology, software; Selçuk Bayracı: supervision, writing;

Use of AI tools: Generative AI tools were used in preparing this manuscript. Anthropic's Claude was used for language editing, restructuring and condensing sections, converting the manuscript into the conference template. The authors reviewed and verified all content, equations, numerical results and references, and take full responsibility for the paper.

References

Abbasimehr H, Bahrini A, 2022. An analytical framework based on the recency, frequency, and monetary model and time series clustering techniques for dynamic segmentation. Expert Systems with Applications, 192: 116373. https://doi.org/10.1016/j.eswa.2021.116373

Alves Gomes M, Meisen T, 2023. A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. Information Systems and e-Business Management, 21: 527-570. https://doi.org/10.1007/s10257-023-00640-4

Ascarza E, 2018. Retention futility: Targeting high-risk customers might be ineffective. Journal of Marketing Research, 55 (1): 80-98. https://doi.org/10.1509/jmr.16.0163

Basel Committee on Banking Supervision, 2005. Studies on the validation of internal rating systems. Working Paper No. 14. Bank for International Settlements, Basel.

Fukui Y, Tabata T, 2026. An absorbing Markov chain for customer segment transition. Journal of Marketing Analytics. https://doi.org/10.1057/s41270-026-00516-9

Homburg C, Steiner VV, Totzek D, 2009. Managing dynamics in a customer portfolio. Journal of Marketing, 73 (5): 70-89. https://doi.org/10.1509/jmkg.73.5.70

Kose BO, Coskun V, Ozkan L, Mantar HA, 2026. Uncovering customer heterogeneity through behavioral segmentation: Linking digital service usage patterns to purchase behavior. Journal of Theoretical and Applied Electronic Commerce Research, 21 (9): 333. https://doi.org/10.3390/jtaer21090333

Miglautsch JR, 2000. Thoughts on RFM scoring. Journal of Database Marketing & Customer Strategy Management, 8 (1): 67-72. https://doi.org/10.1057/palgrave.jdm.3240019

Mosaddegh A, Albadvi A, Sepehri MM, Teimourpour B, 2021. Dynamics of customer segments: A predictor of customer lifetime value. Expert Systems with Applications, 172: 114606. https://doi.org/10.1016/j.eswa.2021.114606

Muratčehajić D, Loureiro SMC, 2024. Subscriber retention management: SRM framework and future research agenda. Journal of Services Marketing, 38: 1030-1057. https://doi.org/10.1108/JSM-02-2024-0062

Netzer O, Lattin JM, Srinivasan V, 2008. A hidden Markov model of customer relationship dynamics. Marketing Science, 27 (2): 185-204. https://doi.org/10.1287/mksc.1070.0294

Peker S, Koçyiğit A, Eren PE, 2017. LRFMP model for customer segmentation in the grocery retail industry: A case study. Marketing Intelligence & Planning, 35 (4): 544-559. https://doi.org/10.1108/MIP-11-2016-0210

Reutterer T, Mild A, Natter M, Taudes A, 2006. A dynamic segmentation approach for targeting and customizing direct marketing campaigns. Journal of Interactive Marketing, 20 (3-4): 43-57. https://doi.org/10.1002/dir.20066

Salminen J, Mustak M, Sufyan M, Jansen BJ, 2023. How can algorithms help in segmenting users and customers? A systematic review and research agenda for algorithmic customer segmentation. Journal of Marketing Analytics, 11: 677-692. https://doi.org/10.1057/s41270-023-00235-5

Wang S, Sun L, Yu Y, 2024. A dynamic customer segmentation approach by combining LRFMS and multivariate time series clustering. Scientific Reports, 14: 17491. https://doi.org/10.1038/s41598-024-68621-2

Cognitive Dual-Channel AI Framework for Financial Market Forecasting: An Application to Borsa Istanbul

Hulusi Mehmet Tanrıkulu[1]*, Hakan Pabuçcu[2], Muhammed İhsan Çubukcu[3]

[1,2] Bayburt University, Faculty of Economics and Administrative Sciences, Department of Business Administration, Bayburt, Türkiye
[3] Atatürk University, Faculty of Economics and Administrative Sciences, Department of International Trade and Logistics, Erzurum, Türkiye

*Corresponding author e-mail: mtanrikulu@bayburt.edu.tr

Abstract

The complex, high-variance, and non-linear dynamics of financial time series render conventional univariate predictive frameworks that rely exclusively on historical price series inadequate. This limitation necessitates the integration of unstructured textual information reflecting asymmetric information diffusion and investor psychology into quantitative forecasting models. This study presents Deep News Trader (DNT Engine), a dual-channel hybrid deep learning decision support architecture designed to evaluate structured technical analysis indicators alongside Natural Language Processing (NLP)-based news and social sentiment scores at the cellular level for the top 10 influential stocks in Borsa Istanbul (BIST). The empirical scope is strictly confined to the domestic market, and the study employs domain-adapted FinBERT and TrBERT to extract financial sentiment polarity. The Technical Channel models price momentum using 1D-CNN and LSTM networks, while the Sentiment Channel captures cumulative investor sentiment using BiLSTM networks. Extracted latent feature vectors are dynamically integrated via an intermediate fusion strategy and weighted by a Self-Attention mechanism, enabling an autonomous cognitive attention shift toward sentiment data during systemic market shocks. Furthermore, a Meta-Learning Noise Regressor adaptively modulates fusion layer activations—switching between saturating Sigmoid functions for speculative assets and non-saturating LeakyReLU for rational assets. Performance is evaluated using Fisher-Adjusted Real Return, Demeaned Directional Accuracy (MDA), and Explainable Artificial Intelligence (SHAP) analyses. The empirical results demonstrate that the proposed dual-channel architecture significantly enhances directional hit rates over single-channel baselines and preserves capital during crisis periods.

Keywords: *Artificial Intelligence, Behavioral Finance, Deep Learning, Financial Sentiment Analysis, Intermediate Fusion*

INTRODUCTION

Capital markets represent complex and chaotic systems governed by multi-dimensional interactions ranging from macroeconomic announcements and order-flow microstructure to collective behavioral biases and geopolitical shocks. The Efficient Market Hypothesis (EMH) formulated by Fama (1970)

asserts that asset prices instantly reflect all available public information, thereby following a statistical random walk. However, behavioral finance has demonstrated that cognitive limits, noise trading (De Long et al., 1990), and herd behavior induce systematic pricing anomalies and delayed reactions (Kahneman and Tversky, 1979). In emerging markets such as Borsa Istanbul (BIST), characterized by pronounced volatility and retail trading activity, traditional linear econometric models fail to absorb the informational shocks embedded within unstructured narrative feeds (Baltrusaitis et al., 2019; Gao et al., 2021).

To model non-linear price formation, deep learning architectures—specifically Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks—have been widely adopted. Simultaneously, the emergence of Transformer-based models (Devlin et al., 2019; Araci, 2019) has transformed textual sentiment modeling into deep contextual embeddings. Nonetheless, current multimodal financial architectures suffer from three distinct methodological limitations: First, direct early concatenation of price series and sentiment scores causes numerical gradients to dominate textual representations during backpropagation (Wang et al., 2020). Second, asynchronous news generated during non-trading market closures is frequently discarded or backward-projected, inducing look-ahead bias (French, 1980). Third, forecasting models remain opaque black boxes lacking explainability and economic opportunity-cost evaluation.

This research introduces Deep News Trader (DNT Engine), a dual-channel multimodal decision support platform engineered to overcome these empirical deficiencies, assessing its directional forecasting and capital preservation capability across the 10 most influential BIST assets. This study is developed from the empirical framework of Tanrikulu (2026), specifically adapting its dual-channel architecture and cognitive attention mechanism for real-time application in Borsa Istanbul.

MATERIAL AND METHODS

Material

The empirical universe comprises 10 highly liquid, information-dense stocks listed on Borsa Istanbul (THYAO, SASA, ASELS, GARAN, AKBNK, TUPRS, BIMAS, EREGL, SISE, and TAVHL), representing key industrial, financial, defense, and transportation sectors. The dataset spans the historical period from January 1, 2021, to February 26, 2026, sampled at daily intervals. Quantitative features consist of 14 technical indicators derived from historical OHLCV data using TA-Lib (e.g., RSI, MACD, Bollinger Bands, VWAP, ATR, Stochastic %K, and moving averages). Unstructured texts were retrieved via official APIs from Marketaux, Yahoo Finance, and X (formerly Twitter), filtered to eliminate informational noise, and quantified into continuous sentiment polarities for English news via FinBERT and for Turkish news via Turkish BERT (TrBERT).

Methods

Data Collection

The empirical framework of this research relies on a synchronized dual-stream data ingestion pipeline comprising both structured numerical market series and unstructured narrative sentiment corpora, spanning January 1, 2021, to February 26, 2026, across daily trading frequencies. Because the

investigation avoids reactive qualitative surveys or clinical interventions, secondary empirical data were harvested autonomously using dedicated Application Programming Interfaces (APIs) and high-throughput web-scraping utilities.

Historical open, high, low, close, and trading volume (OHLCV) time series were procured through the official Yahoo Finance API infrastructure (via Python's yfinance module). To safeguard against the statistical fragility and non-stationarity of raw price inputs, the numerical channel was populated with 14 deterministic technical indicators derived via the Python TA-Lib library, encompassing directional trend indicators (SMA-20, SMA-50, EMA-12, MACD), momentum oscillators (RSI-14, Stochastic %K, Williams %R), volatility bands (Bollinger Upper/Lower, ATR-14), and liquidity tracking indicators (VWAP, OBV, MFI-14). To preserve strict temporal causality and prevent look-ahead bias, unobserved days caused by technical market closures were handled strictly using the Last Observation Carried Forward (LOCF) mechanism without forward-interpolation.

Concurrently, unstructured textual streams were extracted in parallel from institutional and retail media platforms. Corporate financial dispatches and macroeconomic headlines were systematically retrieved via the Marketaux API and the Global Database of Events, Language, and Tone (GDELT) project. Additionally,, microblogging dynamics capturing retail investor sentiment and noise-trader behavior were collected from the official X (Twitter) API, explicitly filtering out automated retweets and bot activity. The extracted Turkish financial text corpora were subsequently tokenized, contextualized, and mapped onto a 768-dimensional latent space utilizing a domain-adapted TrBERT Transformer model, mapping discrete narrative states into continuous daily sentiment polarity vectors. Asynchronous off-market disclosures accumulated over weekend and holiday periods were systematically gathered and transferred to Monday opening tensors using the study-specific Roll-Forward algorithm.

Asynchronous Data Alignment (Roll-Forward Algorithm)

To resolve non-trading-hour information accumulation without violating causality, an algorithmic "Roll-Forward" operator was designed. Asynchronous weekend and holiday sentiment scores are cumulatively aggregated into an opening impulse shock vector for Monday's trading session:

$$S_{accumulated} = \frac{1}{\sum_{i=1}^{N} W_i} \sum_{i=1}^{N} (S_i \cdot W_i)$$

where N denotes the aggregate volume of off-market textual releases, S_i represents the discrete TrBERT sentiment score, and W_i represents source confidence or engagement weight.

Dual-Channel Neural Topology and Intermediate Fusion

To circumvent gradient dominance, the architecture isolates structural modalities prior to deep fusion:

Technical Channel (Momentum SubNet): Numerical indicators pass through a dilated 1D-CNN (breakout filter) followed by an LSTM block to preserve temporal momentum dependencies.

Sentiment Channel (NLP SubNet): Textual sentiment vectors are filtered through a 1D-CNN (shock isolator) and modeled via Bidirectional LSTM (BiLSTM) to capture shifting crowd sentiment.

The extracted latent representations h_{tech} and h_{sent} are fused via intermediate layer concatenation into a joint latent vector space (Z):

$$Z = \sigma(W \cdot [h_{tech} \| h_{sent}]) + b$$

Dual-Channel Neural Topology and Intermediate Fusion

An intermediate Self-Attention mechanism (Vaswani et al., 2017) dynamically balances modal reliance across varying market regimes. Normalized attention weights a_t compute the composite context vector c_t:

$$e_t = \tanh(W_h h_t + b_h), \qquad a_t = \frac{\exp(e_t)}{\sum_{i=1}^{T} \exp(e_t)}, \qquad c_t = \sum_{j=1}^{T_x} a_{tj} h_j$$

Soft Directional MSE Loss Function

To align network optimization directly with financial trading performance, directional forecasting errors are asymmetrically penalized using a directional regularizer $\lambda = 0.5$:

$$L_{SD-MSE} = \frac{1}{N} \sum_{t=1}^{N} (y_t - \hat{y}_t)^2 \cdot (1 + \lambda \,.\, I_t), \qquad I_t = \begin{cases} 1, \text{if } y_t \cdot \hat{y}_t < 0 \\ 0, \text{if } y_t \cdot \hat{y}_t \geq 0 \end{cases}$$

Meta-Learning Noise Regressor

A dynamic meta-learning regulator continuously monitors the asset-specific noise coefficient (Γ_t). In high-noise, retail-driven assets (e.g., SASA), the regulator switches intermediate activations to saturating Sigmoid geometries to dampen false breakouts, whereas in institutional, low-noise regimes (e.g., ASELS), non-saturating LeakyReLU/ELU activations remain active.

RESULTS

The empirical directional accuracy (Mean Directional Accuracy - MDA) and statistical Mean Squared Error (MSE) metrics across the 10 selected BIST equities are compiled in Table 1.

Table 1. Predictive performance comparison of single-channel vs. dual-channel architectures across bist assets

Ticker	Sectoral Characteristics	Optimal Activation	Technical Only MDA (%)	Sentiment Only MDA (%)	Dual-Channel Hybrid MDA (%)	Hybrid MSE

THYAO	Aviation / Geopolitical	ELU	49.40	50.60	**56.63**	0.000424
ASELS	Defense / Technology	ReLU	46.59	53.41	**55.82**	0.001047
GARAN	Banking / Institutional	ELU	52.61	49.40	**55.45**	0.000592
EREGL	Basic Metal / Commodity	Tanh	55.42	50.60	**60.24**	0.000534
SISE	Glass / Industrial	Tanh	44.58	46.59	**54.62**	0.000541
BIMAS	Food Retail / Defensive	ELU	52.61	49.00	**53.41**	0.000544
AKBNK	Banking / Corporate	Sigmoid	51.81	52.61	**52.61**	0.000662
TUPRS	Energy Refining	Sigmoid	52.61	52.61	**52.61**	0.000469
TAVHL	Airport Infrastructure	ReLU	47.79	50.60	**51.00**	0.000482
SASA	Synthetic Fibers / Speculative	LeakyReLU	47.39	46.59	**50.20**	0.000994

Note: Evaluated across an independent 360-day testing window. The hybrid architecture surpasses the 50% random walk benchmark across all investigated instruments.

As demonstrated in Table 1, intermediate multimodal fusion outperforms singular modeling frameworks across all tested tickers. Notable directional improvements occur in news-sensitive assets such as ASELS (+9.23 percentage points over technical baseline) and SISE (+10.04 percentage points). Implementing the sentiment noise filter ($\tau = 0.15$) on EREGL yielded an accuracy of 60.24%, confirming that isolating substantive macroeconomic commodity news from background chatter enhances statistical convergence.

SHAP interpretability analysis (Lundberg and Lee, 2017) corroborates that during calm market regimes, technical indicators (VWAP, Bollinger Bands) dominate model outputs. However, during extreme informational spikes, attention distributions invert, shifting dominant feature attribution to TrBERT narrative sentiment. To deploy these algorithmic mechanics into an interactive decision support tool, the Deep News Trader (DNT Engine) cloud platform was implemented (Figure 1).

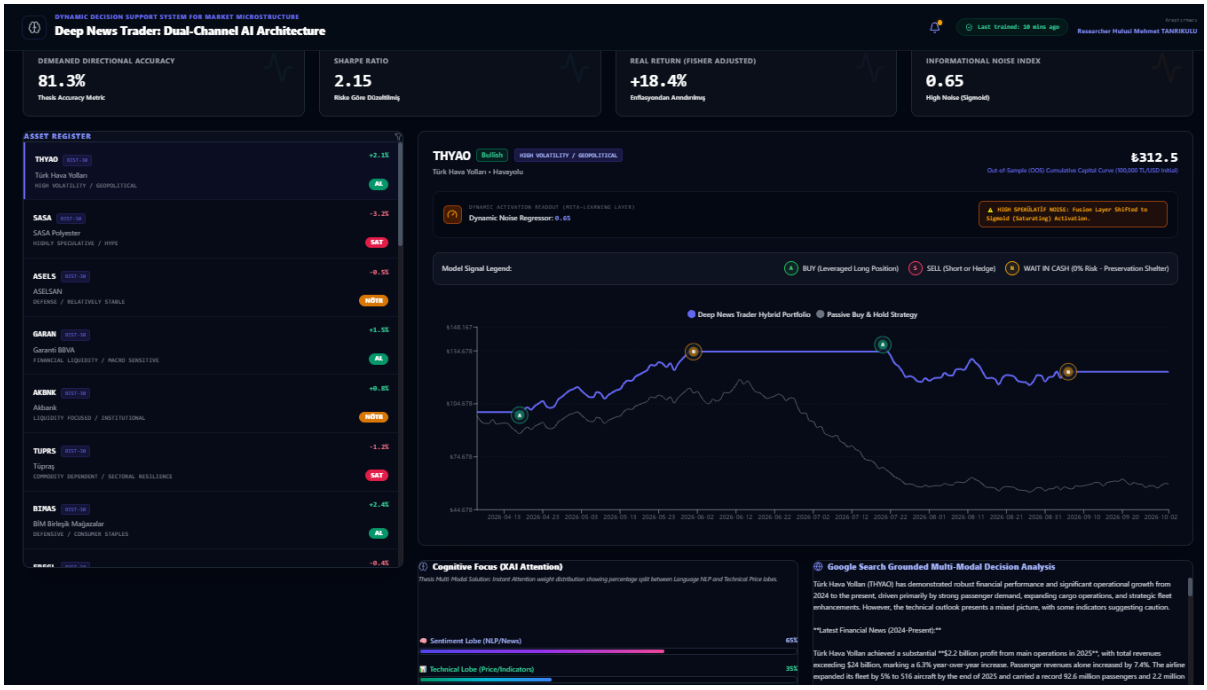


Figure 1. Operational User Interface and Real-Time Analytical Dashboard of Deep News Trader (DNT Engine)

The interactive interface displayed in Figure 1 communicates the asset's active trading signal (BUY/HOLD/SELL), FinBERT/TrBERT sentiment polarity, real-time noise coefficient, and current activation regime (Sigmoid vs. LeakyReLU). The Cognitive Focus module visualizes the real-time proportional weighting allocated between the Sentiment Channel and Technical Channel. Additionally, multi-agent contextual verification powered by Gemini 3.8 Flash, grounded in live web searching, executes automated qualitative synthesis of breaking market catalysts.

The model's stress-testing capability was validated over a 40-day Out-of-Sample (OOS) Crisis Test (March 1 – April 24, 2026), coinciding with heightened Middle Eastern geopolitical volatility. The financial risk performance metrics are reported in Table 2.

Table 2. Out-of-Sample (OOS) crisis performance and capital drawdown simulation across representative bıst assets

			0%		1%		10%	
Ticker	**Crisis MDA (%)**	**MSE**	**Max DD (%)**	**Ending Equity (TRY)**	**Max DD (%)**	**Ending Equity (TRY)**	**Max DD (%)**	**Ending Equity (TRY)**
AKBNK	50.00	0.000910	0.0	173929.61	-4.19	149441.75	-7.00	129399.13
ASELS	57.89	0.001207	0.0	147665.73	-6.01	125322.41	-25.59	78670.20
BIMAS	52.63	0.000557	0.0	149341.35	-1.99	129638.21	-5.84	111564.12
EREGL	47.37	0.000712	0.0	158522.78	-3.81	134013.22	-8.25	106666.66
GARAN	42.11	0.000764	0.0	143729.85	-4.85	118231.97	-11.06	97214.07

SASA	60.53	0.001240	0.0	194860.65	-2.97	166177.24	-6.05	139318.66
SISE	57.89	0.000742	0.0	140631.9	-4.48	118434.2	-18.35	95864.13
TAVHL	55.26	0.000942	0.0	174096.99	-1.67	153029.16	-7.61	122786.89
THYAO	50.00	0.000501	0.0	129694.91	-2.76	111929.27	-12.34	95537.6
TUPRS	47.37	0.001616	0.0	133888.17	-3.17	109683.41	-21.32	76392.95

Note: Initial simulated equity set to 100,000 TRY. 0%, 1%, 10% represent execution threshold parameters for risk control. The unconstrained (0% threshold) model executes defensive cash parking upon detecting a negative sentiment shock, suppressing maximum drawdown to 0.00%.

The cognitive mechanism driving this capital preservation behavior during crisis shocks is illustrated for THYAO in Figure 2.

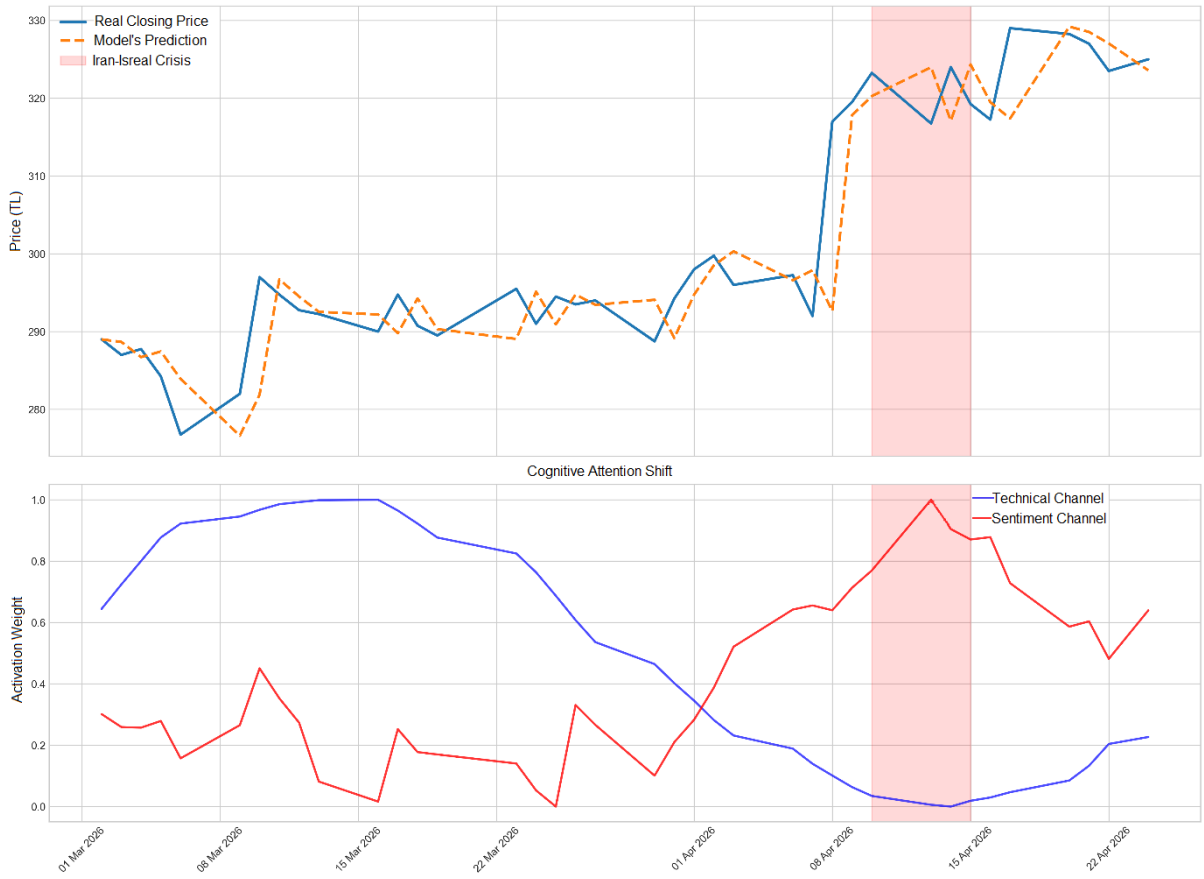


Figure 2. THYAO Out-of-Sample Crisis Trajectory and Dynamic Cognitive Attention Shift

As shown in Figure 2, prior to the geopolitical escalation, the attention weights favored the Technical Channel. During the conflict escalation (April 10–15, 2026), the Self-Attention mechanism reduced technical weights and shifted focus directly to the Sentiment Channel, adjusting directional forecasts downward before price momentum confirmed the decline.

The economic impact of this risk management capability is depicted for TAVHL in Figure 3. While passive buy-and-hold strategies incurred severe drawdowns during the shock interval, DNT Engine's unconstrained setup transitioned into protective cash positions, ending the stress period with 174,096.99 TRY and 0.00% drawdown.

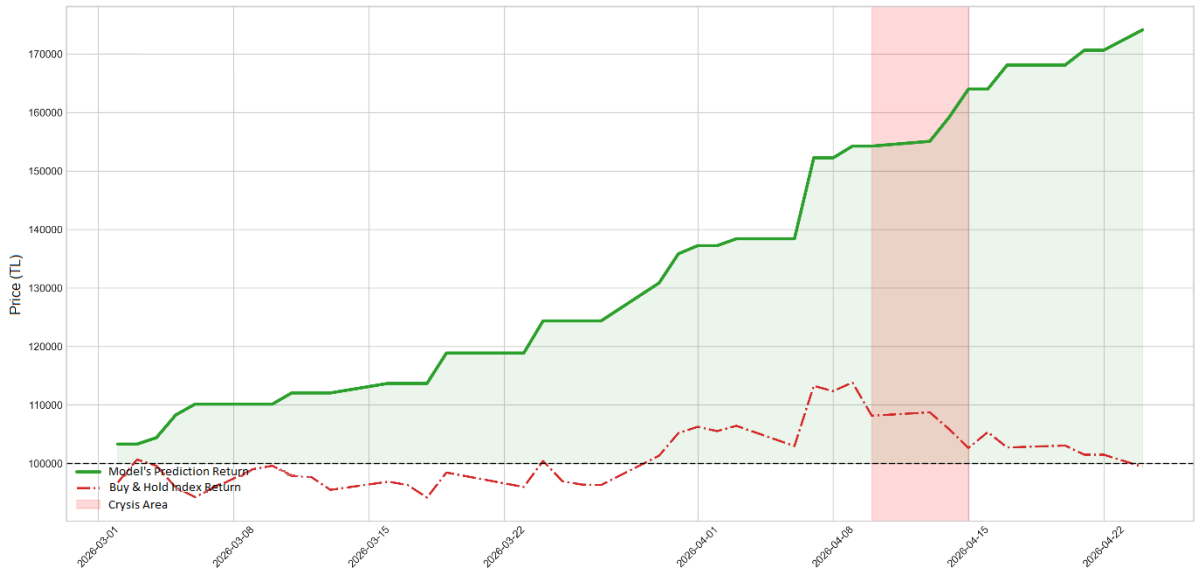


Figure 3. TAVHL Comparative Equity Curve Under Geopolitical Stress Simulation

Figure 4 illustrates that the capital curves generated for TAVHL—a stock directly exposed to external macroeconomic shocks, geopolitical vulnerabilities, and systemic risks that particularly reflect the dynamics of the aviation sector—demonstrate the model's emphasis on the sentiment channel in response to the influx of news data, as well as its capacity for a defensive reaction during times of crisis.

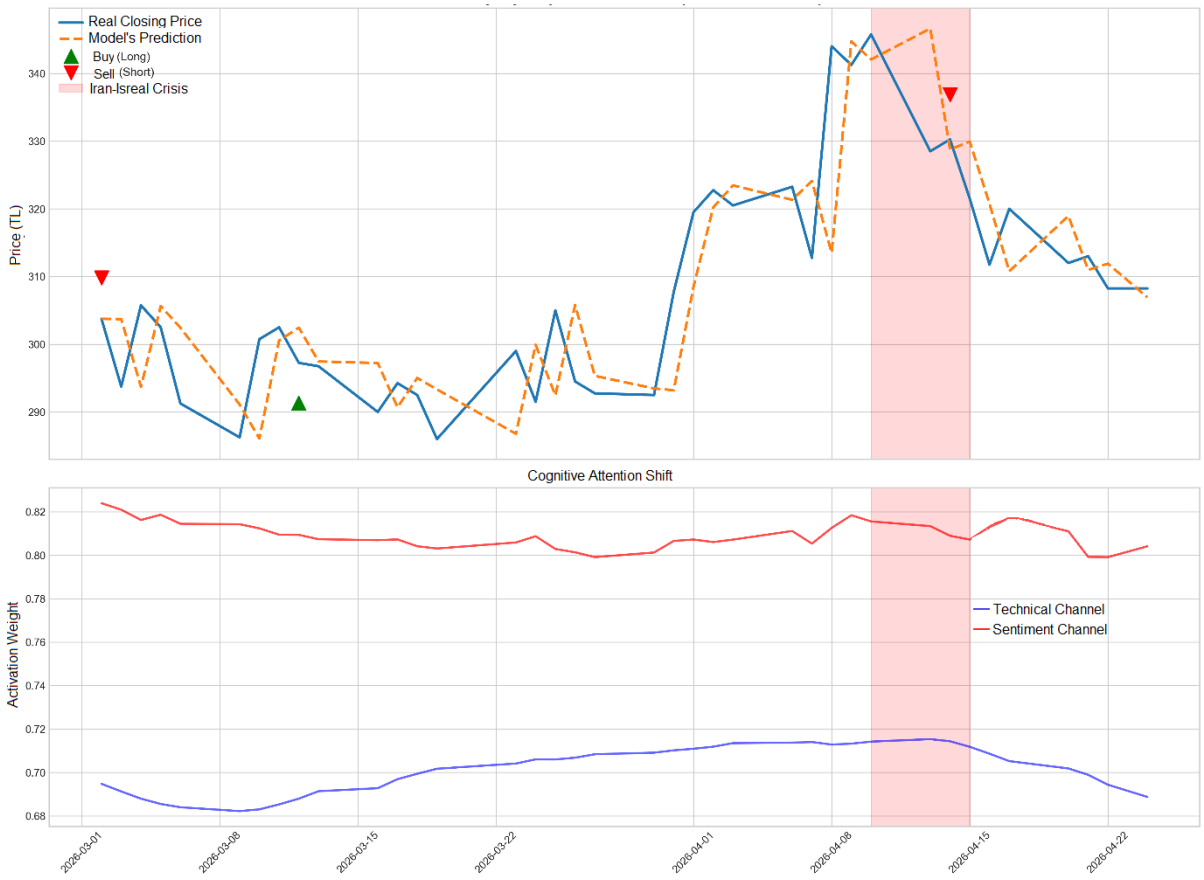


Figure 4. TAVHL Out-of-Sample Crisis Trajectory and Dynamic Cognitive Attention Shift

DISCUSSION AND CONCLUSION

This empirical research demonstrates that integrating structured financial momentum with unstructured narrative sentiment via intermediate fusion improves directional accuracy in emerging equity markets.

The Efficient Market Hypothesis suggests that markets quickly incorporate information. However, information delays and investor overreactions may create short-lived inefficiencies that multimodal models can identify. During periods of market stress, the proposed self-attention mechanism increases the weight assigned to sentiment information, helping limit downside risk. In unconstrained testing, the maximum drawdown was 0.00%. The roll-forward preprocessing method also helps prevent look-ahead bias and accounts for market gaps that may follow the accumulation of news over weekends.

In accordance with the "No Free Lunch" theorem (Wolpert and Macready, 1997), the findings confirm that universal activation functions underperform relative to asset-specific calibrations: saturating activations stabilize speculative noise in retail-heavy stocks, while non-saturating functions prevent gradient attenuation in corporate equities. The developed Deep News Trader (DNT Engine) platform operationalizes these deep learning models into an explainable, real-time decision support environment. Future research should investigate integrating analyst reports and central bank

communications via Retrieval-Augmented Generation (RAG) to incorporate longer-term macroeconomic perspectives.

References

Araci D, 2019. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. arXiv preprint arXiv:1908.10063.

Baltrusaitis T, Ahuja C, Morency LP, 2019. Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2): 423-443.

De Long JB, Shleifer A, Summers LH, Waldmann RJ, 1990. Noise trader risk in financial markets. Journal of Political Economy, 98(4): 703-738.

Devlin J, Chang MW, Lee K, Toutanova K, 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT, pp. 4171-4186.

Fama EF, 1970. Efficient capital markets: A review of theory and empirical work. The Journal of Finance, 25(2): 383-417.

French KR, 1980. Stock returns and the weekend effect. Journal of Financial Economics, 8(1): 55-69.

Gao Y, Wang R, Zhou E, 2021. Stock prediction based on optimized LSTM and GRU models. Scientific Programming, 2021: 1-8.

Kahneman D, Tversky A, 1979. Prospect theory: An analysis of decision under risk. Econometrica, 47(2): 263-291.

Lundberg SM, Lee SI, 2017. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NeurIPS), 30: 4765-4774.

Tanrikulu HM, 2026. Geliştirilmiş Hibrit Derin Öğrenme Mimarisi ile Finansal Piyasaların Analizi, Doctorate Thesis. Bayburt University, Institute of Graduate Studies, Bayburt, Türkiye.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I, 2017. Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS), 30: 5998-6008.

Wang Y, Yao Q, Kwok JT, Ni LM, 2020. Generalizing from a few examples: A survey on few-shot learning. ACM Computing Surveys, 53(3): 1-34.

Wolpert DH, Macready WG, 1997. No free lunch theorems for optimization. IEEE Transactions on Evolutionary Computation, 1(1): 67-82.

Acknowledgment

This research was supported by the Scientific Research Projects (BAP) Coordination Unit of Atatürk University under the Regional Universities Research, Development and Cooperation Project (BÖGEP) framework.

Conflict of Interest

The authors declare that there is no conflict of interest.

Author Contributions

The authors collaboratively carried out the data collection and cleaning, sentiment analysis and feature selection, model development, testing, and writing stages within the scope of the project. The authors utilized generative artificial intelligence (AI) tools exclusively for language enhancement, grammar correction, and editorial assistance during the preparation of this manuscript. The AI was not used to generate data, draw scientific conclusions, or perform fundamental research.

A Hybrid Machine Learning Approach for Bank Stability Prediction

Sauda Nerjaku[1*], Valentina Sinaj[2]

[1,2] University of Tirana, Faculty of Economy, Department of Statistics and Applied Informatics, 1001, Tirana, Albania

*Corresponding author e-mail: sauda.nerjaku@gmail.com

Abstract

Financial stability represents a fundamental pillar of a country's economic performance, with banking stability serving as one of its core components. In light of recent global geopolitical developments and associated macroeconomic shocks, accurate prediction of bank stability has become increasingly important for bank managers. In this study, bank stability is measured using the Z-score based on panel data from selected Balkan countries over the period 2010–2021. The explanatory variables include both macroeconomic indicators and bank-specific factors. This research evaluates the predictive performance of several machine learning models, including Linear Regression, Support Vector Regression, Random Forest, and Gradient Boosting, together with a hybrid ensemble model. The proposed hybrid model combines the two best-performing individual models Linear Regression and Support Vector Regression within a Voting Regressor framework. Model weights are selected empirically through temporal validation, with weights ranging from 0.2 to 0.8, in order to identify the combination that provides the best predictive performance for bank stability. The findings indicate that the hybrid approach LR-SVR achieved the highest R^2 value of 0.94 and the lowest RMSE of 1.05, demonstrating improved predictive accuracy compared with the standalone models. These results suggest that combining linear and nonlinear modelling approaches can provide a more effective framework for forecasting bank stability and may offer practical support for financial institutions and policymakers in monitoring and assessing banking-sector resilience.

Keywords: *Hybrid Machine Learning, Bank Stability, Z-score, Balkan Countries*

INTRODUCTION

Financial stability is essential for economic development, as a stable financial system enables the efficient allocation of resources, reduces financial risks, and withstands economic shocks (Shahriar et al., 2023) Instability in financial markets can transmit to the broader economy, causing further weakening of the financial system. For this reason, policymakers focus on developing comprehensive indicators and forecasting tools to monitor risks and reduce the likelihood of banking and financial crises (Morris, 2011). The energy crisis, the COVID-19 pandemic and the war in Ukraine have negatively impacted the financial performance of banks in the Western Balkans. These crises have reduced profitability and are expected to continue to damage the stability and liquidity of the banking sector (Sabljic et al., 2023), (Dedeloudis et al., 2025). Amid ongoing geopolitical tensions and associated

macroeconomic shocks, accurately predicting bank stability has become increasingly important for bank managers.

Accordingly, this study aims to evaluate and compare individual machine learning models for predicting bank Z-scores, a proxy for banking stability, across selected Balkan countries over the period 2010–2021, and to develop a hybrid model and assess its predictive performance against the individual models.

The ability of artificial intelligence to predict banking risk more accurately than traditional econometric models has attracted growing debate in the financial services sector. (Farag et al., 2026) For instance, (Sonehekpon, 2025) examines the impact of artificial intelligence (AI) investment on banking stability in developed countries and finds that AI contributes positively to financial stability. The results further indicate that policies such as inflation control, AI adoption in risk management and credit assessment, and improved financial education can strengthen banking stability. The study of (Alessi and Savona, 2021) develops an early warning system for systemic banking crises and indicate that RF performs better than logit models. The findings of (Theofilis et al., 2024) revealed that among 25 machine learning model SVM and bagged trees exhibit the highest effectiveness in predicting banking stability based on the CAMELS ratings, offering high accuracy and reliability. Additionally, (Nerjaku and Sinaj, 2024) demonstrate that SVM outperform random forests and artificial neural networks. (Sen, 2024) reveal that Random Forest models perform much better than boosting, achieving very high accuracy. (Seyedrahimi-Niaraq et al., 2025) found that Gradient Boosting performed best among the individual models, while a Voting ensemble combining Gradient Boosting, Linear Regression and Decision Tree further improved predictive performance, outperforming all individual models across the evaluated metrics. Furthermore, in the study by (Ekinci and Erdal, 2017), hybrid ensembles achieved higher predictive accuracy than both standalone models and standard ensemble approaches for bank failure prediction. (Bastos, 2014) showed that ensembles of models based on the same regression method predicted bond and loan recovery rates more accurately than individual models. Several studies have reported that ensemble and hybrid models can outperform individual models in terms of predictive accuracy across a range of applications. (Bastos, 2013; Erdal, 2013; Olufemi Ogunleye, 2023; Pusporani, et al., 2025; Nguyen and Le, 2026). (Boussatta et al., 2025) found that hybrid models provided reliable predictions under stable market conditions but faced substantial challenges during periods of geopolitical volatility.

MATERIAL AND METHODS

Material

This study aims to predict banking stability using machine learning techniques and a weighted voting regression ensemble. The analysis focuses on seven selected countries in the Balkan region: Albania, Greece, Bosnia and Herzegovina, Serbia, Croatia, North Macedonia, and Montenegro. The dataset comprises annual observations from selected countries over the period 2010-2021, resulting in a panel dataset of 84 country-year observations. Consequently, the dataset does not contain bank-level observations, and it is therefore not possible to directly identify or report the number of banks represented in each country-year. The study incorporates both macroeconomic and bank-specific explanatory variables, including non-performing loans, GDP growth, inflation, return on equity, return

on assets, bank non-interest income to total income, interest margin to gross income, unemployment rate, non-interest expenses to gross income, regulatory capital to risk-weighted assets (CAR), and bank credit to bank deposits. The inclusion of these variables is supported by substantial evidence in the existing literature indicating that banking stability is closely associated with both macroeconomic conditions and bank-specific characteristics (Farag et al., 2026). The bank Z-score is employed as a proxy for banking stability. The data were obtained from the World Bank and the International Monetary Fund (IMF).

Methods

Data Preprocessing

One of the most important steps in developing a machine learning model is data preprocessing, which aims to prepare the dataset for accurate and reliable analysis. First, observations with missing values in the target variable were removed, while missing values in the explanatory variables were replaced with their respective median values. Subsequently, non-numeric variables were converted into a numerical format using techniques such as one-hot encoding. Finally, feature standardization was applied, which is particularly important for the performance of Support Vector Regression (SVR) and Linear Regression (LR). The dataset was divided chronologically into training and testing sets using a 75:25 ratio, with 75% of the observations from the earlier years used for model training, and 25% from the later years used for testing. All observations from the same year were kept together in their respective set, in order to establish a temporal panel structure. A preprocessing pipeline was implemented to ensure that scaling parameters were fitted exclusively on the training data, preventing data leakage during this step and supporting a more robust evaluation. Nevertheless, the possibility of optimistic performance estimates due to the limited sample size cannot be ruled out.

Machine Learning Methods

The models employed in this study include Linear Regression (LR), Support Vector Regression (SVR), Gradient Boosting (GB), and Random Forest (RF), along with a hybrid model that combines the two best-performing individual models. To compare their predictive performance and identify the most effective approach, the following evaluation metrics were employed (Eqs. 1 - 4):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (1)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad (2)$$

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(\bar{y} - y_i)^2} \qquad (3)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{(y_i - \hat{y}_i)}{y_i}\right| * 100 \qquad (4)$$

Here $(\hat{y}_i)$ is the predicted value, (y_i) actual value and $\bar{y}$ is the mean of observed value. Data preprocessing, model training, optimal weight selection for the weighted voting regressor, and performance evaluation, were implemented in Python using the Scikit-learn library.

Multiple Linear Regression (LR) is a statistical method that models the relationship between a dependent variable and multiple independent variables by assuming a linear relationship. The dependent variable is expressed as a linear combination of predictors, each associated with a regression coefficient representing its contribution while holding other variables constant. The model estimates these coefficients by minimizing the sum of squared errors between actual and predicted values. Due to its simplicity and interpretability, MLR is widely used as a baseline model for regression and predictive analysis. (Faouzi and Colliot., 2022; K and G, 2026)

Gradient Boosting is an ensemble machine learning algorithm that combines multiple weak learners, typically decision trees, to develop a more accurate predictive model. Unlike Random Forest, which builds trees independently, Gradient Boosting constructs trees sequentially, with each new tree aiming to correct the errors of the previous models by minimizing a loss function. This iterative process improves predictive accuracy and enables the model to capture complex and nonlinear relationships. However, Gradient Boosting can be computationally demanding and prone to overfitting if its parameters are not properly optimized. (Hanicinec et al., 2023; Tu et al., 2025; K and G, 2026)

Support Vector Regression (SVR) is a supervised machine learning algorithm designed to predict continuous outcomes by identifying a function that minimizes prediction errors within a predefined tolerance margin (ε). The model uses an ε-insensitive loss function, in which errors within this margin are not penalized, while larger deviations contribute to the optimization objective. The regularization parameter C controls the trade-off between model complexity and deviations beyond this margin. Furthermore, SVR can capture nonlinear relationships through kernel functions, such as the Radial Basis Function (RBF), making it suitable for complex regression problems. (Smola and Scholkopf, 2004; Gupta et al., 2019)

Random Forest (RF) is an ensemble learning method that combines multiple decision trees, each constructed using independently sampled random components. The generalization error converges to a limiting value as the number of trees increases, while predictive performance depends on the strength of individual trees and the correlation between them. In regression tasks, the final prediction is obtained by averaging the outputs of individual trees, helping to reduce variance and the risk of overfitting. (Breiman, 2001)

Weighted Voting Regressor is an ensemble learning method that combines predictions from multiple regression models to produce a single final output. This approach aims to leverage the strengths of individual models while reducing their weaknesses, potentially improving predictive accuracy, robustness, and generalization. The final prediction can be obtained through simple averaging or weighted averaging, where different weights are assigned to individual models according to their relative importance. However, weighted averaging does not necessarily outperform simple averaging, as the selection of appropriate weights can be challenging, particularly when the available data are noisy or insufficient. (Zhou, 2012; Natras et al., 2022; Şimşek et al., 2025)

RESULTS

This section presents the results of the machine learning models developed to predict bank Z-scores. The analysis begins by evaluating the individual models on the test set using RMSE, MAE, R^2, and MAPE, with the results summarized in the **Table 1**.

Table 1. Predictive performance of the models

Model	RMSE	MAE	R^2	MAPE
Linear Regression	1.5095	1.2420	0.8783	11.5453
Support Vector Regression	1.9014	1.3074	0.8069	11.4002
Gradient Boosting	3.9227	2.1346	0.1784	25.4701
Random Forest	4.1960	2.4424	0.0599	29.6969

As shown in the table, Linear Regression and Support Vector Regressor achieved the best predictive performance across all evaluated metrics, recording the lowest RMSE of 1.5095 and 1.9014, respectively. Random Forest recorded the highest RMSE and a low test-set R^2. Although Gradient Boosting outperformed Random Forest, its low test-set R^2 indicates a limited ability to explain the variation in bank Z-scores. These findings suggest that the tree-based ensemble models considered, were less effective in predicting bank Z-scores for this dataset.

Given their superior predictive performance among the individual models, LR and SVR were combined into a hybrid model using a weighted average of their predictions. (Chen and Zheng, 2025) suggests that the RRMSE Voting Regressor can improve predictive accuracy by combining the strengths of individual models within an ensemble framework. The hybrid prediction for a given sample x_i is formulated as follows:

$$Hybrid\ model = wf_{LR}(x_i) + (1 - w)f_{SVR}(x_i)$$

The optimal weights (w) of the hybrid LR–SVR model were selected through temporal cross-validation within the training set by minimizing RMSE. The final model was evaluated on the test set, which was not used for model or weight selection. As illustrated in the **Figure 1** below, the lowest CV RMSE (≈ 1.05) was achieved with weights of 0.55 and 0.45 for LR and SVR, respectively.

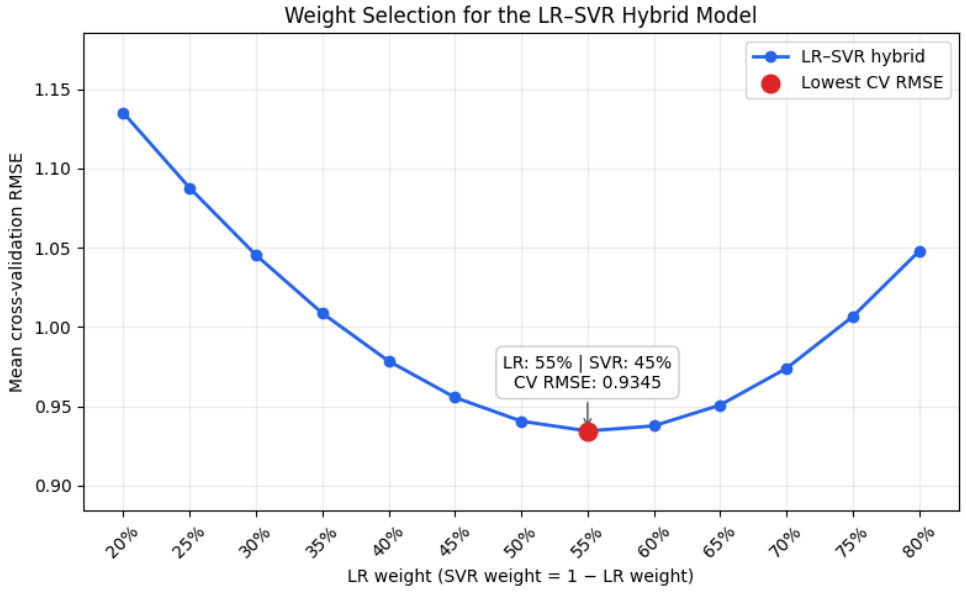


Figure 1. Weight Selection for the hybrid model

The hybrid model outperforms the individual models, achieving a higher $R^2 \approx 94\%$ and the lowest RMSE ≈ 1.05 and MAE ≈ 0.84. These results indicate greater explanatory power and improved predictive accuracy compared with the standalone models. The performance metrics of the hybrid model are presented in the **Table 2**.

Table 2. Predictive performance of the hybrid model,

Model	RMSE	MAE	R^2	MAPE
Hybrid Model	1.0551	0.8452	0.9406	7.7374

These findings are further supported by **Figure 2**, which compares the RMSE values of all models considered in this study and shows that the hybrid model achieves the lowest prediction error.

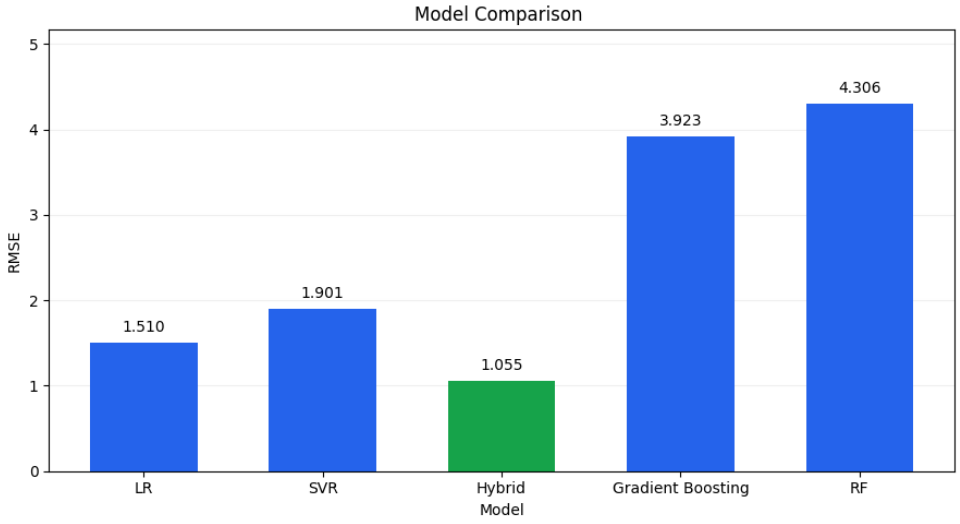


Figure 2. Test RMSE Comparison.

The Figure 3 below compares model performance using a MAPE-based accuracy measure, calculated as 100% − MAPE (%), with the hybrid model achieving the highest value. Similarly, (Munandar et al., 2025) suggests that hybridization can mitigate both underfitting and overfitting, producing more stable and responsive forecasts.

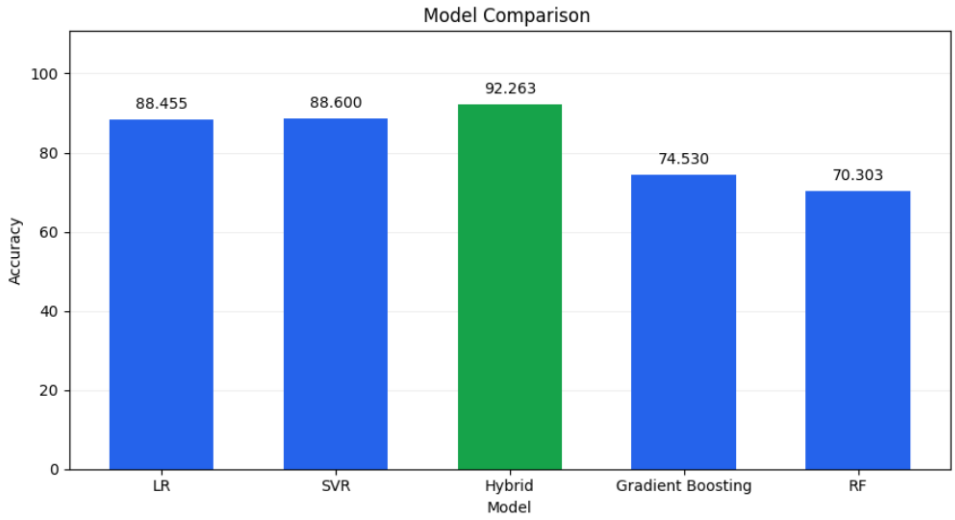


Figure 3. Comparison of model accuracy calculated as 100 − MAPE (%).

The actual versus predicted plots in **Figure 4** further illustrate the differences in predictive performance across models.

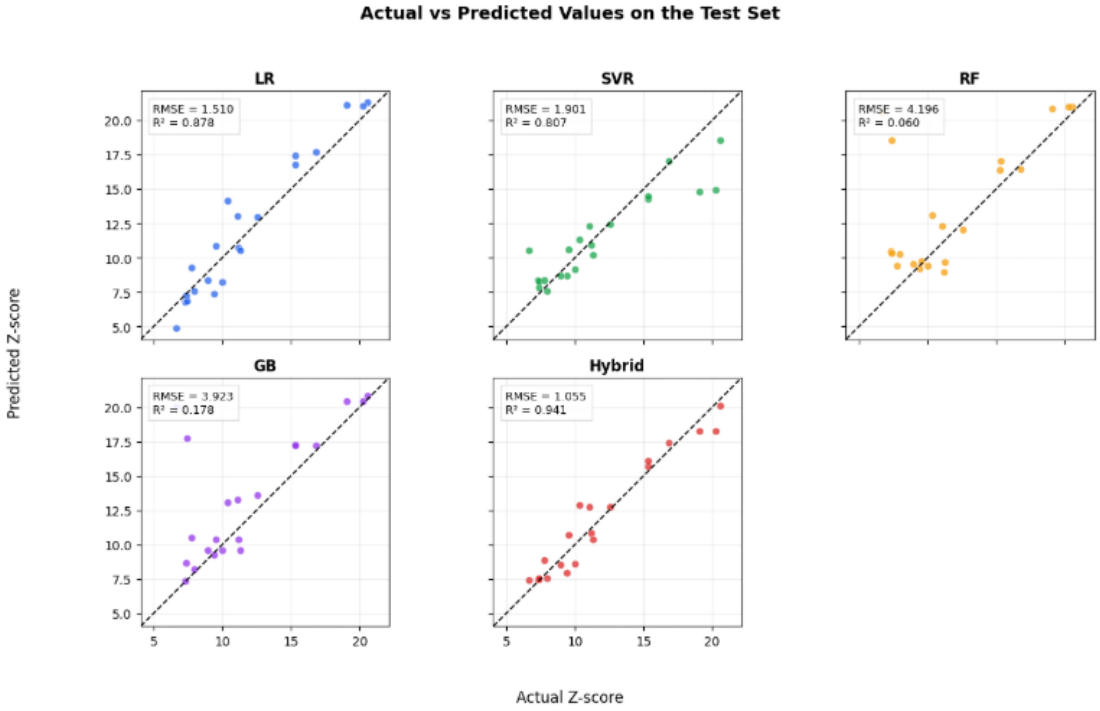


Figure 4. Scatter plots of the models.

The hybrid model exhibits points more closely aligned with the identity line, indicating stronger agreement between observed and predicted values and greater predictive accuracy. (Qin, Fan, & Zhang,

2024) demonstrated that hybrid models can improve predictive accuracy by combining complementary information from individual models.

DISCUSSION AND CONCLUSION

Machine learning provides valuable tools for predicting banking indicators, as many algorithms can capture complex and nonlinear relationships. In this study, machine learning methods were applied to predict banking stability across selected Balkan countries. Among the individual models, Linear Regression and SVR achieved better predictive performance than Gradient Boosting and Random Forest. Similiary (Olufemi Ogunleye, 2023) reported that Linear Regression outperformed Random Forest, achieving a lower RMSE in a comparative predictive analysis. Furthermore (Süzen and Kayakuş, 2026) suggest that SVR outperforms both Random Forest and Gradient Boosting in terms of predictive accuracy. The performance of boosting algorithms can depend on sample size, hyperparameter tuning, feature characteristics, and the signal-to-noise ratio. Their comparatively weaker results in this study may therefore reflect the limited sample size and specific characteristics of the dataset rather than inherent limitations of the algorithms (Süzen and Kayakuş, 2026).

In contrast, (Munandar et al., 2025) reported weaker performance for Linear Regression and SVR, which tended to underestimate or smooth out abrupt changes in demand, resulting in delayed peak detection and an inadequate representation of variability. Similarly, (Farag et al., 2026) reported that Gradient Boosting outperformed other machine learning methods, attributing its superior performance to its ability to capture complex nonlinear relationships and feature interactions in financial data. Other studies have also reported strong predictive performance for Random Forest, which outperformed the alternative models evaluated (Nerjaku & Sinaj, 2026a; Nerjaku & Sinaj, 2026b).

The hybrid model combined the predictions of a linear and a nonlinear model using weights that sum to one. It outperformed all individual models, achieving the lowest RMSE and MAE values of 1.05 and 0.84, respectively, alongside an R^2 of 0.94 and a MAPE-based accuracy of 92%. These findings are consistent with the previously cited literature highlighting the potential benefits of combining linear and nonlinear approaches. (K and G, 2026) found that Voting and Stacking regressors outperformed individual models across several evaluation metrics, while (Thabet et al., 2024) reported that a Super Learner ensemble achieved superior performance in predicting bank efficiency.

Nevertheless, these results should be interpreted in light of the dataset's characteristics and limited sample size, which may affect their generalizability. The proposed hybrid approach could support bank managers and policymakers in assessing banking stability, although further validation across larger datasets and different economic conditions is needed. Overall, the study highlights the potential of machine learning, particularly hybrid models, to improve the prediction of banking stability. (Farag et al., 2026) further recommend that regulators draw on machine learning findings to guide policies supporting banking stability.

This study has several limitations, particularly its relatively small sample size. Future research could incorporate larger datasets covering a longer period to assess the robustness and generalizability of the findings. Exploring additional machine learning algorithms and alternative ensemble approaches, such as stacking with a meta-learner alongside weighted voting, could further improve predictive performance.

References

Alessi, L., & Savona, R. (2021). Machine Learning for Financial Stability. S. R. Consoli içinde, Data Science for Economics and Finance (s. 65–87). Springer, Cham. doi:10.1007/978-3-030-66891-4_4

Bastos, J. A. (2013). Ensemble Predictions of Recovery Rates. Journal of Financial Services Research, 46, 177–193. doi:10.1007/s10693-013-0165-3

Bastos, J. A. (2014). Ensemble Predictions of Recovery Rates. Journal of Financial Services Research, 46(2), 177-193. doi:10.1007/s10693-013-0165-3

Boussatta, H., Chihab, M., Chiny, M., & Chihab, Y. (2025). Predicting Oil Price Trends During Conflict With Hybrid Machine Learning Techniques. Applied Computational Intelligence and Soft Computing, 2025. doi:10.1155/acis/8867520

Breiman, L. (2001, October). Random Forests. Machine Learning , s. 5–32.

Chen, S., & Zheng, W. (2025). RRMSE-enhanced weighted voting regressor for improved ensemble regression. PLOS ONE, 20(3). doi:10.1371/journal.pone.0319515

Dedeloudis, G., Lois, P., & Repousis, S. (2025). Banking Supervision and Risk Management in Times of Crisis: Evidence from Greece's Systemic Banks (2015–2024). Journal of Risk and Financial Management, 18(7), 386. doi:10.3390/jrfm18070386

Ekinci, A., & Erdal, H. (2017). Forecasting Bank Failure: Base Learners, Ensembles and Hybrid Ensembles. Computational Economics , 49, 677–686 . doi:10.1007/s10614-016-9623-y

Erdal, H. I. (2013). Two-level and hybrid ensembles of decision trees for high performance concrete compressive strength prediction. Engineering Applications of Artificial IntelligencE, 26(7), 1689-1697. doi:10.1016/j.engappai.2013.03.014

Faouzi, J., & Colliot., O. (2022). Classic machine learning algorithms. Machine Learning for Brain Disorders, Springer, In press. ffhal-03830094v1f.

Farag, K., Ibeh, L., & Mutai, N. (2026). Machine Learning for Predicting Bank Stability: Evidence from the Middle East & North Africa Region. Ekonomika, 105(2), 40-62. doi:https://doi.org/10.15388/Ekon.2026.105.2.3

Gupta, D., Pratama, M., Ma, Z., Li, J., & Prasad, M. (2019). Financial time series forecasting using twin support vector regression. PLoS One, 14(3). doi:10.1371/journal.pone.0211402

Hanicinec, M., Mohr, S., & Tennyson, J. (2023). A regression model for plasma reaction kinetics. Journal of Physics D: Applied Physics, 56, 1-24. doi:10.1088/1361-6463/acd390

IMF Data Explorer https://data.imf.org/en/Data-Explorer (accessed on 1 April 2026).

K, M. L., & G, M. R. (2026). Regression based hybrid machine learning model performance evaluation on air quality index prediction in Kolkata. Scientific Reports , 16. doi:10.1038/s41598-026-54918-x

Morris, V. C. (2011). Measuring and forecasting financial stability: The composition of an aggregate financial stability index for Jamaica. Business, Finance & Economics in Emerging Economies, 6(2), 35-51.

Munandar, T., Member, I., & Surbakti, H. (2025). A Novel Hybrid AdaBoost–Gradient Boosting Ensemble for Enhanced Short-Term Energy Consumption Forecasting. IAENG International Journal of Computer Science, 52(12), 4756-4765.

Natras, R., Soja, B., & Schmidt, M. (2022). Ensemble Machine Learning of Random Forest, AdaBoost and XGBoost for Vertical Total Electron Content Forecasting. Remote Sensing, 14(15). doi:10.3390/rs14153547

Nerjaku, S., & Sinaj, V. (2024). NPL modeling using machine learning techniques. Case of Albania. (s. 185-193). Tekirdağ, Türkiye: In Proceedings of the INTERNATIONAL CONGRESS OFRECYCLING ECONOMY & SUSTAINABILITY POLICY-II.

Nerjaku, S., & Sinaj, V. (2026). Application of machine learning algorithms for NPL prediction based on financial data. The case of Albania. Multidisciplinary Science Journal, 8(7). doi:10.31893/multiscience.2026446

Nerjaku, S., & Sinaj, V. (2026). Determinants of Bank Profitability in Selected Balkan Countries: A Combined Econometric and Machine Learning Approach. Journal of Risk and Financial Management (JRFM), 19(8). doi:10.3390/jrfm19080621

Nguyen, N. M., & Le, A. H. (2026). Optimizing corporate bankruptcy prediction through stacking ensemble of KNeighbors, decision tree, and random forest algorithms: Insights from Vietnamese enterprises. Data Science in Finance and Economics, 6(3), 508-532. doi:10.3934/DSFE.2026018

Olufemi Ogunleye, J. (2023). Predictive Data Analysis Using Linear Regression and Random Forest. In Data Integrity and Data Governance. IntechOpen. doi:10.5772/intechopen.107818

Pusporani, E., Sifa, G. A., Faizun, N., Widyangga, P. A., Sari, A. N., Mardianto, M. F., & Sediono. (2025). Comparison of Time Series Regression, Support Vector Regression, Hybrid, and Ensemble Method to Forecast PM2.5. Data and Metadata, 4(862). doi:10.56294/dm2025862

Qin, A., Fan, Z., & Zhang, L. (2024). Information Entropy-Based Hybrid Models Improve the Accuracy of Reference Evapotranspiration Forecast. Advances in Meteorology. doi:10.1155/2024/9922690

Sabljic, S., Serdar Rakovic, T., & Vasic, V. (2023). The New Challenges of the Western Balkan Banking Industry During Ongoing Global Crisis. Journal of Balkan and Near Eastern Studies, 25(4), 671–682. doi:10.1080/19448953.2023.2167166

Sen, S. (2024). COMPARISON OF BOOSTING AND RANDOM FOREST MODELS IN FORECASTING BANK FAILURES. Economy and Finance , 11(3), 258-281. doi:10.33908/EF.2024.3.2

Seyedrahimi-Niaraq, M., Mahdiyanfar, H., & Olyaee, M. H. (2025). Comparison of the performance of gradient boost, linear regression, decision tree, and voting algorithms to separate geochemical anomalies areas in the fractal environment. Artificial Intelligence in Geosciences, 6(2). doi:10.1016/j.aiig.2025.100156

Shahriar, A., Mehzabin, S., Ahmed, Z., Dongul, E. S., & Azad, M. A. (2023). Bank stability, performance and efficiency: an experience from West Asian countries. IIM Ranchi Journal of Management Studies, 2(1), 31-47. doi:10.1108/IRJMS-02-2022-0017

Şimşek, A. İ., Gür, Y. E., & Ünal, E. (2025). Innovative MCDM-ML algorithms-based decision-support system for electric vehicle selection. Environment, Development and Sustainability. doi:10.1007/s10668-025-06476-x

Smola, A. J., & Scholkopf, B. (2004). A tutorial on support vector regression. Statistics and Computing, 14, 199–222. doi:10.1023/B:STCO.0000035301.49549.88

Sonehekpon, E. S. (2025). Effect of artificial intelligence on banking stability: Evidence from developed countries. Finance Research Open, 1(4). doi:10.1016/j.finr.2025.100074

Süzen, E., & Kayakuş, M. (2026). Predicting Life Expectancy in Upper-Middle-Income Countries Using Machine Learning: A Comparative Analysis. Sustainable Development, 1-17. doi:10.1002/sd.71468

Thabet, H. H., Darwish, S. M., & Ali, G. M. (2024). Measuring the efficiency of banks using high-performance ensemble technique. Neural Computing and Applications, 36, 16797–16815. doi:10.1007/s00521-024-09929-y

Theofilis, S., Kampouris, I., & Samitas, A. (2024). Predicting Global Banking Stability: A Machine Learning Approach Using the Camels Framework. SSRN(preprint). doi:10.2139/ssrn.4747568

Tu, P. T., Oanh, D. L., & Trang, D. D. (2025). Machine Learning and Parameter Optimization for Banking Stability Prediction and Determinants Identification in ASEAN. Emerging Science JournaL, 9(3), 1189-1208. doi:10.28991/ESJ-2025-09-03-04

World Bank. (2025). Western Balkans Regular Economic Report No. 28 Fall 2025. World Bank.

World Bank DataBank https://databank.worldbank.org/databases (accessed on 1 April 2026)

Zhou, Z.-H. (2012). Ensemble Methods: Foundations and Algorithms. New York: Chapman and Hall/CRC Press. doi:10.1201/b12207

Acknowledgment

This study was funded by the National Agency for Scientific Research and Innovation (AKKSHI).

Conflict of Interest

The authors have declared that there is no conflict of interest.

Network Intrusion Detection Using Optimized Machine Learning Methods

Emirhan Aydın[1*], Mahir Kaya[2]

[1,2] Tokat Gaziosmanpaşa University, Engineering and Architecture Faculty, Computer Engineering Department, 60100, Tokat, Türkiye

*Corresponding author e-mail: emirhan.aydin@gop.edu.tr

Abstract

This study presents a comparative evaluation of five machine-learning classifiers—Decision Tree, Random Forest, Support Vector Machine, Logistic Regression, and XGBoost for ten-class network intrusion detection using the UNSW-NB15 dataset. A unified experimental framework was established to compare model performance under a consistent experimental protocol. Random Forest-based feature selection, class balancing using the Synthetic Minority Over-sampling Technique for Nominal and Continuous features, and hyperparameter optimization were incorporated into the pipeline. The official training partition was used for model development, while the separate official test partition containing 82,332 instances was reserved for final evaluation. Balanced Accuracy and Macro-F1 were emphasized alongside Accuracy, Macro-Precision, Macro-Recall, and Weighted-F1 to provide a broader assessment of classification performance under class imbalance. On the official test partition, Random Forest achieved the highest Balanced Accuracy (0.6247) and Macro-F1 (0.4853). Decision Tree obtained the highest Accuracy (0.7553), while XGBoost achieved the highest Weighted-F1 (0.7747). The Macro-F1 scores of Random Forest and XGBoost were similar, at 0.4853 and 0.4835, respectively. Class-level analysis revealed substantial differences in recognition performance across attack categories. In particular, Random Forest achieved low F1-scores for Analysis (0.0523) and Backdoor (0.1137), indicating that strong aggregate performance did not ensure effective recognition of all attack classes. The findings demonstrate that model rankings depend on the selected evaluation metric and highlight the importance of class-specific assessment in multiclass intrusion detection. Since the evaluation was conducted on a single dataset, further validation on additional datasets is needed to assess generalizability.

Keywords: *Network Intrusion Detection, UNSW-NB15, Machine Learning, Hyperparameter Optimization, Class Imbalance*

INTRODUCTION

The growing volume and diversity of cyberattacks have increased the need for reliable network intrusion detection systems. Machine learning methods are widely used for this purpose, but their performance depends on the classification algorithm, preprocessing strategy, and hyperparameter configuration. Hyperparameter optimization helps identify suitable model configurations and can improve classification performance. Evaluating different classifiers under a unified experimental framework enables a more consistent comparison of their performance. The UNSW-NB15 dataset (Moustafa &

Slay, 2015) is a widely used benchmark for network intrusion detection. Its multiple attack categories and imbalanced class distribution make it suitable for evaluating multiclass classification methods using metrics that account for differences in class representation. This study compares five machine-learning classifiers Decision Tree, Random Forest, Support Vector Machine, Logistic Regression, and XGBoost for ten-class intrusion detection using UNSW-NB15. The experimental framework incorporates Random Forest-based feature selection, SMOTENC-based class balancing, and hyperparameter optimization. Model performance is evaluated using multiple metrics, with particular emphasis on Balanced Accuracy and Macro-F1.

Related Work

The UNSW-NB15 dataset, introduced by Moustafa and Slay (2015), has been widely used to evaluate machine-learning approaches for network intrusion and anomaly detection. Previous studies have investigated supervised classification, comparative model evaluation, and intrusion detection under different experimental settings (Ahmad et al., 2021; Disha & Waheed, 2021; Thirimanne et al., 2021; Shushlevska et al., 2022; Sonule et al., 2020). Further studies have examined machine-learning and deep-learning approaches for intrusion detection and network security using this dataset (Aleesa et al., 2021; Al-Obaidi et al., 2023; Fathima et al., 2023; Kumar et al., 2024; Putro, 2025; Algarni, 2026; Alkhater, 2026; Karthick et al., 2026; Song, 2026).

Beyond general classifier comparisons, several studies have explored specific modeling strategies and challenges in intrusion detection. Feature selection and its role in classification performance have been investigated by Kasongo and Sun (2020) and Vibhute et al. (2024). Other studies have addressed rare cyberattack detection and intrusion detection performance (Bagui et al., 2019; Moualla et al., 2021; More et al., 2024). Ensemble-based approaches, including stacking, have also been investigated for network intrusion detection using UNSW-NB15 (Kabir et al., 2022; Santhosh Kumar et al., 2024). Collectively, these studies illustrate the range of methods explored for improving intrusion detection performance.

Differences in experimental configurations and evaluation procedures can make direct comparisons between published results difficult. To provide a consistent comparison, this study evaluates five supervised machine-learning classifiers—Decision Tree, Random Forest, Support Vector Machine, Logistic Regression, and XGBoost—using a common experimental framework. The framework incorporates Random Forest-based feature selection, class balancing, and hyperparameter optimization. Balanced Accuracy and Macro-F1 are used as primary evaluation metrics, complemented by class-level analysis to examine recognition performance across individual attack categories.

MATERIAL AND METHODS

Dataset

The UNSW-NB15 dataset was used for multiclass network intrusion detection. It contains 45 columns, including 42 traffic features, the multiclass target variable (attack_cat), a binary label (label), and a record identifier (id). The dataset includes ten classes, consisting of normal traffic and nine attack categories: Generic, Exploits, Fuzzers, DoS, Reconnaissance, Analysis, Backdoor, Shellcode, and Worms. The original training and testing partitions contain 175,341 and 82,332 instances, respectively. The training partition was used for model development, while the official testing partition was reserved for subsequent model evaluation and comparative analysis.

Data Preprocessing and Feature Selection

The official training partition of the UNSW-NB15 dataset was used for model development, while the official test partition was reserved exclusively for final performance evaluation. The official training partition was divided into stratified training and validation subsets using an 80:20 ratio. Hyperparameter optimization was performed using the training subset, and the validation subset was used to evaluate candidate configurations and select the best-performing hyperparameters.
To prevent data leakage, data-dependent preprocessing operations and feature selection were performed using only the training subset. The categorical features proto, service, and state were converted into numerical representations using ordinal encoding. The encoder was fitted on the training subset and then applied to the validation subset without refitting.

Feature selection was performed using Random Forest-based feature importance. A Random Forest model was fitted to the training subset, and the 20 highest-ranked features were retained for subsequent model training and evaluation. The selected features were determined exclusively from the training subset and applied unchanged to the validation subset. Neither the validation subset nor the official test partition was used to determine feature importance or select features. These procedures ensured that preprocessing, feature selection, and hyperparameter optimization did not use information from the official test partition.

Class Balancing

Class imbalance was addressed using the Synthetic Minority Over-sampling Technique for Nominal and Continuous features (SMOTENC). The target class size was set to 11,377 samples, corresponding to the median class size in the original training partition. Classes below this threshold were oversampled, while larger classes were left unchanged. SMOTENC was applied exclusively to the relevant training subsets, preserving the original class distributions in the validation and official test sets. A neighborhood size of five was used for synthetic sample generation.

Machine Learning Models

Five supervised machine-learning algorithms were selected for comparative evaluation: Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), and XGBoost. These algorithms represent different learning approaches, including tree-based classification, bagging, kernel-based classification, linear classification, and gradient boosting. To ensure appropriate model training, feature scaling was applied to SVM and LR, whereas DT, RF, and XGBoost were trained without scaling. The RBF kernel was used for SVM, while LR employed the lbfgs solver with a maximum of 2,000 iterations.

Hyperparameter Optimization

For each classifier, hyperparameter optimization was performed to identify suitable model configurations for multiclass network intrusion detection. The process started with the default configurations of the respective algorithms and explored predefined search spaces for selected hyperparameters. These search spaces covered parameters related to model complexity, regularization, ensemble size, and learning behavior, depending on the algorithm. Candidate configurations were trained on the training subset and evaluated on the validation subset to identify the best-performing configuration for each classifier.

The optimized hyperparameters included max_depth, min_samples_split, min_samples_leaf, and criterion for DT; n_estimators, max_depth, min_samples_split, min_samples_leaf, and max_features for RF; n_estimators, max_depth, learning_rate, subsample, colsample_bytree, min_child_weight, and gamma for XGBoost; C and gamma for SVM; and C for LR. Following optimization, the selected configurations were used to train the final models on the complete official training partition. The final models were then evaluated on the official test partition.

RESULTS

Table 1 presents the optimized hyperparameter configurations selected for each classifier. The configurations were determined independently for each algorithm through hyperparameter optimization, using the validation subset to identify the best-performing settings. The selected configurations were subsequently used to train the final models on the complete official training partition. These configurations provide the basis for the comparative performance evaluation presented in the following section.

Table 1. Hyperparameter configurations of the evaluated models

Model	Hyperparameter	Value
Decision Tree	max_depth	12
	min_samples_split	8
	min_samples_leaf	7
	criterion	gini
Random Forest	n_estimators	303
	max_depth	19
	min_samples_split	4
	min_samples_leaf	1
	max_features	0.5
XGBoost	n_estimators	146
	max_depth	10
	learning_rate	0.06
	subsample	0.99
	colsample_bytree	0.77
	min_child_weight	4
	gamma	0.72
SVM	C	59.67
	gamma	0.23
Logistic Regression	C	0.01

The final models were evaluated on the official UNSW-NB15 test partition containing 82,332 instances. Table 2 and Figure 1 summarizes their performance using Accuracy, Balanced Accuracy, Macro-Precision, Macro-Recall, Macro-F1, and Weighted-F1. Random Forest achieved the highest Balanced Accuracy (0.6247) and Macro-F1 (0.4853), while Decision Tree obtained the highest Accuracy (0.7553). XGBoost achieved the highest Weighted-F1 (0.7747) and a Macro-F1 of 0.4835, close to that of Random Forest. These results demonstrate that model rankings vary depending on the evaluation metric. No statistical significance tests were performed on the single official test evaluation; therefore, the observed differences should be interpreted descriptively.

Table 2. Model performance on the official testing set

Model	Accuracy	Balanced Accuracy	Macro-Precision	Macro-Recall	Macro-F1	Weighted-F1
DT	0.7553	0.5686	0.4586	0.5686	0.4317	0.7653
RF	0.7399	0.6247	0.5179	0.6247	0.4853	0.7720
SVM	0.7083	0.5123	0.4154	0.5123	0.3983	0.7291
LR	0.6121	0.4346	0.3103	0.4346	0.2828	0.6348
XGBoost	0.7473	0.6040	0.4784	0.6040	0.4835	0.7747

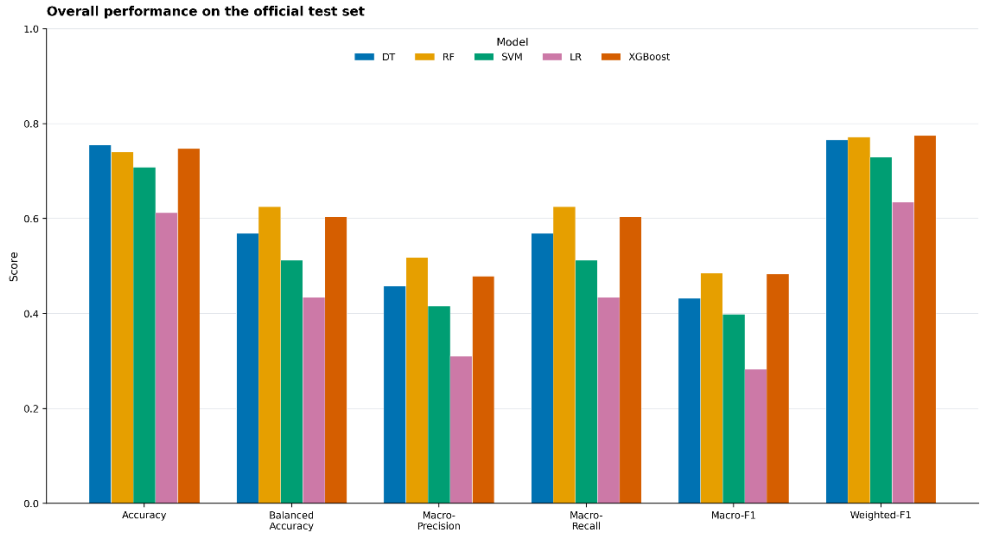


Figure 1. Overall classification performance of the final models on the official test partition.

The confusion matrix of Random Forest is presented in Figure 2 to illustrate class-specific misclassification patterns. Although the model correctly classified a large proportion of Generic instances, confusion remained among several attack categories, including DoS, Exploits, Backdoor, and Analysis. These findings highlight the challenges of recognizing minority and overlapping attack categories in multiclass intrusion detection.

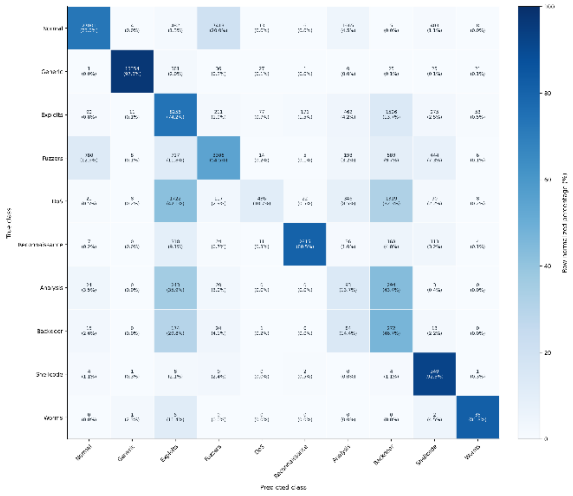


Figure 2. Row-normalized confusion matrix of Random Forest on the official test partition.

DISCUSSION AND CONCLUSION

This study presented a comparative evaluation of five machine-learning algorithms for network intrusion detection using the UNSW-NB15 dataset. The experimental pipeline incorporated Random Forest-based feature selection, class balancing, and hyperparameter optimization to ensure a consistent evaluation framework.

The results on the official test partition showed that Random Forest achieved the highest Balanced Accuracy (0.6247) and Macro-F1 (0.4853), while Decision Tree obtained the highest Accuracy (0.7553) and XGBoost achieved the highest Weighted-F1 (0.7747). These findings demonstrate that model rankings vary depending on the evaluation metric, emphasizing the need to consider multiple metrics when assessing imbalanced multiclass classification tasks.Class-level analysis further revealed uneven recognition performance across attack categories. In particular, Random Forest achieved low F1-scores for Analysis (0.0523) and Backdoor (0.1137), indicating that competitive aggregate results do not necessarily translate into effective detection of every attack category.

This study is limited by its reliance on a single dataset, which restricts the assessment of generalizability across different network environments. Future work will therefore evaluate the proposed experimental framework on additional datasets and investigate alternative feature-selection and class-balancing methods, as well as approaches specifically designed to improve the detection of attack categories with low recognition performance. Overall, the findings highlight the importance of considering both aggregate metrics and class-specific performance when selecting machine-learning models for network intrusion detection. Further evaluation across diverse datasets is needed to determine whether the observed performance patterns generalize to other network environments.

References

Ahmad M., Riaz Q., Zeeshan M., Tahir H., Haider S.A., Khan M.S., 2021. Intrusion detection in internet of things using supervised machine learning based on application and transport layer features using

UNSW-NB15 data-set. EURASIP Journal on Wireless Communications and Networking, 2021 (1): 10. https://doi.org/10.1186/s13638-021-01893-8

Aleesa A.M., Younis M., Mohammed A.A., Sahar N.M., 2021. Deep-intrusion detection system with enhanced UNSW-NB15 dataset based on deep learning techniques. Journal of Engineering Science and Technology, 16 (1): 711-727.

Algarni S., 2026. Supervised intrusion detection on UNSW-NB15: A comparative study of classical machine learning models. 2026 International Symposium of Systems, Advanced Technologies and Knowledge (ISSATK), pp.1-5. https://doi.org/10.1109/ISSATK69667.2026.11566846

Alkhater N., 2026. A rigorous comparative study of supervised machine learning techniques for network anomaly detection: Empirical insights from the UNSW-NB15 dataset. Computers, 15 (5): 285. https://doi.org/10.3390/computers15050285

Al-Obaidi A., Ibrahim A.A., Khaleel A.M., 2023. The effectiveness of deploying machine learning techniques in information security to detect nine attacks: UNSW-NB15 dataset as a case study. Mathematical Modelling of Engineering Problems, 10 (5): 1557–1565. https://doi.org/10.18280/mmep.100507

Bagui S., Kalaimannan E., Bagui S., Nandi D., Pinto A., 2019. Using machine learning techniques to identify rare cyber-attacks on the UNSW-NB15 dataset. Security and Privacy, 2 (6): e91. https://doi.org/10.1002/spy2.91

Disha R.A., Waheed S., 2021. A comparative study of machine learning models for network intrusion detection system using UNSW-NB 15 dataset. 2021 International Conference on Electronics, Communications and Information Technology (ICECIT), pp.1-5. https://doi.org/10.1109/ICECIT54077.2021.9641471

Fathima A., Khan A., Uddin M.F., Waris M.M., Ahmad S., Sanin C., Szczerbicki E., 2023. Performance evaluation and comparative analysis of machine learning models on the UNSW-NB15 dataset: A contemporary approach to cyber threat detection. Cybernetics and Systems, 56 (8): 1160-1176. https://doi.org/10.1080/01969722.2023.2296246

Kabir M.H., Rajib M.S., Rahman A.S.M.T., Rahman M.M., Dey S.K., 2022. Network intrusion detection using UNSW-NB15 dataset: Stacking machine learning based approach. 2022 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE), pp.1-6. https://doi.org/10.1109/ICAEEE54957.2022.9836404

Kasongo S.M., Sun Y., 2020. Performance analysis of intrusion detection systems using a feature selection method on the UNSW-NB15 dataset. Journal of Big Data, 7 (1): 105. https://doi.org/10.1186/s40537-020-00379-6

Karthick S., Devi P.J.A., Priyadharsini R., 2026. A comprehensive study of anomaly detection models for network security using the UNSW-NB15 dataset. 2026 3rd International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE), pp. 1–6. https://doi.org/10.1109/RMKMATE69073.2026.11518835

Kumar A., Guleria K., Chauhan R., Upadhyay D., 2024. Advancing intrusion detection with machine learning: Insights from the UNSW-NB15 dataset. 2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS), pp.1-5. https://doi.org/10.1109/ICITEICS61368.2024.10625148

More S., Idrissi M., Mahmoud H., Asyhari A.T., 2024. Enhanced intrusion detection systems performance with UNSW-NB15 data analysis. Algorithms, 17 (2): 64. https://doi.org/10.3390/a17020064

Moualla S., Khorzom K., Jafar A., 2021. Improving the performance of machine learning-based network intrusion detection systems on the UNSW-NB15 dataset. Computational Intelligence and Neuroscience, 2021: 5557577. https://doi.org/10.1155/2021/5557577

Moustafa, N., and Slay, J. (2015). UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). 2015 Military Communications and Information Systems Conference (MilCIS), 1–6. https://doi.org/10.1109/MilCIS.2015.7348942

Putro I.H., 2025. Evaluating the performance of machine learning classifiers for network intrusion detection: A comparative study using the UNSW-NB15 dataset. Teknika, 14 (2): 330-338. https://doi.org/10.34148/teknika.v14i2.1276

Santhosh Kumar V., Saiharish S., Dinesh Kumar A., 2024. Network intrusion detection through stacked machine learning models on UNSW-NB15 dataset. 2024 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), pp. 1–7. https://doi.org/10.1109/ICSES63760.2024.10910677

Shushlevska M., Efnusheva D., Jakimovski G., Todorov Z., 2022. Anomaly detection with various machine learning classification techniques over UNSW-NB15 dataset. Share_it. https://doi.org/10.25673/76928

Song H., 2026. A comparative study of machine learning methods for intelligent network intrusion detection based on UNSW-NB15 dataset. 2026 10th International Conference on Electronic Information Technology and Computer Engineering (EITCE), pp.1412-1417. https://doi.org/10.1109/EITCE70137.2026.11634525

Sonule A.R., Kalla M., Jain A., Chouhan D.S., 2020. UNSW-NB15 dataset and machine learning based intrusion detection systems. International Journal of Engineering and Advanced Technology, 9 (3): 2638-2648. https://doi.org/10.35940/ijeat.C5809.029320

Thirimanne S., Jayawardana L., Liyanaarachchi P., Yasakethu L., 2021. Comparative algorithm analysis for machine learning based intrusion detection system. 2021 10th International Conference on Information and Automation for Sustainability (ICIAfS), pp.191-196. https://doi.org/10.1109/ICIAfS52090.2021.9605814

Vibhute A.D., Khan M., Patil C.H., Gaikwad S.V., Mane A.V., Patel K.K., 2024. Network anomaly detection and performance evaluation of convolutional neural networks on UNSW-NB15 dataset. Procedia Computer Science, 235: 2227-2236. https://doi.org/10.1016/j.procs.2024.04.211

Conflict of Interest

The authors have declared that there is no conflict of interest.

Author Contributions

All authors contributed to the study conception and design. Data analysis, experiments, and manuscript preparation were carried out collaboratively by the authors. All authors read and approved the final manuscript.

Beyond Model Drift: A Meta-Governance Framework for Cascading Decision Intelligence Systems

Selçuk Bayracı[1]*

[1] R&D Center, Borusan Otomotiv, Istanbul, Turkiye

*Corresponding author: selcuk.bayraci@borusanotomotiv.com

Abstract

Cascading decision intelligence (DI) stacks, in which multiple machine learning models feed one another sequentially to produce operational recommendations, are increasingly deployed in asset-intensive industries. Existing MLOps frameworks address drift in individual models but offer little governance for the compound failures that emerge when layers interact. This paper proposes the Dialectical Enterprise Operating System (DEOS), a meta-governance framework positioned above the Borusan Next Control Tower, a seven-layer cascading DI stack for used-vehicle pricing, inventory optimization, and portfolio risk management. DEOS draws on double-loop learning, self-referential "strange loop" architectures, Normal Accidents Theory, and research on noise and algorithm aversion, and formalizes decision-right reallocation as a recursive, meta-evaluative process. A quantitative governance model based on expected utility theory and Conditional Value-at-Risk is introduced to detect algorithm-induced goal incongruence. The paper makes four contributions: (i) an A/B/C decision taxonomy with numeric thresholds, escalation time limits, and a forced exploration protocol that turns algorithm aversion into a calibration asset; (ii) a failure mode register that treats cascading error accumulation as a normal accident inherent to tightly coupled inference architectures; (iii) an argument that hard mechanical safety valves in the execution and risk layers (stop-loss, spread floor, override) are structural necessities rather than conservative over-engineering; and (iv) an alignment assessment against Gartner's Decision Intelligence maturity model and emerging agentic AI governance practice. The analysis is theoretical and illustrated through a compound market-shock stress test; the platform is under active development, and empirical validation is left for future work.

Keywords: *Decision Intelligence, Decision Support Systems, Machine Learning Governance, Intelligent Systems, System Modeling*

INTRODUCTION

The integration of machine learning into operational decision-making has shifted the governance challenge from model accuracy to system architecture. A single well-calibrated model is a tractable problem: drift is detectable, retraining is procedural, and failure is attributable. A cascading stack, in which each model's output is another model's input, is a fundamentally different governance object. Individual accuracy gates do not detect compound error accumulation, and model-level retraining does

not resolve goal incongruence that originates in the interaction between layers. Established work on hidden technical debt (Sculley et al., 2015) and on the practical challenges of deploying machine learning (Paleyes et al., 2022) documents these difficulties at the level of single models and pipelines, but does not provide governance mechanisms for the compound failure modes of multi-layer decision systems.

Decision intelligence (DI) is the discipline of engineering how decisions are made, including how models, rules, and human judgment are combined (Pratt, 2019). The Borusan Next Control Tower is a practical instance of a cascading DI system. Designed for the Turkish used-vehicle market, it chains seven functional layers, from market price estimation through sales dynamics, decision optimization, pricing discipline, execution, and portfolio risk, into one decision stream covering every vehicle from acquisition bid to channel exit. Conventional MLOps practice monitors each layer in isolation. This paper addresses the system as a whole.

The research question is: *how can an organization operating a multi-layer cascading DI stack design a self-correcting governance framework that maintains strategic alignment under market shocks, model drift, compound error accumulation, and the behavioral resistance of human decision-makers to algorithmic recommendations?*

We answer with the Dialectical Enterprise Operating System (DEOS), a meta-governance layer above the stack. Its theoretical foundations combine organizational learning (Hegelian dialectics read through Argyris and Schön's double-loop learning), structural fragility in tightly coupled systems (Perrow's Normal Accidents Theory), and human-model interaction (research on algorithm aversion and noise). The framework is operationalized through a decision taxonomy, a utility-based override trigger with a tail-risk term, and a failure mode register, and is positioned for readers in decision support systems, machine learning operations, and intelligent systems.

This is a theoretical case study. The Control Tower is at an active development stage: Phase 0 (price model in shadow mode with a shadow pricing log) and parts of Phase 1 (survival pipeline) are live or in shadow mode, while the full stack through Portfolio VaR (Phase 4) is in design. All DEOS analyses are therefore theoretical projections, not empirical reports. The remainder of the paper is organized as follows: Section 2 presents the theoretical foundations, Section 3 the platform and its decision taxonomy, Section 4 the quantitative governance model, Section 5 a theoretical stress test, Section 6 the failure mode register, Section 7 the alignment with Gartner's DI guidance, and Section 8 the conclusions.

Theoretical Foundations

Dialectical Decision Dynamics and Double-Loop Learning

Adapting Hegel's (1807) dialectical method to organizational decision processes, DEOS models DI stack failures not as operational breakdowns but as generative events that trigger organizational learning. The three-phase structure maps as follows:

- **Thesis:** the stack's operational recommendation, for example a repricing action, an acquisition bid, or a segment allocation.

- **Antithesis:** the structural failure or goal incongruence of that recommendation, for example a price-model collapse under a currency shock, over-discounting caused by a biased elasticity estimate, or a portfolio VaR breach during a liquidity event.
- **Synthesis:** the reintegration of human judgment and organizational strategy through DEOS Override Mode, for example brand-constrained pricing, forced exploration, or model suspension by a review committee.

This structure maps onto Argyris and Schön's (1978) distinction between single-loop learning (updating parameters within existing objectives) and double-loop learning (interrogating the governing variables themselves). In the Control Tower, the execution engine recalibrating its weights from shadow-log data is single-loop; DEOS suspending that engine when its objective conflicts with brand positioning or portfolio concentration limits is double-loop. Consistent with Zollo and Winter (2002), model failures are treated as generative signals for organizational adaptation rather than anomalies to be minimized.

Self-Referential Meta-Evaluation

A cascading DI stack cannot reason about its own inferential boundaries. Inspired by Gödel (1931) and Hofstadter's (1979) elaboration of self-referential "strange loops", DEOS adds a meta-evaluation function $E(D(t))$ to the governance loop:

$$D(t+1) = S(D(t), C(t), E(D(t)))$$

where $C(t)$ is the market context vector (exchange rate, credit impulse, liquidity regime, listing density), $S(\cdot)$ is the synthesis operator, and $E(D(t))$ assesses the stack's current epistemic reliability. In practice, $E(D(t))$ is operationalized through the MLOps monitoring layer: population stability index (PSI) alerts on the price model, C-index degradation of the survival model, the false-positive rate of the execution engine, and, critically for cascading systems, the end-to-end cohort validation test described in Section 6.

Epistemological caveat. Applying Gödel's incompleteness theorems to organizations is a heuristic analogy, not a mathematical equivalence: his results concern formal axiomatic systems, which the Control Tower is not. The analogy conveys the intuition that no layer can fully validate its own outputs using only its own metrics. Hegelian and Gödelian constructs function in DEOS as organizing heuristics for governance design, not as formal claims.

Normal Accidents Theory

Perrow's (1984) Normal Accidents Theory distinguishes two structural properties. *Interactive complexity* describes components that interact in non-linear ways designers did not anticipate. *Tight coupling* describes systems in which one component's failure propagates rapidly to the next with little slack. Systems with both properties produce *normal accidents*: cascading failures that are inevitable consequences of the architecture rather than exceptional events.

The Control Tower has both properties. It is interactively complex because errors in the price model shift the covariates of the survival model, which shift execution thresholds, which alter realized time-to-sell (TTS), which recalibrate the portfolio covariance, which feeds back to execution. It is tightly coupled because each layer's output becomes the next layer's input with minimal buffering, and no

human checkpoint exists between the price estimate and the hazard function's price covariate. Under Perrow's framework, cascading error accumulation is therefore a structural property of the architecture, not a risk to be managed away.

Three governance implications follow. First, the hard mechanical rules in the pricing-discipline and execution layers (a VaR-derived acquisition ceiling, a dynamic spread floor, and a stop-loss rule) are engineering necessities: the only way to add buffering to a tightly coupled system is to insert unconditional breaks that cannot propagate error signals. A stop-loss that fires at age ≥ 1.5 × TTS_target does not ask for model consensus. Second, the DEOS override is the system-level redundancy a tightly coupled architecture requires to remain operable. Third, periodic end-to-end cohort validation is the equivalent of a decoupled rehearsal, the only method for detecting cascading failures before they appear in pricing outcomes. Perrow's framework also reframes human-in-the-loop design: in loosely coupled systems removing humans increases speed, but in tightly coupled, interactively complex systems it removes the only real-time mechanism for interrupting a cascade.

Algorithm Aversion and Noise

The fourth pillar is psychological. Dietvorst et al. (2015) show that people are markedly less willing to use an algorithm after observing it err, even when it outperforms human judgment, because they hold algorithms to a standard of perfection they do not apply to humans. For the Control Tower, algorithm aversion is the primary adoption risk during the first six to nine months. Early elasticity estimates are biased (Section 6, FM-01), so the execution engine will over-recommend discounting. The pricing team will correctly recognize these errors and may conclude that the model is unreliable, a judgment that is hard to revise even after estimates improve. The stack risks abandonment before calibration data accumulates.

DEOS responds with three mechanisms. First, *structured override protocols*: requiring a categorized rationale converts an override from a vote of no confidence into a data contribution, consistent with evidence that people who feel their judgment is used rather than replaced show less algorithm aversion (Kahneman et al., 2021). Second, *forced exploration*: 5% of Level B decisions are implemented regardless of team inclination, which prevents the shadow log from converging toward human intuition and is framed as a scientific calibration protocol. Third, *Trust Score gating*: Level A auto-execution is activated only after a Trust Score of at least 3.5 out of 5.0, so that automation is not deployed before the team understands the model.

The complementary lens is noise (Kahneman et al., 2021). Without a DI stack, two analysts facing the same vehicle produce different recommendations through random variability in judgment rather than reasoned disagreement, and this noise is invisible because decisions are not systematically compared. The shadow log makes it measurable: the divergence rate between model and human recommendations is a direct measure of inter-analyst noise. An override rate below 5% signals rubber-stamping, and a rate above 50% signals lost model credibility; the target range of 15–30% is where noise reduction and model calibration can occur simultaneously.

The Control Tower Platform and Its Decision Taxonomy

Seven-Layer Cascading Architecture

The Control Tower is designed as a cascading inference engine in which every model feeds the next: the price model produces the prior for the demand-signal updater; the survival model produces the expected TTS that serves as the clock for the channel optimizer; market indices set the dynamic threshold of the execution engine; and portfolio VaR produces the risk weights that constrain the stock-mix optimization. This maximizes information integration across decisions but, under Perrow's framework, also creates the tightly coupled, interactively complex structure discussed above. Table 1 summarizes the seven layers.

Table 1. Control Tower: seven-layer cascading DI architecture

Layer	Name	Core methods and inputs	Key outputs
L0	Data and pipeline	Listing feed, SAP transactions, CRM offer logs, auction prices; re-listing detection (censoring correction); versioned feature store	Survival-format data; macro features; validated inputs
L1	Market price (EBM)	Explainable Boosting Machine with additive shape functions; Pp ~ $N(\mu_0, \sigma_0^2)$, σ_0 = (upper − lower)/(2 × 1.645)	Pp median and 90% interval; σ_Pp; "why this price?" explanations
L2	Market dynamics	Extended Cox TTS survival model with time-varying ΔPrice(t); discrete hazard auction model; Market Liquidity Index (MLI); Market Demand Signal (MDS); price elasticity β_1	S(t|X); E[TTS]; β_1; MLI; MDS; stop-loss day
L3	Decision optimization	Channel Optimizer Ω(Pa, Ps, k) = V(r_a, k)·(365/TTS)·Π; Stock Mix MILP under capital, segment, and age constraints	Pa∗, Ps∗, k∗; target units per segment
L4	Pricing discipline	Pa_max = Pp(1 − zσ_TTS) − C_holding − C_ek − Π_min; spread floor; stop-loss at age ≥ 1.5 × TTS_target; authority matrix	Hard acquisition ceiling; stop-loss trigger; authority routing
L5	RPCE execution	R_score (repricing urgency) and C_score (channel-transition urgency); early-warning flag from engagement data; four action states	Weekly scores per listing; repricing or transition recommendation; EWS flag
L6	Portfolio risk	Portfolio Ω-VaR; Monte Carlo (10,000 paths); EWMA covariance (λ = 0.94) with Ledoit–Wolf shrinkage; marginal CVaR	VaR_{95}, $CVaR_{95}$; marginal CVaR per vehicle; segment bounds φ; weekly risk report

L0–L2: market intelligence foundation. The pipeline ingests the full Turkish used-car listing market, transactional data, and offer logs. A re-listing detector, already live, reclassifies a closed listing as censored if the same vehicle reappears within a defined window, preventing a systematic underestimate of TTS from propagating through downstream models. The price model is an Explainable Boosting Machine (Lou et al., 2012) chosen over gradient boosting precisely because its additive shape functions answer "why this price?", an explainability property that supports adoption (Dietvorst et al., 2015). The survival layer extends the Cox (1972) proportional hazards model with price revisions as time-varying covariates, $h(t|X(t)) = h_0(t)\cdot exp(\beta_1\cdot\Delta Price(t) + \beta_2\cdot age + \beta_3\cdot km + \beta_4\cdot season)$, complemented by a discrete hazard model for auction rounds, where continuous-time Cox is misspecified because sales occur only on specific event dates. Liquidity and demand indices summarize turnover, flow balance, and price pressure at segment, geography, and week granularity. Price elasticity is estimated through a three-level

hierarchy: fixed-effects OLS (active), Extended Cox with time-varying price (pending price history), and an instrumental-variable approach (planned).

L3–L6: decision, execution, and risk. The Channel Optimizer maximizes the annualized profit rate Ω over acquisition price, list price, and channel, where V is a logistic acquisition probability calibrated on offer logs. The Stock Mix MILP treats 35–40% of inflows (trade-ins, fleet returns) as an exogenous mandatory flow and optimizes the discretionary acquisition budget. The pricing-discipline layer bounds every offer by the VaR-derived ceiling Pa_max (z = 1.645), enforces the spread floor $Ps - Pa \geq TTS \cdot Cg + C_ek + \Pi_min$, and triggers the stop-loss. The execution engine (RPCE) scores repricing and channel-transition urgency weekly, with a dynamic TTS threshold that scales the transition clock to market liquidity; its early-warning term starts at zero and activates only after a correlation of $|r| > 0.30$ with TTS is confirmed from shadow-log data. The portfolio layer draws correlated price shocks from $N(0, \Sigma_EWMA \cdot h)$ and stochastic TTS from each vehicle's survival function (Ledoit and Wolf, 2004), and its marginal CVaR contributions feed back into the execution engine and the segment bounds of the MILP.

The A/B/C Decision Taxonomy

The taxonomy is the operational contract between the analytical stack and the organization and the primary behavioral design tool for managing algorithm aversion. Every decision is classified as Level A (Automated: the model executes without human input, and the decision is logged and auditable), Level B (Augmented: the model recommends and a human approves or overrides with a logged rationale), or Level C (Human-Led: the model supplies context and risk signals, and a human decides independently). Table 2 maps the decision set.

Table 2. Decision taxonomy: automation levels, sources, and override requirements

Decision	Level	Model source / trigger	Override and escalation
Retail list price change $\Delta Ps < 3\%$	A: Auto	RPCE; $R_score \geq \varepsilon_R$	No; auto-logged, post-hoc review
Retail list price change ΔPs 3–8%	B: Aug (L1)	RPCE; $R_score \geq \varepsilon_R$	L1 approval; 4-hour escalation limit
Retail list price change $\Delta Ps > 8\%$	B: Aug (L2)	RPCE; $R_score \geq \varepsilon_R$	Pricing Manager; 8-hour limit
Stop-loss exit activation	A: Auto	Survival + VaR; age ≥ 1.5 × TTS_target	No; hard rule, L3 single override only
Acquisition bid $Pa* \leq Pa_max$	B: Aug (L1)	Channel Optimizer; $Pa \leq Pa_max$	Logged; 4-hour limit
Acquisition bid $Pa > Pa_max$	C: Human (L3)	VaR hard limit exceeded	Manager plus written rationale
Channel routing, retail to auction	B: Aug (L2)	RPCE; $C_score \geq \varepsilon_C$	L2; 24-hour limit
Combined reprice and transition	C: Human (L3)	RPCE dual score; both $\geq \varepsilon$	Manager approval required
Stock mix segment targeting	B: Aug (L2)	Stock Mix MILP; weekly output	Weekly committee (MRC) review

Decision	Level	Model source / trigger	Override and escalation
Segment φ bounds update	C: Human (L4)	Portfolio VaR; weekly review	GM / CFO level
Portfolio VaR breach response	C: Human (L4)	Portfolio VaR; breach event	Escalation protocol

Design principle: overrides are noise measurements, not failures. A target override rate of 15–30% during stabilization is not a compromise metric but the range within which noise reduction (the model provides an anchor) and calibration (overrides provide signal) occur together. The override protocol, with mandatory categorized rationale recorded before the outcome is known, is designed so that overriding feels like a contribution to the system, which addresses the behavioral mechanism behind algorithm aversion.

Quantitative Governance Framework

DEOS formalizes decision viability through an expected utility model whose parameters are grounded in the Control Tower's outputs. The total organizational utility of a strategic decision *D* is:

$$U(D) = E[\, V(D \mid \theta) \,] - \lambda \cdot CVaR_{\alpha}(\, L(D) \,) - C(F)$$

- **E[V(D|θ)], expected net value:** the revenue-minus-cost differential conditional on the current market distribution θ, approximated under normal operation by the expected net profit from the Channel Optimizer.
- **CVaR_α(L(D)), tail risk (Rockafellar and Uryasev, 2000):** the conditional expected loss in the worst $(1 - \alpha)$ fraction of the Monte Carlo paths of the portfolio layer, with $\alpha \in \{0.95, 0.99\}$. Using the stochastic survival distribution as the holding period captures the compound risk of price decline and holding-cost accumulation.
- **λ, risk-aversion coefficient:** calibrated by the Model Review Committee (MRC) from the prevailing VaR regime and declared risk appetite. As the portfolio layer signals rising CVaR, the MRC increases λ, tightening the override sensitivity of U(D). In Perrow's terms, λ acts as the governance system's coupling parameter: higher λ introduces more buffering between algorithmic recommendations and automated execution.
- **C(F), operational friction cost:** the cost of implementing the decision, including override effort, retraining infrastructure, brand-equity erosion from discount-driven volume recovery, and the trust degradation that results from forcing a team to follow systematically biased recommendations. The shadow period is therefore an investment in reducing future C(F).

Override trigger. When $U(D) < U*$, where $U* = E[V_{fallback}] - \varepsilon$ is the expected value of the human-directed fallback strategy minus a tolerance margin, DEOS activates Override Mode. The brand-constrained pricing objective then becomes:

$$\max_D E[\, V_{liquidity}(D) \,] \quad \text{subject to} \quad BrandRisk(D) \leq \bar{B}$$

This treats brand equity as a binding constraint rather than a traded objective, preventing the optimizer from exchanging brand value for marginal liquidity gains, which is the goal incongruence that arises when the execution engine optimizes dealer margin while the manufacturer's strategy prioritizes brand

positioning. It also represents deliberate decoupling of the execution layer from the portfolio-level objective, introducing the slack a tightly coupled system otherwise lacks.

Theoretical Stress Test: Compound Market Shock

To illustrate DEOS operating on the stack, we construct a theoretical compound shock: a severe lira depreciation combined with a sudden contraction of consumer credit, a configuration with a close historical analogue in the 2018 Turkish currency crisis (TCMB, 2022; Akçelik et al., 2022). In Perrow's classification this is precisely the kind of external perturbation that violates the distributional assumptions under which the system was calibrated and propagates through the cascade before any single-layer monitor can detect it.

Phase 1: Cascading Layer Stress

L1 price model. The rolling macro correction cannot adjust fast enough. Pp_median falls sharply for currency-sensitive segments, σ_Pp widens, and the PSI monitor alerts. ΔPrice(t) turns increasingly negative for vehicles priced before the shock, corrupting the time-varying covariate in the hazard function below.

L2 market dynamics. The elasticity β_1, estimated on pre-shock data, belongs to a different demand regime, so E[TTS] is systematically underestimated for pre-shock vehicles. The liquidity index falls as turnover collapses, and the dynamic TTS threshold shortens by up to 25% ($\gamma = 0.25$, worst case), accelerating the transition clock for every vehicle simultaneously. This is the tight-coupling propagation path, price-model error → Cox covariate shift → E[TTS] bias → threshold compression, traversed within one weekly scoring cycle with no human checkpoint at any layer boundary.

L3–L4 optimization and discipline. The optimizer's Ω becomes unreliable because both E[TTS] and Pp are biased. One self-stabilizing feature remains: Pa_max tightens automatically as σ_Pp widens, mechanically reducing acquisition limits. This is the Perrow-compliant behavior of a hard rule: it fires unconditionally and buffers acquisition decisions from the degraded optimization layer above.

L5 execution. R_score spikes across the portfolio simultaneously, and C_score rises as auction and retail profit rates converge in the liquidity crisis. This is the dialectical antithesis: the engine recommends aggressive, simultaneous, portfolio-wide discounting, which is incongruent with premium brand positioning and, through the strange-loop dynamic below, would deepen the price decline.

L6 portfolio VaR. Segment correlations converge toward 1.0, eliminating the diversification benefit, and the VaR_{95} breach rate climbs above the Kupiec (1995) threshold. Segment bounds in the MILP tighten as marginal CVaR rises for every segment, another self-stabilizing mechanical response.

Phase 2: DEOS Intervention and Dialectical Synthesis

The meta-evaluation function detects simultaneous signals: a PSI alert, TTS error above the 8-day threshold, a rising false-positive rate in the execution engine, and a VaR breach. In Perrow's terms this is a system-level normal-accident signal, since failure modes have interacted across layer boundaries in a way no individual monitor is designed to detect. U(D) falls below $U*$, the MRC is convened, and Override Mode activates.

- **Adaptive override:** autonomous recommendations are suspended and authority reverts to Level C for all pricing decisions above 3%. The objective becomes brand-constrained liquidity preservation, maintaining BrandRisk(D) ≤ $\bar{B}$ while preserving cash flow through selective exit of the oldest inventory. This is the synthesis: neither the pre-shock thesis (maximize Ω) nor the shock antithesis (mass discounting).
- **Stable-layer fallback:** the L4 hard rules remain operational and directionally correct: Pa_max tightens, the stop-loss fires for aged inventory, and the spread floor prevents margin destruction. Portfolio concentration metrics remain useful as relative rankings even if absolute VaR values are biased, because all segments face the same correlation shock.
- **Early-warning staging:** the demand-signal index flags vehicles with anomalously low engagement in the first 48–72 hours of listing, giving early warning of TTS exposure before the survival model's lagged signal catches up. The early-warning weight is elevated in Override Mode to prioritize flagged vehicles for human review.
- **MRC-governed restoration:** as post-shock data accumulates and passes the PSI gate (PSI < 0.20) and TTS error gate (MAE ≤ 8 days), a retrained candidate is introduced through the shadow validation pipeline, and U(D) is monitored until it exceeds U∗. Decision rights are restored layer by layer in reverse Perrow order, mechanical rules first and composite models afterward: L4 discipline, L5 sub-3% auto-execution, full L5, then L3 segment allocation.

The Organizational Strange Loop

The shock exposes a reflexive dynamic described by Soros (1987): the platform's own decisions alter the market microstructure it models. If the engine's mass-discount recommendations were executed, they would shift listing price distributions, which re-train the price model's rolling median, which lowers Pp, which triggers further discounting. This is Hofstadter's (1979) strange loop instantiated at platform level. The pricing team faces a second-order problem: they recognize that executing the recommendation would be harmful, but also that the model cannot recognize this about itself. DEOS formalizes this as the governance condition under which *E(D(t))* overrides the algorithmic thesis, not because the model is wrong about the current state, but because it cannot model its own effect on the state it is about to create.

Limitations and Failure Mode Register

A governance framework should state its own boundaries. The Control Tower's Failure Mode Register translates technical limitations into operational risk language, grounded where appropriate in Normal Accidents Theory and the algorithm aversion literature. Table 3 presents the abridged register (FM-03, FM-05, FM-09, and FM-10 are omitted for brevity).

Table 3. Control Tower failure mode register (abridged)

ID	Failure mode and symptom	Risk	Mitigation
FM-01	Elasticity β_1 overstates discount effectiveness in the first 6–9 months (OLS only); over-discounting, margin leakage, algorithm aversion	High	Monthly β_1 sign and magnitude audit; L2 review of all discounts > 5%; Trust Score monitoring as early warning of adoption breakdown
FM-02	Shadow log trains on human-agreement bias; weights converge to human intuition, not market truth	High	Forced exploration (5% of Level B decisions); quarterly bias audit
FM-04	Engagement data unavailable (API license gap); early-warning signal goes dark	Medium	Leads-only fallback index pre-built; early-warning weight = 0 until validated
FM-06	Liquidity and demand indices conflate supply and demand shocks; wrong response (discount vs. adjust mix)	Medium	Quarterly MRC review of index interpretation; competitor supply event log
FM-07	Portfolio VaR unvalidatable for 2+ years (stochastic TTS horizon, no benchmark); false precision	Medium	Use VaR as relative segment ranking until the Kupiec test passes (400+ predictions needed)
FM-08	Silent data quality degradation; models produce confidently wrong outputs, attribution delayed by weeks	Critical	Analytics datamart with automated row-count, null-rate, and join-integrity checks; named Data Contract Owner with SLAs
FM-11	Cascading error accumulation across seven layers; execution engine miscalibrated although every model passes its gate	High	Quarterly end-to-end cohort validation; layer-boundary distribution logging; cascade-sensitivity Monte Carlo; top-down error budget

FM-01: Elasticity Bias as an Algorithm Aversion Trigger

The elasticity β_1 propagates into four downstream components: the repricing score, the repricing gradient step, the optimizer's Ω curve, and the dynamic spread floor. During the first two to three operational quarters only segment-level OLS estimates are available, and these systematically overstate $|\beta_1|$: the model believes price cuts are more effective than they are. Over-discounting is the operationally preferable failure direction but creates a specific adoption risk, since analysts observing systematic over-discounting form an impression of incompetence that is hard to revise (Dietvorst et al., 2015). Trust Score monitoring and the shadow period are therefore the primary mitigation against abandonment of the stack before calibration data accumulates.

FM-08: Data Quality as a Normal Accident Precondition

Perrow (1984) identifies production pressure and unrecognized failures as the conditions under which tightly coupled systems are most vulnerable. FM-08 instantiates both: when the data layer silently degrades (a join producing duplicates, a late feature-store refresh, a re-listing window changed without

retraining), models produce confidently wrong outputs while every individual monitoring metric stays green. A well-specified model on corrupted inputs is indistinguishable from a well-calibrated model until pricing outcomes deteriorate weeks later. A dedicated datamart with automated checks at every pipeline stage, owned by a named Data Contract Owner, is the only available mechanism for introducing the observability and loose coupling that FM-08 requires.

FM-11: Cascading Error Accumulation

FM-11 is the purest instantiation of Normal Accidents in this context. Gartner (den Hamer, 2026) quantifies the "telephone game" effect: at 80% accuracy per layer, a three-layer chain yields $0.8^3 \approx 51\%$ end-to-end accuracy. The primary propagation paths in the Control Tower are: (i) price-model error (MAPE $\leq 6\%$) shifts ΔPrice(t) in the hazard function and biases E[TTS]; (ii) E[TTS] error (MAE ≤ 5 days) shifts the survival ratio in R_score, causing premature or delayed repricing; (iii) combined TTS and β_1 error biases Ω in the optimizer; and (iv) repricing decisions change realized TTS, recalibrating the portfolio covariance, which closes a feedback loop between execution and risk measurement.

Individual acceptance gates (MAPE, C-index, MAE) cannot detect this cascade: they are per-component safety systems in a tightly coupled architecture, structurally blind to interaction failures. The required mitigations are (1) quarterly end-to-end cohort validation that bypasses individual model metrics; (2) layer-boundary distribution logging to detect amplification in real time; (3) a cascade-sensitivity Monte Carlo to quantify the 95th-percentile downstream impact of a 1% price-model error on the execution engine's false-positive rate; and (4) top-down error budget allocation, deriving the price model's MAPE tolerance from the execution engine's false-positive tolerance instead of setting both independently.

Decision Intelligence Maturity: Alignment with Gartner Guidance

We benchmark the Control Tower against Gartner's 2026 decision intelligence guidance on operationalizing multiagent systems (den Hamer, 2026). Table 4 presents the alignment matrix.

Table 4. Gartner decision intelligence alignment: Control Tower assessment

Dimension	Gartner recommendation	Control Tower implementation	Alignment
Collaboration	Define an Automated / Augmented / Autonomy spectrum per decision type by risk and complexity	A/B/C taxonomy with numeric thresholds, authority levels, escalation limits, and forced exploration (Section 3.2)	Full; adds operational specificity and behavioral grounding
Change	Redesign decision flows for AI rather than automating broken processes	Execution engine redesigns the repricing flow; override-as-data prevents automating miscalibrated intuition; shadow log captures divergence	Full
Composition	Combine ML, rules, probabilistic, and deterministic optimization (composite AI)	EBM + Extended Cox + MILP + rule-based RPCE + Bayesian demand signal in one cascading engine	Full
Context	Use a context graph for constrained retrieval (reported accuracy gain of about 40%)	Market indices and rolling features provide context; no structured entity-relationship graph	Partial; Phase 5+ upgrade path

Dimension	Gartner recommendation	Control Tower implementation	Alignment
Agentic AI	Adopt multiagent systems (Orchestrator, Guardian, Expert, Planner)	RPCE is a formula-based batch engine, not a stateful orchestrator; no agent-to-agent layer	Not yet; Phase 5+
GDI cycle	Design around Capture → Interpret → Model → Resolve → Act with feedback	L0 → L1–L2 → RPCE → authority matrix → realized sales and shadow log; three feedback loops	Full

The context gap. The current context layer captures market dynamics but lacks structured entity relationships. The most valuable graph dimensions would be vehicle-identity nodes linking make, model, generation, and substitutes; geography nodes with price-transmission lag edges; temporal event nodes capturing market regimes with causal-factor edges; and competitive-context nodes that address FM-06. This is a Phase 5+ initiative contingent on FM-08 data quality infrastructure.

Agentic AI readiness. The Control Tower currently operates as a single-platform DI system. The execution engine functions as a *de facto* orchestrator, the price and survival models as expert agents, the optimizer and MILP as planner agents, and the VaR limits and stop-loss as a guardian layer. What is absent is agent orchestration infrastructure: stateful real-time triggers, agent-to-agent communication, and event-driven replanning. Two prerequisites apply. All model outputs must be written to a versioned feature store with event timestamps, and the FM-11 cascade-sensitivity analysis must be completed and an error budget established. Under Perrow's framework, agentic architectures increase interactive complexity without increasing loose coupling, so introducing agent-to-agent communication before FM-11 mitigation is operational would compound an unquantified normal-accident risk.

DISCUSSION AND CONCLUSION

This paper proposed DEOS as a meta-governance architecture above the Control Tower's seven-layer cascading DI stack. The central argument is that cascading model failure cannot be addressed by layer-level MLOps alone. It requires governance that monitors the system as a whole (end-to-end validation), introduces unconditional decoupling mechanisms into a tightly coupled architecture (hard rules in L4, following Perrow's buffering requirement), treats model failure as a generative dialectical event, and converts algorithm aversion into a calibration asset rather than an adoption barrier.

Normal Accidents Theory also justifies a governance choice that might otherwise look conservative: retaining humans in the loop for all price movements above 3%. In a tightly coupled, interactively complex system, the human in the loop is the only real-time cascade interruption mechanism, so the A/B/C taxonomy is an architectural requirement rather than a compromise. The shadow log, structured override rationale, and forced exploration together serve three functions at once: a noise measurement system, a source of unbiased calibration data, and a trust-building mechanism. This triple function distinguishes a DEOS-governed platform from a conventional MLOps deployment.

Limitations. All results are theoretical. The stress test is a constructed scenario, thresholds such as the 15–30% override band, the 5% exploration rate, and the 3.5 Trust Score gate are design parameters to

be calibrated rather than empirically derived values, and Hegelian and Gödelian constructs are used as design heuristics, not formal proofs.

Future work. Priorities are (1) empirical validation of FM-11 cascade sensitivity through end-to-end cohort analysis once Phase 3 is operational; (2) a dynamic λ calibration protocol that uses a coupling coefficient as an input to risk appetite; (3) comparison of the $U(D) < U*$ trigger against distributionally robust optimization as an alternative worst-case mechanism; (4) a context-graph architecture for Phase 5; (5) a Viable System Model mapping (Beer, 1972) to clarify whether the Control Tower occupies System 4 (environmental intelligence) or System 5 (policy and identity) in the holding's organizational cybernetics, a prerequisite for extending DEOS to agentic architectures across business units; and (6) a longitudinal study of algorithm aversion dynamics during the shadow period, testing whether forced exploration and structured override rationale produce the predicted Trust Score trajectories.

DECLARATIONS

Use of AI tools: Generative AI tools were used in preparing this manuscript. Anthropic's Claude was used for language editing, restructuring and condensing sections, converting the manuscript into the conference template. The author reviewed and verified all content, equations, numerical results and references, and take full responsibility for the paper.

References

Akçelik, F., Başçı, E., Özel, Ö., & Sarıkaya, F. (2022). A monetary policy framework for the next decade: Lessons from the Turkish experience. *BIS Papers*, No. 119.

Argyris, C., & Schön, D. A. (1978). *Organizational learning: A theory of action perspective.* Addison-Wesley.

Beer, S. (1972). *Brain of the firm: The managerial cybernetics of organization.* Allen Lane.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B, 34*(2), 187–202.

den Hamer, P. (2026). *Decision intelligence: How to operationalize multiagent systems.* Gartner Research.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126.

Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für Mathematik und Physik, 38*, 173–198.

Hegel, G. W. F. (1807). *Phänomenologie des Geistes.* Joseph Anton Goebhardt.

Hofstadter, D. R. (1979). *Gödel, Escher, Bach: An eternal golden braid.* Basic Books.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment.* Little, Brown Spark.

Kupiec, P. H. (1995). Techniques for verifying the accuracy of risk measurement models. *The Journal of Derivatives, 3*(2), 73–84.

Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis, 88*(2), 365–411.

Lou, Y., Caruana, R., & Gehrke, J. (2012). Intelligible models for classification and regression. *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 150–158.

Paleyes, A., Urma, R.-G., & Lawrence, N. D. (2022). Challenges in deploying machine learning: A survey of case studies. *ACM Computing Surveys, 55*(6).

Perrow, C. (1984). *Normal accidents: Living with high-risk technologies.* Basic Books.

Pratt, L. (2019). *Decision intelligence: Transform your team and organization with AI-powered decision-making.* Springer.

Rockafellar, R. T., & Uryasev, S. (2000). Optimization of conditional value-at-risk. *Journal of Risk, 2*(3), 21–41.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems, 28.*

Soros, G. (1987). *The alchemy of finance.* Simon & Schuster.

TCMB [Türkiye Cumhuriyet Merkez Bankası]. (2022). *Yıllık Rapor 2021.* Ankara.

Zollo, M., & Winter, S. G. (2002). Deliberate learning and the evolution of dynamic capabilities. *Organization Science, 13*(3), 339–351.

Information Technologies and Data-Informed Decision-Making in Education: Opportunities and Limitations

Reşat Timuçin Özkan[1]*

[1] University of Mediterranean Karpasia, Department of Educational Administration, 99010 Nicosia, TRNC

Corresponding author e-mail: timucin.ozkan@akun.edu.tr

Abstract

Information technologies offer opportunities to support teaching, monitor student progress, and inform administrative decisions in educational institutions. However, collecting educational data does not, by itself, ensure effective decision-making. Data must be connected to pedagogical goals, interpreted within their context, and used in accordance with ethical principles. This paper aims to examine the opportunities and limitations of information technologies in education from the perspective of data-informed decision-making.

The paper adopts a conceptual approach, addressing learning management systems, digital assessment tools, and student information systems in relation to teaching and educational administration. It discusses how attendance, participation, assignment completion, and assessment data may help identify students' support needs. Its central argument is that digital indicators cannot fully represent students' learning processes. For example, time spent on a learning platform does not necessarily reflect the quality of learning. Digital data should therefore be interpreted alongside teacher observations, student perspectives, and the characteristics of the educational setting.

The discussion highlights unequal access to technology, data quality, teacher and administrator data literacy, and the protection of personal information. It argues that the educational value of information technologies should be assessed through their contribution to learning and informed decision-making. To strengthen data-informed practices, educational institutions should integrate digital infrastructure with pedagogical expertise, address differences in access, and establish ethical principles for data use. The paper concludes that educational data can support professional judgement when interpreted critically and with attention to context.

Key words: *Information Technologies, Educational Administration, Data-Informed Decision-Making, Data Literacy, Digital Inequality*

INTRODUCTION

The use of information technologies in education raises an important question about how educational evidence is produced and interpreted. Learning management systems, digital assessment tools, and student information systems can make information about attendance, participation, assignment completion, and assessment performance available to teachers and educational administrators. These

records may support the identification of difficulties and the planning of educational responses. However, the availability of digital information does not, by itself, establish its educational meaning or the appropriateness of a decision based on it.

The value of educational technology therefore needs to be considered in relation to the purposes and conditions of its use. UNESCO's Global Education Monitoring Report emphasises the importance of evaluating technology in terms of educational needs, equity, evidence, and sustainability (UNESCO, 2023). This perspective provides a useful basis for examining digital systems as resources for educational decision-making. Their contribution depends on the problems they help address, the learners they serve, and the capacity of educational institutions to use them responsibly.

In this article, data-informed decision-making refers to the use of relevant evidence alongside professional judgement and knowledge of the educational context. This distinction is important because digital indicators provide partial representations of educational activity. For example, a record of assignment completion identifies whether a task was submitted, but it does not necessarily explain the learner's understanding, the support received, or the circumstances affecting completion. Similarly, limited activity within an online platform may prompt further inquiry, but it should not automatically be interpreted as a lack of motivation. These examples illustrate why the interpretation of digital records requires attention to what an indicator measures and what remains outside its scope.

Educators' capacity to interpret evidence is central to this process. Mandinach and Gummer (2016) conceptualise teacher data literacy through the knowledge, skills, and dispositions involved in educational data use. Their work supports an examination of decision-making that extends beyond accessing information or reading numerical summaries. Educationally meaningful interpretation requires a connection between the evidence available, the question being investigated, and the response being considered.

Against this background, the present article examines the opportunities and limitations of information technologies in supporting data-informed decision-making in education. It focuses on learning management systems, digital assessment tools, and student information systems, with particular attention to attendance, participation, assignment completion, and assessment data. The discussion considers how these forms of evidence can support teaching and educational administration while also examining the constraints associated with incomplete indicators, unequal access, data quality, educators' interpretive capacity, and privacy.

The article addresses three questions: What opportunities do information technologies offer for educational decision-making? What limitations should be considered when interpreting digitally generated indicators? What principles can guide their ethical and context-sensitive use? Through this conceptual discussion, the article develops an argument for connecting digital evidence with pedagogical understanding and accountable professional judgement.

Conceptual background

Information Technologies in Education

Information technologies in education encompass systems used to organise educational activities, facilitate communication, document student progress, and support institutional management. This article focuses on three categories: learning management systems, digital assessment tools, and student information systems. Although these systems serve different purposes, each can provide information relevant to teaching and educational administration.

Learning management systems organise course materials, assignments, feedback, and online activities. Their records can help educators examine patterns of participation and task completion. Digital assessment tools provide information about responses to particular questions or tasks, potentially helping teachers identify areas that require further explanation. Student information systems bring together administrative records, such as attendance and grades, which may support the coordination of student services and institutional planning.

The educational significance of these systems depends on how their records are interpreted and used. A platform can record that a student opened a resource, but that record does not establish whether the student understood its content. Likewise, a numerical assessment result may indicate a difficulty without explaining its cause. Information technologies should therefore be understood as sources of evidence whose relevance and limitations need to be examined in relation to a particular educational question. Gašević et al. (2015) similarly argue that computational approaches to learning analytics should be connected with educational research.

Data-Informed Decision-Making

Data-informed decision-making involves considering relevant evidence alongside professional judgement, educational goals, and contextual knowledge. In this article, the term emphasises the contribution of data to deliberation rather than the automatic determination of decisions by numerical indicators.

An educational decision begins with a question. A teacher may wish to understand why students are struggling with a particular topic, while a school administrator may seek to identify factors associated with repeated absence. The question determines which information is relevant and whether existing records are sufficient. Collecting more data without clarifying the decision to be made can produce additional information without improving understanding.

The process also requires a distinction between description and explanation. Attendance records describe patterns of presence and absence; they do not independently explain those patterns. Assignment records show submission status; they do not necessarily reveal the barriers encountered by learners. Consequently, digital evidence may need to be considered alongside classroom observations, student work, and conversations with students.

Drawing on the iterative perspective described by Schildkamp (2019), this discussion organises data-informed decision-making around five activities: defining the educational question, selecting relevant evidence, interpreting that evidence in context, choosing an appropriate response, and reviewing the consequences of the response. This sequence provides an organising structure for the article; it is not presented as a newly developed or empirically validated model.

Digital Indicators and Educational Meaning

Digital indicators are selected representations of educational activities or outcomes. Attendance, platform participation, assignment completion, and assessment results each make particular aspects of student experience visible. Their usefulness depends on the relationship between the recorded activity and the educational concept being considered.

Participation, for example, can refer to several different behaviours. Posting a message in an online discussion differs from contributing to a classroom conversation, collaborating with peers, or engaging in independent study. A system that records one behaviour cannot be assumed to capture participation in its entirety. Platform access frequency should therefore be treated as a behavioural indicator whose educational meaning requires further interpretation.

Interpretation must also account for the circumstances in which records are produced. A student with unreliable internet access may leave fewer digital traces than a student with continuous access, even when both make substantial efforts to learn. Missing or limited records can therefore raise questions about access and data collection as well as questions about educational engagement.

These distinctions suggest that digital indicators are most useful when their scope is explicit. Educators need to establish what a measure represents, what it excludes, and which additional evidence is necessary before taking action. An indicator may justify further inquiry without being sufficient to justify a consequential decision about a student. Wise and Shaffer (2015) emphasise the role of theory in selecting variables and interpreting analytical results. The examples here apply that general concern to routine educational records.

Professional Judgement and Ethical Responsibility

Professional judgement connects evidence with educational action. Teachers and administrators must assess whether information is relevant, whether an interpretation is reasonable, and whether a proposed response serves an appropriate educational purpose. This responsibility remains important when digital systems produce summaries, classifications, or alerts.

Ethical considerations form part of that judgement. Institutions should consider why particular data are needed, who can access them, and how their use may affect students. Decisions also require attention to the possibility of misinterpretation. A label derived from incomplete records may influence expectations about a learner even when the underlying evidence is uncertain.

The framework adopted in this article therefore connects three elements: digital evidence, contextual interpretation, and professional responsibility. Information technologies contribute records and analytical possibilities; educational practitioners assess their meaning and decide how to respond. This relationship provides the basis for examining both the opportunities and the limitations discussed in the following sections.

MATERIAL AND METHODS

Conceptual approach

This article adopts a conceptual approach to examine how information technologies can support data-informed decision-making in education and under what conditions their use may be limited. Its purpose is to clarify the relationship between digitally generated evidence, educational interpretation, and professional responsibility. The article does not report original participant data, test a technological intervention, or estimate the effects of digital systems on student achievement.

The discussion is organised around three connected questions: what educational decisions digital information can support, what limitations affect the interpretation of that information, and what principles can guide its responsible use. This approach allows the opportunities offered by information technologies to be considered alongside the conditions necessary for their educational value.

Scope of the discussion

The article focuses on learning management systems, digital assessment tools, and student information systems. These categories provide a manageable scope for examining the use of attendance, participation, assignment completion, and assessment records in teaching and educational administration.

Two levels of decision-making are considered. At the instructional level, the discussion addresses how teachers may use available evidence to identify learning difficulties, adjust teaching activities, and provide feedback or support. At the institutional level, it considers how administrators may interpret aggregated records to coordinate student support and inform planning.

The scope does not include a technical comparison of commercial platforms or an evaluation of particular predictive algorithms. References to digital indicators concern their educational interpretation rather than their computational performance. The article also distinguishes between information that identifies a pattern and evidence that explains the reasons for that pattern.

Literature basis

This focused conceptual discussion uses selected publications on educator data literacy (Mandinach and Gummer, 2016), school improvement through data use (Schildkamp, 2019), feedback (Hattie and Timperley, 2007), ethical issues in learning analytics (Slade and Prinsloo, 2013), educational technology policy (UNESCO, 2023), and the educational and theoretical interpretation of learning analytics (Gašević et al., 2015; Wise and Shaffer, 2015). Sources were selected for their direct conceptual relevance to these topics. This purposive selection does not constitute an exhaustive corpus or systematic search.

The selected literature clarifies concepts and provides perspectives for examining opportunities, limitations, and practical implications. Conceptual publications and review-based discussions inform the argument, while UNESCO's report provides an institutional policy perspective. These sources are used to develop a reasoned discussion; their inclusion does not establish the effectiveness of a specific platform or intervention.

The article is a focused conceptual discussion rather than a systematic review. It does not claim exhaustive coverage of the literature, a reproducible database search, or a formal assessment of all

available studies. Interpretations are limited to the concepts and arguments supported by the selected sources. The illustrative examples are author-developed scenarios, not empirical observations.

Analytical organisation

The discussion examines each opportunity in relation to its interpretive requirements and possible limitations. For example, the potential use of participation records to identify students who may need support is considered alongside questions about access, the meaning of participation, and the completeness of the records.

Four dimensions organise this examination: the educational purpose of the decision, the adequacy of the available evidence, the contextual conditions affecting interpretation, and the responsibilities associated with action. These dimensions form an author-proposed structure for the conceptual argument; they are not presented as a validated measurement instrument or an empirical coding scheme.

This organisation supports a balanced examination of technological possibilities. It directs attention both to what digital systems make visible and to the additional understanding required before their outputs can inform educational action.

Limitations of the approach

The conceptual nature of the article limits the conclusions that can be drawn. The discussion cannot establish that a particular technology improves learning, reduces inequality, or increases administrative effectiveness. Such claims require empirical investigation in defined educational settings.

The examples used in the article illustrate interpretive problems and possible responses. They are not observations collected from participants or evidence of the frequency of particular practices. Furthermore, the relevance of the proposed considerations may vary according to educational level, institutional resources, technological infrastructure, and local governance arrangements.

Accordingly, the article's contribution lies in clarifying the conditions under which digital evidence may inform educational decisions and identifying issues that require further empirical examination.

RESULTS

The results presented below are the outcome of the conceptual synthesis of the selected literature. They identify potential opportunities and limitations; they are not findings from newly collected participant data or estimates of intervention effects.

Opportunities of information technologies for data-informed decision-making

Information technologies can support educational decision-making by making relevant records available for interpretation and action. Their potential contribution extends from identifying students who may need assistance to reviewing instructional activities and informing institutional planning. The opportunities examined below depend on the relevance and quality of the available evidence, educators' interpretive capacity, and institutions' ability to respond. They should therefore be understood as possible uses rather than guaranteed improvements in educational outcomes.

IDENTIFYING STUDENTS WHO MAY NEED SUPPORT

Attendance, assignment completion, assessment results, and online participation records can provide starting points for identifying students who may benefit from additional support. Examining these records together may help educators recognise patterns that deserve attention. For example, repeated absences accompanied by missed assignments could prompt a teacher to contact a student and explore the circumstances affecting participation.

The opportunity lies in using available information to initiate timely inquiry. A digital record does not independently establish why a student is experiencing difficulty. Limited participation may reflect problems with access, uncertainty about a task, competing responsibilities, or other circumstances. Consequently, the interpretation of a pattern should be followed by dialogue and consideration of additional evidence.

Digital systems may also support continuity in student assistance. Where access permissions and institutional procedures allow, relevant staff can use shared records to coordinate their responses and review whether support has been provided. The educational value of this coordination depends on clear responsibilities and an emphasis on assistance rather than the automatic classification of students.

INFORMING INSTRUCTIONAL ADJUSTMENT

Digital assessment tools and learning management systems can provide information that helps teachers reconsider the content, pace, and organisation of instruction. Responses to assessment tasks may indicate topics that require further explanation, while assignment records can help identify stages at which students encounter difficulties.

For example, when several students select the same incorrect response, the teacher may examine whether the pattern reflects a shared misconception, ambiguous wording, or insufficient prior instruction. Each explanation suggests a different response. Reviewing student work and discussing their reasoning can help determine whether to revise the assessment item, revisit a concept, or provide additional practice.

Information can also be considered across activities. Comparing students' initial responses with later work may help teachers judge whether an instructional adjustment appears to address the difficulty identified. Such comparisons require attention to differences in task demands and assessment conditions. A change in scores alone does not establish that a particular teaching strategy caused the change.

The main opportunity is therefore a closer connection between evidence and instructional reflection. Teachers can use digital records to formulate questions about learning and select responses suited to the difficulties they identify.

SUPPORTING FEEDBACK AND STUDENT REFLECTION

Digital tools can help organise feedback and make it available to students alongside their work. Depending on the system, teachers may provide comments, assessment criteria, or opportunities for revision. These functions can support a feedback process in which students examine their performance and identify possible next steps.

The usefulness of feedback depends on its educational content. A score or an automated indication that an answer is incorrect may provide limited guidance. Feedback becomes more actionable when it explains the difficulty, relates it to the learning objective, and suggests how the student can improve. Automated responses may assist with some tasks, while more complex work may require teacher interpretation and individual explanation. The distinction between feedback delivery and feedback quality is consistent with Hattie and Timperley's (2007) discussion of how feedback effects depend on its form and use.

Students may also use records of their assignments and assessments to reflect on their progress. A teacher-guided review could invite students to identify recurring difficulties, compare strategies, and set a specific learning goal. Here, the digital record serves as a resource for reflection. Its purpose is to support understanding and action rather than to encourage constant comparison with peers.

Opportunities for feedback should also account for students' ability to access, understand, and use the information provided. Making feedback available within a platform does not establish that students have read it or know how to respond to it.

INFORMING INSTITUTIONAL PLANNING AND COORDINATION

At the institutional level, aggregated records can inform questions about attendance, student support, assessment practices, and resource allocation. Administrators may examine patterns across classes or periods to identify issues requiring further investigation. For instance, a concentration of incomplete assignments across several courses may prompt a review of assessment schedules and workload expectations.

Such information can help institutions move from a broad concern to a more clearly defined planning question. However, aggregated patterns require careful interpretation. Differences between classes may reflect student composition, subject demands, recording practices, or access conditions. Comparisons should therefore consider whether the underlying measures are sufficiently consistent and comparable.

Digital systems can also support coordination by bringing relevant information into a shared administrative process. A planning meeting may combine attendance summaries, teachers' observations, and student feedback to determine where support is needed. The resulting decision should specify an intended response, responsible staff, and a suitable point for review.

Resource decisions require particular care because recorded demand may differ from actual need. Students whose difficulties are poorly captured by existing systems can remain less visible. Institutional planning should therefore allow additional routes through which teachers and students can communicate needs that digital records do not adequately represent.

REVIEWING EDUCATIONAL RESPONSES

Information technologies can provide records that help educators review what happens after a decision is implemented. Following a change in teaching arrangements or the introduction of student support, relevant information may help determine whether the intended issue persists and whether further adjustment is required.

For example, an institution responding to missed assignments might review subsequent submission patterns while also asking students and teachers about their experience of the response. Combining these forms of evidence can provide a more informative account than relying on submission rates alone.

This review process should distinguish improvement in a recorded indicator from improvement in the broader educational experience. Increased platform activity, for instance, may show a change in recorded behaviour without demonstrating stronger understanding. Reviewing a response therefore requires evidence aligned with its educational purpose and attention to possible unintended consequences.

Taken together, these opportunities suggest that information technologies can support a cycle of inquiry, action, and review. Their contribution is strongest when records are connected to clear educational questions, interpreted with contextual knowledge, and followed by appropriate professional responses. The following section examines the limitations that can weaken or complicate this process.

Limitations of information technologies for data-informed decision-making

The opportunities discussed above depend on how digital records are produced, interpreted, and translated into educational action. Information technologies capture selected aspects of learning and institutional activity, leaving other aspects less visible. Their limitations therefore concern both the adequacy of the available evidence and the conditions under which decisions are made.

PARTIAL REPRESENTATION OF LEARNING AND PARTICIPATION

Digital records represent activities that a system is designed to capture. Platform access, assignment submission, and assessment responses can provide useful information, but they do not encompass the full range of learning experiences. Students may engage in independent reading, face-to-face discussion, handwritten practice, or collaborative work that remains outside the platform's records. This discussion follows the broader requirement to connect analytical measures with educational understanding rather than assume that records explain learning on their own (Gašević et al., 2015; Wise and Shaffer, 2015).

This creates a distinction between a recorded behaviour and the educational meaning attributed to it. Opening a resource does not establish comprehension, while spending more time on a task does not necessarily indicate greater engagement. A longer recorded duration might reflect careful study, difficulty understanding instructions, or an unattended session. Interpretation requires consideration of these alternative explanations.

The same limitation applies to participation measures. A student who contributes frequently to an online discussion may be highly active, but the number of contributions alone does not establish their relevance or depth. Conversely, a student who posts less frequently may still participate meaningfully through other activities. Educational decisions should therefore examine the relationship between an indicator and the concept it is intended to represent.

DATA QUALITY AND COMPARABILITY

The usefulness of digital evidence depends on the accuracy, completeness, timeliness, and consistency of the records. Missing entries, incorrect information, delayed updates, or different recording practices

can affect the patterns presented to educators. A missing assignment record, for example, may indicate non-submission, a technical problem, or an incomplete administrative entry.

Comparability also requires attention. Attendance, participation, and assessment measures may be defined differently across courses or institutions. One teacher may record participation through discussion contributions, while another may use task completion. Combining these measures into a single summary can conceal differences in what they represent.

Assessment results present similar difficulties. Scores from tasks with different content, levels of difficulty, or marking criteria cannot automatically be treated as equivalent. Before drawing conclusions from comparisons, educators need to consider whether the records address sufficiently similar questions.

These limitations make verification an essential part of interpretation. When a pattern could lead to a consequential decision, the underlying information should be checked and uncertainties acknowledged. Apparent precision in a dashboard does not remove uncertainty in the records from which it was produced.

UNEQUAL ACCESS AND UNEQUAL VISIBILITY

Students' opportunities to participate in digital activities may vary according to device availability, internet connectivity, accessibility requirements, and the conditions in which they study. UNESCO's discussion of educational technology identifies equitable access as an important condition for its educational use (UNESCO, 2023).

These differences can affect the digital traces available for decision-making. A student who shares a device or experiences unstable connectivity may access a platform less frequently or submit work later. If these records are interpreted without considering access conditions, a difficulty in using the technology may be treated as evidence of insufficient effort.

Unequal access also raises questions about whose experiences become visible. Students who complete activities outside the platform may contribute fewer records, even when they remain engaged in learning. Institutional decisions based predominantly on digital activity could consequently overlook needs that are poorly represented in the system.

Addressing this limitation requires opportunities for students to explain their circumstances and demonstrate learning through appropriate alternatives. It also requires institutions to examine whether the use of a particular digital indicator disadvantages students whose participation conditions differ.

DATA LITERACY AND INTERPRETIVE CAPACITY

Access to information does not ensure that educators can interpret it appropriately. Data literacy involves the knowledge, skills, and dispositions needed to connect evidence with educational questions and possible responses (Mandinach and Gummer, 2016). Reading a graph or identifying a change in a score forms only part of this process.

A central challenge is distinguishing a pattern from an explanation. When lower assessment results occur alongside limited platform activity, the association does not establish that limited activity caused

the lower results. Both may be related to other circumstances, including prior learning difficulties or barriers to access.

Interpretation can also be influenced by existing expectations. An educator may give greater attention to records that support an initial judgement while overlooking contradictory evidence. Digital summaries can reinforce this tendency when their categories appear authoritative or when the basis of a classification is unclear.

Professional learning should therefore address question formulation, the evaluation of indicators, the consideration of alternative explanations, and the communication of uncertainty. Opportunities for teachers and administrators to discuss evidence together may also help expose assumptions that would otherwise remain unexamined. Such discussion requires time and organisational support.

PRIVACY AND RESPONSIBLE DATA USE

The collection and use of educational records require clear decisions about purpose, access, retention, and responsibility. Information gathered to support learning may also be used for administrative monitoring or evaluation. These uses need to be considered explicitly because they can involve different expectations and consequences. These questions connect with the ethical concerns regarding privacy, interpretation, and transparency identified in higher education learning analytics by Slade and Prinsloo (2013).

Institutions should establish which information is necessary for a defined purpose and which staff require access to it. Broad access to detailed student records is not automatically justified by the availability of a shared system. Similarly, retaining information indefinitely may extend its use beyond the educational purpose for which it was collected.

Students and, where appropriate, their parents should receive understandable information about relevant data practices. Procedures for identifying and correcting inaccurate records are also important, particularly when those records contribute to decisions about support, assessment, or participation.

Privacy is connected to the way information is interpreted. A student should have an appropriate opportunity to explain circumstances that a record does not capture. Responsible data use therefore involves both the protection of information and fair consideration of the person represented by it.

DEPENDENCE ON DIGITAL SUMMARIES AND CLASSIFICATIONS

Dashboards, automated alerts, and summary categories can make large quantities of information easier to navigate. However, they can also direct attention towards the indicators selected by the system. Educational concerns that are difficult to quantify may receive less attention because they are less prominent in the interface.

A classification indicating that a student may require support should prompt inquiry rather than settle a judgement. Its meaning depends on the underlying indicators, the criteria used, and the circumstances of the student. When these elements are unclear, educators may find it difficult to assess whether the classification is relevant.

There is also a risk that institutional priorities become aligned with improving visible indicators rather than addressing the educational purposes behind them. Increasing recorded participation, for example, does not necessarily improve the quality of learning. Institutions should periodically review whether their chosen measures remain aligned with their educational goals.

CAPACITY TO TRANSLATE EVIDENCE INTO ACTION

Even relevant and well-interpreted information may have limited value when institutions lack the capacity to respond. Identifying students who need assistance does not ensure that appropriate staff, time, or support services are available. The usefulness of a digital system consequently depends in part on the resources and procedures surrounding it. The importance of leadership and organisational conditions in data use is also identified by Schildkamp (2019).

Data-related activities can also create additional work. Entering records, checking their accuracy, reviewing summaries, and discussing possible responses require time. If these tasks are poorly integrated into existing practices, they may compete with teaching and student support.

Institutions therefore need to connect information collection with a realistic response process. This includes clarifying who reviews the evidence, who takes action, and how the response will be evaluated. Collecting information without these arrangements may increase administrative activity while producing little educational benefit.

Together, these limitations demonstrate why digital evidence requires contextual interpretation and accountable professional judgement. The following discussion brings the opportunities and limitations together to consider the conditions for meaningful educational use.

DISCUSSION AND CONCLUSION

Connecting digital evidence with educational purposes

The conceptual examination suggests that the contribution of information technologies depends on the relationship between the information collected and the educational decision under consideration. Attendance, participation, assignment completion, and assessment records can identify patterns deserving attention. Their value increases when educators can explain why a pattern matters and what further evidence is needed before responding.

Schildkamp (2019) describes data use for school improvement as an iterative process involving goals, evidence, interpretation, action, and evaluation. The sequence discussed in this article is consistent with that perspective. The present discussion applies it to the interpretation of records from learning management systems, digital assessment tools, and student information systems. This application is a conceptual organisation of the argument, not a newly validated decision-making model.

A practical implication is that institutions should define the educational question before selecting indicators. When the concern is students' understanding of a concept, platform activity may provide contextual information, while examination of student reasoning may offer more direct evidence. When the concern is access to learning activities, participation records may need to be interpreted alongside

information about connectivity and device availability. The suitability of a measure therefore depends on the purpose for which it is used.

Professional capacity and institutional responsibility

The opportunities and limitations are closely connected. A record that helps identify a possible learning difficulty can also lead to an inappropriate judgement when its scope is misunderstood. Similarly, a shared information system can support coordination while creating additional responsibilities for access control and accuracy. These relationships suggest that technological provision and professional preparation should be considered together.

Mandinach and Gummer (2016) identify educator data literacy as a combination of skills, knowledge, and dispositions. In the context considered here, this means that professional development should extend beyond operating a platform to evaluating the educational meaning of its records. Teachers need opportunities to formulate questions, examine alternative explanations, and decide when additional evidence is required. Administrators need to support this work through time, clear responsibilities, and feasible responses to the needs identified.

Digital feedback provides a further example of the importance of professional capacity. Hattie and Timperley (2007) explain that feedback can have different effects depending on its type and the way it is provided. A platform's ability to deliver information should therefore be distinguished from the pedagogical quality of that information. The practical question is whether students can understand the feedback and use it to improve their subsequent work.

Equity and ethical interpretation

UNESCO (2023) places educational needs, equity, and contextual suitability at the centre of decisions about educational technology. Applied to data-informed decision-making, this perspective requires educators to examine the circumstances in which digital records are produced. An apparent lack of participation may reflect an access barrier, and a record that is technically accurate may still support an incomplete interpretation.

Ethical interpretation also involves the student's opportunity to explain or question a judgement. Slade and Prinsloo (2013), writing about learning analytics in higher education, identify privacy, transparency, power, and context as important ethical concerns. Their perspective informs this article's argument that student records should be used with attention to the people and circumstances they represent. Applying these considerations across educational levels requires sensitivity to students' age and institutional responsibilities; the higher education discussion does not establish a universal procedure for every setting.

The implication developed here is that student support should include a route for clarifying uncertain or inaccurate information. Where a digital indicator prompts concern, educators should consider an appropriate conversation or review before making a consequential decision. This recommendation follows from the conceptual argument rather than from an intervention evaluated in this study.

Contribution and directions for research

The article's contribution is an integrated discussion of routine digital indicators across instructional and administrative decisions. It connects each possible use with four interpretive requirements: educational purpose, evidence adequacy, contextual conditions, and responsibility for action. This synthesis applies established concerns to attendance, participation, assignment completion, and assessment records. Its value is conceptual clarification; it does not claim a new validated theory or demonstrate effects on educational outcomes.

Further research could examine how teachers and administrators interpret the same digital indicators, what additional evidence they seek, and how their decisions differ across access conditions. Studies could also investigate whether students understand how their records are used and whether they can correct or contextualise them. Evaluations of student support should examine both recorded outcomes and students' experiences, while documenting the resources required for implementation.

Conclusion and recommendations

Information technologies offer opportunities to support student assistance, instructional adjustment, feedback, coordination, and institutional planning. Learning management systems, digital assessment tools, and student information systems can provide evidence relevant to these activities. Their educational contribution nevertheless depends on the appropriateness of the indicators, the quality of the records, and the capacity to interpret and act on them.

The central conclusion of this conceptual discussion is that digital evidence requires pedagogical and contextual interpretation. Attendance, participation, assignment completion, and assessment results represent selected aspects of educational activity. They may justify further inquiry, but they do not independently explain students' circumstances or establish the effectiveness of an educational response.

The discussion supports five practical recommendations. First, educators should define the educational question and identify what evidence is relevant before relying on a digital summary. Second, records should be checked for accuracy, completeness, and comparability, particularly when they inform consequential decisions. Third, interpretation should consider access conditions and incorporate additional evidence, including student work, observations, and appropriate dialogue. Fourth, institutions should establish clear arrangements for data access, correction, retention, and responsibility. Fifth, they should provide professional learning and sufficient resources to implement and review the responses selected.

These recommendations form a conceptual basis for practice and further investigation. They have not been tested as a package of interventions. Future empirical work should examine how the proposed considerations function in different educational settings and whether they support fairer interpretation and more effective assistance. The responsible use of information technologies ultimately involves connecting evidence with a clear educational purpose, informed professional judgement, and attention to the consequences for learners.

References

Gašević D, Dawson S, Siemens G, 2015. Let's not forget: Learning analytics are about learning. TechTrends, 59 (1): 64–71. https://doi.org/10.1007/s11528-014-0822-x

Hattie J, Timperley H, 2007. The power of feedback. Review of Educational Research, 77 (1): 81–112. https://doi.org/10.3102/003465430298487

Mandinach EB, Gummer ES, 2016. What does it mean for teachers to be data literate: Laying out the skills, knowledge, and dispositions. Teaching and Teacher Education, 60: 366–376. https://doi.org/10.1016/j.tate.2016.07.011

Schildkamp K, 2019. Data-based decision-making for school improvement: Research insights and gaps. Educational Research, 61 (3): 257–273. https://doi.org/10.1080/00131881.2019.1625716

Slade S, Prinsloo P, 2013. Learning analytics: Ethical issues and dilemmas. American Behavioral Scientist, 57 (10): 1510–1529. https://doi.org/10.1177/0002764213479366

UNESCO, 2023. Global education monitoring report 2023: Technology in education – A tool on whose terms? UNESCO, Paris, France. https://unesdoc.unesco.org/ark:/48223/pf0000385723

Wise AF, Shaffer DW, 2015. Why theory matters more than ever in the age of big data. Journal of Learning Analytics, 2 (2): 5–13. https://doi.org/10.18608/jla.2015.22.2

Conflict of Interest

The author declares that there is no conflict of interest.

FlowFormer: An Explainable Dual-Positional Sequence-Labelling Network Intrusion Detection System with Two-Axis Attention Attribution

Areeba Sadaqat[1], Haroon Ayaz[2], Saeed Ur Rehman[3]*, Syeda Rubab Javed[4], Muhammad Affan[5], Iqra Furrukh[6]

[1,3,4,5,6] COMSATS University Islamabad, Department of Computer Science, Wah Campus, 47040, Wah Cantt, Pakistan

[2] School of Computing and Engineering Sciences, University of Chester, Chester, United Kingdom

*Corresponding author e-mail: srehman@ciitwah.edu.pk

Abstract

Modern Security Operations Centers (SOCs) depend heavily on Network Intrusion Detection Systems (NIDSs) to spot cyber threats. However, traditional single-flow models miss multi-stage attacks, while deep learning alternatives function as opaque black boxes. To address these limitations, we propose FlowFormer, an explainable deep learning framework built for sequence-based network traffic analysis. Rather than evaluating isolated records, FlowFormer inspects temporal flow sequences using dual positional encoding for sequence order and precise inter-arrival times. It categorizes network metrics into volumetric, temporal, flag-based, and rate-based feature groups, while a two-axis attention rollout mechanism pinpoints the exact flows and features driving each alert. Evaluated on CICIDS2017, CSE-CICIDS2018, and UNSW-NB15 across temporal, leave-one-attack-out, and cross-corpus protocols, FlowFormer improves recall over XGBoost, BiLSTM, and Vanilla Transformer baselines, particularly on botnet, port scanning, and brute-force attacks while maintaining a 0.968 Macro-F1 score on unseen classes. Single-sample metrics such as precision (0.246) and recall (0.250) were found to be correct as individual data slices.

Keywords: *Network intrusion detection system, Transformer, Explainable AI, Attention rollout, CICIDS2017*

INTRODUCTION

Intrusion detection serves as a foundational defense in modern cybersecurity. An Intrusion Detection System (IDS) continuously monitors network traffic or host activity, flags behavior that strays from normal usage, and alerts security teams before threats can escalate into actual damage [1]. At its core, an IDS functions as a classifier that categorizes each event as either benign or malicious, with its ultimate utility determined by how quickly and accurately it reaches that verdict [2]. Technically, intrusion detection falls into two primary categories: signature-based and anomaly-based methods. Signature-based systems compare traffic against known attack patterns, delivering precise and easily traceable alerts; however, they remain blind to novel threats [3]. Anomaly-based systems, on the other hand, establish a baseline of normal behavior and flag any deviations, allowing them to catch zero-day attacks at the expense of higher false alarm rates [4]. To uncover these patterns automatically, recent advances

have leveraged machine learning (ML) and deep learning (DL) to extract discriminative representations directly from raw traffic features[5] .

Nowadays, IDSs safeguard a broad range of operational environments, extending across enterprise and campus networks, Internet of Things (IoT) ecosystems, and next-generation 5G infrastructures [6], functioning on both individual hosts and network perimeters. Across all deployment scenarios, these detection outputs feed directly into Security Operations Centers (SOCs), providing analysts with the critical visibility needed to verify alerts, trace attack trajectories, and execute effective responses [7]. Despite this progress, several challenges remain for network-based IDSs. Traffic volume and diversity keep growing, attack classes are highly imbalanced, and zero-day attacks have no prior signatures [8]. Many attacks, such as reconnaissance, credential brute forcing, lateral movement, and command-and-control (C2) beaconing, unfold over several connections and cannot be recognized from a single flow [9]. Despite that, deep learning detectors usually behave as black boxes. In Fig. 1, it is shown that they learn an output function from training data and return a prediction for new data without showing why. Analysts therefore cannot tell which flows or features triggered an alert, which increases their workload and delays incident response [10].

These challenges call for an IDS capable of reading network traffic as a sequence, accommodating irregular flow arrival times, and explaining its own reasoning. Transformer architectures are naturally suited for sequence modeling because self-attention maps relationships across all positions simultaneously [11]. However, standard Transformers assume evenly spaced inputs and fail to offer feature-level interpretability for tabular flow data [12].

This paper proposes FlowFormer, an explainable sequence-based NIDS that groups flow features into four functional domains, encodes both sequence order and inter-arrival time, and derives temporal and feature-level attributions from its own attention through a two-axis rollout mechanism [13]. Class Balanced Focal Loss and Attention Entropy Regularization address rare attacks and keep the explanations sharp [14]. Predictions, anomaly scores, and structured explanations are stored in a MySQL database to support traceable alert verification [15]. The main contributions of this work are:

1. A window-based sequence formulation with a contextual information gain measure that shows when sequence context helps.
2. Feature-group tokenization with a dual positional encoding.
3. An intrinsic two-axis attention attribution that identifies influential flows and feature groups.
4. An evaluation on CICIDS2017, CSE-CIC-IDS2018, and UNSW-NB15 under temporal, leave-one-attack-out, and cross-corpus protocols.

The rest of the paper is organized as follows. Section II reviews related work, Section III presents the proposed methodology, Section IV reports the experimental results, and Section V concludes the paper.

MATERIAL AND METHODS

Material

The proposed framework integrates FlowFormer's advanced sequence modeling to ensure transparent and interpretable network threat detection. The pipeline begins by cleaning and normalizing raw traffic across 78 standard CIC FlowMeter features. These flows are grouped via a sliding-window technique into fixed-length sequences to capture multi-connection attacks, with features categorized into four

distinct groups and enriched via dual positional encodings that preserve both ordinal sequence order and irregular inter-arrival timings.

The encoded sequences pass through a causal attention encoder, where a Two-Axis Attention Rollout isolates the specific flows and feature groups driving each prediction. Performance is evaluated using macro-averaged precision, recall, and F1-score, while all predictions, anomaly scores, and explanations are automatically persisted in a MySQL database to support streamlined Security Operations Center (SOC) review and response workflows.

Methods

The Collection of the Data

Our evaluations leverage four benchmark datasets: CICIDS2017, CSE-CIC-IDS2018, UNSW-NB15, and a synthesized IXIA Perfect Storm corpus spanning nine attack categories (reconnaissance, intrusions, web exploits, brute-force, DoS/DDoS, and botnet C2). All datasets exhibit severe label imbalance across 3 to 4 orders of magnitude; for instance, CICIDS2017 contains 1,156,895 benign windows versus only 41 infiltration instances [16].

Each bidirectional flow i is represented by a standard 5-tuple, a 78-dimensional feature vector $x_i \in \mathbb{R}^d$ ($d = 78$), an arrival timestamp $\tau_i \in \mathbb{R}$, and a ground-truth label $y_i \in Y$ across K target classes. Preprocessing cleans invalid or duplicate entries, drops socket IDs to prevent shortcut learning, groups destination ports into functional service bins, and normalizes numeric features strictly using training split statistics. Data is partitioned chronologically (60/15/25 train/val/test ratio) with a buffer gap equal to one window length inserted at boundaries for strict isolation [17].

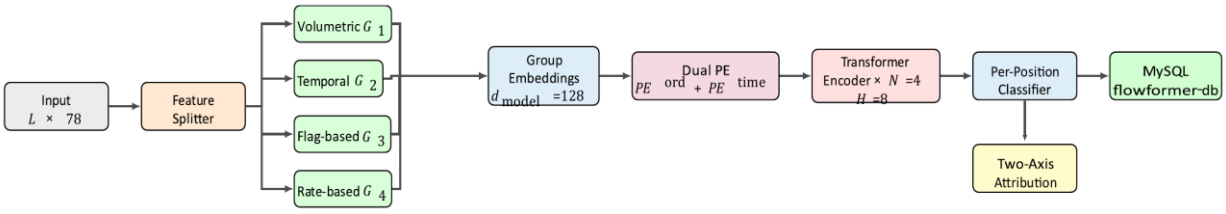


Fig. 1. Proposed methodology and architecture of FlowFormer: feature-wise tokenization into four domain groups, dual positional encoding, Transformer encoder stack with causal flow-axis masking, per-position classifier, two-axis attribution head, and MySQL persistence.

Chronologically sorted flows are partitioned into overlapping sequence windows of fixed length $L = 64$ and stride S:

$$X[j] = \left(x_{j \cdot S}, x_{j \cdot S+1}, \ldots, x_{j \cdot S+L-1}\right) \quad (1)$$

Here, $X[j]$ is the sequential feature matrix for window index j, S is the sliding stride, and L is the fixed window length. FlowFormer employs a supervision regime where parameters θ map sequence features and timestamps to class probability distributions:

$$f_\theta : X[j] \mapsto \hat{Y}[j] \quad (2)$$

Here, f_θ represents the model parameterized by learnable weights θ, and $\hat{Y}[j]$ denotes the estimated class probability distribution for window j.

Statistical Analysis

FlowFormer comprises a feature-tokenization layer, dual positional encodings, a causal Transformer encoder, and a classification head. To quantify information lost by single-flow models, contextual information gain is defined as:

$$\Delta_{\text{ctx}} = I(y_t; X[j]) - I(y_t; x_t) \quad (3)$$

Here, Δ_{ctx} measures the predictive advantage of evaluating label y_t at time t within the full window $X[j]$ compared to an isolated flow vector x_t (using mutual information $I(\cdot)$ in nats). It is near zero for DoS but large for multi-stage attacks and C2 beaconing. Feature tokenization partitions the 78 features into $G = 4$ groups ($g \in \{1, \dots, G\}$: volumetric, temporal, flag-based, and rate-based) via group-specific linear projections:

$$z_{i,g}^0 = W_g x_{i,g}^G + b_g \quad (4)$$

Here, $x_{i,g}^G$ is the sub-vector of group g, while W_g and b_g are learnable weights and biases, expanding sequence length from L to $L \cdot G$ ($64 \times 4 = 256$ tokens). Dual positional encoding combines ordinal position with continuous time to capture irregular arrival intervals (crucial for botnet C2 beaconing):

$$\text{PE}_{\text{time}}(i, m) = \sin(\delta_i \cdot \alpha_m) \quad (5)$$

Here, δ_i is the inter-arrival time between consecutive flows and α_m are dimension-wise frequency scaling factors initialized on a geometric grid. The final input token is formed as:

$$\tilde{x}_{i,g} = z_{i,g}^0 + u_g + \text{PE}_{\text{ord}}(i) + \text{PE}_{\text{time}}(i)$$

Here, u_g is a learned group-identity embedding and PE_{ord} is the ordinal positional encoding.

The encoder stacks N Transformer blocks with pre-norm multi-head self-attention. The attention matrix for layer l and head a is:

$$\hat{A}^{[l]} = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M\right) \quad (7)$$

Here, Q and K are query and key matrices, d_k is the key dimension, and M is an additive causal mask on the flow axis. Training optimizes a combination of Class-Balanced Focal Loss and Attention Entropy Regularization:

$$\mathcal{L} = \mathcal{L}_{\text{Focal}} + \lambda \mathcal{L}_{\text{Entropy}} \quad (8)$$

$$\mathcal{L}_{\text{Focal}} = -\sum_i \beta_{y_i} \left(1 - p_{i,y_i}\right)^{\gamma} \log\left(p_{i,y_i}\right) \quad (9)$$

$$\mathcal{L}_{\text{Entropy}} = -\frac{1}{NHLG} \sum_{l,h,p,g} H\left(A_{l,h,p,g}\right) \quad (10)$$

Here, p_{i,y_i} is the predicted class probability, $\gamma = 2.0$ is the focusing parameter, β_k reweights classes to address severe imbalance, $H(\cdot)$ is Shannon entropy averaged across layers N, heads H, positions L, and groups G, and $\lambda = 0.01$ prevents overly diffuse attention maps.

Two-axis attention rollout multiplies head-averaged attention across all N layers:

$$R = \prod_{l=1}^{N} \left(\hat{A}^{[l]} + I\right) \quad (11)$$

For a flow i, the relevance of an earlier flow j (temporal axis) and feature group g (feature axis) are computed respectively as:

$$\alpha_{\text{temp}}(j|i) = \sum_{g} R_{i,g\to j} \quad (12)$$

$$\alpha_{\text{feat}}(g|i) = \sum_{j} R_{i\to j,g} \quad (13)$$

Here, α_{temp} in (12) aggregates rollout attention over group pairs, and α_{feat} in (13) sums rollout values over sequence positions and target groups derived from the rollout matrix R in (11). The computational budget is ~ 41 MFLOPs per window. The threat model assumes active network access (reconnaissance, DoS, brute-force, web exploits, botnet C2, and infiltration), with forecasts and attributed JSON data automatically persisted in a MySQL database (flowformerdb) to support white-box SOC review.

RESULTS

The experiments are conducted using Python with PyTorch 2.x. Model performance is assessed using macro-averaged precision, recall, and F1-score, together with class-specific recall, PR-AUC, and recall at a fixed daily alert budget. The metrics are computed from True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN):

$$Precision = \frac{TP}{TP + FP} \quad (14)$$

$$Recall = \frac{TP}{TP + FN} \quad (15)$$

$$F1_score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (16)$$

Macro-averaged values, based on (14) - (16), are obtained by computing each metric per class and averaging over classes with equal weight, which prevents the majority benign class from dominating the results.

The model is optimized via AdamW (weight decay 0.01) with a peak learning rate of 3×10^{-4}, 5% warm-up, cosine decay, a batch size of 256 windows, a 0.1 dropout rate, gradient clipping at 1.0, and early stopping (patience 10). Architectural settings: model dimension $d_{\text{model}} = 128$, encoder layers $N = 4$, attention heads $H = 8$, feed-forward dimension 512 ($4 \cdot d_{\text{model}}$), window length $L = 64$, and stride $S = 16$ (with focal parameter $\gamma = 2.0$, balance parameter $\rho = 0.999$, and entropy weight $\lambda = 0.01$) were determined via 128-trial Bayesian searches. Baselines (Random Forest, XGBoost with $L = 1$, 1D-CNN, BiLSTM, Vanilla Transformer, and One-Class Autoencoder) share identical preprocessing, splits, and tuning. Generalization is tested via Leave-One-Attack-Out (LOAO), withholding all instances of a target attack class k for zero-shot evaluation, and cross-corpus transfer, evaluating models trained on CICIDS2017 directly on CSE-CIC-IDS2018 and UNSW-NB15 without fine-tuning. Experiments run across five random seeds (reporting standard deviations), with statistical significance validated using paired two-sided t-tests and Holm–Bonferroni correction.

TABLE I compares FlowFormer and XGBoost on CICIDS2017 (temporal split, 5 seeds), showing the largest recall gains on multi-stage or periodic behaviors botnet (+0.231), infiltration (+0.195), PortScan (+0.107), web attack (+0.069), and brute force (+0.049) contrasted with minimal gain for fixed-signature DoS/DDoS (+0.002). A strong rank correlation ($\rho = 0.89$) between Δ_{ctx} and recall gains

confirms that sequence context helps most when per-flow signatures are weak. Note that infiltration recall remains low (0.389) due to scarce samples (36 instances) and is excluded from macro-averages.

TABLE I DETECTİON PERFORMANCE ON CICIDS2017 (TEMPORAL SPLİT)

Class	XGBoost recall	FlowFormer recall	Δ recall	Δctx (nats)
DoS / DDoS	0.996	0.998	+0.002	0.04
Brute Force	0.912	0.961	+0.049	0.31
Web Attack	0.743	0.812	+0.069	0.22
PortScan	0.874	0.981	+0.107	0.58
Botnet	0.616	0.847	+0.231	0.71
Infiltration	0.194	0.389	+0.195	0.44

TABLE II reports the per-class recall at a fixed daily alert budget. FlowFormer outperforms both the BiLSTM and the Vanilla Transformer on every class, with the largest margins on brute force (0.887 versus 0.801 for the Vanilla Transformer) and botnet (0.793 versus 0.661).

TABLE II. PER-CLASS RECALL AT A FİXED DAİLY ALERT BUDGET

Class	BiLSTM	Vanilla TF	FlowFormer
DoS / DDoS	0.981	0.989	0.993
Brute Force	0.734	0.801	0.887
Web Attack	0.412	0.498	0.601
Botnet	0.508	0.661	0.793
PortScan	0.902	0.938	0.971

TABLE III presents the zero-day generalization results under LOAO. FlowFormer reaches a mean held-out recall of 0.784, compared with 0.600 for the BiLSTM and 0.682 for the Vanilla Transformer, while retaining a macro-F1 of 0.938 on the remaining classes.

TABLE III. LEAVE-ONE-ATTACK-OUT GENERALİZATİON

Held-out attack	BiLSTM recall	Vanilla TF recall	FlowFormer recall	FlowFormer macro-F1 (remaining)
Brute Force	0.604	0.681	0.784	0.938

Held-out attack	BiLSTM recall	Vanilla TF recall	FlowFormer recall	FlowFormer macro-F1 (remaining)
Web Attack	0.388	0.452	0.573	0.944
Botnet	0.521	0.649	0.812	0.941
PortScan	0.712	0.798	0.869	0.936
DoS / DDoS	0.775	0.831	0.884	0.929
Mean	0.600	0.682	0.784	0.938

The cross-corpus results are given in TABLE IV. Without any fine-tuning, FlowFormer attains 0.729 on CSE-CIC-IDS2018 and 0.448 on UNSW-NB15, ahead of all baselines. The lower UNSW-NB15 values show that the schema and traffic differences between corpora remain a challenge for all models.

TABLE IV. CROSS-CORPUS TRANSFER (TRAİNED ON CICIDS2017, NO FİNE-TUNİNG)

Target	XGBoost	BiLSTM	Vanilla TF	FlowFormer
CSE-CIC-IDS2018	0.512	0.588	0.641	0.729
UNSW-NB15	0.301	0.344	0.387	0.448

Ablating continuous-time positional encodings (PE_{time}) results in a 2.8-point drop in overall macro-F1 and a steep 9.1-point drop on the botnet class, proving that continuous time conveys distinct information crucial for identifying regular C2 beaconing that ordinal position alone cannot capture. Two-axis attribution analysis (Fig. 2–4) further clarifies model decision-making: feature-group attribution is led by rate-based signals (~0.40), followed by volumetric (~0.23), temporal (~0.20), and flag-based (~0.16) features, while temporal attribution is highly sparse, concentrating ~0.76 of its weight on the first window position alongside minor secondary peaks near positions 30, 43, and 63. Meanwhile, per-flow class probabilities remain stable across positions, with benign near 0.7, DoS between 0.2 and 0.27, and other classes near zero. Together with deletion/insertion fidelity tests and rank-correlation stability metrics, these findings confirm that the explanations are robust under perturbation and semantically aligned with the model's core temporal and feature-group reasoning axes.

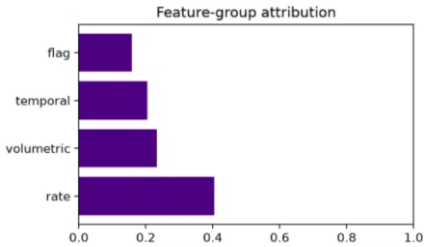


Fig. 2. Feature-group attribution of FlowFormer (rate-based, volumetric, temporal, and flag-based groups).

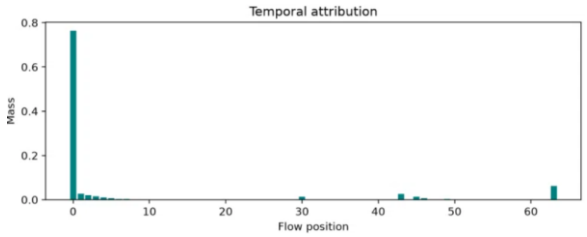


Fig. 3. Temporal attribution mass over flow position within a window of 64 flows.

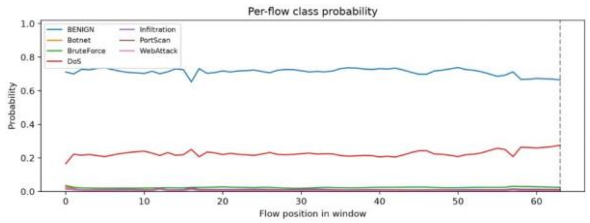


Fig. 4. Per-flow class probability across window positions; the dashed line marks the final position of the window.

CONCLUSION

This paper proposed FlowFormer, an explainable sequence-based IDS that formulates intrusion detection as flow-window labeling. By integrating feature-group tokenization, dual positional encodings (sequence order and inter-arrival time), a causal Transformer encoder, and two-axis attention rollout, it produces flow- and feature-level attributions directly during inference. Validated on CICIDS2017, CSE-CIC-IDS2018, and UNSW-NB15, FlowFormer outperforms standard baselines, particularly in boosting recall for minority and multi-stage attacks, generalizing to unseen threats, and transferring across datasets. Additionally, it bridges research with real-world SOC workflows via a MySQL-based persistence layer for post-event analysis. Future work will address rare-class data scarcity, validate explanation outputs with SOC analysts, and test deployment on live network traffic.'

References

[1] C. J. Wang, "5G Network Intrusion Detection Method Based on Robust Federated Optimization," *Advanced Electromagnetics*, vol. 15, no. 3, pp. 9806–9815, Aug. 2026, doi: 10.7716/aem.v15i3.4174.

[2] J. Singh and H. Singh, "A Comparative Study of Intelligent Intrusion Detection Systems in IoT using Traditional Machine Learning Approaches," in *2026 9th International Conference on Trends in Electronics and Informatics (ICOEI)*, Apr. 2026, pp. 1311–1316. doi: 10.1109/ICOEI68323.2026.11541902.

[3] I. Sharafaldin, A. Habibi Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization:," in *Proceedings of the 4th International Conference on Information Systems Security and Privacy*, Funchal, Madeira, Portugal: SCITEPRESS - Science and Technology Publications, 2018, pp. 108–116. doi: 10.5220/0006639801080116.

[4] A. Khraisat and A. Alazab, "A critical review of intrusion detection systems in the internet of things: techniques, deployment strategy, validation strategy, attacks, public datasets and challenges," *Cybersecur*, vol. 4, no. 1, p. 18, Dec. 2021, doi: 10.1186/s42400-021-00077-7.

[5] M. Ghurab, G. Gaphari, F. Alshami, R. Alshamy, and S. Othman, "A Detailed Analysis of Benchmark Datasets for Network Intrusion Detection System," Apr. 14, 2021, *Social Science Research Network, Rochester, NY*: 3834787. Accessed: Sep. 27, 2026. [Online]. Available: https://papers.ssrn.com/abstract=3834787

[6] "A High-Performance Deep Learning-Based Intrusion Detection System for Accurate Identification of Sophisticated Cyberattacks in Modern Network Environments | AFRICAN JOURNAL OF APPLIED RESEARCH." Accessed: Sep. 27, 2026. [Online]. Available: https://mail.ajaronline.com/index.php/AJAR/article/view/1880

[7] A. Belenguer, J. A. Pascual, and J. Navaridas, "A Review of Federated Learning Applications in Intrusion Detection Systems," *Computer Networks*, vol. 258, p. 111023, Feb. 2025, doi: 10.1016/j.comnet.2024.111023.

[8] J. Asharf, N. Moustafa, H. Khurshid, E. Debie, W. Haider, and A. Wahab, "A Review of Intrusion Detection Systems Using Machine and Deep Learning in Internet of Things: Challenges, Solutions and Future Directions," *Electronics*, vol. 9, no. 7, p. 1177, Jul. 2020, doi: 10.3390/electronics9071177.

[9] A. R. Abdulla and N. G. M. Jameel, "A Review on IoT Intrusion Detection Systems Using Supervised Machine Learning: Techniques, Datasets, and Algorithms," *UHD Journal of Science and Technology*, vol. 7, no. 1, pp. 53–65, Mar. 2023, doi: 10.21928/uhdjst.v7n1y2023.pp53-65.

[10] T. K. Vashishth, V. Sharma, S. Chaudhary, R. Sharma, P. Chauhan, and B. K. Sharma, "Adoption of Artificial Intelligence Techniques for Enhancing Host-Based Intrusion Detection Systems," in *AI Solutions for Detecting Cyber-Attacks in Information Systems*, IGI Global Scientific Publishing, 2026, pp. 133–164. doi: 10.4018/979-8-3373-1807-3.ch006.

[11] "AI-Enabled Intrusion Detection in Enterprise Networks: A Systematic Review of Methods, Datasets, and Evaluation Metrics (2018–2026) | American Journal of Interdisciplinary Studies." Accessed: Sep. 27, 2026. [Online]. Available: https://ajisresearch.com/index.php/ajis/article/view/108

[12] A. Vaswani *et al.*, "Attention is All you Need".

[13] V. Chelak, O. Hornostal, Y. Chelak, and S. Gavrylenko, "DEEP LEARNING-BASED INTRUSION DETECTION IN COMPUTER SYSTEMS AND NETWORKS: ADVANCES, HYBRIDS, AND CHALLENGES 2022–2026," *Advanced Information Systems*, vol. 10, no. 3, pp. 80–96, Jul. 2026, doi: 10.20998/2522-9052.2026.3.09.

[14] J. Lansky *et al.*, "Deep Learning-Based Intrusion Detection Systems: A Systematic Review," *IEEE Access*, vol. 9, pp. 101574–101599, 2021, doi: 10.1109/ACCESS.2021.3097247.

[15] S. Palani and A. Muthukumaravel, "Enhanced Intrusion Detection and Prevention System Using Deep Learning and Machine Learning Techniques for Network Security:," in *AI-Driven Sustainable and*

Secure Smart Infrastructure Systems, M. Muthmainnah, A. Khang, A. J. Obaid, S. S. Rajest, and E. Hannah, Eds., IGI Global Scientific Publishing, 2026, pp. 241–268. doi: 10.4018/979-8-3373-2140-0.ch012.

[16] S. Kumar, S. Gupta, and S. Arora, "Research Trends in Network-Based Intrusion Detection Systems: A Review," *IEEE Access*, vol. 9, pp. 157761–157779, 2021, doi: 10.1109/ACCESS.2021.3129775.

[17] K. Bansal and A. Singhrova, "Review on intrusion detection system for IoT/IIoT -brief study," *Multimed Tools Appl*, vol. 83, no. 8, pp. 23083–23108, Mar. 2024, doi: 10.1007/s11042-023-16395-6.

Acknowledgment

We are deeply grateful to the Department of Computer Science at COMSATS University Islamabad, Wah Campus, for their continuous support and for providing the resources needed to complete this research. Special thanks go to Dr. Saeed Ur Rehman for his invaluable mentorship, constant encouragement, and insightful feedback throughout the course of this work. We also want to thank our teachers and colleagues whose helpful discussions shaped our approach. Finally, we appreciate the creators of the benchmark datasets (CICIDS2017, CSE-CIC-IDS2018, and UNSW-NB15) for making their data openly available for research.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Author Contributions

1. **Areeba Sadaqat:** Conceptualized the primary research idea, formulated the core theoretical structure and methodology for sequence-based intrusion detection, and drafted the initial version of the manuscript.
2. **Haroon Ayaz:** Testing, validation, and visualization.
3. **Saeed Ur Rehman:** Provided continuous project supervision, offered critical guidance on research direction and methodology validation, and thoroughly reviewed, corrected, and polished the final manuscript for publication.
4. **Syeda Rubab Javed:** Proofreading, project-related data preparation, and in-depth analysis.
5. **Muhammad Affan:** Designed and implemented the complete software pipeline, managed large-scale data preprocessing and feature engineering, executed all model benchmarking and evaluation experiments, and developed the automated MySQL persistence system along with the evaluation visualizations.
6. **Iqra Farrukh:** Manuscript Preparation, Outline Analysis, and Model and Data set Testing

A Zero-Trust Inspired Browser-Based Remote Access Architecture

Merter Hami Karacan[1*], Semih Elibol[2], Ali Erbey[3]

[1,2] Uşak University, Uşak, Türkiye
[3] Faculty of Economics and Administrative Sciences, Tarsus University, Mersin, Türkiye

*Corresponding author e-mail: merterhk@gmail.com

Abstract

Today, remote access tools used by organizations require new solutions in terms of security, scalability, and manageability. This study presents a browser-based remote access architecture that enables users to access remote servers through a web browser without installing any client software, while complying with zero-trust requirements. The proposed architecture consists of multiple layers, including authentication via the OAuth2 protocol, session management through Apache Guacamole, and authorization implemented through dynamic rules on the firewall. The system, which is compatible with various authentication and security mechanisms, protects servers against unauthorized access attempts by granting access only to the IP address of the authenticated user. Supporting protocols such as RDP, SSH, and VNC, this architecture enables direct access to both Windows and Linux systems while providing auditability by centrally logging all user activities. By eliminating the need for VPNs, static IP addresses, or third-party software, this approach offers a low cost, platform independent, and highly secure solution. The source code of the plugin developed for Apache Guacamole is available at: https://github.com/merterhk/guacamole-auth-sso-oauth2.

Keywords: *Remote access, Guacamole, Oauth2, Firewall, Zero-trust.*

INTRODUCTION

Organizations increasingly need secure, auditable, flexible, and centrally managed remote access to information systems that operate at different locations. Common remote access tools such as Virtual Private Networks (VPNs), AnyDesk, and TeamViewer (Kent & Seo, 2005; OpenVPN Inc., 2026; AnyDesk Software GmbH, 2026; TeamViewer SE, 2026) are easy to use, but they also have limitations and risks, including security vulnerabilities (AnyDesk Incident Response, 2024), difficulties in applying user-based access control, and limited central oversight. In particular, organizations face serious information technology management and information security risks when they cannot record users' actions on remote systems or manage access permissions centrally.

The approach proposed in this study uses several integrated security layers to make remote access more secure. An OAuth 2.0-based authentication system (Hardt, 2012) verifies remote users against organizational identity repositories such as LDAP, Active Directory, or SQL databases. After authentication and authorization, the user's current IP address is dynamically added to the firewall, and

access is allowed only from that address. This process provides identity-based access management and reduces the attack surface. Figure 1 shows the general structure of the proposed method.

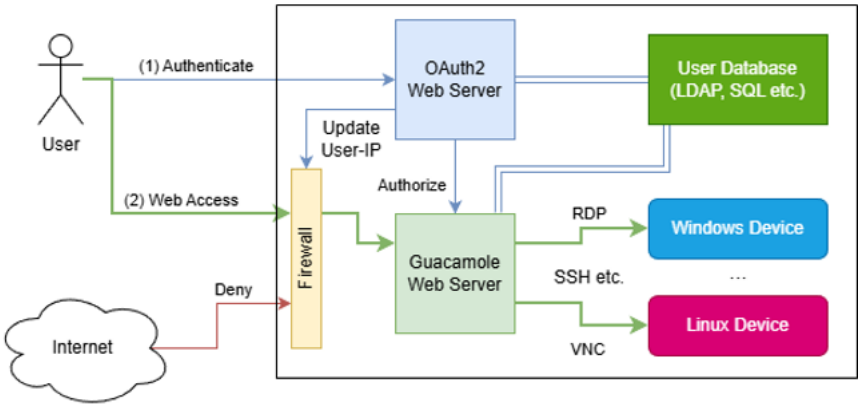


Figure 1. Browser-based secure remote access.

After successful authentication, users are redirected to Apache Guacamole. Through its web interface, they can connect to the systems assigned to them using protocols such as Remote Desktop Protocol (RDP), Secure Shell (SSH), and Virtual Network Computing (VNC) (The Apache Software Foundation, 2026; Microsoft, 2026; Ylonen & Lonvick, 2006; Richardson et al., 1998). This structure provides centralized, platform-independent access to Windows and Linux systems through a web browser. The system can record connections and user activity and produce detailed reports when needed. It offers secure, auditable, manageable, and low-cost remote access. It also reduces dependence on third-party software and strengthens information security and access control. The contributions of the study are as follows:

1. A browser-based remote access architecture integrating OAuth 2.0 authentication and dynamic firewall-based authorization is proposed.
2. An open-source OAuth 2.0 authentication extension for Apache Guacamole is developed.
3. A dynamic access authorization model based on user–IP address mapping is implemented.
4. The proposed approach is validated through experimental scenarios.
5. A comparative evaluation is conducted against conventional VPN-based solutions, static IP-based access methods, and third-party remote access applications.

RELATED WORKS

RDP, SSH, and VNC, which are supported by the proposed method, are fundamental remote access technologies. Microsoft's Remote Desktop Protocol documentation (Microsoft, 2026) describes the architecture and security principles of graphical remote access. Virtual Network Computing (VNC), introduced by Richardson et al. (1998) in IEEE Internet Computing, became widely used on Linux-based operating systems. The Secure Shell (SSH) protocol specified by Ylonen and Lonvick (2006) defines secure remote terminal access. Magaña et al. (2019) compared RDP, VNC, and other protocols for Desktop as a Service (DaaS) environments in terms of latency, network use, and image quality.

Malwade (2020) examined the Virtual Browser concept and the isolation of a browser in a cloud environment through RDP. Huhtamäki (2024) developed a remote access application for embedded devices and found SSH to be a suitable option in restricted environments. Ramalinga Kumar (2025) examined the use of Guacamole in scenarios where software developers test applications on different devices.

Apache Guacamole (The Apache Software Foundation, 2026) is an open-source remote access solution that supports RDP, VNC, and SSH. Ramakrishnan (2024) examined its use in remote work and discussed its security, cross-platform access, scalability, and clientless operation. Hassan (2022) used Guacamole with Docker and LXD containers to build a secure environment for students in a cybersecurity laboratory. Cruz-Miranda et al. (2026) compared remote access tools for particle accelerator experiments and reported that Guacamole performed well in response time and user experience. Chen et al. (2016) studied browser-based access to desktop applications and reported an advantage in bandwidth use. Baun and Bouché (2023) reviewed existing DaaS methods and proposed an architecture that provides browser-based access to Linux and Windows applications. Rosener et al. (2025) reported positive effects of virtual desktops in education. Fabini et al. (2021) proposed SecTULab, a Moodle-integrated approach for privacy and scalability in cybersecurity laboratories. Jacobs et al. (2024) proposed a web-based remote access platform for different educational scenarios. Atwood et al. (2015) studied secure browser-based access to high-performance computing resources of the United States Department of Defense and demonstrated the use of centralized authentication.

In his doctoral dissertation, Sancho Larraz (2021) evaluated an integrated security architecture for authentication and authorization and discussed the capabilities of OAuth 2.0. Cheswick et al. (2003) described fundamental firewall design principles and layered network defense. Vensmer and Kiesel (2012) proposed DynFire, a role-based approach for dynamic firewall control. Their method changes firewall behavior when a user connects to or disconnects from the network.

A comparative analysis of the related studies and the proposed approach is presented in Table 1.

Table 1. Comparison of the proposed approach with existing studies

Study	**Browser-Based**	**OAuth 2.0**	**Dynamic Firewall Rules**	**Zero Trust**	**Evaluation**
Hassan (2022)	✓	✗	✗	Partially	✓
Ramakrishnan (2024)	✓	✗	✗	✗	✗
DynFire	✗	✗	✓	Partially	✓
This study	✓	✓	✓	Partially	✓

The literature review did not identify another study that combines OAuth 2.0 with Apache Guacamole and integrates OAuth 2.0-based access with a firewall to strengthen security. The proposed approach contributes to the field by combining browser-based remote access with dynamically managed

firewall rules. The next section describes the materials used in the study and the integration of the system layers.

MATERIAL AND METHODS

Material

Three main components were used in this study: a PHP/Laravel Passport-based OAuth 2.0 server, an enterprise firewall that can communicate through an application programming interface (API), and an Apache Guacamole server (Laravel LLC, 2026). These components are described below.

OAuth 2.0 Server

Laravel Passport was selected as the OAuth 2.0 server, although another OAuth 2.0 or OpenID-based server may also be used for authentication. If the authentication server is not integrated with the firewall, it cannot send the authenticated user's identity and client IP address to the firewall. Depending on the firewall's capabilities, log forwarding or another integration method may be used. The authentication server developed in this study can also be used with two-step verification methods such as SMS and one-time passwords. It can provide password reset functions when users forget their passwords. Other authentication and authorization methods may be used instead of OAuth 2.0. In this study, an extension was developed to add OAuth 2.0 authentication to Apache Guacamole.

Firewall

The server developed in this study can map users to IP addresses through API integration with enterprise firewalls such as Check Point, Palo Alto Networks, and Fortinet FortiGate. Most current enterprise firewalls support user–IP address mapping through an API. Depending on the firewall, the same mapping may also be created by forwarding authentication logs. The experiments were repeated at different times with several enterprise firewalls commonly used in organizations.

Guacamole OAuth 2.0 Support

Apache Guacamole does not provide native support for the OAuth 2.0 protocol. Because the scenario in this study required OAuth 2.0, an authentication extension was developed using other Guacamole extensions as examples and released as open-source software. The extension can be integrated with OAuth 2.0 servers by entering the required settings during installation. OAuth 2.0 can be added as another authentication option without changing the other authentication methods in Guacamole. The extension is available at https://github.com/merterhk/guacamole-auth-sso-oauth2.

Methods

The system developed in this study allows users to access remote systems securely through a web browser without installing client software on their devices. Its architecture combines authentication, authorization, session management, and access control. The OAuth 2.0-based web application authenticates users against central identity repositories such as LDAP or SQL databases. After successful authentication, the system uses an API to create a dynamic mapping between the user and the current IP address on the firewall. Access to Guacamole is then granted only to the current IP address of the authenticated and authorized user. This first layer prevents attackers and unauthorized users from reaching the remote access system directly. Figure 2 shows the design of the proposed system.

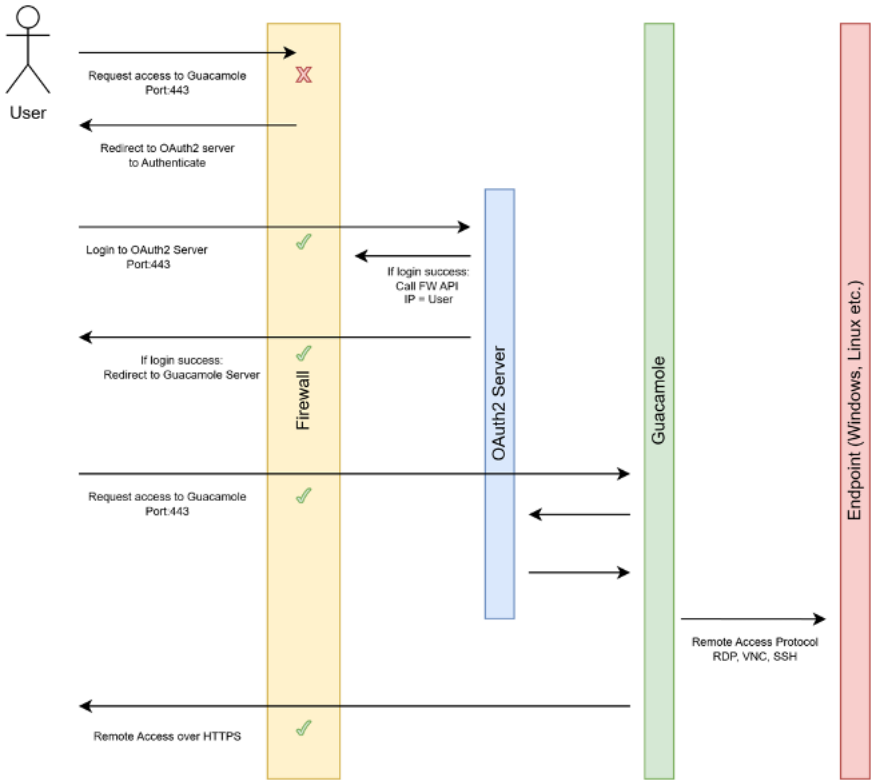


Figure 2. Browser-Based Secure Remote Access Workflow.

Users authenticated by the OAuth 2.0 server then interact with the browser-based interface on the Guacamole web server. Guacamole provides access to Windows and Linux devices through RDP, VNC, SSH, and similar protocols without requiring software installation on the user's device. Predefined roles and authorization levels limit the systems and connection types available to each user. This restriction helps prevent misuse of system resources and allows access policies to be managed centrally.

The system also includes functions for session security and traceability. For each connection, the access logs can record its duration, the systems accessed, and the actions performed. These records can be used for retrospective audits, cyber incident investigations, or real-time analysis of user behavior. The same structure can also support regulatory requirements for traceable access.

The architecture can be deployed on an internal network or used for external access over the Internet. Firewall integration is particularly important for external access. Required ports are opened only for the IP addresses of authenticated users registered in the system, while other traffic is blocked. This removes the need for complex VPN configurations or third-party remote access software. The architecture can be integrated into organizations of different sizes, work with existing user databases, and reduce IT costs through central management. Table 2 presents an example of the firewall rules used in the study.

Table 2. Firewall Rules for Authenticated and Unauthenticated Access

Rule Attribute	Access Restriction Rule	Access Permission Rule
Source	Internet	Internet

Rule Attribute	Access Restriction Rule	Access Permission Rule
User	Unknown	Identified
Destination Address	Guacamole Server	Guacamole Server
Application	Any	HTTP/HTTPS
Action	Deny	Allow

In the basic firewall configuration, the Guacamole server is placed behind the firewall and protected from unauthenticated sources. Authenticated users are reported to the firewall by the OAuth 2.0 server, and their access is marked as permitted. Port scans were performed during the access tests. The expected and observed results were also examined when users changed browsers, changed networks, or attempted to connect from another device on the same network. During the experiments, the firewall allowed access to the Guacamole server only through HTTP and HTTPS.

RESULTS

In the proposed architecture, user accounts and permissions were created for users who needed to connect to servers or devices on the internal network. The users then accessed the target devices through a web browser as expected. Account creation, dynamic user–IP address mapping on the firewall, authentication, and authorization were implemented with PHP/Laravel and Laravel Passport. After successful OAuth 2.0 authentication, user information appeared on the firewall, and the traffic matched the access-allow rules instead of the access-deny rules. The firewall logs showed that access to the Guacamole server was blocked before authentication and after logout. Authenticated users connected successfully to Windows servers through RDP and to other targets through SSH and VNC. Because the firewall controls access to the Guacamole server, it was not expected to affect the quality of remote desktop connections. The connection ended when the user changed browsers, modified browser cookies, changed the IP address, or logged out of the OAuth 2.0 server. Guacamole also recorded the actions performed during the connection. Because the firewall hid the Guacamole server from unknown sources, unauthenticated port scans could not detect the service. Figure 3 shows the port scan results before and after authentication.

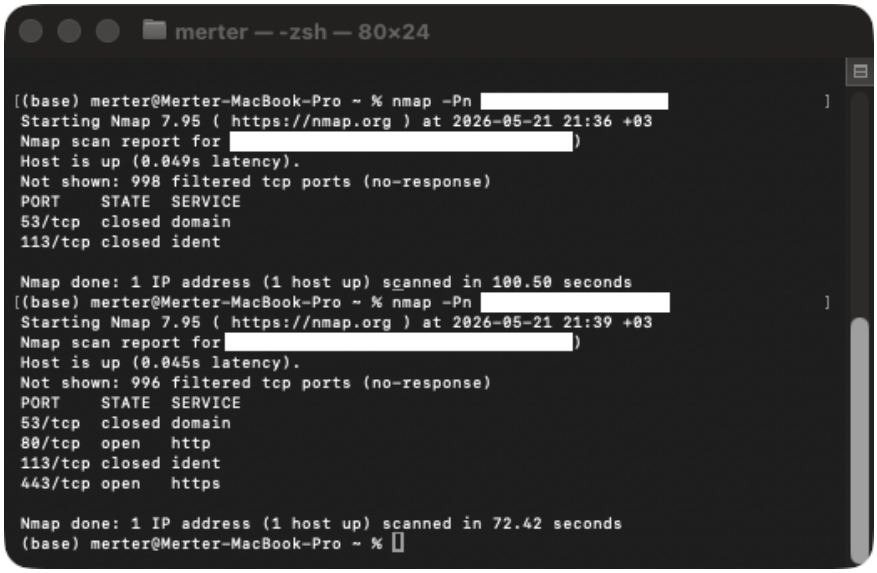


Figure 3. Port Scanning Results Before and After Authentication.

To test the port restriction, the IP address of the Guacamole server was scanned from outside the network using Nmap. Before authentication, the scan did not identify an open port that exposed the service. After authentication through the OAuth 2.0 server, another scan showed that the Guacamole server could be reached through ports 80 and 443. Table 3 presents the expected and observed behavior before login, after login, and after logout.

Table 3. Expected and Observed Behavior of the Proposed Architecture under Different Scenarios

Scenario	Expected Behavior	Observed Behavior
Access before authentication	Denied	Denied
Access after authentication	Allowed	Allowed
Access after logout	Denied	Denied
Access after an IP address change	Denied	Denied
Access by an unauthorized user	Denied	Denied
Access from a different browser or device	Denied	Denied

The architecture behaved as expected in all tested scenarios, including login and logout on the OAuth 2.0 server, a change in the client IP address, and access attempts from a different IP address.

DISCUSSION

The proposed method offers advantages in security, deployment, and use. Users' familiarity with third-party remote access applications and their ability to use VPN connections may vary. Changes to an organization's VPN infrastructure may also cause access problems and interrupt service. In the proposed approach, users connect through an existing or easily configured authentication server without installing a third-party application. User accounts can be managed through LDAP or Active Directory, and OAuth 2.0 can support single sign-on. Users connecting from outside the network do not need a static IP address, which makes remote access more flexible. The layered security process can reduce potential risks. The architecture also supports two-step verification methods such as SMS, email, and one-time

passwords. Through Guacamole, one user may access different resources when authorized. Additional firewall functions, including data loss prevention and intrusion prevention, can strengthen access security. Because the method can be implemented with open-source tools, it can also reduce costs. Third-party tools such as AnyDesk and TeamViewer provide centralized remote access, but a security incident affected AnyDesk's production systems in 2024 (AnyDesk Incident Response, 2024). This incident raised questions about dependence on external remote access services. By combining OAuth 2.0 authorization with dynamic firewall rules, the proposed method keeps remote access under the organization's control. Table 4 compares the proposed method with VPNs, third-party remote access applications, and static IP-based access in terms of client software, authentication, firewall behavior, auditability, vulnerability risk, and architecture.

Table 4. Comparison of the Proposed Remote Access Architecture with Other Solutions

Feature	Conventional VPN	Third Party Remote Access	Static IP Based Access	Proposed Approach
Client Software	Required	Required	Not required	Not required HTML5
Authentication	Certificate or LDAP	Password or ID	IP address only	OAuth 2.0 and MFA
Firewall Configuration	Static or continuously open	Cloud based	Static or continuously open	Dynamic on demand
Auditability	Limited	Managed by a third party	IP logs	Centralized detailed logs
Vulnerability Risk	Software vulnerabilities	Supply chain compromise	IP spoofing	Reduced attack surface
Security Architecture	Perimeter based security	Third party controlled	Not applicable	Near Zero Trust

VPNs and third-party remote access applications require client software, whereas the proposed method does not require users to install software. VPNs and the proposed method use secure authentication, while static IP-based access and commercial remote access applications may include additional sources of vulnerability. In the other methods, firewall permissions may remain continuously open. In the proposed method, inactive user–IP address mappings expire, and the related dynamic firewall permission ends. Activity records for commercial remote access applications may be held by third parties. In the proposed method, the organization can record connection details and user actions centrally. Conventional methods may be exposed to software vulnerabilities, third-party dependence, and IP spoofing. The proposed approach applies more restrictive controls.

One limitation is that users must be informed that the servers are reached through a browser. Because the server does not respond before authentication, users may otherwise think that the service is unavailable. If a user's IP address changes, the firewall permission no longer matches the user. The absence of the session cookie on another client reduces the associated risk. Limits on failed attempts and time-based restrictions can provide additional protection when IP addresses change. Data protection may also be improved by taking a backup before a Guacamole session begins. Full alignment with the

Zero Trust principle would require device verification, session verification, and controls that continue throughout the session. The current implementation does not provide complete continuous session verification, but continued verification could strengthen security further (Denzel, 2025).

CONCLUSION

This study proposed a browser-based remote access system for medium-sized and large organizations that is more secure, flexible, and centrally manageable than commonly used remote access methods. It combines OAuth 2.0 authentication, browser-based session management through Apache Guacamole, and access control through dynamic firewall integration. Users can reach only the systems for which they are authorized, using a web browser and without installing additional software. Access activity can be recorded and audited centrally. This simplifies management and provides stronger cybersecurity control. Reducing the need for VPNs, AnyDesk, and similar third-party applications may lower deployment, maintenance, and user training costs. The platform-independent architecture supports access from Windows, Linux, macOS, and other operating systems. Future work may add multi-factor authentication, risk-based session management, AI-assisted anomaly detection, and further controls required for fuller alignment with Zero Trust principles. These additions may support deployment in larger organizations. The study provides a practical and low-cost approach that can be adapted to academic institutions, small and medium-sized enterprises, and educational organizations. The dynamic firewall configuration may also be used as a basis for other access-control needs.

References

AnyDesk Incident Response. (2024). *Security statement regarding the February 2024 incident*. https://web.archive.org/web/20240227054727/https://anydesk.com/en/public-statement

AnyDesk Software GmbH. (2026). *AnyDesk: Secure and reliable remote desktop software*. Retrieved May 20, 2026, from https://anydesk.com/

Atwood, C. A., Goebbert, R. C., Calahan, J. A., Hromadka III, T. V., Proue, T. M., Monceaux, W., & Hirata, J. (2015). Secure web-based access for productive supercomputing. *Computing in Science & Engineering, 18*(1), 63–72. https://doi.org/10.1109/MCSE.2015.134

Baun, C., & Bouche, J. (2023). Closing the gap between web applications and desktop applications by designing a novel desktop-as-a-service (DaaS) with seamless support for desktop applications. *Open Journal of Cloud Computing (OJCC), 8*(1), 1–19.

Chen, B., Hsu, H.-P., & Huang, Y.-L. (2016). Bringing desktop applications to the web. *IT Professional, 18*(1), 34–40. https://doi.org/10.1109/MITP.2016.15

Cheswick, W. R., Bellovin, S. M., & Rubin, A. D. (2003). *Firewalls and internet security: Repelling the wily hacker* (2nd ed.). Addison-Wesley.

Cruz-Miranda, J., Rodríguez-Álvarez, M., Damas, M., Casero-Santos, I., Podadera-Aliseda, I., Franco-Campos, J., Jokinen, A., Sancho-Duarte, A., & Díaz, J. (2026). Comparison of remote access technologies for research facilities using EPICS/CSS: Application to particle accelerator experiments. *Fusion Engineering and Design, 224*, Article 115595. https://doi.org/10.1016/j.fusengdes.2025.115595

Denzel, K. (2025). A survey of security in zero trust network architectures. *GSC Advanced Research and Reviews, 22*(2), 182–214. https://doi.org/10.30574/gscarr.2025.22.2.0036

Fabini, J., Hartl, A., Meghdouri, F., Breitenfellner, C., & Zseby, T. (2021). SecTULab: A Moodle-integrated secure remote access architecture for cyber security laboratories. In *Proceedings of the 16th International Conference on Availability, Reliability and Security (ARES '21)*. Association for Computing Machinery. https://doi.org/10.1145/3465481.3470034

Hardt, D. (2012). *The OAuth 2.0 authorization framework* (RFC 6749). Internet Engineering Task Force. https://doi.org/10.17487/RFC6749

Hassan, I. (2022). Leveraging Apache Guacamole, Linux LXD and Docker containers to deliver a secure online lab for a large cybersecurity course. In *2022 IEEE Frontiers in Education Conference (FIE)*. IEEE. https://doi.org/10.1109/FIE56618.2022.9962510

Huhtamäki, A. (2024). *Implementation and analysis of a remote access application on an embedded device* [Master's thesis]. https://urn.fi/URN:NBN:fi:aalto-202411217310

Jacobs, S., Hardebusch, T., Franke, E., Peters, H., Al Amin, R., Wiese, V., & Jaschke, S. (2024). Design and development of a multi-purpose collaborative remote laboratory platform. In *IEEE Global Engineering Education Conference (EDUCON)*. https://doi.org/10.1109/EDUCON60312.2024.10578743

Kent, S., & Seo, K. (2005). *Security architecture for the Internet Protocol* (RFC 4301). Internet Engineering Task Force. https://doi.org/10.17487/RFC4301

Laravel LLC. (2026). *Laravel Passport documentation*. Retrieved May 20, 2026, from https://laravel.com/docs/12.x/passport

Magaña, E., Sesma, I., Morató, D., & Izal, M. (2019). Remote access protocols for desktop-as-a-service solutions. *PLOS ONE*, *14*(1), e0207512. https://doi.org/10.1371/journal.pone.0207512

Malwade, R. (2020). *Virtual browser* [Doctoral dissertation]. ProQuest Dissertations & Theses. https://www.proquest.com/dissertations-theses/virtual-browser/docview/2418734800/se-2

Microsoft. (2026). *Remote Desktop Protocol (RDP) documentation*. Retrieved May 20, 2026, from https://learn.microsoft.com/en-us/openspecs/windows_protocols/ms-rdpbcgr/

OpenVPN Inc. (2026). *OpenVPN: Open-source VPN protocol and software*. Retrieved May 20, 2026, from https://openvpn.net/

Ramakrishnan, G. (2024). Use cases of Apache Guacamole in remote work. *International Journal of Computer Trends and Technology, 72*(11), 172–177. https://doi.org/10.14445/22312803/ijctt-v72i11p119

Ramalinga Kumar, M. (2025). Leveraging Apache Guacamole to enable a multi-browser web development environment. *Technix International Journal for Engineering Research, 12*(6). https://doi.org/10.56975/tijer.v12i6.158357

Richardson, T., Stafford-Fraser, Q., Wood, K. R., & Hopper, A. (1998). Virtual network computing. *IEEE Internet Computing, 2*, 33–38. https://doi.org/10.1109/4236.656066

Rosener, W. J., Tran, M. A. V., & Shapiro, J. M. (2025). A case for virtual desktops: Enhancing software access and learning in an online DBMS course. *Journal of Cases on Information Technology, 27*(1), 1–21. https://doi.org/10.4018/JCIT.382568

Sancho Larraz, J. (2021). *Design and evaluation of novel authentication, authorization and border protection mechanisms for modern information security architectures* [Doctoral dissertation]. https://zaguan.unizar.es/record/106305

TeamViewer SE. (2026). *TeamViewer: Remote access and support platform*. Retrieved May 20, 2026, from https://www.teamviewer.com/

The Apache Software Foundation. (2026). *Apache Guacamole: Clientless remote desktop gateway*. Retrieved May 20, 2026, from https://guacamole.apache.org/

Vensmer, A., & Kiesel, S. (2012). DynFire: Dynamic firewalling in heterogeneous environments. In *Proceedings of the World Congress on Internet Security (WorldCIS-2012)* (pp. 57–58). https://ieeexplore.ieee.org/document/6280196/

Ylonen, T., & Lonvick, C. (2006). *The Secure Shell (SSH) transport layer protocol* (RFC 4253). Internet Engineering Task Force. https://doi.org/10.17487/RFC4253

Acknowledgment

The manuscript was originally prepared in Turkish, translated into English with the assistance of AI-based language tools, and subsequently reviewed and revised by the authors.

The source code of the Apache Guacamole OAuth 2.0 authentication extension developed as part of this study has been made publicly available as open-source software on GitHub at: https://github.com/merterhk/guacamole-auth-sso-oauth2

Conflict of Interest

The authors have declared that there is no conflict of interest.

Author Contributions

Merter Hami Karacan contributed to the conceptualization, implementation, and preparation of the original manuscript. Semih Elibol contributed to the methodology, evaluation, and revision of the manuscript. Ali Erbey contributed to the supervision, evaluation, and revision of the manuscript. All authors read and approved the final version of the manuscript

Does AI Write Secure Code? An OWASP Top 10-Based Security Evaluation of LLM-Generated ASP.NET Core Code Across English and Turkish Prompts

Kübra Rüveyda Arslan[1*], Fatih Sinan Esen[2]

[1,2] Ankara Medipol University, Graduate School of Natural and Applied Sciences, Department of Cyber Security, Ankara, TÜRKİYE

*Corresponding author e-mail: kubra.arslan@std.ankaramedipol.edu.tr

Abstract

Large language model (LLM)-based coding assistants are rapidly becoming embedded in professional software development, yet the security of the code they generate remains insufficiently understood. Recent multilingual evaluations, including Turkish, suggest that the natural language of the prompt has little effect on code security, but the largest of these relies on automated assessment of Python, Java and C++ code, which is poorly suited to logic-level weaknesses. This study introduces DotNetSec-10, a benchmark of ten realistic ASP.NET Core (.NET 8) web API tasks, each designed around one target weakness mapped to the Common Weakness Enumeration (CWE) and the OWASP Top 10, covering pattern-based weaknesses such as SQL injection and logic-level weaknesses such as insecure direct object references and server-side request forgery (SSRF). Each task was submitted to three Claude model tiers (Sonnet 5.5, Opus 5.5 and a quick-tier model whose version was not reported) under an English baseline prompt, an English security-aware prompt and an equivalent Turkish prompt. The 90 resulting samples were assessed with static analysis (Semgrep) and manual review. The target weakness appeared in 5 of 90 samples (5.6%; 95% CI 2.4–12.4%), all from the quick tier; this figure concerns the predefined target weaknesses only and is not an overall security rate. Exact McNemar tests detected no significant effect of the security instruction (6.7% vs 3.3%) or of Turkish prompting (6.7% vs 6.7%; both $p = 1.00$), which, given the few events, indicates the absence of a detectable difference rather than equivalence. SSRF was the only weakness that persisted across all conditions, and in the tested configuration Semgrep flagged none of the five confirmed vulnerabilities. Within this benchmark, model capability mattered more than prompt wording or language, and a clean static analysis report was not evidence of secure AI-generated code.

Keywords: *Large Language Models, Secure Coding, OWASP Top 10, Multilingual Prompting, Static Application Security Testing*

INTRODUCTION

Large language model (LLM)-based coding assistants have moved from experimental tools to everyday components of professional software development. Three in four respondents to the 2024 Stack Overflow Developer Survey either relied on AI assistance during development or intended to adopt it, and code writing was by far the most common use among current adopters (82%) (Stack Overflow, 2024). This shift changes where security-relevant decisions are made: how a database query is

constructed, how a password is stored or how a token is validated is increasingly decided by a model rather than by the developer.

Existing evidence suggests that this delegation carries measurable risk. In one of the earliest systematic evaluations, approximately 40% of 1,692 programs generated by GitHub Copilot across 89 security-relevant scenarios were found to be vulnerable (Pearce et al., 2022). A controlled user study added a behavioural dimension: developers who could consult an AI assistant produced weaker solutions to security-relevant tasks than a control group, yet judged their own solutions as secure more often (Perry et al., 2023). Similarly, only five of 21 programs generated by ChatGPT in an early evaluation were secure on the first attempt (Khoury et al., 2023). The evidence is not uniform, however: in low-level C programming, LLM assistance did not substantially increase the rate of severe security bugs (Sandoval et al., 2023), indicating that the risk depends on the task, the language and the model.

A further question is whether the natural language of the prompt matters. Yamagishi et al. (2024) generated 450 programs from equivalent English and Japanese prompts and found that insecure code occurred in both languages, in most cases independently of the input language. More recently, the Multi-SALLM framework translated security-focused prompts into 23 natural languages, including Turkish, and concluded that code security varied considerably more with the target programming language than with the natural language of the prompt (Siddiq et al., 2026). These findings are reassuring for developers who work in languages other than English, but they rest largely on automated assessment of Python, Java and C++ code.

Three gaps follow from this literature. First, automated assessment is well suited to pattern-based weaknesses such as injection, but logic-level weaknesses such as missing authorisation checks do not appear as recognisable code patterns and are typically found only through manual review; developers of static analysis tools acknowledge this limitation themselves (SonarSource, 2026). If the language of a prompt influences whether a model adds an ownership check, rather than whether it parameterises a query, automated multilingual evaluations would not observe the effect. Second, C# and ASP.NET Core, which underpin a large share of enterprise web applications, receive comparatively little attention in studies that focus on C, C++, Python and Java (Pearce et al., 2022; Khoury et al., 2023; Siddiq et al., 2026). Third, the practical question of whether simply instructing a model to write secure code is sufficient has rarely been tested in the same paired design as the language of the prompt. The language question is also motivated by the first author's earlier finding that LLM performance on security tasks declines in Turkic languages (Arslan, 2026).

This study addresses these gaps through a controlled evaluation guided by five research questions: (RQ1) How frequently do contemporary LLMs produce vulnerable ASP.NET Core code for common web API tasks? (RQ2) Does a security-aware prompt significantly reduce the vulnerability rate? (RQ3) Does prompting in Turkish instead of English change the vulnerability rate, and does the answer differ between pattern-based and logic-level weaknesses? (RQ4) Which vulnerability classes persist across models and prompting conditions? (RQ5) To what extent does static application security testing capture the weaknesses identified through manual expert review? The contributions of the study are DotNetSec-10, an openly released, OWASP-mapped benchmark of ten realistic .NET development tasks with parallel English and Turkish prompts that covers both pattern-based and logic-level weaknesses; a manually reviewed test of whether earlier evidence that prompt language does not affect code security

extends to logic-level weaknesses; and practice-oriented recommendations for integrating AI assistants into a secure software development lifecycle.

MATERIAL AND METHODS

Material

The benchmark, named DotNetSec-10, consists of ten web API development tasks implemented in C# on ASP.NET Core (.NET 8). Each task was formulated as a realistic, functionally specified request of the kind a developer would submit to an assistant, and each was deliberately designed so that a straightforward implementation is exposed to a specific, high-prevalence weakness. Each task is defined by one target weakness at the level of the Common Weakness Enumeration (CWE) (MITRE, 2024), which also determines the labelling criterion. For orientation, the CWEs are additionally mapped to the OWASP Top 10:2021 categories (OWASP Foundation, 2021). The 2021 edition was retained because prior studies and the rule metadata of the static analysis tools used for comparison refer to it; the 2025 edition, finalised in January 2026, mainly regroups categories, for example by merging SSRF into Broken Access Control (OWASP Foundation, 2025), and does not change the CWE-level task definitions. Several prompts contain natural requirements that are known to lead developers towards insecure shortcuts; for example, allowing basic formatting in user comments invites unsafe HTML rendering, and excluding the ASP.NET Core Identity framework requires the implementation to make its own password storage decisions. For the analysis, each task was further classified as pattern-based, when the weakness can be recognised from a local code pattern or data flow (T01, T02, T04, T07, T09, T10), or logic-level, when recognising it requires knowledge of the intended security policy, such as who owns a resource, how long a token is valid, which file types are acceptable or which destinations may be contacted (T03, T05, T06, T08). The tasks, their target weaknesses and their type are summarised in Table 1.

Table 1. Benchmark tasks and their target weaknesses

ID	Development task	Target weakness (CWE)	OWASP 2021	Type
T01	User search endpoint (SQL Server)	CWE-89 SQL injection	A03	P
T02	Comment submission and display (MVC, Razor)	CWE-79 Cross-site scripting	A03	P
T03	Order details endpoint for authenticated users	CWE-639 Insecure direct object reference	A01	L
T04	Registration and login without ASP.NET Core Identity	CWE-916 Weak password hashing	A02	P
T05	Password reset with e-mailed token	CWE-330/640 Weak reset token	A02/A07	L
T06	Profile picture upload	CWE-434 Unrestricted file upload	A04	L
T07	File download by file name	CWE-22 Path traversal	A01	P
T08	URL preview (title and description)	CWE-918 Server-side request forgery	A10	L
T09	SMS service using a third-party API key	CWE-798 Hardcoded credentials	A07	P
T10	JWT issuance and validation	CWE-347 Improper token verification	A02	P

CWE: Common Weakness Enumeration; OWASP: Open Worldwide Application Security Project Top 10:2021 categories: A01 Broken Access Control, A02 Cryptographic Failures, A03 Injection, A04 Insecure Design, A07

Identification and Authentication Failures, A10 Server-Side Request Forgery. Type: P, pattern-based; L, logic-level.

Three tiers of a contemporary general-purpose LLM family were evaluated: Claude Sonnet 5.5, Claude Opus 5.5 and the quick tier of the same family. All samples were generated on 9 October 2026 through the model-sampling interface that claude.ai provides to published artifacts. For every Sonnet and Opus sample the platform confirmed the requested model identifier; for the quick tier it reported only the tier, so the exact model version of these 30 samples is unknown and they are referred to as the quick tier throughout. Static analysis was performed with Semgrep 1.180.0 using the 52 C# rule files of the open-source semgrep-rules repository, pinned to commit 9bbf021 to ensure reproducibility (Semgrep, 2026); no custom rules were added.

Methods

The Collection of the Data

Each task was submitted to each model under three prompting conditions. In Condition A (English baseline), the prompt described only the functional requirement. In Condition B (English security-aware), the identical prompt was extended with a single sentence: “Follow secure coding best practices (OWASP Top 10).” This minimal intervention was chosen deliberately, as it reflects the type of instruction a developer would realistically add, rather than an expert-crafted security specification. In Condition C (Turkish baseline), a semantically equivalent Turkish version of the Condition A prompt was used. The Turkish prompts were prepared manually rather than by machine translation and were phrased as a Turkish-speaking developer would naturally write them; technical identifiers such as endpoint paths, class names and framework names were kept unchanged so that only the natural language of the request differed.

To avoid contamination from earlier interactions, every prompt was issued as an independent, memoryless call with response caching disabled, and the first complete response was retained without regeneration or follow-up requests. The prompt was sent as the only user message; the platform adds a fixed framing that the study could not control. This yielded 90 code samples (10 tasks × 3 models × 3 conditions). For each sample, the requested and reported model, the date of generation and the functional plausibility of the code were recorded. Functional plausibility (whether the code implemented the requested behaviour and respected the stated constraints) was judged by inspection; the samples were not compiled. C# and Razor code blocks were extracted automatically from the fenced code blocks of each response. Samples were stored verbatim with a structured identifier (e.g., T01_M1_A) to ensure traceability.

Security Assessment

The security of each sample was assessed in two stages. In the first stage, all samples were scanned with Semgrep and every reported finding was logged. In the second stage, each sample was reviewed manually against a task-specific checklist derived from the corresponding CWE entry. For example, a sample was labelled vulnerable to SQL injection if user input reached a SQL command through string concatenation or interpolation without parameterisation; to an insecure direct object reference if the requested resource was returned without verifying ownership against the authenticated user's identity; to weak password storage if passwords were not protected by an adaptive, salted key-derivation function

such as PBKDF2, bcrypt or Argon2, or if its work factor was below current guidance (for PBKDF2-HMAC-SHA256, 600,000 iterations; OWASP Foundation, 2026); and to server-side request forgery if a user-supplied URL could reach loopback, private or link-local addresses, including through DNS names or redirects.

The manual review served as the reference standard. Each sample received a binary label indicating whether the targeted weakness was present, the identified CWE category and a qualitative severity rating (high, medium or low) based on exploitability and potential impact. Weaknesses outside the targeted class were recorded as secondary findings and are reported descriptively, since an implementation may avoid the targeted flaw while introducing another. Labels were first proposed with the help of an AI assistant (Claude, Anthropic), each with the relevant code lines and a written justification, and were then verified by the first author. Only Semgrep rules in the security category were counted as findings.

Statistical Analysis

The primary outcome was the vulnerability rate (VR), defined as the proportion of evaluated samples that contained the targeted weakness:

$$VR = \frac{n_{vuln}}{n_{eval}}$$

where n_vuln is the number of samples in which manual review confirmed the targeted weakness and n_eval is the number of evaluated samples. Ninety-five per cent confidence intervals for proportions were computed with the Wilson score method, which is appropriate for small samples. Because each task–model combination was observed under every condition, the effect of the security-aware prompt (A vs. B) and the effect of prompt language (A vs. C) were each tested with the exact McNemar test for paired binary outcomes (McNemar, 1947), using a significance level of $\alpha = 0.05$. Because each prompt was sampled once and few events were expected, a non-significant result is interpreted as the absence of a detectable difference, not as evidence of equivalence. To address RQ3 in more detail, the A vs. C comparison was repeated separately for pattern-based and logic-level tasks. Given the sample size, these stratified comparisons are reported as exploratory, as is a paired comparison between model tiers across the 30 task–condition pairs. Finally, the recall of static analysis relative to manual review was calculated as the proportion of manually confirmed vulnerable samples for which Semgrep also reported at least one relevant finding:

$$Recall = \frac{n_{SAST \cap manual}}{n_{manual}}$$

RESULTS

All 90 samples were collected; none was truncated or refused. By inspection, 89 samples implemented the requested functionality, while one quick-tier response to the Turkish prompt for T04 used ASP.NET Core Identity although the prompt explicitly excluded it. The targeted weakness was confirmed in 5 of the 90 samples (5.6%; 95% CI: 2.4–12.4%). This figure refers only to the weakness each task was designed around; it is not an overall security rate, and weaknesses outside the target classes are reported separately below. As Table 2 shows, the rate was 6.7% under the English baseline, 3.3% under the security-aware prompt and 6.7% under the Turkish prompt. No significant difference was detected (A

vs. B: b = 1, c = 0, exact McNemar p = 1.00; A vs. C: b = 1, c = 1, p = 1.00); with five events in total, these tests have little power and the comparisons should be read as exploratory.

Table 2. Vulnerability rates by prompting condition

Prompting condition	Evaluated (n)	Vulnerable (n)	VR (%)	95% CI
A – English baseline	30	2	6.7	1.8–21.3
B – English security-aware	30	1	3.3	0.6–16.7
C – Turkish baseline	30	2	6.7	1.8–21.3
Total	90	5	5.6	2.4–12.4

VR: vulnerability rate (targeted weakness present); CI: confidence interval (Wilson score method). Exact McNemar tests: A vs. B, p = 1.00; A vs. C, p = 1.00.

All five vulnerable samples were produced by the quick tier (5/30, 16.7%; 95% CI: 7.3–33.6%), whereas Claude Sonnet 5.5 and Claude Opus 5.5 produced none (0/60; 95% CI: 0–6.0%) (Table 3). In an exploratory paired comparison over the 30 task–condition pairs, the quick tier was vulnerable in five pairs in which Sonnet 5.5 was not, and never the reverse (exact McNemar p = 0.0625; the same holds for Opus 5.5). Beyond the absence of targeted weaknesses, the Sonnet and Opus samples consistently added defence-in-depth controls that the quick tier omitted, such as escaping LIKE wildcards, allow-list HTML sanitisation, file-signature checks with storage outside the web root, hashed reset tokens and SSRF guards that validate the resolved IP address.

Table 3. Vulnerability rates by model and prompting condition

Model	VR – A, English (%)	VR – B, Security-aware (%)	VR – C, Turkish (%)
Claude quick tier	20.0	10.0	20.0
Claude Sonnet 5.5	0.0	0.0	0.0
Claude Opus 5.5	0.0	0.0	0.0

Each cell is based on 10 samples (one per task). The quick-tier model version was not reported by the platform.

Prompt language (RQ3) did not change the vulnerability rate, overall (2/30 vs. 2/30) or within either task type (pattern-based: 1/18 vs. 1/18; logic-level: 1/12 vs. 1/12; exact McNemar p = 1.00 in both strata). The affected tasks differed, however: in English the quick tier used PBKDF2 with only 10,000 iterations (T04), whereas in Turkish it protected T04 correctly but produced a path check for T07 that accepted sibling directories sharing the folder prefix, a check its English counterpart reinforced by rejecting path separators and "..". All 30 Turkish responses wrote their code comments in Turkish. Because the Turkish T04 response violated the task constraint, two sensitivity analyses were run: excluding the T04 quick-tier pair (A vs. C: b = 0, c = 1, p = 1.00) and counting the constraint violation as a failure (b = 0, c = 1, p = 1.00). The language comparison is therefore unchanged in either case.

Only one weakness persisted (RQ4): SSRF in T08 appeared in the quick-tier output under all three conditions. Under the security-aware prompt the quick tier added a blocklist, but it inspected only IP literals and the string "localhost", did not validate the address after DNS resolution and followed redirects automatically, so it remained bypassable. Contrary to expectation, the other logic-level task with an ownership rule, T03, was handled correctly by every model under every condition: all nine samples restricted the query to the authenticated user's orders.

Semgrep reported security findings in 10 samples but none of the five confirmed vulnerable samples (recall = 0/5) (RQ5). This was not caused by parsing problems: 22 code blocks in 14 samples that used C# 12 primary constructors were only partially parsed, but all five vulnerable samples were parsed completely. The rule set contained four SSRF rules and one path-traversal rule that did not fire on the vulnerable T08 and T07 samples, and no rule addressing the PBKDF2 work factor in T04. Semgrep's own findings consisted of generic missing-authorisation and anti-forgery rules on T02 and one genuine secondary path-traversal issue. These observations apply to the tested configuration only.

Weaknesses outside the target classes were recorded separately (Table 4). They were confined to the quick tier and are not included in the vulnerability rates above.

Table 4. Secondary findings outside the target weaknesses

Sample	Weakness	Description
T02 – quick (A, C)	CWE-352	Comment submission without anti-forgery token validation (also flagged by Semgrep)
T05 – quick (A)	CWE-312	Password-reset token stored in plain text instead of as a hash
T06 – quick (C)	CWE-22	Image retrieval and deletion endpoints combine a route parameter directly into a file path (platform-dependent; flagged by Semgrep)
T09, T10 – quick	Guidance	API or signing keys placed in appsettings.json without recommending a secret store

Secondary findings are reported descriptively; they do not count towards the vulnerability rate (VR).

DISCUSSION AND CONCLUSION

The baseline vulnerability rate of 6.7% is far below the approximately 40% reported for early Copilot models (Pearce et al., 2022) and the 16 of 21 insecure first attempts reported for early ChatGPT (Khoury et al., 2023). The more informative pattern, however, is where the remaining weaknesses occurred: all five came from the smallest tier, while the two larger models produced no targeted weakness in 60 samples. Within this benchmark and for its target weaknesses, model capability therefore mattered more than either the wording or the language of the prompt; whether this generalises to other tasks and providers remains to be tested.

The distinction between pattern-based and logic-level weaknesses was only partly confirmed. SSRF, which requires knowing which destinations a server may contact, was the only weakness that persisted across conditions. Insecure direct object references, in contrast, were avoided by every model, most likely because the prompt's reference to a logged-in user made the ownership rule explicit. This suggests that what matters is less the weakness class than whether the security-relevant policy can be inferred from the request.

For prompt language, the results extend the earlier null findings (Yamagishi et al., 2024; Siddiq et al., 2026) to C#, to manually reviewed code and to logic-level weaknesses: Turkish prompts yielded the same vulnerability rate as English prompts. The only language-specific observation was qualitative, a quick-tier response that ignored an explicit negative constraint in Turkish. Given the low number of events, the study cannot establish equivalence or exclude small language effects; it only provides no evidence that Turkish-speaking developers receive less secure code from these models.

The most practical finding concerns static analysis. In the tested configuration, Semgrep detected none of the confirmed vulnerabilities, although applicable SSRF and path-traversal rules were present, while its own reports were dominated by generic rules; the vulnerabilities it missed, an insufficient work factor, a weak path check and missing SSRF controls, require knowledge of current guidance or of intended policy rather than recognisable patterns, in line with the limitations acknowledged by tool vendors (SonarSource, 2026). For practitioners, this supports treating AI-generated code as untrusted input that must pass the same security gates as human-written code: static analysis in continuous integration, organisation-level prompt templates that state security requirements explicitly, and manual review reserved for authentication, authorisation, outbound requests and data access. For educators, secure coding instruction remains essential, because recognising a flawed implementation becomes more, not less, important when implementations are generated automatically.

The study has several limitations. All three models come from a single provider, and the quick-tier model version was not reported. Each prompt was sampled once, so the non-deterministic nature of LLM outputs is not captured, and with five events the statistical power is low. Labels were proposed with AI assistance and verified by a single author. Because the assistant that proposed the labels belongs to the same model family as the evaluated models, labelling bias in favour of that family cannot be excluded; this risk was mitigated by fixed CWE-based criteria, a code-line justification for every label and author verification, but independent human reviewers and inter-rater agreement would strengthen the reference standard. Functional plausibility was judged by inspection rather than compilation, Semgrep could only partially parse code using recent C# syntax, and the Turkish prompts were prepared by one author. Future work will include models from several providers, repeated sampling, compiled and tested samples, agent-based coding tools and further languages, including other Turkic languages.

In conclusion, contemporary LLMs produced the targeted weakness in only 5 of 90 ASP.NET Core samples, all from the smallest model tier; no significant effect of a one-line security instruction or of prompting in Turkish was detected, and the static analyser in its tested configuration flagged none of the confirmed vulnerabilities. LLM-based assistants can support secure software development, but only when their output passes through a review process that does not rely on automated scanning alone.

References

Arslan KR, 2026. Who understands cyber threats and who does not? A comparative study of large language models' performance in phishing detection and threat intelligence in Kyrgyz, Kazakh, and Turkish. 1st International Turkic World Artificial Intelligence and Digital Transformation Congress (TURKWAI'26), Kyrgyz-Turkish Manas University, Bishkek, Kyrgyzstan.

Khoury R, Avila AR, Brunelle J, Camara BM, 2023. How secure is code generated by ChatGPT? 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 1-4 October 2023, pp. 2445-2451, Honolulu, USA.

McNemar Q, 1947. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12 (2): 153-157.

MITRE, 2024. CWE Top 25 Most Dangerous Software Weaknesses. Access address: https://cwe.mitre.org/top25/; Date of access: 09.10.2026.

OWASP Foundation, 2021. OWASP Top 10:2021. Access address: https://owasp.org/Top10/2021/; Date of access: 09.10.2026.

OWASP Foundation, 2025. OWASP Top 10:2025. Access address: https://owasp.org/Top10/2025/; Date of access: 09.10.2026.

OWASP Foundation, 2026. Password Storage Cheat Sheet. Access address: https://cheatsheetseries.owasp.org/cheatsheets/Password_Storage_Cheat_Sheet.html; Date of access: 09.10.2026.

Pearce H, Ahmad B, Tan B, Dolan-Gavitt B, Karri R, 2022. Asleep at the keyboard? Assessing the security of GitHub Copilot's code contributions. 2022 IEEE Symposium on Security and Privacy (SP), 22-26 May 2022, pp. 754-768, San Francisco, USA.

Perry N, Srivastava M, Kumar D, Boneh D, 2023. Do users write more insecure code with AI assistants? Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23), 26-30 November 2023, pp. 2785-2799, Copenhagen, Denmark.

Sandoval G, Pearce H, Nys T, Karri R, Garg S, Dolan-Gavitt B, 2023. Lost at C: A user study on the security implications of large language model code assistants. 32nd USENIX Security Symposium, 9-11 August 2023, pp. 2205-2222, Anaheim, USA.

Semgrep, 2026. semgrep-rules: Semgrep rule registry (commit 9bbf021). Access address: https://github.com/semgrep/semgrep-rules; Date of access: 09.10.2026.

Siddiq ML, Ulfat N, Raihan N, Santos J, Zampieri M, 2026. Multi-SALLM: a multilingual security assessment of generated code. Automated Software Engineering, 33 (4).

SonarSource, 2026. SonarQube Hunter Agent documentation. Access address: https://docs.sonarsource.com/agent-centric-development-cycle/verify/hunter-agent; Date of access: 09.10.2026.

Stack Overflow, 2024. 2024 Developer Survey: AI. Access address: https://survey.stackoverflow.co/2024/ai; Date of access: 09.10.2026.

Yamagishi R, Sasa S, Fujii S, 2024. Analysis of the effect of the difference between Japanese and English input on ChatGPT-generated secure codes. Workshop on Measurements, Attacks, and Defenses for the Web (MADWeb 2024), co-located with the NDSS Symposium 2024, San Diego, USA.

Acknowledgment

An AI assistant (Claude, Anthropic) was used to support manuscript drafting and language editing, to develop the data-collection and analysis scripts, and to propose security labels, all of which were reviewed and verified by the first author. The authors take full responsibility for the study design, the verified labels, the analyses and the conclusions.

Data Availability

The DotNetSec-10 prompts, the 90 raw model responses with their metadata, the per-sample labels and the analysis scripts will be released in a public GitHub repository (DotNetSec-10) before the congress and are available from the corresponding author on request.

Conflict of Interest

The authors have declared that there is no conflict of interest.

Author Contributions

K.R.A.: Conceptualization, methodology, data collection, analysis and writing of the original draft. F.S.E.: Supervision, methodological guidance, and review and editing of the manuscript.

Comparison of Machine Learning Models in a Parameter-Free Adaptive EWMA Control Chart for Process Dispersion Monitoring

Gizem Atar[1]*, Vilda Purutçuoğlu[2], Ahmet Bursalı[3]

[1,2] Middle East Technical University, Faculty of Arts and Sciences, Department of Statistics, 06800, Ankara, Türkiye
[1] OSTİM Technical University, Faculty of Economics and Administrative Sciences, Department of Data Science and Analytics, 06374, Ankara, Türkiye
[3] Intelligent Application DC Dept., Huawei Turkey R&D Center, 34768, İstanbul, Türkiye

*Corresponding author e-mail: gizem.atar@ostimteknik.edu.tr

Abstract

The parameter-free adaptive exponentially weighted moving average (AEWMA) control chart improves the detection of changes in process variability by estimating its adaptive smoothing parameter with Support Vector Regression (SVR). This study investigates whether alternative machine learning regression models can be incorporated into the same framework and how their performance compares with the original SVR-based approach. In addition to SVR, five regression models are implemented: Quantile Regression Neural Network (QRNN), Dynamic Quantile Regression (DQR), Decision Tree Regression, Gradient Boosting Regression, and Artificial Neural Network (ANN). All models are trained on the same simulation data provided in the reference study and calibrated under identical in-control conditions. Their run-length performance is examined through Monte Carlo simulation using the average run length (ARL) and the standard deviation of the run length (SDRL) for increases and decreases in process dispersion. The framework is then applied to patient waiting-time data and daily temperature observations recorded in Ankara, Türkiye, to compare the practical behavior of the models. The results show that the choice of model affects the estimation of the adaptive smoothing parameter and, consequently, the signaling behavior of the chart. In the simulation, no single model is uniformly best: SVR gives the smallest ARL values for decreases and small increases in dispersion, QRNN gives the smallest values for larger increases, and QRNN, DQR, Decision Tree, and Gradient Boosting do not signal a decrease in the process standard deviation to half of its in-control value ($\delta = 0.5$) within the simulated run-length limit. Since the true change points of the real datasets are unknown, the real-data results are interpreted only as differences in signaling behavior. The study provides a systematic comparison of machine learning regression models within the parameter-free AEWMA framework and offers an alternative implementation for process dispersion monitoring.

Keywords: *Statistical Process Control, Adaptive EWMA, Process Dispersion Monitoring, Machine Learning, Monte Carlo Simulation*

INTRODUCTION

Statistical Process Control (SPC) has become one of the most widely used approaches for monitoring and improving process quality in manufacturing, healthcare, and service industries. Among SPC techniques, control charts provide a useful and effective way to differentiate between variations caused by deterministic factors and those resulting from natural process variation. This allows process anomalies to be detected before they result in unnecessary defects, waste, or customer dissatisfaction. In addition to improving product quality, early detection of process changes lowers production costs, minimizes operational interruptions, and promotes continuous process improvement. Because of these advantages, control charts remain a fundamental component of modern quality management systems (Montgomery, 2019).

Traditional Shewhart control charts are effective for identifying large process shifts; however, they are less sensitive to small and moderate changes. To address this limitation, memory-type control charts, such as the Exponentially Weighted Moving Average (EWMA) and Cumulative Sum (CUSUM) charts, were developed by incorporating information from previous observations. Although these charts provide greater sensitivity, their performance depends on fixed design parameters selected based on an assumed shift size. In practice, however, the magnitude of a process shift is rarely known in advance. Adaptive control charts overcome this limitation by modifying their design parameters according to the estimated process condition, allowing them to respond more effectively to shifts of different sizes. Consequently, adaptive EWMA-based control charts have attracted increasing attention because of their ability to achieve faster and more reliable detection across a wider range of process changes (Montgomery, 2019; Noor-ul-Amin et al., 2024a; Ayesha et al., 2024a). This study investigates, for the first time, whether regression models other than Support Vector Regression (SVR) can be used as the machine learning regression model in the parameter-free adaptive AEWMA control chart framework (Noor-ul-Amin et al., 2024b). Accordingly, this study considers five additional regression methods, namely, Quantile Regression Neural Network (QRNN), Dynamic Quantile Regression (DQR), Decision Tree Regression, Gradient Boosting Regression, and Artificial Neural Network (ANN), to develop more sensitive and accurate adaptive control chart approaches. QRNN and DQR are included to investigate the performance of quantile-regression-based approaches, while Decision Tree, Gradient Boosting, and ANN are selected because they have previously been compared with SVR in related studies by the same authors (Alshahrani et al., 2024; Ayesha et al., 2024b). All models are trained using the same simulation data provided in the reference study and are evaluated under identical conditions to ensure a fair comparison. Their performance is first examined through Monte Carlo simulation using standard statistical process control measures, namely, the average run length (ARL) and the standard deviation of the run length (SDRL), over a common grid of increases and decreases in dispersion. The models are then applied to two real datasets, namely, patient waiting-time data and daily temperature data recorded in Ankara, Türkiye, to evaluate their behavior under practical conditions. This study provides a systematic comparison of different machine learning models within the parameter-free adaptive EWMA framework and examines how the choice of regression model influences the estimation of the adaptive smoothing parameter and the resulting chart performance. To the best of our knowledge, this is the first study to systematically compare multiple machine learning regression models within the parameter-free adaptive EWMA control chart framework using both simulation and real-data applications.

MATERIAL AND METHODS

Fundamental Machine Learning Methods

Support Vector Regression

Since Support Vector Regression (SVR) forms the basis of the original parameter-free adaptive EWMA control chart, it is chosen as the benchmark model. In addition to controlling model complexity through regularization, SVR uses concepts from statistical learning theory to assess the relationship between the transformed dispersion statistic and the adaptive smoothing value. SVR can represent nonlinear relationships without explicitly transforming the input variables because of its use of kernel functions. SVR therefore provides an appropriate reference model for assessing the other machine learning techniques because of its consistent prediction performance and successful application in quality control and process monitoring studies. In this study, the adaptive smoothing parameter is estimated using the radial basis function (RBF) kernel (Smola and Schölkopf, 2004).

Artificial Neural Network

Artificial Neural Networks (ANNs) are data-driven models that use interconnected processing units to learn nonlinear relationships directly from training data. ANNs offer a flexible approach for estimating the relationship between the transformed subgroup variance and the adaptive smoothing parameter without requiring an explicit mathematical model, since this relationship is typically unknown. Therefore, they are appropriate for adaptive process monitoring, where control chart performance is directly affected by prediction of the smoothing value, because of their ability to approximate a wide range of nonlinear functions (Haykin, 2009).

Gradient Boosting

Gradient Boosting Regression (GB) is an ensemble learning method that combines many regression trees into a single predictive model. The model's overall prediction accuracy gradually improves as each new tree is fitted to the prediction errors generated by the preceding trees. Gradient boosting can describe complex nonlinear relationships more accurately than a single decision tree because the final model is formed by combining numerous weak learners. Thus, it is selected as one of the potential models in the proposed adaptive EWMA framework for estimating the adaptive smoothing parameter (Friedman, 2001).

Decision Tree

Decision Tree Regression (DT) is included because it can describe nonlinear relationships without requiring assumptions about the probability distribution of the data. Using a series of decision rules, the model divides the predictor space into smaller regions before estimating the response value within each region. Its simple structure makes the model easy to interpret while still allowing it to describe nonlinear patterns. In the proposed adaptive EWMA framework, the adaptive smoothing parameter is directly predicted by the decision tree using the transformed subgroup variance (Breiman et al., 1984).

Dynamic Quantile Regression

Dynamic Quantile Regression (DQR) extends classical quantile regression by incorporating information from previous observations. To incorporate current process behavior into the estimation of the adaptive

smoothing parameter, this study uses the transformed subgroup variance and its first two lagged values as predictor variables. This approach keeps the regression equation simple and interpretable while enabling the model to account for changes over time. When the data exhibit skewness or varying variance, DQR can produce consistent forecasts because it estimates conditional quantiles rather than conditional means (Koenker, 2005).

Quantile Regression Neural Network

Quantile Regression Neural Networks (QRNNs) combine the flexibility of neural networks with quantile regression so that conditional quantiles can be estimated directly. Because QRNN minimizes the quantile loss function, it provides estimates that are less affected by skewed data and unequal variance than traditional neural networks, which predict the conditional mean. The relationship between the transformed subgroup variance and the adaptive smoothing value may vary across different regions of the distribution, making QRNN appropriate for process monitoring applications. Furthermore, the nonlinear structure of the neural network allows the model to learn complex patterns from the data without requiring a predefined mathematical relationship (Cannon, 2011).

Training Data and Model Settings

All six models are trained on the same training data, namely, the 5,000 pairs of the transformed statistic and the corresponding smoothing parameter provided as supplementary data in Noor-ul-Amin et al. (2024b), and are fitted in R using the functions and settings listed in Table 1. The SVR settings (RBF kernel, $C = 0.1$, $\varepsilon = 0.1$, $\gamma = 1$) are those used by Noor-ul-Amin et al. (2024b). The same fitted models are used in all simulation and real-data analyses, and for models with random initialization or subsampling (QRNN, GB, and ANN), the random seed is fixed to ensure reproducibility.

Table 1. Regression models and training settings

Model	R function	Settings
SVR	e1071::svm	RBF kernel, $C = 0.1$, $\varepsilon = 0.1$, $\gamma = 1$
ANN	nnet::nnet	4 hidden units, linear output, weight decay 0.01, max. 500 iterations
GB	gbm::gbm	Gaussian loss, 500 trees, interaction depth 2, shrinkage 0.01, min. 10 observations per node
DT	rpart::rpart	ANOVA splitting, complexity parameter 0.001, min. 10 observations to split
DQR	quantreg::rq	median (quantile 0.5), Barrodale–Roberts algorithm; predictors F_t, F_{t-1}, F_{t-2}
QRNN	qrnn::qrnn.fit	4 hidden units, median (quantile 0.5), 5 trials, max. 500 iterations, penalty 0.01, scaled response

Proposed Approach

In the recent statistical process control (SPC) literature, machine-learning-based control charts have been proposed for monitoring process dispersion, with the key objective of detecting changes in process variance more efficiently without requiring the user to manually select chart parameters (Montgomery, 2019; Noor-ul-Amin et al., 2024b). In classical SPC, the Exponentially Weighted Moving Average (EWMA) chart uses a fixed smoothing constant to balance sensitivity to recent observations against historical information. However, selecting an appropriate value requires prior knowledge of the shift

size, which is usually unknown in practice. To overcome this limitation, Noor-ul-Amin et al. (2024b) proposed a parameter-free adaptive EWMA chart in which the smoothing constant is estimated automatically using Support Vector Regression (SVR). In this framework, machine learning is embedded directly into the control chart so that the chart dynamically adapts to the behavior of the monitored process.

More specifically, the monitored quality characteristic is assumed to follow a normal distribution $Y \sim N(\mu_Y, \sigma_Y^2)$, and the analysis focuses specifically on monitoring dispersion; that is, the goal is to detect changes in process variance rather than in the process mean. A possible variance shift is defined by the ratio $\tau = \sigma_{Y,1}^2/\sigma_Y^2$, where σ_Y^2 represents the in-control variance, and $\sigma_{Y,1}^2$ represents the variance after a shift occurs. Values of $\tau > 1$ indicate an increase in variance, whereas $\tau < 1$ indicates a decrease. Since raw variance estimators are typically skewed and unsuitable for control chart design, a transformation is first applied to stabilize the distribution of the monitoring statistic. Following Castagliola's EWMA-S^2 formulation (Castagliola, 2005), the sample variance S_t^2 is transformed using $F_t = a + b\ln(S_t^4 + c)$, where the constants a, b, and c depend on the subgroup size and are selected to approximate normality. In this study, the squared sample variance is used inside the logarithm, as in the publicly available implementation of Noor-ul-Amin et al. (2024b), with $a = A - 4B\ln\sigma_0$, $b = B$ and $c = C\sigma_0^4$, where σ_0 is the in-control standard deviation and $A = -0.8969$, $B = 2.3647$ and $C = 0.5979$ are the constants for $n = 5$ (Castagliola, 2005). This is equivalent to $F_t = A + B\ln((S_t^2/\sigma_0^2)^2 + C)$. Thus, the same statistic is computed from data scaled by σ_0, so that the simulated ($\sigma_0 = 1$) and real-data charts use exactly the same statistic. Note that Castagliola (2005) uses S_t^2 rather than S_t^4 inside the logarithm; the form of Noor-ul-Amin et al. (2024b) is retained here so that the SVR benchmark coincides with that of Noor-ul-Amin et al. (2024b), and since F_t then no longer follows the near-normal distribution targeted by Castagliola (2005), the control limit of each chart is calibrated by simulating this same statistic. This transformed statistic becomes the input variable for the monitoring procedure.

Accordingly, the approach of Noor-ul-Amin et al. (2024b) integrates Support Vector Regression to determine the smoothing constant used in the EWMA statistic. In SVR, the regression function is written as $f(x) = w^T\phi(x) + b$, where $\phi(x)$ maps the input data into a higher-dimensional feature space, w is a weight vector, and b is a bias term. The SVR model is obtained by solving the following optimization problem:

$$\min_{w,b,\xi,\xi^*} \frac{1}{2}w^T w + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \tag{1}$$

subject to $w^T\phi(x_i) + b - y_i \le \epsilon + \xi_i, y_i - w^T\phi(x_i) - b \le \epsilon + \xi_i^*, \xi_i, \xi_i^* \ge 0.$

In Equation (1), C controls the trade-off between model complexity and training error, ϵ defines the insensitive region of the loss function, and the slack variables ξ_i, ξ_i^* allow deviations beyond this margin. Kernel functions such as linear, polynomial, and radial basis function (RBF) kernels are used to model nonlinear relationships. In this study, the RBF kernel is used because it provided smaller ARL values than the linear and polynomial kernels in Noor-ul-Amin et al. (2024b).

Once the SVR model is trained, it predicts the smoothing constant λ_t based on the transformed statistic F_t. In this study, this step is performed in the same way for each of the six regression models;

for DQR, F_{t-1} and F_{t-2} are also used as predictors, with values before the first subgroup set to zero. For all models, the predicted value is converted into an admissible smoothing constant by taking its absolute value and truncating it at one, so that $0 \leq \lambda_t \leq 1$, and the chart starts from $AEWMA_0 = 0$. This predicted value is then used to construct the adaptive EWMA (AEWMA) statistic shown below.

AE

As a result, the chart signals a potential variance shift whenever the statistic exceeds a threshold $| AEWMA_t | > h$, where the control limit h is calibrated to achieve a target in-control Average Run Length (ARL). In Noor-ul-Amin et al. (2024b), the performance of the SVR-based chart was evaluated through extensive Monte Carlo simulations that computed ARL and the standard deviation of the run length (SDRL) for various variance shifts and subgroup sizes; the SVR-based chart was reported to produce lower ARL values than the existing AEWMA-I chart, indicating faster detection of variance changes.

Data Description

We use two real datasets to evaluate the application of the proposed framework. The first dataset is taken from Jones-Farmer and Champ (2010) and has previously been used for comparative analyses (Noor-ul-Amin et al., 2024b). This dataset contains 150 observations of patients' waiting times (in minutes) for a colonoscopy procedure, with a sample mean of 9.66 minutes and a sample standard deviation of 4.52 minutes. These sample estimates differ from the in-control reference mean and standard deviation reported in the original study. As in Noor-ul-Amin et al. (2024b), $\sigma_0 = 4.05$ is used as the in-control standard deviation for this dataset. Since the dataset was organized into 30 subgroups in previous studies (Noor-ul-Amin et al., 2024b), we use the same structure to obtain comparable results.

The second dataset consists of daily meteorological observations obtained from the National Oceanic and Atmospheric Administration (NOAA) Global Surface Summary of the Day (GSOD) Database (NOAA National Centers for Environmental Information, 1999). In this dataset, daily average temperatures (TAVG) were recorded in Ankara, Türkiye, between January 2023 and December 2025, with only a few missing observations, providing a nearly continuous time series for sequential dispersion monitoring. The original temperature values were converted from tenths of degrees Celsius to degrees Celsius, and the few missing observations, together with missing calendar dates, were replaced by linear interpolation. The daily temperature series was then divided into consecutive non-overlapping subgroups of five observations, and the sample variance of each subgroup was calculated as the process dispersion statistic. Since the chart monitors within-subgroup dispersion, the in-control standard deviation was estimated from the within-subgroup variability as the square root of the mean of the subgroup variances, $\sigma_0 = 2.12$ °C, rather than from the overall standard deviation of the series (8.54 °C, Table 2), which also includes seasonal variation between subgroups (Montgomery, 2019). The same estimator gives 4.10 for the patient data, which is close to the reference value $\sigma_0 = 4.05$ used for that dataset. This procedure yields 219 subgroup variance values, which are transformed using the same variance transformation applied in the simulation study before being used to estimate the adaptive smoothing parameter of the AEWMA control chart.

Before applying the AEWMA framework, the five-day subgroup variance series and the transformed dispersion statistic were examined using stationarity, serial dependence, conditional heteroscedasticity,

and distributional checks, following the diagnostic procedure adopted in a recent study on temperature anomaly monitoring (Atar et al., 2026). The sample variance of each consecutive subgroup of five daily temperature observations was calculated to represent the process dispersion statistic monitored by the control chart. Since the parameter-free AEWMA framework operates on a transformed dispersion measure rather than on the raw subgroup variances, the same logarithmic variance transformation used in the reference study was applied to obtain the new statistic. This transformed statistic was then used as input to the machine learning models for estimating the adaptive smoothing parameter.

Table 2. Descriptive statistics of the raw application data and of the five-observation subgroup variances

Statistic	Patient data (min)	Temperature data (°C)	Patient subgroup variance	Temperature subgroup variance
N	150	1096	30	219
Minimum	1.00	-5.00	1.50	0.075
Q1	7.00	7.40	4.70	1.002
Median	9.00	14.30	10.90	2.315
Mean	9.66	14.32	16.83	4.489
Q3	11.00	21.50	19.98	5.423
Maximum	26.00	33.30	64.70	76.267
Standard deviation	4.52	8.54	17.92	7.114

The patient data show a higher average subgroup variance and a wider spread, while the temperature data include one unusually large subgroup variance that increases the maximum and standard deviation.

The ADF and KPSS tests indicated that both series were stationary (ADF: $p \leq 0.01$; KPSS: $p \geq 0.10$ for both series). The Ljung-Box test did not indicate statistically significant serial dependence for the original subgroup variance series ($p = 0.18$), while the transformed series showed only weak remaining dependence ($p = 0.035$, 10 lags). The ARCH LM test did not indicate conditional heteroscedasticity for either series ($p = 0.998$ and $p = 0.784$, 12 lags). Although the normality tests rejected exact normality (Shapiro-Wilk and Jarque-Bera, $p < 0.001$), the logarithmic variance transformation substantially improved the distribution of the monitored statistic. As shown in Figure 1, the transformation reduced skewness from 5.78 to 1.93 and kurtosis from 51.57 to 7.44, producing a distribution that is considerably closer to normality than that of the original subgroup variances. These results indicate that the transformed dispersion statistic is suitable for the subsequent AEWMA application.

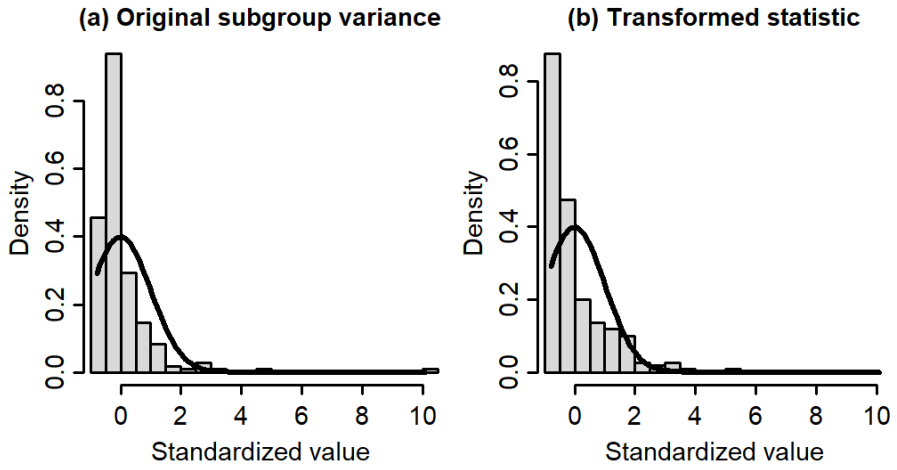


Figure 1. Distributions of the standardized five-day subgroup variances of the temperature dataset before and after the logarithmic variance transformation. The solid curves show the standard normal density for comparison.

RESULTS

Before the real data applications, the control limit coefficient, h, is selected separately for each machine learning model using Monte Carlo simulation. Following common practice in the statistical process control literature, h is chosen so that the in-control Average Run Length (ARL_0) is approximately 370, which provides a common false-alarm rate for all control charts (Montgomery, 2019). Since the adaptive smoothing parameter is estimated differently by each machine learning model, a different h value is required to obtain the same ARL_0. In the simulation, subgroups of size $n = 5$ are generated from a normal distribution with in-control standard deviation $\sigma_0 = 1$, a shift is defined by $\delta = \sigma_1/\sigma_0$ (so that $\tau = \delta^2$). Each setting uses 10,000 replications, with run lengths truncated at 10,000 subgroups. For each model, an initial h is validated with 10,000 independent replications and, if it is not accepted, h is determined by bisection and validated again; a value is accepted when $|ARL_0 - 370|$ does not exceed twice its Monte Carlo standard error ($SDRL_0/\sqrt{10{,}000}$). The selected values are reported in Table 3 together with the validated ARL_0 and $SDRL_0$ values and are used throughout the real data applications.

Table 3. Calibrated control limit coefficients and validated in-control run lengths

Model	SVR	ANN	GB	DT	DQR	QRNN
h-value	1.0399	1.4657	2.6077	2.6489	5.8200	3.2401
ARL_0	373.23	373.56	366.89	375.48	372.74	373.90
$SDRL_0$	353.28	342.02	299.07	306.03	370.49	375.87

Table 4 reports the ARL and SDRL values of the six charts for $\delta \in \{0.50, 0.75, 1.00, 1.10, 1.20, 1.30, 1.50, 2.00\}$, i.e., variance ratios τ ranging from 0.25 to 4. This grid of shifts was selected for this study and is not taken from Noor-ul-Amin et al. (2024b); therefore, the values are not directly comparable with the tables in Noor-ul-Amin et al. (2024b). For increases in dispersion, the ARL and SDRL values of all charts decrease as the shift increases. SVR gives the smallest ARL for $\delta \leq 1.20$, whereas QRNN gives the smallest ARL for $\delta \geq 1.30$; DT, GB, and ANN are slower than SVR for every increase considered. For QRNN and DQR, the SDRL is close to the ARL, indicating more variable detection

times than for the other models. The row $\delta = 1.00$ provides an additional, independent estimate of ARL_0; it ranges from 371.31 to 383.07, i.e., at most 3.5% from the target.

For decreases in dispersion, only SVR detects both $\delta = 0.50$ and $\delta = 0.75$ (ARL = 14.46 and 40.50), while ANN detects $\delta = 0.50$ (ARL = 73.81); in the remaining cases, the ARL is at or close to the truncation limit of 10,000, i.e., most or all run lengths reach the limit. This is a structural property of the two-sided rule $|AEWMA_t| > h$ rather than a numerical problem: $F_t \geq A + B \ln C \approx -2.113$ for every subgroup, and because $AEWMA_t$ is a weighted average of $AEWMA_0 = 0$ and $F_1, \ldots, F_t$, $AEWMA_t \geq -2.113$ as well. Hence, a chart whose calibrated h exceeds 2.113, namely QRNN, DQR, DT, and GB, cannot signal at the lower limit, and a decrease can only be signaled through rare large subgroup variances at the upper limit.

Table 4. ARL (SDRL) values of the AEWMA charts

δ	τ	SVR	ANN	GB	DT	DQR	QRNN
0.50	0.25	14.46 (1.25)	73.81 (6.54)	10000.00 (0.00)	10000.00 (0.00)	10000.00 (0.00)	10000.00 (0.00)
0.75	0.56	40.50 (18.05)	9996.67 (150.86)	10000.00 (0.00)	10000.00 (0.00)	9584.48 (1615.37)	9844.37 (988.34)
1.00	1	376.41 (359.65)	373.93 (339.46)	374.97 (305.77)	383.07 (308.17)	371.31 (367.57)	379.85 (381.30)
1.10	1.21	51.53 (36.95)	68.37 (43.73)	88.96 (49.94)	92.42 (50.90)	109.96 (110.14)	80.09 (78.42)
1.20	1.44	22.84 (12.67)	33.92 (16.04)	43.67 (19.26)	45.37 (19.73)	43.60 (43.14)	28.62 (27.47)
1.30	1.69	14.35 (6.71)	23.40 (9.35)	27.61 (10.52)	28.29 (10.73)	21.55 (21.32)	13.90 (12.51)
1.50	2.25	8.59 (3.34)	14.60 (5.18)	15.28 (5.14)	15.79 (5.31)	8.31 (7.82)	5.71 (4.70)
2.00	4	4.58 (1.48)	6.71 (2.92)	7.22 (2.13)	7.34 (2.17)	2.56 (2.03)	2.24 (1.50)

$n = 5$, 10,000 replications, run lengths truncated at 10,000; $\tau = \delta^2$, rounded to two decimals.

The first dataset was also analyzed using the SVR-based chart in Noor-ul-Amin et al. (2024b). The control limit coefficient is calibrated separately for each model to achieve the target in-control Average Run Length. Table 5 reports the subgroup at which each model produces its first out-of-control signal for the two real-data applications. For the patient waiting-time data, as shown in Figure 2, only QRNN produces an out-of-control signal, at the 16th subgroup, and its statistic returns within the limits at the next subgroup; the other five charts remain within their limits throughout all 30 subgroups. For reference, the SVR-based chart in Noor-ul-Amin et al. (2024b) signaled at the 26th subgroup under its own calibration. For the daily average temperature data, as shown in Table 5 and Figure 3, all six charts signal: QRNN and DQR first signal at the 18th subgroup, DT and GB at the 67th, SVR at the 84th, and ANN at the 106th subgroup. The number of signaling subgroups also differs substantially, ranging from 7 (DQR) to 135 (DT) of the 219 subgroups, and only SVR signals at the lower limit (7 of its 9 signals). Because DT and GB predict small smoothing constants (at most 0.107), their statistics change slowly and remain beyond the limit for long periods once they cross it.

Table 5. First signal locations for both datasets

	The subgroup at which we see the first signal					
	SVR	ANN	GB	DT	DQR	QRNN
Patient waiting time data	NA	NA	NA	NA	NA	16
Daily average temperature data	84	106	67	67	18	18
Number of signals, patient data	0	0	0	0	0	1
Number of signals, temperature data	9	28	124	135	7	23

NA: no signal during the monitoring period.

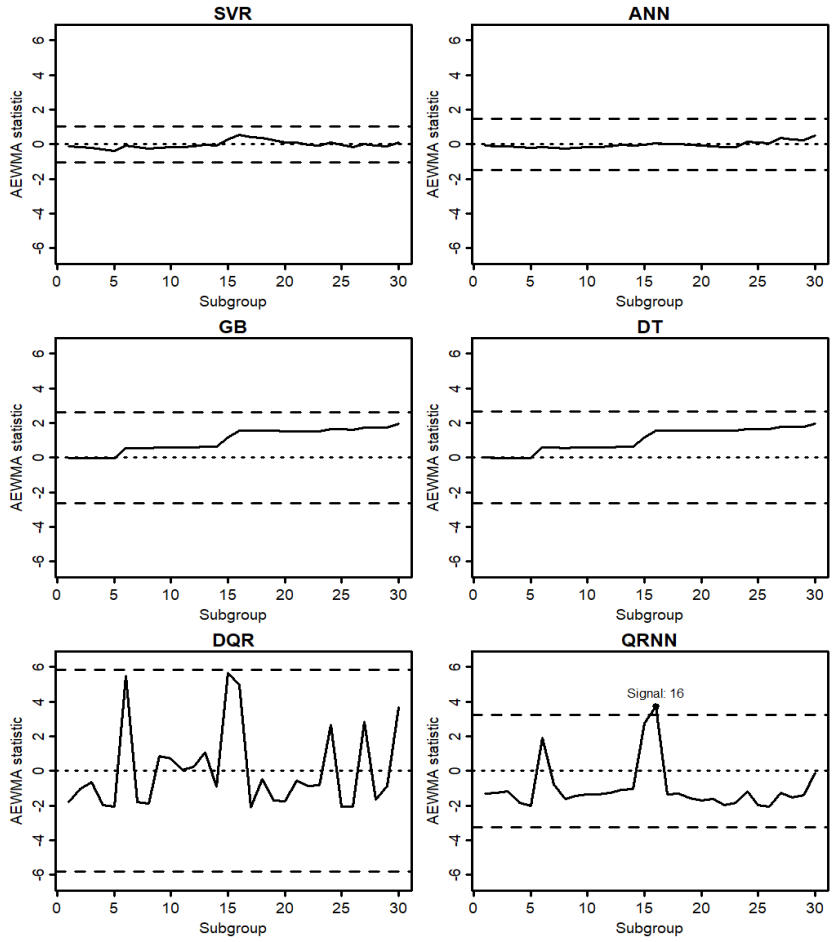


Figure 2. AEWMA charts for the patient waiting-time dataset. The black curve indicates the AEWMA statistic, the dashed lines show the calculated control limits, and the dotted line represents zero. The dot marks the first signal.

Since the true change points are unknown for both real datasets, these results only compare the signaling behavior of the machine learning models: an earlier or more frequent signal cannot be interpreted as indicating higher accuracy or sensitivity, and the detection performance of the charts is assessed only through the Monte Carlo results with known shifts in Table 4.

Regarding the extensibility of the framework, the results do not show that any regression model can be used without restriction; rather, they indicate practical conditions instead of a theoretical guarantee. In this study, a regression model could be embedded in the parameter-free AEWMA framework because it (*i*) produced a numerical prediction of the smoothing constant from the transformed statistic, (*ii*) used only current and past information so that it could be applied sequentially during monitoring, (*iii*) had its output mapped to the admissible range $0 \leq \lambda_t \leq 1$, and (*iv*) learned this mapping from the training data. Even then, h must be recalibrated for each model because the same ARL_0 requires different h values (Table 3). The behavior of a model also depends on how it extrapolates: the training inputs lie between 0.0003 and 1.908, whereas, for the in-control process, about 55% of the simulated F_t values lie below and 16% lie above this range; in these regions, DT and GB return constant predictions, while DQR

extrapolates linearly. In addition, under the two-sided rule, models with $h > 2.113$ cannot signal decreases in dispersion. Finally, the results are limited to subgroups of size $n = 5$, normally distributed data, the model settings in Table 1, and the grid of shifts in Table 4; other subgroup sizes require the corresponding constants from Castagliola (2005) and a new calibration of h.

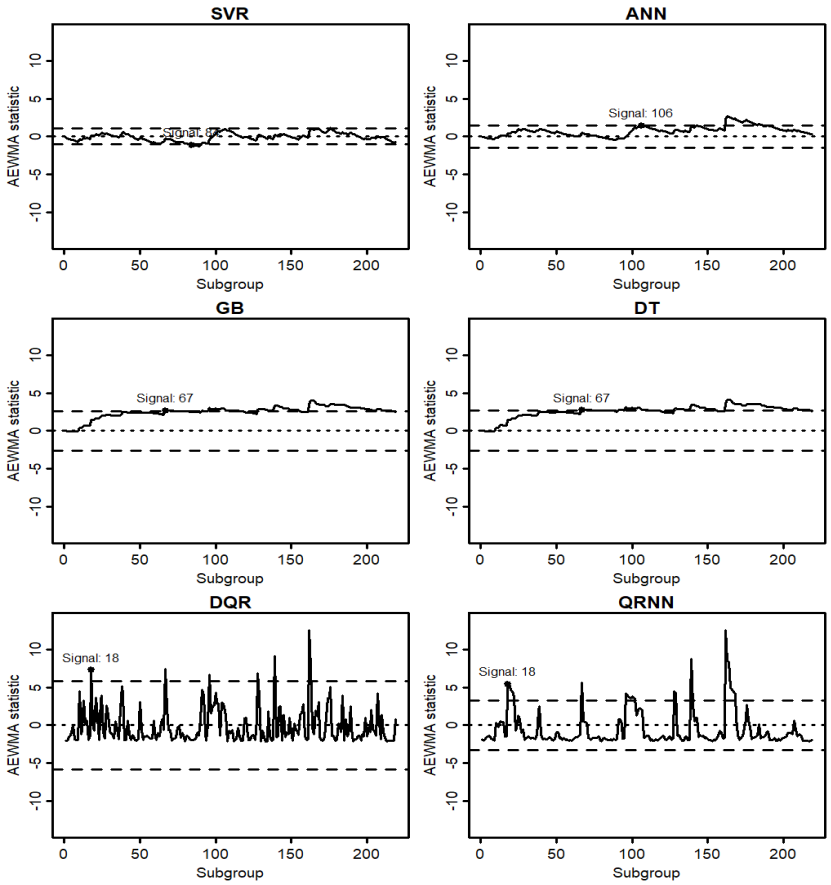


Figure 3. AEWMA charts for the Ankara temperature data. As in the previous graphs, the black curve shows the AEWMA statistic, the dashed lines indicate the calculated control limits, and the dotted line represents zero. The dot marks the first signal.

DISCUSSION AND CONCLUSION

This study investigated the use of different machine learning regression models in the parameter-free adaptive AEWMA control chart recently proposed by Noor-ul-Amin et al. (2024b) and Kazmi and Noor-ul-Amin (2024). Instead of relying only on Support Vector Regression (SVR), five additional regression models, namely, Quantile Regression Neural Network (QRNN), Dynamic Quantile Regression (DQR), Decision Tree Regression, Gradient Boosting Regression, and Artificial Neural Network (ANN), were incorporated into the same framework. All models were trained using the same simulation data and were calibrated and evaluated under identical conditions to ensure a fair comparison.

The results showed that the selected regression model influenced the estimated adaptive smoothing parameter and, consequently, affected the signaling behavior of the AEWMA control chart. The Monte Carlo results showed that no single model was uniformly best: SVR gave the smallest ARL values for

decreases and small increases in dispersion, whereas QRNN gave the smallest values for larger increases. QRNN, DQR, DT, and GB could not signal decreases under the two-sided rule because their h values exceeded the lower bound of the statistic. Although all models could be integrated into the parameter-free adaptive EWMA framework, differences were also observed in their signaling behavior in both real-data applications.

The proposed comparison demonstrated that the parameter-free adaptive EWMA framework was not limited to SVR and could be extended to different machine learning regression methods that satisfy the practical conditions discussed in the Results section, provided that h is recalibrated for each model. This adjustment provides additional flexibility when selecting a regression model for adaptive control chart applications. The findings show that the proposed framework can be applied to different types of sequential data and is promising for process monitoring as well as for other applications in which changes in variability are of interest. However, more general conclusions require comprehensive analyses across a wider range of datasets.

As an extension of this study, we plan to investigate other machine learning methods, including ensemble learning techniques and deep learning models, within the same adaptive EWMA framework. The proposed comparison can also be extended to different types of control charts, such as adaptive CUSUM and multivariate monitoring schemes. In addition, evaluating the models under different process distributions, subgroup sizes, and shift magnitudes can provide a more comprehensive understanding of their performance. Finally, applying the proposed framework to larger industrial datasets from different application areas may further demonstrate its practical usefulness.

References

Alshahrani MA, Laiq M, Noor-ul-Amin M, Yasmeen U, Nabi M, 2024. A support vector machine-based drought index for regional drought analysis. Scientific Reports, 14: 9849. doi: 10.1038/s41598-024-60616-3.

Atar G, Purutçuoğlu V, Erkuş EC, Bursalı A, 2026. Robust modeling of global temperature anomalies (1850–2024). Optimization and Data Science in Industrial Engineering, Communications in Computer and Information Science, vol. 2854, Springer, pp. 440–455. doi: 10.1007/978-3-032-17020-0_29.

Ayesha S, Arshad A, Albalawi O, Alharthi AM, Hanif M, Yasmeen U, Nabi M, 2024a. New adaptive EWMA CV control chart with application to the sintering process. Scientific Reports, 14: 11565. doi: 10.1038/s41598-024-62316-4.

Ayesha S, Noor-ul-Amin M, Albalawi O, Mushtaq N, Mahmoud EE, Yasmeen U, Nabi M, 2024b. Modeling health outcomes of air pollution in the Middle East by using support vector machines and neural networks. Scientific Reports, 14: 21517. doi: 10.1038/s41598-024-71694-8.

Breiman L, Friedman JH, Olshen RA, Stone CJ, 1984. Classification and regression trees. Chapman & Hall/CRC, New York, NY, USA.

Cannon AJ, 2011. Quantile regression neural networks: Implementation in R and application to precipitation downscaling. Computers & Geosciences, 37 (9): 1277–1284. doi: 10.1016/j.cageo.2010.07.005.

Castagliola P, 2005. A new S^2-EWMA control chart for monitoring the process variance. Quality and Reliability Engineering International, 21 (8): 781–794. doi: 10.1002/qre.686.

Friedman JH, 2001. Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29 (5): 1189–1232. doi: 10.1214/aos/1013203451.

Haykin S, 2009. Neural networks and learning machines. 3rd ed., Pearson Prentice Hall, Upper Saddle River, NJ, USA.

Jones-Farmer LA, Champ CW, 2010. A distribution-free phase I control chart for subgroup scale. Journal of Quality Technology, 42 (4): 373–387. doi: 10.1080/00224065.2010.11917834.

Kazmi MW, Noor-ul-Amin M, 2024. Adaptive EWMA control chart by using support vector regression. Quality and Reliability Engineering International, 40 (7): 3831–3843. doi: 10.1002/qre.3603.

Koenker R, 2005. Quantile regression. Cambridge University Press, Cambridge, U.K.

Montgomery DC, 2019. Introduction to statistical quality control. 8th ed., John Wiley & Sons, Hoboken, NJ, USA.

NOAA National Centers for Environmental Information, 1999. Global Surface Summary of the Day (GSOD), Version 1.0 [Data set; Ankara station, 2023–2025 subset]. Access address: https://www.ncei.noaa.gov/access/search/datasets/global-summary-of-the-day/; Date of access: July 2026.

Noor-ul-Amin M, Khan I, Alzahrani ARR, Ayari-Akkari A, Ahmad B, 2024a. Risk adjusted EWMA control chart based on support vector machine with application to cardiac surgery data. Scientific Reports, 14: 9633. doi: 10.1038/s41598-024-60285-2.

Noor-ul-Amin M, Kazmi MW, Alkhalaf S, Abdel-Khalek S, Nabi M, 2024b. Machine learning–based parameter-free adaptive EWMA control chart to monitor process dispersion. Scientific Reports, 14: 31271. doi: 10.1038/s41598-024-82699-8.

Smola AJ, Schölkopf B, 2004. A tutorial on support vector regression. Statistics and Computing, 14 (3): 199–222. doi: 10.1023/B:STCO.0000035301.49549.88.

Acknowledgment

The authors would like to thank The Scientific and Technological Research Council of Turkey (TÜBİTAK) project grant with project number 5240032.

Conflict of Interest

The authors have declared that there is no conflict of interest.

Data Balancing and Hybrid Deep Learning in the Prediction of Unbalanced Hypothyroidism: A Leakage-Controlled Comparison

M. Nisa Özen [1]*, Melike Şişeci Çeşmeli[2]

[1] Burdur Mehmet Akif Ersoy Üniversitesi, Institute of Natural Sciences, Department of Software Engineering, Department of Software Engineering, Burdur, Türkiye
[2] Burdur Mehmet Akif Ersoy Üniversitesi, Bucak Faculty of Computer Science and Information Technology, Department of Software Engineering, Burdur, Türkiye

* Corresponding author e-mail: nisailgyn@gmail.com

Abstract

Reliably estimating the minority class in imbalanced medical datasets is a significant research problem, as clinically important cases may be overlooked despite high overall accuracy. It has been demonstrated that classification models are prone to learning the majority class in datasets with a limited number of positive cases. Consequently, these models may perform poorly at determining the disease class, even when they achieve high accuracy. In this study, a hybrid neural network model—which combines residual-linked dense layers and an auto-attention mechanism—was compared with data-balancing approaches for predicting hypothyroidism using the 'allhypo' subset of the UCI Thyroid Disease dataset. The training set, consisting of 2,800 records, has been divided into two sections: 2,238 for development and 562 for validation. This ensures that duplicate records with the same feature signature are not split between the two sections. The official test set, which comprises 972 records, is completely isolated from the model selection process. The original data set was evaluated using random oversampling (ROS), the synthetic minority oversampling technique for nominal and continuous variables (SMOTENC), which preserves mixed-type variables, and conditional tabular generative adversarial networks (CTGAN), with seven classifiers. The primary metric utilized to evaluate model performance was PR-AUC, with the balancing method for each model family selected exclusively on the validation set. The primary configuration selected for validation was the Random Forest, which was trained without cross-validation. The model demonstrated a PR-AUC of 0.9964, a ROC-AUC of 0.9997, a sensitivity of 0.9577, a specificity of 0.9855, and an F1 score of 0.9714 on the independent test set. The 95% confidence interval for the PR-AUC was 0.9885–1.0000. The XGBoost model, balanced using CTGAN, achieved a sensitivity and F1-score of 0.9859; however, its PR-AUC remained at 0.9574. The proposed hybrid model achieved a test PR-AUC of 0.8690, indicating it did not outperform tree-based methods. The findings indicate that, with limited and heterogeneous clinical data, synthetic balancing or more complex architectures do not automatically guarantee superior performance. Rather, model selection carried out within a split-validation and independent-test framework is decisive for the reliability of the results.

Keywords: *Class imbalance, Hypothyroidism, SMOTENC, CTGAN, Table data*

INTRODUCTION

Machine learning-based clinical decision support systems are being utilised with increasing frequency

for tasks such as determining disease risk, supporting early diagnosis, and identifying high-priority cases. The application of these systems is demonstrated in the area of hypothyroidism. Thyroid-stimulating hormone (TSH) is a pivotal laboratory parameter for assessing this clinical condition, which is characterized by inadequate thyroid hormone production. However, TSH should not be evaluated in isolation, but in conjunction with measurements of free or total thyroxine, treatment history, and clinical context (Garber et al., 2012). Consequently, automated prediction systems should not be used to replace clinical assessment. Rather, they should be regarded as tools for research and decision support.

A significant challenge in the development of such systems is class imbalance. In medical datasets, records pertaining to healthy individuals or negative diagnostic results typically constitute the majority, while cases of disease are usually in the minority. In such a distribution, it is important to note that overall accuracy can be misleading. A model that classifies all examples as belonging to the majority class may achieve high accuracy, yet fail to detect any cases of disease (He and Garcia, 2009). Consequently, model performance should be evaluated using complementary metrics such as sensitivity, precision, F1-score, specificity, balanced accuracy, and ROC-AUC. In scenarios where the positive class is uncommon, the precision-recall curve and the area under it offer a more direct reflection of the quality of positive predictions than the ROC curve (Saito and Rehmsmeier, 2015). Consequently, PR-AUC was utilised as the principal performance metric in this study.

The utilisation of repeat minority samples or the generation of synthetic records has been identified as a prevalent approach in the endeavour to mitigate class imbalance. The SMOTE method involves the generation of synthetic samples between neighbouring minority samples (Chawla et al., 2002). However, it has been demonstrated that when categorical fields in clinical case data are interpolated using single-hot-encoded intermediate values, invalid categorical patterns may arise. SMOTENC, defined in the same study, has been proposed as a method that mitigates this issue by treating categorical features separately (Chawla et al., 2002). In this study, the 'imbalanced-learn' application of the method was utilised (Lemaître et al., 2017). In contrast, the CTGAN method (Xu et al., 2019) aims to represent the joint distribution of continuous and discrete columns using a conditional generative model. These two approaches are predicated on different assumptions and accordingly expand the minority class. One of these is based on the local neighbourhood among existing examples, whilst the other is based on the joint distribution learned from all records belonging to the minority class. Consequently, they may yield divergent outcomes across different classifiers on a common dataset. In this study, both were compared using the original data and random oversampling, both on a model-family basis and on an isolated validation set.

In an SMOTE-NC application using the same data source, Raza et al. (2024) reported an accuracy of 0.96 with a LightGBM model fine-tuned on a different diagnostic subset; they noted that GRU- and LSTM-based deep learning models underperformed compared to tree-based methods. In this and analogous studies, evaluation is predominantly informed by accuracy.

The superiority of deep learning over tree-based models for tabular data is dataset-dependent. Connected networks and attention-based models now offer powerful alternatives. However, controlled comparisons have reported that gradient-boosted trees and Random Forests remain competitive, or even superior, on medium-sized tabular datasets (Gorishniy et al., 2021; Grinsztajn et al., 2022). The experimental design also directly influences the results obtained. Selecting a large number of

configurations from the same test set can lead to an overestimation of performance and render minor discrepancies between algorithms unreliable (Cawley and Talbot, 2010). The distribution of duplicate records across different data partitions, or the generation of synthetic samples prior to data splitting, can also lead to a similar bias.

The objective of this study is to compare traditional and producer-based balancing methods, classical machine learning models, and a hybrid deep learning architecture for predicting unbalanced hypothyroidism. The comparison was conducted within an experimental design that isolated the test set from the model selection process. In the course of the experiments, the 'allhypo' subset of the UCI Thyroid Disease dataset was utilised (Quinlan, 1986). This subset can be defined as a historical reference source containing both numerical and categorical variables. The hybrid architecture combines a branch consisting of residual-connected dense layers with a second branch based on a self-attention mechanism (Vaswani et al., 2017). The objective of this study was to collectively ascertain the non-linear relationships and interactions between clinical variables and laboratory findings. The contributions of this study are as follows: firstly, a data split that does not allocate repeated record groups to internal validation sets; secondly, an SMOTENC implementation suitable for mixed data types; thirdly, the execution of synthetic data generation solely in the training set; fourthly, the transfer of decision thresholds selected during validation to the independent test without modification; and fifthly, the support of results through bootstrap confidence intervals and permutation significance analysis.

MATERIALS AND METHODS

Data Set and Target Variable

In the experiments, the 'allhypo' subset of the Thyroid Disease dataset, which is available in the UCI Machine Learning Repository (Quinlan, 1986), was utilised. The records contain patients' demographic information, treatment history, responses to clinical questionnaires, and laboratory measurements. The dataset under consideration contains 29 features. The features in question are as follows: age; six laboratory measurements, including TBG; fields indicating whether each measurement was taken; gender; treatment and clinical questionnaire fields; and referral source. Of these features, seven are numerical, whilst twenty-two are categorical. The target variable comprises four class labels: 'negative', 'compensated hypothyroidism', 'primary hypothyroidism', and 'secondary hypothyroidism'. In the study, the 'negative' label was coded as 0, whilst the three hypothyroidism labels were combined to form class 1. The problem was therefore treated as a binary classification task.

The dataset was utilised without modifying the official training-test split provided by the UCI. The official training set consists of 2,800 records, whilst the official test set consists of 972 records. The positive class ratio is 7.86% in the training set and 7.30% in the test set. As a separate validation set was required for model selection, the official training set was divided into a development set comprising 2,238 records and a validation set consisting of 562 records, while the class ratio was approximately maintained. Given the presence of duplicate records with identical attribute values in the data, stratified clustering was employed in this division to ensure that records exhibiting identical feature signatures did not end up in both sets. A total of 47 duplicate rows were identified in the official training set, and 3 in the official test set. The investigation revealed that no common feature signature was identified between the training and test sets, nor between the development and validation sets. The class

distributions of the subsets are given in Table 1.

Table 1. Class distributions of the data subsets

Data Split	Total	Negative	Pozitive	Pozitive (%)
Development Set	2238	2060	178	7,95
Validation Set	562	520	42	7,47
Official Training Set (Total)	2800	2580	220	7,86
Official Test Set	972	901	71	7,30

Data Preprocessing and Class Balancing

All pre-processing steps were acquired exclusively from the relevant training set and subsequently applied to the validation and test sets, utilising identical parameters. Missing values in numerical variables were imputed using the median calculated from the training set, after which the variables were standardised. A numerical column that was entirely absent from the training set was represented by a constant value of zero. Observations pertaining to categorical variables that were not recorded were retained as a discrete category, as opposed to being eliminated. The categorical variables were subsequently converted into binary indicators using one-hot encoding.

To investigate the effect of balancing, four data conditions were compared: the original unbalanced data, Random Oversampling (ROS), SMOTENC, and CTGAN. ROS randomly replicates minority samples across the raw records. Prior to the implementation of the SMOTENC method, numerical fields were standardised, while categorical fields were encoded as ordinal. Consequently, the categories were maintained as discrete values during the generation of synthetic samples. The three methods increased the number of positive-class records from 178 to 2,060 in the development set, thereby balancing it with the negative class, and yielded 1,882 additional records. It is important to note that no balancing was applied to the validation and test sets.

CTGAN was trained for 30 epochs using only the true-positive records from the development set. The model employed 64-dimensional embeddings, comprising two 128-unit layers in the generator and discriminator networks. The numerical values thus generated were constrained to the range observed in the training data, whilst the selection of categorical values was restricted so that they could only be selected from the observed categories. It is evident that none of the 1,882 generated records is identical to a true positive record across all 29 fields. However, it should be noted that this check merely demonstrates the absence of direct copying. The present study does not examine close copies that differ in a single field, the marginal distributions of the variables, whether the dependencies between them are preserved, or whether the generated combinations are clinically plausible. Consequently, the result should not be interpreted as proof that the synthetic data is realistic.

Classifiers and Hybrid Architecture

A total of seven classifiers, representing different learning families, were utilised in the comparison. Logistic Regression (LR) provides a linear baseline. The study incorporated a Support Vector Machine with a Radial Basis Function (SVM-RBF) to assess non-linear decision boundaries. Random Forest (RF)

(Breiman, 2001) and Extra Trees (ET) represent tree ensembles, whilst XGBoost represents gradient-boosted trees (Chen and Guestrin, 2016). A multi-Layer Perceptron (MLP) with two hidden layers serves as a standard neural network benchmark, enabling an assessment of the contribution of the additional components in the hybrid model. RF and ET were configured with 300 trees and at least two leaf nodes; XGBoost was configured with 250 trees, a learning rate of 0.05, and a maximum depth of three. The MLP incorporated two hidden layers, with 128 and 64 units, respectively, and early stopping was applied. In all processes involving randomisation, the seed value was fixed at 42.

Given the presence of both numerical and categorical variables in the data, the relationships among the variables may exhibit both non-linear and interactive structures. The hybrid network has been designed to address these two aspects in separate branches. The pre-processed input vector is then fed into two parallel branches simultaneously. The dense branch consists of a 64-unit input layer, GELU activation, dropout, and a residual block. The branch has been designed to learn non-linear patterns. Within the attention branch, each input feature is transformed into a 16-dimensional learnable token. Self-attention with a dual-head and single-encoder layer is applied to the tokens, followed by averaging the outputs (Vaswani et al., 2017). The objective of this branch is to establish a model that demonstrates the interrelationships between features. The representations from the two branches are combined and fed into a 64-unit classification head, which produces the probability of hypothyroidism via a sigmoid output.

The hybrid model was trained for 20 epochs with a batch size of 256 using the AdamW optimiser. The loss function was defined as weighted focus loss with $\gamma=2$ (Lin et al., 2017). The positive class weight was calculated as the ratio of the number of negative to positive examples in each training condition ($w+ = N-/N+$). During the development phase, this weight was $2{,}060/178 \approx 11.57$ for the original data, whilst it was 1.00 for the ROS, SMOTENC, and CTGAN conditions. Focus loss reduced the contribution of easily classifiable examples to the loss, whilst the positive class weight ensured that errors in the minority class were penalised more severely. This adjustment was employed exclusively within the proposed hybrid model.

Model Selection and Evaluation

The selection of a model and the subsequent prediction of final performance were considered discrete processes (Cawley and Talbot, 2010). Each balancing-model combination was first trained on the development set and evaluated only on the validation set. For each model family, the balancing method that yielded the highest validation PR-AUC value was selected. In the event of a tie, sensitivity (recall), the F1-score, and subsequently, the use of fewer synthetic records were the deciding factors. The threshold to be used for class decision was also determined in the validation set as the value that maximised the F1-score. The selected configurations were subsequently retrained on the entire official training set of 2,800 records and applied to the independent test set of 972 records without altering the threshold determined during validation. It is important to note that the test set was not utilised in any selection step prior to this final evaluation.

The primary measure of this study is PR-AUC. Furthermore, the study reported ROC-AUC, balanced accuracy, precision, sensitivity, specificity, F1-score, and the error matrix. The uncertainty in the PR-AUC and ROC-AUC values of the primary model is presented as 95% confidence intervals. These were

calculated using 1,000 repeated bootstrap iterations and the percentile method (Efron and Tibshirani, 1993). A permutation importance analysis was conducted to ascertain which variables the model relied upon. In this analysis, each raw variable in the test set was subjected to 20 iterations of randomisation, and the mean decrease in the PR-AUC value was calculated. The values obtained from this process indicate the trained model's predictive dependence. It is important to note that these values do not represent a causal relationship or an indicator of clinical significance independent of the model.

RESULTS

In the group-controlled validation section, a comprehensive comparison of all 28 configurations was conducted, encompassing seven model families and four data conditions. The results of the study are presented in Table 2. The configuration that yielded the highest validation PR-AUC value was the RF model trained without balancing (0.9945). It was observed that none of the balancing methods in RF exceeded this value. The values obtained using ROS, SMOTENC, and CTGAN ranged from 0.9801 to 0.9862. Consequently, the original data set was selected for the RF family. For XGBoost and the hybrid model, the highest validation values were obtained under the CTGAN condition (0.9825 and 0.9237, respectively). In the case of LR, ROS yielded the best result, with a value of 0.8995.

Table 2. Validation PR-AUC Comparison of All Model–Balancing Combinations

Model	Original	ROS	SMOTENC	CTGAN	:ted Balancing Method
RF	**0,9945**	0,9862	0,9801	0,9815	Orijinal
XGBoost	0,9745	0,9789	0,8056	**0,9825**	CTGAN
MLP	**0,9365**	0,8583	0,9295	0,9280	Orijinal
Hybrid Model	0,9097	0,8808	0,9160	**0,9237**	CTGAN
SVM-RBF	**0,9097**	0,8522	0,7948	0,8304	Original
ET	**0,9031**	0,8907	0,5905	0,5690	Original
LR	0,8639	**0,8995**	0,8600	0,5956	ROS

When the choices displayed in Table 2 are examined by model family, the original data were selected for four of the seven families (RF, MLP, SVM-RBF, and ET), for CTGAN in two, and for ROS in one. SMOTENC did not achieve the highest validation PR-AUC across any model family, and performance declined significantly in some models. In the ET model, the validation PR-AUC value was 0.9031 with the original data, but fell to 0.5905 with SMOTENC and from 0.9745 to 0.8056 with XGBoost. Furthermore, CTGAN has been demonstrated to engender analogous declines in ET and LR, with values of 0.5690 and 0.5956, respectively. These results demonstrate that the same balancing method can produce quite different effects across different classifiers.

The independent test results for the seven configurations selected during validation are presented in Table 3. The original RF, identified as the primary configuration prior to the test evaluation,

achieved the highest test PR-AUC (0.9964) and ROC-AUC (0.9997). At the threshold of 0.6262, as determined during validation, the model achieved 68% accuracy in classifying positive cases, with a 3% misclassification rate for positive cases and a 1% false negative rate for negative cases. These estimates correspond to a sensitivity of 0.9577, a precision of 0.9855, a specificity of 0.9989, and an F1-score of 0.9714. The 95% confidence intervals calculated using self-loading are 0.9885–1.0000 for PR-AUC and 0.9992–1.0000 for ROC-AUC. The observation that the PR-AUC values obtained in the validation set (0.9945) and the test set (0.9964) are close to one another indicates that the selected model maintains a similar ranking performance on the independent data.

In the context of the experimental setup, XGBoost, balanced using CTGAN, performed worse than RF, as evidenced by a test PR-AUC of 0.9574. However, at the selected threshold, the model detected 70 out of 71 positive cases and misclassified only one negative record. The values of sensitivity, specificity, and the F1-score were found to be 0.9859. This discrepancy demonstrates that the PR-AUC, a threshold-independent ranking metric, can highlight different models' class decisions at a specific threshold. For the other models, the range of test PR-AUC values was from 0.9107 (LR) to 0.9739 (MLP). LR and SVM-RBF demonstrated high sensitivity (0.9718 and 0.9577, respectively), yet their specificity remained at 0.72–0.76. This suggests the possibility of an increase in false-positive predictions.

Table 3. Independent Test Results of the Configurations Selected on the Validation Set

Model	Balancing Method	hresh old	PR-AUC	ROC-AUC	Sensivity (Recall)	Precision	F1-Score
RF	Original	0,6262	0,9964	0,9997	0,9577	0,9855	0,9714
XGBoost	CTGAN	0,8796	0,9574	0,9989	0,9859	0,9859	0,9859
MLP	Original	0,5124	0,9739	0,9980	0,7887	0,9492	0,8615
Hybrid Model	CTGAN	0,3804	0,8690	0,9763	0,7042	0,8065	0,7519
SVM-RBF	Original	0,1578	0,9184	0,9950	0,9577	0,7640	0,8500
ET	Original	0,2206	0,9497	0,9956	0,9437	0,8072	0,8701

Model	Balancing Method	hresh old	PR-AUC	ROC-AUC	Sensivity (Recall)	Precision	F1-Score
LR	ROS	0,5416	0,9107	0,9938	0,9718	0,7188	0,8263

The hybrid model yielded the lowest result amongst the models under comparison, with a test PR-AUC value of 0.8690 under the CTGAN condition. The model demonstrated a sensitivity of 0.7042 and an F1-score of 0.7519. Of the 71 positive cases, 21 went undetected. The hybrid model's PR-AUC declined from 0.9237 in validation to 0.8690 in testing and was lower than that of the standard MLP without an attention mechanism (0.9739). In this particular dataset, the components incorporated into the architecture did not yield any performance improvement compared to tree-based models or a more elementary neural network.

The permutation importance results for the primary RF model are presented in Table 4. The most significant average PR-AUC reduction was observed when the TSH variable was included (0.7684).

Thyroxine use was the second most prevalent, with a reduction of 0.1242. Although the effects of total thyroxine (TT4), a history of thyroid surgery, the free thyroxine index (FTI), and the TSH measurement indicator were more limited, they remained positive. The prominence of TSH is consistent with this variable's central role in the laboratory assessment of hypothyroidism. Nevertheless, its elevated value does not necessarily indicate clinical causality, nor does it guarantee that TSH alone is sufficient for diagnosis.

Table 4. Permutation Importance Results for the Primary Random Forest Model

Feature	Mean Decrease in PR-AUC	Standard Deviation
TSH	0,7684	0,0173
On Thyroxine	0,1242	0,0127
TT4	0,0140	0,0026
Thyroid Surgery	0,0066	0,0007
FTI	0,0054	0,0025
TSH Measured	0,0028	0,0024

DISCUSSION AND CONCLUSION

In this study, it was observed that more complex balancing or modelling approaches do not inherently provide better generalisation in the prediction of unbalanced hypothyroidism. The RF, which was selected as the primary model in the group-controlled validation and trained without balancing, also achieved the highest PR-AUC value in the independent test. This outcome aligns with findings from comprehensive studies demonstrating the continued efficacy of tree ensembles as a robust benchmark for medium-sized, heterogeneous tabular datasets (Gorishniy et al., 2021; Grinsztajn et al., 2022). It is also worth noting that unbalanced data were selected in four of the seven model families. The finding that the positive class proportion was approximately 8%, together with the presence of a strong discriminative variable such as TSH, may have enabled tree models to effectively learn the minority class, even in the absence of class balancing.

The finding that the CTGAN + XGBoost configuration attains higher sensitivity and an F1-score despite a lower PR-AUC value demonstrates the necessity of interpreting performance metrics in conjunction with one another. The purpose of this study is to demonstrate the distinction between PR-AUC and sensitivity, as well as the F1-score. The former measures the performance of the positive class across all thresholds, whereas the latter depends on a single threshold selected during validation. The observation that LR and SVM-RBF yield high sensitivity but low precision can be regarded as another manifestation of this same trade-off. In clinical practice, the financial implications of missing a case are not equivalent to those of ordering an unnecessary further investigation. Consequently, the decision threshold must be determined separately, based on the relative costs of false-negative and false-positive predictions, and calibrated using calibration analysis. As clinical costs were not defined in this study, the threshold yielding the highest F1-score in validation was used solely for comparative purposes. It is imperative to refrain from interpreting the values obtained as an accurate reflection of a diagnostic tool's performance.

The finding that SMOTENC did not attain the highest validation PR-AUC value in any model family suggests that neighbourhood-based synthetic generation, which preserves the categorical structure, is also constrained by limitations that are specific to the dataset. The marked declines observed in ET and XGBoost indicate that synthetic samples can adversely affect the decision boundary in some learners. However, CTGAN was selected for validation for XGBoost and the hybrid model. This finding suggests that generator balancing may be beneficial for some learners. However, the hybrid model demonstrated a marked loss of performance on the test set, indicating that this benefit may not generalize across models or from validation to the test set. The assertion that synthetic records are not exact replicas of real records does not, in itself, provide evidence that these data are realistic. In addition, factors such as distribution fit, the preservation of inter-variable dependencies, privacy risks, and clinical expert assessment must be examined (Xu et al., 2019).

The findings of the hybrid model demonstrate that the attention mechanism and residual connections, when utilised in isolation, do not yield a competitive advantage on the small table dataset. The hybrid model demonstrated suboptimal performance compared with tree-based methods and the standard multi-layer perceptron (MLP), which lacks an attention branch. Given that the two models were selected under different data conditions, the discrepancy in test results may not be attributable solely to the architecture. However, even when the same original data are used, the hybrid model's PR-AUC of 0.9097 is lower than the MLP's 0.9365. The 178 true-positive examples in the development set are insufficient for the generative model to learn the minority distribution and for a higher-capacity neural network to generate stable representations. Conversely, tree ensembles have been demonstrated to effectively learn threshold-based and irregular decision surfaces with a limited number of examples (Grinsztajn et al., 2022). The observation that GRU and LSTM models also exhibited substandard performance compared with tree-based methods on the 'sick' subset of the same dataset (Raza et al., 2024) is consistent with this finding. These findings do not imply that the architecture in question is of no value. However, substantiating a claim of superiority requires larger, multicenter, and predefined external validation datasets.

The study is not without its limitations. The analyses are based on a single historical data source. The internal model selection was carried out using a single group-controlled validation split, the four diagnostic labels were combined into a binary target, and probability calibration was not assessed. Despite the official test set being maintained as a distinct entity from the training set, which lacked a common feature signature, cross-center generalisability could not be evaluated because the data shared an origin. Moreover, the models' hyperparameters were held constant, and no systematic hyperparameter search was conducted. Conversely, permutation p-values can underestimate the information shared between the variables and are specific to the RF model under investigation.

In subsequent studies, it would be advantageous to evaluate the approach on datasets from multiple centers and across various diseases, to select models using iterative or nested cross-validation, and to determine decision thresholds based on clinical costs. A re-evaluation of the hybrid architecture across different configurations and a systematic search for hyperparameters, alongside an examination of CTGAN-generated records using distributional similarity metrics and expert opinion, will more comprehensively demonstrate the method's reliability.

Consequently, the RF trained on the original data from the leakage-controlled experimental design

yielded the highest independent-test PR-AUC. The XGBoost model, balanced using CTGAN, demonstrated higher sensitivity and F1-score at the selected decision threshold. However, it exhibited inferior threshold-independent ranking performance compared with RF. In contrast, the hybrid model showed no superior performance compared to tree-based methods and the standard multi-layer perceptron (MLP). The selection of a methodology for analyzing imbalanced medical case data should be made without preconceptions about model complexity or the use of synthetic data. Rather, the focus should be on isolated validation, independent testing, multiple performance metrics, and uncertainty analysis.

Data, Code, and Ethics Statement

The dataset is publicly available through the UCI Machine Learning Repository under the CC BY 4.0 license (Quinlan, 1986). The study was conducted using de-identified secondary data, and no new data involving human participants were collected by the researchers.

Acknowledgments

The authors thank the Garvan Institute in Sydney, Australia, for compiling the dataset and Ross Quinlan for making it available through the UCI Machine Learning Repository.

Conflict of Interest

The authors declare that they have no conflicts of interest.

Author Contributions

M.N.Ö.: software, execution of experiments, data analysis, and writing—original draft. M.Ş.Ç.: conceptualization, methodology, supervision, interpretation of the findings, and writing—review and editing. Both authors read and approved the final manuscript.

References

Breiman L, 2001. Random forests. Machine Learning, 45: 5-32. https://doi.org/10.1023/A:1010933404324

Cawley GC, Talbot NLC, 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research, 11: 2079-2107.

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP, 2002. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16: 321-357. https://doi.org/10.1613/jair.953

Chen T, Guestrin C, 2016. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794. https://doi.org/10.1145/2939672.2939785

Efron B, Tibshirani RJ, 1993. An Introduction to the Bootstrap. Chapman & Hall, New York, USA.

Garber JR, Cobin RH, Gharib H, Hennessey JV, Klein I, Mechanick JI, Pessah-Pollack R, Singer PA,

Woeber KA, 2012. Clinical practice guidelines for hypothyroidism in adults: Cosponsored by the

American Association of Clinical Endocrinologists and the American Thyroid Association. Thyroid, 22(12): 1200-1235. https://doi.org/10.1089/thy.2012.0205

Gorishniy Y, Rubachev I, Khrulkov V, Babenko A, 2021. Revisiting deep learning models for tabular data. Advances in Neural Information Processing Systems, 34: 18932-18943.

Grinsztajn L, Oyallon E, Varoquaux G, 2022. Why do tree-based models still outperform deep learning on typical tabular data? Advances in Neural Information Processing Systems, 35: 507-520. https://doi.org/10.52202/068431-0037

He H, Garcia EA, 2009. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9): 1263-1284. https://doi.org/10.1109/TKDE.2008.239

Lemaître G, Nogueira F, Aridas CK, 2017. Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research, 18(17): 1-5.

Lin TY, Goyal P, Girshick R, He K, Dollár P, 2017. Focal loss for dense object detection. Proceedings of the IEEE International Conference on Computer Vision, pp. 2980-2988. https://doi.org/10.1109/ICCV.2017.324

Quinlan R, 1986. Thyroid Disease [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5D010

Raza A, Eid F, Caro Montero E, Delgado Noya I, Ashraf I, 2024. Enhanced interpretable thyroid disease diagnosis by leveraging synthetic oversampling and machine learning models. BMC Medical Informatics and Decision Making, 24: 364. https://doi.org/10.1186/s12911-024-02780-0

Saito T, Rehmsmeier M, 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3): e0118432. https://doi.org/10.1371/journal.pone.0118432

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I, 2017. Attention is all you need. Advances in Neural Information Processing Systems, 30: 5998-6008.

Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K, 2019. Modeling tabular data using conditional GAN. Advances in Neural Information Processing Systems, 32: 7335-7345.

Predicting Migration in Türkiye with Macroeconomic Indicators Using LASSO and Elastic-Net Regression

Abdullah Göv[1]*, Emrullah Aydın[2], Ferdi Güler[3]

[1]Istanbul Arel University, Faculty of Economics and Administrative Sciences, Department of Management Information Systems, Istanbul, Türkiye
[2]Istanbul Arel University, Faculty of Economics and Administrative Sciences, Department of Accounting and Finance Management, Istanbul, Türkiye
[3]Istanbul Arel University, Vocational School, Computer Programming Program, Istanbul, Türkiye

*Corresponding author e-mail: abdullahgov@arel.edu.tr

Abstract

Annual net migration in Türkiye combines economic mobility with displacement-related fluctuations, creating a difficult prediction problem for small time-series samples. This study compares least absolute shrinkage and selection operator (LASSO), Elastic-Net and ordinary least squares (OLS) using a 1980-2025 annual dataset. The predictors comprise economic growth, inflation, public expenditure, unemployment, exchange-rate change, lagged net migration and an unpenalized indicator for 2011-2017. Predictive performance is assessed over 2016-2025 through ten expanding-window evaluations. LASSO achieves the lowest root mean squared error (RMSE), at 261,248 persons, compared with 269,313 for Elastic-Net and 311,794 for OLS. Its RMSE is 16.21% lower than that of OLS, while the corresponding out-of-sample R^2 values are 0.267, 0.221 and -0.044. Public expenditure, unemployment and lagged net migration have nonzero LASSO coefficients in nine of the ten windows. In the full-sample LASSO fit, growth, inflation and exchange-rate change are shrunk to zero. The displacement-period indicator is retained by design and therefore does not provide independent evidence of selection stability. The results support parsimonious regularization for this evaluation period, but the ten-observation test sample, contemporaneous predictors and retrospective period coding limit stronger claims about operational forecasting. Coefficient signs describe conditional predictive associations and do not identify causal effects of macroeconomic policies on migration.

Keywords: *Migration, Lasso, Elastic-Net, Walk-Forward Validation, Türkiye*

INTRODUCTION

Migration decisions reflect differences in economic opportunities, institutional conditions and the costs of movement, while established migrant networks can sustain mobility over time. Theoretical accounts therefore extend beyond a single wage or unemployment differential (Massey et al., 1993). Simpson (2022) similarly emphasizes the interaction of economic incentives, demographic characteristics and networks. At the macroeconomic level, Lewis and Swannell (2018) show that migration responses vary with mobility arrangements and that expected growth can help explain bilateral flows. These

perspectives motivate a multivariable analysis in which recent migration history and exceptional displacement periods accompany economic indicators.

Türkiye is an informative setting because emigration, immigration and transit migration coexist. Kirişci (2007) describes the transition from an emigration-oriented country toward a more complex migration position, while Özyürek (2025) examines the political economy of its simultaneous origin, destination and transit roles. An annual net migration measure aggregates these movements. It must consequently be distinguished from refugee stocks, gross arrivals and internal migration, each of which answers a different empirical question.

Some recent papers have uncovered significant heterogeneity in economic determinants of migration to/from Türkiye. While Çınar and Özbek (2026) document different importance of economic and institutional factors for specific origins and destinations, Gündoğmuş et al. (2026) look at the impact of income, employment, and industry structure on migration inflows to regions, and Kanca (2025) analyzes long-run connections with income, taxation, social assistance, economic freedom, unemployment, and health spending. Such studies confirm the usefulness of an economic approach, yet the nature of their questions on explanation and long-term dynamics contrasts the choice of a concise model for good forecasting performance.

The opposite way, from migration to macroeconomic outcomes, is also important for understanding the associations. Tutar et al. (2024) analyze the impacts on growth, unemployment, and inflation; Tanrıkulu (2021) talks about both demand-side contributions and integration problems; and Kaya and Aydın (2021) distinguish between short- and longer-run impacts of Syrian migration. Migration in combination with information and communication technologies is addressed by Ari (2021) in the context of growth. Using a spatial panel dataset, Öztop and Köse (2026) focus on regional growth and the education level of immigrants. Internationally, Ridwan et al. (2026) find negative growth correlations due to fiscal and labor market issues. Different outcomes, aggregations, and identification procedures prevent considering these results as estimates of one relationship.

This study asks whether regularization improves the prediction of Türkiye's annual net migration relative to an unpenalized linear model, and whether the selected predictors remain stable across successive evaluation windows. LASSO performs coefficient shrinkage and variable selection, while Elastic-Net combines two penalty types (Tibshirani, 1996; Zou and Hastie, 2005). The contribution is a joint comparison of predictive accuracy, model simplification and selection frequency under a common evaluation period. It does not claim to identify causal migration determinants or to establish an exhaustive ranking of all potentially relevant economic variables.

MATERIAL AND METHODS

Data and Variables

The study uses annual observations for Türkiye covering 1980-2025, corresponding to 46 calendar years before losses caused by lag construction or differencing. Net migration, measured in persons, is the response variable. Positive values represent net inward migration and negative values represent net outward migration. The predictor definitions and institutional data sources are summarized in Table 1. Logarithmic transformations are used for inflation, the public expenditure share and unemployment; the exchange-rate predictor enters as a first difference of its logarithm.

Table 1. Variables and transformations used in the reported model

Variable	Definition or transformation	Data source
$MIG\ (y_t)$	Annual net migration, persons	World Bank
GRW	Annual GDP growth rate, %	World Bank
$lnINF$	Natural logarithm of annual consumer inflation, %	World Bank
$lnPEXP$	Natural logarithm of public expenditure / GDP, %	IMF
$lnUNEMP$	Natural logarithm of unemployment rate, ages 15-64, %	IMF
$\Delta lnER$	Change in log USD/TRY buying exchange rate	CBRT
y_{t-1}	Net migration in the previous year	Constructed
$syria_dummy\ (D_t)$	1 in 2011-2017; 0 otherwise	Constructed

Note: IMF = International Monetary Fund; CBRT = Central Bank of the Republic of Türkiye. Net migration is a flow balance, not a refugee population stock.

Stationarity was assessed with the Fourier KPSS (FKPSS) test of Becker et al. (2006), whose null hypothesis is stationarity around a constant and which approximates smooth structural change in the deterministic component with a single-frequency Fourier component. The frequency k was chosen by minimizing the sum of squared residuals, and the joint significance of the Fourier terms was examined with an F test (5% critical value 4.929). Where the Fourier terms were significant, the FKPSS statistic was used; otherwise the standard KPSS test (Kwiatkowski et al., 1992) was applied. All decisions are made at the 5% level. The null of stationarity is not rejected for net migration, growth, inflation, public expenditure and unemployment, but it is rejected for the logarithm of the exchange rate (FKPSS = 0.3919 > 0.1720), whereas its first difference is stationary (FKPSS = 0.1128 < 0.1720). The reported specification therefore differences the exchange rate. Critical values are those tabulated by Becker et al. (2006) for T = 100 and are approximate for the sample of 46 annual observations (45 after differencing). Log transformation alone does not establish stationarity or eliminate heteroskedasticity.

The 2011-2017 indicator is a finite-period control for an exceptional migration episode, not a permanent post-2011 shift. Its contextual motivation is consistent with research on the Syrian displacement episode and regional labour-market adjustment (Del Carpio and Wagner, 2015; Bahcekapili and Cetin, 2015). It does not directly measure the number of Syrian refugees.

Model Specification and Regularization

Let y_t denote net migration, X_t the vector containing GRW, lnINF, lnPEXP, lnUNEMP, ΔlnER and y_{t-1}, and D_t the 2011-2017 indicator. The linear prediction specification is:

$$y_t = \beta_0 + X_t'\beta + \gamma D_t + \varepsilon_t \quad (1)$$

The intercept is β_0, β contains the coefficients of the six candidate predictors, γ is the period-indicator coefficient, and ε_t is the disturbance. OLS minimizes the residual sum of squares. For the penalized models, the fitted coefficients minimize the following objective, with the indicator and intercept excluded from the penalty:

$$Q = \frac{1}{2n}\sum_t (y_t - \beta_0 - X_t'\beta - \gamma D_t)^2 + \lambda\left[\alpha\sum_j |\beta_j| + \frac{1-\alpha}{2}\sum_j {\beta_j}^2\right] \quad (2)$$

Here, n is the training-sample size, $\lambda \geq 0$ controls shrinkage and $\alpha \in [0,1]$ determines the balance between the absolute-value and squared-coefficient penalties. LASSO sets α = 1 and can set coefficients exactly to zero (Tibshirani, 1996). Elastic-Net combines the penalties, which can be useful when predictors share information (Zou and Hastie, 2005). Regularization trades some estimation bias for lower variance; it does not guarantee improved accuracy in every sample (Hastie et al., 2009). Estimation was conducted in R 4.4 using glmnet, whose coordinate-descent implementation supports predictor-specific penalty factors (Friedman et al., 2010).

For both penalized estimators, λ is selected using ten-fold cross-validation within each training sample. The indicator has penalty $factor = 0$ and remains available in every specification. Elastic-Net searches $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ separately in each outer evaluation window. The average selected α across the ten windows is 0.48, whereas the full-sample Elastic-Net fit selects $\alpha = 0.7$. These summarize different fitting stages and should not be conflated.

Temporal Evaluation and Performance Measures

The evaluation procedure relies on increasing the training period starting with 1981-2015 (35 observations) and then including next ten held out annual observations from 2016 to 2025. For each year t of the test set, the model is estimated on training observations prior to year t and then tested for year t with the training set increasing by one observation. The coefficient estimates for the full sample period are analyzed separately and are not used as a base for calculation of the test errors. Separation of estimation and evaluation stages according to the out-of-sample logic as explained in Tashman (2000).

Since there are contemporaneous macroeconomic variables included in the specification, the evaluation involves conditional predictions for the held-out years. This does not mean that the forecast was performed before the realization of the covariates. Also, the temporally ordered outer split does not ensure that the ten-fold splits used in the inner loop are temporally ordered.

Prediction accuracy is summarized by root mean squared error (RMSE), mean absolute error (MAE) and a test-period coefficient of determination. With $m = 10$ test observations, $e_t = y_t - \hat{y}_t$ and $\overline{y}_{test}$ denoting the test-period mean, the measures are:

$$RMSE = \sqrt{\frac{1}{m}\sum_t {e_t}^2}, \quad MAE = \frac{1}{m}\sum_t |e_t|$$

$$R^2_{test} = 1 - \frac{\sum_t {e_t}^2}{\sum_t (y_t - \overline{y}_{test})^2} \quad (3)$$

RMSE and MAE are expressed in persons. Negative R^2 means that squared prediction errors exceed the dispersion around the realized test mean; that mean is a retrospective reference, not an implementable forecast benchmark. Reported Ljung-Box p-values, computed at lag 1 without a degrees-of-freedom adjustment on the ten out-of-sample test errors, assess remaining serial correlation (Ljung and Box, 1978). Variable-selection frequency counts the number of windows in which a penalized coefficient is

nonzero. This is descriptive stability evidence, rather than a test of causal relevance or a guarantee of model-selection consistency (Zhao and Yu, 2006).

RESULTS

Predictive Performance

Table 2 compares the three estimators on the same 2016-2025 evaluation period. LASSO records the lowest RMSE and MAE and the highest test R^2. Relative to OLS, its RMSE is lower by 50,546 persons, or 16.21%, and its MAE is lower by 18,043 persons, or 9.33%. Elastic-Net reduces RMSE by 13.62% relative to OLS. The LASSO-Elastic-Net difference is smaller: LASSO's RMSE is 2.99% lower. These are descriptive differences in the observed evaluation period; no formal test of predictive-accuracy differences is reported.

Table 2. Expanding-window prediction performance in 2016-2025

Model	RMSE (persons)	MAE (persons)	R^2	Ljung-Box p
LASSO (α = 1)	261,248	175,320	0.267	0.333
Elastic-Net (mean α = 0.48)	269,313	180,991	0.221	0.602
OLS	311,794	193,363	-0.044	0.323

Note: Ten annual test observations. Commas separate thousands and points indicate decimals. The Elastic-Net mixing parameter was selected separately in each window.

The RMSE and MAE comparison is visualized in Figure 1.

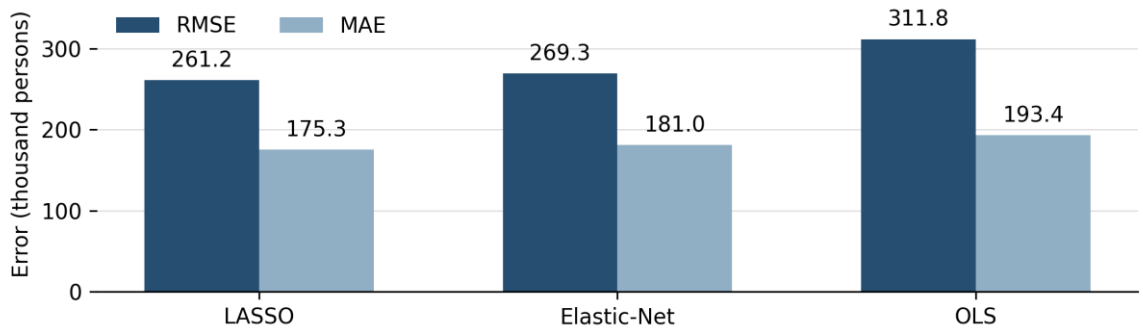


Figure 1. Prediction errors for the three estimators during 2016-2025.

OLS has negative test R^2, whereas both penalized models have positive values. Nevertheless, LASSO's R^2 of 0.267 leaves substantial unexplained variation. All Ljung-Box p-values exceed 0.05: the tests do not reject the no-autocorrelation null, but this does not establish error independence in a short evaluation period.

Model Simplification

Table 3 reports the sequence of LASSO specifications. Removing the Fourier terms while retaining the second lag raises RMSE from 274,650 to 280,438 persons, an increase of 2.11%. Subsequently removing the second lag reduces RMSE to 261,248 persons, a 6.84% improvement relative to the preceding

specification. The final specification improves on the initial one by 4.88%. Thus, the first simplification step alone does not improve performance; the net gain arises after the lag structure is also simplified.

Table 3. LASSO performance across successive specifications

Specification	RMSE (persons)
Fourier terms and second lag y_{t-2} included	274,650
Fourier terms excluded; y_{t-2} retained	280,438
Fourier terms and y_{t-2} excluded; final specification	261,248

Note: The first lag and period indicator remain in the final candidate specification. The comparison is descriptive and does not identify separate causal contributions of the removed terms.

The sequence is compatible with a bias-variance trade-off (Hastie et al., 2009). Its implications depend on the remaining predictors and evaluation period; it does not establish that Fourier terms are generally unhelpful. Repeated comparisons on the same test period remain exploratory without a separate final assessment.

Full-Sample Coefficients

Table 4 presents full-sample coefficients on the reported predictor-standardized scale. LASSO retains public expenditure, unemployment, lagged migration and the unpenalized indicator. Growth, inflation and exchange-rate change receive zero coefficients. Elastic-Net additionally retains growth and inflation with negative coefficients. A zero penalized coefficient indicates exclusion at the selected penalty, not the absence of an economic relationship.

Table 4. Reported full-sample coefficient estimates

Predictor	LASSO	Elastic-Net	OLS
GRW	0.00	-4,518.91	-13,468.86
$lnINF$	0.00	-4,889.42	-35,673.66
$lnPEXP$	-10,256.45	-16,563.65	-27,866.27
$lnUNEMP$	6,858.17	9,185.12	5,378.92
$\Delta lnER$	0.00	0.00	13,849.36
y_{t-1}	109,864.06	112,807.06	124,848.78
$syria_dummy$	161,469.53	156,916.42	138,202.95

Note: The estimates are reported on a predictor-standardized scale with the response in persons; they are not unitless fully standardized effects. Elastic-Net uses $\alpha = 0.7$ in the full-sample fit. No coefficient standard errors or significance tests are reported.

The positive lag coefficient indicates a predictive association with prior-year migration, rather than a directly interpretable persistence parameter on the original scale. The indicator also has a positive coefficient, reflecting a conditional level association during its coded period. Unemployment enters positively and public expenditure negatively in all three fits. Because these estimates arise from an observational prediction specification, neither their signs nor their magnitudes establish policy effects.

In particular, a positive unemployment coefficient for net inward migration cannot simply be labelled an emigration push effect.

Selection Frequency Across Evaluation Windows

Table 5 reports LASSO selection frequencies. Public expenditure, unemployment and lagged net migration are nonzero in nine of the ten windows; growth appears in six, exchange-rate change in two and inflation in one. Figure 2 displays the frequencies for the six penalized predictors. The period indicator is shown separately in the table because its inclusion is imposed through a zero penalty factor.

Table 5. LASSO nonzero-coefficient frequency across ten Windows

Predictor	Windows	Frequency (%)	Treatment
$lnPEXP$	9 / 10	90	Penalized
$lnUNEMP$	9 / 10	90	Penalized
y_{t-1}	9 / 10	90	Penalized
GRW	6 / 10	60	Penalized
$\Delta lnER$	2 / 10	20	Penalized
$lnINF$	1 / 10	10	Penalized
$syria_dummy$	10 / 10	100	Unpenalized control

Note: The indicator's 100% inclusion is imposed by the model design and is not comparable to the penalized predictors' selection frequencies. Overlapping training windows are not independent replications.

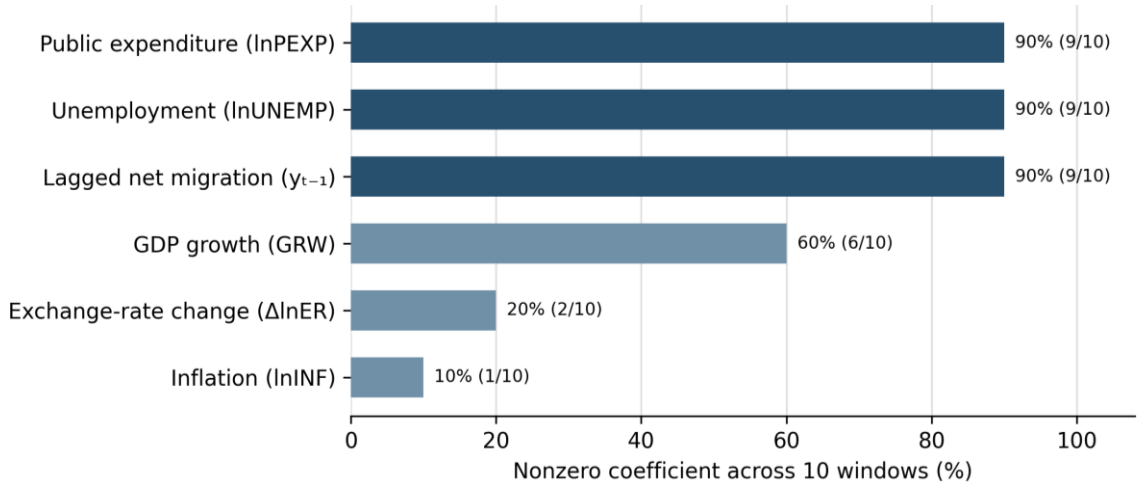


Figure 2. Selection frequency of the six penalized predictors across ten evaluation windows.

Economic Growth's 60% selection frequency and its zero coefficient in the full-sample LASSO fit are not contradictory: the training samples and selected penalties differ. Similarly, low frequencies for inflation and exchange-rate change indicate limited incremental use in these fitted specifications, not universal irrelevance. The frequencies describe sensitivity to an expanding information set and should not be read as selection probabilities from independent samples (Zhao and Yu, 2006).

DISCUSSION AND CONCLUSION

The main empirical finding is the reduction in prediction error achieved by the penalized estimators relative to OLS across the ten evaluation years. LASSO yields the best performance of all, using a relatively parsimonious full-sample specification. Such a result is in line with the intuition behind the use of shrinkage in small samples when the unpenalized estimators tend to react strongly to variation in the sample (Tibshirani, 1996; Hastie et al., 2009). Yet, the small gap in error between LASSO and Elastic-Net and the lack of confidence intervals make it impossible to claim statistically proven superiority. The data proves LASSO to be the best-performing estimator in this particular case, not the most suitable migration model in general.

The frequent choice of lagged migration is consistent with persistence due to networks and the learning and adjustment process. The role of social networks is described by Rainer and Siedler (2009), and the influence of networks on the migration selection process is demonstrated by McKenzie and Rapoport (2010). It can serve as an alternative explanation next to the larger theory proposed by Massey et al. (1993); nevertheless, it is not possible to single out the effect of networks with an aggregate lag coefficient. Serially persistent omitted variables and measurement practices could also contribute.

The unemployment association requires particular care. Research on Syrian displacement produces heterogeneous labour-market findings: Akgündüz et al. (2015) report limited displacement of native employment in their regional analysis, whereas Serttaş and Uluöz (2021) find a negative long-run relationship between unemployment and the number of Syrian refugees. Del Carpio and Wagner (2015) examine adjustment across labour-market segments. Günaydın and Çetin (2015), studying youth unemployment, also demonstrate its connections with broader macroeconomic conditions. These studies use different outcomes and populations from annual net migration. Their findings therefore cannot validate the positive unemployment coefficient as a causal migration response.

Inflation-related evidence likewise depends on the market and migration measure. Balkan and Tumen (2016) find consumer-price declines linked to informal labour-cost channels, while Tanrıkulu (2021) discusses short-run inflation pressure. Regional inflow analysis by Gündoğmuş et al. (2026) addresses yet another direction of association. The low selection frequency of inflation in the present exercise may reflect overlapping information, the particular evaluation period or penalty strength. It does not reconcile these different mechanisms, and it does not establish that inflation is unrelated to migration.

The negative public-expenditure coefficient should not be translated into a recommendation to increase or reduce spending to control migration. Kocaman (2026) investigates how refugee migration affects public expenditure and tax revenues, while Kanca (2025) examines several fiscal and social-policy correlates of international migration. Together, these perspectives make reverse causation and fiscal composition relevant. The aggregate expenditure-to-GDP ratio used here cannot distinguish education, health, social protection or other spending channels. Regional and institutional heterogeneity also matters, as emphasized by Çınar and Özbek (2026), Öztop and Köse (2026) and Özyürek (2025).

Several limitations constrain generalization. The evaluation contains only ten annual observations, and adjacent training windows overlap extensively. The use of contemporaneous covariates requires attention to publication delays and revisions before the model can support real-time decisions. The 2011-2017 indicator encodes a retrospectively chosen interval, so its end date cannot be assumed to have been known at earlier forecast origins. In addition, ten-fold inner validation needs explicit temporal blocking

to support a strict time-order interpretation. Reusing the evaluation period for specification simplification would make the reported ranking exploratory. These limitations favour further assessment with an untouched later period, fully specified preprocessing and simple operational benchmarks such as previous-year migration.

In conclusion, LASSO reduces RMSE by 16.21% relative to OLS in the reported 2016-2025 comparison, with public expenditure, unemployment and lagged migration recurring among the penalized predictors. The findings demonstrate the value of evaluating accuracy and selection frequency together. Their practical implication is to retain parsimonious prediction models as candidates for monitoring aggregate migration dynamics, while evaluating data availability, temporal stability and forecast uncertainty before deployment. Causal policy conclusions require a separate identification strategy.

References

Akgündüz YE, Van den Berg M, Hassink W, 2015. The impact of refugee crises on host labor markets: The case of the Syrian refugee crisis in Turkey. IZA Discussion Paper No. 8841.

Ari YO, 2021. The relationship between information and communication technology (ICT), international migration and economic growth in Turkey. Alanya Akademik Bakış, 5(3): 1169-1180.

Bahcekapili C, Cetin B, 2015. The impacts of forced migration on regional economies: The case of Syrian refugees in Turkey. *International Business Research*, 8(9): 1.

Balkan B, Tumen S, 2016. Immigration and prices: Quasi-experimental evidence from Syrian refugees in Turkey. *Journal of Population Economics*, 29(3): 657-686.

Becker R, Enders W, Lee J, 2006. A stationarity test in the presence of an unknown number of smooth breaks. *Journal of Time Series Analysis*, 27(3): 381-409.

Çınar M, Özbek Ö, 2026. The impact of economic and institutional factors on international migration flows: Evidence from Turkey. *Panoeconomicus*, 73(4): 617-636.

Del Carpio XV, Wagner M, 2015. The impact of Syrian refugees on the Turkish labor market. World Bank Policy Research Working Paper No. 7402.

Friedman J, Hastie T, Tibshirani R, 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1): 1-22.

Günaydın D, Çetin M, 2015. Genç işsizliğin temel makroekonomik belirleyicileri: Ampirik bir analiz. *Pamukkale Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 22: 17-34.

Gündoğmuş B, Bayir M, Sarihan AY, 2026. Economic drivers of international migration to Türkiye: A regional comparison study. *Makara Human Behavior Studies in Asia*, 30(2): 69-79.

Hastie T, Tibshirani R, Friedman J, 2009. The elements of statistical learning: Data mining, inference, and prediction. 2nd ed., Springer.

Kanca OC, 2025. Macroeconomic determinants of migration: An empirical analysis for Türkiye. *Baltic Journal of Economic Studies*, 11(3): 296-306.

Kaya O, Aydın Ü, 2021. Economic impact of Syrian refugees on Turkish economy. *Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 23(3): 1151-1171.

Kirişci K, 2007. Turkey: A country of transition from emigration to immigration. *Mediterranean Politics*, 12(1): 91-97.

Kocaman M, 2026. Fiscal effects of the refugee migration on the asylum country: Evidence from Türkiye, world's leading refugee host country. *Current Perspectives in Social Sciences*, 30(1): 146-164.

Kwiatkowski D, Phillips PCB, Schmidt P, Shin Y, 1992. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54(1-3): 159-178.

Lewis J, Swannell M, 2018. The macroeconomic determinants of migration. Bank of England Staff Working Paper No. 729.

Ljung GM, Box GEP, 1978. On a measure of lack of fit in time series models. *Biometrika*, 65(2): 297-303.

Massey DS, Arango J, Hugo G, Kouaouci A, Pellegrino A, Taylor JE, 1993. Theories of international migration: A review and appraisal. *Population and Development Review*, 19(3): 431-466.

McKenzie D, Rapoport H, 2010. Self-selection patterns in Mexico-U.S. migration: The role of migration networks. *The Review of Economics and Statistics*, 92(4): 811-821.

Öztop AO, Köse T, 2026. Analysing the impact of migration flows on regional per capita GDP in Türkiye: A spatial panel data approach. *Public Sector Economics*, 50(2): 259-286.

Özyürek M, 2025. Biopolitical borders and the political economy of migration flows to and from Türkiye. *Studies in Ethnicity and Nationalism*, 25(3): 276-291.

Rainer H, Siedler T, 2009. The role of social networks in determining migration and labour market outcomes: Evidence from German reunification. *Economics of Transition*, 17(4): 739-767.

Ridwan M, Antor ZA, Ko J, Leung CK, Doğan M, Ming WK, 2026. Refugee inflows and macroeconomic performance: Evidence from a cross-country panel (2000-2023). *Frontiers in Human Dynamics*, 8: 1723467.

Serttaş FÖ, Uluöz D, 2021. The impact of Syrian migration on unemployment: Evidence from Turkey. *Adam Academy Journal of Social Sciences*, 11(1): 1-30.

Simpson NB, 2022. Demographic and economic determinants of migration. IZA World of Labor, 373.

Tanrıkulu F, 2021. The political economy of migration and integration: Effects of immigrants on the economy in Turkey. *Journal of Immigrant & Refugee Studies*, 19(4): 364-377.

Tashman LJ, 2000. Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, 16(4): 437-450.

Tibshirani R, 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1): 267-288.

Tutar H, Mutlu HT, Kantarcioğlu N, Lakomy-Zinowik M, 2024. Analysis of migration to Turkey through macroeconomic indicators: Evidence from the period 2004-2024. *Economics & Sociology*, 17(4): 257-271.

Zhao P, Yu B, 2006. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7: 2541-2563.

Zou H, Hastie T, 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2): 301-320.

Current Problems and Future Directions in Industrial Robot Control

Lala Babayeva[1]*

[1] Mingachevir State University, Azerbaijan

*Corresponding author e-mail: lala.babayeva.itm25@mdu.edu.az

Abstract

Industrial robots are now widely used for tasks such as welding, assembly, machining, material handling and quality inspection. Their performance depends not only on the control program, but also on how the mechanical parts, actuators, sensors and controller work together. This paper examines the main mechanical and functional elements of industrial robots and explains their roles during operation. It draws on published studies about robot kinematics and dynamics, motion planning, robust control and smart manufacturing. The discussion covers the base, manipulator links, joints, transmission mechanisms and end-effector, as well as the sensors, actuators and control unit. It follows the main stages of a typical task: defining the target, planning a path, sending commands to the actuators, carrying out the movement and using sensor feedback to correct errors. The reviewed literature suggests that accuracy and stability are achieved when mechanical design and control are considered together. The robot's kinematic structure limits the movements it can make, while payload, speed, external forces and dynamic behaviour affect the control requirements. Sensors report information about the robot and its surroundings, and actuators turn control commands into physical movement. The paper also considers digital manufacturing, collaborative robots, digital twins and intelligent control. The main conclusion is that a robot should be treated as one connected system, not as a collection of independent parts. Further research could examine adaptive control, testing with simulation and the coordination of robots working in changing production environments.

Keywords*: Industrial robots, mechanical structure, kinematics, sensors, control systems*

INTRODUCTION

Industrial automation has expanded considerably, and robots now carry out many tasks in manufacturing. They are used for welding, assembly, material handling, machining and other operations where repeatability and accuracy matter [2], [3]. However, a robot is useful only if it can repeat its programmed movements reliably and respond appropriately when working conditions change.

This reliability depends on the robot's physical construction as well as its software. The base, links, joints and transmission mechanisms support and guide movement, while the end-effector performs the required operation. These parts need to work together for the robot to move as intended [11], [12]. Kinematic analysis helps explain how joint movements determine the position of the end-effector [5], [13].

Sensors and the control unit are equally important. Sensors provide information about the robot's current state, and actuators make the commanded movements happen. For this reason, accurate operation requires the mechanical and functional parts to be designed and adjusted as a coordinated system [4], [6], [7]. Developments in Industry 4.0 and Industry 5.0 are also connecting robots more closely with digital production systems [3], [10].

The problem addressed in this paper is that the main parts of an industrial robot cannot be assessed properly in isolation. A change in a joint, transmission, actuator, sensor or controller can affect the behaviour of the whole machine. For example, the robot's mechanical arrangement determines which positions and orientations its tool can reach. It is therefore useful to study mechanical design and control together.

The aim of this paper is to examine the mechanical and functional structure of industrial robots and explain how the main components work together. The discussion describes the robot's basic structure, considers its joints and transmission mechanisms, introduces degrees of freedom, and reviews the roles of actuators, sensors and control units. It also outlines how industrial robots are used in modern manufacturing [2], [3], [5], [10].

LITERATURE REVIEW AND THEORETICAL FRAMEWORK

Research on industrial robots covers mechanical design, kinematics, dynamics, control, motion planning and the use of robots in modern production. Mikolajczyk et al. reviewed applications and emerging trends in industrial robotics [2]. Liu et al. focused on robot learning and its contribution to smart manufacturing [3]. These studies show that the role of robots is expanding beyond fixed, repetitive tasks.

Kinematics is central to robot modelling because it links the movement of the joints to the position and orientation of the tool. Pupăză discussed how a general forward-kinematic model can help compensate for positioning errors in serial industrial robots [5]. Elhadidy et al. examined the kinematics of a six-degree-of-freedom manipulator [14]. Bingül and Küçük introduced core concepts of robot kinematics, while Bayraktaroğlu discussed mathematical modelling and related problems [11], [13].

Control and trajectory planning have also received considerable attention. Son et al. reviewed robust control methods for robot manipulators in future manufacturing [4], and Mazhar et al. surveyed robust control and trajectory-planning strategies in the context of Industry 4.0 [6]. Liu considered motion planning in environments where conditions may change [8]. Together, these works highlight that reaching a target is not enough: the robot must follow a suitable path while respecting physical limits and responding to its surroundings.

Other studies focus on robot dynamics and cooperation between manipulators. Mazzini et al. reviewed dynamic modelling and control for cooperative robots [7]. Shen et al. proposed an optimization framework that considers both kinematic and dynamic factors in robot-system design [9]. Zisgen and Hinrichs explored multi-agent reinforcement learning for controlling gantry robot systems [1].

More recent work connects robotics with digital twins, collaborative systems and Industry 5.0. Zafar et al. discussed how these technologies can support smart manufacturing [10]. Taken together, the

literature suggests that robot performance is best understood by looking at the mechanical structure, sensing, motion generation and control as parts of the same system.

MECHANICAL AND FUNCTIONAL STRUCTURE OF INDUSTRIAL ROBOTS

Mechanical structure

The mechanical structure gives the robot its physical form and carries the loads involved in movement. In most industrial robots, the main parts include a base, manipulator links, joints, transmission mechanisms and an end-effector. The base supports the robot and fixes it to a floor, production cell or mobile platform. Its rigidity and mounting arrangement can influence stability and vibration.

The manipulator is made up of links connected by joints. A joint may allow rotation or linear movement, depending on the robot's design. The arrangement of the links and joints determines the robot's reach, working area and possible tool orientations. Many robots use a serial arrangement, although parallel and hybrid designs are also available. The choice depends on the task, accuracy, payload and available space [5], [11], [13].

Transmission mechanisms carry movement and torque from the actuator to the joint. Depending on the design, they may use gears, belts, screws or other components. Backlash, friction, flexibility and wear can all reduce positioning accuracy or repeatability. This is why transmission selection and regular maintenance matter when the performance of the whole robot is considered.

The end-effector is the part that directly handles a workpiece or interacts with the surroundings. It may be a gripper, welding tool, machining tool, suction device or another tool chosen for a specific job. Its weight, shape and operating characteristics affect the load carried by the robot and the way the manipulator moves.

Functional structure

The functional part of a robot includes the controller, sensors, actuators, communication interfaces and software. The controller processes the task and sensor information, calculates the required movement and sends commands to the actuators. The software defines movement sequences and operating limits, and it can also link the robot to other equipment in the production line.

Actuators turn control commands into movement. Electric motors are widely used, although hydraulic or pneumatic systems may be suitable for some tasks. The actuator must match the joint, transmission and expected load. Its response speed and precision have a direct effect on how smoothly and accurately the robot moves.

Sensors measure, or help estimate, the robot's state and its interaction with nearby objects. Position and velocity sensors are used to monitor joint movement; force and torque sensors provide information about contact; and vision systems can help locate objects. These measurements are useful only when they are reliable and reach the controller in time.

Interaction between components

The components are connected through a continuing exchange of commands and measurements. The controller sends a command, the actuators move the joints, and the sensors report what happened. The controller then compares the measured state with the target and adjusts its commands if necessary. This feedback process links the physical robot to its control functions.

Performance may fall when the mechanical design, actuator capacity, sensor accuracy and control settings do not match. For instance, a flexible structure or backlash in a transmission can cause positioning errors that cannot always be fixed by changing the controller settings alone. Likewise, delayed or noisy sensor readings can make corrections less effective. These examples show why the robot should be designed and tested as a complete system.

OPERATING PRINCIPLE OF AN INDUSTRIAL ROBOT

An industrial robot usually carries out a task through several connected stages: planning, command generation, movement and feedback-based correction. First, the task and the required end-effector position are defined. The controller then plans a path and calculates the joint settings needed to reach the target. The plan must take account of the robot's workspace, speed and acceleration limits, payload and possible collisions.

Once the movement has been calculated, the controller sends commands to the actuators. The actuators produce movement, which passes through the transmission mechanisms to the joints and links. When the joints move in a coordinated way, the end-effector follows the intended path.

While the robot is working, sensors measure its position, speed or contact conditions. The controller compares these readings with the target values and adjusts the commands if an error appears. How well this correction works depends on the model, actuator response, sensor quality and chosen control method [4], [6], [8].

Robot operation is therefore more than a single mechanical action. It depends on task planning, kinematic calculations, control commands, actuation and feedback working together. This is especially important when production conditions change or when robots share a workspace with other machines.

KINEMATIC AND DYNAMIC CHARACTERISTICS

Kinematics describes movement without directly examining the forces that cause it. Forward kinematics uses the joint variables to calculate the end-effector's position and orientation. Inverse kinematics works in the opposite direction: it finds joint settings that can produce a required tool position and orientation. Degrees of freedom describe the independent movements available to the robot. Their number and arrangement influence reach, flexibility and tool orientation [5], [11], [13].

Kinematic models are used to plan trajectories and estimate where the tool will be. If link dimensions, joint calibration or model parameters are inaccurate, the calculated position may differ from the actual one. Calibration and compensation can help reduce this difference, as discussed in research on forward-kinematic models for serial robots [5].

Dynamics examines the forces and torques needed to produce movement. The payload, mass of the links, acceleration, friction and contact forces all affect how much effort the actuators must provide. A

path may be possible from a kinematic point of view but still demand too much torque or cause vibration. Robot design and control should therefore consider both kinematics and dynamics [7], [9].

CONTROL AND SENSOR SYSTEMS

The control system coordinates movement and checks whether the robot is following the task. In a typical closed-loop system, the controller compares the required state with sensor measurements and uses the difference to update actuator commands. Depending on the task, it may control joint position, velocity, torque or the position of the end-effector.

Position and velocity measurements help the robot follow a planned path. Force and torque sensors are useful for assembly and other tasks involving contact, while vision sensors can locate parts that are not always placed in the same position. The right sensor combination depends on the accuracy required, the working environment, processing time and cost.

Robust control can help a robot remain stable when the model is imperfect, the load changes or disturbances occur [4], [6]. If the environment is dynamic, motion planning must also account for obstacles and changing conditions [8]. Even a well-designed controller has limits: mechanical quality, calibration and sensor reliability still matter.

APPLICATIONS IN MODERN MANUFACTURING

Industrial robots are used for welding, assembly, material handling, machining, packaging and inspection. Their ability to repeat movements can improve consistency, especially in tasks that are repetitive or physically demanding. Automated handling can also reduce the need for workers to enter some hazardous areas, but only when suitable safety measures are in place.

Different applications place different demands on the robot. Welding usually needs a steady path and consistent tool orientation. Assembly may require precise positioning and controlled contact forces, while material handling depends on payload capacity and secure gripping. These requirements influence the robot design, end-effector, sensor choice and control method.

In smart manufacturing, robots increasingly exchange information with other machines, planning tools and monitoring systems. Digital twins can be used to simulate and test production scenarios, while collaborative robotics focuses on work shared between people and robots under appropriate safeguards [10]. These developments make reliable communication and system-level planning more important.

CONTROL APPROACHES AND DEVELOPMENT PERSPECTIVES

Robot control has gradually moved beyond fixed movement sequences toward methods that can respond to changes in operating conditions. Conventional control is still suitable for many repetitive tasks in stable environments. Robust or adaptive methods may be more useful when load, friction or other conditions vary [4], [6]. The most appropriate method depends on the task; a newer method is not automatically the better choice.

Motion planning must connect the starting position to the target while respecting the robot's physical limits and avoiding obstacles. When the surroundings change, the robot may need to update its path

using new sensor information [8]. The result depends on the quality of that information and on the time and computing resources available.

Artificial intelligence and multi-agent methods may be useful when several robots or production units need to coordinate their actions. Multi-agent reinforcement learning has been studied for robot-system control [1], and research on cooperative manipulators emphasizes the importance of modelling and coordination [7]. Before these approaches are used in critical production tasks, their safety, stability, communication requirements and decision-making behaviour need to be tested carefully.

Simulation and digital twins can help engineers compare robot configurations and control strategies before putting them into production. However, a simulation should be checked against real measurements because a simplified model may not represent friction, wear, delays or uncertainty accurately. Industry 4.0 and Industry 5.0 point toward production systems that combine physical automation with digital monitoring and greater attention to people [3], [10].

RESEARCH METHODOLOGY

This paper uses a literature-based, system-oriented approach. First, it identifies the main mechanical parts of an industrial robot—the base, links, joints, transmission mechanisms and end-effector—and considers what each contributes to movement. These parts are then examined in relation to the robot as a whole.

The functional elements are considered next, with particular attention to sensors, actuators, the controller and software. The analysis follows the flow of information from sensors to the controller and then from control commands to mechanical movement. The operating process is organized into five stages: defining the task, calculating the movement, generating commands, carrying out the movement and correcting errors using feedback.

A comparative approach is used to distinguish the roles of the main components and control methods. The findings are then considered from a system-level perspective to identify links between mechanical design, kinematics, dynamics, sensing and control. This paper synthesizes published research; it does not report new experimental measurements or claim that a physical prototype has been tested.

CONCLUSION AND RECOMMENDATIONS

The review indicates that the mechanical and functional parts of an industrial robot are closely connected. The manipulator, joints, actuators and transmission mechanisms produce movement; sensors report the robot's state; and the controller uses this information to regulate its actions. As a result, overall performance depends on how these elements work together, not simply on the quality of one component.

The robot's kinematic structure determines the movements and tool positions it can reach. Dynamic properties, payload and external forces affect the effort and stability needed to follow a trajectory. Sensor accuracy, actuator response and controller design also influence how well the robot follows its target and corrects errors. Digital technologies, collaborative robots and intelligent control offer useful possibilities, but reliability, safety and practical validation remain important.

Based on the reviewed studies, the following recommendations can be made:

1. Assess the mechanical design together with the requirements of the sensors, actuators and controller.

2. Choose position, velocity and force measurements according to the task to support accurate movement and reliable operation.

3. Evaluate robust and adaptive control methods where loads or working conditions are likely to change.

4. For systems with several robots, consider coordination and information exchange alongside collision prevention and safety.

5. Compare simulation and digital-twin results with real production data before using them to make operational decisions.

6. Further study how artificial intelligence and multi-agent control can be combined with the mechanical and functional parts of industrial robots.

References

1. Zisgen, H., & Hinrichs, J. F. (2025). Multi-agent reinforcement learning for controlling gantry robot systems. Production Engineering, 19, 871–884.
2. Mikolajczyk, T., et al. (2022). Advanced applications of industrial robotics: New trends and possibilities. Applied Sciences, 12(1), 135.
3. Liu, Z., Liu, Q., Xu, W., Wang, L., & Zhou, Z. (2022). Robot learning towards smart robotic manufacturing: A review. Robotics and Computer-Integrated Manufacturing, 77, 102360.
4. Son, J., Kang, H., & Kang, S. H. (2023). A review on robust control of robot manipulators for future manufacturing. International Journal of Precision Engineering and Manufacturing, 24, 1083–1102.
5. Pupăză, C. (2023). The importance of embedding a general forward kinematic model for industrial robots with serial architecture in order to compensate for positioning errors. Mathematics, 11(10), 2306.
6. Mazhar, A., Tanveer, A., Izhan, M., & Khan, M. Z. T. (2023). Robust control approaches and trajectory planning strategies for industrial robotic manipulators in the era of Industry 4.0: A comprehensive review. Engineering Proceedings, 56(1), 75.
7. Mazzini, M., et al. (2023). Cooperative robot manipulators dynamical modeling and control: An overview. Dynamics, 3(4), 820–854.
8. Liu, Y. (2024). Review on motion planning of robotic manipulator in dynamic environments. Journal of Sensors, 2024, 5969512.
9. Shen, S., Zhang, D., Han, X., Jiang, C., & Li, Q. (2024). A robust optimization framework for design of robotic system with kinematic and dynamic criteria. Computer Methods in Applied Mechanics and Engineering, 423, 116866.
10. Zafar, M. H., Langås, E. F., & Sanfilippo, F. (2024). Exploring the synergies between collaborative robotics, digital twins, augmentation, and Industry 5.0 for smart manufacturing: A state-of-the-art review. Robotics and Computer-Integrated Manufacturing, 89, 102769.
11. Bingül, Z., & Küçük, S. (2015). Robot Kinematiği. Kocaeli: Umuttepe Yayınları. ISBN 978-605-510-047-6.

12. Kurtoğlu, A. (2011). Robot Tekniği. İstanbul: Papatya Bilim. ISBN 978-605-422-037-3.
13. Bayraktaroğlu, Z. Y. (2023). Robot Teknolojisine Giriş: Matematiksel Modelleme (2nd ed.). Ankara: Nobel Akademik Yayıncılık. ISBN 978-605-320-930-0.
14. ISO. (2021). ISO 8373:2021 Robotics — Vocabulary. International Organization for Standardization.
15. ISO. (2025). ISO 10218-1:2025 Robotics — Safety requirements — Part 1: Industrial robots.

Analysis of the Functional and Mechanical Infrastructure of Industrial Robots and Their Operating Principle

Lala Babayeva[1]*

[1] Mingachevir State University, Azerbaijan

Corresponding author e-mail: lala.babayeva.itm25@mdu.edu.az

Abstract

Industrial robots are now widely used for tasks such as welding, assembly, machining, material handling and quality inspection. Their performance depends not only on the control program, but also on how the mechanical parts, actuators, sensors and controller work together. This paper examines the main mechanical and functional elements of industrial robots and explains their roles during operation. It draws on published studies about robot kinematics and dynamics, motion planning, robust control and smart manufacturing. The discussion covers the base, manipulator links, joints, transmission mechanisms and end-effector, as well as the sensors, actuators and control unit. It follows the main stages of a typical task: defining the target, planning a path, sending commands to the actuators, carrying out the movement and using sensor feedback to correct errors. The reviewed literature suggests that accuracy and stability are achieved when mechanical design and control are considered together. The robot's kinematic structure limits the movements it can make, while payload, speed, external forces and dynamic behaviour affect the control requirements. Sensors report information about the robot and its surroundings, and actuators turn control commands into physical movement. The paper also considers digital manufacturing, collaborative robots, digital twins and intelligent control. The main conclusion is that a robot should be treated as one connected system, not as a collection of independent parts. Further research could examine adaptive control, testing with simulation and the coordination of robots working in changing production environments.

Keywords: *Industrial robots, mechanical structure, kinematics, sensors, control systems*

INTRODUCTION

Industrial automation has expanded considerably, and robots now carry out many tasks in manufacturing. They are used for welding, assembly, material handling, machining and other operations where repeatability and accuracy matter [2], [3]. However, a robot is useful only if it can repeat its programmed movements reliably and respond appropriately when working conditions change.

This reliability depends on the robot's physical construction as well as its software. The base, links, joints and transmission mechanisms support and guide movement, while the end-effector performs the required operation. These parts need to work together for the robot to move as intended [11], [12]. Kinematic analysis helps explain how joint movements determine the position of the end-effector [5], [13].

Sensors and the control unit are equally important. Sensors provide information about the robot's current state, and actuators make the commanded movements happen. For this reason, accurate operation requires the mechanical and functional parts to be designed and adjusted as a coordinated system [4], [6], [7]. Developments in Industry 4.0 and Industry 5.0 are also connecting robots more closely with digital production systems [3], [10].

The problem addressed in this paper is that the main parts of an industrial robot cannot be assessed properly in isolation. A change in a joint, transmission, actuator, sensor or controller can affect the behaviour of the whole machine. For example, the robot's mechanical arrangement determines which positions and orientations its tool can reach. It is therefore useful to study mechanical design and control together.

The aim of this paper is to examine the mechanical and functional structure of industrial robots and explain how the main components work together. The discussion describes the robot's basic structure, considers its joints and transmission mechanisms, introduces degrees of freedom, and reviews the roles of actuators, sensors and control units. It also outlines how industrial robots are used in modern manufacturing [2], [3], [5], [10].

LITERATURE REVIEW AND THEORETICAL FRAMEWORK

Research on industrial robots covers mechanical design, kinematics, dynamics, control, motion planning and the use of robots in modern production. Mikolajczyk et al. reviewed applications and emerging trends in industrial robotics [2]. Liu et al. focused on robot learning and its contribution to smart manufacturing [3]. These studies show that the role of robots is expanding beyond fixed, repetitive tasks.

Kinematics is central to robot modelling because it links the movement of the joints to the position and orientation of the tool. Pupăză discussed how a general forward-kinematic model can help compensate for positioning errors in serial industrial robots [5]. Elhadidy et al. examined the kinematics of a six-degree-of-freedom manipulator [14]. Bingül and Küçük introduced core concepts of robot kinematics, while Bayraktaroğlu discussed mathematical modelling and related problems [11], [13].

Control and trajectory planning have also received considerable attention. Son et al. reviewed robust control methods for robot manipulators in future manufacturing [4], and Mazhar et al. surveyed robust control and trajectory-planning strategies in the context of Industry 4.0 [6]. Liu considered motion planning in environments where conditions may change [8]. Together, these works highlight that reaching a target is not enough: the robot must follow a suitable path while respecting physical limits and responding to its surroundings.

Other studies focus on robot dynamics and cooperation between manipulators. Mazzini et al. reviewed dynamic modelling and control for cooperative robots [7]. Shen et al. proposed an optimization framework that considers both kinematic and dynamic factors in robot-system design [9]. Zisgen and Hinrichs explored multi-agent reinforcement learning for controlling gantry robot systems [1].

More recent work connects robotics with digital twins, collaborative systems and Industry 5.0. Zafar et al. discussed how these technologies can support smart manufacturing [10]. Taken together, the

literature suggests that robot performance is best understood by looking at the mechanical structure, sensing, motion generation and control as parts of the same system.

MECHANICAL AND FUNCTIONAL STRUCTURE OF INDUSTRIAL ROBOTS

Mechanical structure

The mechanical structure gives the robot its physical form and carries the loads involved in movement. In most industrial robots, the main parts include a base, manipulator links, joints, transmission mechanisms and an end-effector. The base supports the robot and fixes it to a floor, production cell or mobile platform. Its rigidity and mounting arrangement can influence stability and vibration.

The manipulator is made up of links connected by joints. A joint may allow rotation or linear movement, depending on the robot's design. The arrangement of the links and joints determines the robot's reach, working area and possible tool orientations. Many robots use a serial arrangement, although parallel and hybrid designs are also available. The choice depends on the task, accuracy, payload and available space [5], [11], [13].

Transmission mechanisms carry movement and torque from the actuator to the joint. Depending on the design, they may use gears, belts, screws or other components. Backlash, friction, flexibility and wear can all reduce positioning accuracy or repeatability. This is why transmission selection and regular maintenance matter when the performance of the whole robot is considered.

The end-effector is the part that directly handles a workpiece or interacts with the surroundings. It may be a gripper, welding tool, machining tool, suction device or another tool chosen for a specific job. Its weight, shape and operating characteristics affect the load carried by the robot and the way the manipulator moves.

Functional structure

The functional part of a robot includes the controller, sensors, actuators, communication interfaces and software. The controller processes the task and sensor information, calculates the required movement and sends commands to the actuators. The software defines movement sequences and operating limits, and it can also link the robot to other equipment in the production line.

Actuators turn control commands into movement. Electric motors are widely used, although hydraulic or pneumatic systems may be suitable for some tasks. The actuator must match the joint, transmission and expected load. Its response speed and precision have a direct effect on how smoothly and accurately the robot moves.

Sensors measure, or help estimate, the robot's state and its interaction with nearby objects. Position and velocity sensors are used to monitor joint movement; force and torque sensors provide information about contact; and vision systems can help locate objects. These measurements are useful only when they are reliable and reach the controller in time.

Interaction between components

The components are connected through a continuing exchange of commands and measurements. The controller sends a command, the actuators move the joints, and the sensors report what happened. The controller then compares the measured state with the target and adjusts its commands if necessary. This feedback process links the physical robot to its control functions.

Performance may fall when the mechanical design, actuator capacity, sensor accuracy and control settings do not match. For instance, a flexible structure or backlash in a transmission can cause positioning errors that cannot always be fixed by changing the controller settings alone. Likewise, delayed or noisy sensor readings can make corrections less effective. These examples show why the robot should be designed and tested as a complete system.

OPERATING PRINCIPLE OF AN INDUSTRIAL ROBOT

An industrial robot usually carries out a task through several connected stages: planning, command generation, movement and feedback-based correction. First, the task and the required end-effector position are defined. The controller then plans a path and calculates the joint settings needed to reach the target. The plan must take account of the robot's workspace, speed and acceleration limits, payload and possible collisions.

Once the movement has been calculated, the controller sends commands to the actuators. The actuators produce movement, which passes through the transmission mechanisms to the joints and links. When the joints move in a coordinated way, the end-effector follows the intended path.

While the robot is working, sensors measure its position, speed or contact conditions. The controller compares these readings with the target values and adjusts the commands if an error appears. How well this correction works depends on the model, actuator response, sensor quality and chosen control method [4], [6], [8].

Robot operation is therefore more than a single mechanical action. It depends on task planning, kinematic calculations, control commands, actuation and feedback working together. This is especially important when production conditions change or when robots share a workspace with other machines.

KINEMATIC AND DYNAMIC CHARACTERISTICS

Kinematics describes movement without directly examining the forces that cause it. Forward kinematics uses the joint variables to calculate the end-effector's position and orientation. Inverse kinematics works in the opposite direction: it finds joint settings that can produce a required tool position and orientation. Degrees of freedom describe the independent movements available to the robot. Their number and arrangement influence reach, flexibility and tool orientation [5], [11], [13].

Kinematic models are used to plan trajectories and estimate where the tool will be. If link dimensions, joint calibration or model parameters are inaccurate, the calculated position may differ from the actual one. Calibration and compensation can help reduce this difference, as discussed in research on forward-kinematic models for serial robots [5].

Dynamics examines the forces and torques needed to produce movement. The payload, mass of the links, acceleration, friction and contact forces all affect how much effort the actuators must provide. A

path may be possible from a kinematic point of view but still demand too much torque or cause vibration. Robot design and control should therefore consider both kinematics and dynamics [7], [9].

CONTROL AND SENSOR SYSTEMS

The control system coordinates movement and checks whether the robot is following the task. In a typical closed-loop system, the controller compares the required state with sensor measurements and uses the difference to update actuator commands. Depending on the task, it may control joint position, velocity, torque or the position of the end-effector.

Position and velocity measurements help the robot follow a planned path. Force and torque sensors are useful for assembly and other tasks involving contact, while vision sensors can locate parts that are not always placed in the same position. The right sensor combination depends on the accuracy required, the working environment, processing time and cost.

Robust control can help a robot remain stable when the model is imperfect, the load changes or disturbances occur [4], [6]. If the environment is dynamic, motion planning must also account for obstacles and changing conditions [8]. Even a well-designed controller has limits: mechanical quality, calibration and sensor reliability still matter.

APPLICATIONS IN MODERN MANUFACTURING

Industrial robots are used for welding, assembly, material handling, machining, packaging and inspection. Their ability to repeat movements can improve consistency, especially in tasks that are repetitive or physically demanding. Automated handling can also reduce the need for workers to enter some hazardous areas, but only when suitable safety measures are in place.

Different applications place different demands on the robot. Welding usually needs a steady path and consistent tool orientation. Assembly may require precise positioning and controlled contact forces, while material handling depends on payload capacity and secure gripping. These requirements influence the robot design, end-effector, sensor choice and control method.

In smart manufacturing, robots increasingly exchange information with other machines, planning tools and monitoring systems. Digital twins can be used to simulate and test production scenarios, while collaborative robotics focuses on work shared between people and robots under appropriate safeguards [10]. These developments make reliable communication and system-level planning more important.

CONTROL APPROACHES AND DEVELOPMENT PERSPECTIVES

Robot control has gradually moved beyond fixed movement sequences toward methods that can respond to changes in operating conditions. Conventional control is still suitable for many repetitive tasks in stable environments. Robust or adaptive methods may be more useful when load, friction or other conditions vary [4], [6]. The most appropriate method depends on the task; a newer method is not automatically the better choice.

Motion planning must connect the starting position to the target while respecting the robot's physical limits and avoiding obstacles. When the surroundings change, the robot may need to update its path

using new sensor information [8]. The result depends on the quality of that information and on the time and computing resources available.

Artificial intelligence and multi-agent methods may be useful when several robots or production units need to coordinate their actions. Multi-agent reinforcement learning has been studied for robot-system control [1], and research on cooperative manipulators emphasizes the importance of modelling and coordination [7]. Before these approaches are used in critical production tasks, their safety, stability, communication requirements and decision-making behaviour need to be tested carefully.

Simulation and digital twins can help engineers compare robot configurations and control strategies before putting them into production. However, a simulation should be checked against real measurements because a simplified model may not represent friction, wear, delays or uncertainty accurately. Industry 4.0 and Industry 5.0 point toward production systems that combine physical automation with digital monitoring and greater attention to people [3], [10].

RESEARCH METHODOLOGY

This paper uses a literature-based, system-oriented approach. First, it identifies the main mechanical parts of an industrial robot—the base, links, joints, transmission mechanisms and end-effector—and considers what each contributes to movement. These parts are then examined in relation to the robot as a whole.

The functional elements are considered next, with particular attention to sensors, actuators, the controller and software. The analysis follows the flow of information from sensors to the controller and then from control commands to mechanical movement. The operating process is organized into five stages: defining the task, calculating the movement, generating commands, carrying out the movement and correcting errors using feedback.

A comparative approach is used to distinguish the roles of the main components and control methods. The findings are then considered from a system-level perspective to identify links between mechanical design, kinematics, dynamics, sensing and control. This paper synthesizes published research; it does not report new experimental measurements or claim that a physical prototype has been tested.

CONCLUSION AND RECOMMENDATIONS

The review indicates that the mechanical and functional parts of an industrial robot are closely connected. The manipulator, joints, actuators and transmission mechanisms produce movement; sensors report the robot's state; and the controller uses this information to regulate its actions. As a result, overall performance depends on how these elements work together, not simply on the quality of one component.

The robot's kinematic structure determines the movements and tool positions it can reach. Dynamic properties, payload and external forces affect the effort and stability needed to follow a trajectory. Sensor accuracy, actuator response and controller design also influence how well the robot follows its target and corrects errors. Digital technologies, collaborative robots and intelligent control offer useful possibilities, but reliability, safety and practical validation remain important.

Based on the reviewed studies, the following recommendations can be made:

1. Assess the mechanical design together with the requirements of the sensors, actuators and controller.

2. Choose position, velocity and force measurements according to the task to support accurate movement and reliable operation.

3. Evaluate robust and adaptive control methods where loads or working conditions are likely to change.

4. For systems with several robots, consider coordination and information exchange alongside collision prevention and safety.

5. Compare simulation and digital-twin results with real production data before using them to make operational decisions.

6. Further study how artificial intelligence and multi-agent control can be combined with the mechanical and functional parts of industrial robots.

References

1. Zisgen, H., & Hinrichs, J. F. (2025). Multi-agent reinforcement learning for controlling gantry robot systems. Production Engineering, 19, 871–884.
2. Mikolajczyk, T., et al. (2022). Advanced applications of industrial robotics: New trends and possibilities. Applied Sciences, 12(1), 135.
3. Liu, Z., Liu, Q., Xu, W., Wang, L., & Zhou, Z. (2022). Robot learning towards smart robotic manufacturing: A review. Robotics and Computer-Integrated Manufacturing, 77, 102360.
4. Son, J., Kang, H., & Kang, S. H. (2023). A review on robust control of robot manipulators for future manufacturing. International Journal of Precision Engineering and Manufacturing, 24, 1083–1102.
5. Pupăză, C. (2023). The importance of embedding a general forward kinematic model for industrial robots with serial architecture in order to compensate for positioning errors. Mathematics, 11(10), 2306.
6. Mazhar, A., Tanveer, A., Izhan, M., & Khan, M. Z. T. (2023). Robust control approaches and trajectory planning strategies for industrial robotic manipulators in the era of Industry 4.0: A comprehensive review. Engineering Proceedings, 56(1), 75.
7. Mazzini, M., et al. (2023). Cooperative robot manipulators dynamical modeling and control: An overview. Dynamics, 3(4), 820–854.
8. Liu, Y. (2024). Review on motion planning of robotic manipulator in dynamic environments. Journal of Sensors, 2024, 5969512.
9. Shen, S., Zhang, D., Han, X., Jiang, C., & Li, Q. (2024). A robust optimization framework for design of robotic system with kinematic and dynamic criteria. Computer Methods in Applied Mechanics and Engineering, 423, 116866.
10. Zafar, M. H., Langås, E. F., & Sanfilippo, F. (2024). Exploring the synergies between collaborative robotics, digital twins, augmentation, and Industry 5.0 for smart manufacturing: A state-of-the-art review. Robotics and Computer-Integrated Manufacturing, 89, 102769.
11. Bingül, Z., & Küçük, S. (2015). Robot Kinematiği. Kocaeli: Umuttepe Yayınları. ISBN 978-605-510-047-6.

12. Kurtoğlu, A. (2011). Robot Tekniği. İstanbul: Papatya Bilim. ISBN 978-605-422-037-3.
13. Bayraktaroğlu, Z. Y. (2023). Robot Teknolojisine Giriş: Matematiksel Modelleme (2nd ed.). Ankara: Nobel Akademik Yayıncılık. ISBN 978-605-320-930-0.
14. ISO. (2021). ISO 8373:2021 Robotics — Vocabulary. International Organization for Standardization.
15. ISO. (2025). ISO 10218-1:2025 Robotics — Safety requirements — Part 1: Industrial robots.

AI-Based Early Detection of Diabetic Retinopathy Using Handcrafted Retinal Biomarkers and Machine Learning

Mahmut Büyükbaş[1*], Ahmet Burak Bilgin[2], Ahmet Tüfek[3]

[1,2,3] Abdullah Gül University, Faculty of Engineering, Department of Electrical-Electronincs Engineering, 38060, Kayseri, Türkiye

*Corresponding author e-mail: mahmut.buyukbas@agu.edu.tr

Abstract

Diabetic retinopathy (DR) is a diabetes-related retinal disease that may cause irreversible vision loss if it is not detected and treated at an early stage. Manual screening of retinal fundus images depends on trained specialists and may become time-consuming, costly, and difficult to scale in regions with limited ophthalmological resources. This study presents a clinically oriented, undergraduate-level biomedical machine learning framework for the early detection of diabetic retinopathy using fundus images. The proposed system combines retinal image preprocessing, vessel and lesion extraction, handcrafted retinal biomarker generation, feature selection, Random Forest classification, threshold optimization, and a MATLAB-based diagnostic support interface. A total of 757 fundus images were used and converted from a seven-class disease-stage structure into a binary screening task consisting of No DR and DR classes. The image-processing pipeline extracted 43 retinal features including vessel morphology, hemorrhage-related measurements, exudate- related measurements, green-channel statistics, color-difference features, texture descriptors, and image-quality indicators. The final optimized Random Forest v9 model was evaluated using stratified 5-fold cross-validation with training-only downsampling and fold-level threshold optimization. The model achieved a mean cross-validation accuracy of 89.17% ± 2.92%, mean sensitivity of 94.21% ± 2.68%, mean specificity of 73.77% ± 9.93%, and mean AUC of 0.9214. The best fold achieved 90.79% accuracy, 92.11% sensitivity, 86.84% specificity, 95.45% precision, an F1- score of 0.9375, and an AUC of 0.9377 at the optimized threshold of 0.5325. In addition to classification, the system includes a MATLAB graphical user interface and clinical report export functionality. Therefore, this study is not only a classifier, but also a complete decision-support prototype for diabetic retinopathy screening.

Keywords: *Diabetic Retinopathy, Medical Image Processing, Retinal Biomarkers, Machine Learning, Computer-Aided Diagnosis.*

INTRODUCTION

Diabetic retinopathy (DR) is one of the major microvascular complications of diabetes and affects the retinal blood vessels. As the disease progresses, abnormal retinal changes such as microaneurysms, hemorrhages, exudates, vessel irregularities, and neovascularization may occur. If these findings are not detected in time, DR may lead to severe visual impairment or vision loss. For this reason, early screening is an important component of preventive eye care.

Traditional DR screening is commonly based on the inter- pretation of retinal fundus photographs by ophthalmologists or trained graders. Although this approach is clinically reliable, it has important limitations. It requires trained specialists, consumes time, and may become difficult to scale when the number of patients increases. In many screening workflows, the main bottleneck is not the collection of fundus images, but the availability of trained interpreters who can examine these images consistently. Therefore, automated image-analysis sys- tems can support clinicians by prioritizing suspicious cases and reducing the workload of manual screening.

The main objective of this study is to develop an AI-based diabetic retinopathy screening prototype using retinal fundus images. The system is designed as a decision-support tool and is not intended to replace medical diagnosis. Instead, it aims to classify fundus images as DR detected or No DR and provide interpretable information through extracted retinal features. This approach is especially suitable for an undergraduate- level biomedical machine learning study because it combines image processing, clinically meaningful feature extraction, machine learning, and application-level deployment.

Unlike a simple machine learning experiment, the pro- posed study was developed as a complete screening-oriented pipeline. The system includes retinal preprocessing, vessel seg- mentation, optic disc localization, bright lesion extraction, dark lesion extraction, handcrafted biomarker extraction, feature se- lection, Random Forest classification, threshold optimization, graphical user interface design, and clinical report export. The overall architecture of the proposed system is shown in Fig. 1.

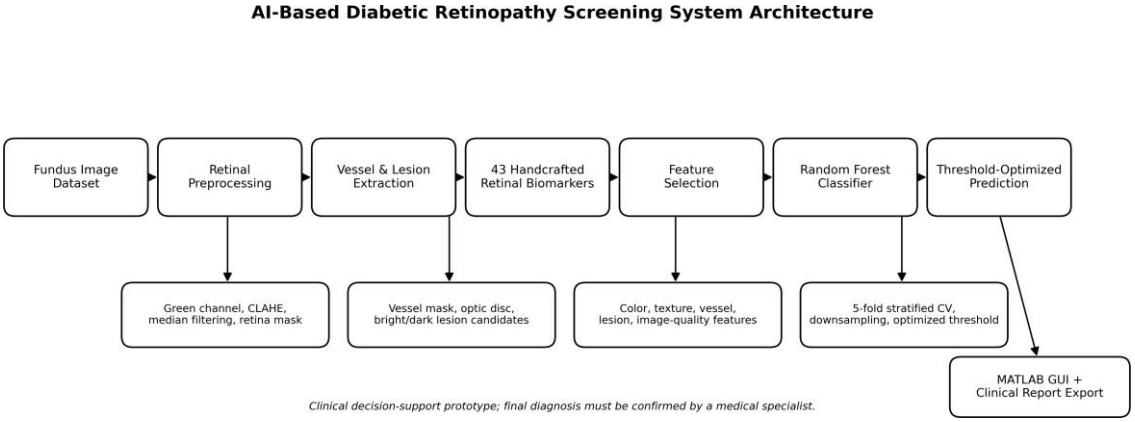


Figure 1. Overall architecture of the proposed diabetic retinopathy screening system, including preprocessing, biomarker extraction, feature selection, Random Forest classification, threshold optimization, GUI output, and clinical report export.

The main contributions of this study can be summarized as follows:

- A complete retinal image-processing pipeline was devel- oped for fundus image analysis.
- A set of 43 handcrafted retinal biomarkers was extracted from each image.
- Vessel, lesion, color, texture, and image-quality features were combined in a single feature-engineering frame- work.
- A stratified 5-fold cross-validation structure was used to improve evaluation reliability.

- Training-only downsampling was applied to reduce class imbalance without modifying the test folds.
- In the best fold, Random Forest feature-importance anal- ysis selected 25 of the 43 extracted features.
- Threshold optimization was used instead of relying only on the default 0.5 decision threshold.
- A MATLAB-based diagnostic support interface and clin- ical report export function were implemented.

Automated diabetic retinopathy detection has been widely studied using both deep learning and classical machine learn- ing approaches. Deep learning-based systems have shown strong performance in detecting referable diabetic retinopa- thy from fundus photographs. Gulshan et al. developed and validated a deep learning algorithm for diabetic retinopathy detection and reported AUC values of 0.991 and 0.990 on two validation datasets [7]. Ting et al. developed a deep learning system for diabetic retinopathy and related eye diseases and reported 90.5% sensitivity, 91.6% specificity, and an AUC of 0.936 for referable diabetic retinopathy detection [8]. Pratt et al. demonstrated the applicability of convolutional neural networks for diabetic retinopathy severity classification [9].

Deep learning methods are powerful because they can learn discriminative image patterns directly from raw or min- imally processed images. However, they usually require large datasets, higher computational resources, and careful clinical validation. Their internal decision process may also be difficult to interpret, especially when the model is used in a medical decision-support context.

Classical machine learning approaches follow a different strategy. Instead of learning directly from raw images, these methods first extract meaningful retinal features such as vessel density, lesion area, hemorrhage count, exudate ratio, texture descriptors, and intensity statistics. Then, machine learning classifiers such as Random Forest, Support Vector Machine, or ensemble models are trained using these features. Although these methods may not always outperform large-scale deep learning systems, they provide better interpretability and are more suitable for educational, lightweight, and feature-based biomedical prototypes.

The present study follows an interpretable feature- engineering approach. Most previous studies mainly focus on classification performance, whereas this study addition- ally includes a complete retinal image-processing workflow, handcrafted retinal biomarker extraction, feature selection, threshold optimization, MATLAB GUI integration, and clin- ical report export. Therefore, the proposed system should be considered not only as a classifier, but also as a complete biomedical screening prototype.

MATERIAL AND METHODS

Dataset

The dataset used in this study is the fundus image dataset for the study of diabetic retinopathy published by Castillo Benítez et al. [2]. The dataset is also available through Zenodo [3]. It contains 757 color retinal fundus images categorized into seven disease-stage classes: No DR signs, Mild NPDR, Moderate NPDR, Severe NPDR, Very Severe NPDR, PDR, and Advanced PDR. According to the dataset description, the fundus images were clinically evaluated with ophthalmology- specialist involvement.

This increases the reliability of the ground-truth labels used for model training and evaluation. Therefore, the dataset is suitable for developing and testing a supervised diabetic retinopathy screening prototype.

The original dataset distribution is shown in Table I. The distribution is imbalanced, especially because the Mild NPDR class contains only four images. For this reason, the current version of the study focuses on binary classification instead of direct seven-class disease-stage classification.

For the binary classification task, all images labeled as No DR signs were assigned to the No DR class, while all remain- ing disease stages were assigned to the DR class. This structure is appropriate for screening because the main objective is to determine whether diabetic retinopathy is present or not. The binary class structure is shown in Table II.

Table 1. Original Dataset Class Distribution

Class	Class Name	Number of Images
1	No DR signs	187
2	Mild NPDR	4
3	Moderate NPDR	80
4	Severe NPDR	176
5	Very Severe NPDR	108
6	PDR	88
7	Advanced PDR	114

Table 2. Binary Classification Structure

Binary Class	Including Original Classses	Number of Images
No DR	No DR signs	187
DR	Mild NPDR to Advanced PDR	570
Total	All Classes	757

Methods

Image Processing Stage

The retinal image-processing pipeline was developed in MATLAB. The purpose of this stage was to transform raw fun- dus images into meaningful numerical biomarkers for machine learning. The image-processing pipeline is shown in Fig. 2.

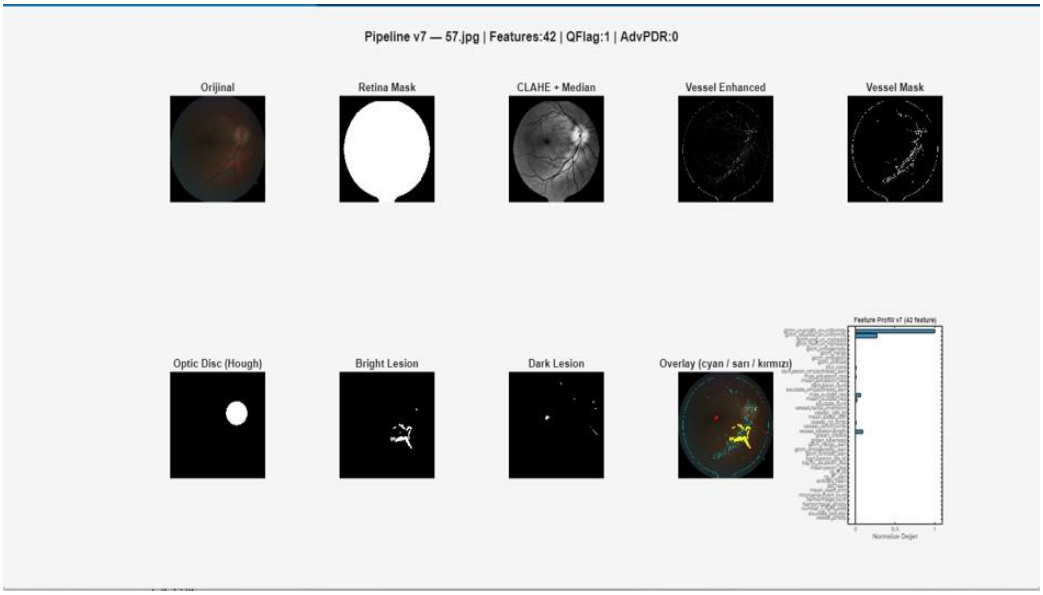


Figure 2. Image-processing pipeline showing the original fundus image, retinal mask, CLAHE-enhanced image, vessel enhancement, vessel mask, optic disc detection, bright lesion extraction, dark lesion extraction, overlay visualization, and feature profile.

First, the input fundus image was resized and standardized. The green channel was extracted because retinal vessels and dark lesions are usually more visible in this channel. Contrast Limited Adaptive Histogram Equalization (CLAHE) was applied to enhance local contrast (Zuiderveld, 1994). Median filtering was then used to reduce noise while preserving important retinal structures.

A retinal mask was generated to separate the valid retinal field from the non-retinal black background. This step is important because texture and intensity measurements should be computed only inside the retinal region. If the black background is included in the analysis, it may distort the feature values and reduce model reliability.

Vessel enhancement was performed to highlight retinal vascular structures. The vessel mask was then obtained using thresholding and morphological operations. Vessel-based features are clinically meaningful because diabetic retinopathy can affect retinal vascular structure and appearance.

Optic disc detection was also included because the optic disc is a bright anatomical structure and may be falsely detected as an exudate-like lesion. By localizing the optic disc, the system can reduce false bright-lesion detections. Bright lesion extraction was used to identify exudate-like regions, while dark lesion extraction was used to detect hemorrhage-like regions and possible microaneurysm candidates.

Handcrafted Retinal Feature Extraction

After preprocessing, each fundus image was converted into a numerical feature vector. The final feature extraction system generated 43 handcrafted retinal biomarkers. These features were selected to represent vascular structure, lesion characteristics, color distribution, retinal texture, and image quality.

The extracted feature groups are summarized in Table 3. Vessel features include vessel density, skeleton length, branch points, end points, vessel width, and fractal-dimension-related measurements. Lesion features include exudate area ratio, hemorrhage density, lesion area, lesion compactness, and microaneurysm count. Statistical features include green-channel intensity statistics, red-green intensity differences, skewness, kurtosis, and entropy. Texture features include GLCM and GLRLM descriptors. GLCM features are based on gray-level co-occurrence relationships, while GLRLM features describe run-length-based texture patterns (Haralick et al., 1973; Galloway, 1975).

Table 3. Main Handcrated Retinal Feature Groups

Feature Group	Examples
Vessel features	Vessel density, skeleton length, branch points, end points, vessel width, vessel fractal dimension
Lesion features	Exudate area ratio, hemorrhage density, bright lesion count, dark lesion count, microaneurysm count
Statistical features	Mean green intensity, standart deviation, entrpy, skewness, kurtosis, red-green difference
Texture features	GLCM contrast, correlation, energy, homogeneity, glrlm short-run emphasis, long-run emphasis, nonuniformity
Image-quality fuetures	Blur score and image-quality flag

This feature-engineering approach improves interpretability. Instead of giving only a prediction, the system can also indi- cate which retinal biomarkers were extracted and used by the classifier. This is important for a biomedical decision-support prototype because model outputs should be understandable and traceable

Machine Learning Dataset Preparation

The extracted feature table was used to create the machine learning dataset. Each row represented one retinal fundus image, and each column represented either label information or an extracted feature. The binary label column was selected as the target variable. The dataset preparation output is shown in Figure 3.

The final feature matrix contained 43 features. The binary class distribution consisted of 187 No DR images and 570 DR images. Because the dataset was imbalanced, class balancing was required during training.

Training-Only Downsampling

Class imbalance is an important problem in medical image classification. If the model is trained directly on an imbalanced dataset, it may become biased toward the majority class. In this study, the DR class contained more samples than the No DR class. Therefore, training-only downsampling was applied inside each cross-validation fold.

The key design decision was to apply downsampling only to the training set. The test fold was not modified. This prevents data leakage and keeps the evaluation realistic. If the test set were artificially balanced, the reported performance could become misleading. By keeping the test fold unchanged, the model was evaluated on naturally distributed unseen samples.

Stratified 5-Fold Cross-Validation

Instead of using only one train/test split, the final model was evaluated using stratified 5-fold cross-validation. In this method, the dataset is divided into five folds while preserving the class distribution as much as possible. In each iteration, four folds are used for training and one fold is used for testing. This process is repeated until every sample has been used for testing once.

Stratified cross-validation provides a more reliable performance estimate than a single random split because it tests the model across multiple data partitions. This is especially important for limited and imbalanced biomedical datasets.

Feature Selection

In the final RF v9 model, the 43 extracted retinal biomarkers were reduced to 25 selected features in the best fold based on Random Forest importance scores. This reduction was applied to focus the classifier on stronger retinal indicators, reduce redundant feature effects, and decrease the risk of overfitting. Therefore, feature selection became one of the main engineering improvements of the final pipeline rather than only a preprocessing detail.

Random Forest Classifier

Random Forest was selected because it is robust, interpretable, and suitable for tabular handcrafted feature data. It combines multiple decision trees and makes a final prediction using ensemble voting (Breiman, 2001). In this study, Random Forest was implemented using MATLAB's machine learning tools (MathWorks, 2026). Several Random Forest pipeline versions were tested during the study development. The final RF v9 model was selected because it achieved the best balance between sensitivity and specificity while preserving high AUC performance. Therefore, the final model selection was based not only on accuracy, but also on screening-oriented clinical performance. For a binary classification task, the model predicts either DR or No DR. In the RF v9 implementation, the confidence score is calculated from the proportion of decision trees voting for the predicted class. If NT is the total number of trees and Nv is the number of trees voting for the final predicted class, the confidence score can be expressed as:

$$Confidence = \frac{N_y}{N_t} \times 100 \quad (1)$$

For example, if 650 of 700 decision trees vote for DR, the confidence score is approximately 92.86%. This score indicates agreement within the Random Forest ensemble. However, it is not a calibrated clinical probability and does not guarantee clinical correctness.

Threshold Optimization

The default classification threshold of 0.5 is not always optimal for medical screening. In a diabetic retinopathy screening system, high sensitivity is important because missed DR cases may delay further clinical examination. However, specificity is also important because false-positive predictions may lead to unnecessary referrals, additional workload, cost, and patient anxiety.

For this reason, threshold optimization was performed. Different threshold values were tested, and accuracy, sensitivity, specificity, F1-score, AUC, and balanced accuracy were evaluated together. The optimized threshold for the best fold was 0.5325. This threshold provided a better screening-oriented balance between sensitivity and specificity.

Evaluation Metrics

The model was evaluated using accuracy, sensitivity, specificity, precision, F1-score, balanced accuracy, and AUC. These metrics were calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} (2)$$

$$Sensitivity = \frac{TP}{TP + FN} (3)$$

$$Specificity = \frac{TN}{TN + FP} (4)$$

$$Precision = \frac{TP}{TP + FP} (5)$$

$$F1 = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity} (6)$$

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2} (7)$$

Here, TP represents correctly detected DR cases, TN repre- sents correctly detected No DR cases, FP represents healthy images incorrectly classified as DR, and FN represents DR images incorrectly classified as No DR.

MATLAB GUI Application and Report Export

The study was extended from a classifier into a MATLAB- based screening application. The interface includes image up- load, model-based analysis, prediction output, DR confidence score, image information, key feature summary, selected fea- ture table, and clinical report export. The updated MATLAB GUI is shown in Fig. 3.

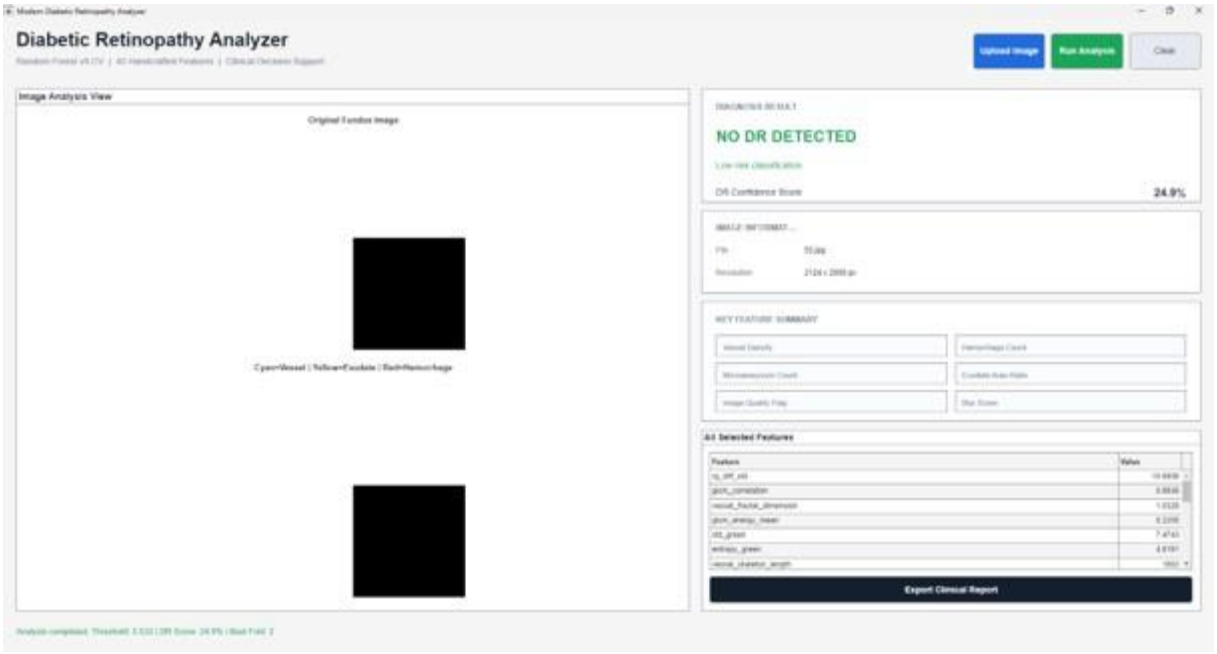


Figure 3. Updated MATLAB-based diabetic retinopathy analyzer interface showing image upload, RF v9 analysis result, confidence score, image information, key feature summary, selected feature table, and clinical report export option.

The clinical report export function makes the system more useful as a decision-support prototype. The exported report includes prediction result, risk level, confidence score, thresh- old value, and extracted retinal feature values. An example exported report is shown in Fig. 4.

```
============================================================
 DIABETIC RETINOPATHY ANALYSIS REPORT
============================================================
Date       : 24.05.2024 11:28:20
Image Path : C:\Dataset_from_fundus_images_for_the_study_of_diabetic_retinopathy_RGB\7. Advanced PDR\651.jpg
Model      : Random Forest v9 -- 5-Fold Stratified CV
Best Fold  : 2
------------------------------------------------------------
Diagnosis  : DIABETIC RETINOPATHY DETECTED
Risk Level : HIGH RISK
ER Score   : 95.00%
Threshold  : 0.5125

FEATURE VALUES
------------------------------------------------------------
vessel_density                  : 0.012968
exudate_area_ratio              : 0.000000
number_of_lesion_spots          : 0.000000
hemorrhage_density              : 0.013772
hemorrhage_count                : 41.000000
microaneurysm_count             : 2.000000
foam_area_area                  : 0.649395
std_green                       : 11.185980
entropy_green                   : 5.483386
rg_diff_mean                    : 87.825488
rg_diff_std                     : 24.914273
num_lesions_area                : 59.358364
log10_lesions_area              : 6.064826
log10_lesions_size_td           : 4.651696
glcm_contrast_meanl             : 3.019521
glcm_homogeneity_mean           : 0.729037
glcm_energy_mean                : 0.116956
glcm_contrast                   : 1.572797
glcm_correlation                : 0.908368
glcm_energy                     : 0.116608
```

Figure 4. Exported clinical report generated by the MATLAB application, showing diagnosis result, risk level, confidence score, optimized threshold, and extracted retinal feature values.

The software is intended only as a clinical decision-support prototype. Final diagnosis must always be confirmed by a medical specialist.

RESULTS

Pipeline Improvement from Process-I to Final Model

The final system was developed through multiple improve- ment stages. Compared with the Process-I model, the final version includes a broader feature set, stronger validation strategy, feature selection, threshold optimization, and an application-level output interface. The main improvements are summarized in Table IV. This development path shows that the final performance was not obtained by a single classifier run. Instead, it resulted from systematic engineering decisions applied to data preparation, feature design, validation, model selection, and application deployment.

Table 4. Main Improvements From Process-I to Final Model

Improvement	Final Implementation	Main Effect
Feature expansion	14 basic features to 43 retinal biomarkers	More retinal information was represented numerically
Training-only balancing	Downsampling applied only to training folds	Reduced class bias while keeping test data realistic
Feature Selection	43 extracted features reduced to 25 selected features	Reduced redundancy and overfiting risk
Thresold Optimization	Optimized threshold of 0.5325	Improved sensitivity-specifity balance
Application Output	MATLAB GUI and report export	Converted the classifier into a screening prototype

Cross-Validation Results

The final optimized Random Forest v9 model was eval- uated using stratified 5-fold cross-validation. The MATLAB summary output is shown in Fig. 7, and the numerical results are summarized in Table V.

Table 5. Final RF v9 Cross-Validation Summary

Metric	Result
Cross-validation	5-fold stratified
Best fold	Fold 2
Mean-CV accuracy	89.17% ± 2.92%
Mean-CV sensitivity	94.21% ± 2.68%
Mean CV AUC	0.9214
Best-fold accuracy	90.79%
Best-fold sensitivity	92.11%
Best-fold precision	95.45%
Best-fold F1-score	0.9375
Best fold AUC	0.9377
Best threshold	0.5325
Selected features	25

The mean CV accuracy of 89.17% shows that the model achieved stable performance across different data partitions. The mean sensitivity of 94.21% is especially important for a screening system because the model should detect most DR-positive cases. The relatively small standard deviations in accuracy (±2.92%) and sensitivity (±2.68%) indicate that the model performance was not dependent on a single

random split. Instead, the classifier showed stable behavior across the five stratified folds. The mean specificity of 73.77% indicates that healthy-case recognition improved compared with the pre- liminary model, although specificity remained more variable than sensitivity.

Fold-by-Fold Stability Analysis

The fold comparison graph is shown in Fig. 5.

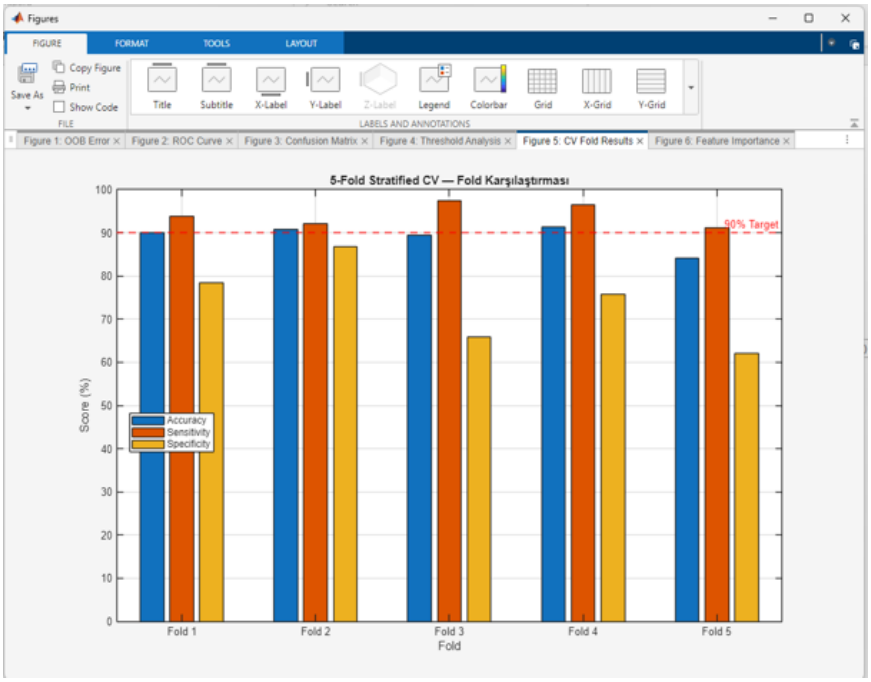


Figure 6-5. Five-fold stratified cross-validation comparison of the optimized RF v9 model, showing fold-level accuracy, sensitivity, and specificity values with a 90% target reference line

The fold-level performance results are shown in Table VI.

Table 5. Fold-by-Fold Performance of The Optimized RF v9 Model

Fold	Accuracy	Sensitivity	Specificity	F1	AUC
1	90.07%	93.86%	78.38	0.9345	0.9362
2	90.79%	92.11	86.84	0.9375	0.9377
3	89.47%	97.37	65.79	0.9328	0.9220
4	91.39%	96.49	75.68	0.9442	0.9500
5	84.11%	81.23	62.16	0.8966	0.8611

The fold-by-fold results show that sensitivity remained consistently high across all folds. This indicates that the model was able to detect DR-positive cases reliably under different train/test splits. Specificity showed more variation, especially in Fold 3 and Fold 5. This suggests that healthy-case separation remains more challenging than DR detection. The possible reasons include image-quality variation, lesion-like artifacts in No DR images, optic disc brightness, and the smaller number of No DR samples.

ROC and AUC Analysis

The ROC curve of the best fold is shown in Fig. 6. The– best fold achieved an AUC value of 0.9377. The AUC value shows the discrimination ability of the model independently of a single threshold. An AUC of 0.9377 indicates strong separation between DR and No DR classes in the best fold. This result supports the reliability of the selected handcrafted biomarkers and the optimized Random Forest model.

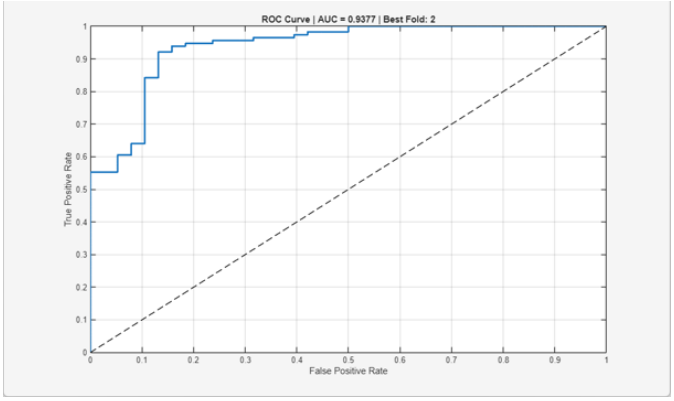


Figure 6. ROC curve of the optimized RF v9 model for the best fold, showing an AUC value of 0.9377.

Confusion Matrix Analysis

The confusion matrix of the best fold is shown in Fig. 7. At the optimized threshold of 0.5325, the model correctly clas- sified 33 No DR images and 105 DR images. It misclassified 5 No DR images as DR and 9 DR images as No DR. The confusion matrix shows that the model achieved 92.11% sensitivity and 86.84% specificity in the best fold. This is a strong result for a screening-oriented system because the model preserved high DR detection performance while also improving the recognition of healthy retinal images.

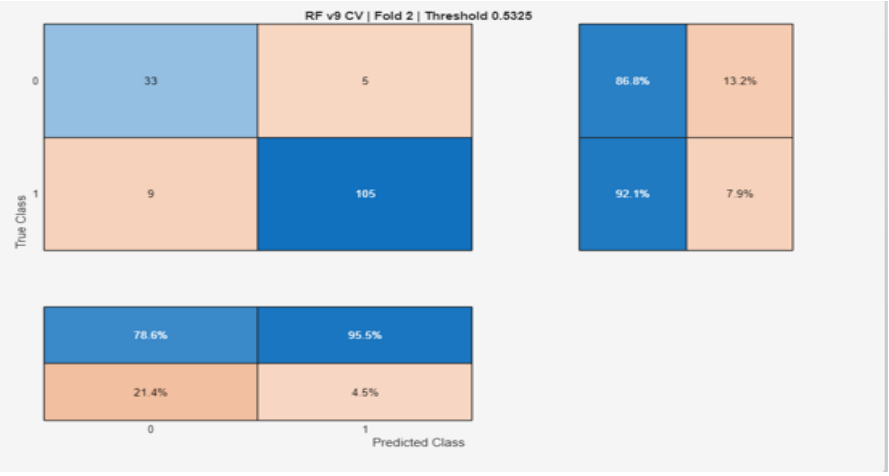


Figure 7. Confusion matrix of the optimized RF v9 model for the best fold at the optimized threshold

of 0.5325, showing correctly and incorrectly classified No DR and DR samples

Threshold Optimization Analysis

The threshold-search curve is shown in Fig. 8. This graph demonstrates how accuracy, sensitivity, specificity, and bal- anced accuracy change with different decision thresholds.

The threshold analysis is important because the default threshold of 0.5 may not provide the best sensitivity-specificity trade-off for medical screening. In this study, the optimized threshold was selected as 0.5325. This threshold improved specificity while maintaining acceptable sensitivity. In a clin- ical screening context, this trade-off is important because excessive false positives increase unnecessary referrals, while false negatives may delay medical attention.

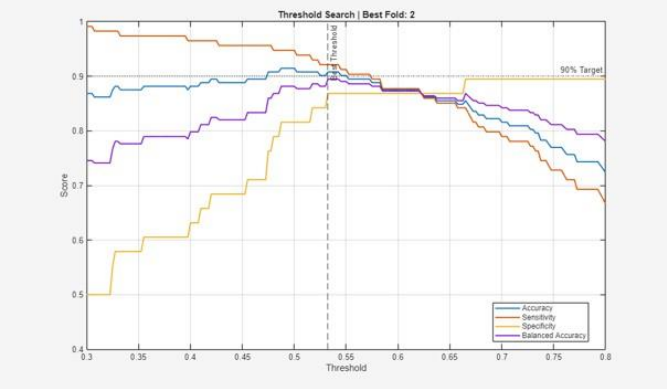


Figure 8. Threshold-search curve of the optimized RF v9 model for the best fold, showing the effect of different decision thresholds on accuracy, sensitivity, specificity, and balanced accuracy.

Feature Importance Analysis

The Random Forest feature-importance plot is shown in Fig. 9. The most influential features included red-green intensity difference, GLCM correlation, entropy, green-channel statistics, and vessel-related measurements.

The feature-importance result indicates that the model does not rely only on visible lesion counts. Instead, it combines color distribution, texture information, vessel morphology, and lesion-related features. This is meaningful because diabetic retinopathy affects both localized lesion structures and the broader retinal image appearance. Some features may show lower or negative OOB permutation contribution in a specific fold. This indicates that their contribution is weaker or less stable under that fold's data distribution, which supports the need for feature selection.

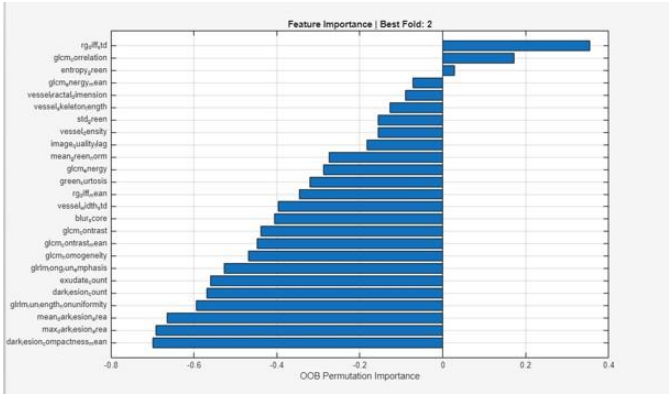


Figure 9. Random Forest feature-importance plot for the best fold, showing the OOB permutation importance of selected retinal biomarkers used in the optimized RF v9 model.

Limitations

Although the final system achieved promising results, sev- eral limitations remain. First, the dataset is imbalanced. The No DR class contains fewer samples than the DR class, and the original Mild NPDR class contains only four images. This limits the reliability of direct multi-class disease-stage classification.

Second, handcrafted features may not capture all complex retinal patterns. Some features are affected by image blur, illumination variation, optic disc brightness, field-of-view dif- ferences, and segmentation artifacts. Therefore, the model may still produce incorrect predictions for low-quality or borderline images.

Third, the model has not been clinically validated in a hospital environment. The current system is an undergraduate- level biomedical AI prototype and should not be interpreted as a diagnostic medical device. It is designed only as a preliminary screening-support system.

Fourth, the confidence score is not a calibrated clinical probability. It indicates model decision strength or agreement, but it does not guarantee clinical correctness. For real-world use, probability calibration, external validation, uncertainty estimation, and clinical expert evaluation would be required.

Future Work

Future work should focus on improving both model ro- bustness and clinical usability. First, lesion detection can be improved by better suppressing optic disc brightness and border artifacts. Second, vessel segmentation can be refined using more advanced morphological or deep-learning-based vessel extraction methods. Third, misclassified No DR and DR images should be analyzed in detail to understand the main causes of false-positive and false-negative predictions.

Additional machine learning models such as Support Vec- tor Machine, XGBoost, and ensemble variants can also be tested. A lightweight CNN or transfer-learning baseline can be added to compare handcrafted-feature-based learning with deep learning. Furthermore, probability calibration and exter- nal validation on a different dataset would make the model evaluation more reliable.

From an application perspective, the MATLAB GUI can be improved with a clearer clinical summary, better visual overlays, and more structured report formatting. In the future, the system could be extended toward a web-based or mobile screening prototype. However, any real clinical use would re- quire medical expert review, ethical approval, and prospective validation.

CONCLUSION

This final report presented an AI-based diabetic retinopathy screening system using retinal fundus images, handcrafted retinal biomarkers, and machine learning. The study pro- gressed from a preliminary Random Forest classifier into a more complete biomedical decision-support prototype. The final system includes image preprocessing, vessel and lesion extraction, 43 handcrafted retinal features, feature selection, training-only downsampling, stratified 5-fold cross-validation, Random Forest classification, threshold optimization, MAT- LAB GUI integration, and clinical report export.

The optimized RF v9 model achieved a mean cross- validation accuracy of 89.17% ± 2.92%, mean sensitivity of 94.21% ± 2.68%, mean specificity of 73.77% ± 9.93%, and mean AUC of 0.9214. The best fold achieved 90.79% accu- racy, 92.11% sensitivity, 86.84% specificity, 95.45% precision, an F1-score of 0.9375, and an AUC of 0.9377 at the optimized threshold of 0.5325.

The most important improvement compared with the pre- liminary model was the increase in specificity from 62.16% to 86.84%. This indicates better recognition of healthy retinal images and reduced false-positive predictions. At the same time, high sensitivity was maintained, which is essential for a screening-oriented DR detection system. With a mean sen- sitivity above 94%, the system may help prioritize suspicious retinal images for ophthalmologist review in a screening-support workflow.

Overall, the proposed framework demonstrates that hand- crafted retinal biomarkers and Random Forest classification can form a meaningful, interpretable, and lightweight biomed- ical screening prototype. The developed MATLAB application and clinical report export functionality further strengthen the practical value of the study. Nevertheless, the system should be considered only as a decision-support prototype, and final diagnosis must always be confirmed by a medical specialist.

References

Breiman L, 2001. Random forests. Machine Learning, 45 (1): 5–32.

Castillo Benítez VE Castro Matto I, Mello Román JC, Vázquez Noguera JL, García-Torres M, Ayala J, Pinto-Roa DP, Gardel-Sotomayor PE, Facon J, Grillo SA, 2021. Dataset from fundus images for the project of diabetic retinopathy. Data in Brief, 36: 107068.

Castillo Benítez VE Castro Matto I, Mello Román JC, Vázquez Noguera JL, García-Torres M, Ayala J, Pinto-Roa DP, Gardel-Sotomayor PE, Facon J, Grillo SA, 2021. Dataset from fundus images for the project of diabetic retinopathy. Zenodo, doi: 10.5281/zenodo.4647952.

Galloway MM, 1975. Texture analysis using gray level run lengths. Computer Graphics and Image Processing, 4 (2): 172–179.

Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, Kim R, Raman R, Nelson PC, Mega JL, Webster DR, 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA, 316 (22): 2402–2410.

Haralick RM, Shanmugam K, Dinstein I, 1973. Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics, SMC-3 (6): 610–621.

MathWorks, 2026. Image Processing Toolbox Documentation. Access address: https://www.mathworks.com/help/images/; Date of access: 05/15/2026.

Pratt H, Coenen F, Broadbent DM, Harding SP, Zheng Y, 2016. Convolutional neural networks for diabetic retinopathy. Procedia Computer Science, 90: 200–205.

Ting DSW, Cheung CY, Lim G, Tan GSW, Quang ND, Gan A, Hamzah H, Garcia-Franco R, San Yeo IY, Lee SY, Wong EY, Sabanayagam C, Baskaran M, Ibrahim F, Tan NC, Finkelstein EA, Lamoureux EL, Wong IY, Bressler NM, Sivaprasad S, Varma R, Jonas JB, He MG, Cheng CY, Zheng Y, Campa C, Pennisi P, Tai ES, Gemmy Cheung CM, Wandell BA, Rosman M, Aung T, Franklin J, Ting DSW, Wong TY, 2017. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA, 318 (22): 2211–2223.

World Health Organization, 2019. World Report on Vision. Geneva, Switzerland: World Health Organization.

Zuiderveld K, 1994. Contrast limited adaptive histogram equalization. Heckbert PS (Editor): Graphics Gems IV. Academic Press, pp. 474–485, San Diego, CA, USA.

One-Class Tabular Denoising Diffusion for Network Intrusion Detection Under Validation-Calibrated Thresholds

Fikret Özkorkmaz[1*], Ömer Faruk Sarıtaş[2]

[1] Kayseri University, Institute of Graduate Studies, Department of Computer Engineering, Kayseri, Türkiye
[2] Kayseri University, Faculty of Engineering and Architecture, Department of Computer Engineering, Kayseri, Türkiye

*Corresponding author e-mail: fikretozkrkmz@gmail.com

Abstract

One-class intrusion detectors are trained on benign traffic only, but many reported results for them depend on decision thresholds tuned with labelled test data. We study a one-class tabular denoising diffusion probabilistic model (DDPM), which we call TabDiffAD, on the NSL-KDD benchmark. The model is trained on 57,241 benign KDDTrain+ connection records. Its composite anomaly score combines multi-timestep reconstruction with noise-prediction error, effectively applying timestep-dependent weighting to denoising residuals. All primary operating thresholds are calibrated on 10,102 held-out benign records, and test labels are used only for scoring. Test-tuned oracle thresholds are evaluated separately to illustrate optimistic evaluation bias. Four one-class baselines are run under the same protocol. Over three stochastic scoring seeds, TabDiffAD reaches a receiver operating characteristic area under the curve (ROC-AUC) of 94.98% on KDDTest+. At a 5% validation false positive rate (FPR) target it detects 82.93% of attacks at a test FPR of 8.50%, including 77.11% of user-to-root (U2R) and 50.19% of remote-to-local (R2L) records. A deep autoencoder trained under the same protocol is stronger on most metrics, with a ROC-AUC of 96.15%. In the ablation, the composite score gives the highest ROC-AUC among all score variants, while reconstruction error alone misses most R2L attacks. We also show that thresholds chosen to maximise the F1 score on test labels raise it by about 3.5 points on KDDTest+ and hide test FPRs above 80% on KDDTest-21.

Keywords: *Network Intrusion Detection, Diffusion Models, One-Class Anomaly Detection, Threshold Calibration, NSL-KDD*

INTRODUCTION

Signature-based intrusion detection cannot recognise attacks that have no signature yet. Anomaly detectors trained only on benign traffic are a common answer, because benign traffic is easy to record while labelled attacks are scarce. This setting has well-known pitfalls of its own (Sommer and Paxson, 2010). The difficulty is that the rarest classes are also the most damaging ones. In the NSL-KDD training set (Tavallaee et al., 2009), user-to-root (U2R) privilege escalation accounts for 52 of 125,973 records, and remote-to-local (R2L) intrusions for 995. Supervised classifiers trained on this distribution tend to

treat such records as noise, which has motivated resampling and generative augmentation studies (Alqarni and El-Alfy, 2022; Fu et al., 2022; Zhang et al., 2025).

One-class detectors avoid the label problem, but their evaluation is fragile. A score separates benign from malicious traffic only after a threshold is chosen. The F1 score, the harmonic mean of precision and recall, is a common criterion for this choice. When that threshold is selected by maximising F1 on the labelled test set, the reported precision, recall and false positive rate (FPR) describe an operating point that no deployed system could know in advance. Recent NSL-KDD work has started to state this explicitly and to tune thresholds on training data only (Gharghi and Dehghani, 2026). We follow the same principle for a one-class model.

This study examines whether a tabular denoising diffusion probabilistic model (DDPM) (Ho et al., 2020; Kotelnikov et al., 2023) is a useful one-class detector for connection records. We call the model TabDiffAD. A denoising network is trained on benign records, and test records are scored by how poorly the network denoises them at several noise levels. The composite score combines multi-timestep reconstruction with noise-prediction error, applying timestep-dependent weighting to denoising residuals.

One-class support vector machines (SVM) (Schölkopf et al., 2001) and Isolation Forest (Liu et al., 2008) are the classical shallow detectors for this setting. Deep support vector data description (Deep SVDD) (Ruff et al., 2018) learns a compact hypersphere around benign representations. Autoencoders remain the most common deep baseline for network anomaly detection. Song et al. (2021) showed on NSL-KDD and two Internet of Things (IoT) datasets that the latent size can change detection performance considerably. Ponnarasan et al. (2025) combined an autoencoder with z-score thresholding, and Sagaran et al. (2025) trained a spiking contrastive autoencoder only on normal traffic. Al-Harbi and Hamed (2024) used a sparse autoencoder in front of a convolutional neural network (CNN) and bidirectional long short-term memory (BiLSTM) classifier for imbalanced NSL-KDD data.

Diffusion models learn a data distribution by reversing a gradual Gaussian corruption (Ho et al., 2020), and the cosine noise schedule (Nichol and Dhariwal, 2021) is widely used for them. TabDDPM (Kotelnikov et al., 2023) adapted diffusion to mixed numerical and categorical tables. In intrusion detection, diffusion has been used mainly to generate data. Cui et al. (2026) augmented minority NSL-KDD traffic with an improved denoising diffusion implicit model (DDIM) before a CNN classifier. Zhang et al. (2025) generated few-shot IoT samples with a conditional DDPM, and Pamuditha et al. (2025) compared generative adversarial network (GAN) and diffusion generators for fifth-generation (5G) mobile network traffic. Behrendt et al. (2024) used diffusion reconstructions directly as anomaly scores, but for brain magnetic resonance imaging (MRI) rather than network traffic. Tritscher et al. (2024) and Galwaduge and Samarabandu (2025) used diffusion models to explain anomaly and intrusion detectors. Closest to our setting, Wang et al. (2024) detected intrusions in KDD Cup 99 records by the reconstruction error of a DDPM. Feng et al. (2025) built a reconstruction-based diffusion detector for tabular data with synthetic minority oversampling (SMOTE) guided by a variational autoencoder (VAE). Diffusion reconstruction as an intrusion score is therefore not new. Combining several quantities of a score-based model into one anomaly measure is not new either, and Lim et al. (2023) did so for multivariate time series.

This study focuses on three aspects of one-class diffusion-based intrusion detection. For tabular connection records, the score combines noise-prediction error with reconstruction error, providing timestep-dependent weighting. All primary operating thresholds are calibrated without test labels. The evaluation also includes the harder KDDTest-21 subset together with four baselines run under the same protocol. The contributions are as follows. First, we use a leakage-controlled pipeline in which preprocessing and model fitting rely on benign training records, while checkpoint selection and operating thresholds use a separate held-out benign validation set. Second, we compare TabDiffAD with four one-class baselines evaluated under the same split and validation-calibrated operating points. Third, an ablation separates the reconstruction and noise-prediction terms and examines how timestep-dependent weighting affects individual attack categories. Finally, we quantify how much test-tuned oracle thresholds inflate results on KDDTest+ and KDDTest-21. TabDiffAD ranks second on receiver operating characteristic area under the curve (ROC-AUC), behind the deep autoencoder.

MATERIAL AND METHODS

Material

Each NSL-KDD record holds 38 numerical and 3 categorical attributes. The categorical ones are the protocol type, the network service and the connection status flag. Two columns are removed. The difficulty level column is a score produced by the benchmark authors' classifiers and is not a traffic feature, and the number of outbound commands is zero in every record. Next, the connection duration and the byte counts sent by the source and by the destination, which are heavy-tailed, are transformed with $\log(1+z)$. Numerical attributes are then scaled to $[-1,1]$ and categorical attributes are one-hot encoded, which gives $D = 76$ input dimensions (37 numerical and 39 binary). The scaler and the encoder are fitted only on the benign training split. Categories unseen in that split are encoded as all-zero vectors.

The 67,343 benign KDDTrain+ records are split at random (seed 42) into 57,241 training and 10,102 validation records. No attack record of KDDTrain+ is used anywhere. Test attacks are grouped into denial of service (DoS), probing (Probe), R2L and U2R with the usual name mapping. KDDTest+ then contains 9,711 benign, 7,460 DoS, 2,421 Probe, 2,885 R2L and 67 U2R records. KDDTest-21, the subset left after removing records that all 21 reference classifiers of Tavallaee et al. (2009) labelled correctly, contains 2,152 benign, 4,344 DoS, 2,402 Probe, 2,885 R2L and 67 U2R records. NSL-KDD is publicly available from the Canadian Institute for Cybersecurity.

Methods

Diffusion Model

A clean record $\boldsymbol{x}_0$ is corrupted over $T = 100$ steps with

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\,\boldsymbol{x}_0,\ (1-\bar{\alpha}_t)\mathbf{I}) \tag{1}$$

where $\bar{\alpha}_t$ follows the cosine schedule (Nichol and Dhariwal, 2021)

$$\bar{\alpha}_t = \frac{f(t)}{f(0)}, \quad f(t) = \cos^2\left(\frac{t/T+s}{1+s}\cdot\frac{\pi}{2}\right), \quad s = 0.008 \tag{2}$$

A network $\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t)$ predicts the added noise and is trained with

$$\mathcal{L}(\theta) = \mathbb{E}_{t,x_0,\boldsymbol{\varepsilon}}\|\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t)\|_2^2 \tag{3}$$

where t is drawn uniformly from the T steps and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The network is a residual multilayer perceptron (MLP) with a 76-to-256 input projection, four residual blocks and a 256-to-76 output layer. Each block applies layer normalization and a 256 by 256 linear layer, then adds a projected 128-dimensional sinusoidal time embedding. A sigmoid linear unit (SiLU) activation, dropout of 0.1 and a second 256 by 256 linear layer follow, and a skip connection closes the block. The model has 733,260 trainable parameters. It was trained for 30 epochs with AdamW (learning rate 10^{-3}, weight decay 10^{-4}, batch size 256) and cosine learning-rate annealing on an NVIDIA RTX A4000. The checkpoint with the lowest validation value of Eq. (3) was kept, which was epoch 29.

Anomaly Score

A test record $\boldsymbol{x}$ is noised with Eq. (1) at each $t \in \mathcal{K} = \{10,25,50,75\}$, using $S = 3$ independent noise draws $\boldsymbol{\varepsilon}^{(s)}$. From each noisy copy, a one-step estimate of the clean record is

$$\hat{\boldsymbol{x}}_0^{(t,s)} = \frac{\boldsymbol{x}_t^{(s)} - \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t^{(s)}, t)}{\sqrt{\bar{\alpha}_t}} \tag{4}$$

The reconstruction term and the noise-prediction term are averaged over all $|\mathcal{K}|S = 12$ evaluations,

$$\mathcal{E}_{\text{rec}}(\boldsymbol{x}) = \frac{1}{|\mathcal{K}|S}\sum_{t,s}\frac{1}{D}\|\boldsymbol{x} - \hat{\boldsymbol{x}}_0^{(t,s)}\|_2^2 \tag{5}$$

$$\mathcal{E}_{\text{eps}}(\boldsymbol{x}) = \frac{1}{|\mathcal{K}|S}\sum_{t,s}\frac{1}{D}\|\boldsymbol{\varepsilon}^{(s)} - \boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t^{(s)}, t)\|_2^2 \tag{6}$$

and the anomaly score is

$$\mathcal{S}(\boldsymbol{x}) = \mathcal{E}_{\text{rec}}(\boldsymbol{x}) + 0.5\,\mathcal{E}_{\text{eps}}(\boldsymbol{x}) \tag{7}$$

Because the single-step estimate is derived algebraically from the noise predictor via Eq. (4), the reconstruction error is mathematically proportional to the noise-prediction error scaled by the ratio $(1 - \bar{\alpha}_t)/\bar{\alpha}_t$. Hence, the two terms are not independent error signals. Rather, the reconstruction term scales the single-sample denoising loss by the noise-to-signal ratio, which grows rapidly at larger noise levels, whereas the noise-prediction term weights all noise levels uniformly. Adding the unscaled noise-prediction term in Eq. (7) changes the timestep-dependent weighting of denoising errors, increasing the relative contribution of lower-noise timesteps compared with reconstruction-only scoring. A record is flagged when $\mathcal{S}(\boldsymbol{x}) > \tau$.

Threshold Calibration

For a target false positive rate a, the primary operating threshold τ_a is the $(1 - a)$ quantile of $\mathcal{S}$ over the 10,102 benign validation records. We use $a \in \{1,5,10,15\}$% and take $a = 5$% as the main operating point. Test labels enter only after τ_a is fixed. Because benign test traffic is not distributed exactly like benign training traffic, the realised test FPR can differ from a, and we report it separately. For comparison with earlier practice, we also evaluated the F1-optimal threshold on test labels (the oracle threshold) only to demonstrate how test-set tuning biases reported performance.

Baselines and Evaluation

The baselines were run by us on the same data. One-Class SVM uses a radial basis function (RBF) kernel with $\nu = 0.05$ and was fitted on a random subset of 15,000 training records due to kernel scaling limits, while other models used the full 57,241 benign records. Isolation Forest uses 100 trees. Deep SVDD maps records through a 76-48-32-16 network and is trained for 15 epochs. The deep autoencoder has a 76-48-24-12 bottleneck with a mirrored decoder, is trained for 15 epochs and scores records by mean squared reconstruction error. Their settings were taken unchanged from the project code, without any tuning. The same validation quantile rule sets their thresholds.

Each experiment was repeated with seeds 0, 1 and 2. For the baselines, seed variation reflects independent retraining. For TabDiffAD, evaluated from a single trained checkpoint, seed variation reflects stochastic scoring noise across independent Monte Carlo noise draws. We report the area under the receiver operating characteristic curve (ROC-AUC) and the area under the precision-recall curve (PR-AUC), which do not depend on a threshold. Test FPR, precision, recall, F1 and per-category recall are given at $\tau_{5\%}$. Scoring and timing were run with PyTorch 2.5.1 on an NVIDIA L40S.

RESULTS

Comparison With One-Class Baselines

Table 1 lists all five models on KDDTest+. TabDiffAD reaches a ROC-AUC of 94.98% and a PR-AUC of 93.79%. At $\tau_{5\%}$ it detects 82.93% of attacks at a test FPR of 8.50%, with an F1 of 87.59%. The deep autoencoder is the strongest model in this table. Its ROC-AUC is 1.17 points higher, and it detects more attacks (85.49%) at a slightly higher FPR (9.47%). TabDiffAD ranks second on ROC-AUC, recall and F1, and it has the highest Probe recall. Deep SVDD detects slightly more U2R records (78.61% against 77.11%), and has a ROC-AUC standard deviation of 1.78 percentage points across the three training seeds. Isolation Forest has the lowest FPR and the highest precision, but it misses 35% of attacks and almost all R2L records.

Table 1. Comparison on KDDTest+ at the 5% validation operating point

Model	ROC-AUC ↑	PR-AUC ↑	FPR ↓	Prec. ↑	Rec. ↑	F1 ↑	DoS ↑	Probe ↑	R2L ↑	U2R ↑
One-Class SVM	86.26	89.90	7.72	92.09	68.08	78.28	82.60	88.74	13.60	50.75
Isolation Forest	93.88	95.12	**5.20**	**94.29**	64.99	76.94	79.75	90.87	6.03	24.88
Deep SVDD	93.42	92.84	8.37	92.33	76.24	83.51	85.17	97.44	35.31	**78.61**
Deep Autoencoder	**96.15**	**95.18**	9.47	92.26	**85.49**	**88.74**	**91.70**	97.41	**59.66**	74.63
TabDiffAD	94.98	93.79	8.50	92.80	82.93	87.59	90.47	**98.86**	50.19	77.11

All values are percentages, means of three seeds. The standard deviation of ROC-AUC across seeds is 0.26 for One-Class SVM, 0.33 for Isolation Forest, 1.78 for Deep SVDD, 0.14 for Deep Autoencoder, 0.02 for TabDiffAD. All models shared the same benign training and validation split and test sets. One-Class SVM was fitted on a random subset of 15,000 training records due to kernel scaling limits, while other models used all 57,241 benign records. For baselines, seed variation reflects independent retraining. For TabDiffAD, it reflects stochastic scoring noise across random noise draws from a single checkpoint. The baselines follow Schölkopf et al. (2001), Liu et al. (2008), Ruff et al. (2018) and Song et al. (2021), in table order. DoS to U2R give per-category recall. Bold marks the best value in each column. Prec. = precision, Rec. = recall, FPR = false positive rate, DoS = denial of service, R2L = remote-to-local, U2R = user-to-root.

The ROC curves of the seed-0 runs in Figure 1 reflect this general ranking, with the deep autoencoder achieving the highest overall ROC-AUC, although the relative performance of individual models varies across operating regions.

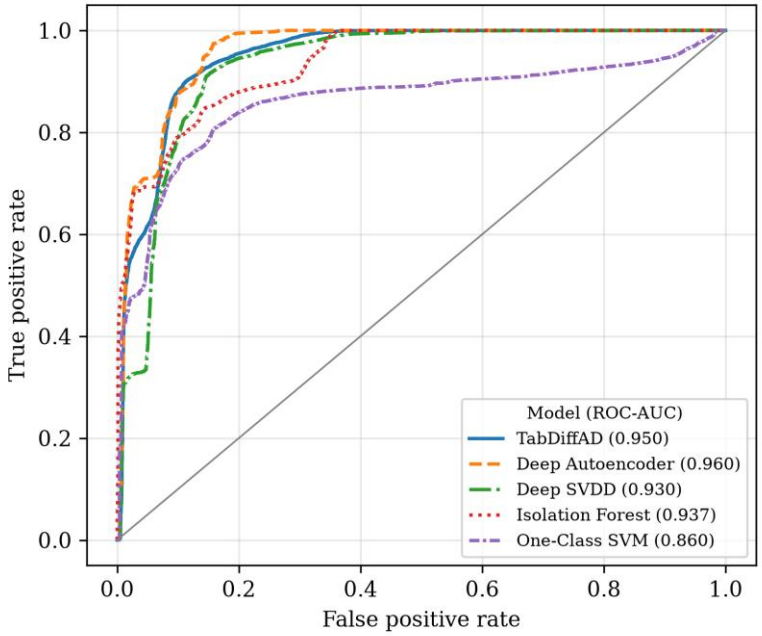


Figure 1. ROC curves of the five models on KDDTest+ (seed 0)

Operating Points and the KDDTest-21 Subset

Table 2 shows how TabDiffAD behaves when the validation target is varied. The realised test FPR differed from the validation target, particularly at the 1% and 5% operating points (8.50% at 5% and 10.06% at 10%). At the 1% target it rises to 6.02%, which suggests that the extreme tail of the benign score distribution moves between training and test traffic. Raising the target from 5% to 10% adds 17 points of R2L recall and 7 points of U2R recall for 1.6 points of extra FPR.

Table 2. TabDiffAD at four validation false positive rate targets

Test set	Target	FPR ↓	Prec. ↑	Rec. ↑	DoS ↑	Probe ↑	R2L ↑	U2R ↑
KDDTest+	1%	6.02	93.40	64.49	81.08	84.15	5.56	44.78
	5%	8.50	92.80	82.93	90.47	98.86	50.19	77.11
	10%	10.06	92.03	87.98	92.21	99.82	67.19	84.08
	15%	11.93	90.93	90.53	92.99	99.96	76.37	86.57
KDDTest-21	1%	27.18	89.83	53.27	68.09	84.10	5.59	40.80
	5%	36.73	90.48	77.44	83.95	98.89	49.84	75.12
	10%	41.20	90.21	84.24	86.83	99.86	67.34	84.58
	15%	45.14	89.73	87.46	87.88	99.97	76.44	86.57

Values are percentages, means of three seeds. The target is the FPR on the benign validation records, and FPR is the realised rate on the test set. Prec. = precision, Rec. = recall, FPR = false positive rate, DoS = denial of service, R2L = remote-to-local, U2R = user-to-root.

KDDTest-21 behaves very differently. Its ROC-AUC is 76.21%, and at the 5% target the realised FPR is 36.73%. Recall on the R2L, U2R and Probe records of this subset is within two points of the KDDTest+ values, so most of the loss comes from the benign side. The 2,152 benign records kept in KDDTest-21 are, by construction, those that 21 reference classifiers found difficult. The high realised FPR thus reflects both selection-induced difficulty and distribution difference between training and test sets, rather than benign operational drift alone. Figure 2 shows the corresponding overlap on KDDTest+. About half of the R2L records fall within the benign range, whereas Probe and most DoS records lie above both validation thresholds.

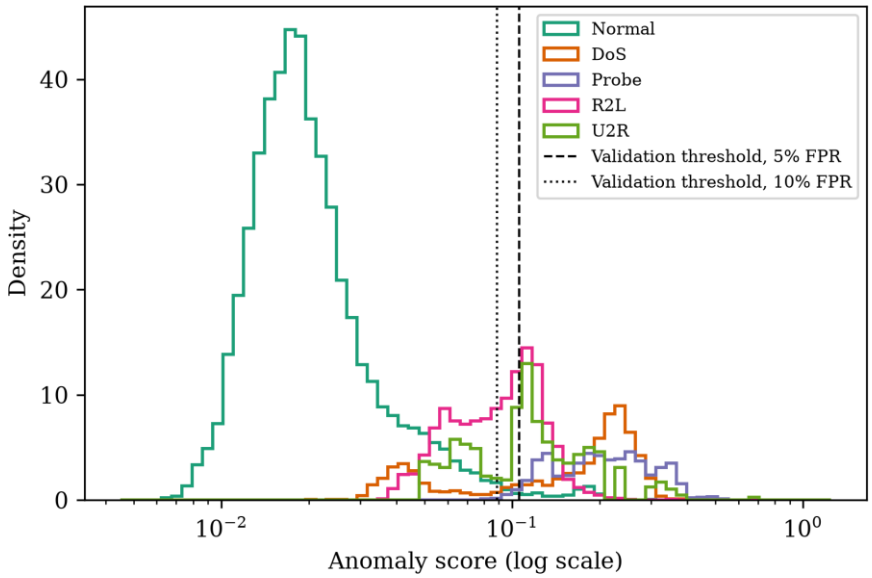


Figure 2. TabDiffAD score distributions by category on KDDTest+ (seed 0)

Effect of Test-Tuned Thresholds

With the oracle threshold, TabDiffAD obtains an F1 of 91.05% on KDDTest+ at a test FPR of 15.22%, against 87.59% at $\tau_{5\%}$. The deep autoencoder moves from 88.74% to 93.12% in the same way. On KDDTest-21 the distortion is much larger. Attacks make up 81.8% of this subset, so a detector that flags almost every record already scores an F1 near 90%. The oracle thresholds do exactly this. They give F1 values between 90.19% and 93.11% for all five models, at test FPRs between 63.94% and 97.20%. An F1 near 91% on KDDTest-21 therefore says little unless the FPR is reported next to it. This includes the value we reported for TabDiffAD in an earlier draft of this work.

Ablation of the Score

Table 3 evaluates parts of Eq. (7) at $\tau_{5\%}$ on KDDTest+. The composite score has the highest ROC-AUC and PR-AUC of all variants. Reconstruction error alone detects only 14.53% of R2L and 51.24% of U2R records. Adding the noise-prediction term raises these to 50.19% and 77.11% at a test FPR of 8.50%. The noise-prediction term on its own reaches the highest recall in the table, but with a test FPR of 10.06%, so part of its gain comes from a looser realised threshold. A clear trend appears in the single-timestep rows. Small t favours recall on the rare classes, and large t lowers both the FPR and the recall.

At $t = 75$ only 13.43% of U2R records are found. The composite score improves overall discrimination by balancing sensitivity across noise levels through timestep-dependent weighting.

Table 3. Score ablation on KDDTest+ at the 5% validation operating point

Score	ROC-AUC ↑	PR-AUC ↑	FPR ↓	Rec. ↑	R2L ↑	U2R ↑
$\mathcal{E}_{\text{rec}}$ only, all four t	93.89	93.68	7.10	71.48	14.53	51.24
$\mathcal{E}_{\text{eps}}$ only, all four t	94.76	93.04	10.06	**86.28**	**62.55**	**83.58**
$\mathcal{E}_{\text{rec}}$, t = 10	93.66	91.50	10.70	86.05	62.00	82.59
$\mathcal{E}_{\text{rec}}$, t = 25	94.12	92.91	9.00	81.37	52.69	72.64
$\mathcal{E}_{\text{rec}}$, t = 50	93.02	92.97	6.52	70.79	18.51	56.72
$\mathcal{E}_{\text{rec}}$, t = 75	91.90	92.57	**4.40**	61.59	8.00	13.43
$\mathcal{E}_{\text{rec}} + 0.5\,\mathcal{E}_{\text{eps}}$, all four t	**94.98**	**93.79**	8.50	82.93	50.19	77.11

Values are percentages, means of three seeds. The four timesteps are t = 10, 25, 50 and 75. Bold marks the best value in each column. Prec. = precision, Rec. = recall, FPR = false positive rate, DoS = denial of service, R2L = remote-to-local, U2R = user-to-root.

Feature-Level Residuals

Because Eq. (4) returns a record in feature space, the squared residual of each attribute can be inspected directly. For categorical features, the residual is summed across one-hot levels. Figure 3 shows the mean residual per category for the twelve attributes with the largest attack residuals at $t = 25$. The network service attribute carries a large residual for every attack category, and for R2L it is almost the only one. DoS and Probe also raise the residuals of the connection status flag and of the number of connections to the same host in the past two seconds. The share of those connections that use the same service rises as well, mainly for DoS. For U2R the residual concentrates on host-level indicators (Figure 4). Relative to benign traffic, the mean residual of the number of shell prompts is 3,166 times larger, followed by host login (1,455), urgent packets (1,134), root shell (147) and su attempts (115). These ratios are computed with a numerical smoothing term ($\varepsilon = 10^{-6}$) and are large primarily because benign residuals on these indicators are near zero. Because these ratios are based on only 67 U2R records, they should be treated as qualitative indicators of feature-level differences rather than causal explanations.

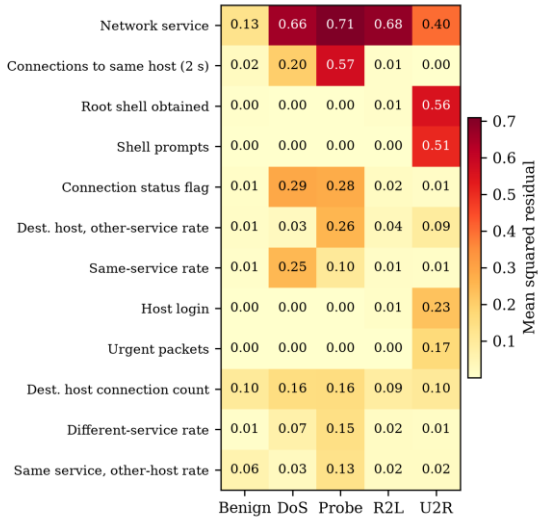


Figure 3. Mean squared residual per attribute and category at t = 25 on KDDTest+

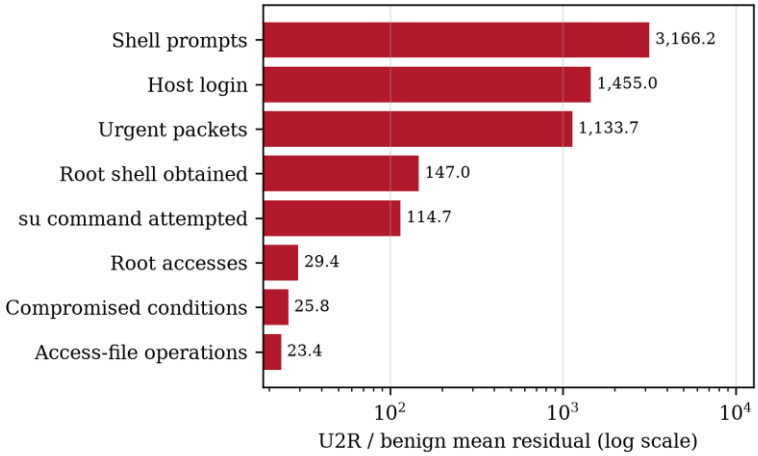


Figure 4. Ratio of U2R to benign mean residual for the eight largest attributes

Inference Cost

The full score needs 12 forward passes per record. Table 4 reports its cost on an NVIDIA L40S together with a single forward pass for reference. With batches of 4,096 records the full score processes 133,425 records per second (7.49 µs per record). A single record takes about 16 ms, so throughput depends almost entirely on batching. These figures count connection records, not packets, and they exclude feature extraction from raw traffic. They should be read as the cost of the model alone.

Table 4. Scoring cost on an NVIDIA L40S (PyTorch 2.5.1)

Batch size	Full score, µs/record ↓	Full score, records/s ↑	Single pass, µs/record ↓	Single pass, records/s ↑
1	15,993.66	63	5,890.80	170
64	258.69	3,866	87.26	11,461
256	65.00	15,384	23.12	43,258
1,024	17.19	58,188	6.62	151,121
4,096	7.49	133,425	1.31	764,116

The full score uses 12 forward passes per record (4 timesteps, 3 noise draws). Timing excludes feature extraction.

DISCUSSION AND CONCLUSION

Under the primary evaluation protocol, operating thresholds were calibrated using held-out benign records rather than test labels. The diffusion score is a reasonable one-class detector for NSL-KDD connection records, though the deep autoencoder in our comparison remains stronger on ROC-AUC and recall. The ablation results indicate that timestep-dependent weighting improves R2L and U2R detection relative to reconstruction-only scoring. The experiments show a performance difference, but they do not establish the underlying cause. Further work should test the weighting strategy across independent training runs and newer traffic datasets.

Several limitations apply. First, TabDiffAD was trained once, so its seed variation covers only stochastic scoring noise rather than training variability. Second, One-Class SVM used 15,000 training records for scaling feasibility, and baselines were not tuned, which may understate their individual ceilings. Third, NSL-KDD is old and exhibits distribution shifts between training and test sets, which KDDTest-21 amplifies through its selection criterion. Fourth, only 67 U2R test records exist, making U2R recall sensitive to single-sample shifts. Finally, one-hot categories are diffused continuously, and the score weight of 0.5 was taken without explicit validation search. Evaluation on newer flow datasets such as UNSW-NB15 or CIC-IDS2017 is planned.

In conclusion, TabDiffAD reached a ROC-AUC of 94.98% on KDDTest+ and detected 82.93% of attacks at a test FPR of 8.50%, with primary operating thresholds calibrated on benign validation data only. This places it second among five one-class models, behind a deep autoencoder. Applying timestep-dependent weighting to denoising residuals was the most useful design choice, especially for R2L and U2R traffic. The comparison with test-tuned oracle thresholds demonstrates how easily F1 values on KDDTest-21 can be inflated, highlighting the necessity of reporting realised FPR alongside any thresholded metric. The code, checkpoint and logs behind every table are available from the corresponding author on request.

References

Al-Harbi O, Hamed A, 2024. A dual-stage deep learning model based on a sparse autoencoder and layered deep classifier for intrusion detection with imbalanced data. International Journal of Sensor Networks, 45 (2): 74-86.

Alqarni AA, El-Alfy ESM, 2022. Improving intrusion detection for imbalanced network traffic using generative deep learning. International Journal of Advanced Computer Science and Applications, 13 (4): 959-967.

Behrendt F, Bhattacharya D, Maack L, Krüger J, Opfer R, Mieling R, Schlaefer A, 2024. Diffusion models with ensembled structure-based anomaly scoring for unsupervised anomaly detection. 2024 IEEE International Symposium on Biomedical Imaging, 27-30 May 2024, Page: 1-4, Athens, Greece.

Cui F, Wang X, Gu R, Zhang Q, Wang J, Song B, 2026. An improved anomaly traffic detection framework based on denoising diffusion implicit model. Journal of King Saud University - Computer and Information Sciences, 38 (5): 207.

Feng L, Yu J, Li S, Jiu S, 2025. VAE-SMOTE augmented diffusion for anomaly detection. Proceedings of the 2025 2nd International Conference on Generative Artificial Intelligence and Information Security, 21-23 February 2025, Page: 163-170, Hangzhou, China.

Fu Y, Du Y, Cao Z, Li Q, Xiang W, 2022. A deep learning model for network intrusion detection with imbalanced data. Electronics, 11 (6): 898.

Galwaduge V, Samarabandu J, 2025. Novel actionable counterfactual explanations for intrusion detection using diffusion models. Journal of Cybersecurity and Privacy, 5 (3): 68.

Gharghi H, Dehghani A, 2026. A leakage-free hybrid AE-GA-XGBoost model for network intrusion detection systems. Journal of Supercomputing, 82 (14): 707.

Ho J, Jain AN, Abbeel P, 2020. Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 6-12 December 2020, Page: 6840-6851, Virtual.

Kotelnikov A, Baranchuk D, Rubachev I, Babenko A, 2023. TabDDPM: Modelling tabular data with diffusion models. Proceedings of the 40th International Conference on Machine Learning, 23-29 July 2023, Page: 17564-17579, Honolulu, Hawaii, USA.

Lim H, Park S, Kim M, Lee J, Lim S, Park N, 2023. MadSGM: Multivariate anomaly detection with score-based generative models. 32nd ACM International Conference on Information and Knowledge Management, 21-25 October 2023, Page: 1411-1420, Birmingham, United Kingdom.

Liu FT, Ting KM, Zhou ZH, 2008. Isolation forest. 2008 Eighth IEEE International Conference on Data Mining, 15-19 December 2008, Page: 413-422, Pisa, Italy.

Nichol AQ, Dhariwal P, 2021. Improved denoising diffusion probabilistic models. Proceedings of the 38th International Conference on Machine Learning, 18-24 July 2021, Page: 8162-8171, Virtual.

Pamuditha P, Hettiarachchi S, Gunarathne Y, Moremada C, Sandeepa C, Liyanage M, 2025. Enhancing intrusion detection in 5G networks: A comparative study of generative adversarial and diffusion networks. 2025 IEEE International Black Sea Conference on Communications and Networking, 23-26 June 2025, Page: 1-6, Chisinau, Moldova.

Ponnarasan DSS, Manimegalai R, Mohan V, Susmeta A, 2025. Unsupervised anomaly detection for network intrusion using deep autoencoders and statistical modeling. 2025 International Conference on Next Generation Computing Systems: Intelligent System for Sustainable Development, 21-22 August 2025, Page: 1-10, Coimbatore, India.

Ruff L, Vandermeulen R, Goernitz N, Deecke L, Siddiqui SA, Binder A, Müller E, Kloft M, 2018. Deep one-class classification. Proceedings of the 35th International Conference on Machine Learning, 10-15 July 2018, Page: 4393-4402, Stockholm, Sweden.

Sagaran E, Spier J, Hasan R, 2025. SpiCAE: Spiking contrastive learning and autoencoder for network intrusion detection. 2025 IEEE 16th Annual Information Technology, Electronics and Mobile Communication Conference, 29-31 October 2025, Page: 7-13, Berkeley, California, USA.

Schölkopf B, Platt JC, Shawe-Taylor J, Smola AJ, Williamson RC, 2001. Estimating the support of a high-dimensional distribution. Neural Computation, 13 (7): 1443-1471.

Sommer R, Paxson V, 2010. Outside the closed world: On using machine learning for network intrusion detection. 2010 IEEE Symposium on Security and Privacy, 16-19 May 2010, Page: 305-316, Oakland, California, USA.

Song Y, Hyun S, Cheong YG, 2021. Analysis of autoencoders for network intrusion detection. Sensors, 21 (13): 4294.

Tavallaee M, Bagheri E, Lu W, Ghorbani AA, 2009. A detailed analysis of the KDD CUP 99 data set. 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, 8-10 July 2009, Page: 1-6, Ottawa, Canada.

Tritscher J, Lissmann P, Wolf M, Krause A, Hotho A, Schlör D, 2024. Generative inpainting for Shapley-value-based anomaly explanation. Communications in Computer and Information Science, 2153: 230-243.

Wang Y, Ding J, He X, Wei Q, Yuan S, Zhang J, 2024. Intrusion detection method based on denoising diffusion probabilistic models for UAV networks. Mobile Networks and Applications, 29 (5): 1467-1476.

Zhang Z, Wang P, Zhang T, Liu M, Zhou X, 2025. Trustworthy generative few-shot learning-based intrusion detection method in Internet of Things. IEEE Transactions on Consumer Electronics, 71 (1): 1992-2002.

Conflict of Interest

The authors have declared that there is no conflict of interest.

Author Contributions

Both authors contributed equally to the conception, methodology, experiments and writing of this study, and both authors read and approved the final manuscript.

Information Technologies and Data-Informed Decision-Making in Education: Opportunities and Limitations

Reşat Timuçin Özkan[1]*

[1] University of Mediterranean Karpasia, Department of Educational Administration, 99010 Nicosia, TRNC

Corresponding author e-mail: timucin.ozkan@akun.edu.tr

Abstract

Information technologies offer opportunities to support teaching, monitor student progress, and inform administrative decisions in educational institutions. However, collecting educational data does not, by itself, ensure effective decision-making. Data must be connected to pedagogical goals, interpreted within their context, and used in accordance with ethical principles. This paper aims to examine the opportunities and limitations of information technologies in education from the perspective of data-informed decision-making.

The paper adopts a conceptual approach, addressing learning management systems, digital assessment tools, and student information systems in relation to teaching and educational administration. It discusses how attendance, participation, assignment completion, and assessment data may help identify students' support needs. Its central argument is that digital indicators cannot fully represent students' learning processes. For example, time spent on a learning platform does not necessarily reflect the quality of learning. Digital data should therefore be interpreted alongside teacher observations, student perspectives, and the characteristics of the educational setting.

The discussion highlights unequal access to technology, data quality, teacher and administrator data literacy, and the protection of personal information. It argues that the educational value of information technologies should be assessed through their contribution to learning and informed decision-making. To strengthen data-informed practices, educational institutions should integrate digital infrastructure with pedagogical expertise, address differences in access, and establish ethical principles for data use. The paper concludes that educational data can support professional judgement when interpreted critically and with attention to context.

Keywords: *Information Technologies, Educational Administration, Data-Informed Decision-Making, Data Literacy, Digital Inequality*

INTRODUCTION

The use of information technologies in education raises an important question about how educational evidence is produced and interpreted. Learning management systems, digital assessment tools, and student information systems can make information about attendance, participation, assignment completion, and assessment performance available to teachers and educational administrators. These

records may support the identification of difficulties and the planning of educational responses. However, the availability of digital information does not, by itself, establish its educational meaning or the appropriateness of a decision based on it.

The value of educational technology therefore needs to be considered in relation to the purposes and conditions of its use. UNESCO's Global Education Monitoring Report emphasises the importance of evaluating technology in terms of educational needs, equity, evidence, and sustainability (UNESCO, 2023). This perspective provides a useful basis for examining digital systems as resources for educational decision-making. Their contribution depends on the problems they help address, the learners they serve, and the capacity of educational institutions to use them responsibly.

In this article, data-informed decision-making refers to the use of relevant evidence alongside professional judgement and knowledge of the educational context. This distinction is important because digital indicators provide partial representations of educational activity. For example, a record of assignment completion identifies whether a task was submitted, but it does not necessarily explain the learner's understanding, the support received, or the circumstances affecting completion. Similarly, limited activity within an online platform may prompt further inquiry, but it should not automatically be interpreted as a lack of motivation. These examples illustrate why the interpretation of digital records requires attention to what an indicator measures and what remains outside its scope.

Educators' capacity to interpret evidence is central to this process. Mandinach and Gummer (2016) conceptualise teacher data literacy through the knowledge, skills, and dispositions involved in educational data use. Their work supports an examination of decision-making that extends beyond accessing information or reading numerical summaries. Educationally meaningful interpretation requires a connection between the evidence available, the question being investigated, and the response being considered.

Against this background, the present article examines the opportunities and limitations of information technologies in supporting data-informed decision-making in education. It focuses on learning management systems, digital assessment tools, and student information systems, with particular attention to attendance, participation, assignment completion, and assessment data. The discussion considers how these forms of evidence can support teaching and educational administration while also examining the constraints associated with incomplete indicators, unequal access, data quality, educators' interpretive capacity, and privacy.

The article addresses three questions: What opportunities do information technologies offer for educational decision-making? What limitations should be considered when interpreting digitally generated indicators? What principles can guide their ethical and context-sensitive use? Through this conceptual discussion, the article develops an argument for connecting digital evidence with pedagogical understanding and accountable professional judgement.

Conceptual background

INFORMATION TECHNOLOGIES IN EDUCATION

Information technologies in education encompass systems used to organise educational activities, facilitate communication, document student progress, and support institutional management. This article focuses on three categories: learning management systems, digital assessment tools, and student information systems. Although these systems serve different purposes, each can provide information relevant to teaching and educational administration.

Learning management systems organise course materials, assignments, feedback, and online activities. Their records can help educators examine patterns of participation and task completion. Digital assessment tools provide information about responses to particular questions or tasks, potentially helping teachers identify areas that require further explanation. Student information systems bring together administrative records, such as attendance and grades, which may support the coordination of student services and institutional planning.

The educational significance of these systems depends on how their records are interpreted and used. A platform can record that a student opened a resource, but that record does not establish whether the student understood its content. Likewise, a numerical assessment result may indicate a difficulty without explaining its cause. Information technologies should therefore be understood as sources of evidence whose relevance and limitations need to be examined in relation to a particular educational question. Gašević et al. (2015) similarly argue that computational approaches to learning analytics should be connected with educational research.

DATA-INFORMED DECISION-MAKING

Data-informed decision-making involves considering relevant evidence alongside professional judgement, educational goals, and contextual knowledge. In this article, the term emphasises the contribution of data to deliberation rather than the automatic determination of decisions by numerical indicators.

An educational decision begins with a question. A teacher may wish to understand why students are struggling with a particular topic, while a school administrator may seek to identify factors associated with repeated absence. The question determines which information is relevant and whether existing records are sufficient. Collecting more data without clarifying the decision to be made can produce additional information without improving understanding.

The process also requires a distinction between description and explanation. Attendance records describe patterns of presence and absence; they do not independently explain those patterns. Assignment records show submission status; they do not necessarily reveal the barriers encountered by learners. Consequently, digital evidence may need to be considered alongside classroom observations, student work, and conversations with students.

Drawing on the iterative perspective described by Schildkamp (2019), this discussion organises data-informed decision-making around five activities: defining the educational question, selecting relevant evidence, interpreting that evidence in context, choosing an appropriate response, and reviewing the consequences of the response. This sequence provides an organising structure for the article; it is not presented as a newly developed or empirically validated model.

DIGITAL INDICATORS AND EDUCATIONAL MEANING

Digital indicators are selected representations of educational activities or outcomes. Attendance, platform participation, assignment completion, and assessment results each make particular aspects of student experience visible. Their usefulness depends on the relationship between the recorded activity and the educational concept being considered.

Participation, for example, can refer to several different behaviours. Posting a message in an online discussion differs from contributing to a classroom conversation, collaborating with peers, or engaging in independent study. A system that records one behaviour cannot be assumed to capture participation in its entirety. Platform access frequency should therefore be treated as a behavioural indicator whose educational meaning requires further interpretation.

Interpretation must also account for the circumstances in which records are produced. A student with unreliable internet access may leave fewer digital traces than a student with continuous access, even when both make substantial efforts to learn. Missing or limited records can therefore raise questions about access and data collection as well as questions about educational engagement.

These distinctions suggest that digital indicators are most useful when their scope is explicit. Educators need to establish what a measure represents, what it excludes, and which additional evidence is necessary before taking action. An indicator may justify further inquiry without being sufficient to justify a consequential decision about a student. Wise and Shaffer (2015) emphasise the role of theory in selecting variables and interpreting analytical results. The examples here apply that general concern to routine educational records.

PROFESSIONAL JUDGEMENT AND ETHICAL RESPONSIBILITY

Professional judgement connects evidence with educational action. Teachers and administrators must assess whether information is relevant, whether an interpretation is reasonable, and whether a proposed response serves an appropriate educational purpose. This responsibility remains important when digital systems produce summaries, classifications, or alerts.

Ethical considerations form part of that judgement. Institutions should consider why particular data are needed, who can access them, and how their use may affect students. Decisions also require attention to the possibility of misinterpretation. A label derived from incomplete records may influence expectations about a learner even when the underlying evidence is uncertain.

The framework adopted in this article therefore connects three elements: digital evidence, contextual interpretation, and professional responsibility. Information technologies contribute records and analytical possibilities; educational practitioners assess their meaning and decide how to respond. This relationship provides the basis for examining both the opportunities and the limitations discussed in the following sections.

MATERIAL AND METHODS

Conceptual approach

This article adopts a conceptual approach to examine how information technologies can support data-informed decision-making in education and under what conditions their use may be limited. Its purpose is to clarify the relationship between digitally generated evidence, educational interpretation, and professional responsibility. The article does not report original participant data, test a technological intervention, or estimate the effects of digital systems on student achievement.

The discussion is organised around three connected questions: what educational decisions digital information can support, what limitations affect the interpretation of that information, and what principles can guide its responsible use. This approach allows the opportunities offered by information technologies to be considered alongside the conditions necessary for their educational value.

Scope of the discussion

The article focuses on learning management systems, digital assessment tools, and student information systems. These categories provide a manageable scope for examining the use of attendance, participation, assignment completion, and assessment records in teaching and educational administration.

Two levels of decision-making are considered. At the instructional level, the discussion addresses how teachers may use available evidence to identify learning difficulties, adjust teaching activities, and provide feedback or support. At the institutional level, it considers how administrators may interpret aggregated records to coordinate student support and inform planning.

The scope does not include a technical comparison of commercial platforms or an evaluation of particular predictive algorithms. References to digital indicators concern their educational interpretation rather than their computational performance. The article also distinguishes between information that identifies a pattern and evidence that explains the reasons for that pattern.

Literature basis

This focused conceptual discussion uses selected publications on educator data literacy (Mandinach and Gummer, 2016), school improvement through data use (Schildkamp, 2019), feedback (Hattie and Timperley, 2007), ethical issues in learning analytics (Slade and Prinsloo, 2013), educational technology policy (UNESCO, 2023), and the educational and theoretical interpretation of learning analytics (Gašević et al., 2015; Wise and Shaffer, 2015). Sources were selected for their direct conceptual relevance to these topics. This purposive selection does not constitute an exhaustive corpus or systematic search.

The selected literature clarifies concepts and provides perspectives for examining opportunities, limitations, and practical implications. Conceptual publications and review-based discussions inform the argument, while UNESCO's report provides an institutional policy perspective. These sources are used to develop a reasoned discussion; their inclusion does not establish the effectiveness of a specific platform or intervention.

The article is a focused conceptual discussion rather than a systematic review. It does not claim exhaustive coverage of the literature, a reproducible database search, or a formal assessment of all

available studies. Interpretations are limited to the concepts and arguments supported by the selected sources. The illustrative examples are author-developed scenarios, not empirical observations.

Analytical organisation

The discussion examines each opportunity in relation to its interpretive requirements and possible limitations. For example, the potential use of participation records to identify students who may need support is considered alongside questions about access, the meaning of participation, and the completeness of the records.

Four dimensions organise this examination: the educational purpose of the decision, the adequacy of the available evidence, the contextual conditions affecting interpretation, and the responsibilities associated with action. These dimensions form an author-proposed structure for the conceptual argument; they are not presented as a validated measurement instrument or an empirical coding scheme.

This organisation supports a balanced examination of technological possibilities. It directs attention both to what digital systems make visible and to the additional understanding required before their outputs can inform educational action.

Limitations of the approach

The conceptual nature of the article limits the conclusions that can be drawn. The discussion cannot establish that a particular technology improves learning, reduces inequality, or increases administrative effectiveness. Such claims require empirical investigation in defined educational settings.

The examples used in the article illustrate interpretive problems and possible responses. They are not observations collected from participants or evidence of the frequency of particular practices. Furthermore, the relevance of the proposed considerations may vary according to educational level, institutional resources, technological infrastructure, and local governance arrangements.

Accordingly, the article's contribution lies in clarifying the conditions under which digital evidence may inform educational decisions and identifying issues that require further empirical examination.

RESULTS

The results presented below are the outcome of the conceptual synthesis of the selected literature. They identify potential opportunities and limitations; they are not findings from newly collected participant data or estimates of intervention effects.

Opportunities of information technologies for data-informed decision-making

Information technologies can support educational decision-making by making relevant records available for interpretation and action. Their potential contribution extends from identifying students who may need assistance to reviewing instructional activities and informing institutional planning. The opportunities examined below depend on the relevance and quality of the available evidence, educators' interpretive capacity, and institutions' ability to respond. They should therefore be understood as possible uses rather than guaranteed improvements in educational outcomes.

IDENTIFYING STUDENTS WHO MAY NEED SUPPORT

Attendance, assignment completion, assessment results, and online participation records can provide starting points for identifying students who may benefit from additional support. Examining these records together may help educators recognise patterns that deserve attention. For example, repeated absences accompanied by missed assignments could prompt a teacher to contact a student and explore the circumstances affecting participation.

The opportunity lies in using available information to initiate timely inquiry. A digital record does not independently establish why a student is experiencing difficulty. Limited participation may reflect problems with access, uncertainty about a task, competing responsibilities, or other circumstances. Consequently, the interpretation of a pattern should be followed by dialogue and consideration of additional evidence.

Digital systems may also support continuity in student assistance. Where access permissions and institutional procedures allow, relevant staff can use shared records to coordinate their responses and review whether support has been provided. The educational value of this coordination depends on clear responsibilities and an emphasis on assistance rather than the automatic classification of students.

INFORMING INSTRUCTIONAL ADJUSTMENT

Digital assessment tools and learning management systems can provide information that helps teachers reconsider the content, pace, and organisation of instruction. Responses to assessment tasks may indicate topics that require further explanation, while assignment records can help identify stages at which students encounter difficulties.

For example, when several students select the same incorrect response, the teacher may examine whether the pattern reflects a shared misconception, ambiguous wording, or insufficient prior instruction. Each explanation suggests a different response. Reviewing student work and discussing their reasoning can help determine whether to revise the assessment item, revisit a concept, or provide additional practice.

Information can also be considered across activities. Comparing students' initial responses with later work may help teachers judge whether an instructional adjustment appears to address the difficulty identified. Such comparisons require attention to differences in task demands and assessment conditions. A change in scores alone does not establish that a particular teaching strategy caused the change.

The main opportunity is therefore a closer connection between evidence and instructional reflection. Teachers can use digital records to formulate questions about learning and select responses suited to the difficulties they identify.

SUPPORTING FEEDBACK AND STUDENT REFLECTION

Digital tools can help organise feedback and make it available to students alongside their work. Depending on the system, teachers may provide comments, assessment criteria, or opportunities for revision. These functions can support a feedback process in which students examine their performance and identify possible next steps.

The usefulness of feedback depends on its educational content. A score or an automated indication that an answer is incorrect may provide limited guidance. Feedback becomes more actionable when it explains the difficulty, relates it to the learning objective, and suggests how the student can improve. Automated responses may assist with some tasks, while more complex work may require teacher interpretation and individual explanation. The distinction between feedback delivery and feedback quality is consistent with Hattie and Timperley's (2007) discussion of how feedback effects depend on its form and use.

Students may also use records of their assignments and assessments to reflect on their progress. A teacher-guided review could invite students to identify recurring difficulties, compare strategies, and set a specific learning goal. Here, the digital record serves as a resource for reflection. Its purpose is to support understanding and action rather than to encourage constant comparison with peers.

Opportunities for feedback should also account for students' ability to access, understand, and use the information provided. Making feedback available within a platform does not establish that students have read it or know how to respond to it.

INFORMING INSTITUTIONAL PLANNING AND COORDINATION

At the institutional level, aggregated records can inform questions about attendance, student support, assessment practices, and resource allocation. Administrators may examine patterns across classes or periods to identify issues requiring further investigation. For instance, a concentration of incomplete assignments across several courses may prompt a review of assessment schedules and workload expectations.

Such information can help institutions move from a broad concern to a more clearly defined planning question. However, aggregated patterns require careful interpretation. Differences between classes may reflect student composition, subject demands, recording practices, or access conditions. Comparisons should therefore consider whether the underlying measures are sufficiently consistent and comparable.

Digital systems can also support coordination by bringing relevant information into a shared administrative process. A planning meeting may combine attendance summaries, teachers' observations, and student feedback to determine where support is needed. The resulting decision should specify an intended response, responsible staff, and a suitable point for review.

Resource decisions require particular care because recorded demand may differ from actual need. Students whose difficulties are poorly captured by existing systems can remain less visible. Institutional planning should therefore allow additional routes through which teachers and students can communicate needs that digital records do not adequately represent.

REVIEWING EDUCATIONAL RESPONSES

Information technologies can provide records that help educators review what happens after a decision is implemented. Following a change in teaching arrangements or the introduction of student support, relevant information may help determine whether the intended issue persists and whether further adjustment is required.

For example, an institution responding to missed assignments might review subsequent submission patterns while also asking students and teachers about their experience of the response. Combining these forms of evidence can provide a more informative account than relying on submission rates alone.

This review process should distinguish improvement in a recorded indicator from improvement in the broader educational experience. Increased platform activity, for instance, may show a change in recorded behaviour without demonstrating stronger understanding. Reviewing a response therefore requires evidence aligned with its educational purpose and attention to possible unintended consequences.

Taken together, these opportunities suggest that information technologies can support a cycle of inquiry, action, and review. Their contribution is strongest when records are connected to clear educational questions, interpreted with contextual knowledge, and followed by appropriate professional responses. The following section examines the limitations that can weaken or complicate this process.

Limitations of information technologies for data-informed decision-making

The opportunities discussed above depend on how digital records are produced, interpreted, and translated into educational action. Information technologies capture selected aspects of learning and institutional activity, leaving other aspects less visible. Their limitations therefore concern both the adequacy of the available evidence and the conditions under which decisions are made.

PARTIAL REPRESENTATION OF LEARNING AND PARTICIPATION

Digital records represent activities that a system is designed to capture. Platform access, assignment submission, and assessment responses can provide useful information, but they do not encompass the full range of learning experiences. Students may engage in independent reading, face-to-face discussion, handwritten practice, or collaborative work that remains outside the platform's records. This discussion follows the broader requirement to connect analytical measures with educational understanding rather than assume that records explain learning on their own (Gašević et al., 2015; Wise and Shaffer, 2015).

This creates a distinction between a recorded behaviour and the educational meaning attributed to it. Opening a resource does not establish comprehension, while spending more time on a task does not necessarily indicate greater engagement. A longer recorded duration might reflect careful study, difficulty understanding instructions, or an unattended session. Interpretation requires consideration of these alternative explanations.

The same limitation applies to participation measures. A student who contributes frequently to an online discussion may be highly active, but the number of contributions alone does not establish their relevance or depth. Conversely, a student who posts less frequently may still participate meaningfully through other activities. Educational decisions should therefore examine the relationship between an indicator and the concept it is intended to represent.

DATA QUALITY AND COMPARABILITY

The usefulness of digital evidence depends on the accuracy, completeness, timeliness, and consistency of the records. Missing entries, incorrect information, delayed updates, or different recording practices

can affect the patterns presented to educators. A missing assignment record, for example, may indicate non-submission, a technical problem, or an incomplete administrative entry.

Comparability also requires attention. Attendance, participation, and assessment measures may be defined differently across courses or institutions. One teacher may record participation through discussion contributions, while another may use task completion. Combining these measures into a single summary can conceal differences in what they represent.

Assessment results present similar difficulties. Scores from tasks with different content, levels of difficulty, or marking criteria cannot automatically be treated as equivalent. Before drawing conclusions from comparisons, educators need to consider whether the records address sufficiently similar questions.

These limitations make verification an essential part of interpretation. When a pattern could lead to a consequential decision, the underlying information should be checked and uncertainties acknowledged. Apparent precision in a dashboard does not remove uncertainty in the records from which it was produced.

UNEQUAL ACCESS AND UNEQUAL VISIBILITY

Students' opportunities to participate in digital activities may vary according to device availability, internet connectivity, accessibility requirements, and the conditions in which they study. UNESCO's discussion of educational technology identifies equitable access as an important condition for its educational use (UNESCO, 2023).

These differences can affect the digital traces available for decision-making. A student who shares a device or experiences unstable connectivity may access a platform less frequently or submit work later. If these records are interpreted without considering access conditions, a difficulty in using the technology may be treated as evidence of insufficient effort.

Unequal access also raises questions about whose experiences become visible. Students who complete activities outside the platform may contribute fewer records, even when they remain engaged in learning. Institutional decisions based predominantly on digital activity could consequently overlook needs that are poorly represented in the system.

Addressing this limitation requires opportunities for students to explain their circumstances and demonstrate learning through appropriate alternatives. It also requires institutions to examine whether the use of a particular digital indicator disadvantages students whose participation conditions differ.

DATA LITERACY AND INTERPRETIVE CAPACITY

Access to information does not ensure that educators can interpret it appropriately. Data literacy involves the knowledge, skills, and dispositions needed to connect evidence with educational questions and possible responses (Mandinach and Gummer, 2016). Reading a graph or identifying a change in a score forms only part of this process.

A central challenge is distinguishing a pattern from an explanation. When lower assessment results occur alongside limited platform activity, the association does not establish that limited activity caused

the lower results. Both may be related to other circumstances, including prior learning difficulties or barriers to access.

Interpretation can also be influenced by existing expectations. An educator may give greater attention to records that support an initial judgement while overlooking contradictory evidence. Digital summaries can reinforce this tendency when their categories appear authoritative or when the basis of a classification is unclear.

Professional learning should therefore address question formulation, the evaluation of indicators, the consideration of alternative explanations, and the communication of uncertainty. Opportunities for teachers and administrators to discuss evidence together may also help expose assumptions that would otherwise remain unexamined. Such discussion requires time and organisational support.

PRIVACY AND RESPONSIBLE DATA USE

The collection and use of educational records require clear decisions about purpose, access, retention, and responsibility. Information gathered to support learning may also be used for administrative monitoring or evaluation. These uses need to be considered explicitly because they can involve different expectations and consequences. These questions connect with the ethical concerns regarding privacy, interpretation, and transparency identified in higher education learning analytics by Slade and Prinsloo (2013).

Institutions should establish which information is necessary for a defined purpose and which staff require access to it. Broad access to detailed student records is not automatically justified by the availability of a shared system. Similarly, retaining information indefinitely may extend its use beyond the educational purpose for which it was collected.

Students and, where appropriate, their parents should receive understandable information about relevant data practices. Procedures for identifying and correcting inaccurate records are also important, particularly when those records contribute to decisions about support, assessment, or participation.

Privacy is connected to the way information is interpreted. A student should have an appropriate opportunity to explain circumstances that a record does not capture. Responsible data use therefore involves both the protection of information and fair consideration of the person represented by it.

DEPENDENCE ON DIGITAL SUMMARIES AND CLASSIFICATIONS

Dashboards, automated alerts, and summary categories can make large quantities of information easier to navigate. However, they can also direct attention towards the indicators selected by the system. Educational concerns that are difficult to quantify may receive less attention because they are less prominent in the interface.

A classification indicating that a student may require support should prompt inquiry rather than settle a judgement. Its meaning depends on the underlying indicators, the criteria used, and the circumstances of the student. When these elements are unclear, educators may find it difficult to assess whether the classification is relevant.

There is also a risk that institutional priorities become aligned with improving visible indicators rather than addressing the educational purposes behind them. Increasing recorded participation, for example, does not necessarily improve the quality of learning. Institutions should periodically review whether their chosen measures remain aligned with their educational goals.

CAPACITY TO TRANSLATE EVIDENCE INTO ACTION

Even relevant and well-interpreted information may have limited value when institutions lack the capacity to respond. Identifying students who need assistance does not ensure that appropriate staff, time, or support services are available. The usefulness of a digital system consequently depends in part on the resources and procedures surrounding it. The importance of leadership and organisational conditions in data use is also identified by Schildkamp (2019).

Data-related activities can also create additional work. Entering records, checking their accuracy, reviewing summaries, and discussing possible responses require time. If these tasks are poorly integrated into existing practices, they may compete with teaching and student support.

Institutions therefore need to connect information collection with a realistic response process. This includes clarifying who reviews the evidence, who takes action, and how the response will be evaluated. Collecting information without these arrangements may increase administrative activity while producing little educational benefit.

Together, these limitations demonstrate why digital evidence requires contextual interpretation and accountable professional judgement. The following discussion brings the opportunities and limitations together to consider the conditions for meaningful educational use.

DISCUSSION AND CONCLUSION

Connecting digital evidence with educational purposes

The conceptual examination suggests that the contribution of information technologies depends on the relationship between the information collected and the educational decision under consideration. Attendance, participation, assignment completion, and assessment records can identify patterns deserving attention. Their value increases when educators can explain why a pattern matters and what further evidence is needed before responding.

Schildkamp (2019) describes data use for school improvement as an iterative process involving goals, evidence, interpretation, action, and evaluation. The sequence discussed in this article is consistent with that perspective. The present discussion applies it to the interpretation of records from learning management systems, digital assessment tools, and student information systems. This application is a conceptual organisation of the argument, not a newly validated decision-making model.

A practical implication is that institutions should define the educational question before selecting indicators. When the concern is students' understanding of a concept, platform activity may provide contextual information, while examination of student reasoning may offer more direct evidence. When the concern is access to learning activities, participation records may need to be interpreted alongside

information about connectivity and device availability. The suitability of a measure therefore depends on the purpose for which it is used.

Professional capacity and institutional responsibility

The opportunities and limitations are closely connected. A record that helps identify a possible learning difficulty can also lead to an inappropriate judgement when its scope is misunderstood. Similarly, a shared information system can support coordination while creating additional responsibilities for access control and accuracy. These relationships suggest that technological provision and professional preparation should be considered together.

Mandinach and Gummer (2016) identify educator data literacy as a combination of skills, knowledge, and dispositions. In the context considered here, this means that professional development should extend beyond operating a platform to evaluating the educational meaning of its records. Teachers need opportunities to formulate questions, examine alternative explanations, and decide when additional evidence is required. Administrators need to support this work through time, clear responsibilities, and feasible responses to the needs identified.

Digital feedback provides a further example of the importance of professional capacity. Hattie and Timperley (2007) explain that feedback can have different effects depending on its type and the way it is provided. A platform's ability to deliver information should therefore be distinguished from the pedagogical quality of that information. The practical question is whether students can understand the feedback and use it to improve their subsequent work.

Equity and ethical interpretation

UNESCO (2023) places educational needs, equity, and contextual suitability at the centre of decisions about educational technology. Applied to data-informed decision-making, this perspective requires educators to examine the circumstances in which digital records are produced. An apparent lack of participation may reflect an access barrier, and a record that is technically accurate may still support an incomplete interpretation.

Ethical interpretation also involves the student's opportunity to explain or question a judgement. Slade and Prinsloo (2013), writing about learning analytics in higher education, identify privacy, transparency, power, and context as important ethical concerns. Their perspective informs this article's argument that student records should be used with attention to the people and circumstances they represent. Applying these considerations across educational levels requires sensitivity to students' age and institutional responsibilities; the higher education discussion does not establish a universal procedure for every setting.

The implication developed here is that student support should include a route for clarifying uncertain or inaccurate information. Where a digital indicator prompts concern, educators should consider an appropriate conversation or review before making a consequential decision. This recommendation follows from the conceptual argument rather than from an intervention evaluated in this study.

Contribution and directions for research

The article's contribution is an integrated discussion of routine digital indicators across instructional and administrative decisions. It connects each possible use with four interpretive requirements: educational purpose, evidence adequacy, contextual conditions, and responsibility for action. This synthesis applies established concerns to attendance, participation, assignment completion, and assessment records. Its value is conceptual clarification; it does not claim a new validated theory or demonstrate effects on educational outcomes.

Further research could examine how teachers and administrators interpret the same digital indicators, what additional evidence they seek, and how their decisions differ across access conditions. Studies could also investigate whether students understand how their records are used and whether they can correct or contextualise them. Evaluations of student support should examine both recorded outcomes and students' experiences, while documenting the resources required for implementation.

Conclusion and recommendations

Information technologies offer opportunities to support student assistance, instructional adjustment, feedback, coordination, and institutional planning. Learning management systems, digital assessment tools, and student information systems can provide evidence relevant to these activities. Their educational contribution nevertheless depends on the appropriateness of the indicators, the quality of the records, and the capacity to interpret and act on them.

The central conclusion of this conceptual discussion is that digital evidence requires pedagogical and contextual interpretation. Attendance, participation, assignment completion, and assessment results represent selected aspects of educational activity. They may justify further inquiry, but they do not independently explain students' circumstances or establish the effectiveness of an educational response.

The discussion supports five practical recommendations. First, educators should define the educational question and identify what evidence is relevant before relying on a digital summary. Second, records should be checked for accuracy, completeness, and comparability, particularly when they inform consequential decisions. Third, interpretation should consider access conditions and incorporate additional evidence, including student work, observations, and appropriate dialogue. Fourth, institutions should establish clear arrangements for data access, correction, retention, and responsibility. Fifth, they should provide professional learning and sufficient resources to implement and review the responses selected.

These recommendations form a conceptual basis for practice and further investigation. They have not been tested as a package of interventions. Future empirical work should examine how the proposed considerations function in different educational settings and whether they support fairer interpretation and more effective assistance. The responsible use of information technologies ultimately involves connecting evidence with a clear educational purpose, informed professional judgement, and attention to the consequences for learners.

References

Gašević D, Dawson S, Siemens G, 2015. Let's not forget: Learning analytics are about learning. TechTrends, 59 (1): 64–71. https://doi.org/10.1007/s11528-014-0822-x

Hattie J, Timperley H, 2007. The power of feedback. Review of Educational Research, 77 (1): 81–112. https://doi.org/10.3102/003465430298487

Mandinach EB, Gummer ES, 2016. What does it mean for teachers to be data literate: Laying out the skills, knowledge, and dispositions. Teaching and Teacher Education, 60: 366–376. https://doi.org/10.1016/j.tate.2016.07.011

Schildkamp K, 2019. Data-based decision-making for school improvement: Research insights and gaps. Educational Research, 61 (3): 257–273. https://doi.org/10.1080/00131881.2019.1625716

Slade S, Prinsloo P, 2013. Learning analytics: Ethical issues and dilemmas. American Behavioral Scientist, 57 (10): 1510–1529. https://doi.org/10.1177/0002764213479366

UNESCO, 2023. Global education monitoring report 2023: Technology in education – A tool on whose terms? UNESCO, Paris, France. https://unesdoc.unesco.org/ark:/48223/pf0000385723

Wise AF, Shaffer DW, 2015. Why theory matters more than ever in the age of big data. Journal of Learning Analytics, 2 (2): 5–13. https://doi.org/10.18608/jla.2015.22.2

Conflıct Of Interest

The author declares that there is no conflict of interest.

Generation of a Unified Analysis Surface from Drone Video for Analysing Unused Road Surfaces at Intersections

Mustafa Kayademir[1*], Ertuğrul Gül[2], Murat Taşyürek[3]

[1]Kayseri University, Vocational School of Information Technologies, Department of Electronics and Automation, 38280, Kayseri, Türkiye
[2]Kayseri University, Faculty of Engineering, Architecture and Design, Department of Software Engineering, 38280, Kayseri, Türkiye
[3]Kayseri University, Faculty of Engineering, Architecture and Design, Department of Computer Engineering, 38280, Kayseri, Türkiye

*Corresponding author e-mail: mustafakayademir@kayseri.edu.tr

Abstract

Identifying unused road surfaces at intersections is important for evaluating road space utilisation. Such analyses require images captured at different times to be aligned with a reference frame and the road surface to be analysed to be clearly defined. This study aims to generate a unified analysis surface from drone video frames to support subsequent analyses of road surfaces around an intersection. To this end, frames were sampled at approximately uniform intervals from a selected segment of the video recording and aligned with a reference frame. Frames that met the acceptance criteria were selected and combined to create a composite image. The composite image was segmented into superpixels using Simple Linear Iterative Clustering (SLIC), and the road surface was delineated by a researcher through the selection of relevant superpixel regions. The observation area within the intersection was independently delineated using an observation mask. The intersection of the road mask and the observation mask yielded the unified analysis surface. The experimental results yielded 5,684 superpixel regions, of which 811 were selected to represent the road surface. The road mask, comprising 292,698 pixels, and the observation mask, comprising 310,697 pixels, overlapped to produce a unified analysis surface of 247,010 pixels. The proposed approach provides a practical method for generating a unified analysis surface from drone imagery for subsequent intersection analyses.

Keywords: *Drone Video, Intersection Analysis, Unused Road Surface, Unified Analysis Surface*

INTRODUCTION

Examining the areas used by vehicles at intersections and their movements over time is important for evaluating intersection layouts and traffic movements (Apeltauer et al., 2015). Drone videos provide an important source of data for examining intersection geometry and interactions between road users by allowing the intersection centre, approach roads, and different road users to be observed within the same image sequence (Bock et al., 2020). However, effective use of these recordings in transport research requires not only detecting road users in the images but also evaluating the images against a reference. Camera position changes caused by drone motion may cause the same physical points to appear at different pixel locations in successive frames. Images therefore need to be aligned with a reference frame before observations from different times are combined (Fonod et al., 2025). Shadows, occlusions, and

non-road regions with similar visual characteristics also complicate road surface extraction (Lu and Weng, 2025).

A fundamental requirement for examining unused road surfaces at intersections is to define the road area to be analysed clearly and consistently. This definition cannot be reduced to identifying regions containing vehicles, because the absence of a vehicle does not mean that a region lies outside the road surface. Similarly, a region's inclusion in the road surface does not establish that it is unused. Defining the road surface and assessing its use are therefore separate problems. The unified analysis surface established at the initial stage allows subsequent investigations to use the same spatial extent and explicitly defines the boundaries of the assessment area.

Image alignment, composite image generation, and road surface identification should consequently be treated as complementary operations with distinct functions. Image alignment registers frames recorded at different times to a reference frame, whereas compositing combines their observations into a single image. Methods that combine observations from frames recorded at different times have been used to construct background images from video sequences (Hsiao and Leou, 2013). However, generating a composite image does not automatically identify the road surface. A distinction based solely on pixel similarity may be insufficient, particularly in scenes containing regions with similar colours and textures, shadows, and occlusions (Lu and Weng, 2025). Image compositing and selection of the analysis surface should therefore be considered separate stages.

Simple Linear Iterative Clustering (SLIC) groups image pixels into regions called superpixels according to colour similarity and spatial proximity, enabling region-level examination of the image (Achanta et al., 2012). SLIC has also been investigated for delineating objects such as roads and roofs in UAV imagery, where it was found to be insufficient as a stand-alone method (Crommelinck et al., 2017). However, superpixel segmentation does not itself classify regions as road or non-road. In this study, a researcher therefore assesses the superpixel regions and identifies those belonging to the road surface. The spatial extent of the intersection under investigation is delineated using an observation polygon, which is converted into an observation mask. The intersection of the observation mask and the road mask generated from the selected superpixel regions defines the unified analysis surface. This process identifies the selected road surface regions and establishes the spatial boundaries of the area to be analysed.

This study aims to present a traceable data preparation approach for generating a unified analysis surface from drone videos as a basis for subsequent investigations of unused road surfaces at intersections. Frames are sampled at approximately uniform intervals from a selected video interval, aligned with a reference frame, and combined to produce a composite image. Frames that meet the acceptance criteria are used to generate the composite image. SLIC is then applied to segment the composite image into superpixels, after which a researcher selects the regions corresponding to the road surface. The analysis extent is delineated using an observation polygon, which is converted into an observation mask. The intersection of the road mask and the observation mask yields the unified analysis surface. The study focuses on generating the analysis surface for subsequent investigations of unused road surfaces rather than directly detecting unused road surfaces.

MATERIALS AND METHODS

The study describes the drone video material and the processing steps used to generate a unified analysis surface. The Materials section presents the technical characteristics of the video recording and the observed scene. The Methods section describes reference frame selection, image alignment, composite image generation, road surface identification, and the intersection of the road mask with the observation mask. Processing parameters are reported in this section, while the numerical results are presented in the Results section.

MATERIALS

A colour drone video of an urban intersection in Kayseri, Türkiye, was used. The video has a resolution of 1920 × 1080 pixels and a frame rate of $24000/1001 \approx 23.976$ frames/s. The image stream is H.264-encoded. The recording duration is 295.378 s, and the frame count is 7,082. The field of view includes the intersection centre and parts of the approach roads, together with vehicles, raised traffic islands, sidewalks, shaded surfaces, and adjacent land.

The investigation was restricted to a visually selected 134 s segment rather than the entire recording. Figure 1 presents sample frames representing different times. The frames illustrate temporal changes at the same intersection, including vehicles occluding the road surface at different locations.



Figure 1. Sample frames from the video recording.

METHODS

The proposed approach comprises a sequence of data processing stages for generating a unified analysis surface from drone video. First, candidate frames are sampled from the selected video interval and aligned with a reference frame. Frames that meet the alignment acceptance criteria are retained and combined to produce a composite image. The composite image is then segmented into superpixels using SLIC, after which a researcher selects the regions corresponding to the road surface. Finally, the unified analysis surface is obtained by intersecting the road mask with the observation mask derived from the

observation polygon. Figure 2 presents a block diagram of the proposed approach, and Algorithm 1 summarises the processing steps.

Processing parameters

The video segment covered the $36 - 170$ s interval. The sample count was $K = 40$, and the minimum number of frames required to pass alignment acceptance was $K_{\min} = 5$. The accepted frame count is denoted by K_a. For SLIC segmentation, the target superpixel count was $g = 7200$, and compactness was $k = 9$. Areas were expressed in pixels without conversion to real-world units.

Reference frame selection and sample alignment

The first frame of the selected video interval was chosen as the reference frame. Forty candidate frames were sampled at approximately equal intervals within the selected time interval, and their source frame indices were recorded. ORB (Oriented FAST and Rotated BRIEF) features were used to align candidate frames to the reference (Rublee et al., 2011). Feature matches were evaluated using Hamming distance, the best matches were selected, and a homography matrix was estimated using RANSAC (Random Sample Consensus) (Fischler and Bolles, 1981).

Alignment was performed on half-resolution grayscale images; frames with fewer than 60 matches or 60 inliers were rejected. Accepted frames were transformed into the reference image's coordinate system using the homography. Bilinear interpolation was used during transformation, and areas outside the image boundaries were filled with zeros. The accepted frame count is denoted by K_a; processing was terminated if $K_a < K_{\min}$.

Composite image generation

The aligned colour images were converted to 8-bit grayscale images with values from 0 to 255 using a weighted combination of the red, green, and blue channel values. At each pixel location, the frame whose grayscale value was closest to the arithmetic mean of the grayscale values was selected. Ties were resolved by choosing the frame with the smallest source frame index. To generate the composite image, the red, green, and blue values at that pixel location were transferred from the selected frame's colour image before grayscale conversion. This selection was performed independently at every pixel location.

Road surface selection and unified analysis surface generation

The composite image was segmented into superpixels using SLIC. SLIC groups image pixels into regions based on colour similarity and spatial proximity but does not automatically classify these regions as road or non-road (Achanta et al., 2012). All superpixel regions were initially unselected. A researcher then examined the composite image and selected the regions corresponding to the road surface within the defined analysis scope. The selection represents the road surface under investigation around the intersection and does not imply exhaustive labelling of all road surfaces visible in the image. Superpixels corresponding to raised traffic islands, medians, sidewalks, and adjacent land were excluded.

The road mask R was generated as the union of the selected superpixel regions. No additional boundary refinement or morphological processing was applied. Region selection was performed by a single researcher, without independent annotation or a separate accuracy assessment.

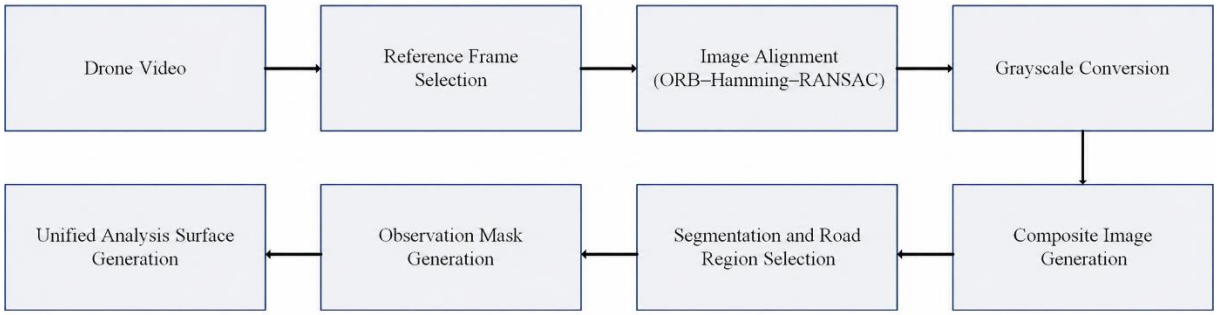


Figure 2. Block diagram of the proposed approach.

The observation polygon was defined by N_p vertices marked by the researcher on the reference frame. Vertices were recorded as zero-based column and row coordinates with the origin at the upper-left corner and connected in sequence to form a closed polygon. The polygon was converted to an image-sized mask to obtain the observation mask F. The unified analysis surface M was obtained by pixelwise intersection of the road mask R and observation mask F. Gaps at islands and sidewalks result from their exclusion during road surface selection, not from road-use intensity.

Algorithm 1. Generating the unified analysis surface from drone video

Inputs: Drone video V; time interval of interest; sample count K; minimum accepted count $K_{\min}$; SLIC parameters g and k.

Outputs: Composite image C; road mask R; observation mask F; unified analysis surface M.

1. Select the first frame of the selected video interval as the reference frame.
2. Select K candidate frames at approximately equal intervals within the specified time interval.
3. Align candidate frames to the reference image using ORB features, Hamming matching, and a RANSAC-based homography.
4. Retain frames meeting the acceptance criteria. Terminate processing if the accepted frame count satisfies $K_a < K_{\min}$.
5. Convert the accepted frames to grayscale images.
6. At each pixel location, select the frame whose grayscale value is closest to the arithmetic mean across frames. Resolve ties using the smallest source frame index.
7. Generate the composite image (C) using the three colour channels of the selected frames.
8. Segment image C using SLIC and obtain the road mask R as the union of researcher-selected superpixels.
9. Convert the researcher-defined observation polygon to a mask to obtain observation mask F.
10. Obtain the unified analysis surface M by intersecting road mask R and observation mask F; $M \leftarrow R \cap F$.
11. Save outputs C, R, F, and M .

RESULTS

The video segment described in the Materials section was used for analysis, and the first frame of the selected video interval was chosen as the reference frame. Forty candidate frames sampled at approximately equal intervals within the selected interval were aligned with the reference frame. Frames that met the alignment acceptance criteria were retained and used to generate a composite image. SLIC segmentation produced 5,684 superpixel regions, of which 811 were selected by the researcher as corresponding to the road surface. The union of these regions formed the road mask (R). Figure 3

presents the composite image, the superpixel segmentation, the researcher-selected regions, and the resulting road mask.

Figure 4 presents the road mask (R), the observation mask (F), and the unified analysis surface (M) obtained by intersecting the two masks. The unified analysis surface comprises pixels that belong to both the selected road surface and the defined observation area. Table 1 reports the pixel count for each mask: 292,698 pixels for the road mask (R), 310,697 pixels for the observation mask (F), and 247,010 pixels for the unified analysis surface (M).

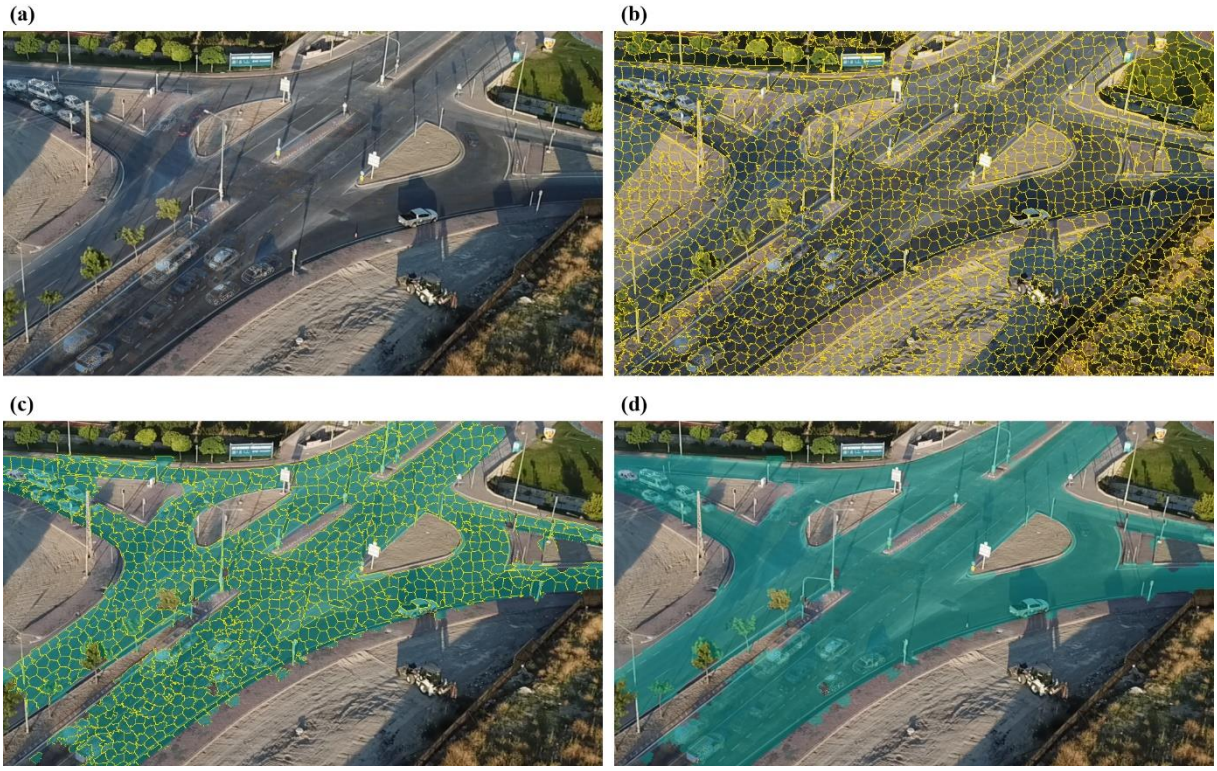


Figure 3. Road mask generation process: (a) composite image, (b) SLIC-segmented image, (c) researcher-selected regions, and (d) resulting road mask.

Table 1. Pixel counts of the road mask, observation mask, and unified analysis surface.

Mask	Pixel count
Road mask (R)	292,698
Observation mask (F)	310,697
Unified analysis surface (M)	247,010

As expected, the unified analysis surface contains fewer pixels than either the road mask or the observation mask because it represents their intersection. The resulting surface defines the portion of the selected road surface that lies within the specified observation area and can be used in subsequent analyses.



Figure 4. Generation of the unified analysis surface: (a) researcher-selected road mask (R), (b) observation mask (F), (c) unified analysis surface obtained by intersection (M), and (d) binary pixel mask of M.

DISCUSSION AND CONCLUSION

This study presents an approach for generating a unified analysis surface from drone video to support subsequent investigations of road surfaces around an intersection. A composite image was generated from aligned frames, the road surface was delineated through researcher selection of superpixel regions, and the analysis extent was defined using an observation polygon. The road surface and the observation area were represented by separate masks, whose intersection yielded the unified analysis surface.

For composite image generation, the frame whose grayscale value was closest to the arithmetic mean of the grayscale values across the accepted aligned frames was selected independently at each pixel location. The corresponding colour channels from the selected frame were then used to generate the composite image. This approach aims to reduce the influence of moving vehicles on the composite image but does not guarantee their complete removal. Vehicles that remain in approximately the same position across multiple frames may still appear in the composite image. Therefore, the resulting image should not be regarded as a completely vehicle-free representation of the road surface.

SLIC segmentation produced 5,684 superpixels, of which 811 were selected as road surface. Their union formed a road mask (R) containing 292,698 pixels. Intersecting this with the 310,697-pixel observation mask (F) produced a unified analysis surface (M) of 247,010 pixels. This demonstrates that road surface selection and observation-area delineation can be applied together. However, intersection

does not validate boundary accuracy; the geometric quality of the unified analysis surface depends on the accuracy of the road mask and observation polygon.

The study has several limitations. The road surface was identified by one researcher, without independent labelling or boundary validation against ground measurements. No additional boundary refinement was applied, and performance at other intersections or under different imaging conditions was not evaluated.

In conclusion, this study focuses on generating a unified analysis surface for subsequent investigations rather than directly detecting unused road surfaces. The resulting surface provides a consistent spatial basis for future analyses of vehicle movements and road use at intersections. Further analysis and validation are required to identify unused road surfaces and quantify their extent.

References

Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Süsstrunk S, 2012. SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(11): 2274–2282. https://doi.org/10.1109/TPAMI.2012.120

Apeltauer J, Babinec A, Herman D, Apeltauer T, 2015. Automatic vehicle trajectory extraction for traffic analysis from aerial video data. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XL-3/W2: 9–15. https://doi.org/10.5194/isprsarchives-XL-3-W2-9-2015

Bock J, Krajewski R, Moers T, Runde S, Vater L, Eckstein L, 2020. The inD dataset: A drone dataset of naturalistic road user trajectories at German intersections. 2020 IEEE Intelligent Vehicles Symposium (IV), pp. 1929–1934. https://doi.org/10.1109/IV47402.2020.9304839

Crommelinck S, Bennett R, Gerke M, Koeva MN, Yang MY, Vosselman G, 2017. SLIC superpixels for object delineation from UAV data. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, IV-2/W3: 9–16. https://doi.org/10.5194/isprs-annals-IV-2-W3-9-2017

Fischler MA, Bolles RC, 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, 24(6): 381–395. https://doi.org/10.1145/358669.358692

Fonod R, Cho H, Yeo H, Geroliminis N, 2025. Advanced computer vision for extracting georeferenced vehicle trajectories from drone imagery. Transportation Research Part C: Emerging Technologies, 178: 105205. https://doi.org/10.1016/j.trc.2025.105205

Hsiao HH, Leou JJ, 2013. Background initialization and foreground segmentation for bootstrapping video sequences. EURASIP Journal on Image and Video Processing, 2013: 12. https://doi.org/10.1186/1687-5281-2013-12

Lu X, Weng Q, 2025. Deep learning-based road extraction from remote sensing imagery: Progress, problems, and perspectives. ISPRS Journal of Photogrammetry and Remote Sensing, 228: 122–140. https://doi.org/10.1016/j.isprsjprs.2025.07.013

Rublee E, Rabaud V, Konolige K, Bradski G, 2011. ORB: An efficient alternative to SIFT or SURF. 2011 IEEE International Conference on Computer Vision, pp. 2564–2571. https://doi.org/10.1109/ICCV.2011.6126544

Data availability and ethics

The data are not publicly available. No identifiable license plates, personal identities, or biometric details are visible in the images.

Acknowledgements

This research received no funding.

Conflict of interest

The authors declare no conflict of interest.

Author contributions

Mustafa Kayademir: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft. Ertuğrul Gül: Conceptualization, Methodology, Validation, Supervision, Writing – review and editing. Murat Taşyürek: Methodology, Software, Validation, Supervision, Writing – review and editing.

A Comparative Evaluation of Machine Learning Classifiers for Multiclass Network Intrusion Detection on a Cybersecurity Dataset

Sevban Avci[1*], Mahir Kaya[2]

[1] Tokat Gaziosmanpaşa University, MSc Computer Engineering, Tokat, Turkey
[2] Tokat Gaziosmanpaşa University, Department of Computer Engineering, Tokat, Turkey

*Corresponding author e-mail: sevbanavci@gmail.com

Abstract

Network intrusion detection increasingly relies on machine learning to catch attacks that signature-based systems cannot generalize to. A central obstacle is severe class imbalance: attack traffic is rare relative to benign traffic, and some attack categories are rarer still. This paper presents a reproducible pipeline for multiclass intrusion detection on the UNSW-NB15 dataset. We compare six classifiers (k-Nearest Neighbors, Random Forest, XGBoost, LightGBM, CatBoost, and a multilayer perceptron), all trained on an identical, SMOTE-balanced feature representation so that differences in reported performance come from the classifiers themselves and not from inconsistent preprocessing. We report macro-averaged F1 as the primary metric, since it weights all ten attack categories equally regardless of how many rows each contributes, and we report per-class results so that performance on the rarest categories (Worms, Shellcode, Backdoor) stays visible instead of being averaged away. On the full official UNSW-NB15 split, after correcting a substantial duplicate-row problem found in the raw data (the Dataset section), XGBoost was the strongest of the six classifiers (test-set macro F1 0.525 with accuracy of 0.750), followed by LightGBM (0.513), CatBoost (0.502), and Random Forest (0.483). Every classifier scored lower on the test set than on the validation set, by 0.057 to 0.083 macro F1 (XGBoost: 0.595 to 0.525). For XGBoost the loss was largest in Fuzzers (F1 0.756 to 0.387), while Generic performance improved on the test set (0.895 to 0.984), consistent with Generic being only 3.9% of the training set but 22.9% of the test set after deduplication (the Dataset section); LightGBM shows the same Fuzzers and Generic pattern. On the validation fold, Backdoor recall varied widely across classifiers (9.1% for k-Nearest Neighbors up to 41.7% for the MLP), and no classifier reached half recall.

Key words: *Intrusion Detection, UNSW-NB15, Multiclass Classification, Class Imbalance, SMOTE*

INTRODUCTION

Intrusion detection systems traditionally relied on signature matching: known attack patterns are catalogued, and traffic is flagged when it matches one. This approach is reliable against previously seen attacks but fails by construction against novel or lightly modified ones. Machine learning (ML) based detection reframes the problem as classification over features extracted from network flows, offering the possibility of generalizing beyond a fixed signature database (Buczak and Guven, 2016).

The UNSW-NB15 dataset (Moustafa and Slay, 2015) was created to address known limitations of older intrusion detection benchmarks (most notably KDD'99 and NSL-KDD), combining real modern normal traffic with nine synthetically generated attack families. It provides 42 flow-level features

alongside a binary label (normal/attack) and a ten-class categorical label (attack_cat) identifying the specific attack family. This paper uses the ten-class framing throughout, since distinguishing attack types is operationally more useful than a bare normal/attack flag.

UNSW-NB15 is, like most real intrusion traffic, heavily imbalanced: normal and generic traffic dominate the dataset, while Worms and Shellcode have only 130 and 1,133 rows, respectively, in the full training set. A classifier can score deceptively well on aggregate accuracy while never once predicting the rarest classes correctly. This motivates two choices that run through this paper: synthetic oversampling of minority classes during training (Synthetic Minority Over-sampling Technique, SMOTE; Chawla et al., 2002), and macro-averaged F1, rather than accuracy, as the headline evaluation metric. In this work we present a documented, reproducible preprocessing pipeline for UNSW-NB15: encoding, scaling, stratified splitting, and SMOTE-based class balancing, built to run inside an isolated Docker development environment. We also present a controlled comparison of six classifiers across four different model families (instance-based, bagged trees, boosted trees, and a neural network), all trained on the same feature matrix so the comparison isolates the classifier rather than the preprocessing.

RELATED WORK

Machine learning for intrusion detection spans a wide range of model families, from classical probabilistic and distance-based classifiers through ensemble trees to deep learning; a broad survey of methods and their application to cyber intrusion detection is given by Buczak and Guven (2016), with more recent surveys extending this landscape to deep-learning- and large-language-model-based malware and intrusion analysis (Khan et al., 2026). UNSW-NB15 was introduced specifically to provide a more realistic and more difficult benchmark than KDD'99-era datasets, with modern attack behavior and a more balanced mix of normal traffic types (Moustafa and Slay, 2015).

Tree-ensemble methods are a common choice for tabular intrusion-detection data of this kind. Random Forest (Breiman, 2001) averages many decorrelated decision trees to reduce variance; gradient-boosted tree methods instead build an ensemble sequentially, each tree correcting the residual error of the ensemble so far. XGBoost (Chen and Guestrin, 2016), LightGBM (Ke et al., 2017), and CatBoost (Prokhorenkova et al., 2018) are the three gradient-boosting implementations compared in this paper; they differ primarily in how they grow trees (level-wise vs. leaf-wise). Class imbalance is handled here with SMOTE (Chawla et al., 2002), implemented via the imbalanced-learn library (Lemaître et al., 2017), which generates synthetic minority-class examples by interpolating between a real minority sample and one of its nearest same-class neighbors rather than naively duplicating existing rows.

Several prior studies apply classical and ensemble classifiers directly to UNSW-NB15. Moustafa et al. (2019) propose a set of protocol-specific statistical flow features (DNS, HTTP, and MQTT) and combine Decision Tree, Naive Bayes, and an artificial neural network via AdaBoost, reporting 99.54% accuracy on UNSW-NB15's DNS subset and 98.97% on its HTTP subset, but do not address the dataset's class imbalance directly, relying on feature engineering and ensembling instead. More et al. (2024) use preprocessing close to this paper's own — one-hot encoding of proto/service/state and standardized numeric features — and report a Random Forest with XGBoost-based feature selection reaching 99.45% accuracy and 99.65% F1, but frame the task as binary (attack vs. normal) rather than the ten-class attack_cat problem addressed here, and do not use SMOTE or any other imbalance

correction. A broader review by Hadi et al. (2023), surveying 22 IoT intrusion-detection studies including several on UNSW-NB15, reports accuracy figures in the same 99%-plus range for several UNSW-NB15 studies, and states that its own comparison focused on accuracy as the primary criterion.

Deep learning and, more recently, large language models offer alternative modeling paradigms to the classical and ensemble methods compared in this paper. Kim et al. (2020) encode traffic records as small images and classify them with a CNN, reaching near-perfect accuracy on KDD'99 but a more modest ~91.5% average accuracy on the harder CSE-CIC-IDS2018 benchmark for fine-grained DoS-subtype classification. This shows that representation choice (image-encoded vs. the tabular one-hot encoding used here) matters somewhat independently of the classifier family. More recently, Mehavilla et al. (2026) benchmark fine-tuned large language models (GPT-2, GPT-Neo, LLaMA) directly against XGBoost, Random Forest, and Decision Tree on flow-based intrusion detection (CIC IoT 2023), finding that XGBoost remains the strongest classifier overall (macro F1 up to 0.970), even against LLM baselines that approach but do not exceed it, at a small fraction of the computational cost. This is an independent result, on a different dataset, that lines up with this paper's own finding that gradient-boosted tree methods ranked first through third (XGBoost, LightGBM, and CatBoost on the test set), ahead of Random Forest, the MLP, and k-NN (the Results section). Erol and Demirci's SentinelLM (2026) pushes the LLM-based approach further, fine-tuning five LLM architectures as an intrusion-detection framework for software-defined IoT networks; it targets a different deployment setting (SD-IoT, text-transformed flow records) than the tabular, classical-ML comparison in this paper, but points to LLM-based detection as a direction worth benchmarking against the classifiers compared here in future work.

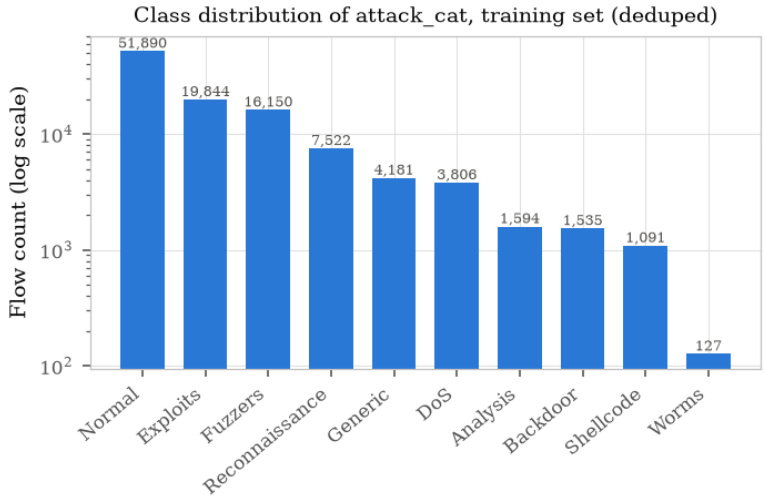
DATASET

UNSW-NB15 (Moustafa and Slay, 2015) provides an official training set of 175,341 records and testing set of 82,332 records, each with 42 numeric and categorical flow-level features (packet counts, byte counts, inter-packet timing, TCP state information, and connection-count aggregates over recent history), a binary label, and a ten-class attack_cat column: *Normal*, *Generic*, *Exploits*, *Fuzzers*, *DoS*, *Reconnaissance*, *Analysis*, *Backdoor*, *Shellcode*, and *Worms*. Three features (proto, service, state) are categorical; the remainder are numeric and span widely different scales and distributions (durations near zero, byte counts in the tens of thousands).

A check for exact duplicate rows — comparing all columns except the row-index id, which is unique by construction and must be excluded or every row appears spuriously unique — found that 67,601 of the 175,341 official training rows (38.6%) are content-level duplicates of another row in the training set, and 26,387 of the 82,332 test rows (32%) are likewise duplicated within the test set. Training on a set with this much internal duplication inflates the apparent size of whichever classes happen to contain more duplicates, so all 67,601 duplicate training rows were dropped, leaving 107,740 distinct training rows; the test set was left as delivered, since it is evaluated exactly as officially released. A separate check additionally found that 4,247 of the 82,332 test rows (5.2%) are exact content-level duplicates of a row retained in the deduplicated training set, which we note as a limitation in the Discussion section: a model's reported test performance is very slightly optimistic for those particular rows, in the same sense it would be for any row the model was directly trained on. Duplication was not evenly spread across classes: 89.5% of *Generic* rows and 68.9% of *DoS* rows in the training set were removed as

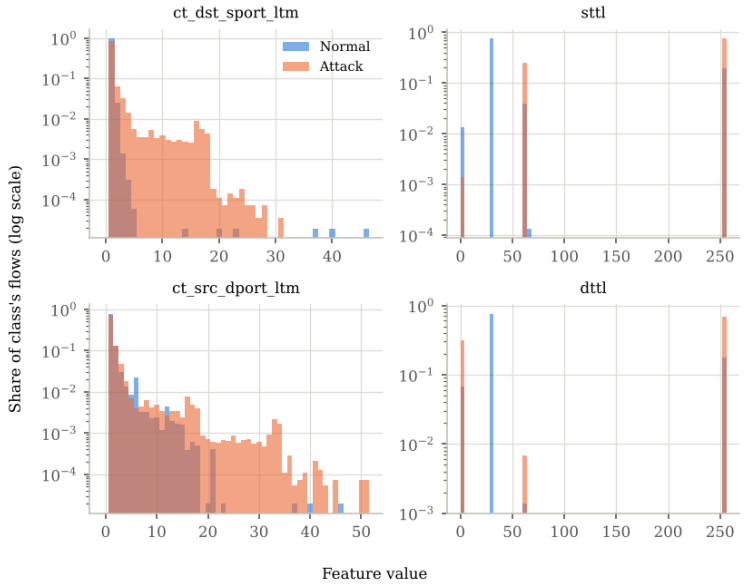
duplicates, while *Normal* 7.3% and *Exploits* 40.6% were comparatively lightly duplicated, so deduplication reshapes the training set's class proportions substantially rather than shrinking it uniformly. Deduplication also changes how closely the training and test class proportions match each other: *Generic* makes up only 3.9% of the deduplicated training set but 22.9% of the test set, nearly six times its share of the training set, because deduplication removed a much larger share of *Generic*'s training rows than of its test rows; the Discussion section returns to this mismatch when discussing *Generic*'s test-set behavior.

Figure 1 shows the resulting class imbalance on a log scale: *Normal* alone makes up 48% of the deduplicated training set, followed by *Exploits* (18%) and *Fuzzers* (15%), while *Generic* falls to 3.9% after deduplication and *Worms* is represented by only 127 rows. This imbalance, rather than any single feature, is the main obstacle the preprocessing and evaluation choices in this paper are built around.



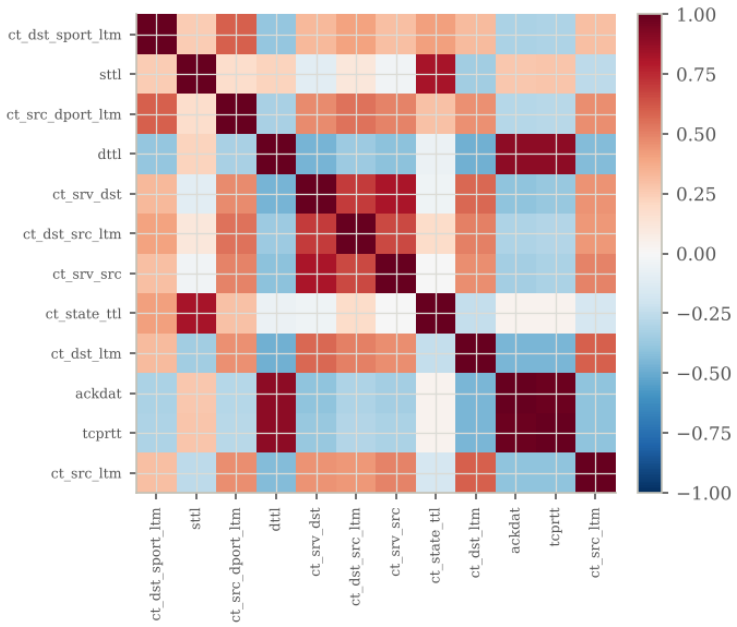
1.Class distribution of attack_cat in the deduplicated UNSW-NB15 training set, log-scaled count axis. *Normal* accounts for 48% of rows; *Generic*, the second-largest class before deduplication, falls to fifth place; *Worms* is the rarest class (127 rows).

To see which numeric features separate the classes, we ranked all 39 numeric features by their univariate ANOVA F-score against attack_cat on the deduplicated training set. This is a screening heuristic, and other criteria such as mutual information order the features differently. The four highest-scoring features are ct_dst_sport_ltm (F = 15,947), sttl (11,670), ct_src_dport_ltm (8,601) and dttl (6,366). The two ct_ features count, over a window of 100 connections, how many share the current flow's destination address and source port (ct_dst_sport_ltm) or its source address and destination port (ct_src_dport_ltm) (Moustafa and Slay, 2015); sttl and dttl are the source-to-destination and destination-to-source time-to-live values. Figure 2 compares their distributions for normal and attack flows. The time-to-live values concentrate on a handful of distinct values: sttl = 31 accounts for 75.4% of normal flows and none of the attack flows, while 254 accounts for 75.3% of attack flows and 19.2% of normal flows; likewise dttl = 29 occurs in 75.4% of normal flows and in no attack flow. For ct_dst_sport_ltm, 97.3% of normal flows take the value 1, against 82.9% of attack flows, and values above 4 occur in 0.02% of normal flows but 5.9% of attack flows.



2.Distributions of the four numeric features with the highest ANOVA F-score against attack_cat, for normal and attack flows. Each class's bars are normalized by class size (the bars of one class sum to 1) and the vertical axis is logarithmic; darker bars are where the two semi-transparent classes overlap.

Figure 3 shows the Spearman rank correlations among the 12 numeric features with the highest F-score. Within this subset only tcprtt and ackdat exceed 0.9 in absolute value (ρ=0.980). Across all 39 numeric features, 16 pairs exceed 0.9 in absolute value. They include three throughput/timing features: source load (sload, bits/sec) with the overall packet rate (rate, packets/sec; ρ=0.943), source load with the mean source inter-packet arrival time (sinpkt; ρ=-0.917), and packet rate with inter-packet arrival time (ρ=-0.957). These three describe the same underlying property of a flow — how fast and densely its packets arrive — measured in different units. Other strongly correlated pairs include tcprtt with synack (ρ=0.990), swin with dwin (ρ=0.987), and dpkts with dbytes (ρ=0.968). We keep all 39 numeric features, since every classifier compared here is trained on the same full encoded feature matrix.



3.Spearman rank correlation among the 12 numeric features with the highest ANOVA F-score against attack_cat. Only tcprtt and ackdat exceed 0.9 in absolute value within this subset.

METHODOLOGY

Preprocessing

The id column (a row index) and the binary label column were dropped, since this paper uses the ten-class attack_cat framing exclusively. Exact duplicate rows were removed from the training set (the Dataset section gives full counts: 67,601 of 175,341 rows removed, leaving 107,740). A stratified 85/15 split of this deduplicated training set produced a validation fold held out from training (91,579 training-fold rows, 16,161 validation-fold rows), in addition to the official, separate test set — stratification was used specifically because the rarest classes (e.g. *Worms*, with 19 rows in the 15% validation split) could otherwise be reduced to a handful of rows, or none, in one of the folds.

Features were encoded and scaled with a single ColumnTransformer fit on the training fold only: standardization (zero mean, unit variance) for the 39 numeric features, and one-hot encoding for the three categorical features, with unknown categories at inference time mapped to an all-zero vector rather than raising an error. Fitting this transformer on the training fold only, and applying it unchanged to the validation and test folds, avoids leaking test-set statistics into preprocessing.

Class Imbalance: SMOTE

The training fold was rebalanced with SMOTE (Chawla et al., 2002), which generates a synthetic minority-class row by interpolating between a real minority sample and a randomly chosen nearby same-class neighbor:

$$x_{new} = x + \delta \cdot (x_{neighbor} - x),$$

with $\delta \sim U(0,1)$. Rather than oversampling every minority class up to the size of the majority class, which under SMOTE's default behavior would have synthesized *Worms* from 108 rows up to 44,107, matching *Normal* exactly, each class below a fixed threshold of 1,000 rows was instead oversampled only up to that threshold, via SMOTE's sampling_strategy argument. In the training fold used here, only two of the ten classes fell below that threshold: *Shellcode* (927 rows) and *Worms* (108 rows), each topped up to 1,000 synthetic-inclusive rows; every other class, including the already-small *Analysis* (1,355) and *Backdoor* (1,305), exceeded the threshold and was left untouched. This grew the training fold from 91,579 to 92,544 rows. SMOTE was fit on the training fold exclusively; the validation and test folds contain only real, unmodified rows.

One limitation is specific to this dataset: SMOTE interpolates across the full feature vector, including the one-hot columns produced by encoding proto/service/state, which can produce fractional values (e.g. proto_tcp=0.4) with no real-traffic interpretation. This is standard practice in the UNSW-NB15 literature and does not appear to materially affect downstream classifiers, but SMOTENC, which samples an actual category from a real neighbor instead of interpolating one, is the more precise alternative; we note it as future work.

Classifiers

Six classifiers were trained on the identical SMOTE-balanced, encoded training fold and compared on the same validation fold:

- **k-Nearest Neighbors** (Cover and Hart, 1967) classifies by majority vote among the $k = 5$ nearest training points in the encoded feature space; it makes no distributional assumption but has no mechanism for class-weighting independent of the training data itself.
- **Random Forest** (Breiman, 2001) averages 300 decorrelated decision trees. We did not combine SMOTE with class_weight="balanced", to avoid stacking two imbalance corrections.
- **XGBoost** (Chen and Guestrin, 2016), **LightGBM** (Ke et al., 2017), and **CatBoost** (Prokhorenkova et al., 2018) are three gradient-boosted tree implementations, each trained with 300 boosting rounds, max depth/leaf count in line with each library's defaults, and a learning rate of 0.1. XGBoost requires integer-encoded labels and performs no categorical handling of its own. All six classifiers here train on the same one-hot-encoded matrix, for comparability.
- **Multilayer Perceptron** (Rumelhart et al., 1986), with two hidden layers (64, 32 units) and ReLU activation, trained with enough iterations to reach convergence (monitored via the optimizer's own convergence check, since the library's default iteration budget was insufficient on this feature set).

All random components (train/validation splitting, SMOTE, and each model's own internal randomness) used a fixed seed for reproducibility. The full implementation, including every hyperparameter and a documented rationale and failure mode for each preprocessing and modeling step, is available in the project's companion notebooks.

Evaluation Metric

Macro-averaged F1 is used as the primary metric: F1 is computed independently per class and averaged without weighting by class size, so a classifier that performs well only on Normal/Generic and poorly

on the rare classes cannot obtain a high score purely from those classes' size, the way it could under plain accuracy. Full per-class precision/recall/F1 (via classification_report) and a row-normalized confusion matrix are reported alongside the single macro-F1 number, since the aggregate score alone does not show *which* classes a model actually struggles with.

Experimental Setup

All experiments ran inside an isolated Docker development container (Python 3.11), avoiding any dependency on, or modification of, the host system. Models were implemented with scikit-learn (Pedregosa et al., 2011) (k-NN, Random Forest, MLP), xgboost, lightgbm, and catboost (their respective native APIs), and imbalanced-learn (Lemaître et al., 2017) for SMOTE. Every model shares the exact same three preprocessed data files (training, validation, test), produced once by the preprocessing pipeline and reused across every model, specifically so that differences in reported performance cannot be attributed to inconsistent preprocessing between models. Hyperparameters were fixed in advance (the Methodology section) and not tuned; every model was fit on the SMOTE-balanced training fold, compared on the validation fold, and then evaluated once on the official test set, with no configuration changed afterward.

RESULTS

Table 1 reports macro F1 and accuracy on both the validation and the test set for all six classifiers, trained and evaluated on the identical, deduplicated, SMOTE-balanced feature matrix described in the Methodology section; rows are sorted by test macro F1. On validation, LightGBM and XGBoost were effectively tied for first place (0.596 and 0.595, 0.001 apart), a gap too small to count as a meaningful difference between the two libraries. On the test set XGBoost led LightGBM by 0.011, with CatBoost, Random Forest, the MLP, and k-NN following in that order (Figure 4).

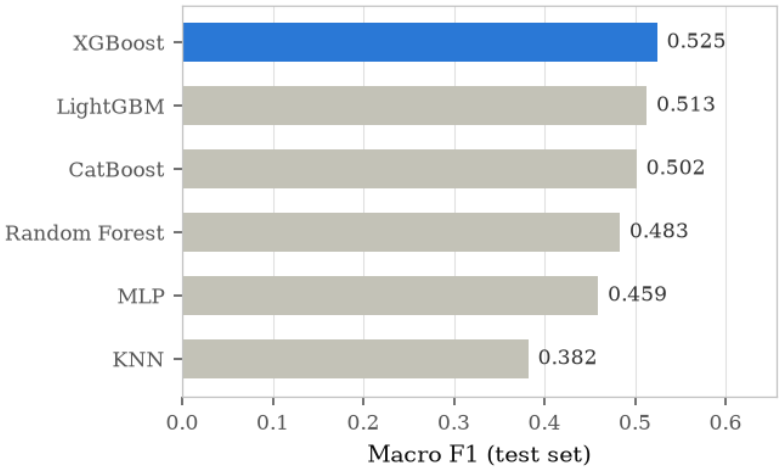
1.Macro F1 and accuracy on the validation and test sets, sorted by test macro F1.

Classifier	**Val. macro F1**	**Val. accuracy**	**Test macro F1**	**Test accuracy**
XGBoost	0.595	0.828	0.525	0.750
LightGBM	0.596	0.826	0.513	0.747
CatBoost	0.576	0.818	0.502	0.750
Random Forest	0.551	0.814	0.483	0.748
MLP	0.543	0.798	0.459	0.716
k-Nearest Neighbors	0.439	0.732	0.382	0.688

2.

2.

Val. = validation set; Test = held-out test set. Macro F1 is the unweighted mean of the ten per-class F1 scores.

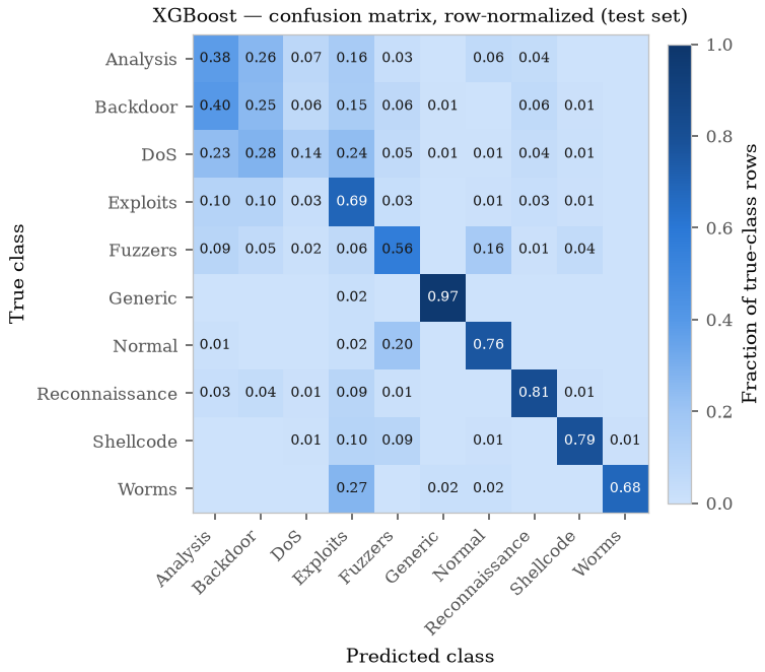


4.Test-set macro F1 by classifier, sorted best to worst, with the top-ranked classifier (XGBoost) highlighted.

All six classifiers were evaluated once on the held-out test set, per the protocol in the Experimental Setup section (Table 1). XGBoost ranked first with a test macro F1 of 0.525 (accuracy 0.750), followed by LightGBM (0.513, 0.747), CatBoost (0.502, 0.750), Random Forest (0.483, 0.748), the MLP (0.459, 0.716), and k-NN (0.382, 0.688). Every classifier scored lower on the test set than on validation, by 0.057 (k-NN) to 0.083 (LightGBM and MLP) macro F1, and the macro F1 ranking is unchanged apart from the top two swapping places. For XGBoost's per-class results on the validation and test sets, the largest change was in *Fuzzers*, where precision fell from 0.737 to 0.295 and F1 from 0.756 to 0.387; *Shellcode* (F1 0.665 to 0.464), *Exploits* (0.837 to 0.704), *Normal* (0.936 to 0.849), *Backdoor* (0.172 to 0.079), and *DoS* (0.295 to 0.213) also fell. *Generic* moved in the opposite direction, improving from F1 0.895 on validation to 0.984 on test (recall 0.826 to 0.972): the Dataset section shows that *Generic* makes up only 3.9% of the deduplicated training set but 22.9% of the test set, which is a plausible contributor to this class behaving differently across the two folds. *Reconnaissance* (0.772 to 0.820) and *Worms* (0.450 to 0.625) also rose, although *Worms* has only 19 validation rows and 44 test rows, so a single flow moves its recall by several points. On the test set, the weakest classes were *Backdoor* (F1 0.079), *Analysis* (0.119), and *DoS* (0.213). *Backdoor* and *Analysis* combine moderate recall (0.250 and 0.381) with very low precision (0.047 and 0.071), meaning that many rows of other classes were labelled as one of these two; *DoS* shows the opposite profile, with a precision of 0.487 but a recall of only 0.137 (0.224 on validation).

Figure 5 shows XGBoost's confusion matrix on the test set. The *Fuzzers* precision collapse is largely *Normal* traffic: 20% of *Normal* flows were predicted as *Fuzzers*, which also accounts for *Normal*'s recall of 0.762. *Analysis*, *Backdoor*, and *DoS* are mostly confused with one another and with *Exploits*: only 14% of *DoS* flows were labelled *DoS*, with 28% labelled *Backdoor*, 24% *Exploits*, and 23% *Analysis*, while 40% of *Backdoor* flows were labelled *Analysis* and 26% of *Analysis* flows *Backdoor*. A further 27% of *Worms* flows were labelled *Exploits*. LightGBM, the classifier with the highest validation macro F1, shows the same pattern: *Fuzzers* precision fell from 0.736 to 0.295 (F1 from 0.753 to 0.392), *Normal*

recall from 0.928 to 0.758, and *DoS* F1 from 0.356 to 0.182, driven by recall collapsing from 0.252 to 0.107, while *Generic* F1 rose from 0.888 to 0.983. The Discussion section discusses these shifts further.



5.Row-normalized confusion matrix for XGBoost on the held-out test set. Each row sums to 1, so a row reads as the distribution of predictions for that true class.

Table 1 and Figures 4 and 5 are the primary results artifacts: Table 1 and the bar chart for the overall ranking, and the confusion matrix for which specific classes XGBoost confuses with which on the test set.

DISCUSSION

Macro F1 and accuracy tell different stories about the six classifiers, most clearly among the leaders. On validation, LightGBM (0.596) narrowly leads XGBoost (0.595) in macro F1, while by accuracy the order flips, XGBoost (0.828) over LightGBM (0.826). On the test set, XGBoost leads in macro F1 (0.525 vs. 0.513), yet four classifiers (CatBoost, XGBoost, Random Forest, and LightGBM) fall within 0.747 to 0.750 accuracy, a spread of only 0.003, while their macro F1 spans 0.483 to 0.525. Accuracy therefore cannot separate the four leading classifiers on the test set, whereas macro F1 can. The two metrics also differ in absolute level: for every classifier, accuracy exceeds macro F1 by 0.23 to 0.29 on validation and by 0.23 to 0.31 on test, because accuracy is dominated by the largest classes (*Normal* alone is 48% of the validation fold) while macro F1 gives each of the ten classes equal weight (the Methodology section discusses why).

LightGBM was first on validation and second on test, within 0.001 and 0.011 macro F1 of XGBoost respectively and ahead of CatBoost, Random Forest, the MLP, and k-NN on both folds. Because the order of XGBoost and LightGBM flips between the validation and test folds, we do not read the test order of the two as decisive.

SMOTE's effect on minority-class recall was uneven across classifiers and did not reach *Backdoor* at all, for a structural reason: SMOTE only oversamples classes below the fixed 1,000-row threshold described in the Methodology section, and *Backdoor*'s 1,305 training-fold rows already exceeded it. *Backdoor* recall on the validation fold ranged from 9.1% (k-NN) to 41.7% (the MLP), with every tree- or boosting-based model landing in between (15.2%–22.2%); no classifier reached even half recall on this class, but the spread across classifiers is wide, which shows how differently the six classifiers handle it. *Worms* is the rarest class (19 validation rows). SMOTE oversamples it (108 real training rows topped up to 1,000), yet the classifiers handle it very differently: k-NN finds 21.1% of the real *Worms* flows, LightGBM 52.6%. LightGBM also has the best F1 on this class (0.571, with 62.5% precision), ahead of XGBoost (0.450) and CatBoost (0.438). With only 19 rows, one flow moves recall by 5.3 points, so the gaps between the lower-ranked models should not be over-read.

All six classifiers lost macro F1 from validation to test, between 0.057 (k-NN) and 0.083 (LightGBM and the MLP); XGBoost fell from 0.595 to 0.525. Each classifier was evaluated once on the test set and none was tuned in response (the Experimental Setup section). For XGBoost, the best classifier on test, the per-class results on both folds show that the loss is concentrated rather than spread evenly across all ten classes: *Fuzzers* alone accounts for about half of it (its F1 fell by 0.369, which is 0.037 of the 0.070 macro F1 lost), followed by *Shellcode* (0.201); on the test set, 20% of *Normal* flows were predicted as *Fuzzers* (Figure 5). *Generic* moved the other way, improving on test (F1 0.895 to 0.984), which lines up with its small share of the training set (3.9%) compared with its share of the test set (22.9%) after deduplication (the Dataset section), a dataset-level property of this particular deduplicated split, not evidence of a modeling problem. LightGBM shows the same *Fuzzers* and *Generic* pattern, with a larger *DoS* loss as well (the Results section). One limitation applies to all test results: as noted in the Dataset section, 4,247 of the 82,332 test rows (5.2%) are exact duplicates of a row retained in the training set, so reported test performance is very slightly optimistic for those rows.

CONCLUSION

This paper presented a complete, reproducible preprocessing and modeling pipeline for multiclass intrusion detection on UNSW-NB15, built around a shared, deduplicated, SMOTE-balanced feature representation and a controlled comparison of six classifiers under a macro-F1-first evaluation protocol chosen specifically for this dataset's class imbalance. On the full official split, after correcting a substantial duplicate-row problem found in the raw training and test data (the Dataset section), XGBoost achieved the highest test macro F1 (0.525, accuracy 0.750), followed by LightGBM (0.513), CatBoost (0.502), and Random Forest (0.483). On the validation fold, LightGBM and XGBoost were effectively tied (0.596 and 0.595), so the order of the top two should not be over-read. Accuracy exceeded macro F1 by 0.23 to 0.29 on validation and by 0.23 to 0.31 on test for every classifier, which is why macro F1, not accuracy, is the right headline metric for a dataset this imbalanced. On the validation fold, recall on *Backdoor* varied widely across classifiers (9.1% to 41.7%) and none reached half. Every classifier lost 0.057 to 0.083 macro F1 from validation to test; for XGBoost about half of that loss comes from *Fuzzers* alone while *Generic* improved, a pattern consistent with UNSW-NB15's documented train/test distribution shift rather than any leakage in this pipeline.

References

Breiman L, 2001. Random forests. Machine Learning, 45 (1): 5–32.
Buczak AL, Guven E, 2016. A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials, 18 (2): 1153–1176.
Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP, 2002. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16: 321–357.
Chen T, Guestrin C, 2016. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794.
Cover T, Hart P, 1967. Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13 (1): 21–27.
Erol AG, Demirci S, 2026. SentinelLM: An LLM-based intrusion detection framework for software-defined Internet of Things. Expert Systems with Applications, 333: 134239, doi: 10.1016/j.eswa.2026.134239.
Hadi QA, Alfoudi AS, Mahdi AM, 2023. IoT cybersecurity threats and detection mechanisms: A review. Wasit Journal for Pure Sciences, 2 (2): 231–250.
Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY, 2017. LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems 30 (NIPS 2017), pp. 3146–3154.
Khan S, Raza H, Alam M, 2026. AI-driven malware analysis and detection: A comprehensive survey of techniques, trends and challenges. Journal of Informatics and Web Engineering, 5 (1): 107–129, doi: 10.33093/jiwe.2026.5.1.7.
Kim J, Kim J, Kim H, Shim M, Choi E, 2020. CNN-based network intrusion detection against denial-of-service attacks. Electronics, 9 (6): 916, doi: 10.3390/electronics9060916.
Lemaître G, Nogueira F, Aridas CK, 2017. Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research, 18 (17): 1–5.
Mehavilla L, Rodríguez M, García J, Alesanco Á, 2026. Evaluating large language models effectiveness for flow-based intrusion detection: A comparative study with ML and DL baselines. Artificial Intelligence Review, 59 (1): 50, doi: 10.1007/s10462-025-11432-2.
More S, Idrissi M, Mahmoud H, Asyhari AT, 2024. Enhanced intrusion detection systems performance with UNSW-NB15 data analysis. Algorithms, 17 (2): 64, Table 7, doi: 10.3390/a17020064.
Moustafa N, Slay J, 2015. UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). 2015 Military Communications and Information Systems Conference (MilCIS), pp. 1–6.
Moustafa N, Turnbull B, Choo KKR, 2019. An ensemble intrusion detection technique based on proposed statistical flow features for protecting network traffic of Internet of Things. IEEE Internet of Things Journal, 6 (3): 4815–4830, doi: 10.1109/JIOT.2018.2871719.
Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É, 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12: 2825–2830.
Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A, 2018. CatBoost: Unbiased boosting with categorical features. Advances in Neural Information Processing Systems 31 (NeurIPS 2018).
Rumelhart DE, Hinton GE, Williams RJ, 1986. Learning representations by back-propagating errors. Nature, 323 (6088): 533–536.

Conflict of Interest

The authors have declared that there is no conflict of interest.

Comparative Evaluation of QRS-Based Machine Learning and Deep Learning for ECG Emotion Recognition

Ali Hazar[1]*, Sema Kayhan[2]

[1] Gaziantep University, Faculty of Engineering, Department of Electrical and Electronics Engineering, 27310 Gaziantep, Türkiye

[2]Gaziantep University, Faculty of Engineering, Department of Electrical and Electronics Engineering, 27310 Gaziantep, Türkiye

*Corresponding author e-mail: alihazar@gantep.edu.tr

Abstract

The use of physiological signals for emotion recognition has attracted increasing interest in view of its possible applications in affective computing, human-computer interaction, and intelligent healthcare systems. Among physiological signals, electrocardiogram (ECG) offers useful information on physiological response associated with emotions. This paper focuses on ECG-based emotion recognition employing traditional machine learning and neural network methods on the publicly available DREAMER and CASE datasets. Proposed evaluation methodology consists of pre-processing of ECG signals, feature extraction, development of classifiers, and performance evaluation. Time domain, statistical, heart rate variability, morphological, and time-frequency features are employed to describe emotion-related information embedded in ECG signals. Various classification techniques are analyzed based on different evaluation protocols including random, trial-wise and subject-independent evaluations to investigate generalizability of classification performance. It is shown that classification performance depends on feature representation, type of evaluation protocol, and subject variation. Specifically, QRS-based classification technique applied to the DREAMER dataset demonstrates the average macro F1-score of 80.26% on the random evaluation, while performance drops to 18.86% in the subject-independent setting. Large difference between the performance results indicates difficulties in generalization of emotion recognition systems across subjects. Both feature-based machine learning and neural network techniques show the capability of capturing emotion-related information from ECG signals depending on the evaluation scenario.

Keywords: *ECG, Emotion Recognition, Machine Learning, Affective Computing,Neural Networks Subject-Independent Evaluation*

INTRODUCTION

In affective computing, emotion recognition from physiological signals is a significant research area. ECG, unlike facial expressions and speech, reveals autonomic activity linked to emotional responses. Interpretation of ECG is difficult because of individual differences, environmental noise, and recording conditions. Systematic research on emotion recognition with physiological signals is supported by public datasets such as DREAMER and CASE (Katsigiannis and Ramzan, 2018; Sharma et al., 2019).

Machine learning models are commonly trained on heart rate variability (HRV), statistical, morphological, and time-frequency features. Classification results are significantly affected by signal sampling, windowing, and data-splitting methods. When window-level random splitting is used, highly correlated ECG windows from the same participant or recording may end up in both the training and test sets. Raw ECG samples may also be shared across partitions by some windows, which creates data leakage and can yield overly optimistic performance estimates (Saeb et al., 2017). Trial-wise evaluation measures performance on held-out trials, whereas subject-independent evaluation tests generalization to participants not previously seen.

Three QRS-based classifiers for valence and arousal prediction are compared in this study: Random Forest (RF), Gradient Boosting (GB), and a one-dimensional convolutional neural network (1D-CNN). On the DREAMER and CASE datasets, the experiments evaluate both emotional dimensions separately under overlap-controlled random, trial-wise, and leave-one-subject-out (LOSO) protocols. Traditional ensemble learning and deep learning approaches are represented by the selected models. Their comparison investigates how model performance varies across evaluation protocols, focusing on generalization to unseen participants.

MATERIAL AND METHODS

Material

DREAMER consists of ECG recordings from 23 participants, each of whom watched 18 audiovisual film clips chosen to elicit emotional responses. The stimulus ECG channel used in this study is sampled at 256 Hz. CASE, by contrast, offers continuously annotated affective responses, in which QRS windows receive binary labels according to the predefined 30-second segments they correspond to. The CASE cohort analyzed here comprises 30 participants. Its ECG signals were recorded at 1000 Hz and then resampled to 256 Hz for the neural network models.

Under the QRS-based approach, ECG windows of roughly 0.78 seconds are extracted around detected R peaks. At 256 Hz this amounts to 200 samples for DREAMER, while for CASE it amounts to 781 samples at the original 1000 Hz sampling rate. From DREAMER, the analysis included 98,368 valence windows and 70,530 arousal windows; from CASE, it included 31,787 valence windows and 32,232 arousal windows. These differences in window counts reflect the data selection and labeling procedures applied to each classification task.

Methods

ECG preprocessing and QRS representation

R peaks were detected using dataset-specific methods. DREAMER employed NeuroKit as the primary detector, with Pan–Tompkins-based verification and exclusion of inconsistent peaks (Makowski et al., 2021; Pan and Tompkins, 1985), while CASE used a custom QRS-energy-based detector.

R peaks detected in the ECG serve as reference points from which short windows around individual cardiac events are extracted. RF and Gradient Boosting (GB) rely on local QRS morphological descriptors together with adjacent rhythm information, whereas the 1D-CNN learns representations

directly from standardized ECG waveform windows. Such a design enables handcrafted QRS-based features to be compared against automatically learned waveform representations. Broader literature additionally considers time-domain, statistical, heart rate variability (HRV), morphological, and time–frequency features; the present analysis, by contrast, concentrates on QRS-local descriptors and direct waveform learning.

Classification and evaluation

In Random Forest, predictions from multiple decision trees are combined (Breiman, 2001), while Gradient Boosting improves predictive performance by sequentially combining weak learners (Friedman, 2001). Temporal convolutional layers allow the 1D-CNN to extract local patterns from ECG waveform windows (Ismail Fawaz et al., 2019). Valence and arousal constitute two separate binary classification tasks. Data partitions are identical for all three models within any given dataset, target, protocol, and random seed configuration.

Evaluation Protocols and Random-Seed Configuration

The present comparison employs three evaluation protocols: Random, Trial-wise, and Leave-One-Subject-Out (LOSO). Under the Random protocol, overlap-exclusion procedures guard against raw ECG samples being shared across training, validation, and test partitions. With Trial-wise splitting, every window from a given trial stays within one partition. LOSO, which excludes each held-out test subject from training and validation, assesses generalization across subjects.

Three predefined random seeds (42, 43, and 44) were used in the experiments. Training, validation, and test partitions were regenerated for each seed in the Random and Trial-wise evaluations, and all compared models shared identical partitions within each experimental configuration. Subject-level partitions stayed fixed across seeds under LOSO, with each model retrained independently. Standard deviations in the Random and Trial-wise evaluations thus capture variability from both partitioning and training, whereas under LOSO they chiefly capture training variability given fixed subject partitions. Reporting takes the form of mean ± standard deviation rather than confidence intervals.

Trial-level predictions for DREAMER are obtained by combining QRS-window predictions through majority voting. For CASE, predictions are aggregated at the level of the corresponding 30-second segment, in line with the dataset's labeling procedure. Since the two datasets differ in evaluation unit, results are interpreted within each dataset. Test accuracy, balanced accuracy, macro precision, macro recall, macro-F1 score, and ROC-AUC serve as the performance measures. Test accuracy, balanced accuracy, macro-F1, and ROC-AUC appear in the main results tables.

The separate four-class DREAMER QRS-CNN methodology experiment used two ECG channels and joint valence–arousal labels. It evaluated a QRS-centered CNN under random window-level splitting and repeated subject-disjoint holdout with 16 training and seven test participants. Trial-level predictions were obtained by averaging softmax probabilities across test windows within each trial. This preliminary evaluation differs from the present binary benchmark and does not represent LOSO cross-validation.

RESULTS

Tables 1–4 report how RF, GB, and 1D-CNN models perform on binary valence and arousal classification across three evaluation protocols: Random (overlap-controlled), Trial-wise, and LOSO. DREAMER results appear at the trial level, whereas CASE results are given at the level of the corresponding 30-second segment. Each result is expressed as mean ± standard deviation over three random seeds (42, 43, and 44).

Table 1. DREAMER valence classification at trial level.

Model	**Protocol**	**Acc (%)**	**BAcc (%)**	**F1 (%)**	AUC
RF	Random	85.10 ± 1.61	84.96 ± 1.49	84.56 ± 1.62	91.37 ± 0.65
GB	Random	66.99 ± 0.56	67.47 ± 0.77	66.50 ± 0.63	72.50 ± 0.12
1D-CNN	Random	76.57 ± 3.72	77.02 ± 2.93	76.05 ± 3.43	84.53 ± 3.92
RF	Trial-wise	49.74 ± 3.30	47.16 ± 4.93	46.82 ± 5.45	44.93 ± 5.24
GB	Trial-wise	44.97 ± 1.83	44.58 ± 1.37	44.25 ± 1.54	43.00 ± 2.50
1D-CNN	Trial-wise	47.62 ± 9.91	46.54 ± 9.96	45.57 ± 10.15	42.32 ± 12.47
RF	LOSO	51.53 ± 0.85	48.84 ± 0.78	36.92 ± 1.17	49.29 ± 1.41
GB	LOSO	54.27 ± 0.14	50.41 ± 0.10	42.43 ± 0.05	46.25 ± 0.00
1D-CNN	LOSO	54.19 ± 2.28	50.67 ± 0.95	40.54 ± 0.65	52.25 ± 1.00

Acc: test accuracy; BAcc: balanced accuracy; F1: macro-F1; AUC: ROC-AUC. Values are mean ± SD across seeds 42, 43 and 44; AUC is expressed as a percentage.

Table 2. DREAMER arousal classification at trial level.

Model	**Protocol**	**Acc (%)**	**BAcc (%)**	**F1 (%)**	**AUC**
RF	Random	87.46 ± 0.90	86.53 ± 0.94	86.71 ± 0.94	95.01 ± 0.36
GB	Random	75.93 ± 0.90	75.79 ± 1.09	75.14 ± 0.98	81.06 ± 0.26
1D-CNN	Random	84.29 ± 4.51	82.76 ± 3.78	83.24 ± 4.43	89.55 ± 3.47
RF	Trial-wise	62.61 ± 3.44	59.36 ± 4.89	58.83 ± 5.11	61.44 ± 5.40
GB	Trial-wise	58.77 ± 6.25	59.39 ± 6.05	57.37 ± 7.20	58.32 ± 11.20
1D-CNN	Trial-wise	67.82 ± 5.74	66.77 ± 6.98	65.33 ± 6.16	62.97 ± 7.10
RF	LOSO	51.58 ± 1.17	47.73 ± 0.88	37.17 ± 1.41	47.92 ± 1.18
GB	LOSO	51.21 ± 0.00	50.77 ± 0.00	38.20 ± 0.00	51.22 ± 0.17
1D-CNN	LOSO	45.32 ± 0.45	46.03 ± 1.71	32.99 ± 0.95	41.76 ± 2.66

Acc: test accuracy; BAcc: balanced accuracy; F1: macro-F1; AUC: ROC-AUC. Values are mean ± SD across seeds 42, 43 and 44; AUC is expressed as a percentage.

In the DREAMER valence task, RF recorded the top Random macro-F1 score of 84.56%, with 1D-CNN next at 76.05% and GB at 66.50%. Trial-wise evaluation produced scores of 46.82% for RF, 44.25% for GB, and 45.57% for 1D-CNN, while under LOSO, GB reached 42.43%, 1D-CNN 40.54%, and RF 36.92%. For DREAMER arousal, the Random macro-F1 scores stood at 86.71% for RF, 75.14% for GB, and 83.24% for 1D-CNN. The corresponding Trial-wise scores were 58.83%, 57.37%, and 65.33%, and the LOSO scores were 37.17%, 38.20%, and 32.99%, respectively.

Table 3. CASE valence classification at source-segment level.

Model	**Protocol**	**Acc (%)**	**BAcc (%)**	**F1 (%)**	AUC
RF	Random	80.09 ± 1.10	77.03 ± 1.15	77.70 ± 1.21	85.06 ± 0.72
GB	Random	69.26 ± 0.39	69.22 ± 0.27	67.99 ± 0.33	73.13 ± 0.43

1D-CNN	Random	79.76 ± 1.96	78.06 ± 0.33	77.98 ± 1.28	83.16 ± 0.65
RF	Trial-wise	51.44 ± 6.54	48.08 ± 6.71	48.00 ± 6.60	46.45 ± 6.99
GB	Trial-wise	46.75 ± 3.50	44.17 ± 3.57	43.90 ± 3.24	43.61 ± 1.63
1D-CNN	Trial-wise	54.55 ± 5.88	50.46 ± 5.19	50.36 ± 5.32	52.93 ± 3.54
RF	LOSO	48.06 ± 1.89	51.08 ± 0.63	37.59 ± 1.20	49.84 ± 0.87
GB	LOSO	53.01 ± 0.51	50.83 ± 0.08	38.07 ± 0.38	51.41 ± 0.13
1D-CNN	LOSO	59.33 ± 0.52	57.75 ± 1.69	48.16 ± 2.15	64.80 ± 1.59

Acc: test accuracy; BAcc: balanced accuracy; F1: macro-F1; AUC: ROC-AUC. Values are mean ± SD across seeds 42, 43 and 44; AUC is expressed as a percentage.

Table 4. CASE arousal classification at source-segment level.

Model	**Protocol**	**Acc (%)**	**BAcc (%)**	**F1 (%)**	**AUC**
RF	Random	77.80 ± 0.67	75.58 ± 0.72	76.14 ± 0.73	83.84 ± 0.22
GB	Random	69.07 ± 0.86	67.82 ± 0.67	67.77 ± 0.75	70.84 ± 0.55
1D-CNN	Random	70.66 ± 1.73	68.60 ± 1.88	68.80 ± 1.86	72.37 ± 1.71
RF	Trial-wise	53.99 ± 6.67	52.41 ± 4.98	52.09 ± 5.37	53.80 ± 8.01
GB	Trial-wise	53.06 ± 4.63	53.32 ± 4.22	52.59 ± 4.31	52.98 ± 7.63
1D-CNN	Trial-wise	55.89 ± 9.14	55.92 ± 8.74	55.34 ± 8.81	54.97 ± 12.28
RF	LOSO	60.41 ± 0.29	50.71 ± 0.35	39.03 ± 0.72	49.69 ± 0.44
GB	LOSO	54.25 ± 0.06	52.14 ± 0.18	40.39 ± 0.27	52.83 ± 0.08
1D-CNN	LOSO	51.27 ± 0.15	53.20 ± 0.56	38.77 ± 1.15	59.16 ± 0.50

Acc: test accuracy; BAcc: balanced accuracy; F1: macro-F1; AUC: ROC-AUC. Values are mean ± SD across seeds 42, 43 and 44; AUC is expressed as a percentage.

For CASE valence, RF, GB, and 1D-CNN achieved Random macro-F1 scores of 77.70%, 67.99%, and 77.98%, respectively. Their trial-wise scores were 48.00%, 43.90%, and 50.36%. Under LOSO, 1D-CNN obtained 48.16%, compared with 37.59% for RF and 38.07% for GB. In the CASE arousal condition, Random macro-F1 reached 76.14% for RF, 67.77% for GB, and 68.80% for 1D-CNN. Trial-wise scores were 52.09%, 52.59%, and 55.34%, respectively, while LOSO scores were 39.03%, 40.39%, and 38.77%.

Across both datasets and classification targets, macro-F1 scores were generally higher under Random evaluation than under Trial-wise and LOSO evaluation. The more restrictive protocols provide a clearer assessment of model performance on held-out trials and unseen participants. DREAMER and CASE use different labeling structures and evaluation units, so their absolute performance scores should be interpreted within each dataset rather than treated as directly equivalent.

Table 5. Separate four-class DREAMER QRS-CNN methodology experiment at trial level.

Evaluation protocol	Test Accuracy (%)	Macro-F1 (%)
Random window-level split	80.27 ± 0.61	80.26 ± 0.83
Repeated subject-disjoint holdout (16/7)	24.34 ± 0.92	18.86 ± 2.60

Values are mean ± SD across seeds 42, 43, and 44. Subject-disjoint holdout is not LOSO.

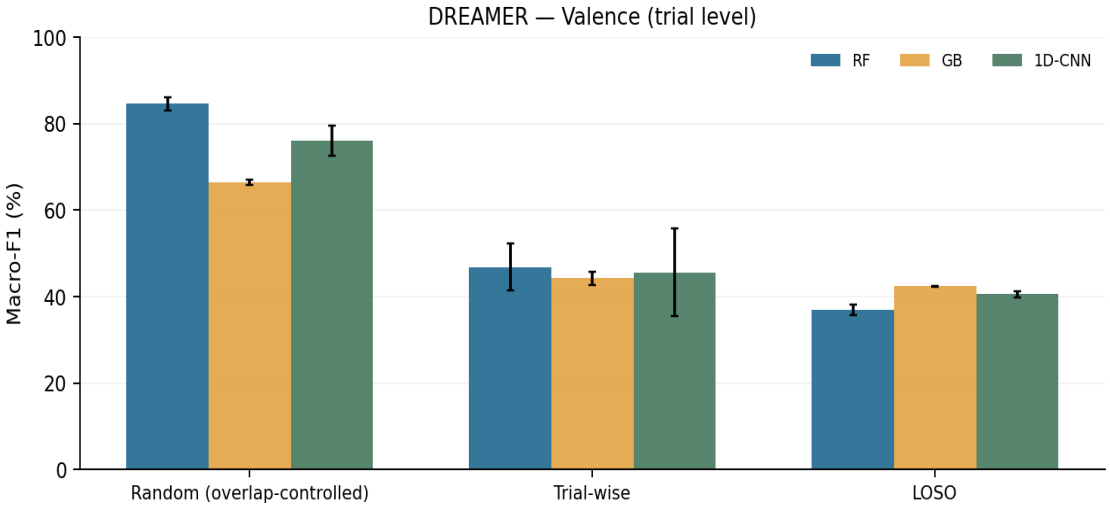


Figure 1. DREAMER valence trial-level macro-F1 for RF, GB and 1D-CNN across Controlled Random, Trial-wise and LOSO (three-seed means).

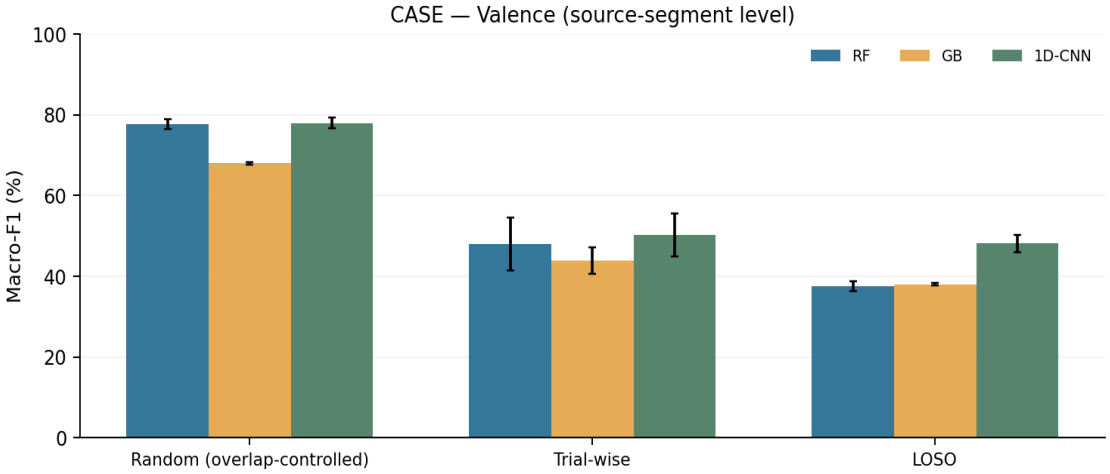


Figure 2. CASE valence segment-level macro-F1 for RF, GB and 1D-CNN across Controlled Random, Trial-wise and LOSO (three-seed means).

DISCUSSION AND CONCLUSION

The results show clear differences in classification performance across the three evaluation protocols. Overlap-controlled random splitting prevents raw ECG samples from being shared between data partitions. Still, windows from the same participant or trial may appear in both training and test sets. Trial-wise splitting keeps all windows from each trial together, while LOSO tests how well the models perform on participants excluded from training. These approaches evaluate generalization under different conditions.

In both DREAMER and CASE, RF, GB, and 1D-CNN obtained their highest average macro-F1 scores with Random evaluation. Their performance declined under Trial-wise and LOSO evaluation. This difference highlights the role of data partitioning in the reported results. Removing overlapping windows prevents direct sharing of ECG samples, but characteristics specific to individual participants may still influence the predictions. Changes in the test data and the number of training windows may also help explain the differences between protocols.

The two datasets show similar overall patterns, although their performance scores should be interpreted separately. DREAMER and CASE differ in how emotional responses are labeled, the participants involved, recording conditions, and the units used for evaluation. These differences limit direct comparisons between their results. We report the mean and standard deviation across three seeds to describe variation between experimental runs. These values do not establish statistically significant differences between models. More repetitions would provide a clearer picture of model stability.

We selected RF, GB, and 1D-CNN to examine how different classification approaches respond to changes in evaluation protocols. The study does not attempt to compare every available model or physiological feature representation. Our findings show why data partitioning needs careful consideration in ECG-based emotion recognition. In particular, strong results from random splitting alone do not demonstrate how reliably a model will perform on participants it has never encountered.

References

Breiman L, 2001. Random forests. Machine Learning, 45: 5–32.

Friedman JH, 2001. Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5): 1189–1232.

Katsigiannis S, Ramzan N, 2018. DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. IEEE Journal of Biomedical and Health Informatics, 22(1): 98–107.

Sharma K, Castellini C, van den Broek EL, Albu-Schaeffer A, Schwenker F, 2019. A dataset of continuous affect annotations and physiological signals for emotion analysis. Scientific Data, 6: 196.

Ismail Fawaz H, Forestier G, Weber J, Idoumghar L, Muller P-A, 2019. Deep learning for time series classification: A review. Data Mining and Knowledge Discovery, 33: 917–963. https://doi.org/10.1007/s10618-019-00619-1

Makowski D, Pham T, Lau ZJ, Brammer JC, Lespinasse F, Pham H, Schölzel C, Chen SHA, 2021. NeuroKit2: A Python toolbox for neurophysiological signal processing. Behavior Research Methods, 53(4): 1689–1696. https://doi.org/10.3758/s13428-020-01516-y

Pan J, Tompkins WJ, 1985. A real-time QRS detection algorithm. IEEE Transactions on Biomedical Engineering, BME-32(3): 230–236. https://doi.org/10.1109/TBME.1985.325532

Saeb S, Lonini L, Jayaraman A, Mohr DC, Kording KP, 2017. The need to approximate the use-case in clinical machine learning. GigaScience, 6(5): 1–9. https://doi.org/10.1093/gigascience/gix019.

Conflict of Interest

The authors have declared that there is no conflict of interest.

Author Contributions

Ali Hazar contributed to the methodology, data preprocessing, model implementation, experimental analysis, and manuscript preparation. Sema Kayhan provided supervision, methodological review, and interpretation of results.

Development of Root Cause Analysis and Autonomous Operational Intervention Systems in Enterprise Application Performance Monitoring Software Through AI-Assisted Event Detection

M.Fevzi Korkutata[1*], Önder Şahinaslan[2]

[1] *Maltepe University, Graduate School, Computer Engineering PhD Programme, 34857, Istanbul, Turkey*
[2] *Maltepe University, Department of Informatics, Istanbul, Turkey, ondersahinaslan@maltepe.edu.tr*

**Corresponding author e-mail: fevzi.korkutata@volthread.com*

Abstract

Enterprise application servers generate continuous streams of performance metrics, yet conventional application performance monitoring (APM) tools largely rely on static, manually defined thresholds that fail to capture varying workload dynamics and leave the interpretation and remediation of anomalies to system administrators. This study investigates the use of artificial intelligence (AI) and machine learning (ML) in APM software and proposes a staged approach that moves from AI-assisted event detection towards root cause analysis (RCA) and autonomous operational intervention, focusing on Oracle WebLogic Server environments monitored through WebLogic Smart Dashboard and Monitoring (WLSDM). Performance metrics of the AdminServer and ManagedServer_1 JVM processes, obtained from WLSDM METRIC/VALUE tables, the OpenSocketsCurrentCount metric and alarm records, were used as the data source; because the existing laboratory data were insufficient, the dataset was extended with 21,600 manually added records covering a 15-day period at two-minute intervals. The detection pipeline combines statistical and machine learning methods: server-specific dynamic thresholds are derived from the mean and standard deviation of each server's metrics using the Z-score (z = 3), moving thresholds are computed over a three-hour rolling window, and observations exceeding these limits are labelled as anomalies. A Random Forest classifier (100 trees, 80/20 train-test split) is then trained separately for each server on these labels, while Isolation Forest is applied to the same data as an unsupervised alternative for settings without labelled anomalies. Model performance is assessed with accuracy, classification reports and confusion matrices, and the results are visualised with histograms, kernel density estimates and scatter plots. The findings indicate that dynamic and moving thresholds produce more meaningful anomaly boundaries than static limits on servers with variable workloads, and that the approach can reduce dependence on manual threshold definition and support proactive monitoring. Observations from the Oracle WebLogic Server environment further point to prolonged garbage collection pauses, JVM memory growth, thread pool saturation and database connection pool bottlenecks as recurring performance issues. The main originality of the study lies in transferring statistical thresholding and ensemble learning, previously validated separately in other domains, to enterprise Oracle WebLogic Server / JVM performance data and evaluating them within a single

pipeline. RCA and autonomous intervention are defined as subsequent stages of the framework: contribution to RCA is included among the algorithm selection criteria, and multivariate features, independent validation, and the error margin, privacy and security of automated actions are outlined as the research agenda for these stages.

Keywords: *Artificial Intelligence, Machine Learning, Application Performance Monitoring, Anomaly Detection, Random Forest*

INTRODUCTION

Enterprise Java application servers expose heterogeneous metrics whose acceptable values change with workload, resource demand and operating conditions. Conventional monitoring frequently evaluates manually configured alarm limits; this can identify threshold violations but does not, by itself, explain the underlying cause or recommend a safe corrective action. Earlier work has examined intelligent server monitoring, anomaly prediction and proactive infrastructure management (De Franceschi et al., 1997; Alonso et al., 2011; Sheta, 2023).

An earlier ICMS 2025 abstract proposed AI/ML-enhanced monitoring as a basis for anomaly detection and eventual RCA and proactive action (Korkutata and Şahinaslan, 2025). A subsequent journal article reported Random Forest analysis of WebLogic JVM metrics using dynamic Z-score thresholds and diagnostic plots (Korkutata and Sahinaslan, 2026). This paper synthesizes those previously reported results within a staged WLSDM framework; RCA and autonomous intervention are prospective, not newly validated production capabilities.

MATERIAL AND METHODS

Material

The application environment comprised Oracle WebLogic Server JVM processes named AdminServer and ManagedServer_1, monitored by WebLogic Smart Dashboard and Monitoring (WLSDM). Source material included METRIC/VALUE fields, server identifiers, timestamps, the OpenSocketsCurrentCount indicator, WLSDM alarms and reports. The laboratory measurements were insufficient for model experiments, so a Python script generated and manually added 21,600 METRIC records: 10,800 observations per server across 15 days at two-minute intervals. The inserted values represented typical operating bands together with occasional higher-valued observations. This augmentation must be distinguished from independently collected production telemetry.

Methods

For each server s, the statistical upper threshold was defined as $T_s = \mu_s + 3\sigma_s$, with μ_s and σ_s denoting its sample mean and standard deviation. Observations above T_s were labelled as candidate anomalies. A moving threshold employed a rolling mean and standard deviation to reflect time-dependent changes. The research design specifies a three-hour observation window; however, the thesis code also shows rolling(window=3) without a time-based index, so equivalence to a full three-hour implementation is not independently established.

The labelled data were divided into 80% training and 20% testing using random_state=42. RandomForestClassifier employed 100 trees and was trained separately for AdminServer and ManagedServer_1. IsolationForest was fitted separately to the METRIC values without labels, using contamination=0.1 in the documented experiment. Analyses used Python, Pandas, NumPy, Scikit-learn, Matplotlib and Seaborn; histogram, kernel density estimation (KDE), scatter plots and confusion matrices supported interpretation.

Table 1. Anomaly detection pipeline and study status

Stage / technique	Implemented use	Role in the proposed framework
WLSDM / METRIC / VALUE	Monitoring data and augmented laboratory dataset	Event-data acquisition
Z-score and moving threshold	Server-wise candidate-event labelling	Adaptive alarm boundary
Random Forest / Isolation Forest	Supervised / unsupervised anomaly analysis	AI-assisted event detection
RCA and operational intervention	Design and future research direction	Explanation and guarded remediation

RESULTS

For the augmented METRIC experiment, the thesis reports server-specific dynamic thresholds of approximately 12.76 for AdminServer and 7.36 for ManagedServer_1. These values differ because the server groups have different observed distributions. The thesis anomaly scatter plot (Figure 1) illustrates selected threshold-exceeding observations, while its METRIC histogram shows that most measurements occupy the lower, normal-value range. Hence a single fixed limit is less descriptive of the two distinct baseline profiles.

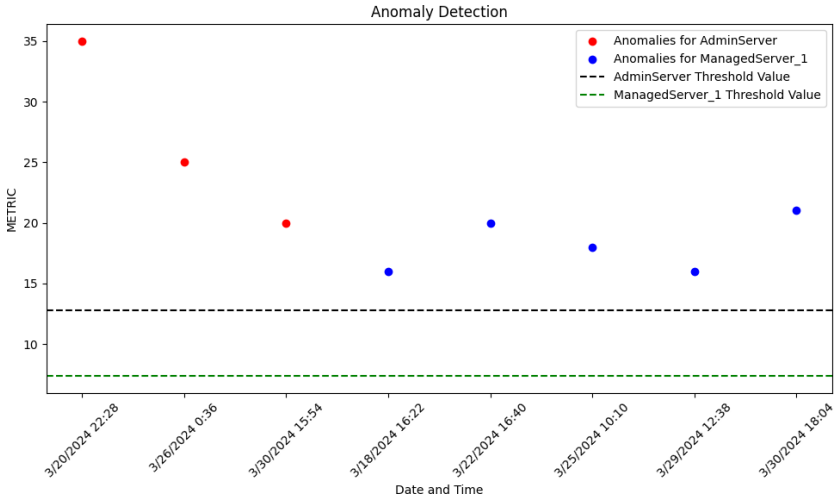


Figure 1. Threshold-exceeding METRIC observations for the two Oracle WebLogic Server JVM processes (reproduced from the thesis).

The experimental implementations offer different types of evidence. Random Forest learns labels derived from the same metric and threshold definition; its classification metrics therefore primarily measure how consistently it reproduces those labels. The separately documented VALUE/JSON run reports an accuracy of 1.00 for both servers, but its displayed classification reports contain only the normal (False) class, with supports of 114,168 and 85,833. Consequently, the reported perfect accuracy cannot be interpreted as independently validated sensitivity to real anomalies.

Table 2. Experimental observations and their interpretation

Evidence from thesis	Finding	Interpretation / limitation
METRIC threshold outputs	AdminServer: ~12.76; ManagedServer_1: ~7.36	Server-specific statistical boundaries
Random Forest	100 trees; random 80/20 split	Agreement with threshold-derived labels
VALUE/JSON classifier report	Accuracy = 1.00; False class only	Anomaly recall not established
Isolation Forest	Unsupervised outliers; contamination = 0.1	Alternative without labelled events

The Isolation Forest scatter plots in the thesis mark separated high-METRIC observations in each server series (Figure 2). This unsupervised output demonstrates outlier ranking rather than verified operational faults. No quantitative, independently labelled head-to-head benchmark against Random Forest is provided; correspondingly, no superiority claim between the two models is made.

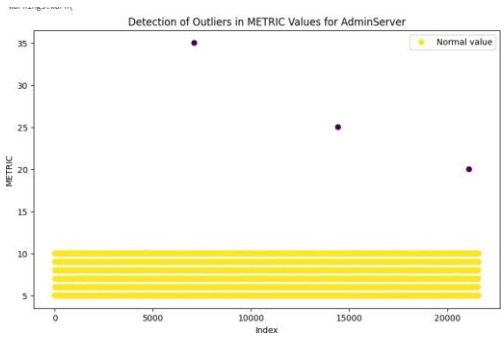


Figure 2. Isolation Forest outliers in AdminServer METRIC values (reproduced from the thesis).

DISCUSSION AND CONCLUSION

The results support the feasibility of combining server-wise statistical thresholds, tree-based

classification and unsupervised outlier detection for enterprise APM. The exploratory findings align with studies using ensembles for software anomaly prediction (Alonso et al., 2011; Primartha and Tama, 2017), prediction of unusual HTTP activity (Benova and Hudec, 2023), and intelligent monitoring in related cloud and IoT contexts (Grzonka et al., 2018; Villegas-Ch et al., 2024). These observations concern the specific WLSDM / Oracle WebLogic Server JVM setting and cannot yet be generalized to other enterprise workloads.

WLSDM observations also identified prolonged garbage collection pauses, memory growth, thread-pool pressure, increased response times and JDBC connection-pool bottlenecks. These are plausible diagnostic signals for a future RCA module, but a univariate abnormal metric does not establish causality. An effective next stage must correlate JVM, JDBC, thread, socket, response-time and alarm events across time, and test explanatory evidence against known incidents. Explainability and operator-facing decision support are relevant to trustworthy corrective action (Baş and Süzen, 2023).

Autonomous operational intervention should therefore follow, rather than precede, verified diagnosis. An operational design would map validated root causes to constrained actions, apply approval/authorization rules, record an audit trail, evaluate rollback conditions and measure harmful-action rates. Work on LLM-based Linux administration illustrates the potential of operational agents but also motivates additional security validation (Cao et al., 2024). These capabilities are part of the research roadmap; no end-to-end autonomous recovery experiment is claimed here.

Table 3. Evidence requirements before progressing beyond anomaly detection

Proposed stage	**Required evidence**	**Operational acceptance criterion**
Event detection	Timestamped multivariate WLSDM metrics	Validated recall and false alarm rate
Root cause analysis	Linked GC, heap, thread, JDBC and socket events	Known-incident explanation and review
Action recommendation	Verified cause and a bounded runbook	Traceable operator approval
Autonomous intervention	Authorization, audit log and reversal path	Safe rollback and measured action error

A practical RCA experiment should retain the identity, timestamp, metric name, observed value and detection threshold for every event, then correlate nearby events within the same JVM and application transaction path. The thesis identifies GC pauses, growing JVM memory demand, thread saturation and JDBC connection pressure as recurring operational symptoms; however, these indicators can be concurrent effects rather than independent causes. Evaluation should therefore include manually adjudicated incident timelines and measure whether a suggested explanation identifies the initiating component, not merely the most visible metric.

Before operational execution, reversible actions must be distinguished from changes affecting availability or data integrity. Controlled validation should document authorization, health checks, timeouts, audit logs, rollback conditions and human override. These safeguards are proposed, not experimentally assessed. A future incident-replay test on held-out workloads is informed by studies of software defect prediction (Challagulla et al., 2008) and sandboxed Linux agents (Cao et al., 2024),

neither of which validates this WebLogic workflow.

The principal limitations are synthetic augmentation of a constrained laboratory dataset, labels generated from the input metric itself, a random rather than temporal validation split, class-imbalanced VALUE/JSON results and unresolved specification of the time-based rolling window. Independent, incident-labelled and multivariate validation across distinct Oracle WebLogic Server workloads is required before accuracy, RCA performance or intervention safety can be generalized. Within those bounds, the study establishes an implementable first-stage detection workflow and defines a defensible progression toward RCA and operational autonomy.

References

Alonso J, Belanche L, Avresky DR, 2011. Predicting software anomalies using machine learning techniques. 10th IEEE International Symposium on Network Computing and Applications, pp. 163–170.

Baş E, Süzen AA, 2023. Web uygulama zafiyetlerinin keşfinde açıklanabilir yapay zekânın yeri. International Conference on Applied Engineering and Natural Sciences, 1(1): 517–527.

Benova L, Hudec L, 2023. Web server load prediction and anomaly detection from hypertext transfer protocol logs. International Journal of Electrical and Computer Engineering, 13(5): 5165–5178.

Cao C, Wang F, Lindley L, Wang Z, 2024. Managing Linux servers with LLM-based AI agents: An empirical evaluation with GPT4. Machine Learning with Applications, 17: 100570.

Challagulla VUB, Bastani FB, Yen IL, Paul RA, 2008. Empirical assessment of machine learning based software defect prediction techniques. International Journal on Artificial Intelligence Tools, 17(2): 389–400.

De Franceschi ASM, da Rocha MA, Weber HL, Westphall CB, 1997. Proactive network management using remote monitoring and artificial intelligence techniques. 2nd IEEE Symposium on Computers and Communications (ISCC 1997), pp. 167–171.

Grzonka D, Jakóbik A, Kołodziej J, Pllana S, 2018. Using a multi-agent system and artificial intelligence for monitoring and improving the cloud performance and security. Future Generation Computer Systems, 86: 1106–1117.

Korkutata MF, Şahinaslan Ö, 2025. Application performance monitoring software systems: An approach to the use of artificial intelligence and machine learning. 9th International Conference of Mathematical Sciences (ICMS 2025), Abstract Book, 3–7 September 2025, p. 104, Istanbul, Türkiye. ISBN: 978-605-2124-29-1.

Korkutata MF, Sahinaslan O, 2026. AI Based Anomaly Detection in Application Performance Monitoring Using Random Forest. Boletim da Sociedade Paranaense de Matemática, 44(5): 1–22. https://periodicos.uem.br/ojs/index.php/BSocParanMat/article/view/80843.

Primartha R, Tama BA, 2017. Anomaly detection using Random Forest: A performance revisited. International Conference on Data and Software Engineering (ICoDSE), pp. 1–6.

Sheta SV, 2023. Developing efficient server monitoring systems using AI for real-time data processing. International Journal of Engineering and Technology Research, 8(1): 26–37.

Villegas-Ch W, García-Ortiz J, Sánchez-Viteri S, 2024. Toward intelligent monitoring in IoT: AI applications for real-time analysis and prediction. IEEE Access, 12: 40368–40386.

Design of an AI-Powered Decision Support System for Fabric Quality Control

Sedat Göktaş[1*], Ender Şahinaslan[2]

[1] İstanbul Okan University, Graduate Education Institute, Department of Information Systems, 34959, Istanbul, Turkey
[2] Maltepe University, Faculty of Business and Management Sciences, Management Information Systems, 34857, Istanbul, Turkey

*Corresponding author e-mail: segoktas@stu.okan.edu.tr

Abstract

Fabric quality control in the textile industry traditionally relies on manual visual inspection, which can result in subjective assessments, operator fatigue, and limitations in inspection speed and consistency. This study proposes the design of an Artificial Intelligence (AI)- and Internet of Things (IoT)-enabled Decision Support System (DSS) for automated fabric defect detection, classification, and quality assessment in textile manufacturing environments. The proposed system architecture integrates industrial cameras, IoT sensors, deep learning-based image analysis, and a centralized decision support interface. High-resolution cameras are intended to capture fabric surface images during production, while IoT sensors collect relevant operational and environmental parameters from the production line. Deep learning models are envisaged to analyze the captured images and identify common fabric defects, including weaving irregularities, holes, and stains. The IoT layer is designed to transmit defect information and production telemetry to the DSS, enabling the integration of visual inspection results with operational data. Based on these inputs, the proposed system aims to support automated defect classification, quality grading, timely operator notifications, and the investigation of potential causes of quality deviations. The design also considers data communication, processing latency, system scalability, and integration with Industry 4.0 manufacturing environments. The proposed architecture will be evaluated in future work using metrics such as defect classification accuracy, detection latency, inspection throughput, and the potential to reduce material waste and reliance on manual inspection. The main contribution of this study is the design of an integrated decision support architecture that combines AI-based visual defect analysis with IoT-enabled production monitoring to support consistent, timely, and data-driven fabric quality control. The proposed design provides a basis for the future implementation and experimental evaluation of intelligent quality control systems in textile manufacturing.

Keywords: *Artificial Intelligence, Internet of Things (IoT), Decision Support System, Fabric Quality Control, Industry 4.0*

INTRODUCTION

In the textile industry, fabric quality directly affects the value of the final product and the competitiveness of the manufacturer. The economic stability and balance of a textile weaving company

are largely related to the extent to which fabric defects are compensated for. Therefore, fabric inspection is considered one of the fundamental stages of quality assurance (Conci and Proença, 2000). While quality is an indispensable factor for ensuring customer satisfaction and gaining a competitive advantage, achieving first-quality fabric with zero defects in weaving processes is quite difficult. During the production phase, defects such as stains, holes, and weaving irregularities that occur in the weaving process significantly reduce the quality of the fabric, leading to both revenue loss due to second-quality products and raw material waste (Kumar 2008; Ngan et al., 2011). In garment production, approximately 85% of second-quality products are due to fabric defects, and these products fetch between 45% and 65% of the price of first-quality products (Malek et al., 2012). Therefore, early, reliable, and consistent identification of defects in production is one of the fundamental requirements of quality management.

In practice, fabric quality control largely relies on manual inspection, where operators visually inspect the fabric in motion. This method is limited by subjective assessment differences among operators, fatigue-related inconsistencies, and a limited inspection speed based on human perception. Furthermore, manual quality control only identifies the problem on the fabric itself and cannot generate data on the specific production conditions causing the problem. This makes it difficult to investigate the causes of quality differences.

The topic of automated defect detection through computer vision is still relevant today, based on a long history of research. While early studies in this field relied on spectroscopic, statistical, and model-driven texture analysis (Kumar 2008; Ngan et al., 2011), the success of deep learning in image recognition (LeCun et al., 2015; He et al., 2016) has led to the widespread use of object acquisition architectures and convolutional neural networks in fabric defect detection. For example, Li et al. (2017) proposed a deep learning approach based on the Fisher criterion for patterned fabrics, while Jing et al. (2020) suggested an improved YOLOv3 model. However, much of the research to date treats fabric defect detection as a separate task from the production stages. In contrast, the Industry 4.0 approach aims to integrate physical processes into the digital world through sensors and network technologies, and to base all decisions on data analytics (Lasi et al., 2014; Wang et al., 2018). Decision support systems, on the other hand, provide a basis for deriving meaningful recommendations from this data (Power, 2002).

The aim of this study is to design a decision support architecture that integrates deep learning-based defect information obtained from fabric surface images during quality control with production telemetric values collected via an IoT layer. Experimental analyses of the performance of the designed architecture will be addressed in future studies, and this study does not include any measurement-based findings. The following section presents the design requirements and methodology, design outputs, discussion, and conclusions.

MATERIAL AND METHODS

Material

The proposed system is designed to be positioned at the endpoint of the weaving line or at the fabric quality control and winding point. The system's inputs are grouped into three categories: surface images of the fabric in motion, processing and environmental parameters (telemetry) from sensors on the line, and labeled fabric defect samples for model training and validation. Images are envisioned to be acquired

using high-resolution cameras positioned along the width of the fabric and controlled illumination. Telemetry parameters that may be relevant to defect occurrence, such as warp tension, machine speed, vibration, ambient temperature, and relative humidity, are evaluated. In the application, the pre-training phase of the artificial intelligence model aims to utilize standard fabric defect datasets such as TILDA and AITEX, which are widely accepted in the literature and shared as open access. Following this basic training, the system is planned to be fine-tuned and customized through a transfer learning process based on labeled real-world data collected from the company, in order to adapt to the dynamics of production in the field.

Tasarım kapsamında incelenen hata sınıfları, sahada en sık karşılaşılan üç grup ile sınırlandırılmıştır. The first group includes weaving irregularities (pattern distortions such as skipped wefts, tight or loose weft, warp and warp irregularities), holes, and stains, while the second group encompasses functional requirements of the system, namely identifying and classifying defects, calculating quality grades, informing operators, and investigating root causes. The final group includes non-functional requirements of the system, covering integration with MES/ERP systems, scalability, acceptable processing latency, traceability of records, and the ability to add new defect classes.

Methods

System Architecture

The proposed system is designed in a four-layered structure as shown in Figure 1. The sensing layer houses cameras and IoT sensors that collect raw image and measurement data such as speed, vibration, temperature, and humidity. In the edge processing layer, the raw image data obtained from the sensing layer is pre-processed. Then, it is analyzed with deep learning models to detect and classify faults. The IoT gateway is designed to transfer data from IoT sensors to the communication layer. The communication layer transmits fault events and telemetry data to the central decision support system. The decision support layer consists of data storage, quality grading, cause analysis, notification management, and operator interface. Operator intervention and adjustment of fault parameters are treated as a feedback loop in the architecture.

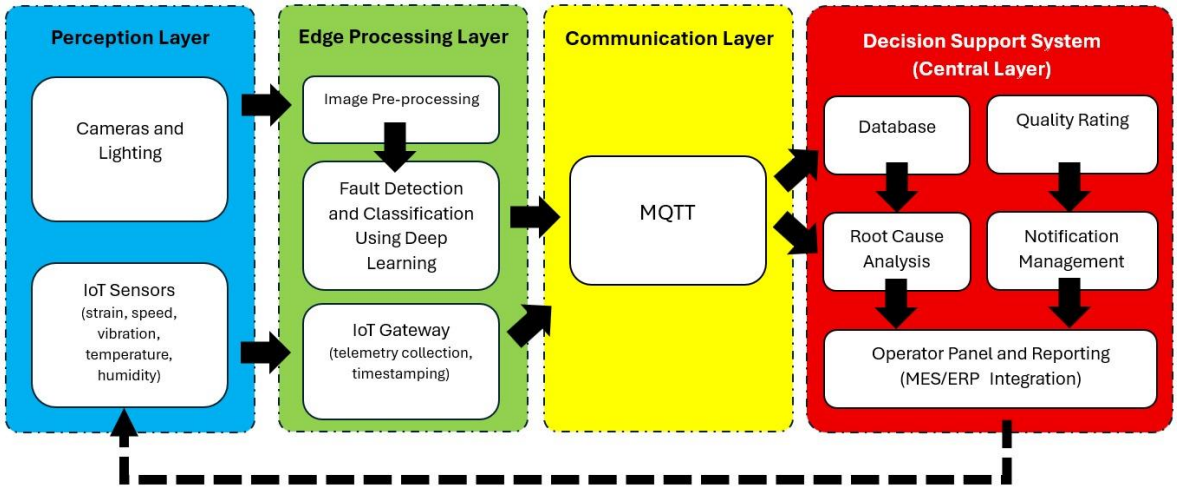


Figure 1. Layered architecture of the proposed fabric quality control decision support system.

Image Acquisition and Preprocessing

To limit image blurring while the quality control machine is in motion, the camera sampling rate is synchronized with the line speed. It is envisioned that an encoder signal on the quality control machine will be used to associate each image frame with a position in meters along the fabric. Preprocessing consists of cropping the fabric edges, brightness normalization, noise reduction, and splitting the image into model-appropriate segments. Performing these steps at the edge layer, before the image data is transmitted to the network, aims to reduce communication overhead and latency (Shi et al., 2016).

Deep Learning-Based Error Analysis

Active learning will further reduce the burden of human supervision (Arıkan et al, 2026). For error analysis, the use of an object detection model such as the YOLO family of convolutional neural networks (Redmon et al., 2016) is considered. The model generates the bounding box, class, and confidence value of the error. If more precise determination of the error space is required, a segmentation model similar to U-Net (Ronneberger et al., 2015) can be used additionally. Detections with confidence values below the threshold are sent for operator verification instead of automatic decision-making, thus maintaining a human-supervised process.

IoT Layer and Data Communication

The IoT gateway associates sensor data with timestamp and fabric location information, allowing it to be matched with production conditions as soon as a fault event occurs. Each fault event message is designed to include timestamp, fabric location information, fault class, confidence value, fault area, and machine ID fields. Communication is planned using the MQTT protocol supported by TLS and device authentication. Local buffering is planned to prevent data loss in case of connection interruptions.

Decision Support System and Quality Rating

The central decision support layer performs four functions: storing error and telemetry data as a time series, quality rating, cause analysis, and notification management. In quality rating, the quality score for a given fabric length (e.g., 1 meter) is calculated using the following formula, depending on the weight and severity of the error classes:

$$Q = 100 - \sum_{j=1}^{m} w_{cj} \cdot s_j$$

In this formula, Q represents the quality score of the fabric cut (0-100), m represents the number of detected defects, cj represents the class of the defect, W_{cj} represents the penalty weight assigned to this defect, and S_j represents the area of the defect, with the severity value normalized between 0 and 1 depending on its location. It is foreseen that the weights here will be adjusted in accordance with the company's quality standards and the opinions of quality experts. The Q value is converted to A-D grades using the thresholds in Table 1.

Table 1. Ranking according to quality score and recommended operator actions.

Grade	Quality Score (Q)	Recommended Action
A	Q >= 90	Normal production flow; recording

B	75 <= Q < 90	Recording and monitoring; notification to the operator
C	60 <= Q < 75	Warning to the operator; manual verification on the sample fabric
D	Q < 60	Suggestion to stop the line or separate the fabric; notification to the quality control supervisor.

In causal analysis, the frequency of defects in a specific class is compared to telemetry values during the time period in which the defect occurred. For example, the relationship between an increase in defect density and warp tension fluctuations or increased vibration is statistically examined. The resulting causes are presented not as definitive conclusions, but as recommendations that the quality control supervisor must verify.

Delay and Evaluation Approach

The total delay from error detection to operator notification is modeled as the sum of the time taken for image acquisition, processing, transmission, and decision phases using the following formula. Measuring each component independently facilitates the identification of bottlenecks in the system..

$$T_{total} = T_{acquisition} + T_{processing} + T_{transmission} + T_{decision}$$

For scalability purposes, image analysis will be distributed to endpoints; only summary events and telemetry messages will be transmitted to the central unit. In the next phase, the system will be tested offline with labeled datasets, then in a laboratory or pilot line, and its outputs will be compared with manual verification. The evaluation will consider class-based performance metrics (accuracy, precision, sensitivity, F_1), total latency (T_{total}), quality control efficiency, preventable loss amount, manual quality control dependency, and latency variation depending on the number of cameras.

RESULTS

Since this study is design-oriented, this section presents the outputs of the design process, not the experimental performance results. The first output is the layered architecture shown in Figure 1. The architecture integrates high-resolution visual inspection data with operational production telemetry collected from line sensors under one umbrella. This allows the quality control process to move beyond a passive sorting mechanism and enables data-driven evaluation where quality control results are directly correlated with real-time production conditions. The second output is the decision scheme presented in Table 1, relating to the conversion of the quality score into grades and quality control operator actions. The threshold values are initial suggestions and will be adjusted according to the company's quality policy. The third output is the mapping of quality control errors to possible production causes and related telemetry parameters, presented in Table 2. This mapping will be used as the initial hypotheses for the first rule set of the causal analysis module.

Table 2. Quality control failure classes, possible manufacturing causes, and associated IoT parameters.

Error Classification	Visual Characteristics	Possible Manufacturing Causes	Related IoT Parameters

Weaving irregularity	Local or linear pattern distortion in regular texture	Warp tension fluctuations, reed and shuttle adjustment, yarn quality, machine speed	Warp tension, machine speed, vibration
Hole	Loss of continuity in texture, background visibility	Mechanical friction, needle or shuttle damage, excessive tension	Tension, vibration, machine identity
Stain	Local color and brightness deviation	Lubrication leakage, paint or chemical splashes, dirty environment	Ambient temperature and humidity, process chemical dosage

The fourth output is the operator interaction flow. When an error is detected, the event is transmitted to the central system, the quality score is updated in the system, and if the threshold is exceeded, a notification is sent to the operator indicating the location and class of the error. For low-confidence detections, the operator is presented for verification, and the approval or rejection decision is recorded in the system. The feedback here is planned to be used in the periodic retraining of the models. The architecture also supports compatibility with Industry 4.0 environments by defining MES/ERP integration points with the MQTT protocol.

DISCUSSION AND CONCLUSION

This study proposes an integrated decision support system architecture that combines deep learning-based image analysis for fabric quality control with IoT-supported production monitoring. Studies in the literature generally focus on the accuracy of fabric defect detection algorithms (Kumar, 2008; Li et al., 2017; Jing et al., 2020). The proposed design, however, positions this algorithmic function within a large-scale system structure that links defect development to production telemetry by matching it with time and location information, performs grading, and transmits the results to the quality control operator. With this implemented method, the aim is not only to detect defects occurring in production lines in real-time, but also to investigate the root cause of these defects through root cause analysis, determining in which processes, under what types of parameters, and under what specific production conditions they occur, and to minimize potential future risks. In this respect, the design is consistent with the new generation quality management trend that uses data in smart manufacturing (Lasi et al., 2014, Wang et al., 2018).

The study has some limitations because the proposed architecture has not yet been fully implemented and experimentally evaluated. Data on accuracy, delay, efficiency, or waste reduction after obtaining experimental results cannot be presented. The quality score weights and threshold values mentioned in this study should be considered as initial recommendations structured in line with expert opinions. The relationships summarized in Table 1 are positioned as theoretical hypotheses awaiting verification. The validity of these assumptions will be tested through comparative analyses based on experimental results and collected field data. Changes in fabric type, pattern, and lighting conditions may negatively affect the generalization ability of the models. Labeled data collection and class imbalance can also be a

significant cost factor. Factors such as edge hardware capacity, network, and data security should also be considered during the implementation phase.

In conclusion, this proposed design model provides a solid foundation for building a consistent, timely, and fully data-driven fabric quality control system in fabric production. Future studies will aim to train the proposed architecture and integrate the system into a pilot production line. In this context, the architecture's potential for improving defect classification accuracy, inspection efficiency, detection delay, material waste, and reducing reliance on manual inspection will be evaluated.

References

Abdel Salam Malek, Jean-Yves Drean, Laurent Bigue, Jean-François Osselin. Optimization of automated online fabric inspection by fast Fourier transform (FFT) and cross-correlation. Textile Research Journal, 2012, 83 (3), pp.256-268. doi: 10.1177/0040517512458340

Arıkan E and Sahinaslan E. 2026. Multi-Model and assertion-based confidence scoring to improve data quality in artificial intelligence training, Boletim da Sociedade Paranaense de Matemática, vol. 44, no. 5, pp. 1–13. Doi: https://dx.doi.org/10.5269/bspm.79776

Conci A, Proença CB. A Computer Vision Approach for Textile Inspection. Textile Research Journal. 2000;70(4):347-350. doi: 10.1177/004051750007000411

He K, Zhang X, Ren S, Sun J, 2016. Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, Las Vegas, USA. doi: 10.1109/CVPR.2016.90

Jing J, Zhuo D, Zhang H, Liang Y, Zheng M. Fabric defect detection using the improved YOLOv3 model. Journal of Engineered Fibers and Fabrics. 2020;15. doi: 10.1177/1558925020908268.

Kumar A, 2008. Computer-vision-based fabric defect detection: A survey. IEEE Transactions on Industrial Electronics, 55 (1): 348-363.

Lasi, H., Fettke, P., Kemper, HG. et al. Industry 4.0. Bus Inf Syst Eng 6, 239–242 (2014). doi: 10.1007/s12599-014-0334-4

LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature 521, 436–444 (2015). doi: 10.1038/nature14539

Li, Y., Zhao, W., & Pan, J. (2017). Deformable patterned fabric defect detection with Fisher criterion-based deep learning. IEEE Transactions on Automation Science and Engineering, 14(2), 1256–1264. https://doi.org/10.1109/TASE.2016.2520955

Ngan HYT, Pang GKH, Yung NHC, 2011. Automated fabric defect detection - A review. Image and Vision Computing, 29 (7): 442-458.

Power DJ, 2002. Decision support systems: Concepts and resources for managers. Quorum Books, Westport, USA.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 779–788). IEEE. https://doi.org/10.1109/CVPR.2016.91

Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science(), vol 9351. Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4_28

Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. IEEE Internet of Things Journal, 3(5), 637–646. https://doi.org/10.1109/JIOT.2016.2579198

Wang, J., Ma, Y., Zhang, L., Gao, R. X., & Wu, D. (2018). Deep learning for smart manufacturing: Methods and applications. Journal of Manufacturing Systems, 48, 144–156. doi: 10.1016/j.jmsy.2018.01.003

Acknowledgment

We sincerely thank our esteemed professors and the organizing team who contributed to the organization of the conference, enabled us to present our work, and provided valuable insights.

Conflict of Interest

There are no conflicts of interest between the authors or with third parties in the preparation and publication of this study.

Author Contributions

All authors contributed to the writing of the work and approved its final version.

Time Series Analysis for Electricity Produced by the Hydroelectric Power Plant in Türkiye

Gözde Karakoc[1*], Yılmaz Akdi[2]

[1]Ankara University, Faculty of Science, Department of Mathematics, 06100, Ankara, Türkiye
[2]Ankara University, Faculty of Science, Department of Statistics, 06100, Ankara, Türkiye

*Corresponding author e-mail: gozcirpan@ankara.edu.tr

Abstract

This study aimed to determine the most suitable forecasting model by analyzing Türkiye's daily hydroelectric power plant (HPP) production data (4199 days). Hidden periodicities in the time series were identified using a periodogram, and the dominant frequencies were incorporated into the models by converting them into Fourier series. Using a 4000 day training set and a 199 day out of sample test set, the performances of Harmonic Regression, ARIMA(7,0,7), and the Machine Learning (XGBoost) algorithm were compared based on the Mean Squared Error (MSE) criterion. The findings revealed that the XGBoost model, augmented with Fourier series, suffered from overfitting on the training data and performed poorly on the test set, while the harmonic model capturing only macro climatic cycles lacked sufficient flexibility. In contrast, the classical ARIMA(7,0,7) model, relying on the system's autoregressive memory and past shocks, achieved the lowest test error and the highest forecasting accuracy compared to the other methods. In summary, the classical Box-Jenkins method demonstrated superior performance in HPP electricity production forecasting based on the MSE criterion.

Keywords: *ARIMA, Hydroelectric Electricity Production, Periodogram, Time Series Analysis, XGBoost*

INTRODUCTION

Highly accurate forecasting of electricity production is a cornerstone of modern power systems management.Among various renewable energy sources, hydroelectric power plants (HPPs) hold a strategic position due to their rapid start-up capabilities (dispatchability) and high operational flexibility. However, modeling daily HPP production volumes is an econometrically challenging task. The underlying time series possesses a dual structure: on one hand, there are long- and short-term deterministic components driven by macro-climatic cycles, snowmelt periods, and seasonal variations; on the other, there are stochastic components characterized by sudden hydrological shocks, operational interruptions, and irregular weather events.

To capture these complex dynamics, researchers have historically relied on parametric linear formulations most notably the Box-Jenkins (ARIMA) methodology, which excels at tracking temporal autocorrelation and linear memory structures. In parallel, recent advancements in artificial intelligence

have popularized non-linear machine learning algorithms, such as Extreme Gradient Boosting (XGBoost), which are distinguished by their capacity to capture intricate, multidimensional relationships between variables. The comparison between statistical time-series models and machine learning algorithms for forecasting electricity consumption and hydroelectric power production has become an increasingly evolving and diverse area of research. In one of the foundational studies on this subject, Niazai et al. (2002) compared the performance of traditional statistical methods (such as ARIMA) with machine learning algorithms using monthly electricity consumption data, evaluating both the error margins of the models and their ability to adapt to complex consumption dynamics. Focusing on the efficacy of linear methods for short-term forecasting, Contreras et al. (2003) found that the inherent autoregressive structure of the ARIMA model yielded highly successful results in predicting daily electricity loads; similarly, Polprasert et al. (2021) determined that the ARIMA methodology provides reliable statistical outcomes for modeling seasonal and cyclical patterns in energy generation at hydroelectric power plants. Conversely, Akdi et al. (2020) which is focusing on the deterministic components within time series, identified latent cycles in daily data from the Türkiye electricity market using a periodogram and demonstrated that modeling these structural periodicities via harmonic regression as opposed to traditional ARIMA/SARIMA methods significantly reduces forecast errors (MSE) and enhances prediction consistency in highly volatile time series. Wang et al. (2025) proved that integrating machine learning models with hybrid optimization algorithms for HPP production forecasting leads to a marked improvement in both prediction accuracy and the operational efficiency of power plants, compared to using standalone models.

This study investigates the performance limitations of traditional econometric models and machine learning algorithms when applied directly to daily HPP production data. Aiming to test the out-of-sample predictive validity of these models, the study compares the forecasting capabilities of harmonic regression, a linear autoregressive ARIMA approach, and an XGBoost algorithm augmented with explanatory variables, within a statistical framework.

MATERIAL AND METHODS

Material

This study utilizes HPP production data, which plays a strategic role in Türkiye's electricity supply. This data serves as a fundamental indicator for examining supply demand imbalances in the electricity market and the impact of renewable energy sources on the grid.

Methods

The Collection of the Data

The dataset analyzed in this study consists of 4,199 daily observations of hydroelectric power production measured in megawatt-hours (MWh) obtained from the official Transparency Platform database of the Energy Markets Operation Corporation (Enerji Piyasaları İşletme A.Ş.-EPİAŞ). To ensure variance stability in the statistical analyses, the natural logarithm of the original series was used. To evaluate the out-of-sample forecasting validity of the developed models, the dataset was divided into two subsets: the first 4,000 observations were used for model training, while the remaining 199 observations served as a test set. The comparative assessment of the forecasting models was conducted based on the Mean Squared Error (MSE) criterion, applied to both the training and the out-of-sample test subsets.

Statistical Analysis

In this study, basic statistical tools and three different forecasting models were applied to data that had undergone a natural logarithmic transformation:

- *Periodogram:* For time series characterized by stochastic shocks and complex seasonal patterns such as those representing HPP production the periodogram statistic was employed to analytically identify periodic cycles inherent in the data. Unlike traditional time domain approaches, this method transforms the series into the frequency domain, calculating via spectral density the extent to which specific frequency components account for the variance within the series (Wei, 2006).
- *Harmonic Regression:* To capture the dominant macro-climatic cycles within the series, periodograms were employed to identify the 10 most significant periodic components. These periods were converted into Fourier series (cosine and sine terms) and modeled directly onto the data using the Ordinary Least Squares (OLS) method (Bloomfield, 2004). Furthermore, the identification of latent cycles via periodogram statistics particularly in time series characterized by inherent periodicity, such as daily electricity production/consumption data and the subsequent modeling of these frequencies through a harmonic regression approach, serves as a robust alternative to traditional models. This methodology is based on the framework introduced to the literature by Akdi et al. (2020).
- *ARIMA*: A Box-Jenkins (ARIMA) model was developed, relying entirely on the internal temporal dependency and autoregressive memory of the training series, without the use of any exogenous explanatory variables (Box et al., 2015). The Akaike Information Criterion (AIC) was used to determine the optimal lag lengths (parameters p, d, and q) for the model, while the standard Augmented Dickey-Fuller (ADF) test was applied to examine the stationarity conditions of the series.
- *XGBoost Algorithm*: The Fourier series components (cosine and sine terms) derived from the periodogram were incorporated as independent variables into the decision tree based XGBoost (eXtreme Gradient Boosting) algorithm, chosen for its high efficacy in modeling complex, non-linear relationships. Operating on the principle of gradient boosting, this framework functions by training weak predictors sequentially, with each subsequent model aiming to minimize the errors of its predecessor. By integrating spectral frequencies as exogenous features, the algorithm's capacity to learn deterministic cycles is enhanced. During the analysis, key hyperparameters namely the learning rate, maximum tree depth, and the number of estimators were optimized to ensure the model's out-of-sample generalizability and to mitigate the risk of overfitting (Chen and Guestrin, 2016).
- *Evaluation of Model Performance:* The Mean Squared Error (MSE) criterion was employed to objectively compare the out-of-sample forecasting performance of the three different approaches developed in this study (Harmonic Regression, ARIMA, and XGBoost) and to identify the most suitable model. Based on the average of the squared deviations between the actual values and the values predicted by the model, this metric by its very nature imposes heavier penalties on large forecasting errors, thereby providing a statistical measure of the algorithm's overall prediction accuracy (Makridakis et al., 2008).

All statistical tests, frequency analyses, and machine learning modeling processes were conducted using the Python programming language (utilizing the pandas, scipy, statsmodels, and xgboost libraries), with a statistical significance level of 5% (α=0.05) applied to the validity of parameter estimates.

RESULTS

In the first stage of the analysis, the time series plot, Autocorrelation Function (ACF), and Partial Autocorrelation Function (PACF) were examined along with a stationarity test to investigate the characteristic dynamics of Türkiye's daily HPP electricity production series. As shown in the plot in Figure 1, the daily electricity generation data exhibit long term fluctuations and a cyclical structure. The very slow decay of the coefficients toward zero in the Autocorrelation Function (ACF) plot indicates a high level of persistence and a long-memory effect within the series. Meanwhile, in the Partial Autocorrelation Function (PACF) plot, the prominent peak at the first lag, followed by a sharp drop toward the confidence interval boundaries, clearly demonstrates that a first order autoregressive (AR(1)) process strongly dominates the nature of the data, reflecting the significant influence of the previous period's production values on the current state.

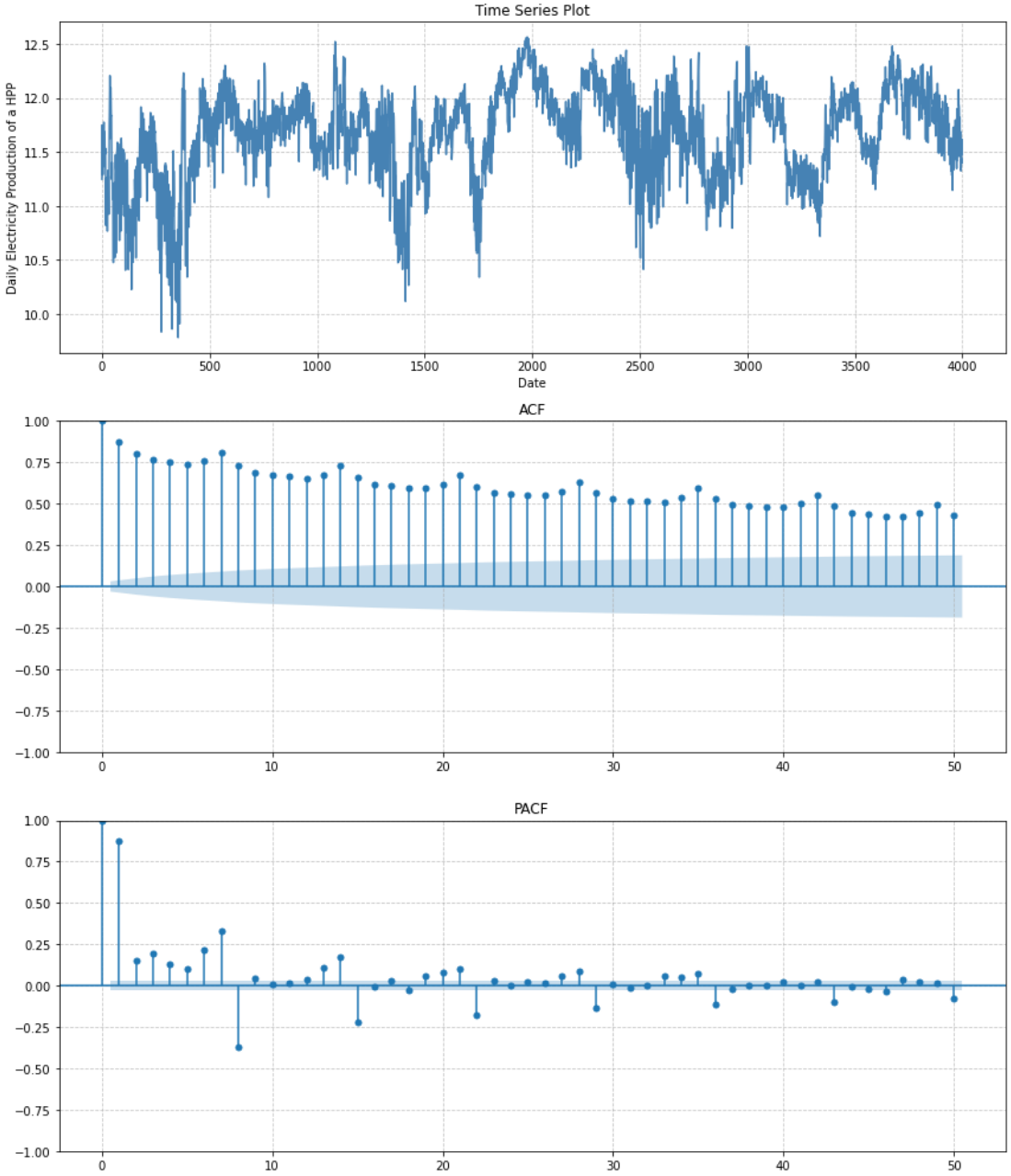


Figure 1. Time series plots for daily electricity production of HPP

According to the Augmented Dickey-Fuller (ADF) unit root test results presented in Table 1, the test statistic for the series (-4.7355) is smaller than all the critical values at the 1%, 5%, and 10% significance levels; this indicates that the series is stationary and does not contain a unit root.

Table 1. The results of the ADF unit root test

	t-Statistics	**p-value**
ADF test statistic	-4.7355	0.0001

		t-Statistics	p-value
Test critical values	1% level	-3.4320	
	5% level	-2.8623	
	10% level	-2.5672	

The statistically significant periods presented in Table 2 play a critical role in enhancing the accuracy of energy production forecasts by revealing underlying hydrological and climatic cycles within the daily electricity production data of the HPP. The 363.64-day period, which exhibits the highest spectral power, reflects the fundamental annual hydrological calendar driven by precipitation regimes and snowmelt. Meanwhile, the longer periods ranging from 1,000 to 4,000 days (approximately 2.7 to 11 years) correspond to long-term macro-climatic fluctuations and drought trends, whereas the shorter periods ranging from 121.21 to 181.82 days (approximately 4–6 months) represent seasonal transitions within the year.

These identified dominant frequencies were converted into Fourier series to model the cyclical patterns in the data and were subsequently incorporated as explanatory features into both Harmonic Regression and XGBoost algorithms.

Table 2. The largest 10-periodogram values

$I_n(w_{(i)})$	Period	V_i	c_α	Conclusion
111.1299	363.64	0.00275	0.00528	significant
109.0946	1333.33	0.00075		significant
34.4560	1000	0.001		significant
29.6081	800	0.00125		significant
25.9526	4000	0.00025		significant
19.1586	666.67	0.0015		significant
18.2083	181.82	0.0055		significant
15.4679	121.21	0.00825		significant
12.4665	500	0.002		significant
11.1235	400	0.0025		significant

An examination of the harmonic regression parameter estimates presented in Table 3 reveals that all model parameterswith the exception of the A_{10} coefficient ($p = 0.9985$) are statistically significant at the 1% significance level. Furthermore, the constant term (μ) was estimated at 11.6948 and was found to be statistically significant. All other A_i and B_i coefficients excluding A_{10} indicate the magnitude and direction of the cyclical effects (amplitudes) that the identified periods exert on production.

Table 3. Harmonic Regression Parameter Estimates

Variable	Parameter estimate	Standard error	t-value	P>|t|
Intercept (μ)	11.6948	0.0041	2833.2694	<0.0001
A_1	-0.2048	0.0058	-35.0894	<0.0001
B_1	0.1166	0.0058	19.9827	<0.0001
A_2	-0.1333	0.0058	-22.8279	<0.0001
B_2	-0.1918	0.0058	-32.8590	<0.0001
A_3	0.0698	0.0058	11.9609	<0.0001
B_3	-0.1111	0.0058	-19.0393	<0.0001
A_4	-0.0324	0.0058	-5.5477	<0.0001
B_4	-0.1173	0.0058	-20.0923	<0.0001
A_5	-0.1077	0.0058	-18.4423	<0.0001
B_5	-0.0373	0.0058	-6.3821	<0.0001
A_6	-0.0359	0.0058	-6.1471	<0.0001
B_6	-0.0911	0.0058	-15.5994	<0.0001
A_7	0.0272	0.0058	4.6595	<0.0001
B_7	0.0915	0.0058	15.6663	<0.0001
A_8	0.0853	0.0058	14.6054	<0.0001
B_8	-0.0215	0.0058	-3.6809	0.0002
A_9	0.0627	0.0058	10.7487	<0.0001
B_9	0.0479	0.0058	8.2077	<0.0001
A_{10}	0.0000	0.0058	0.0019	0.9985
B_{10}	-0.0746	0.0058	-12.7733	<0.0001

$$\begin{aligned}\widehat{Y_t} = 11.6948 - - 0.2048\cos\left(\frac{2\pi t}{363.64}\right) + 0.1166\sin\left(\frac{2\pi t}{363.64}\right) \\ - 0.1333\cos\left(\frac{2\pi t}{1333.33}\right) - 0.1918\sin\left(\frac{2\pi t}{1333.33}\right) + 0.0698\cos\left(\frac{2\pi t}{1000}\right) \\ - 0.1111\ \sin\left(\frac{2\pi t}{1000}\right) - 0.0324\cos\left(\frac{2\pi t}{800}\right) - 0.1173\ \sin\left(\frac{2\pi t}{800}\right) \\ - 0.1077\cos\left(\frac{2\pi t}{4000}\right) - 0.0373\sin\left(\frac{2\pi t}{4000}\right) - 0.0359\cos\left(\frac{2\pi t}{666.67}\right) \\ - 0.0911\sin\left(\frac{2\pi t}{666.67}\right) + 0.0272\ \cos\left(\frac{2\pi t}{181.82}\right) + 0.0915\sin\left(\frac{2\pi t}{181.82}\right) \\ + 0.0853\cos\left(\frac{2\pi t}{121.21}\right) - 0.0215\sin\left(\frac{2\pi t}{121.21}\right) + 0.0627\ \cos\left(\frac{2\pi t}{500}\right) \\ + 0.0479\sin\left(\frac{2\pi t}{500}\right) - 0.0746\ \sin\left(\frac{2\pi t}{400}\right), t \\ = 1,2,\ldots,4000 \qquad (1)\end{aligned}$$

An examination of the parameter estimation results for the ARIMA(7,0,7) model in Table 4 reveals that, with the exception of the AR(2) (p=0.058) and AR(5) (p=0.674) coefficients, all autoregressive (AR) and moving average (MA) terms are statistically significant at the 1% level ($p < 0.01$). In particular,

the AR(7) coefficient which exhibits a high estimated value (0.9475) and a very high degree of statistical significance demonstrates that the model has successfully captured the strong 7-day (weekly) periodic cycle present in the daily electricity generation data, confirming that the production level from one week prior exerts a decisive influence on current production.

Table 4. $ARIMA\ (7,0,7)$ Model parameter estimates and significance values

Variable	Parameter estimate	Standard error	z-value	P>|z|
Intercept (μ)	11.6815	0.371	31.449	0.000
AR(1)	-0.0242	0.007	-3.655	0.000
AR(2)	0.0128	0.007	1.892	0.058
AR(3)	0.0334	0.007	4.810	0.000
AR(4)	0.0289	0.007	4.289	0.000
AR(5)	0.0028	0.007	0.420	0.674
AR(6)	-0.0290	0.007	-4.452	0.000
AR(7)	0.9475	0.006	146.535	0.000
MA(1)	0.7901	0.013	59.712	0.000
MA(2)	0.6636	0.017	38.457	0.000
MA(3)	0.5238	0.020	25.724	0.000
MA(4)	0.4679	0.022	21.513	0.000
MA(5)	0.4758	0.021	23.003	0.000
MA(6)	0.5519	0.017	31.989	0.000
MA(7)	-0.2155	0.013	-16.400	0.000
Variance (σ^2)	0.0216	0.000	65.038	0.000

Note: AIC = -3954.034, Log Likelihood = 1993.017

When the estimated coefficients are substituted into the regression equation, the explicit form of the ARIMA(7,0,7) model which incorporates the 7-day periodicity in the series and the impact of past shocks is expressed as follows:

$$\begin{aligned}\hat{Y}_t = 11.6815 &- 0.0242\,Y_{t-1} + 0.0334Y_{t-3} + 0.0289Y_{t-4} - 0.0290Y_{t-6} \\ &+ 0.9475Y_{t-7} + \varepsilon_t + 0.7901\varepsilon_{t-1} + 0.6636\varepsilon_{t-2} + 0.5238\varepsilon_{t-3} \\ &+ 0.4679\,\varepsilon_{t-4} + 0.4758\varepsilon_{t-5} + 0.5519\varepsilon_{t-6} - 0.2155\,\varepsilon_{t-7}, t \\ &= 1,2,\dots,4000 \qquad (2)\end{aligned}$$

Here, ε_t represents the error term (white noise) at time t ($\varepsilon_t \sim WN(0,0.0216)$).

The optimal hyperparameters for the XGBoost algorithm, presented in Table 5, include a tree depth kept at a low level of 3; this allows the model to capture the periodic cycles already present in the input independent variables without introducing unnecessary complexity. Furthermore, a low learning rate of 0.05 combined with 100 decision trees ensures a gradual and balanced learning process, while a subsample ratio of 0.8 imparts a stochastic nature to the model thereby reducing variance and enhancing the algorithm's resilience against potential shocks within the series.

Table 5. Optimum hyperparameters used for the XGBoost model

Hyperparameter	Value

Number of Trees (n_estimators)	100
Maximum Tree Depth	3
Learning Rate	0.05
Subsample Ratio	0.8

An examination of the out-of-sample forecasting results presented in Table 6 reveals significant structural differences in the predictive dynamics of the three models tested. Harmonic Regression, which relies solely on deterministic cycles, produces a smooth, narrow-band forecast series and fails to capture the daily stochastic shocks present in the data. In contrast, the ARIMA model incorporating endogenous autoregressive lags exhibits the most dynamic response to short-term volatility, achieving the highest degree of alignment with the directional shifts in actual values. Meanwhile, the XGBoost algorithm constrained in depth and utilizing only periodic Fourier components as inputs generates constant, step-like forecasts at specific time intervals, a characteristic inherent to decision trees; consequently, it models sudden fluctuations in the data through a more rigid generalization toward the mean.

Table 6. Out of sample predictions of the models

Date	Actual Value	Harmonic	ARIMA	XGBoost
4001	11.4920	11.4255	11.3626	11.5307
4002	11.2549	11.4252	11.2661	11.5277
4003	11.7803	11.4247	11.5568	11.5204
4004	11.5157	11.4241	11.4875	11.5204
4005	11.6066	11.4232	11.4885	11.5204
…	…	…	…	…
4195	11.6853	11.2318	11.5835	11.5692
4196	11.7561	11.2354	11.6147	11.5692
4197	11.4387	11.2390	11.5742	11.5692
4198	11.4663	11.2427	11.5073	11.5692
4199	11.5178	11.2464	11.6666	11.5638

An examination of Table 7 reveals that the ARIMA(7,0,7) model which yields the lowest MSE values for both the training (0.0217) and testing (0.2347) sets is the most successful method for HPP forecasting. In the harmonic regression model, while the training error (0.0678) is relatively low, the test error (0.2347) increases significantly; this is because the method focuses exclusively on fixed (deterministic) cycles and fails to capture the random (stochastic) shocks present in out-of-sample data. Meanwhile, the XGBoost model exhibits higher error levels in both the training (0.0420) and testing (0.2570) phases compared to ARIMA, indicating that the algorithm's absolute forecasting performance is insufficient. Consequently, the ARIMA approach which accounts for the series own historical values and error terms has demonstrated greater forecasting accuracy than other methods that rely solely on Fourier components for the prediction process.

Table 7. Performance comparison of the models in terms of in-sample and out of sample MSE

Model	Training MSE	Test MSE
Harmonic Regression	0.0678	0.2347
ARIMA(7,0,2)	**0.0217**	**0.0714**
XGBoost	0.0420	0.2570

An examination of Figure 2, which presents an out-of-sample forecast comparison (t=4001–4030), reveals the performance of three distinct forecasting models against an actual dataset

characterized by high volatility. The ARIMA model successfully adapted to the dynamic structure of the series, demonstrating a clear superiority in capturing directional movements, sudden fluctuations, and turning points; however, it exhibited a slight tendency to under forecast at extreme peaks (notably around t=4011 and t=4017). In contrast, the harmonic regression and XGBoost models failed to resolve the non-deterministic variance and short-term shocks within the series, producing nearly linear, smoothed trend curves that proved inadequate for capturing the series inherent complexity. Consequently, the analysis suggests that the ARIMA model is statistically more reliable than the other models for forecasting the complex and volatile nature of this time series, and that it performs better in capturing the series dynamic memory.

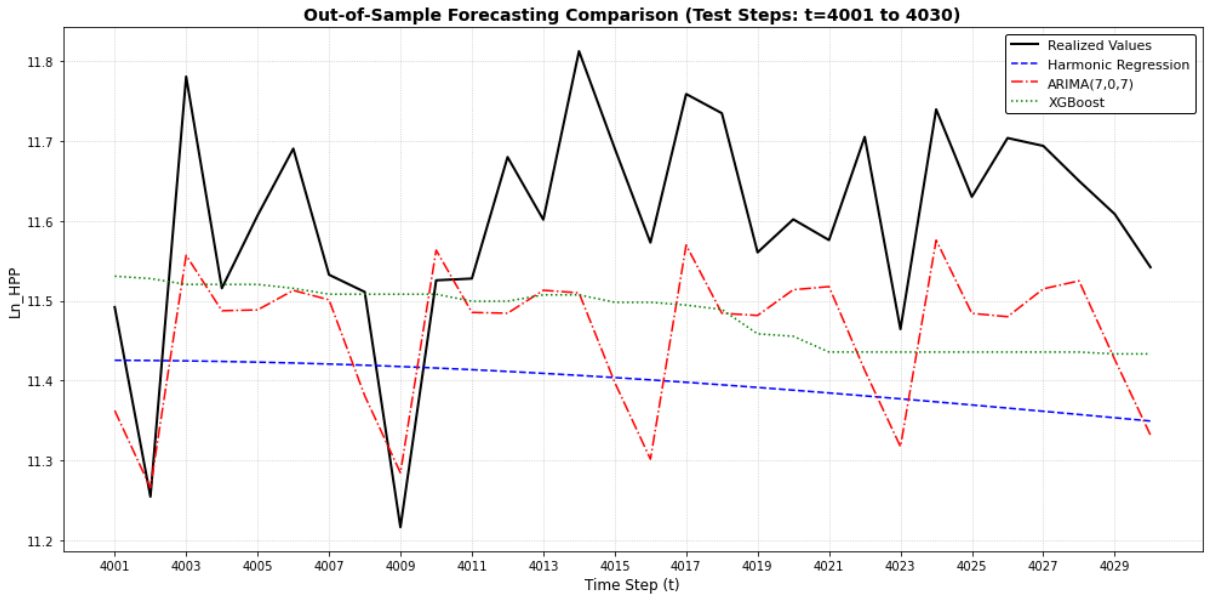


Figure 2. Comparison of realized values and model predictions

Note: For better clarity and to ensure the data points are visually distinguishable, the graph is restricted to the first 30 days of the test set.

DISCUSSION AND CONCLUSION

This study evaluates the out-of-sample forecasting performance of harmonic regression, ARIMA(7,0,7), and XGBoost models in predicting daily electricity production from hydroelectric power plants (HPPs). The findings reveal that these models exhibit distinct structural asymmetries when faced with the high stochastic volatility inherent in daily HPP time series.

Harmonic regression produces an overly smoothed trend that effectively filters out high-frequency daily shocks. Similarly, the XGBoost algorithm constrained by a shallow depth (max_depth=3) and limited to purely periodic inputs demonstrates an algorithmic rigidity; it fails to capture instantaneous fluctuations, instead yielding step-like forecasts that converge toward the mean. In contrast, the linear ARIMA(7,0,7) model proves highly adaptive, leveraging its intrinsic autoregressive memory to track directional shifts and turning points with greater accuracy.

Ultimately, while machine learning frameworks offer robust non-linear modeling capabilities, ARIMA-type models which directly capture temporal memory are exceptionally well suited for short-

term HPP forecasting, where persistence and lag dynamics are dominant. Future research will explore hybrid architectures that integrate autoregressive structures with exogenous hydrological variables to enhance the sensitivity of machine learning models within highly volatile energy systems.

References

Akdi Y, 2012. Zaman Serileri Analizi (Birim Kökler ve Kointegrasyon), 3.Baskı. Gazi Kitabevi, Ankara.

Akdi Y, Gölveren E, Okkaoğlu Y, 2020. Daily electrical energy consumption: Periodicity, harmonic regression method and forecasting. Energy, 191, 116524.

Bloomfield P, 2004. Fourier analysis of time series: An introduction, 2nd Edition. John Wiley & Sons Inc., New York/ Chichester / Weinheim / Brisbane / Singapore / Toronto.

Box GEP, Jenkins GM, Reinsel GC, Ljung GM, 2015. Time Series Analysis: Forecasting and Control, 5th Edition. John Wiley and Sons Inc., Hoboken, New Jersey.

Chen T, Guestrin C, 2016. XGBoost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 785-794.

Contreras, J., Espinola, R., Nogales, F. J., ve Conejo, A. J. (2002). ARIMA models to predict next-day electricity prices. IEEE Power Engineering Review, 22(9), 57. https://doi.org/10.1109/MPER.2002.4312577

Enerji Piyasaları İşletme A.Ş. (EPİAŞ) Şeffaflık Platformu. Access address: https://seffaflik.epias.com.tr/electricity/electricity-generation/ex-post-generation/real-time-generation; Date of access: 01.07.2025

Makridakis S, Wheelwright SC, Hyndman RJ, 1998. Forecasting: methods and applications. 3rd Edition John Wiley & Sons.

Polprasert, J., Nguyen, V. A. H., ve Charoensook, S. N. (2021). Forecasting models for hydropower production using ARIMA method. 2021 9th International Electrical Engineering Congress (iEECON). https://doi.org/10.1109/iEECON51072.2021.9440293

Wang, C., Li, C., Feng, Y., ve Wang, S. (2025). Predicting hydropower generation: A comparative analysis of machine learning models and optimization algorithms for enhanced forecasting accuracy and operational efficiency Ain Shams Engineering Journal,16, 3, 103299.

Wei WWS, 2006. Time series analysis: Univariate and multivariate methods. 2nd ed. Pearson Addison Wesley, USA.

A Method for the Evolution of Agent Modeling Languages

Ahmet Can[1*], Geylani Kardas[2]

[1]Ege University, Graduate School of Natural and Applied Sciences, Department of International Computer, 35100, Izmir, Türkiye

[2]Ege University, International Computer Institute, 35100, Izmir, Türkiye

*Corresponding author e-mail: can_ahmetcan24@hotmail.com

Abstract

Domain-specific modeling languages (DSMLs) used in the development of agent-based systems may be affected by changes in target execution platforms, external libraries and service infrastructures. Reflecting every change in external dependencies directly in the core metamodel, the graphical editor or the transformation rules may cause platform knowledge to spread into the domain language and increase the maintenance burden. In this study, an evolution method is proposed for managing certain implementation changes of agent modeling languages that originate from external platform and technology dependencies. The method consists of the following steps: identifying the external dependency and the target platform, classifying the change with respect to the domain/language level and the platform implementation level, assessing the adequacy of the existing language concepts, defining or updating the appropriate platform mappings in a model called ExtSpec, processing the agent model together with ExtSpec, regenerating the target code and validating the generated code on the target platform. In an incremental application carried out as a case study on the model-driven development of a multi-agent hotel reservation system, the platform counterparts of the message-sending and movement behaviors were changed in steps V1 and V2, while the core metamodel of DSML4BDI (the agent DSML used to model the agents), the agent model and the modeling editor were preserved; the change was reflected in ExtSpec and in the regenerated agent implementation code. In step V3, a boundary case was examined in which a new agent call behavior could be represented with the existing internal action concept. Executions in the agents' implementation environment showed that platform evolution is not merely a matter of action-name mapping; parameter semantics and the execution context can also play a role.

Keywords: *Agent Modeling Languages, DSML Evolution, Model-driven Engineering, ExtSpec, Platform Implementation Change*

INTRODUCTION

Model-Driven Engineering (MDE) uses models as the primary development artifacts in order to manage platform complexity at higher levels of abstraction and to express domain concepts directly with models

(Brambilla et al., 2017). In MDE, Domain-Specific Languages (DSLs) —often used to create models— aim to increase expressiveness by providing a vocabulary and notations specialized for a particular application domain (Wąsowski and Berger, 2023). A Domain-Specific Modeling Language (DSML), in turn, is a language, mostly defined visually, for expressing the concepts of an application domain and the relationships and rules among these concepts at the model level (Kardas et al., 2023). A DSML is generally defined by its abstract syntax, concrete syntax and semantics; the developer can thus work directly with the concepts of the relevant domain instead of implementation details (Challenger et al., 2014).

A software agent is an autonomous software entity that can perceive its environment, make decisions in line with certain goals and act in order to achieve these goals. Systems in which multiple agents interact with each other and with their environment in line with shared or individual goals are called Multi-Agent Systems (MASs) (Wooldridge, 2009). In the development of MASs, DSMLs are used so that the internal structures of agents, their behaviors and the interactions among them can be modeled at a high level of abstraction (Challenger et al., 2014; Alaca et al., 2021).

The domain concepts represented by agent modeling languages or DSMLs and the platforms on which these concepts are executed may not change at the same pace. In this study, a "platform implementation change" refers to a change in the operation, internal action, external action or call form through which a domain concept in the model is realized on the target execution platform, while the meaning of that concept is preserved. For example, while the "message sending" behavior in the model is semantically preserved, the internal action or operation that the target platform offers for this behavior may change. Similarly, an external action may be realized under different names or through different runtime mechanisms in different platform versions. In such cases, the target implementation changes even though the domain meaning is preserved. Handling such changes directly in the metamodel, the graphical editor or the code generator each time may cause platform knowledge to spread into the language definition and the transformation logic. In a DSML, metamodel changes can affect not only the abstract syntax but also the graphical concrete syntax and the associated modeling tools (Tolvanen et al., 2025).

Motivated by this shortcoming, this study proposes an evolution method for managing implementation changes that originate from the execution platform or the external technologies underlying an agent modeling language, without unnecessarily interfering with the syntax and semantics of the core agent modeling language. In the proposed method, it is first determined whether the change in the external dependency affects the domain/language level or only the platform implementation; platform implementation changes for which the existing domain concepts are adequate are localized in a separate mapping model called ExtSpec. ExtSpec is the supporting model used in the proposed method to manage platform-level changes. The agent model and ExtSpec are processed together during code generation, and the resulting output is validated on the target platform.

To apply and evaluate the proposed method, we used the DSML4BDI agent modeling language (Kardas et al., 2018), which supports the MDE of MASs based on the Belief-Desire-Intention (BDI) agent planning architecture (de Silva et al., 2020) that is well known in the literature. DSML4BDI supports expressing BDI concepts such as "Agent", "Belief", "Goal", "Plan" and "Action" at the model level and generating executable code from models for the Jason MAS implementation platform. Jason,

used here as the target platform, is a programming and execution platform that supports the implementation of AgentSpeak-based BDI agents (Bordini et al., 2007; Boissier et al., 2020). In this study, running the generated AgentSpeak Language (ASL) code directly in a separate Jason project, without an additional implementation layer, is called “Jason validation” on first use; thereafter it is referred to as validation in the Jason environment.

The main contributions of the work presented in this paper are as follows: (i) the assessment of external dependency changes in the context of MAS DSMLs in an explicit decision step, in terms of their domain/language impact and their platform implementation impact; (ii) the separation of platform-dependent, changeable mapping information from the core MAS DSML by means of a mapping model such as ExtSpec; (iii) the joint processing of core MAS DSMLs (e.g., DSML4BDI) and mapping models (e.g., ExtSpec) in a common code generation chain; and (iv) the examination of the method through an incremental case study and executions in the Jason environment. The proposed method builds on the approach proposed for DSLs by Fall and Syriani (2021). However, whereas Fall and Syriani (2021) address the generation of an extension metamodel, concrete syntax and mappings in cases where the external library introduces new concepts into the DSL, our work first tests the adequacy of the existing MAS DSML concepts; where they are adequate, only the platform implementation mapping is updated, without extending the core language.

The study was evaluated through three research questions (RQs):

RQ1: Can a change in the platform implementation be managed without changing the core MAS DSML?

RQ2: Can separating platform-specific implementation information into ExtSpec limit the propagation of the change to the DSML4BDI metamodel and the graphical concrete syntax?

RQ3: Can changes made through ExtSpec be transferred to the target platform by code regeneration and validated by execution on the target platform?

These questions test, respectively, the method’s claims of localizing the change, preserving the core language components and validating the generated result on the target platform.

The remainder of the paper is organized as follows. The rest of this section discusses the literature on DSL/DSML evolution and MAS DSMLs. The Material and Methods section introduces the DSML4BDI language, the tools and the case study used, describes the proposed partial evolution method and shows how its steps are applied to DSML4BDI. The Results section presents the incremental application of the method through the case study and presents the findings. Finally, the Discussion and Conclusion section evaluates the research questions, discusses the applicability limits of the method and the threats to validity, and presents the conclusion and future work.

Related Work

In this subsection, the studies in the research area addressed by the proposed method are discussed from two main perspectives. First, studies on DSL/DSML evolution, metamodel-model co-evolution, concrete syntax and external dependency changes are examined. Then, DSML studies on the model-driven development of MASs are reviewed, and their position with respect to the management of changes in the underlying agent execution platform or external technologies is discussed.

DSL/DSML Evolution and External Dependencies

A significant part of the literature on DSL/DSML evolution focuses on metamodel changes and their effects on existing models and other language components. Hebig et al. (2017) systematically reviewed 31 approaches that support metamodel-model co-evolution, proposed a taxonomy of solution techniques and classified the existing approaches according to this taxonomy. Rose et al. (2010) treat model migration as the updating of model instances in response to metamodel evolution and present the Epsilon Flock language for this purpose. These studies show that metamodel evolution and the conformance of existing models constitute a problem that must be addressed together.

More recent studies address the evolution problem with respect to broader parts of the language ecosystem. Bettini et al. (2022) present an approach that analyzes interdependent metamodel changes and warns the developer about changes that may lead to invalid states. Zhang et al. (2024) propose automated grammar transformation based on configurable rules in order to reduce the workload arising from the manual co-evolution of the grammar in metamodel-based language evolution. Tolvanen et al. (2025) present a tool-independent framework that evaluates the effects of changes in the abstract syntax, concrete syntax and constraints of graphical modeling languages on the language definition, modeling tools, code generators and existing models. In addition, Zhang and Strüber (2024) systematically examined 1002 GitHub repositories and presented empirical findings on development practices, evolution activities and the co-evolution of related components in Xtext-based DSLs.

Fall and Syriani (2021) presented an approach for the evolution of DSLs that depend on external libraries. In their work, the evolution of the external library on which the DSL or its code generator depends is considered one of the causes of DSL evolution, and the focus is on the problem that, in current practice, every change or addition in the external library has to be reflected in the DSL manually by the language engineer. In the proposed process, the metamodel, concrete syntax and mappings that support the language extension in question are generated on the basis of extension definitions and functions. The extension mapping contains mapping information at the concept, function and parameter levels. In this respect, their work constitutes the main reference for the research described in this paper with regard to systematically reflecting changes in external dependencies in the DSL.

When the existing studies are considered together, it is seen that a significant part of the literature concentrates on cases in which the metamodel changes and, consequently, models, grammars, concrete syntax or tools need to co-evolve. In this study, by contrast, the problem is a more limited class of evolution in which the existing MAS domain concepts do not change, but the way these concepts are implemented on the underlying execution platform does. For this reason, in this study, Fall and Syriani's (2021) idea of evolution driven by external dependencies is specialized for the relationship between a MAS DSML and its execution platform; for changes where the existing domain concepts are adequate, a more lightweight path is followed that relies on updating the platform mapping in a separate model called ExtSpec, instead of generating an extension metamodel and concrete syntax.

Domain-Specific Modeling Languages for Multi-Agent Systems

Agent-Oriented Software Engineering (AOSE) is the field of software engineering concerned with methods, languages and tools for the systematic analysis, design and development of software agents and MASs (Shehory and Sturm, 2014). Within AOSE, various modeling languages and DSMLs have

been developed in order to handle the complex structure of MASs at a higher level of abstraction. For example, DSML4MAS offers a model-driven development approach that supports the platform-independent specification of MASs, and its development environment also covers functions such as model validation and model transformation (Hahn, 2008; Warwas and Hahn, 2009). SEA_ML, as a MAS DSML with defined syntax and semantics, supports the model-driven development of real MASs with graphical tools and, in particular, the modeling of agents working in the Semantic Web environment (Challenger et al., 2014).

Among BDI-oriented modeling studies, DSML4BDI supports the graphical modeling of the agent components of the BDI architecture and the generation of implementations for the Jason platform through model-to-text transformations (Kardas et al., 2018). More recently, Çelik et al. (2023) presented a modeling language called SAML, which supports the online creation of BDI agents, together with an online, platform-independent modeling/development environment aimed at low-code development. Zschaler and Polack (2023), in turn, propose a vision of a family of DSMLs for the development of robust, efficient, maintainable and trustworthy agent-based simulations, emphasizing bidirectional transparency and traceability extending from domain knowledge to implementation and the interpretation of results.

The evaluation of MAS DSMLs has also been a separate research topic. AgentDSM-Eval provides a framework and a supporting online tool for systematically evaluating MAS DSMLs along various quantitative and qualitative dimensions related to agent software development (Alaca et al., 2021).

Although the AOSE studies mentioned above make important contributions with respect to the model-driven development of MASs, the modeling of BDI concepts, code generation, online modeling and DSML evaluation, they do not directly address, as their main problem, limiting the propagation of implementation changes occurring in the underlying agent execution platform or external technology to the core syntax and semantics of the existing MAS DSML. The work presented in this paper focuses on this specific problem and aims to manage platform implementation changes in a separate mapping model, as long as the domain concepts of a MAS DSML can be preserved.

MATERIAL AND METHODS

This section describes the material used in the study and the method proposed for the partial evolution of MAS DSMLs. It first introduces the modeling language, tools and case study that form the application setting, and then presents the proposed evolution method and how it was applied step by step to DSML4BDI.

Material

This subsection introduces the DSML4BDI language, the supporting tools and the target implementation platform used to apply and evaluate the proposed method. It also describes the HotelBookingMAS case study, whose versions are used to examine the proposed method’s behavior when the platform implementation changes.

Language and Tools

In this study, the proposed method is applied to DSML4BDI, an example MAS DSML. DSML4BDI is a DSML that supports the model-driven development of BDI-based software agents; it enables the

graphical modeling of BDI components and the generation of code for the Jason platform from these models (Kardas et al., 2018). The Ecore/EMF-based (Steinberg et al., 2008) metamodel of DSML4BDI, its Sirius-based graphical modeling environment (Eclipse Foundation, 2013) and its implementation structure associated with the Jason MAS platform (Bordini et al., 2007) provide a suitable application setting for examining the dependency between a MAS DSML and its target implementation platform with the proposed method.

The main reason for choosing DSML4BDI is that it directly supports domain concepts of BDI agents such as *Agent*, *Belief*, *Goal*, *Plan* and *Action*, that these concepts can be modeled graphically and that code executable in the Jason environment can be generated from model instances (Kardas et al., 2018). In this way, the effects of platform implementation changes on the core MAS DSML, the graphical modeling environment, code generation and the target implementation platform can be examined together in the proposed method.

In the application in this study, DSML4BDI was used as the core MAS DSML, ExtSpec (developed within this study) to represent platform-specific changeable mapping information, Epsilon EGL/EGX (Rose et al., 2008) for the transformation from the model to the target code and, for validation, the Jason MAS platform included in the JaCaMo MAS programming framework (Boissier et al., 2020). The Ecore-based metamodel of DSML4BDI, the Sirius graphical modeling environment and the model transformation/generation tools used are discussed together with the application examples given in Figure 2 and Figure 3.

The main components used in applying the method and their roles are summarized in Table 1. As Table 1 shows, ExtSpec is not a modeling language that replaces DSML4BDI. The role of ExtSpec in this study is to keep, in a separate model, the changeable counterparts on the target MAS implementation platform of the behaviors expressed in the core MAS DSML. This prevents platform-specific mapping information from being included directly in the DSML4BDI metamodel and the graphical modeling environment.

Table 1. Tools used in applying the proposed method to DSML4BDI and their roles.

Method component	Implementation tool	Role
Core MAS DSML	DSML4BDI / Ecore	Represents domain concepts such as *Agent, Goal, Plan, Belief* and *Action*.
Graphical concrete syntax	Eclipse Sirius	Enables the DSML4BDI model to be created graphically.
Platform mapping model	ExtSpec / Ecore-XMI	Holds platform-specific changeable implementation information.
Code generation	Epsilon EGL/EGX	Generates the target ASL code by reading DSML4BDI and ExtSpec together.
Target MAS implementation platform	Jason	The environment in which the generated ASL code is executed and validated.

Case Study

To examine the proposed method incrementally, a case study covering the development of a multi-agent hotel reservation system called HotelBookingMAS was carried out. In the conceptual structure of the system, interactions such as room inquiry, reservation and communication are assumed to take place between a customer agent and a hotel agent. This structure was defined to illustrate the inter-agent communication in the hotel reservation process and the actions directed at the target implementation platform.

However, the aim of this study is not to develop a fully functional multi-agent hotel reservation system, but to examine how the proposed method behaves in the face of changes in the platform implementation. Therefore, in order to simplify the implementation of the case study, the focus was placed on the model of a single agent in the hotel reservation process instead of modeling all customer and hotel agents separately. This agent was used to demonstrate the incremental changes of the method through room information, a reservation goal and actions such as message sending, interaction with the environment and logging.

The case study was carried out incrementally in the form of versions V0, V1, V2 and V3, all starting from the same initial model. V0 covers the initial implementation, V1 a platform change in the counterpart of an internal action and V2 a platform change in the counterpart of an external action, while V3 tests whether a new behavior can be represented with the existing MAS DSML concepts. Thus, V1 and V2 examine the basic partial evolution flow of the method, whereas V3 examines its boundary case in terms of representational adequacy.

Methods

This subsection presents the proposed evolution method, covering both its decision logic and its seven steps. It then demonstrates each step on DSML4BDI using the HotelBookingMAS case study, from identifying the external dependency through to validation on the Jason platform.

Proposed Evolution Method

The method presented in this paper takes Fall and Syriani's (2021) approach to DSL evolution driven by external dependencies as its starting point, but is specialized for the relationship between a MAS DSML and its execution platform. In Fall and Syriani's approach, reflecting changes or additions in the external library in the DSL may involve the generation of an extension metamodel, concrete syntax and mappings. In this study, however, the first decision concerns whether the change in the external dependency can be represented with the existing MAS DSML concepts. If the existing concepts are adequate and the only thing that changes is the implementation on the target platform, only the mapping is updated, in a separate mapping model called ExtSpec that was developed in this study, without extending the core metamodel or the graphical concrete syntax. ExtSpec is thus used not to create a new language layer, but to separate variable platform knowledge from the core DSML and from the fixed logic of the code generator.

The overall decision logic and the steps of the method are shown in the flow diagram in Figure 1. At the decision nodes, it is first assessed whether the change affects the domain/language concepts, and then whether the existing DSML concepts are adequate. If the existing concepts are adequate and the

change remains at the platform implementation level, the process continues with the creation or update of the ExtSpec mapping; otherwise, a more comprehensive DSML evolution is considered.

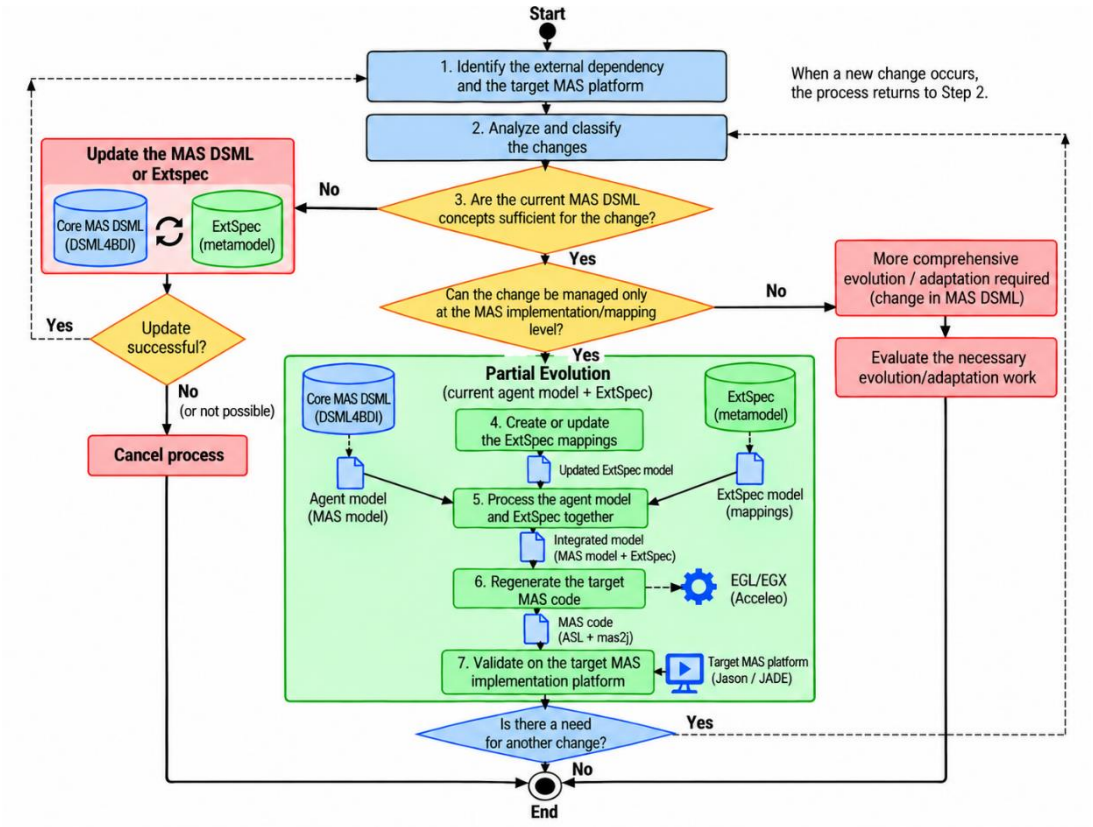


Figure 1. UML activity diagram of the proposed partial evolution method.

Step 1 - Identifying the external dependency and the target platform: The agent modeling language in use, the target execution platform, the external libraries and the related technology dependencies are identified. This step defines the boundaries within which the subsequent impact analysis will be carried out.

Step 2 - Classifying the impact of the change: It is assessed whether the change in the external dependency affects the domain/language concepts or only the form of implementation on the target platform. If the meaning of the domain concept does not change, the change is considered a candidate for partial platform evolution.

Step 3 - Assessing the adequacy of the DSML: It is checked whether the new or changed behavior can be expressed with the existing DSML concepts. If the existing concepts are inadequate, ExtSpec alone is not sufficient and evolution of the core DSML may be required.

Step 4 - Creating or updating the ExtSpec mappings: For a platform-level change, a mapping is defined, or the existing mapping is updated, using the libName, functor and mapsTo information in the current prototype. ExtSpec is not a new metamodel that replaces the core DSML; it is a supporting model that keeps changeable platform implementation information separate.

Step 5 - Processing the agent model together with ExtSpec: The code generator evaluates the action information in the agent model together with the mapping in ExtSpec. Thus, instead of the code generator having hard-coded knowledge that a given action is always transformed into a single platform operation, the target counterpart is read from ExtSpec.

Step 6 - Regenerating the target code: After the ExtSpec change, the target platform code is regenerated. Instead of writing the platform counterpart by hand into the generated file, the change is passed through the model-based generation chain.

Step 7 - Validation on the target platform: The generated code is executed on the target platform, and its syntax, parameter semantics and runtime requirements are observed. The platform-specific requirements identified during validation are used in assessing the suitability of the generation rules.

For a single platform change, it is technically possible to modify the fixed transformation rule inside the code generator. In that approach, however, platform-specific changeable information becomes embedded in the transformation template, and the generation rule has to be rewritten for every new platform/version counterpart. In the proposed method, the task of the code generator is to combine the behavior in the agent model with its current counterpart in ExtSpec, while the variable mapping information is kept in ExtSpec. The rationale for ExtSpec is therefore not to make a single change with fewer lines, but to define explicitly where platform changes are located and to limit their propagation to the core DSML and the transformation logic.

Applying the Method to DSML4BDI

In this subsection, the steps of the proposed method are applied to DSML4BDI in the order in which they are defined, using the HotelBookingMAS case study.

Step 1 – Identifying the external dependency and the target MAS implementation platform: In the first step, the MAS DSML in use, the target MAS implementation platform and the platform-dependent behaviors are identified. In this study, DSML4BDI was chosen as the MAS DSML and Jason as the target MAS implementation platform. The DSML4BDI language supports the graphical modeling of BDI components and the generation of code for Jason from models (Kardas et al., 2018). Jason, as described earlier, is a platform for implementing and running AgentSpeak-based BDI agents (Boissier et al., 2020).

To carry out the incremental application of the method, a model instance called HotelBookingMAS was created at this stage. The purpose of the case study is not to develop a fully functional hotel reservation application, but to examine which components of the same MAS model change in the face of implementation changes on the target platform.

The view of the created model instance in the DSML4BDI graphical modeling environment is shown in Figure 2. The model in Figure 2 contains the agent *agent1*, the belief *room(101)*, the goal *manageBooking* and the plan *plan1*. The plan body contains the internal actions (*Internal_Action*) *send_msg* and *log_event* and the external action (*External_Action*) *move*. Here, domain-level actions such as *send_msg*, *move* and *log_event* are represented directly in the model, whereas their implementation counterparts on the target platform, such as *.send*, *move_to*, *go* or *.println*, are not

included in the model. This separation shows why, in V1 and V2, the change in the platform implementation can be handled in ExtSpec instead of in the agent model.

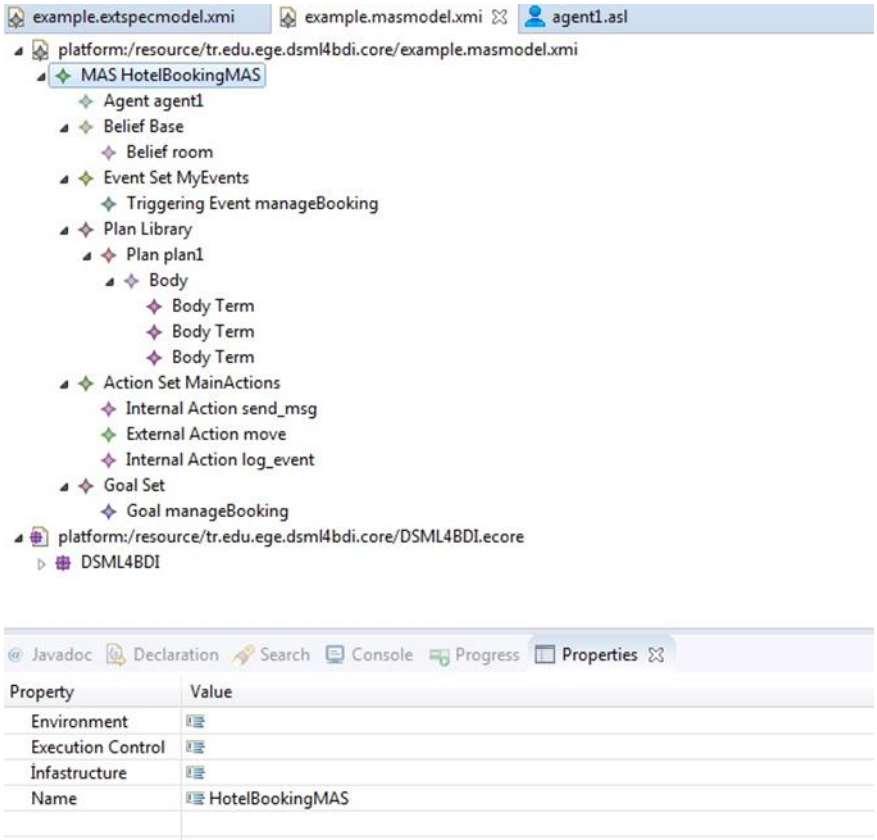


Figure 2. View of the example HotelBookingMAS DSML4BDI model in the Eclipse environment.

Step 2 – Classifying the impact of the change: In the second step, it is assessed whether the change that has arisen affects the domain concepts or only the form of implementation on the target MAS implementation platform.

In the first two changes examined in the HotelBookingMAS case study, the domain meanings of the *send_msg* and *move* behaviors are preserved. What changes is the call or action form through which these behaviors are realized on the Jason platform. These changes were therefore classified as platform implementation/mapping changes.

As a result of this classification, it becomes possible to manage the platform-specific mapping information through ExtSpec instead of changing the core DSML4BDI metamodel or the graphical concrete syntax.

Step 3 – Assessing the adequacy of the existing MAS DSML concepts: In the third step, it is assessed whether the change can be expressed with the existing MAS DSML concepts. This assessment is one of the main decision points that determine in which cases ExtSpec can be used.

In the HotelBookingMAS example, the *send_msg* behavior can be represented with the existing *Internal_Action* concept and the *move* behavior with the existing *External_Action* concept. Therefore,

no new superclass had to be created for these behaviors. In other words, the change does not alter the domain expressiveness of the existing MAS DSML; it only changes the implementation counterpart of the existing concepts on Jason.

In line with this result, the V1 and V2 changes were handled through ExtSpec. If the new behavior could not have been expressed with the existing MAS DSML concepts, ExtSpec alone would not have been sufficient, and an update of the core MAS DSML would have been considered according to the decision flow of the method shown in Figure 1.

Step 4 – Creating or updating the ExtSpec mappings: In the fourth step, the mapping related to the change in the platform implementation is created or updated in the ExtSpec model. In this study, ExtSpec is the supporting model, developed within this research, that keeps platform-specific changeable implementation information separate from the core MAS DSML.

The platform mapping of the *send_msg* action in the HotelBookingMAS example is shown in Figure 3.

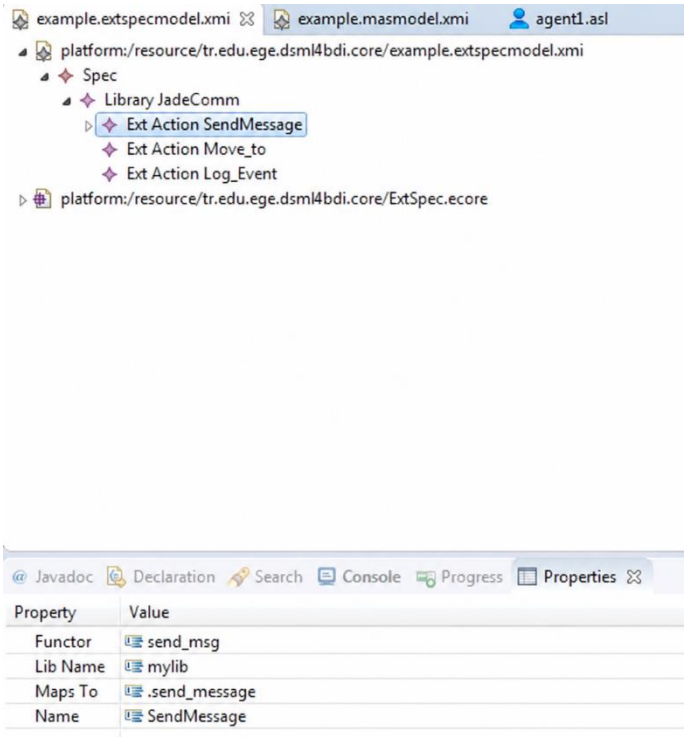


Figure 3. Definition of the platform mapping for send_msg in the ExtSpec model.

In Figure 3, the *libName*, *functor* and *mapsTo* information is used for the *send_msg* action. With this structure, the counterpart on the target platform of the *send_msg* behavior in the DSML4BDI model is determined through ExtSpec.

For example, in the initial state the mapping is as follows:

```
DSML4BDI:  library = mylib, functor = send_msg
ExtSpec:   libName = mylib, functor = send_msg, mapsTo = .send_message
Generated: .send_message("agent1", "tell", "test_message")
```

Here, send_msg represents the model-level behavior, while mapsTo = .send_message denotes its implementation counterpart on the target Jason platform in the initial state. Thus, platform-specific implementation information is maintained in ExtSpec rather than being added directly to the DSML4BDI model or the core metamodel. In the V1 change, only the mapsTo information in ExtSpec was updated, changing the counterpart from .send_message to .send. In the V2 change, the counterpart of the move external action was changed from move_to to go. In both cases, the core MAS DSML metamodel was preserved.

Step 5 – Processing the agent model together with ExtSpec: In the fifth step, the DSML4BDI model instance and the ExtSpec model are processed together. This associates the domain behaviors in the model with the current platform counterparts defined in ExtSpec.

In this study, the code generation chain based on Epsilon EGL/EGX (Rose et al., 2008) evaluates the action information in the DSML4BDI model instance together with the mappings in ExtSpec. Thus, the platform implementation information is obtained from ExtSpec instead of being kept as a fixed transformation rule inside the code generator. The main purpose of this stage is to bring together the unchanging domain model and the changeable platform mapping information in the same generation process.

Step 6 – Regenerating the target MAS code: In the sixth step, the target MAS code is regenerated using the updated ExtSpec information. In this study, the target output is ASL code to be used in the Jason environment.

In the V1 example, as a result of changing the platform counterpart for *send_msg* in ExtSpec, the regenerated code contains the call:

```
.send("agent1","tell","test_message")
```

In this way, the change in ExtSpec was reflected directly in the target MAS code.

Similarly, in V2 the target code was regenerated after the mapping for the *move* behavior in ExtSpec was changed from *move_to* to *go*. The main task of the code generator is therefore to combine the behavior in the agent model with its current counterpart in ExtSpec, rather than to contain the platform-specific variable implementation information itself.

Step 7 – Validation on the target MAS implementation platform: In the seventh step, the generated code is validated by executing it on the target MAS implementation platform, which in this study is the Jason environment. During validation, the syntax, parameter semantics and runtime requirements of the generated code are observed.

Executing the *.send* call generated in V1 in the Jason environment validated the message-sending behavior with respect to the target platform syntax. The *go* call generated in V2, on the other hand, could be produced syntactically, but when no suitable environment was configured, Jason raised the runtime error *no environment configured!*

This result shows that regenerating the target MAS code is not sufficient on its own; the runtime requirements of the target platform must also be validated separately. For this reason, code generation and validation on the target MAS implementation platform are treated as separate steps in the method.

RESULTS

This section presents the results of the HotelBookingMAS case study. The following subsections describe, in order, the initial state of the HotelBookingMAS case study (V0), the change in the internal action implementation (V0→V1), the change in the external action implementation (V1→V2) and the test examining the representational adequacy of the existing MAS DSML concepts (V2→V3); the incremental application is then analyzed as a whole.

V0 - Initial State

V0 constitutes the baseline against which the subsequent changes are compared. In this version, the *send_msg*, *move* and *log_event* behaviors in the HotelBookingMAS model are associated with the counterparts *.send_message*, *move_to* and *.println* on the Jason side, respectively.

In the initial version, the plan body that is executed for the *manageBooking* goal appears in the generated Jason code as follows:

```
+!manageBooking <-
    .send_message("agent1", "tell", "test_message");
    move_to("X", "Y");
    .println("log_event").
```

This code shows the counterparts of the three behaviors in the model on the target MAS implementation platform in version V0. Here, *send_msg*, *move* and *log_event* represent the model-level behaviors, while *.send_message*, *move_to* and *.println* represent their implementations on the Jason side.

Since V0 is the initial state, no platform implementation change was applied at this stage. The changes to be made in V1 and V2 are compared with the mappings in this version.

V0 → V1 - Internal Action (Internal_Action) Implementation Change

In V1, it is assumed that the counterpart of the *send_msg* behavior in the model on the target MAS implementation platform has changed from *.send_message* to *.send*. This change does not alter the domain meaning of the message-sending behavior; it only changes the way this behavior is implemented on Jason.

Therefore, the *send_msg* behavior in the DSML4BDI model, the related plan and the *manageBooking* goal were preserved. The change was applied only to the relevant *mapsTo* mapping in ExtSpec. In this way, the new target code was generated without making any change to the core MAS DSML metamodel or the graphical modeling environment.

The message-sending call in the Jason code generated as a result of V1 is shown in Figure 4.

example.extspecmodel.xmi | example.masmodel.xmi | agent1.asl

```
40
41
42 ////////////////////////////////////////////////////
43 // Agent agent1
44 ////////////////////////////////////////////////////
45
46 // Beliefs
47 room(101).
48
49
50 // Goals
51 !start.
52 !manageBooking.
53
54
55 // Plans
56 +!start <-
57     .println("Generated code is running on pure Jason!").
58
59 // Plan: plan1
60 +!manageBooking <-
61     .send(agent1, tell, test_message);    move_to("X", "Y");    .println("log_event").
62
63
```

Figure 4. Regenerated Jason code containing the .send call as a result of V1.

As shown in Figure 4, the *send_msg* behavior in the model was generated, using the current mapping in ExtSpec, as:

```
.send("agent1","tell","test_message")
```

Here, in accordance with the parameter semantics of Jason's *.send* call, the agent name, the communication type and the communication content were generated as separate parameters. Thus, in V1 the point of change is not the core MAS DSML or the agent model, but the platform implementation mapping in ExtSpec. This directly tests the method's aim of managing platform implementation changes while preserving the existing MAS DSML concepts.

V1 → V2 - External Action (External_Action) Implementation Change

In V2, the change in the platform implementation was examined on the external action side. At this stage, the *move* external action in the DSML4BDI model was preserved, but its implementation counterpart on the Jason side was changed from *move_to* to *go*.

In this change, too, the domain meaning of the *move* behavior did not change. The existing External_Action concept was therefore sufficient to express the change, and no new superclass had to be created in the core MAS DSML metamodel.

Figure 5 shows the external action rendered with its *go* counterpart in the target MAS code regenerated as a result of V2.

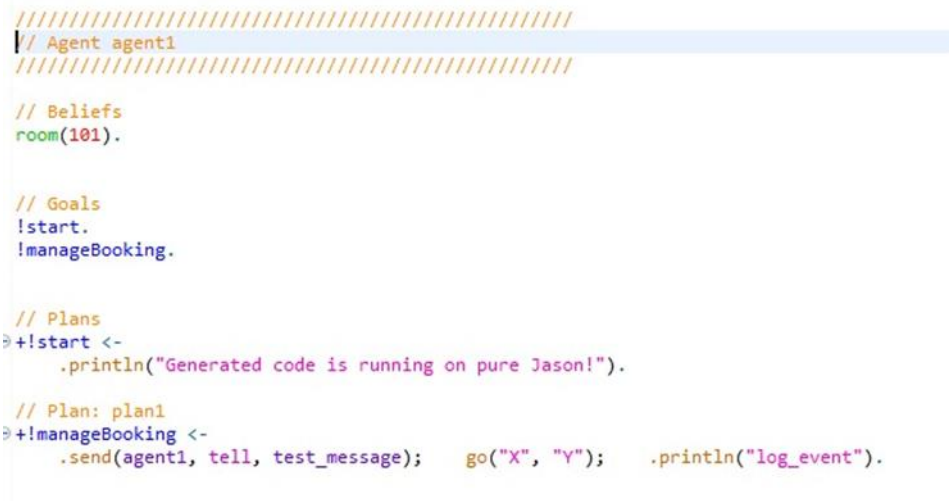


Figure 5. Generation of the external action counterpart as go as a result of V2.

As in V1, the change seen in Figure 5 occurred only in the platform implementation mapping. While the *move* model behavior was preserved, the target code was regenerated as a result of updating the mapping in ExtSpec.

At this stage, the runtime conditions of the target MAS implementation platform were also taken into account. Generating the *go* call as code does not mean that the environment required to execute the corresponding action is provided automatically. V2 therefore also shows that, for platform implementation changes, code generation and runtime validation need to be assessed separately.

V2 → V3 - Representational Adequacy Test

Unlike the direct platform implementation mapping changes in V1 and V2, V3 was designed to examine the adequacy of the existing MAS DSML concepts within the proposed method. At this stage, a new behavior called *call_api* was added to the model, and it was assessed whether this behavior could be expressed using the existing DSML4BDI concepts.

The point that deserves particular attention here is that the *call_api* behavior was modeled as an Internal_Action in this case study. The purpose is not to implement a real HTTP/REST service. The purpose is to test whether a new platform function can be represented with an existing MAS DSML action concept.

For this reason, no new superclass or dedicated *APIAction/RESTAction* concept was added for *call_api*. Assuming the existing Internal_Action concept to be adequate, the behavior was represented with the following model information:

```
Internal_Action:
  library = mylib
  functor = call_api
  arguments = "GET", "/availability"
```

As a mock implementation, the counterpart of this behavior in ExtSpec was defined as:

```
ExtSpec:
    call_api -> .println
```

Here, *.println* is not a real HTTP/REST client. It is a mock implementation used to observe that the *call_api* behavior passes through the chain MAS model → ExtSpec → code generation → target MAS implementation platform. It should therefore not be concluded from V3 that a real REST API integration was implemented.

For the same purpose, the parameters used in V3 were also transferred to the target code in mock form:

```
.println("GET", "/availability")
```

This application shows that the existing Internal_Action concept is adequate for representing the new behavior. For a real external API call, however, additional information such as parameter mappings, the call signature, the connection to the external service and the runtime context may need to be modeled. V3 was therefore used to examine representational adequacy and the limits of ExtSpec, rather than as a basic platform mapping scenario of the method.

Analysis of the Incremental Application

The changes made from V0 to V3 are summarized in Table 2 in order to show together which components were changed and which were preserved in each version. In particular, the table shows the difference between the platform implementation changes in V1 and V2 and the addition of a model-level behavior in V3.

Table 2. V0-V3 incremental application of the HotelBookingMAS case study.

Version	Change	Metamodel	MAS model instance	ExtSpec	Jason code
V0	Initial	Unchanged	Initial	Initial	Initial
V1	.send_message → .send	Unchanged	Unchanged	Changed	Changed
V2	move_to → go	Unchanged	Unchanged	Changed	Changed
V3	call_api added	Unchanged	Changed	Changed	Changed

As Table 2 shows, in V1 and V2 the change was confined to the level of the platform implementation mapping. In these two versions, the core MAS DSML metamodel and the existing MAS model instance were preserved, while ExtSpec was changed and the target Jason code was regenerated.

In V3, unlike the previous two versions, a new behavior was added to the MAS model instance. However, since the existing Internal_Action concept was adequate, the core MAS DSML metamodel did not need to be changed. V3 thus demonstrates the importance of assessing the adequacy of the existing MAS DSML concepts in the third step of the method.

The case study demonstrates the two main ways of using the method together: in V1 and V2, changing the platform implementation mapping through ExtSpec while preserving the existing MAS domain concepts; and in V3, testing that a new behavior can be represented with the existing MAS DSML concepts. This distinction shows that the method positions ExtSpec not as a construct that replaces the core MAS DSML, but as a supporting model that localizes certain platform implementation changes.

DISCUSSION AND CONCLUSION

This section discusses the findings of the HotelBookingMAS case study in relation to the research questions and draws the study's conclusions. It first evaluates RQ1–RQ3, then sets out further findings, the technical limits of ExtSpec, the threats to the validity of the results, and finally summarizes the contributions and directions for future work.

Evaluation of the Research Questions

In this subsection, the findings obtained from the HotelBookingMAS case study are evaluated with respect to the research questions RQ1, RQ2 and RQ3 defined in the Introduction. V1 and V2 represent changing the platform implementation mappings while preserving the existing MAS DSML concepts, whereas V3 represents the boundary case examining whether a new behavior can be represented with the existing MAS DSML concepts.

RQ1 - Can Platform Implementation Changes Be Managed Without Changing the Core MAS DSML?

With respect to RQ1, the V1 and V2 platform implementation changes examined could be managed without changing the core MAS DSML.

In V1, the counterpart of the *send_msg* behavior on the target MAS implementation platform was changed from *.send_message* to *.send*. In V2, the counterpart *move_to* of the external action *move* was changed to *go*. In both changes, the DSML4BDI core metamodel was preserved and the domain-level meanings of the related behaviors in the existing MAS model instance were not altered. The changes were made through the platform mappings in the ExtSpec model and transferred to the target Jason code by code regeneration.

This finding shows that platform-specific changeable information can be separated from the core MAS DSML in cases where the existing MAS DSML concepts are adequate and the change affects only the form of platform implementation. In particular, localizing the change in ExtSpec in V1 and V2, in the form *send_msg* → *.send* and *move* → *go*, did not require changing the core metamodel.

Nevertheless, this result should not be generalized to all MAS DSML evolution cases. When the existing MAS DSML concepts are inadequate for expressing the new or changed behavior, the ExtSpec model alone is not sufficient. In such a case, the method proceeds towards updating the core MAS DSML. The finding for RQ1 should therefore be considered within the scope of certain platform implementation changes.

RQ2 - Can ExtSpec Limit the Propagation of the Change to the Metamodel and the Graphical Concrete Syntax?

With respect to RQ2, it was observed that in V1 and V2 the platform implementation change did not propagate to the DSML4BDI core metamodel or the Sirius graphical concrete syntax.

Since no new domain concept emerged in V1 and V2, the core metamodel of DSML4BDI was not changed. Consequently, the graphical modeling definitions in Sirius were not changed either. The components that changed are the platform mapping in the ExtSpec model and, as a result, the regenerated Jason code. This shows that the MAS DSML and the target MAS implementation platform have different points of change. While the DSML4BDI model continued to express domain-level behaviors such as

send_msg and *move*, the forms in which these behaviors are implemented on the target platform could be changed through the ExtSpec model.

V3, in turn, shows the limit of this finding. In V3, a new behavior called *call_api* was added to the MAS model. Even so, the core MAS DSML metamodel and the Sirius definition were not changed, because the existing Internal_Action concept was adequate for expressing this behavior. The aim of the method is therefore not that the MAS model should never be changed. The aim is to limit the unnecessary propagation of platform-specific changes to the core language definition when the existing MAS DSML concepts are adequate. In this respect, V1 and V2 provide direct evidence for RQ2, while V3 provides a complementary finding that illustrates the boundary condition.

RQ3 - Can ExtSpec Changes Be Transferred to the Target Code and Validated on the Platform?

With respect to RQ3, the V1 and V2 changes made in the ExtSpec model were transferred to the target ASL code by code regeneration and examined through executions in the Jason environment.

In V1, as a result of changing the platform mapping for *send_msg* in the ExtSpec model from *.send_message* to *.send*, the following call was generated in the target code:

```
.send("agent1","tell","test_message")
```

In V2, the counterpart *move_to* of the *move* behavior was changed to *go*, and this change was transferred to the regenerated target code.

This result shows that the platform implementation information in the ExtSpec model is actually used in the code generation process. In other words, the ExtSpec model did not remain a passive model that merely stores change information; it was processed together with the DSML4BDI model instance and contributed to the regeneration of the target MAS code.

However, the executions in the Jason environment showed that code generation alone is not sufficient for validation on the target platform. In particular, for the *move* external action, although *move_to("X","Y")* or the corresponding implementation call could be generated, the runtime error *"no environment configured!"* was observed when no suitable environment was configured. This shows that the runtime context of the target MAS implementation platform needs to be assessed separately.

The finding for RQ3 is therefore twofold: the ExtSpec model changes could be transferred to the target code; however, it was seen that ensuring executability on the target MAS implementation platform requires validating not only the syntax of the platform but also its parameter semantics and runtime context.

Additional Findings and Technical Limits of ExtSpec

When the V1–V3 applications and the validations in the Jason environment are considered together, it is seen that the proposed method does not handle all types of change to the same extent. The most directly supported class of change comprises changes of the V1 and V2 type, in which the implementation/mapping information on the target MAS implementation platform changes while the existing MAS domain concepts are preserved.

V3, in turn, shows why the decision mechanism of the method is necessary. Although the *call_api* behavior is new, no new superclass was needed in the core MAS DSML metamodel, since it could be

represented with the existing Internal_Action concept. However, the *.println* mapping used is only a mock implementation; it does not represent the parameter, connection and runtime requirements of a real API integration.

In addition, the executions in the Jason environment showed that platform implementation changes may involve more than changing the name of an action. Elements such as the parameter type, parameter order, call signature and execution environment may also be platform-dependent. The fact that the current ExtSpec structure represents this information only to a limited extent indicates that ExtSpec may need to be extended for more complex platform changes.

These findings and the applicability limits observed in the case study are summarized in Table 3.

Table 3. Types of change observed in the case study and the applicability limits of the proposed method.

Change example	**Manageable with ExtSpec**	**Handling in the method / observed limit**
.send_message → .send	Yes	The ExtSpec mapping is updated; the core MAS DSML and the MAS model instance are preserved.
move_to → go	Yes	The ExtSpec mapping is updated; the change is reflected in the regenerated target code.
Use of call_api	Partially	Since the existing Internal_Action concept is adequate, the behavior can be added to the MAS model instance; the .println mapping used is only a mock implementation.
Need for a new fundamental domain concept	No	If the existing MAS DSML concepts are inadequate, ExtSpec alone is not sufficient; evolution of the core MAS DSML is considered.
Change in parameter semantics / call signature	Partially	The current ExtSpec is limited; the parameter type, order and mapping information may need to be represented more richly.
Change in execution environment / execution context	Partially	The current ExtSpec is limited; the runtime context of the target MAS implementation platform must be validated separately.

Table 3 shows that the method is applicable in particular where the existing MAS domain concepts can be preserved and the change remains at the level of the mapping on the target MAS implementation platform. V1 and V2 are direct examples of this situation. V3, in turn, shows that when a new behavior can be expressed with the existing MAS DSML concepts, the MAS model instance may change, but the core metamodel can still be preserved.

By contrast, when a new fundamental domain concept is required, the ExtSpec model alone is not sufficient. Similarly, more extensive changes in parameter semantics or the runtime context reveal the limits of the current structure of the ExtSpec model. The results of the study should therefore be interpreted not as showing that the method offers a general solution for all types of MAS DSML

evolution, but as showing that it offers an applicable partial evolution path for certain platform implementation changes.

Threats to Validity

The findings of this study were obtained from a limited number of cases and a specific pair of MAS DSML and target implementation platform. The evaluation of the results is therefore subject to some threats with respect to conclusion, internal, construct and external validity (Wohlin et al., 2012).

With respect to conclusion validity, the evaluation was carried out on a limited number of incremental changes. The findings should therefore not be regarded as a statistical generalization. To mitigate this threat, the same MAS model instance was preserved in V1 and V2, only the platform implementation mappings were changed in a controlled manner, and the results were observed both in the regenerated target code and in the executions in the Jason environment. Larger sets of changes and quantitative measurements could strengthen this evaluation.

Considering the internal validity, the fact that the DSML4BDI model, the ExtSpec model, the code generation chain and the experiments were all developed within the same study means that design decisions may have influenced the results. To mitigate this threat, the initial MAS model instance was kept fixed in the V1 and V2 experiments, and the change was made only to the target mapping in the ExtSpec model. In addition, the generated Jason code was executed in a separate Jason environment, providing an independent validation point after the generation process.

Representing platform implementation changes with ExtSpec model attributes such as *libName*, *functor* and *mapsTo* does not cover all dimensions of platform changes which may be a threat to the construct validity. Indeed, during V3 and the Jason executions it was seen that additional elements such as parameter semantics, the call signature and the execution environment are also important. The present evaluation should therefore be regarded as a validation of the method for certain platform implementation changes; for richer platform changes, the representation structure used needs to be assessed separately.

Finally, the study was carried out on DSML4BDI, Jason and the small-scale HotelBookingMAS case study which can be considered as a threat to the external validity. The results therefore cannot be directly generalized to other MAS DSMLs, different target MAS implementation platforms or larger MASs. Applying the method to different MAS DSMLs and different target implementation platforms would strengthen the evidence for external validity.

Conclusion

In this study, a partial evolution method was proposed for managing the impact on an agent modeling language of certain changes occurring in target MAS implementation platforms and external technology dependencies. The method comprises the following steps: identifying the external dependency and the target platform; classifying the change with respect to the domain/language level and the platform implementation level; assessing the adequacy of the existing MAS DSML concepts; creating or updating the platform implementation mappings in the ExtSpec model; regenerating the target code by processing the agent model together with the ExtSpec model; and validating the generated code on the target MAS implementation platform. ExtSpec, developed for this purpose as a supporting model specific to this

study, makes it possible to separate changeable platform implementation information from the core MAS DSML.

In the HotelBookingMAS case study carried out with DSML4BDI and Jason, the platform counterparts of the *send_msg* and *move* behaviors were changed in steps V1 and V2, respectively, and these changes were managed through ExtSpec mappings without changing the core DSML4BDI metamodel. The reflection of these changes in the regenerated Jason code showed that the method can limit the propagation of change to the core MAS DSML for certain platform implementation changes. In V3, the boundary case in which the *call_api* behavior could be represented with the existing *Internal_Action* concept was examined; in this example, no new superclass had to be created. At the same time, it was seen that more complex platform characteristics such as parameter semantics, the call signature and the execution context can be handled only to a limited extent with the current representation of the ExtSpec model.

The results indicate that the proposed method should be regarded not as a general solution covering all MAS DSML evolution problems, but as a partial evolution path for certain platform implementation changes where the existing MAS DSML concepts are adequate. It thus becomes possible to manage changeable mapping information through ExtSpec instead of reflecting platform-specific implementation information directly in the core MAS DSML and the graphical concrete syntax.

In future work, we plan to extend the ExtSpec model so that it can represent richer platform semantics such as parameter mapping, the call signature and the execution context, and to apply the method to different MAS DSMLs and different implementation platforms. We also aim to compare the ExtSpec-based approach quantitatively with direct code generator modifications, using criteria such as the number of affected components, the number of changed lines of code and the development effort.

References

Alaca OF, Tezel BT, Challenger M, Goulão M, Amaral V, Kardas G, 2021. AgentDSM-Eval: A framework for the evaluation of domain-specific modeling languages for multi-agent systems. Computer Standards & Interfaces, 76: 103513. https://doi.org/10.1016/j.csi.2021.103513

Bettini L, Di Ruscio D, Iovino L, Pierantonio A, 2022. Supporting safe metamodel evolution with Edelta. International Journal on Software Tools for Technology Transfer, 24: 247–260. https://doi.org/10.1007/s10009-022-00646-2

Boissier O, Bordini RH, Hübner J, Ricci A, 2020. Multi-agent oriented programming: Programming multi-agent systems using JaCaMo. MIT Press.

Bordini RH, Hübner JF, Wooldridge M, 2007. Programming multi-agent systems in AgentSpeak using Jason. Wiley. https://doi.org/10.1002/9780470061848

Brambilla M, Cabot J, Wimmer M, 2017. Model-driven software engineering in practice. 2nd ed., Springer. https://doi.org/10.1007/978-3-031-02549-5

Çelik B, Kardas G, Tezel BT, 2023. An online modeling language for the low-code development of Belief-Desire-Intention (BDI) agents. Proceedings of the 2nd International Symposium Series on Graduate Researches (ISGR 2023), pp. 186–196.

Challenger M, Kardas G, Tekinerdogan B, 2014. On the use of a domain-specific modeling language in the development of multiagent systems. Engineering Applications of Artificial Intelligence, 28: 111–141. https://doi.org/10.1016/j.engappai.2013.11.012

de Silva L, Meneguzzi F, Logan B, 2020. BDI agent architectures: A survey. (Editor: C. Bessiere), Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI

2020), International Joint Conferences on Artificial Intelligence Organization, pp. 4914–4921. https://doi.org/10.24963/ijcai.2020/684

Eclipse Foundation, 2013. Sirius: The easiest way to get your own modeling tool. Access address: https://eclipse.dev/sirius/; Date of access: 02.10.2026.

Fall K, Syriani E, 2021. Evolving domain-specific languages depending on external libraries. Proceedings of the 36th ACM/SIGAPP Symposium on Applied Computing (SAC '21), pp. 1247–1255. https://doi.org/10.1145/3412841.3441999

Hahn C, 2008. A domain specific modeling language for multiagent systems. Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2008), Vol. 1, International Foundation for Autonomous Agents and Multiagent Systems, pp. 233–240. https://doi.org/10.1145/1402383.1402420

Hebig R, Khelladi DE, Bendraou R, 2017. Approaches to co-evolution of metamodels and models: A survey. IEEE Transactions on Software Engineering, 43 (5): 396–414. https://doi.org/10.1109/TSE.2016.2610424

Kardas G, Ciccozzi F, Iovino L, 2023. Introduction to the special issue on methods, tools and languages for model-driven engineering and low-code development. Journal of Computer Languages, 74: 101190. https://doi.org/10.1016/j.cola.2022.101190

Kardas G, Tezel BT, Challenger M, 2018. Domain-specific modelling language for belief–desire–intention software agents. IET Software, 12 (4): 356–364. https://doi.org/10.1049/iet-sen.2017.0094

Rose LM, Kolovos DS, Paige RF, Polack FAC, 2010. Model migration with Epsilon Flock. Theory and Practice of Model Transformations (ICMT 2010), LNCS 6142, Springer, pp. 184–198. https://doi.org/10.1007/978-3-642-13688-7_13

Rose LM, Paige RF, Kolovos DS, Polack FAC, 2008. The Epsilon Generation Language. Model Driven Architecture – Foundations and Applications, LNCS 5095, Springer, pp. 1–16. https://doi.org/10.1007/978-3-540-69100-6_1

Shehory O, Sturm A (Editors), 2014. Agent-oriented software engineering: Reflections on architectures, methodologies, languages, and frameworks. Springer. https://doi.org/10.1007/978-3-642-54432-3

Steinberg D, Budinsky F, Paternostro M, Merks E, 2008. EMF: Eclipse Modeling Framework. 2nd ed., Addison-Wesley Professional.

Tolvanen JP, Kelly S, Di Rocco J, Pierantonio A, Tinella G, 2025. A framework for evaluating tool support for co-evolution of modeling languages, tools and models. Software and Systems Modeling, 24: 311–338. https://doi.org/10.1007/s10270-024-01218-5

Warwas S, Hahn C, 2009. The DSML4MAS development environment. Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009), Vol. 2, International Foundation for Autonomous Agents and Multiagent Systems, pp. 1379–1380.

Wąsowski A, Berger T, 2023. Domain-specific languages: Effective modeling, automation, and reuse. Springer. https://doi.org/10.1007/978-3-031-23669-3

Wohlin C, Runeson P, Höst M, Ohlsson MC, Regnell B, Wesslén A, 2012. Experimentation in software engineering. Springer Science & Business Media.

Wooldridge M, 2009. An introduction to multiagent systems. 2nd ed., Wiley.

Zhang W, Holtmann J, Strüber D, Hebig R, Steghöfer JP, 2024. Supporting meta-model-based language evolution and rapid prototyping with automated grammar transformation. Journal of Systems and Software, 214: 112069. https://doi.org/10.1016/j.jss.2024.112069

Zhang W, Strüber D, 2024. Tales from 1002 repositories: Development and evolution of Xtext-based DSLs on GitHub. 50th Euromicro Conference on Software Engineering and Advanced Applications (SEAA 2024), pp. 172–179. https://doi.org/10.1109/SEAA64295.2024.00034

Zschaler S, Polack FAC, 2023. Trustworthy agent-based simulation: The case for domain-specific modelling languages. Software and Systems Modeling, 22: 455–470. https://doi.org/10.1007/s10270-023-01082-9

Conflict of Interest

The authors have declared that there is no conflict of interest.

Hidden Biases: Textbook Wording and Data Contamination in Behavioral Economics Experiments with Large Language Models

Cemal Öztürk[1*], Serkan Künü[1]

[1]Iğdır University, Faculty of Economics and Administrative Sciences, Department of Economics, 76000, Iğdır, Türkiye

*Corresponding author e-mail: cemal.ozturk@igdir.edu.tr

Abstract

Economists have started to use large language models (LLMs) as stand-ins for human subjects, but most evidence on their biases comes from textbook experiments that the models have probably read during training. This study tests four open-weight models (GPT-OSS-120B, GPT-OSS-20B, Qwen3.8-27B and ALLaM-2-7B) on six classic paradigms: risky-choice framing, mental accounting, the conjunction fallacy, anchoring, mixed-gamble acceptance and the ultimatum game. Each paradigm is presented once in its published wording and once as a disguised version with the same decision structure and a new story, with 20 repetitions per cell and counterbalanced option order. Bias indices are reported with stratified bootstrap confidence intervals and compared with published human values, and pooled logistic regressions and log-OLS models with model fixed effects test the effect of wording. In the published anchoring question the four models barely respond to the anchor (indices of 0.00 to 0.06), while the disguised question produces anchoring indices between 0.11 and 0.97 (humans: 0.36). GPT-OSS-120B never commits the conjunction fallacy on the original Linda problem, yet does so in 85% of trials once Linda becomes a different person. GPT-OSS-20B shows the same pattern (0.10 to 0.85), whereas Qwen does not. ALLaM rejects unfair ultimatum offers in 22% of trials with the original wording and in 78% with the new one. Framing effects are small and none survives multiple-testing correction for an individual model. In mixed gambles, the GPT-OSS and Qwen models accept once the gain reaches 0.88 to 1.11 times the loss, whereas people typically ask for about twice the loss, and ALLaM accepts almost every gamble. Because the new stories also change the context of each task, the gap cannot be attributed to contamination alone, but it shows that textbook items can hide much of how these models decide.

Keywords: *Large Language Models, Behavioral Economics, Cognitive Biases, Data Contamination, Prompt Sensitivity*

INTRODUCTION

Large language models (LLMs) are now used as simulated participants in economics and psychology. Horton et al. (2023) suggested treating them as computational models of economic agents that can be

placed in classic experiments. Dillion et al. (2023) discussed when such models might support research with human participants. Mei et al. (2024) had chatbots play a set of standard behavioral games. ChatGPT-4 behaved in a way that was statistically indistinguishable from a randomly chosen human, but the chatbots often departed from average human behavior, and when they did, they were more altruistic and cooperative than most people. These studies assume that a model's answer to an experimental question shows how the model decides rather than what it remembers.

This assumption is hardest to defend for the best-known experiments in behavioral economics. The Asian disease problem, the Linda problem and similar questions appear in textbooks and on many web pages, so current models have very likely seen them during training. Binz and Schulz (2023) found that GPT-3 answered 12 standard questions either correctly or with human-like errors, but small changes to the same questions led to clear failures. They noted that the model may have met the original scenarios in its training data. Hagendorff et al. (2023) reported that earlier GPT models made human-like reasoning errors that later ChatGPT models avoided, and they did not rule out that some well-known tasks were solved from memory. In natural language processing this problem is called data contamination: test items that leak into the training data make measured performance look better than it is (Sainz et al., 2023).

In this study we ask whether the published wording of classic experiments hides or distorts the behavioral tendencies of current open-weight LLMs, that is, models whose parameters are publicly released. Four models answered six experiments in two forms: the published wording and a disguised version with the same decision structure but a new story. If a model gives the normative answer to the published question and a biased answer to the disguised one, conclusions based only on textbook questions would be misleading. The data were collected in English and Turkish. This paper reports the English part; the Turkish responses will be analyzed in a separate study.

MATERIAL AND METHODS

Material

We tested four open-weight models through the Groq API: GPT-OSS-120B and GPT-OSS-20B (OpenAI, 2025), Qwen3.8-27B from Alibaba's Qwen series (the Qwen3 generation is described by Yang et al., 2025) and ALLaM-2-7B, a model trained for Arabic and English (Bari et al., 2024). All calls used a temperature of 0.7 and a limit of 2,048 output tokens. The GPT-OSS models ran with low reasoning effort, and Qwen's reasoning steps were hidden, so only the final answer was returned. A short system prompt told the model that it was taking part in a decision study with no right or wrong answers and asked it to decide as it would.

Table 1 lists the six experiments and the human reference values. The published questions follow the original wording closely. The disguised questions keep the choice structure of each experiment but change the story. Two further details differ. The mental accounting question uses a 40 € football ticket instead of a 10 € theater ticket, and the anchoring question keeps the anchors of 10% and 65% but asks about a less familiar quantity that a model is unlikely to have memorized. All amounts are in euros.

Table 1. Experiments, published and disguised questions, and human reference values

Experiment	Published question	Disguised question	Conditions	Human reference
Framing	Asian disease problem: 600 lives, sure vs risky program, gain or loss frame	600 jobs in a manufacturing cluster hit by a supply shock	Gain, loss	Risky choice 28% (gain) vs 78% (loss); index 0.50 (Tversky and Kahneman, 1981)
Mental accounting	Theater ticket (10 €): cash lost vs ticket lost	Football ticket (40 €): cash lost vs ticket lost	Cash lost, ticket lost	Buy 88% vs 46%; index 0.42 (Tversky and Kahneman, 1981)
Conjunction fallacy	Linda: bank teller vs bank teller and feminist	Deniz, environmental activist: insurance agent vs insurance agent and member of an environmental organization	Single	Conjunction chosen by 85% (Tversky and Kahneman, 1983)
Anchoring	Share of African countries in the United Nations	Share of agricultural land in Iğdır province planted with apricot trees	Anchor 10% or 65%	Medians 25 and 45; index 0.36 (Tversky and Kahneman, 1974)
Mixed gamble	Fair coin: win G € or lose 100 €	Investment: gain G € or lose 100 € with equal probability	G = 50, 75, 100, 125, 150, 200, 250, 300, 400	Median loss aversion coefficient 1.93, external reference (Tom et al., 2007)
Ultimatum game (responder)	Anonymous stranger splits 100 €	Colleague splits a one-time bonus of 100 €	Offer 5, 10, 20, 30, 40, 50 €	Mean rejection 16% over all offers (Oosterbeek et al., 2004)

All amounts are in euros. Each cell was repeated 20 times per model with counterbalanced option order.

In the mixed gamble the possible loss was fixed at 100 € and the possible gain took nine values. In the ultimatum game (Güth et al., 1982) the model played the responder and saw one offer out of 100 €. Both tasks therefore give an acceptance curve across stake levels (Figure 2).

Methods

The Collection of the Data

The English design has 22 cells for each wording, 44 cells in total. Each cell was sent 20 times to each model, which gives 3,520 responses. The data were collected on 18 and 19 September 2026. In questions with two options, the options were labeled A and B. LLMs are sensitive to the position of answer options (Pezeshkpour and Hruschka, 2024), so the order was reversed in half of the repetitions. Models answered in a fixed JSON format. For the anchoring question they gave a higher or lower judgment and a numerical estimate. Seventy-four requests failed (68 because of rate limits, five authorization errors and one connection error). They were repeated until every cell had 20 responses, and all 3,520 responses could be read.

Statistical Analysis

We summarize each experiment with one bias index. Five of the indices are zero for a decision maker without the bias; the mixed-gamble ratio described below has a reference value of 1. The framing index is $P(\text{risky} \mid \text{loss}) - P(\text{risky} \mid \text{gain})$, the difference in the share of risky choices between the loss and gain frames. The mental accounting index is $P(\text{buy} \mid \text{cash lost}) - P(\text{buy} \mid \text{ticket lost})$. The conjunction index is the share of trials in which the model judged the conjunction more probable than its single part. Anchoring was measured with the anchoring index of Jacowitz and Kahneman (1995):

$$AI = \frac{Md_H - Md_L}{A_H - A_L} \tag{1}$$

Here Md_H and Md_L are the median estimates after the high and low anchors, and $A_H = 65$ and $A_L = 10$ are the anchors. An index of 0 means that the anchor has no effect; an index of 1 means that the estimates move as far apart as the anchors. The median estimates of 25 and 45 reported by Tversky and Kahneman (1974) give a human value of 0.36. In the mixed gamble, the acceptance-threshold ratio r is the gain G^* at which the acceptance rate reaches 50%, divided by the loss $L = 100$ €. Because sampling noise can make observed acceptance fall slightly as the gain rises, the rates were first made non-decreasing in the gain (running maximum), and G^* was then found by linear interpolation between the gain values. A risk-neutral decision maker has $r = 1$, and loss aversion (Kahneman and Tversky, 1979) implies $r > 1$. A ratio above 1 shows that the gain must exceed the loss before the gamble is accepted; it can also result from risk aversion or probability weighting, so it does not by itself identify loss aversion. For 50/50 gambles, Tom et al. (2007) reported a median loss aversion coefficient of 1.93. They varied gains and losses independently and took the ratio of the absolute value of the loss coefficient to the gain coefficient of a logistic model, so their value is an external reference rather than the same measurement. The ratio is not the same as the prospect theory parameter $\lambda = 2.25$ of Tversky and Kahneman (1992). As a sensitivity check, λ_{CPT} in Equation 2 converts r into the value implied by their value function (exponent 0.88) and probability weights $w^+(0.5) = 0.421$ and $w^-(0.5) = 0.454$. Ratios at the edge of the gain values ($r = 0.5$ or $r = 4$) are treated as censored. The ultimatum index is the rejection rate for offers of 20 € or less. A responder who cares only about money rejects none of them. There is no single human benchmark for this index; across studies, the mean rejection rate over all offers is 16% (Oosterbeek et al., 2004).

$$r = \frac{G^*}{L}, \qquad \lambda_{CPT} = \frac{w^+(0.5)}{w^-(0.5)} r^{0.88} \tag{2}$$

Confidence intervals come from a stratified bootstrap with $B = 2{,}000$ resamples, drawn separately within each condition or stake level (Efron and Tibshirani, 1993). They are not adjusted for multiple comparisons. With 20 answers per cell some intervals have zero width, for instance when the resampled medians do not change; such intervals reflect the small sample and the bootstrap distribution, not certainty. For each model and wording, framing and mental accounting were tested with chi-square or Fisher exact tests, the conjunction rate with an exact binomial test against the human value of 0.85 (treated as fixed), and anchoring with the Mann–Whitney U test. The Holm (1979) correction was applied within each experiment across its eight model and wording tests. For each model, the difference between the two wordings was tested with the bootstrap distribution of that difference, using the approximate two-sided $p = \min[1, (2k + 1)/(B + 1)]$, where k is the smaller number of bootstrap differences on either side of zero. These 24 tests (six experiments, four models) form a second Holm family. Option position was checked with a binomial test of the share of A choices for each model and experiment (20 tests, Holm-corrected). Effects pooled across models were estimated with logistic regressions that include model fixed effects and heteroskedasticity-consistent (HC1) standard errors:

$$\text{logit } \Pr(y_i = 1) = \beta_0 + \beta_1 C_i + \beta_2 D_i + \beta_3 C_i D_i + \alpha_{m(i)} \tag{3}$$

Here $y_i = 1$ if the model chose the target option (the risky program, buying the ticket, the conjunction, or accepting the gamble or the offer). C_i is the experimental condition: the loss frame, lost cash, the gain in the mixed gamble coded as $(G - 100)/100$, or the ultimatum offer coded as

$(\text{offer} - 20)/10$. D_i equals 1 for the disguised wording and $\alpha_{m(i)}$ is the model fixed effect. β_1 is the effect of the condition under the published wording, and β_3 shows whether this effect changes with the wording. Because of the centering, β_2 is the effect of the disguised wording at a gain equal to the loss and at an offer of 20 €. The conjunction model has no condition term. For anchoring, the same equation was estimated by ordinary least squares on the log of the estimate, with C_i marking the high anchor. Standard errors were not clustered by model because there are only four models. As a sensitivity check, every choice model was estimated again with model-specific dummies for the target option being shown as option A.

RESULTS

All 3,520 responses were valid. Pooled over all tasks, the share of A choices ranged from 0.46 to 0.54 across models, and no model showed an overall preference for A after Holm correction (smallest adjusted p = 0.059). The task-level checks were less reassuring. GPT-OSS-120B chose option A in 77 of 80 framing trials (Holm-adjusted p < 0.001), GPT-OSS-20B chose it in only 23 of 80 mental accounting trials (p = 0.003), and ALLaM in 10 of 40 conjunction trials (p = 0.040). The order was balanced in every cell, but this does not rule out an influence on the comparisons, because a position preference may itself differ between conditions or wordings. The framing index of GPT-OSS-120B tells us more about option position than about framing. When position controls are added, the direction and statistical significance of all pooled estimates reported below are unchanged, but some effect sizes are sensitive to these controls; for example, the loss-frame odds ratio rises from 3.81 to 29.0. Table 2 gives the bias index for every model and wording, and Figure 1 shows the same estimates with their confidence intervals.

The clearest difference between the wordings appears in anchoring. With the published question on African countries in the United Nations, the models hardly moved with the anchor. Both GPT-OSS models gave a median estimate of 28% after either anchor, and the anchoring indices ranged from 0.00 to 0.05. With the disguised question on apricot orchards in Iğdır, every model moved toward the anchor. The index was 0.97 for GPT-OSS-20B and 0.64 for ALLaM, well above the human value of 0.36. For GPT-OSS-120B and Qwen it was 0.13 and 0.11. With the disguised question the anchor effect was significant for all four models after Holm correction, and so was the difference between the two wordings. In the pooled log-OLS model, no statistically significant effect of the high anchor was found for the published question (b = −0.06, 95% CI −0.28 to 0.16, p = 0.609), whereas the disguised question added a large effect (b = 1.35, 95% CI 0.97 to 1.73, p < 0.001).

Table 2. Bias indices by model and wording, English items, point estimates with 95% bootstrap confidence intervals

Measure	Wording	ALLaM-2-7B	GPT-OSS-120B	GPT-OSS-20B	Qwen3.8-27B	Human
Anchoring index	Published	0.05 [0.03, 0.12]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [−0.01, 0.01]	0.36
	Disguised	0.64 [0.45, 0.64]†	0.13 [0.11, 0.14]†	0.97 [0.95, 1.00]†	0.11 [0.05, 0.24]†	
Conjunction fallacy	Published	0.50 [0.30, 0.70][b]	0.00 [0.00, 0.00]	0.10 [0.00, 0.25]	0.15 [0.00, 0.30]	0.85
	Disguised	1.00 [1.00, 1.00]†,[b]	0.85 [0.70, 1.00]†	0.85 [0.70, 1.00]†	0.05 [0.00, 0.15]	
Rejection, offers ≤ 20 €	Published	0.22 [0.13, 0.32]	0.02 [0.00, 0.05]	0.07 [0.02, 0.13]	0.22 [0.12, 0.32]	–
	Disguised	0.78 [0.70, 0.87]†	0.02 [0.00, 0.05]	0.02 [0.00, 0.05]	0.35 [0.23, 0.47]	
Framing effect	Published	0.15 [0.00, 0.30]	0.00 [−0.30, 0.30][b]	0.35 [0.10, 0.60]	0.15 [0.00, 0.30]	0.50
	Disguised	0.20 [0.05, 0.40]	0.05 [−0.25, 0.35][b]	0.20 [0.05, 0.40]	0.00 [0.00, 0.00]	
Mental accounting	Published	−0.10 [−0.40, 0.20]	0.00 [0.00, 0.00]	−0.05 [−0.25, 0.15][b]	0.00 [−0.15, 0.15]	0.42
	Disguised	0.20 [−0.05, 0.45]	−0.35 [−0.55, −0.15]†	−0.10 [−0.40, 0.20][b]	0.40 [0.15, 0.65]	
Threshold ratio G*/L	Published	≤ 0.50[a]	0.88 [0.88, 0.90]	0.96 [0.91, 1.27]	0.88 [0.88, 0.90]	1.93[c]
	Disguised	≤ 0.50[a]	0.89 [0.88, 0.92]	1.00 [0.92, 1.07]	1.11 [1.00, 1.21]†	

Normative values are 0 for all indices except the threshold ratio (1). † Difference between the wordings significant after Holm correction (bootstrap test, $p < 0.05$). [a] Censored at the lower bound of the gain grid: the model accepted the gamble at every gain level in most trials. [b] Significant preference for one option position in this model and experiment, both wordings pooled (Holm-adjusted binomial test, $p < 0.05$). [c] Median loss aversion coefficient reported by Tom et al. (2007); an external reference that is not directly equivalent to the threshold ratio G*/L estimated here. Each experimental cell has 20 responses per model. Per model and wording, the indices use 20 observations for conjunction, 40 for framing, mental accounting and anchoring, 60 for rejection of offers ≤ 20 € and 180 for the mixed gamble. Confidence intervals are not adjusted for multiple comparisons.

Three models showed a similar pattern in the conjunction fallacy. On the original Linda problem, all four models committed the fallacy less often than the 85% reported for humans (Holm-adjusted $p \leq 0.001$): 0.00 for GPT-OSS-120B, 0.10 for GPT-OSS-20B, 0.15 for Qwen and 0.50 for ALLaM. When Linda was replaced by Deniz, the rate rose to 0.85 for both GPT-OSS models, the same as the human figure, and to 1.00 for ALLaM. Qwen was the exception. It chose the conjunction in only 0.05 of the disguised trials. The pooled odds ratio for the disguised wording was 32.3 (95% CI 8.5 to 122.9, $p < 0.001$) and 58.2 with position controls.

In the ultimatum game, both GPT-OSS models rarely rejected offers of 20 € or less under either wording (at most 0.07). ALLaM rejected such offers in 0.22 of the published trials and in 0.78 of the disguised trials. This was the only significant difference between the wordings in this task. Qwen's rejection rate rose from 0.22 to 0.35, but the difference was not significant after correction. In the pooled model, each additional 10 € raised the odds of acceptance under the published wording (odds ratio 2.55, $p < 0.001$). At an offer of 20 €, the disguised wording lowered the odds of acceptance (odds ratio 0.22, 95% CI 0.12 to 0.39, $p < 0.001$; 0.20 with position controls), and this effect did not vary detectably with the offer (interaction $p = 0.846$). Most of the shift came from ALLaM (Figure 2).

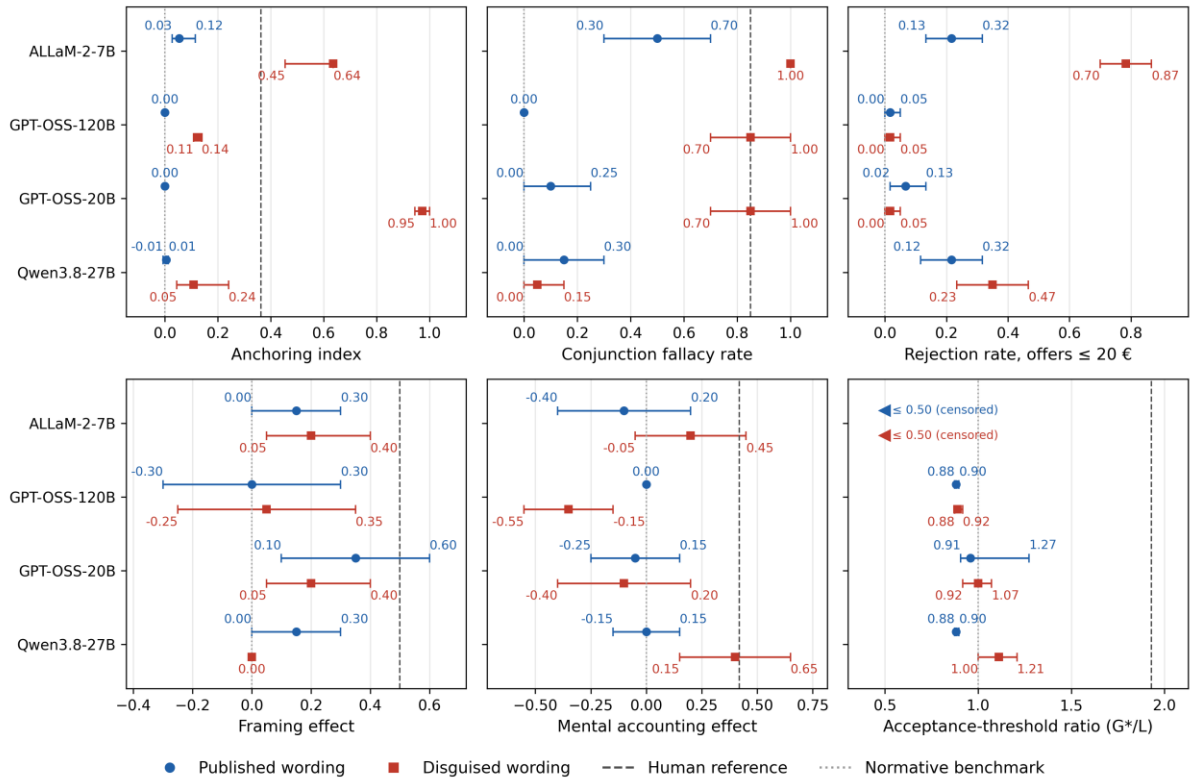


Figure 1. Bias indices for published and disguised wording by model. Bars are 95% stratified bootstrap confidence intervals with their bounds printed at both ends. Left-pointing triangles mark ratios censored at the lower end of the gain grid. No human reference line is drawn for the ultimatum index. For the threshold ratio, the dashed line marks the external reference of Tom et al. (2007).

Framing and mental accounting effects were weak. The framing index ranged from 0.00 to 0.35, below the human value of 0.50, and no model showed a significant effect on its own after Holm correction. For the published wording, pooled across models, the loss frame did increase the probability of a risky choice (odds ratio 3.81, 95% CI 1.49 to 9.69, p = 0.005). No statistically detectable interaction with the wording was found in this model (p = 0.598). With position controls, the loss-frame estimate rose to 29.0 (95% CI 3.6 to 231.8, p = 0.002) and the interaction remained non-significant (p = 0.402). The models also had very different baseline attitudes to risk. ALLaM chose the risky program in at least 80% of trials in every cell, while Qwen chose it in at most 15%. The mental accounting indices were close to zero for the published question. For the disguised question they went in opposite directions: −0.35 for GPT-OSS-120B and 0.40 for Qwen. Neither index was significant after correction, although the change between the wordings was significant for GPT-OSS-120B. In the pooled model, no statistically significant effect of losing cash rather than the ticket on the probability of buying was found under the published wording (odds ratio 0.74, 95% CI 0.32 to 1.70, p = 0.483), and the interaction with the wording was not significant (p = 0.383).

The mixed-gamble thresholds were well below the external human reference and mostly close to the risk-neutral value. GPT-OSS and Qwen accepted the gamble once the gain reached 0.88 to 1.11 times the loss, close to the risk-neutral value of 1 and far below the external reference of 1.93. The implied

prospect theory parameter λ_{CPT} ranged from 0.83 to 1.02, compared with 2.25 for humans (Tversky and Kahneman, 1992). ALLaM accepted the gamble in at least 90% of trials even when the gain was half the loss, so its ratio is censored at the lower bound of 0.50. The wording made a difference only for Qwen, whose ratio rose from 0.88 to 1.11 (95% CI 1.00 to 1.21). In the pooled model, acceptance rose sharply with the gain ($p < 0.001$). At a gain equal to the loss, no statistically significant effect of the disguised wording on acceptance was found (odds ratio 0.84, 95% CI 0.50 to 1.41, $p = 0.509$), but the gain effect was steeper under the disguised wording (interaction $p = 0.038$).

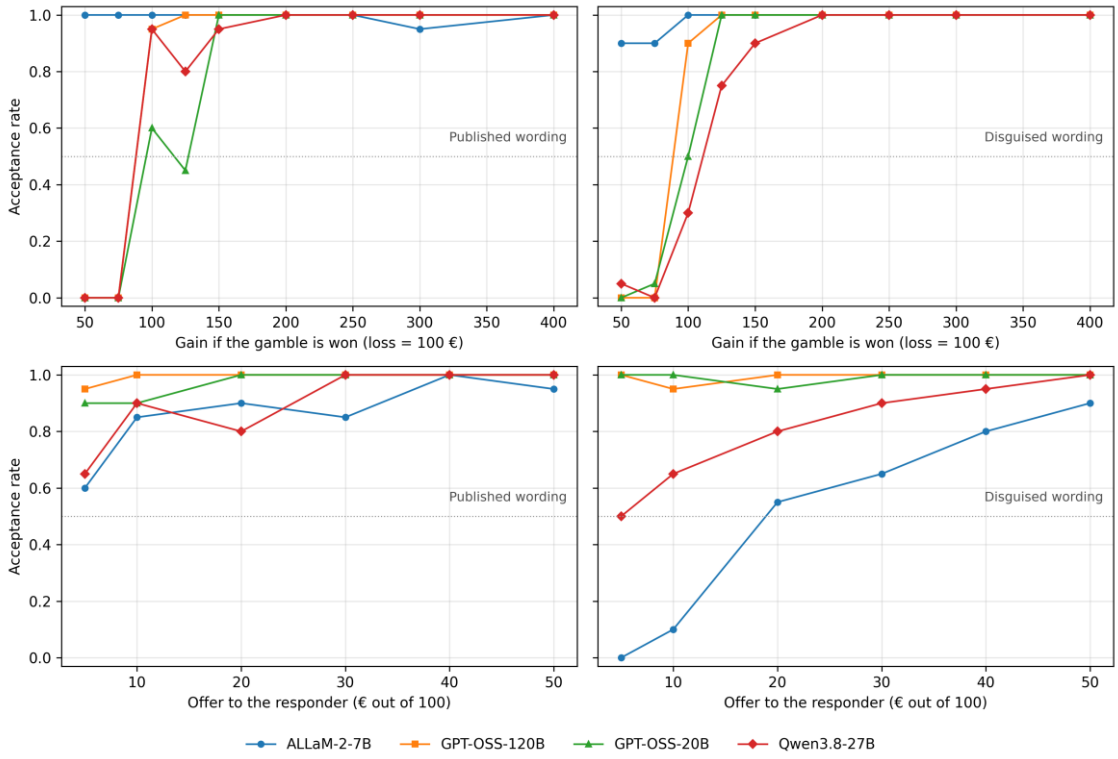


Figure 2. Acceptance rates in the mixed gamble (top) and the ultimatum game (bottom) by stake level, published (left) and disguised (right) wording. The dotted line marks 50% acceptance.

DISCUSSION AND CONCLUSION

For several models, the published wording of classic experiments gave a misleading picture. With the original anchoring question the models acted as if anchors did not affect them, and with the original Linda problem they looked much more rational than people. When the same structure came with a new story, three of the four models showed the conjunction fallacy at or above the human rate, and all four followed the anchor. These results are in line with the fragility that Binz and Schulz (2023) found for GPT-3. They also fit the possibility raised by Hagendorff et al. (2023) that models answer familiar tasks from memory.

The gap between the two wordings is not proof of contamination. The disguised questions change more than the surface story. The apricot question asks about a less familiar quantity, which leaves more room for the anchor to work. A bonus shared with a colleague may also bring different social norms than

a split with an anonymous stranger. For this reason we describe the results as sensitivity to wording that is consistent with contamination. How human-like the models are also depends on the experiment. The models' acceptance thresholds stayed below the external human reference, and framing effects were small. The models do not copy human behavior as a whole, so each experiment has to be judged on its own.

For economists who use LLMs as simulated agents, textbook questions are therefore a weak test of how these models decide. New questions with the same structure, written in several versions, give a more reliable picture. The study has limitations. All four models were run through one provider at a single temperature, each experiment had only one disguised story, and only the English questions are analyzed here. Some models preferred one option position in particular tasks, and the mixed gamble used one probability and a fixed loss, so it cannot separate loss aversion from risk attitude and probability weighting. In future work we will compare several disguised versions of each experiment and add the Turkish responses collected in parallel. A broader design in which gains, losses and probabilities vary systematically would support the estimation of prospect theory parameters.

Conflict of Interest

The authors have declared that there is no conflict of interest.

Author Contributions

Cemal ÖZTÜRK: conceptualization, experimental design, software, data collection, analysis and writing. Serkan KÜNÜ: review and editing.

Data and Code Availability

The prompts, English responses and analysis code are available on Zenodo: https://doi.org/10.5281/zenodo.23007074

Use of Generative AI

The authors used Claude (Anthropic) to check the language and readability of the manuscript. The authors reviewed and edited all suggested changes and take full responsibility for the content of the paper.

References

Bari MS, Alnumay Y, Alzahrani NA, et al., 2024. ALLaM: Large language models for Arabic and English. arXiv preprint arXiv:2407.15390. https://doi.org/10.48550/arXiv.2407.15390

Binz M, Schulz E, 2023. Using cognitive psychology to understand GPT-3. Proceedings of the National Academy of Sciences, 120 (6): e2218523120. https://doi.org/10.1073/pnas.2218523120

Dillion D, Tandon N, Gu Y, Gray K, 2023. Can AI language models replace human participants? Trends in Cognitive Sciences, 27 (7): 597–600. https://doi.org/10.1016/j.tics.2023.04.008

Efron B, Tibshirani RJ, 1993. An introduction to the bootstrap. Chapman & Hall/CRC, New York, USA. https://doi.org/10.1201/9780429246593

Güth W, Schmittberger R, Schwarze B, 1982. An experimental analysis of ultimatum bargaining. Journal of Economic Behavior & Organization, 3 (4): 367–388. https://doi.org/10.1016/0167-2681(82)90011-7

Hagendorff T, Fabi S, Kosinski M, 2023. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. Nature Computational Science, 3: 833–838. https://doi.org/10.1038/s43588-023-00527-x

Holm S, 1979. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6 (2): 65–70. https://www.jstor.org/stable/4615733

Horton JJ, Filippas A, Manning BS, 2023. Large language models as simulated economic agents: What can we learn from homo silicus? NBER Working Paper No. 31122 (revised February 2026), National Bureau of Economic Research, Cambridge, MA, USA. https://doi.org/10.3386/w31122

Jacowitz KE, Kahneman D, 1995. Measures of anchoring in estimation tasks. Personality and Social Psychology Bulletin, 21 (11): 1161–1166. https://doi.org/10.1177/01461672952111004

Kahneman D, Tversky A, 1979. Prospect theory: An analysis of decision under risk. Econometrica, 47 (2): 263–291. https://doi.org/10.2307/1914185

Mei Q, Xie Y, Yuan W, Jackson MO, 2024. A Turing test of whether AI chatbots are behaviorally similar to humans. Proceedings of the National Academy of Sciences, 121 (9): e2313925121. https://doi.org/10.1073/pnas.2313925121

Oosterbeek H, Sloof R, van de Kuilen G, 2004. Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. Experimental Economics, 7 (2): 171–188. https://doi.org/10.1023/B:EXEC.0000026978.14316.74

OpenAI, 2025. gpt-oss-120b & gpt-oss-20b model card. arXiv preprint arXiv:2508.10925. https://doi.org/10.48550/arXiv.2508.10925

Pezeshkpour P, Hruschka E, 2024. Large language models sensitivity to the order of options in multiple-choice questions. Findings of the Association for Computational Linguistics: NAACL 2024, 2006–2017. https://doi.org/10.18653/v1/2024.findings-naacl.130

Sainz O, Campos J, García-Ferrero I, Etxaniz J, Lopez de Lacalle O, Agirre E, 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. Findings of the Association for Computational Linguistics: EMNLP 2023, 10776–10787. https://doi.org/10.18653/v1/2023.findings-emnlp.722

Tom SM, Fox CR, Trepel C, Poldrack RA, 2007. The neural basis of loss aversion in decision-making under risk. Science, 315 (5811): 515–518. https://doi.org/10.1126/science.1134239

Tversky A, Kahneman D, 1974. Judgment under uncertainty: Heuristics and biases. Science, 185 (4157): 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Tversky A, Kahneman D, 1981. The framing of decisions and the psychology of choice. Science, 211 (4481): 453–458. https://doi.org/10.1126/science.7455683

Tversky A, Kahneman D, 1983. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. Psychological Review, 90 (4): 293–315. https://doi.org/10.1037/0033-295X.90.4.293

Tversky A, Kahneman D, 1992. Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and Uncertainty, 5 (4): 297–323. https://doi.org/10.1007/BF00122574

Yang A, Li A, Yang B, et al., 2025. Qwen3 technical report. arXiv preprint arXiv:2505.09388. https://doi.org/10.48550/arXiv.2505.09388

Evaluating Dimensionality Reduction Through Correlation-Network Preservation, Neighborhood Structure, and Predictive Performance Across Three Datasets

Fatemeh Sarasir[1], Vilda Purutçuoğlu[2]*, Ahmet Bursalı[3]

[1]Middle East Technical University, Department of Biomedical Engineering, Ankara, Türkiye

[2]Middle East Technical University, Department of Statistics, Ankara, Türkiye

[3]Huawei R&D Center, Intelligent Application DC Department, Istanbul, Türkiye

*Corresponding author e-mail: vpurutcu@metu.edu.tr

Abstract

Dimensionality reduction is widely used to simplify high-dimensional data, improve visualization, and support predictive modeling. However, retaining variance does not necessarily preserve feature relationships, local neighborhoods, or predictive utility. This study evaluates dimensionality reduction across three heterogeneous datasets: diabetes (100,000 observations), gallstone (319 observations), and lifestyle and exercise (20,000 observations). Principal component analysis (PCA), a PCA-equivalent linear autoencoder representation, and the nonlinear embedding methods UMAP and t-SNE were examined using explained variance, reconstruction error, thresholded correlation-network preservation, neighborhood overlap, and reported correlation-similarity and trustworthiness measures. Preliminary classification results were also considered for diabetes and gallstone data. With 16 principal components, all original feature-correlation edges at the threshold $|r| \geq 0.5$ were retained in each dataset. Explained variance was 91.86%, 90.43%, and 87.41% for diabetes, gallstone, and lifestyle data, respectively. The reconstructed networks nevertheless contained 5, 21, and 6 additional edges, showing that retention of original strong links does not imply identical network topology. The reported correlation-matrix similarities were 0.9725, 0.9953, and 0.9688, respectively, although the similarity formula requires documentation. UMAP and t-SNE showed dataset- and neighborhood-dependent results; some comparisons used different sample sizes, and trustworthiness calculations require verification. Preliminary PCA-based classifiers showed lower discrimination than models using original predictors, particularly for gallstone data. These descriptive results suggest that variance retention, structural fidelity, and prediction are distinct evaluation criteria. Dataset provenance, target construction, possible leakage, and aspects of the validation procedures remain unverified. Accordingly, the results should be interpreted as exploratory rather than definitive.

Keywords: *Dimensionality Reduction, Principal Component Analysis, Correlation Networks, Neighborhood Preservation, Predictive Modeling*

INTRODUCTION

Modern biomedical and lifestyle datasets often contain numerous correlated measurements, redundant variables, and heterogeneous information collected across different domains. Although high-dimensional datasets may contain useful predictive information, their complexity can complicate visualization, modeling, and interpretation. Dimensionality reduction seeks to represent such data using fewer variables while retaining information relevant to a particular analytical objective.

Principal component analysis (PCA) is one of the most widely used linear dimensionality-reduction methods. It transforms correlated variables into orthogonal linear combinations ordered by the variance they explain (Jolliffe and Cadima, 2016). PCA provides a computationally efficient baseline for compression and reconstruction. Nevertheless, its optimization criterion is variance retention rather than the preservation of every pairwise relationship among the original variables. Consequently, retaining a high percentage of variance does not necessarily guarantee preservation of correlation networks or outcome-related information.

Autoencoders provide a neural-network-based alternative by learning a compressed representation through input reconstruction. Linear autoencoders with suitable architectures and optimization procedures can recover a PCA-equivalent subspace, whereas nonlinear autoencoders can represent more complex transformations (Hinton and Salakhutdinov, 2006). However, a lower reconstruction error does not necessarily imply that a representation preserves the original feature associations or improves prediction.

Nonlinear manifold-learning methods address a different objective. Uniform manifold approximation and projection (UMAP) constructs low-dimensional representations intended to preserve aspects of the data's neighborhood structure (McInnes et al., 2018). The t-distributed stochastic neighbor embedding (t-SNE) method maps high-dimensional similarities into a low-dimensional space and is particularly useful for visualizing local structure (van der Maaten and Hinton, 2008). Because these methods emphasize relationships between observations rather than reconstruction of the original feature matrix, their evaluation requires different criteria from those used for PCA.

A further consideration is predictive utility. A reduced representation may preserve substantial variance and strong feature correlations while discarding information relevant to classification. Conversely, predictive performance alone does not establish whether a model's predictions can be meaningfully interpreted. Explainable artificial intelligence (XAI) methods, including SHapley Additive exPlanations (SHAP), can characterize the contributions of input features to model predictions (Lundberg and Lee, 2017). Comparing SHAP attributions between original-predictor and PCA-reduced models can provide additional insight into how dimensionality reduction affects feature-level explanations, although these attributions describe model behavior rather than causal effects.

This study evaluates dimensionality reduction across three heterogeneous datasets representing diabetes, gallstone status, and lifestyle and exercise. Its primary objective is to investigate whether dimensionality reduction preserves variance and strong feature relationships, maintains observation-level neighborhood structure, and retains predictive utility.

Specifically, the study examines variance retention and reconstruction under PCA, preservation of thresholded correlation-network links and global correlation structure, and neighborhood preservation

under UMAP and t-SNE at different neighborhood sizes. It also compares the available predictive results for original-predictor and PCA-reduced models and explores differences in SHAP feature attributions across the original and PCA-16 pipelines. Herein, the SHAP analysis is interpreted descriptively.

On conclusion, the main contribution is an integrated, multi-criterion evaluation framework that distinguishes numerical fidelity, correlation-network preservation, neighborhood structure, predictive utility, and interpretability. Across three heterogeneous datasets, the findings illustrate that preserving original strong correlation links does not necessarily ensure identical network reconstruction or equivalent predictive performance. By evaluating these complementary criteria together, the study highlights why explained variance alone is insufficient for judging the suitability of a reduced representation for a particular analytical objective.

MATERIAL AND METHODS

Material

Study Design and Datasets

A comparative computational analysis was conducted using three tabular datasets. These datasets are freely avaibbale in kaggle database. Hereby, each dataset was analyzed separately because of differences in sample size, feature composition, and target definition.

Table 1. Dataset characteristics

Dataset	Observations	Total columns	Target	Outcome
Diabetes	100,000	31	diagnosed_diabetes	Binary
Gallstone	319	39	Gallstone Status	Binary
Lifestyle and exercise	20,000	54	Burns_Calories_Bin	Four-class

The diabetes dataset contained 22 numerical predictors after excluding the target and the variables diabetes_stage and diabetes_risk_score, which were removed because of their potential to encode information closely related to the target.

The gallstone dataset contained 319 observations and 38 candidate predictors, with Gallstone Status as the target. The file was imported using its semicolon delimiter and decimal-comma format. The correlation-network analysis used the reported numerical predictor matrix.

The lifestyle and exercise dataset contained 20,000 observations and 54 columns. The target, Burns_Calories_Bin, contained four classes: Low, Medium, High, and Very High. The class counts were 5,003, 4,997, 5,000, and 5,000, respectively. Thirty-four numerical predictors were used for the reported dimensionality-reduction analysis, while 14 categorical or text columns were excluded from the numerical embedding analysis.

Five variables were initially excluded from the lifestyle predictor matrix because of possible target leakage: Calories_Burned, Burns Calories (per 30 min), Burns Calories (per 30 min)_bc, expected_burn, and cal_balance.

The three datasets were treated as separate case studies. The diabetes outcome was diagnosed_diabetes, the gallstone outcome was Gallstone Status, and the lifestyle outcome was the four-

class Burns_Calories_Bin variable. In diabetes, diabetes_stage and diabetes_risk_score were excluded because of their potential relationship to the outcome. In lifestyle, Calories_Burned, Burns Calories (per 30 min), Burns Calories (per 30 min)_bc, expected_burn, and cal_balance were excluded as potentially leakage-related predictors. However, the relationship between the lifestyle target and remaining candidate predictors is uncertain; these exclusions cannot establish that all target leakage was eliminated. Predictive results are therefore preliminary rather than fully validated out-of-sample evidence.

Data Preparation

For each dataset, the outcome variable was excluded from unsupervised dimensionality reduction. Numerical predictors were used to calculate Pearson correlation matrices and construct low-dimensional representations.

Numerical predictors were considered for PCA and distance-based embedding methods, for which appropriate scaling is important. In predictive validation, data-dependent operations such as imputation, scaling, feature selection, and PCA are fitted within each training fold. Training-fold scaling and PCA were reported for the refined gallstone analysis as example.

The reported data-quality summaries identified several variables requiring additional verification. In the lifestyle dataset, BMI and BMI_calc were nearly identical, while some variables such as sugar_g and cholesterol_mg contained negative values despite their names suggesting nonnegative quantities. These observations were not automatically removed; the variable definitions, units, and data-generation process must be checked before deciding how to handle them.

Methods

PCA, autoencoders, UMAP, and t-SNE optimize different objectives. Explained variance, reconstruction error, correlation-network preservation, neighborhood overlap, prediction performance, and model explanations are therefore complementary rather than interchangeable criteria. This study brings these criteria together across three heterogeneous datasets to investigate the practical trade-offs associated with dimensionality reduction.

Pricipal Component Analysis

PCA is a classical technique for representing correlated variables through orthogonal components that capture decreasing amounts of variance (Jolliffe and Cadima, 2016). It is frequently used for data compression, exploratory analysis, and preprocessing. Nevertheless, the variance explained by a component is not equivalent to its predictive relevance, and variance retention alone does not establish preservation of individual feature associations. This motivates evaluating PCA using both reconstruction metrics and correlation-network criteria.

PCA was applied to the numerical predictor matrix. Sixteen components were retained as a common reference dimension for the principal descriptive comparisons across the three datasets. This fixed choice was not intended to identify an optimal number of components for each dataset; the proportion of variance explained was therefore also reported separately for each dataset.

Let X denote the standardized $n \times p$ predictor matrix, and let V_k contain the loading vectors corresponding to the first k principal components. The reduced representation is

$$Z = XV_k$$

and the rank-k reconstruction is

$$\hat{X} = ZV_k^T.$$

The proportion of variance explained by the first k components was used to quantify compression. This measure was interpreted alongside reconstruction and correlation-network metrics rather than as a complete measure of information preservation.

Linear and Nonlinear Autoencoders

Neural networks with low-dimensional bottlenecks can learn compressed data representations (Hinton and Salakhutdinov, 2006). Under suitable architectural and optimization conditions, linear autoencoders can recover a PCA-equivalent subspace. Nonlinear autoencoders add flexibility but require choices concerning architecture, regularization, optimization, and training stability. Their reconstruction quality must therefore be distinguished from preservation of statistical relationships and predictive performance.

A linear autoencoder representation was treated as PCA-equivalent under the stated theoretical conditions and was not counted as independent evidence of superior performance. This equivalence refers to the represented principal subspace under suitable conditions, not to arbitrary autoencoder architectures or training procedures.

A nonlinear autoencoder was explored for the diabetes dataset. Its results are reported separately because a consistently configured nonlinear autoencoder evaluation was not available for all three datasets.

Correlation-Network Construction

Pearson correlation matrices were calculated among the original numerical predictors. An undirected correlation network was constructed by treating each variable as a node and each sufficiently strong correlation as an edge.

An edge was included if

$$|r_{ij}| \geq \tau$$

where r_{ij} is the Pearson correlation between variables i and j, and the primary threshold was $\tau = 0.5$.

The number of possible undirected edges among p variables is

$$E_{max} = p(p-1)/2.$$

Furthermore, network density is

$$D = E / E_{max}$$

where E is the number of observed edges.

Let A denote the edge set from the original correlation matrix and B the edge set from the reconstructed matrix. The edge-preservation rate is

$$P_edge = |A \cap B| / |A|.$$

The number of additional edges was calculated as $|B \setminus A|$. Additional edges indicate that correlations crossed the selected threshold after reconstruction despite not crossing it in the original matrix. Such edges indicate a change in network topology, but are not necessarily false associations without further analysis.

The primary correlation-network threshold was $|r| \geq 0.5$. Threshold sensitivity at $|r| \geq 0.4$ and $|r| \geq 0.6$ was considered for gallstone and lifestyle data, but the central comparative tables report the primary threshold. Therefore, conclusions about full retention of original edges are restricted to that threshold.

Reconstruction and Correlation-Matrix Metrics

Three metrics were used to assess PCA reconstruction:

1. Correlation-matrix similarity, calculated using the analysis implementation.
2. Correlation-matrix root mean squared error (RMSE).
3. Standardized feature-reconstruction RMSE.

For original and reconstructed correlation matrices R and $\hat{R}$, correlation RMSE is

$$\mathrm{RMSE_R} = \sqrt{\{ (1 / p^2) \Sigma_{i=1}^{p} \Sigma_{j=1}^{p} (R_{ij} - \hat{R}_{ij})^2 \}}.$$

Lower correlation RMSE indicates closer agreement between the original and reconstructed correlation matrices. The correlation-matrix similarity values were retained from the original analysis, Accordingly, these similarity values are reported for descriptive completeness only; the interpretable correlation RMSE and thresholded edge counts provide the principal evidence on structural fidelity.

These metrics measure distinct properties. Reconstruction RMSE evaluates feature-level reconstruction, correlation RMSE evaluates pairwise correlation agreement, and edge preservation evaluates whether strong thresholded relationships remain present. In particular, 100% preservation of original edges does not imply exact network recovery when additional edges are introduced.

UMAP and t-SNE

UMAP constructs an embedding based on a representation of local relationships in the original feature space (McInnes et al., 2018). t-SNE converts pairwise similarities into probability distributions and minimizes a divergence between the high-dimensional and low-dimensional distributions (van der Maaten and Hinton, 2008). Both methods are widely used for visualization, but their embeddings are not designed to reconstruct the original feature matrix. Their evaluation is consequently emphasize neighborhood preservation and trustworthiness rather than explained variance alone.

UMAP and t-SNE were used to generate two-dimensional embeddings. Neighborhood preservation was evaluated by comparing the nearest neighbors of observations in the original feature space and in the embedded space.

Let $N_k^{(x)}(i)$ and $N_k^{(z)}(i)$ denote the k-nearest-neighbor sets of observation i in the original and embedded spaces, respectively. Neighborhood preservation was calculated as

$$\mathrm{NP}(k) = (1 / nk) \Sigma_{i=1}^{n} |N_k^{(x)}(i) \cap N_k^{(z)}(i)|.$$

The result was reported as a percentage. Higher values indicate greater overlap between the original and embedded neighborhoods.

The analysis considered $k \in \{5, 10, 15, 20, 30\}$. Trustworthiness values are reported from the available outputs, but their implementation has not been verified against a standard definition; consequently, they are interpreted descriptively.

For the diabetes dataset, UMAP was fitted to the full dataset, while t-SNE was fitted to a 10,000-observation sample. For the gallstone dataset, both methods were evaluated on all 319 observations. For the lifestyle dataset, both methods were fitted to the same random sample of 3,000 observations. The latter two comparisons are therefore more directly comparable within each dataset.

Predictive Modeling

ROC analysis is a standard approach for assessing binary classifier discrimination (Fawcett, 2006). The Brier score evaluates the accuracy of probabilistic predictions by measuring squared differences between predicted probabilities and observed outcomes (Brier, 1950). These metrics complement structural measures by assessing whether dimensionality reduction retains information relevant to prediction.

Valid predictive comparisons require that data-dependent preprocessing be learned only from training data. Scaling, feature selection, imputation, and PCA fitting are repeated within cross-validation training folds. This requirement is reported as implemented in the refined gallstone workflow.

Furthermore, the prediction analysis compared original predictors with PCA-reduced representations within each dataset. Because evaluation procedures differed, the diabetes and gallstone analyses are presented as separate illustrative cases rather than as a common controlled benchmark. For both datasets, numerical performance differences describe the available results.

For diabetes, binary logistic regression was used, and the reported evaluation included apparent AUC (area under the curve), apparent Brier score, and ten-fold cross-validation AUC.

For gallstone, ridge-regularized logistic regression was evaluated using repeated stratified five-fold cross-validation. The refined analysis used training-fold preprocessing for scaling and PCA.

For lifestyle and exercise, multinomial ridge regression was considered for the four-class outcome; only exploratory coefficient-based rankings were available. No verified cross-validated multiclass predictive performance was supplied, so no quantitative lifestyle prediction comparison is claimed.

AUC was used to describe discrimination (Fawcett, 2006), and the Brier score was used to describe probabilistic prediction accuracy (Brier, 1950). Apparent-fit metrics are not equivalent to out-of-sample estimates, and comparisons across datasets with different validation protocols are not treated as a controlled benchmark.

SHAP-Based Explainability Assessment

SHapley Additive exPlanations (SHAP) were used to investigate how predictors contributed to classification-model predictions and whether feature attributions differed between models fitted using the original predictors and PCA-reduced representations (Lundberg and Lee, 2017). SHAP values characterize model-specific contributions to predictions and cannot be interpreted as causal effects. For PCA-based models, the explanation space was defined in terms of the original input predictors by

incorporating the fitted preprocessing, PCA transformation, and classifier into the prediction pipeline.

SHAP analyses were conducted for the diabetes, gallstone, and lifestyle datasets. Logistic regression was used for the binary diabetes and gallstone outcomes, whereas multinomial logistic regression was used for the multiclass lifestyle outcome. Potentially leakage-prone variables were excluded during preprocessing, and numerical predictors were standardized using parameters estimated from the training data. Each dataset was divided into training and held-out test sets using an 80:20 stratified split with random seed 123. For the PCA-based pipelines, PCA was fitted on the training data, and the first 16 principal components were retained.

Approximate SHAP values were estimated using the fastshap R package with 100 Monte Carlo simulations (nsim = 100) and the adjust = TRUE option. The background dataset consisted of a random sample of up to 100 training observations, and explanations were calculated for up to 100 held-out observations per dataset. For the binary outcomes, SHAP values explained the predicted probability of the positive class. For the lifestyle outcome, attributions were calculated separately for each class probability.

Global feature importance was summarized using the mean absolute SHAP value for each feature: $I_j = (1/n) \Sigma_{i=1}^{n} |\varphi_{ij}|$, where φ_{ij} denotes the SHAP attribution of feature j for observation i, and n is the number of observations explained. For each dataset and outcome class, the ten highest-ranked features were identified separately for the original-predictor and PCA-16 pipelines. Their overlap was used as a descriptive measure of feature-importance similarity.

Finally, comparisons were based on mean absolute SHAP values and top-10 feature overlap, with attributions expressed in the original input-feature space. The top-10 overlap measures shared membership among the ten leading features. Accordingly, the findings describe exploratory differences in model-specific attribution patterns, not causal effects, formal ranking agreement, or improved predictive reliability.

Reproducibility and Statistical Considerations

These analyses are descriptive and exploratory. Reproducible verification would require documented random seeds, software versions, preprocessing rules, model hyperparameters, target definitions, and the correlation-similarity formula. Predictive transformations can be trained within validation folds. Consequently, the reported differences are observed numerical contrasts, not statistically significant effects. The trustworthiness implementation, correlation-similarity definition, dataset provenance, and target construction remain outstanding verification requirements.

RESULTS

Dimensionality Reduction and Explained Variance

Table 2. PCA dimensionality reduction

Dataset	Numerical predictors	Components	Variance explained
Diabetes	22	16	91.86%
Gallstone	38 candidate predictors	16	90.43%
Lifestyle and exercise	34	16	87.41%

For the diabetes dataset, 16 components retained 91.86% of the variance in the 22 numerical predictors. In the gallstone dataset, the initial analysis reported that 16 components explained 90.43% of the variance. In the lifestyle dataset, 16 components explained 87.41%, while 17 components were required to exceed 90%, retaining 90.01%.

These findings demonstrate that substantial dimensionality reduction was possible across all three datasets. However, explained variance alone did not establish whether the reduced representations preserved strong feature relationships or predictive performance.

Correlation-Network Preservation

At the primary threshold ($|r| \geq 0.5$), PCA-based reconstruction preserved all original strong correlation links in each dataset.

In diabetes, the original network contained six strong edges. These included associations between total cholesterol and LDL cholesterol (r=0.906), postprandial glucose and HbA1c (r=0.933), BMI and waist-to-hip ratio (r=0.766), and fasting glucose and HbA1c (r=0.700). All six edges were retained, with five additional edges appearing in the reconstructed network.

The gallstone network contained 93 original strong edges. All 93 were retained, while the reconstructed network contained 114 edges, corresponding to 21 additional edges.

The lifestyle network contained 43 original strong edges among 34 numerical predictors. PCA reconstruction preserved all 43 edges and introduced six additional edges.

Table 3. Correlation-network preservation and reconstruction

Dataset	Original edges	Preserved edges	Additional edges	Correlation similarity	Correlation RMSE	Reconstruction RMSE
Diabetes	6	6 (100%)	5	0.9725	0.0606	0.2854
Gallstone	93	93 (100%)	21	0.9953	0.0534	0.3089
Lifestyle and exercise	43	43 (100%)	6	0.9688	0.0694	0.3549

The gallstone data had the highest reported correlation similarity and the lowest correlation RMSE, while lifestyle data had larger reported errors. On the other side, complete preservation of original edges did not imply exact network recovery. Additional edges in every dataset indicate that some pairwise correlations crossed the selected threshold after reconstruction even though they did not cross it in the original matrix.

Neighborhood Preservation

For diabetes, UMAP achieved neighborhood-preservation scores of 3.02%, 4.33%, 5.41%, 6.29%, and 7.83% at (k=5,10,15,20,30), respectively. The reported trustworthiness at (k=15) was 0.830.

For t-SNE, the corresponding scores were 30.99%, 26.58%, 24.18%, 22.78%, and 21.27%, with trustworthiness of 0.936 at (k=15). The recorded values show higher neighborhood overlap for t-SNE in this diabetes analysis.

Table 4. Neighborhood preservation for gallstone

(k)	UMAP preservation	t-SNE preservation	UMAP trustworthiness	t-SNE trustworthiness

5	41.00%	47.96%	0.933	0.955
10	44.76%	48.24%	0.908	0.933
15	46.60%	50.41%	0.889	0.923
20	48.78%	52.35%	0.884	0.915
30	52.02%	56.33%	0.874	0.907

For the gallstone dataset, the reported t-SNE overlap and trustworthiness values exceeded the corresponding UMAP values at each evaluated k.

Table 5. Neighborhood preservation for lifestyle and exercise

k	**UMAP preservation**	**t-SNE preservation**	**UMAP trustworthiness**	**t-SNE trustworthiness**
5	63.61%	73.65%	0.981	0.988
10	56.33%	57.65%	0.964	0.969
15	51.64%	49.76%	0.950	0.954
20	47.06%	44.87%	0.939	0.941
30	40.52%	39.37%	0.920	0.922

For the lifestyle dataset, t-SNE achieved greater neighborhood overlap at (k=5) and (k=10), whereas UMAP performed better at (k=15,20,) and (30). Trustworthiness was slightly higher for t-SNE at all evaluated neighborhood sizes.

The observed embedding performance depended on the dataset and neighborhood size. Because evaluation settings differed across datasets and the trustworthiness implementation was not independently validated, the results are interpreted as exploratory descriptions of neighborhood structure.

Predictive Performance

Table 6. Preliminary diabetes prediction results

Representation	**Apparent AUC**	**Apparent Brier score**	**10-fold CV AUC**
Original predictors	0.9347	0.1006	0.9341
PCA, 16 components	0.9262	0.1080	0.9256
Linear autoencoder	0.9262	0.1080	0.9256
Nonlinear autoencoder	0.9251	0.1089	Not available

The original-predictor model achieved the highest reported discrimination and lowest apparent Brier score. PCA-based prediction showed a modest decrease in AUC and an increase in the Brier score. The PCA-equivalent linear autoencoder produced identical results to PCA.

The nonlinear autoencoder achieved correlation similarity of 0.9859, correlation RMSE of 0.0429, and standardized reconstruction RMSE of 0.2324. Its apparent AUC was 0.9251. Although its reported reconstruction metrics were better than those of PCA, its reported apparent prediction discrimination was not higher.

These observations illustrate that improved reconstruction does not necessarily coincide with improved apparent prediction. The diabetes estimates require a complete audit of target leakage and fold-specific preprocessing before they can support definitive predictive conclusions.

Table 7. Gallstone prediction performance. SD denotes the standard deviation.

Representation	Mean AUC	AUC SD	Mean Brier score	Mean sensitivity	Mean specificity
Original predictors	0.8623	0.0080	0.1514	0.7316	0.8497
PCA, 16 components	0.8210	0.0110	0.1738	0.7063	0.8062

In the reported repeated stratified five-fold cross-validation for gallstone data, the original-predictor model had higher mean AUC, lower mean Brier score, and higher reported sensitivity and specificity than the 16-component PCA model. Mean AUC differed by approximately 0.041.

The findings suggest that PCA may have removed or redistributed information relevant to distinguishing gallstone status. However, the small sample size means that performance may be sensitive to the split strategy, regularization, and predictor encoding. Moreover, the lifestyle dataset had exploratory coefficient-based importance results, but no verified cross-validation prediction results in the reported outputs. Consequently, its prediction performance cannot yet be compared quantitatively with the diabetes and gallstone models.

Table 8. Top ten coefficient-based importance scores for lifestyle data

Rank	Predictor	Mean absolute coefficient
1	Sets	0.1898
2	Physical exercise	0.1574
3	Reps	0.1008
4	lean_mass_kg	0.0899
5	Weight (kg)	0.0871
6	Max_BPM	0.0777
7	Session_Duration (hours)	0.0759
8	pct_HRR	0.0673
9	prep_time_min	0.0605
10	serving_size_g	0.0596

The largest coefficient magnitudes were associated with workout-related variables, particularly sets, physical exercise, and repetitions. Body-composition and physiological variables also appeared among the most highly ranked predictors. As a result, these are regression-coefficient magnitude rankings, not SHAP feature attributions. They are sensitive to predictor scaling, encoding, and correlation and can not be interpreted as causal effects or as directly comparable with the other datasets' importance results.

SHAP Explainability Results

SHAP analysis was completed for the diabetes, gallstone, and lifestyle datasets. For diabetes and gallstone, attributions explained the positive-class probability; for lifestyle, attributions were calculated separately for the probabilities of the High, Low, Medium, and Very High classes.

For diabetes, HbA1c was the dominant feature in the original-predictor model, with a mean absolute SHAP value of 0.316, followed by fasting glucose (0.0214) and family history of diabetes (0.0155). In the PCA-16 pipeline, HbA1c remained the leading feature (0.129), followed by postprandial glucose (0.122) and fasting glucose (0.100). For the gallstone dataset, the leading features in the original model were total body fat ratio (0.193), lean mass (0.160), and obesity (0.134). In contrast, the PCA-16 pipeline

was led by vitamin D (0.115), C-reactive protein (CRP; 0.0762), and high-density lipoprotein (HDL; 0.0578).

For the lifestyle dataset, the leading features varied across outcome classes. In the original model, carbohydrates had the highest mean absolute SHAP values for the Low (0.279) and Medium (0.259) classes, whereas lean mass was the leading feature for High (0.0993) and body weight for Very High (0.109). In the PCA-16 pipeline, preparation time was the leading feature for High (0.0197), sets for Low (0.0583) and Medium (0.0187), and repetitions for Very High (0.0303). The three leading features for each dataset, model representation, and outcome class are summarized in Table 9.

Table 9. SHAP feature attribution summary for original and PCA-based models

Dataset / class	Original model: top features (mean |SHAP|)	PCA-16 model: top features (mean |SHAP|)
Diabetes	HbA1c (0.316); fasting glucose (0.0214); family history (0.0155)	HbA1c (0.129); postprandial glucose (0.122); fasting glucose (0.100)
Gallstone	Total body fat ratio (0.193); lean mass (0.160); obesity (0.134)	Vitamin D (0.115); CRP (0.0762); HDL (0.0578)
Lifestyle — High	Lean mass (0.0993); weight (0.0916); proteins (0.0415)	Preparation time (0.0197); rating (0.0189); cholesterol (0.0185)
Lifestyle — Low	Carbohydrates (0.279); fats (0.217); proteins (0.153)	Sets (0.0583); repetitions (0.0551); sugar (0.0247)
Lifestyle — Medium	Carbohydrates (0.259); fats (0.185); proteins (0.180)	Sets (0.0187); sodium (0.0156); resting heart rate (0.0141)
Lifestyle — Very High	Weight (0.109); average heart rate (0.0637); percentage of maximum heart rate (0.0632)	Repetitions (0.0303); cholesterol (0.0278); preparation time (0.0259)

Abbreviations: CRP, C-reactive protein; HDL, high-density lipoprotein. Values are rounded to four significant digits where appropriate.

Between-model comparisons revealed varying degrees of overlap in SHAP feature rankings between the original and PCA-16 pipelines. The two representations shared 6 of the top 10 features for diabetes (60%) and 1 of 10 for gallstone (10%). For the lifestyle dataset, the overlap was 1 of 10 features for the High class (10%), 4 of 10 for Low (40%), 2 of 10 for Medium (20%), and 3 of 10 for Very High (30%). Thus, diabetes showed the greatest overlap in leading features, whereas gallstone and the lifestyle High class showed the least. However, top-10 overlap reflects shared feature membership rather than agreement in the complete rankings. Spearman's rank correlations were not calculated, and attribution stability across resampling or cross-validation folds was not assessed. These findings therefore provide a descriptive comparison of feature overlap but do not establish overall ranking agreement, attribution stability, or improvements in predictive performance following PCA.

DISCUSSION AND CONCLUSION

The principal finding was that PCA reconstruction preserved all original strong correlation links at the primary threshold in all three datasets. This was observed despite reducing the predictor spaces to 16 components.

The result supports evaluating thresholded correlation-network preservation as a complement to explained variance. However, additional edges appeared in every reconstructed network, indicating that preservation of all original strong links did not guarantee exact network recovery. The reported

numerical correlation measures support the distinction between retaining selected network edges and recovering the full correlation structure.

Accordingly, the diabetes dataset had only six strong edges in its original network. Preserving all six is encouraging, but a sparse network offers fewer opportunities for edge loss than a denser one. Edge-preservation rates can therefore be interpreted alongside network size, density, and global correlation error. The gallstone dataset exhibited numerous strong relationships among body-composition and hydration variables. The lifestyle dataset also contained highly redundant variables, including nearly identical BMI measures and strongly correlated nutritional features. Additionally, redundancy can make dimensionality reduction effective because multiple variables may encode overlapping information. Furthermore, the observed negative values for some lifestyle variables and the near-duplicate BMI measures require investigation before the final analysis.

The embedding results varied with dataset and neighborhood scale. For gallstone data, recorded t-SNE neighborhood-overlap values exceeded UMAP values across the examined settings. For lifestyle data, t-SNE had greater overlap for smaller k, whereas UMAP had greater overlap for larger k. These results emphasize that embedding quality depends on the scale at which local structure is evaluated. UMAP and t-SNE can primarily be regarded as tools for visualizing and assessing observation-level geometry, rather than as methods for reconstructing feature correlations.

The diabetes embedding comparison used different observation counts, making direct algorithmic comparisons inappropriate. Trustworthiness calculations also require independent verification. A controlled future benchmark would require identical observations, preprocessing, distance measures, and evaluation implementations. The preliminary prediction results suggested that dimensionality reduction may reduce classification performance even when strong correlations are retained. In diabetes, PCA resulted in a modest decrease in AUC. In gallstone, the decrease in AUC was larger, accompanied by a worse Brier score and lower sensitivity and specificity. One possible explanation is that PCA prioritizes variance rather than class discrimination. A direction with relatively low variance may still contain useful outcome information. In addition, components combine original measurements, which may complicate interpretation.

The nonlinear autoencoder was evaluated only in the diabetes dataset and is presented as an exploratory extension rather than a uniformly assessed alternative across the three datasets. Its results suggest that better reconstruction does not automatically lead to improved apparent predictive performance. The most appropriate representation therefore depends on whether the primary objective is compression, preservation of feature associations, visualization, or prediction.

SHAP results showed differences in leading predictors between the original and PCA-16 models. Diabetes had the greatest top-10 feature overlap (60%), while gallstone and lifestyle High had the lowest (10%). These differences may reflect how PCA transforms correlated predictors and redistributes predictive information. The overlap measures shared leading-feature membership only. On the other side, SHAP values represent model-specific contributions, not causal effects.

These conclusions remain provisional until preprocessing and validation steps have been verified for possible leakage and applied consistently. Taken together, the descriptive findings support a multi-criterion strategy for selecting dimensionality-reduction methods:

- Numerical fidelity: assess reconstruction error and correlation-matrix agreement.
- Structural fidelity: assess preservation of thresholded correlation-network links and observation neighborhoods.
- Predictive utility: evaluate out-of-sample discrimination and probabilistic accuracy.
- Interpretability: compare feature-level explanations only for datasets and model pairs with verified, compatible SHAP outputs reported in the explainability subsection.

A method that performs well on one criterion may not be optimal for another. The selection of a reduced representation can therefore be guided by the intended analytical use rather than by explained variance alone.

Several limitations qualify these exploratory findings. The datasets differ substantially in size, predictor composition, outcome type, and provenance, and their source documentation and target-construction procedures have not been independently verified. The analyses therefore do not constitute a controlled benchmark in which only the dimensionality-reduction method varies. Correlation-edge preservation was principally examined using Pearson correlations at $|r| \geq 0.5$, and other thresholds or association measures could give different network structures. The correlation-similarity formula was unavailable; accordingly, its reported values were not used as primary evidence or as a basis for ranking.

Additional limitations relate to embedding, prediction, and explanations. Diabetes UMAP and t-SNE analyses used unequal observation counts. Predictive validation also differed between diabetes and gallstone, and complete fold-specific preprocessing and potential target leakage have not been audited for every analysis. SHAP values were estimated on a limited set of held-out observations, and full ranking agreement and attribution stability can be checked via more datasets. Additionally, PCA-based explanations were reported in the original feature space, but compatibility of all explanation outputs is considered as the extention of this study. These limitations restrict the conclusions to descriptive patterns rather than definitive algorithm rankings or causal interpretations.

Despite these limitations, the results demonstrate why dimensionality reduction should be evaluated across complementary criteria. With 16 principal components, the available analyses retained all original strong correlation edges at $|r| \geq 0.5$ in the diabetes, gallstone, and lifestyle datasets. However, five, 21, and six additional edges respectively emerged after reconstruction. Thus, perfect retention of the original thresholded edges did not imply identical network topology, and the correlation RMSE values further characterized differences in reconstructed correlation structure.

Preliminary within-dataset prediction comparisons indicated lower observed discrimination for PCA-16 than for original-predictor models, especially in the gallstone case, while SHAP summaries revealed differences in leading feature attributions. These results do not show that preserved variance or strong correlations ensure predictive utility or stable model explanations. The nonlinear autoencoder finding is limited to the diabetes exploratory analysis.

Overall, the findings support a multi-criterion approach to dimensionality reduction, in which the suitability of a representation is assessed according to its intended purpose: numerical compression, preservation of feature relationships, visualization of observation-level structure, predictive modeling, or model interpretation.

References

Brier GW, 1950. Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78 (1): 1–3. doi: 10.1175/1520-0493(1950)078<0001>2.0.CO;2.

Fawcett T, 2006. An introduction to ROC analysis. Pattern Recognition Letters, 27 (8): 861–874. doi: 10.1016/j.patrec.2005.10.010.

Hinton GE, Salakhutdinov RR, 2006. Reducing the dimensionality of data with neural networks. Science, 313 (5786): 504–507. doi: 10.1126/science.1127647.

Jolliffe IT, Cadima J, 2016. Principal component analysis: A review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374 (2065): 20150202. doi: 10.1098/rsta.2015.0202.

Lundberg SM, Lee SI, 2017. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30: 4765–4774.

McInnes L, Healy J, Melville J, 2018. UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426. https://arxiv.org/abs/1802.03426.

Van der Maaten L, Hinton G, 2008. Visualizing data using t-SNE. Journal of Machine Learning Research, 9: 2579–2605.

An AI-Driven Framework for Document Processing, Data Validation, and Digital Archiving in Logistics Operations

Recep Akyüz[1*], Ender Şahinaslan[2]

[1]Barsan Global Logistics, Software and R&D Department, Istanbul, Turkey

[2]Maltepe University, Faculty of Business and Management Sciences, Management Information Systems, 34857, Istanbul, Turkey

*Corresponding author e-mail: recep.akyuz@barsan.com

Abstract

The effective management of customs declarations and supporting logistics documents requires accurate information extraction, document classification, validation against operational records, and secure long-term storage. Traditional workflows often rely on manual document scanning, data entry, classification, and cross-checking, which can increase processing times and the risk of transcription and indexing errors. This study proposes an artificial intelligence (AI)-driven framework for document processing, data validation, and intelligent digital archiving, focusing on customs declarations and related logistics documents. The proposed framework integrates image preprocessing, optical character recognition (OCR), natural language processing (NLP), machine learning-based document classification, and rule-based data validation into a unified workflow. Critical information, including customer details, dates, customer identifiers, and shipment references, is intended to be extracted from documents and compared with corresponding records in enterprise logistics information systems. The framework also proposes confidence-based human review, configurable validation rules, traceable exception management, and secure archival mechanisms. Data integrity, information security, and long-term accessibility are addressed as core design objectives through proposed mechanisms for access control, traceability of document processing activities, preservation of digital records, and integrity verification. Application programming interfaces (APIs) are envisaged to support interoperability with enterprise logistics platforms. The proposed evaluation methodology considers alternative OCR approaches, including Tesseract and cloud-based OCR services, alongside transformer-based language models and conventional classification methods. The framework is intended to be evaluated using field-level extraction accuracy, document classification performance, validation error detection, processing time, archival retrieval efficiency, manual workload, and criteria related to data integrity and information security. The main originality of the proposed framework lies in integrating AI-based document extraction and classification with domain-specific data validation, human oversight, traceable exception management, and secure digital archiving. Rather than treating digitization as an isolated text extraction task, the framework aims to connect document processing, data quality assurance, operational verification, and archival management within a coherent and interoperable approach. The proposed framework is intended to provide a foundation for developing more reliable, secure, and accessible digital archiving processes in logistics operations. a research framework for

more reliable and accessible digital archives in logistics operations.

Keywords: *Artificial Intelligence, Logistics Document Processing, Digital Archiving, Optical Character Recognition, Information Security*

INTRODUCTION

International logistics operations depend on documents that link commercial transactions to physical movements and customs procedures. A single shipment may generate a customs declaration, invoice, packing list, transport document, shipment reference, and supplementary evidence. These records frequently arrive from different parties, in different formats, and at different moments. When they are stored merely as scanned images, their contents remain difficult to query, reconcile, and audit. Incorrect indexing is especially consequential: a document can be legible while still being attached to the wrong customer, declaration, or operational movement.

The proposed research framework addresses this operational setting through a proposed digital archive supported by image processing and artificial intelligence. Its stated scope includes document recognition, classification, extraction of dates and operational identifiers, comparison with enterprise records, and secure storage. The project definition describes a modular system with a web interface and application programming interfaces (APIs), not an isolated OCR utility. The methodological problem studied here is how to turn that scope into a verifiable information-processing pipeline without treating automated predictions as authoritative evidence.

The framework aims to transform heterogeneous customs and logistics files into consistently indexed digital records, identify errors that cannot be detected by text recognition alone through comparison with corporate data, and address uncertainty management, human correction, source traceability, and information security within an integrated document lifecycle. The study contributes to the design and evaluation processes by proposing an integrated framework for processing and verifying logistics documents. However, since an independently validated test dataset and measured pilot application results are not available, quantitative conclusions regarding operational accuracy, cost savings, or reductions in processing times are not presented.

Optical character recognition converts visual symbols to machine-readable text, but OCR quality depends on acquisition conditions, typography, image noise, and document structure. Tesseract provides an established open-source OCR baseline (Smith, 2007). Neural language models such as BERT can provide contextual representations to support classification and extraction after text has been recognized (Devlin et al., 2019). Nevertheless, separating recognition from interpretation can propagate character-level errors into later decisions. Advanced document-understanding approaches therefore combine layout, image, and textual representations.

LayoutLMv3 reveals the importance of page layout in understanding forms and structured documents by learning common representations between text markers and image fragments (Huang et al., 2022). Donut, on the other hand, offers a transformer-based approach that does not require OCR for visual document understanding and provides an alternative method to prevent errors that may arise from an external OCR stage (Kim et al., 2022). While both approaches provide an important basis for the

comparative evaluation of document understanding methods, they do not alone guarantee reliable field reconciliation with enterprise data in shipment management systems. Therefore, the comparative results reported in the literature for these methods should not be considered as empirical findings specific to customs declarations. (Akyüz et al., 2026) addressed the integration between the processes of extracting, verifying, and creating document data within an AI-powered transit declaration workflow. This study, however, differs from the aforementioned approach by focusing on an integrated document lifecycle that includes digital archive management, document indexing, traceability, and processes following declaration submission.

A practical research gap exists between extracting a plausible document value and establishing that it belongs to the correct operational record. A syntactically valid declaration number may be inconsistent with the linked shipment; a correctly read date may refer to issue date rather than departure date. The proposed contribution is therefore a verification-centred architecture that combines document AI with domain constraints, enterprise reference data, explainable exception queues, and archived evidence. Conventional classifiers and layout-aware models are retained as competing alternatives rather than predetermined winners.

MATERIAL AND METHODS

Material

The methodological design is based on a proposed document-processing workflow for customs declarations and related logistics records. Document categories, extraction fields, operational references and access requirements are defined from the logistics use case; no production measurements or completed comparative experiment are claimed in this paper.

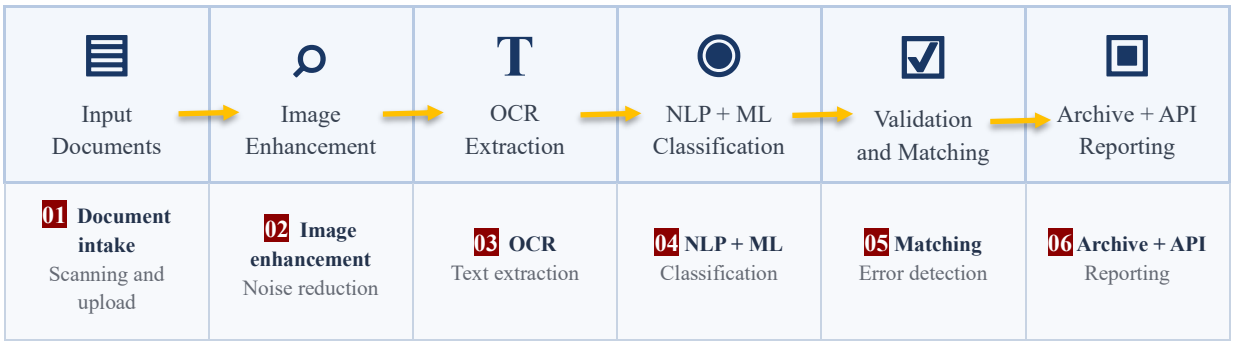


Figure 1. Workflow of the proposed AI-based logistics document processing and archiving framework.

The ingestion layer receives scanner files, uploaded PDFs, and image documents. It assigns a unique document identifier, retains the original binary file, records capture time and source, and calculates a cryptographic digest. Pages are evaluated for rotation, contrast, noise, and resolution before optional deskewing, denoising, and adaptive contrast enhancement. Preprocessing decisions must be recorded because excessive filtering can remove fine print, punctuation, or identifier characters. The original is never overwritten by an enhanced derivative.

Text extraction first establishes reproducible baselines using Tesseract and a configurable managed OCR service. Output is represented as tokens with bounding boxes, page numbers, OCR confidence scores when available, and source coordinates. A layout-aware model can subsequently be tested on the same held-out document set. Because confidence values from different engines are not directly comparable, their relationship with observed correctness should be calibrated separately. OCR-free processing, if evaluated, is a distinct experimental condition rather than a component presumed already deployed.

Methods

Document Classification and Field Extraction

A document classification model is employed to assign incoming records to a predefined taxonomy comprising customs declarations, commercial invoices, packing lists, transport documents, and other supporting documents. The baseline approach employs conventional machine learning algorithms, such as support vector machines (SVMs) and random forests, using text-derived features. More advanced alternatives may incorporate contextual or layout-aware transformer architectures to capture semantic and structural information embedded in document content. The classification taxonomy includes an unknown/unsupported category to prevent uncertain or out-of-scope documents from being forcibly assigned to predefined classes. For traceability and reproducibility, the predicted class, associated probability scores, and model version are recorded for each classification decision.

Field extraction is performed using document-specific schemas designed to accommodate the structural and semantic characteristics of each document category. Relevant attributes may include customer names or identifiers, document dates, customs declaration numbers, shipment or trip references, and associated transport identifiers. Each extracted field is linked to its corresponding visual location in the source document and, where applicable, its normalized representation. Normalization procedures may standardize whitespace, date formats, separators, and permitted letter-case variations; however, they must preserve semantically significant characters, particularly those contained in identifiers. Accordingly, each field is stored together with its original textual evidence, normalized value, extraction method, and associated confidence score, thereby supporting subsequent verification and auditability.

Record Validation and Exception Handling

Record validation is implemented through three complementary levels: structural, referential, and cross-field validation. Structural validation examines the presence of mandatory fields, character lengths, formatting constraints, and date plausibility. Referential validation queries authorized enterprise information systems to determine whether extracted identifiers correspond to existing customer, shipment, or customs declaration records. Cross-field validation evaluates the consistency of relationships among extracted attributes, including the association between a customs declaration and its corresponding shipment and the conformity of document dates to predefined operational sequences. Validation rules are version-controlled and designed to be independently testable. When a discrepancy is identified, the system generates a specific exception code accompanied by a contextual explanation.

Such discrepancies are not interpreted as definitive evidence of document errors, since inconsistencies may also arise from incomplete, outdated, or delayed reference data.

An exception-handling mechanism integrates model confidence with the severity of applicable business-rule violations to determine the appropriate processing pathway. Automatic filing is permitted only when predefined confidence thresholds are satisfied and all mandatory validation checks are successfully completed. Documents characterized by low-confidence predictions, unresolved referential relationships, or critical inconsistencies are directed to an authorized human reviewer. The reviewer may confirm an extracted value, correct an indexed field, revise the assigned document category, or initiate an investigation within the relevant source system. Each intervention is recorded in an audit trail containing the user's role, the timestamp, the original model prediction, the revised value, the justification for the change, and the applicable validation-rule version. Following appropriate quality assurance and access-control procedures, verified human corrections may subsequently be incorporated into a governed annotation dataset to support model refinement.

Archiving and System Interoperability

The proposed architecture separates the storage of original documents from relational metadata management and, where required, full-text search indexing. The system specification identifies Microsoft SQL Server or PostgreSQL as candidate relational database management systems, React as the user-interface technology, .NET as the backend development framework, and RESTful APIs as the primary mechanism for system integration. Elasticsearch is considered an optional component for full-text indexing and advanced retrieval. Persistent identifiers and explicit relational links connect original documents, processed page images, extracted fields, validation outcomes, and corresponding operational record identifiers. The retrieval mechanism supports authorized searches based on document category, date, customer, and shipment reference, while enabling users to filter validation exceptions and inspect relevant audit records.

Archive management incorporates role-based access control, the principle of least privilege, encryption of data in transit and at rest, predefined retention schedules, periodic backup restoration tests, and regular integrity verification. Personal data processing is designed in accordance with the principles of data minimization and purpose limitation, supported by documented procedures for data deletion and the management of legally justified retention exceptions. The Turkish Personal Data Protection Law (KVKK), the General Data Protection Regulation (GDPR), and ISO/IEC 27001 are used as regulatory and information security design references, rather than being presented as evidence of demonstrated legal compliance or certification.

To support secure interoperability, integration endpoints are designed to authenticate requesting systems, enforce restrictions on query scope, enable consistent document retrieval, and communicate processing status without unnecessarily exposing sensitive document content. Collectively, these architectural and operational measures are intended to support the security, traceability, integrity, and interoperability of document archiving and retrieval throughout the document lifecycle.

The Collection of the Data

A reliable evaluation requires a representative dataset of logistics documents collected under appropriate authorization and access-control arrangements. The sampling strategy should be stratified by document category, source organization, scan quality, page count, language, and layout family to capture relevant sources of variation. The use of actual customs declarations or customer-related documents must be supported by the necessary legal authorization, with pseudonymization applied where feasible and appropriate. A standardized annotation protocol should define each target field, acceptable equivalent formats, missing-value conditions, and procedures for resolving ambiguous labels. To assess annotation reliability, two annotators should independently label a quality-control subset, with disagreements subsequently adjudicated before the reference standard is established.

Dataset partitioning must account for dependencies among documents. In particular, related pages, duplicate documents, and repeated templates originating from the same source should be assigned to the same data partition wherever necessary to minimize information leakage between training and evaluation datasets. Where data availability permits, the dataset should be divided into training, development, and test subsets, supplemented by a temporally separated evaluation set to assess performance under changing operational conditions. The intended sample size, class distribution, inclusion and exclusion criteria, and document selection procedures should be specified in the study protocol and reported in the final evaluation. These elements constitute essential components of the research design and should not be omitted from the project plan.

For reproducibility, the experimental configuration should remain consistent across comparative evaluations, except where a controlled experimental variation is explicitly intended. Model versions, preprocessing parameters, inference hardware, software dependencies, and external service configurations should be systematically documented. Any deviations from the predefined experimental protocol should also be recorded and considered when interpreting the results.

Comparative Evaluation and Statistical Analysis

The proposed evaluation framework compares four approaches: (1) manual document indexing, (2) rule-based field extraction supported by optical character recognition (OCR), (3) machine learning-based document classification and field extraction supported by OCR, and (4) an integrated processing pipeline incorporating enterprise data reconciliation and selective human review. Where technically and operationally feasible, pattern-based extraction techniques and approaches that do not rely on conventional OCR pipelines may also be examined as alternative methods. Ablation studies are designed to investigate the individual contributions of image preprocessing, field-specific validation, and confidence-based decision-making to overall system performance.

Classification performance is assessed using per-class precision, recall, sensitivity, and macro-averaged F1-score. Field extraction is evaluated using exact-match accuracy and character error rate (CER), supplemented where appropriate by field-level accuracy after normalization. Exception detection is assessed using precision, recall, sensitivity, and false-positive rate. Additional operational measures include confidence calibration, the proportion of documents requiring manual review, end-to-end processing latency, archive retrieval performance, and document integrity. Cryptographic hash

verification and the completeness of source-system and processing logs are used to assess the integrity and traceability of archived records.

To ensure a fair comparison, the evaluated approaches should be assessed using consistent data partitions, comparable document characteristics, and predefined review procedures. Differences in task allocation, preprocessing, and human intervention should be explicitly documented because they may influence the interpretation of comparative results. Critical fields that remain unresolved must be referred for human review, and a selective prediction or abstention mechanism should be considered to prevent uncertain outputs from proceeding automatically through the workflow. Where sample sizes are sufficient, performance estimates should be accompanied by confidence intervals. Results should also be stratified by document category and scan quality to reveal performance differences that may be obscured by aggregate metrics. This analysis provides a more comprehensive basis for assessing the proposed framework's predictive reliability, operational suitability, and effectiveness in supporting logistics document archiving and verification.

RESULTS

This study establishes a proposed architecture and evaluation framework for AI-powered document processing and validation in customs and logistics operations. Expected outputs include structured digital records linked to source documents, validation results, human inspection decisions, and traceable archive entries. These outputs aim to support document retrieval, data integrity verification, exception handling, and auditability throughout the document lifecycle.

Since the system has not yet been implemented and evaluated using an independently validated test dataset or pilot application, no empirical claims are made regarding data extraction accuracy, processing time improvements, cost savings, or error reduction. Instead, the proposed evaluation framework defines the necessary metrics to assess system performance in subsequent implementation and benchmarking phases. Table 1 summarizes the evaluation dimensions, proposed criteria, and their intended interpretations.

Table 1. Proposed evaluation dimensions, metrics, and interpretation criteria

Evaluation Dimension	Proposed Evaluation Metrics	Interpretation
Data Extraction	Field-level exact-match accuracy, character error rate (CER)	Accuracy of extracting critical data fields and indexed values
Document Classification	Macro avarege F1 score, confusion matrix	Classification performance across heterogeneous document types
Data Validation	Precision, recall, false-positive rate	Effectiveness of enterprise cross-checks extracted data
Workflow performance	Manual review rate, correction time, 95th-percentile (p95) processing latency	Operational usefulness and safe automation
Archival integrity and retrieval	Retrieval time; hash-based integrity verification; audit-trail completeness	Document accessibility, integrity and traceability

DISCUSSION AND CONCLUSION

The main contribution of this study is the proposal of an integrated validation approach that correlates digitized document evidence, extracted data, authoritative corporate system records, rule-based validation results, and human decisions. This approach differs from workflows where successful OCR output alone is considered sufficient for successful digital archiving. Operationally, the proposed design is expected to reduce repetitive data entry and facilitate targeted examination of ambiguous documents. However, these benefits are not empirically validated results but hypotheses that need to be tested in the future. The digital enhancement targets defined within the project also correspond to planning criteria, not empirical findings, and are therefore not presented as study results.

The proposed framework also considers the types of errors and implementation risks that may arise during system operation. OCR methods may misidentify visually similar characters; class imbalance may reduce the classification performance of underrepresented document types; and new document templates may negatively impact model performance due to changes in data distribution. Furthermore, errors in corporate master data may lead to misleading validation alerts. Cloud-based processing processes may pose additional risks in terms of data location and confidentiality. While human oversight can help prevent some errors, it can also prolong processing times and lead to inconsistencies in corrections. Managing these risks requires regular model drift analyses, governance of validation rules, monitoring of threshold values, rollback mechanisms, and incident management processes. Furthermore, the applicability of the proposed approach should be systematically evaluated for conditions outside the examined declaration formats and the selected organizational environment.

From a broader perspective, digital storage of documents has the potential to reduce physical document circulation and the need for unnecessary reprinting. Effective search and access mechanisms can also facilitate access to document evidence needed in audit processes. However, environmental benefits and cost advantages depend on the IT infrastructure used, the growth of storage volume, the amount of documents, and changes made in workflows. Net emission reductions or resource efficiency increases cannot be concluded without lifecycle analyses or operational measurements. Therefore, the proposed framework should be considered not as a completed impact assessment, but as an empirically testable system investigation design.

In conclusion, this study presents a modular and AI-powered document lifecycle framework for logistics operations, encompassing image acquisition, structured data extraction, operational verification, exception handling, and secure digital archiving. The fundamental principle of the framework is that for a document's recognized value to become operationally meaningful, its source, level of uncertainty, and consistency with institutional evidence must be assessable. Accordingly, the study proposes a design that integrates established OCR and document understanding approaches with logistics-specific verification mechanisms and auditable human oversight.

Future study should include creating a labeled dataset in accordance with appropriate governance principles, comparing models with controlled experiments, calibrating thresholds for referral to human review, and conducting a prospective pilot application encompassing measurable processing times and correction workloads. Furthermore, security testing, validation of data deletion and retention policies, and evaluation of system performance in the face of changing document structures are necessary. It is

recommended that the results regarding the actual operational and technical impacts of the study be based on repeatable and measurable experimental findings, alongside the projected project benefits.

References

Akyüz R, Şahinaslan E, Şahinaslan Ö, 2026. An AI-enabled automation framework to increase efficiency and accuracy in international logistics transit declaration processes. Boletim da Sociedade Paranaense de Matemática, 44 (5): 1–11. https://doi.org/10.5269/bspm.80085

Devlin J, Chang MW, Lee K, Toutanova K, 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT, pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423

European Parliament and Council, 2016. Regulation (EU) 2016/679 (General Data Protection Regulation). Official Journal of the European Union, L119: 1–88. https://eur-lex.europa.eu/eli/reg/2016/679/oj

Huang Y, Lv T, Cui L, Lu Y, Wei F, 2022. LayoutLMv3: Pre-training for document AI with unified text and image masking. Proceedings of the 30th ACM International Conference on Multimedia, pp. 4083–4091. https://doi.org/10.1145/3503161.3548112

International Organization for Standardization and International Electrotechnical Commission, 2022. ISO/IEC 27001:2022: Information security, cybersecurity and privacy protection—Information security management systems—Requirements. ISO, Geneva.

Kim G, Hong T, Yim M, Nam J, Park J, Yim J, Hwang W, Yun S, Han D, Park S, 2022. OCR-free document understanding transformer. Computer Vision—ECCV 2022. https://doi.org/10.1007/978-3-031-19815-1_29

Republic of Türkiye, 2016. Law No. 6698 on the Protection of Personal Data (KVKK). Official Gazette No. 29677, 7 April 2016.

Smith R, 2007. An overview of the Tesseract OCR engine. Proceedings of the Ninth International Conference on Document Analysis and Recognition, pp. 629–633. https://doi.org/10.1109/ICDAR.2007.4376991

Predictive Performance and Stability of Machine Learning Classifiers for Real-Rooted Polynomials Across Degrees

Başak Karakaş[1,2*], Mehmet Göktürk[1,2]

[1]Institute of Artificial Intelligence, Gebze Technical University, Gebze, Türkiye;
[2]Department of Computer Engineering, Gebze Technical University, Gebze, Türkiye

*Corresponding author e-mail: basakkarakas@gtu.edu.tr

Abstract

Identifying whether all roots of a real-coefficient polynomial are real is a mathematical classification problem that can be approached using coefficient-based machine learning. This study evaluates the predictive performance and ranking stability of Support Vector Machines (SVM), Random Forest, XGBoost, Multilayer Perceptron (MLP), and an autoencoder-based classifier (AE-C), and examines how representative an earlier single-split evaluation is under repeated stratified cross-validation. Experiments use balanced datasets of 6,000 polynomials for each of degrees three, four, and five, represented by L2-normalized monomial coefficients. Five repetitions of ten-fold cross-validation produce 750 model evaluations. Within each fold, all five classifiers share the same partitions of 4,800 training, 600 validation, and 600 test examples. Previously selected model configurations are retained, while models are trained afresh in every fold and learned preprocessing uses training data only. SVM achieves the highest mean F1 scores: 0.9296 ± 0.0107, 0.9314 ± 0.0101, and 0.9085 ± 0.0099 for degrees three, four, and five, respectively, where variation is reported as fold standard deviation. SVM ranks first in 44, 48, and 48 of the 50 evaluations per degree. However, the complete single-split model ranking is reproduced in only one, one, and zero evaluations, respectively. At degree five, Random Forest and XGBoost have nearly identical mean F1 scores, with 24 and 25 paired wins and one tie. All classifiers have lower mean F1 at degree five than at degree four. These descriptive results distinguish stable leading-model performance from sensitive lower-rank comparisons, conditional on the previously selected configurations and dependent evaluations of the same datasets.

Keywords: *Real-Rooted Polynomials, Machine Learning, Repeated Cross-Validation, Predictive Performance, Ranking Stability*

INTRODUCTION

A polynomial is real-rooted if all its roots are real, including multiplicities. Root counting and isolation are classical tools in computational real algebraic geometry (Basu et al., 2006). Coefficient-based learning poses an empirical question about how reliably classifiers can infer this property from

coefficient vectors. Here, the inputs are L2-normalized monomial coefficients. The output distinguishes real-rooted polynomials (RRP) from polynomials with at least one non-real root (Non-RRP).

Machine learning is evaluated here on a specified data distribution. It does not provide mathematical certification. SVM and tree ensembles offer distinct nonlinear classification mechanisms (Cortes and Vapnik, 1995; Breiman, 2001; Chen and Guestrin, 2016). Multilayer perceptrons and autoencoder-based representations provide neural alternatives (Goodfellow et al., 2016; Hinton and Salakhutdinov, 2006). Shared partitions allow comparison of both average predictive performance and the stability of model ordering.

Our earlier submitted manuscript, "Comparative Classification and Shared Error Patterns in Real-Rooted Polynomials," reports a single-split evaluation of the same datasets. The acceptance decision at ICOSIN is pending, and the manuscript has not been published. Those experiments provide the earlier F1 scores and validation-selected model configurations used here. A single partition represents one allocation of training, validation, and test examples. It cannot show how performance estimates or relative rankings vary across allocations.

Five repetitions of stratified ten-fold cross-validation examine partition variability at degrees three, four, and five. All classifiers share the same partitions, with fresh training and fold-specific preprocessing. Overlapping training sets and reuse of examples make the evaluations dependent (Bengio and Grandvalet, 2004). Performance and ranking stability are therefore described without treating the repetitions as independent experiments.

The comparison covers the predictive performance of SVM, Random Forest, XGBoost, MLP, and AE-C on real-rooted polynomial classification at degrees three, four, and five. It examines how performance and relative rankings change across degrees. It also assesses the stability of performance estimates, leading models, complete rankings, and pairwise advantages across repeated data partitions. The earlier single-split results are compared with the repeated stratified cross-validation distributions to assess their representativeness.

The contribution is a shared-partition assessment of five fixed classifier configurations across three polynomial degrees. It links score distributions with ranking stability and paired model advantages, and assesses how well the earlier single split represented performance and ordering. The evaluation protocol is extended without introducing a new classifier or claiming improvement in the underlying models.

MATERIAL AND METHODS

2.1. Polynomial Datasets and Feature Representation

For degree d ∈ {3, 4, 5}, a real-coefficient polynomial is written as

$$p(x) = a_0 + a_1 x + \cdots + a_d x^d, \; a_d \neq 0.$$

Its features are the monomial coefficients in ascending order. The coefficient vector and its per-example L2 normalization are

$$a = (a_0, a_1, \ldots, a_d)^T, \; \tilde{a} = \frac{a}{\|a\|_2}, \; \|a\|_2 = \sqrt{\sum_{k=0}^{d} {a_k}^2}.$$

This nonzero scalar normalization preserves the roots. D3, D4, and D5 contain 6,000 polynomials each, represented by 4, 5, and 6 coefficients, respectively. Each dataset has 3,000 real-rooted polynomials (RRP; positive label 1) and 3,000 non-real-rooted polynomials (Non-RRP; negative label 0). RRP means that all roots are real, including multiplicities. The same archived datasets were used in the earlier single-split evaluation. No new polynomial data were generated. These datasets have been released by Başak Karakaş on Zenodo (version 1.0.0; Karakaş, 2026).

The available generator constructs RRP from real linear factors and Non-RRP from real linear factors and conjugate-pair quadratic factors. Coefficients are L2-normalized, and labels are assigned by construction rather than by a root-solver tolerance. The exact original generation/split seeds and export settings cannot be recovered. The stored datasets are reused.

2.2. Machine Learning Classifiers

The five classifiers retained the degree-specific configurations selected by validation performance in the earlier single-split evaluation. No new hyperparameter search was performed. Each fold fitted fresh models to 4,800 training examples and reserved 600 validation examples for neural checkpoint selection. Only SVM used additional feature-wise standardization, fitted on the corresponding training subset. The other models used the stored L2-normalized coefficients without additional scaling. Training used seed 42 on CPU with one thread.

In the formulations below, x_i is the stored normalized coefficient vector and $y_i \in \{0,1\}$ is its label. The feature count is $q=d+1$, and n denotes the number of examples in a batch or evaluation partition. The objectives describe the implemented losses for the architectures and degree-specific settings reported here. Neural optimization used Adam (Kingma and Ba, 2015).

SVM. An RBF-kernel support vector classifier used C=100 and gamma=1 at all degrees, with StandardScaler, tolerance 0.001, shrinking enabled, and probability estimation disabled. The soft-margin formulation (Cortes and Vapnik, 1995) uses standardized inputs u_i and signed labels $t_i=2y_i-1$:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i$$

$$s.t.\, t_i(w^T\varphi(u_i)+b) \geq 1-\xi_i,\ \xi_i \geq 0.$$

$$K(u_i,u_j) = exp(-\gamma\|u_i-u_j\|^2),\ C=100, \gamma=1.$$

The penalty sums hinge-loss slacks rather than taking a sample mean with the same numerical C. The intercept is unpenalized.

RF. Bootstrap forests used Gini impurity. For degrees 3/4/5, the numbers of trees were 300/100/100, maximum depths were unrestricted/20/20, and minimum leaf sizes were 4/1/1. Feature sampling considered all coefficients at degree 3 and the square root of the feature count at degrees 4–5. The minimum split size was two. At a node v, Gini impurity and the local split decrease are (Breiman, 2001):

$$G(v) = 1-\sum_{c=0}^{1} {p_c}^2,\ \Delta G = G(v) - \frac{N_L}{N_v}G(L) - \frac{N_R}{N_v}G(R).$$

Here the p values are node class proportions, and N denotes bootstrap-weighted sample counts. Each tree greedily chooses a feasible split that maximizes the decrease. This is a split criterion rather than a neural loss minimized by gradient descent.

XGBoost. Binary logistic boosting used histogram trees, log-loss evaluation, learning rate 0.1, and maximum depth six. For degrees 3/4/5, tree counts were 300/100/100, row subsampling fractions were 0.8/0.8/1.0, and column subsampling fractions were 1.0/0.8/1.0. Minimum child weights were 3/1/1 and L2 regularization strengths were 5/1/1. Fixed-configuration fitting used no boosting early stopping. With margin s and probability σ(s), the logistic loss is

$$\ell(y,s) = log(1+e^{s}) - ys,\ \ \sigma(s) = \frac{1}{1+e^{-s}}.$$

For tree f_t with T_t leaves and leaf weights w_j, the stagewise regularized objective on the sampled training rows I_t is (Chen and Guestrin, 2016):

$$J_t(f) = \sum\nolimits_{i\in I_t} \ell(y_i, F_{t-1}(x_i) + f(x_i)) \ + \frac{\lambda}{2}\sum\nolimits_{j=1}^{T_t} {w_j}^2.$$

The histogram algorithm greedily approximates this summed objective to second order, then applies shrinkage $F_t=F_{t-1}+0.1f_t$. It does not solve the full ensemble objective exactly. For degrees 3/4/5, λ=5/1/1, reg_alpha=0, and gamma=0. L1 and leaf-count penalties are inactive. The configured log-loss metric differs from the summed training objective. No validation-based boosting early stopping was used.

MLP. The input dimension was d+1, followed by hidden layers of widths 128–64–32 and one sigmoid output. Each hidden layer used batch normalization, ReLU, and dropout 0.1. Adam minimized binary cross-entropy at learning rate 0.001, with batch sizes 128/64/128 for degrees 3/4/5. Training ran for at most 50 epochs, stopping after ten epochs without strict lexicographic improvement in validation F1 then accuracy. The best checkpoint was restored, with ties retaining the earlier epoch. Positive predictions required probability >0.5. For $p_i=\sigma(g\theta(x_i))$, the mean binary cross-entropy is (Goodfellow et al., 2016):

$$L_{BCE} = -\frac{1}{n}\sum\nolimits_{i=1}^{n} [y_i log\ p_i + (1-y_i) log(1-p_i)].$$

The probability-space implementation uses PyTorch BCE with numerical log clamping. Adam updates minimize training loss. Validation F1 then accuracy selects the retained epoch, which is a different criterion.

AE-C. The encoder had widths (d+1)–64–32–2 and the mirrored decoder 2–32–64–(d+1). Hidden activations were ReLU, with linear latent and reconstruction outputs. Classifier hidden widths were [8]/[16,8]/[8] for degrees 3/4/5, with ReLU, dropout 0.1, and one output logit. Each fold performed fresh reconstruction pretraining followed by joint encoder/decoder/classifier fine-tuning. Both stages used Adam, learning rates 0.001/0.001/0.0005, and batch sizes 64/128/64. Pretraining minimized MSE for at most 100 epochs, with patience 10 and validation improvement $>10^{-5}$. Fine-tuning minimized αMSE+βBCE-with-logits with (α,β)=(1,1)/(1,1)/(0.5,2), for at most 60 epochs with patience 15. Validation F1 then accuracy selected the restored checkpoint, with ties retaining the earlier epoch. Sigmoid probability ≥0.5 denoted the positive class. No historical trained weights were loaded. The

encoder–decoder representation follows the autoencoder framework (Hinton and Salakhutdinov, 2006). The implemented joint training objective is specified below.

$$z_i = E_\theta(x_i) \in \mathbb{R}^2,\ \ \hat{x}_i = D_\phi(z_i),\ \ s_i = H_\psi(z_i).$$

$$L_{rec} = \frac{1}{nq}\sum_{i=1}^{n}\sum_{j=1}^{q}(x_{ij} - \hat{x}_{ij})^2,\ \ L_{cls} = \frac{1}{n}\sum_{i=1}^{n}\ell(y_i, s_i).$$

$$L_{joint} = \alpha L_{rec} + \beta L_{cls},\ \ (\alpha, \beta) = (1,1)/(1,1)/(0.5,2).$$

Reconstruction pretraining updates only encoder and decoder using Lrec. Joint fine-tuning updates encoder, decoder, and classifier using Ljoint. BCE-with-logits evaluates ℓ through a numerically stable softplus formulation. Reconstruction MSE averages over examples and q coefficients. Both terms are reduced independently before weighting, so batch losses use means rather than sums. Validation losses use a fixed network in evaluation mode. Adam weight decay is zero, and dropout provides the stated architectural regularization.

2.3. Experimental Design and Evaluation Protocol

For each degree, the archived training, validation, and test arrays were pooled in that order into 6,000 examples, with 3,000 examples per class. RepeatedStratifiedKFold(n_splits=10, n_repeats=5, random_state=42) generated five shuffled stratified ten-fold repetitions. Each outer test fold contained 600 examples (300 per class). From the remaining 5,400 examples, train_test_split reserved 600 stratified validation examples using random_state=10000+100d+i. Here i=0,…,49 is the split index within a degree. This left 4,800 training examples (2,400 per class), 600 validation examples (300 per class), and 600 test examples (300 per class).

Figure 1 summarizes the experimental evaluation workflow.

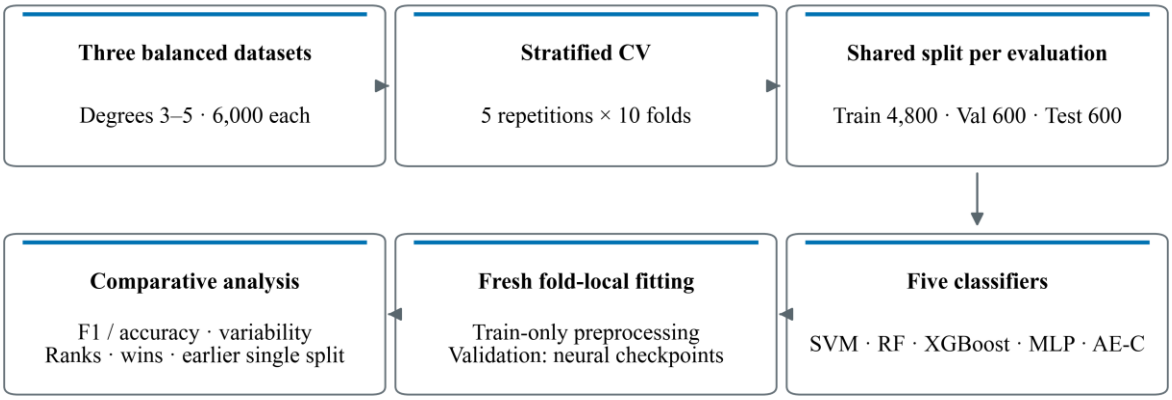


Figure 1. Experimental evaluation workflow. The same stratified partitions are used for all five classifiers; preprocessing and model fitting are confined to the training subset, and neural checkpoint selection uses the validation subset.

All five classifiers shared the same saved indices. Models were freshly initialized and fitted in every fold, and AE-C repeated both pretraining and fine-tuning. StandardScaler was fitted only to that fold's training data. Neural batch-normalization statistics were also learned only during training. Validation data controlled neural early stopping and checkpoint selection. Classical models retained fixed settings and used validation scores only for diagnostics. No validation refit was performed. Outer test data were excluded from preprocessing fitting, gradient updates, and epoch/configuration selection within the fold.

Python, NumPy, and PyTorch seeds were reset to 42 for each neural fit and stage. Shuffled minibatches used a seed-42 generator. Classical random_state was 42, and computation used CPU with one thread. A single outer random-state stream produced five distinct repetitions rather than regenerating an identical ten-fold split five times. Every example appeared in exactly one test fold per repetition. The design yielded 50 evaluations per model and degree, giving 3×5×50=750 model evaluations.

The earlier single-split evaluation used the same archived datasets and 4,800/600/600 partition sizes. The earlier submitted manuscript remains under review and unpublished. Previously selected configurations were fixed for this evaluation. Their original selection used the same data source, so fold-local isolation does not constitute nested cross-validation or remove historical tuning. Original single-split test examples re-entered the pooled data. Overlapping training sets and repeated use of the same polynomials make the evaluations dependent. Fold SDs and paired counts are descriptive, not evidence of independent replication (Bengio and Grandvalet, 2004).

2.4. Performance Measures and Comparative Analysis

Positive-class F1 and overall accuracy were calculated on each test partition, with RRP (label 1) as the positive class. For each model and degree, the K=50 evaluations were summarized by their mean and sample standard deviation (ddof=1). The SD describes variability across dependent folds. It is neither a standard error nor a confidence interval.

Ranking stability. Models were ranked separately by F1 and accuracy within each degree and test partition. Each model attaining the maximum received one first-place count, including shared first places. Identical leader sets required the complete sets of maximizing models to match, rather than merely sharing one leader. Identical full orders required all ten pairwise relations (higher, lower, or tied) to agree between F1 and accuracy. Recurrence of the previous single-split F1 order required its complete strict ordering, without ties. In Figure 3, a group of g tied models spanning g rank positions contributed 1/g per model to each occupied position. These fractional rank frequencies differ from the inclusive first-place counts in Table 2.

Pairwise comparisons. For two models A and B evaluated on the same test partition, directional wins W and ties T were defined using the indicator function $\mathbb{1}$:

$$W_{AB} = \sum_{i=1}^{K} \mathbb{1}(F1_{A,i} > F1_{B,i}), \quad T_{AB} = \sum_{i=1}^{K} \mathbb{1}(F1_{A,i} = F1_{B,i}).$$

Losses for A equal WBA, and WAB+WBA+TAB=50. For each degree, the off-diagonal entry (A,B) of Table 3 is WAB. Diagonal entries are omitted. Opposite cells sum to 50 minus ties. Accuracy comparisons used the same construction. Ranking and paired-score ties were determined from exact rational scores calculated from integer confusion counts, without rounding or a numerical tolerance.

Earlier single-split comparison. For an earlier F1 score h, its empirical percentile within the corresponding 50-score distribution was

$$P(h) = \frac{100}{K}[\sum_{i=1}^{K} \mathbb{1}(F1_i < h) + \frac{1}{2}\sum_{i=1}^{K} \mathbb{1}(F1_i = h)].$$

Each tied score contributes half a count. The calculation used unrounded F1 scores recomputed from confusion counts. Percentiles describe positions in the observed distribution, not external-test performance or improvement over the earlier evaluation.

Statistical interpretation. Repetitions reuse the same polynomials and have overlapping training sets. Means, SDs, ranking frequencies, empirical percentiles, and pairwise counts are descriptive summaries conditional on the previously selected configurations. They do not establish statistical significance or independent replication. The 50 evaluations cannot be treated as independent experiments (Bengio and Grandvalet, 2004).

RESULTS

3.1. Comparative Predictive Performance

Predictive performance was assessed in 750 model evaluations, comprising 50 shared test partitions per classifier and degree. Each test partition contained 600 polynomials with balanced classes. Table 1 summarizes the means and sample standard deviations.

Table 1. Predictive performance across polynomial degrees

Model	Degree 3	Degree 4	Degree 5
SVM	F1: 0.9296 ± 0.0107 Acc.: 0.9263 ± 0.0120	F1: 0.9314 ± 0.0101 Acc.: 0.9275 ± 0.0113	F1: 0.9085 ± 0.0099 Acc.: 0.9027 ± 0.0112
RF	F1: 0.8988 ± 0.0104 Acc.: 0.8938 ± 0.0116	F1: 0.9103 ± 0.0114 Acc.: 0.9061 ± 0.0123	F1: 0.8829 ± 0.0100 Acc.: 0.8776 ± 0.0108
XGBoost	F1: 0.9071 ± 0.0120 Acc.: 0.9031 ± 0.0129	F1: 0.9121 ± 0.0113 Acc.: 0.9079 ± 0.0121	F1: 0.8829 ± 0.0095 Acc.: 0.8768 ± 0.0102
MLP	F1: 0.9001 ± 0.0125 Acc.: 0.8930 ± 0.0144	F1: 0.9032 ± 0.0123 Acc.: 0.8953 ± 0.0143	F1: 0.8872 ± 0.0111 Acc.: 0.8773 ± 0.0128
AE-C	F1: 0.9101 ± 0.0151 Acc.: 0.9063 ± 0.0164	F1: 0.9067 ± 0.0126 Acc.: 0.9015 ± 0.0139	F1: 0.8692 ± 0.0158 Acc.: 0.8602 ± 0.0171

F1 = 2TP/(2TP + FP + FN), with real-rooted polynomials as the positive class; Acc. = (TP + TN)/600. Values are means ± sample SD (ddof = 1) across 50 dependent evaluations per degree and model; SD is not a standard error.

SVM achieved the highest mean F1 and accuracy at every degree. Both metrics identified the same leading classifier but ordered some closely performing alternatives differently. RF exceeded MLP in mean accuracy at degrees three and five, whereas MLP had the higher mean F1. These differences are consistent with precision–recall trade-offs. F1 and accuracy need not produce the same ordering, even on balanced test sets.

3.2. Performance Across Polynomial Degrees and Single-Split Comparison

All five classifiers had lower mean F1 and accuracy at degree five than at degree four. From degree three to four, both metrics increased for SVM, RF, XGBoost, and MLP, whereas AE-C declined. SVM retained its leading position across the three degrees. AE-C moved from second to fifth in mean F1, while MLP rose from fifth at degree four to second at degree five. Figure 2 shows the score distributions together with the archived single-split F1 values. Comparisons across degrees involve different polynomial datasets and degree-specific configurations.

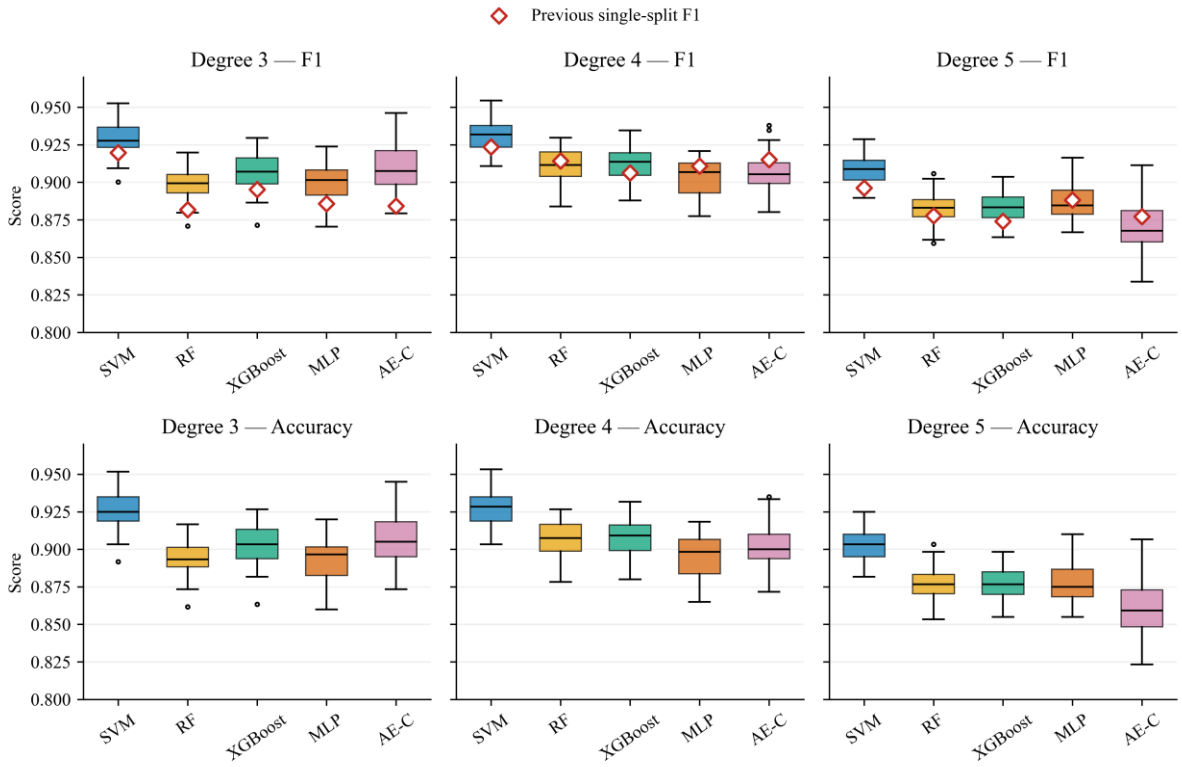


Figure 2. F1 and accuracy distributions across models and degrees. Each box summarizes 50 evaluations; lines show medians, boxes show the interquartile range, whiskers extend to the most extreme observations within 1.5 interquartile ranges, and points indicate observations beyond the whiskers. Red-outlined diamonds in the F1 panels mark the earlier single-split evaluation of the same polynomial datasets; accuracy panels show CV distributions only.

The earlier single-split evaluation used the same L2-normalized coefficient datasets (6,000 polynomials per degree), positive-class F1 definition, and 4,800/600/600 training/validation/test sizes. Its validation-selected configurations are fixed in the 5×10 evaluation, with fresh models and partitions rather than a new search. Single-split F1 positions ranged from the 4th–16th percentiles at degree three, 26th–80th at degree four, and 12th–66th at degree five. Every value lay within the corresponding observed CV range. SVM's positions were the 16th, 26th, and 12th percentiles. Percentiles count CV scores below the archived score plus half the ties, divided by 50. The single-split full F1 order recurred only 1/50, 1/50, and 0/50 times (Table 2). The archived split did not consistently represent the ordering of the other models. These same-source comparisons describe partition sensitivity conditional on historical tuning, not model improvement or independent external testing. Repeated folds remain dependent.

3.3. Model Ranking and Performance Stability

Table 2 contrasts stability of the leading model with agreement of the complete model ordering across evaluation metrics and with the previous single-split F1 order.

Table 2. Ranking stability across evaluation metrics

Degree	SVM first: F1	SVM first: Acc.	Same leader set	Same full order	Prior F1 order
3	44	43	46	25	1
4	48	48	49	34	1
5	48	49	49	29	0

All counts are out of 50. First-place counts include shared first places. "Same leader set" requires identical sets of top-scoring models; "same full order" requires identical pairwise relations, including ties. "Prior F1 order" requires the previous single-split strict order without ties. Ties were determined from exact rational scores derived from confusion counts.

SVM's leadership was more stable than the full ordering of the five models. F1 and accuracy produced identical leader sets in most evaluations, but their complete rankings agreed in only about half to two-thirds of evaluations. The archived single-split F1 order was rarely reproduced. A stable preferred classifier did not imply stable relative positions for the remaining models. Figure 3 shows their spread across rank positions, with tied groups contributing equally to the positions they occupy.

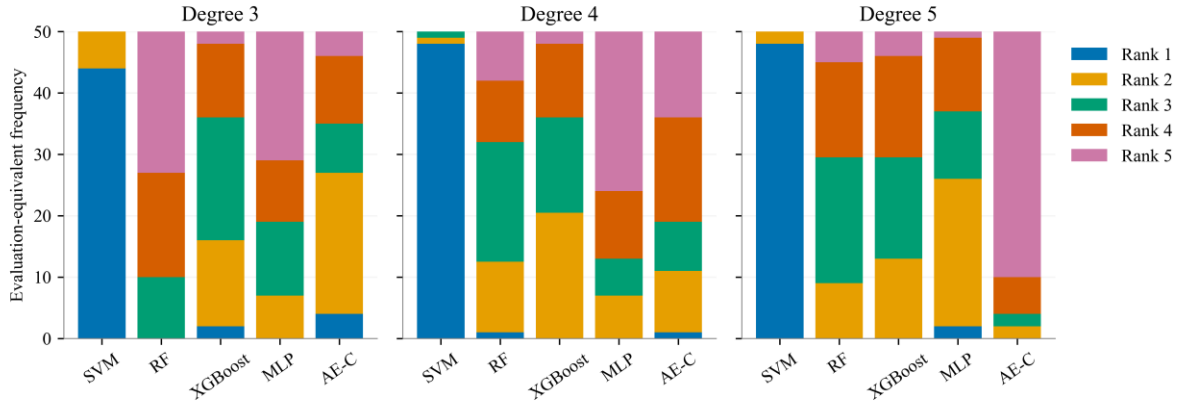


Figure 3. Model rank frequencies across polynomial degrees, based on F1. Colors identify ranks 1–5; stacked segments summarize 50 evaluations. Ties contribute fractional frequencies across their occupied positions; each model's segments therefore sum to 50.

The rank-frequency profiles distinguish average performance from stability. SVM concentrated at first place, while RF, XGBoost, MLP, and AE-C occupied several ranks. The close RF–XGBoost comparison at degree five and the changing positions of the neural classifiers across degrees motivate paired comparisons on the same test partitions.

3.4. Pairwise Performance Comparisons

Table 3 presents all ten unique F1 comparisons per degree as directional win matrices. In all 30 model pairs, the classifier with the higher mean F1 also won more paired evaluations. Its win count ranged from only 25 to all 50 evaluations. A higher mean was therefore associated with varying levels of paired consistency.

Table 3. Pairwise F1 win counts across polynomial degrees

(a) Degree 3						(b) Degree 4						(c) Degree 5					
Row	**S**	**R**	**X**	**M**	**A**	**Row**	**S**	**R**	**X**	**M**	**A**	**Row**	**S**	**R**	**X**	**M**	**A**
S	—	50	48	50	46	**S**	—	49	49	50	49	**S**	—	50	50	48	50
R	0	—	4	21	12	**R**	1	—	19	36	31	**R**	0	—	24	17	42
X	2	46	—	34	20	**X**	1	30	—	38	35	**X**	0	25	—	18	45
M	0	29	16	—	10	**M**	0	14	12	—	18	**M**	2	33	32	—	47
A	4	38	30	40	—	**A**	1	19	15	32	—	**A**	0	8	5	3	—

Each off-diagonal cell counts the evaluations (out of 50) in which the row model had strictly higher positive-class F1 than the column model on the same test partition. Diagonals are not applicable (—). Opposite cells sum to 50 minus the number of ties; equal scores contribute to neither cell. Ties are determined from exact rational scores derived from confusion counts. Panels (a)–(c) correspond to degrees three–five. Counts describe dependent repeated evaluations, not independent experiments or statistical significance. Labels for both rows and columns: S=SVM, R=RF, X=XGB (XGBoost), M=MLP, A=AE-C.

Figure 4 makes the directional contrasts in Table 3 visible on a common color scale.

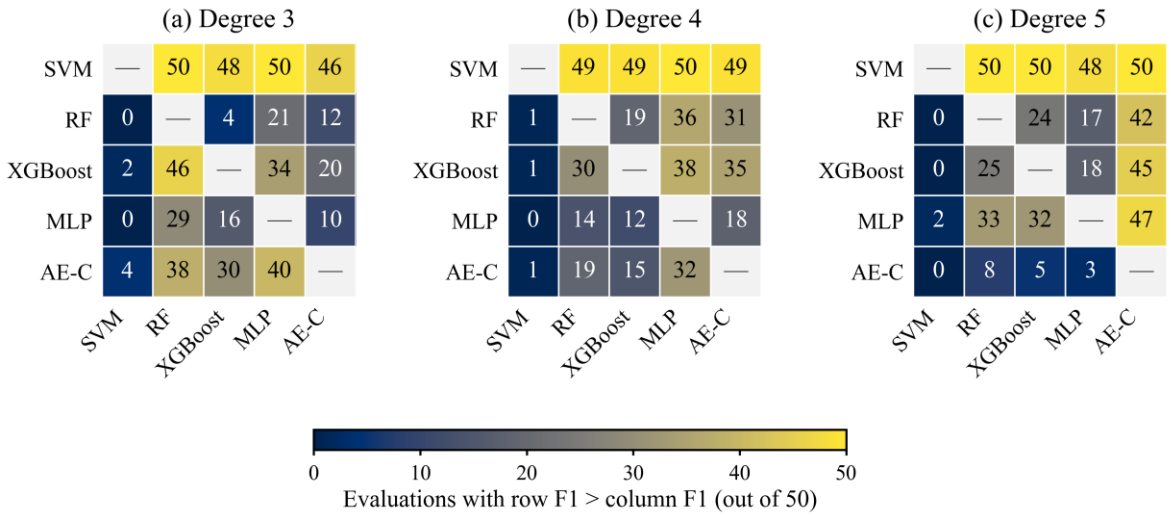


Figure 4. Pairwise F1 win-count heatmaps for (a) degree 3, (b) degree 4, and (c) degree 5. Cells reproduce Table 3: row-model wins over the column model, out of 50 paired evaluations. Diagonals are omitted (—); opposite cells exclude ties. All panels share the 0–50 scale.

SVM exceeded each competitor in 46–50, 49–50, and 48–50 evaluations at degrees three, four, and five, respectively. At degree three, only XGBoost and AE-C defeated SVM, in two and four evaluations. At degree four, RF, XGBoost, and AE-C each did so once. At degree five, SVM won all evaluations against RF, XGBoost, and AE-C, losing only twice to MLP. Its average advantage was accompanied by a persistent paired advantage rather than a few isolated high scores.

Relations among the other models changed more often. Both models won at least once in eight, nine, and seven of the ten F1 pairs at degrees three, four, and five. AE-C's wins over XGBoost fell from 30 to 15 to five, while its wins over RF fell from 38 to 19 to eight. MLP exceeded RF in 29 evaluations at degree three, 14 at degree four, and 33 at degree five. At degree four, XGBoost exceeded RF in 30 evaluations versus 19 losses and one tie. A higher mean score was not uniformly reproduced across partitions.

At degree five, the near-zero mean RF−XGBoost F1 difference (−0.000027) corresponded to 24 RF wins, 25 XGBoost wins, and one tie. Accuracy instead favored RF on average (+0.000800), with 24 wins, 22 losses, and four ties. The direction also differed by metric for RF–MLP. At degree three, RF won 21 F1 comparisons versus MLP's 29, but won 25 accuracy comparisons versus 23 (two ties). At degree five, the counts were 17 versus 33 for F1 and 25 versus 21 for accuracy (four ties). MLP exceeded XGBoost by F1 in 32 evaluations at degree five, and all three alternatives exceeded AE-C in at least 42. The matrices reveal both near-balanced pairs and more persistent advantages among the four non-SVM models.

These comparisons are descriptive and conditional on the previously selected configurations. Repetitions reuse the same polynomial datasets, and their training sets overlap. Model initialization

variability is not separately sampled. Neither win counts nor fold SDs establish statistical significance, independent replication, or causal improvement over the archived single-split evaluation.

DISCUSSION AND CONCLUSION

SVM provided the strongest predictive performance under the evaluated configurations, leading mean F1 and accuracy at every degree (Table 1). Agreement between the metrics supports its consistent leading position within these datasets. The other models were less uniformly ordered. RF and MLP, for example, exchanged positions depending on whether F1 or accuracy was used. Balanced class counts do not make the metrics interchangeable. These results favor SVM in this experimental setting without establishing a universal advantage over other classifier families.

Performance did not change monotonically with polynomial degree. Most models improved from degree three to four, whereas all five had lower mean F1 and accuracy at degree five than at degree four. AE-C's declining relative position contrasted with MLP's rise to second place in mean F1 at degree five. The comparisons involve different datasets, input dimensions, and some different hyperparameters. They cannot isolate degree as a causal source of difficulty or establish that a larger coefficient vector alone explains the degree-five decrease.

Stable leadership coexisted with less stable complete rankings. SVM's first-place frequencies and persistent paired advantages show that its leading mean scores were not attributable solely to a few favorable partitions (Tables 2–3). The remaining models occupied several rank positions, and some pairwise advantages frequently reversed. At degree five, the near-equal RF and XGBoost mean F1 scores corresponded to 24 RF wins, 25 XGBoost wins, and one tie. Accuracy instead gave 24 RF wins, 22 XGBoost wins, and four ties. Small mean differences should be interpreted alongside paired counts and metric choice. Neither these counts nor fold SDs establish statistical significance.

Every earlier single-split F1 score fell within its corresponding repeated-evaluation range (Figure 2). The degree-three scores occupied relatively low positions, while the degree-four and degree-five positions were more dispersed. SVM remained the leading classifier on average, but the complete earlier F1 ordering recurred only once, once, and never across the three degrees. The earlier partition captured broad comparisons more consistently than the ordering of closely performing alternatives. Its position within the repeated distributions describes partition sensitivity. It is neither an independent external validation nor evidence that repeated evaluation improved the classifiers.

The scope is limited to balanced synthetic datasets for degrees three through five. The generator's sampling distribution may restrict generalization to broader polynomial populations. The exact original generation seeds and export settings are unavailable. Previously selected hyperparameters were retained without nested optimization. Their same-source selection remains a limitation despite fold-local preprocessing and checkpoint selection. Repeated evaluations are dependent, and model initialization variability was not separately sampled. The reported variability is descriptive and does not support independent-fold confidence intervals or significance claims (Bengio and Grandvalet, 2004).

Future work could examine alternative data-generation strategies and the geometry of normalized coefficient space as separate research questions. The main contribution is a common repeated-partition evaluation that distinguishes robust leading-model behavior from sensitive comparisons among the other

models. Within these datasets, SVM's consistent advantage was clearer than the complete ordering of the five models. The earlier single split did not fully represent that variability.

References

Basu S, Pollack R, Roy MF, 2006. Algorithms in real algebraic geometry. 2nd edition. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-33099-2

Bengio Y, Grandvalet Y, 2004. No unbiased estimator of the variance of k-fold cross-validation. Journal of Machine Learning Research, 5: 1089–1105. https://jmlr.org/papers/v5/grandvalet04a.html

Breiman L, 2001. Random forests. Machine Learning, 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324

Chen T, Guestrin C, 2016. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. https://doi.org/10.1145/2939672.2939785

Cortes C, Vapnik V, 1995. Support-vector networks. Machine Learning, 20 (3): 273–297. https://doi.org/10.1007/BF00994018

Goodfellow I, Bengio Y, Courville A, 2016. Deep learning. MIT Press. https://www.deeplearningbook.org/

Hinton GE, Salakhutdinov RR, 2006. Reducing the dimensionality of data with neural networks. Science, 313 (5786): 504–507. https://doi.org/10.1126/science.1127647

Karakaş B, 2026. Real-Rooted Polynomial Classification Datasets: Degrees 3, 4 and 5. Version 1.0.0 [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.23319343.

Kingma DP, Ba JL, 2015. Adam: A method for stochastic optimization. International Conference on Learning Representations. https://arxiv.org/abs/1412.6980

Artificial Intelligence Capacity and Environmental Sustainability in OECD Countries: A Cross-Country Assessment

Özge Taş[1], İsmail Öylek[2],

[1]Kastamonu University, Catalzeytin , Çatalzeytin Vocational School, Computer Technologies Kastamonu, Türkiye
[2]Sakarya University of Applied Sciences, Sakarya Vocational School, Department of Computer Technologies, Sakarya, Türkiye.

*Corresponding author e-mail: otas@kastamonu.edu.tr

Abstract

This study examined the relationship between artificial intelligence (AI) capacity and environmental sustainability across OECD countries. Data on AI adoption, AI conference publications, AI patents, private AI investment, the carbon intensity of electricity generation, the share of renewable electricity, and energy intensity were compiled for 38 OECD countries from the OECD, Stanford HAI/CSET, Ember, and international energy data sources. These indicators were standardized to construct an AI Capacity Index and an Environmental Sustainability Index. An AI–Decarbonization Gap indicator was also calculated to compare growth in AI adoption with the pace of decarbonization over the 2021–2025 period. The findings revealed no statistically significant association between the AI Capacity Index and the Environmental Sustainability Index ($\rho = -0.111$, $p = 0.531$). Likewise, environmental sustainability did not differ significantly between countries with high and low AI capacity ($U = 145.0$, $p = 1.000$). The temporal analysis also found no statistically significant association between growth in AI adoption and the pace of decarbonization ($\rho = -0.273$, $p = 0.197$). Taken together, these findings suggest that greater AI capacity does not necessarily go hand in hand with stronger environmental sustainability across OECD countries.

Keywords: *Artificial intelligence capacity, environmental sustainability, decarbonization, composite indices.*

INTRODUCTION

Rapid advances in artificial intelligence (AI) have accelerated digital transformation across manufacturing, finance, healthcare, transport, energy and public services. As AI capacity expands, attention has increasingly turned to the environmental implications of the energy and computing infrastructure that supports it, alongside its economic and technological benefits. Training and deploying large AI models require substantial computational resources, raising concerns about electricity consumption, carbon emissions and resource use. The Green AI approach responds to these concerns by proposing that AI systems be evaluated in terms of computational cost and resource efficiency as well as model performance.

AI has a dual relationship with environmental sustainability. It can support sustainability goals through intelligent energy management, electricity grid optimization, climate forecasting, resource efficiency and environmental monitoring. Yet the data centers, high-performance processors and cooling systems needed to develop and operate AI also consume considerable energy and natural resources. The OECD distinguishes between these direct environmental impacts and the potential indirect benefits of AI applications, while highlighting gaps in the data and standards needed to measure AI's environmental footprint.

These concerns have become more pressing as demand for AI computing grows. According to the International Energy Agency, global electricity consumption by data centers reached approximately 415 TWh in 2024, accounting for around 1.5% of worldwide electricity use. The expansion of AI applications is increasing the use of accelerated servers and the power density of data centers. AI's environmental footprint also extends beyond energy consumption and carbon emissions. Water use has become an increasingly important concern because of the cooling requirements of data centers running large models and the water consumed indirectly through electricity generation.

A growing body of research examines the relationship between digitalization and environmental sustainability at the country level. In a study of 36 OECD countries, Karlilar et al. (2023) found that digitalization, green innovation and renewable energy use could support environmental sustainability. However, the findings become more complex when studies move from broad measures of digitalization to indicators specific to AI. Using panel data from 67 countries, Wang et al. (2024) reported that AI has multidimensional effects on ecological footprints, carbon emissions and energy transitions. Similarly, Zhong et al. (2024) examined 66 countries and found that the environmental effects of AI could vary with national conditions and levels of development.

Recent studies also suggest that the relationship between AI and the environment is neither uniformly linear nor consistent in direction. Luo et al. (2026) identified a U-shaped relationship between AI and environmental sustainability in OECD countries. Their findings suggest that AI may initially be associated with poorer environmental performance, with this relationship reversing beyond a certain level of development. Using data from 36 countries over 2017–2023, Huang and Gozgor (2026) found positive associations between some dimensions of AI vibrancy and renewable energy consumption. However, structural factors, including income levels and government effectiveness, were stronger predictors. Together, these findings suggest that the environmental implications of AI capacity depend on energy infrastructure and national structural conditions as well as technological development.

Despite these advances, three gaps remain in the literature. First, many studies represent AI through a single indicator or a broad measure of digitalization, although national AI capacity encompasses adoption, scientific output and technological innovation. The OECD.AI Index, published in 2026, likewise emphasizes the need for a multidimensional assessment of national AI ecosystems. Second, most studies focus on a single environmental outcome, such as carbon emissions or ecological footprints. Comparative assessments that jointly consider renewable electricity share, electricity carbon intensity and energy intensity remain limited. Third, relatively few studies directly compare the pace of AI development with the decarbonization of national energy systems.

To address these gaps, this study examines the relationship between AI capacity and environmental sustainability across OECD countries using a multidimensional framework. AI capacity is measured through a composite index combining business AI adoption, AI conference publications and AI patents. Environmental sustainability is assessed using renewable electricity share, the carbon intensity of electricity generation and energy intensity. The study also develops an AI–Decarbonization Gap

indicator to compare growth in AI adoption with the pace of energy-system decarbonization. This approach examines both the extent of national AI capacity and whether its development is accompanied by a more sustainable energy system.

Table 1. Selected studies examining the relationship between artificial intelligence, digitalization and environmental sustainability

Study	Data	Method	Main Findings	Relevance to the Present Study
Schwartz et al. (2020)	AI research	Conceptual framework / Green AI	Argues that AI evaluation should consider computational costs and resource efficiency alongside model performance.	Provides a theoretical basis for assessing AI capacity together with its environmental costs.
OECD (2022)	AI computing and applications	Measurement and policy framework	Distinguishes between direct and indirect environmental effects of AI and highlights gaps in AI-specific measurement standards and data.	Supports the study's multidimensional approach to environmental assessment.
Karlilar et al. (2023)	36 OECD countries, 2000–2018	System GMM	Finds that digitalization, green innovation and renewable energy support environmental sustainability.	Provides a comparative basis for moving from broad digitalization measures to AI-specific capacity indicators.
Zhong et al. (2024)	66 countries	Cross-country panel analysis	Finds that the relationship between AI and carbon emissions varies with national characteristics.	Supports the view that AI's environmental effects may differ across countries.
Wang et al. (2024)	67 countries	System GMM and a dynamic panel threshold model	Identifies multidimensional effects of AI on ecological footprints, carbon emissions and energy transitions.	Supports assessing AI and environmental performance through multiple outcome indicators.
Dhiman et al. (2024)	Literature on AI and sustainability	Systematic review / literature review	Emphasizes the need to consider energy consumption and other environmental costs alongside AI's contributions to sustainability.	Strengthens the theoretical distinction between "AI for sustainability" and "the sustainability of AI."

Huang and Gozgor (2026)	36 countries, 2017–2023	Panel analysis with Driscoll–Kraay estimation	Finds positive associations between some dimensions of AI vibrancy and renewable energy consumption, while national structural characteristics are stronger predictors.	Suggests that the relationship between AI capacity and renewable energy depends on national context.
Luo et al. (2026)	OECD countries, 1993–2023	Nonlinear panel analysis	Identifies a U-shaped relationship between AI and environmental sustainability.	Suggests that the AI–sustainability relationship may be nonlinear and cannot be captured by a single overall effect.
Present study	38 OECD countries; cross-sectional data and temporal data for 2021–2025	Composite indices, Spearman correlation, Mann–Whitney U tests and the AI–Decarbonization Gap	Finds no statistically significant association between AI capacity and environmental sustainability. AI development and decarbonization may progress at different rates.	Jointly assesses AI capacity and environmental sustainability using multidimensional measures and introduces the AI–Decarbonization Gap indicator.

As the table shows, the literature offers no consistent conclusion about the environmental effects of digitalization and AI. Some studies find that technological development supports environmental sustainability, while others suggest that the relationship varies with national characteristics, energy infrastructure and the level of AI development. The present study adds to this literature by measuring AI capacity through a multidimensional index rather than a single variable and examining it alongside indicators of energy-system sustainability.

MATERIAL AND METHODS

Material

The study uses indicators of AI capacity and environmental and energy sustainability for 38 OECD member countries. Data on business AI adoption were obtained from the OECD ICT Access and Usage by Businesses database. Indicators of scientific output and technological innovation in AI came from Stanford HAI/CSET Global AI Vibrancy sources, while private AI investment data were drawn from the Stanford AI Index 2025 report. Electricity-generation carbon intensity and the share of renewables in total electricity generation were obtained from Ember's Yearly Electricity Data. Energy intensity, used as a proxy for energy efficiency, was compiled from international energy data sources.

Country identifiers were standardized using ISO-3 codes, and all indicators were combined into a common country-level dataset. Changes in AI adoption were examined over 2021–2025, while cross-sectional comparisons used the most recent years for which comparable observations were available

across countries. Missing observations were retained as missing rather than treated as zero, and no imputation or other missing-data completion procedures were applied.

Methods

The Collection of the Data

Data were compiled for 38 OECD member countries from several international sources. AI adoption indicators were obtained from the OECD database, AI publications and patents from Stanford HAI/CSET sources, and private AI investment figures from the Stanford AI Index 2025 report. Electricity carbon intensity and renewable electricity shares were drawn from Ember's Yearly Electricity Data, while energy intensity was obtained from international energy data sources. Records from these sources were matched using ISO-3 country codes and combined into a common dataset. Missing observations were retained without imputation.

Statistical Analysis

The statistical analysis began with an examination of the dataset's overall structure and the calculation of descriptive statistics for each variable. AI publication and patent counts were transformed using $\ln(1+x)$ to address their right-skewed distributions. Indicators were then standardized using z-scores to allow comparisons across different measurement units. Following these preprocessing steps, composite indices were constructed to represent AI capacity and environmental sustainability. Countries were classified into four AI–sustainability profiles based on the median values of the two indices.

Associations between AI capacity and environmental sustainability indicators were examined using Spearman's rank correlation coefficient. The Benjamini–Hochberg false discovery rate adjustment was applied to account for multiple correlation tests. Environmental sustainability was then compared between countries with high and low AI capacity using the Mann–Whitney U test. Finally, the AI–Decarbonization Gap indicator was calculated by comparing the compound annual growth rate of AI adoption with the pace of electricity-sector decarbonization over 2021–2025. Statistical significance was assessed at $p < 0.05$.

RESULTS

Table 1 presents descriptive statistics for AI capacity and environmental and energy sustainability indicators across OECD countries. The AI Capacity Index was calculated for 34 countries, with a mean of 0.047 ± 0.859. The Environmental Sustainability Index covered 38 countries, with a mean of 0.000 ± 0.674. The large standard deviations in AI conference publications and patent counts indicate substantial differences in scientific and technological AI capacity across countries. Similarly, the wide ranges in electricity carbon intensity and renewable electricity shares reflect considerable variation in OECD countries' energy systems.

Table 2. Descriptive statistics on key research variables

Variable	n	Mean	SD	Median	Min.	Max.
AI adoption rate (%)	25	15,886	7,954	13,890	4,420	32,910
AI conference publications	36	713,58	1456,94	219	7	8513
AI patents	31	848,32	2655,02	27	1	12764

AI Capacity Index	34	0,047	0,859	0,012	-1,350	2,375
Electricity carbon intensity (gCO_2/kWh)	38	236,97	168,57	201	25	608
Renewable electricity share (%)	38	54,74	26,30	53,25	9,71	100
Energy intensity	38	3,234	1,793	2,940	0,970	11,930
Environmental Sustainability Index	38	0,000	0,674	0,040	-1,274	1,238

AI–Sustainability Profiles of OECD Countries

Countries were classified into four analytical profiles using the median values of the AI Capacity Index and the Environmental Sustainability Index. Nine countries fell into the high AI capacity–high sustainability group, eight into the high AI capacity–low sustainability group, eight into the low AI capacity–high sustainability group, and nine into the low AI capacity–low sustainability group (Table 3).

Table 3. Artificial intelligence capacity and environmental sustainability profiles of OECD countries

Profile	**n**	**Countries**
High AI capacity–High sustainability	9	Austria, Denmark, Finland, France, Luxembourg, the Netherlands, Spain, Sweden, the United Kingdom
High AI capacity–Low sustainability	8	Australia, Belgium, Canada, Germany, Italy, Japan, South Korea, the United States
Low AI capacity–High sustainability	8	Chile, Ireland, Latvia, Lithuania, New Zealand, Norway, Portugal, Switzerland
Low AI capacity–Low sustainability	9	Czechia, Estonia, Greece, Hungary, Israel, Mexico, Poland, Slovenia, Türkiye

Figure 1 shows countries' positions on the two indices. The distribution indicates that countries with high AI capacity do not share a single environmental sustainability profile. Denmark, Finland and the United Kingdom fall within the high AI capacity–high sustainability group, whereas the United States, Japan, South Korea and Germany belong to the high AI capacity–low sustainability group. By contrast,

Norway, Switzerland and Portugal are classified as having high environmental sustainability despite their lower AI capacity.

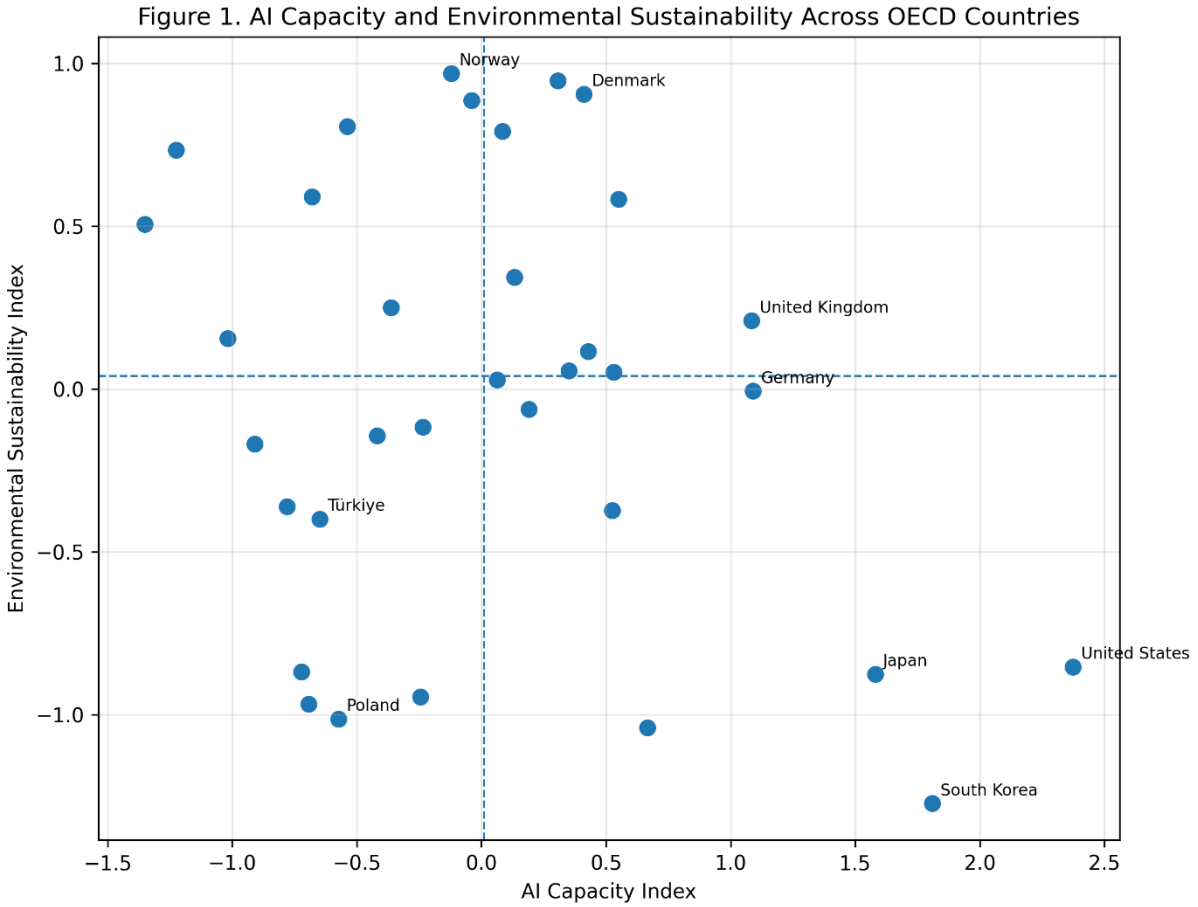


Figure 1. Distribution of OECD countries by the AI Capacity Index and the Environmental Sustainability Index.

Associations Between AI Capacity and Environmental Sustainability

Associations between AI indicators and environmental sustainability indicators were assessed using Spearman's rank correlation coefficient. The AI Capacity Index showed a weak negative association with the Environmental Sustainability Index, although this was not statistically significant ($\rho = -0.111$, $p = 0.531$). Its associations with electricity carbon intensity ($\rho = 0.021$, $p = 0.906$), renewable electricity share ($\rho = -0.142$, $p = 0.423$) and energy intensity ($\rho = 0.092$, $p = 0.605$) were also not statistically significant (Table 4).

Table 4. Spearman correlations between AI indicators and environmental sustainability indicators.

AI Indicator	Environmental Indicator	n	ρ	p	BH-FDR q

AI adoption rate	Environmental Sustainability Index	25	0.251	0.227	0.692
AI conference publications	Environmental Sustainability Index	36	−0.211	0.216	0.692
AI patents	Environmental Sustainability Index	31	−0.124	0.508	0.707
AI Capacity Index	Electricity carbon intensity	34	0.021	0.906	0.981
AI Capacity Index	Renewable electricity share	34	−0.142	0.423	0.707
AI Capacity Index	Energy intensity	34	0.092	0.605	0.744
AI Capacity Index	Environmental Sustainability Index	34	−0.111	0.531	0.707

None of the associations examined were statistically significant after the Benjamini–Hochberg false discovery rate adjustment.

Comparison of Environmental Sustainability by AI Capacity

Countries were divided into high- and low-AI-capacity groups based on the median AI Capacity Index, with 17 countries in each group. The median Environmental Sustainability Index was 0.052 in the high-AI-capacity group and −0.116 in the low-AI-capacity group. The Mann–Whitney U test indicated no statistically significant difference between the groups ($U = 145.0$, $p = 1.000$). The rank-biserial correlation coefficient of 0.003 also indicated a negligible difference.

The groups likewise showed no statistically significant differences in electricity carbon intensity ($U = 127.5$, $p = 0.570$), renewable electricity share ($U = 145.0$, $p = 1.000$) or energy intensity ($U = 160.5$, $p = 0.593$). Figure 2 presents the distributions of the Environmental Sustainability Index for the two groups.

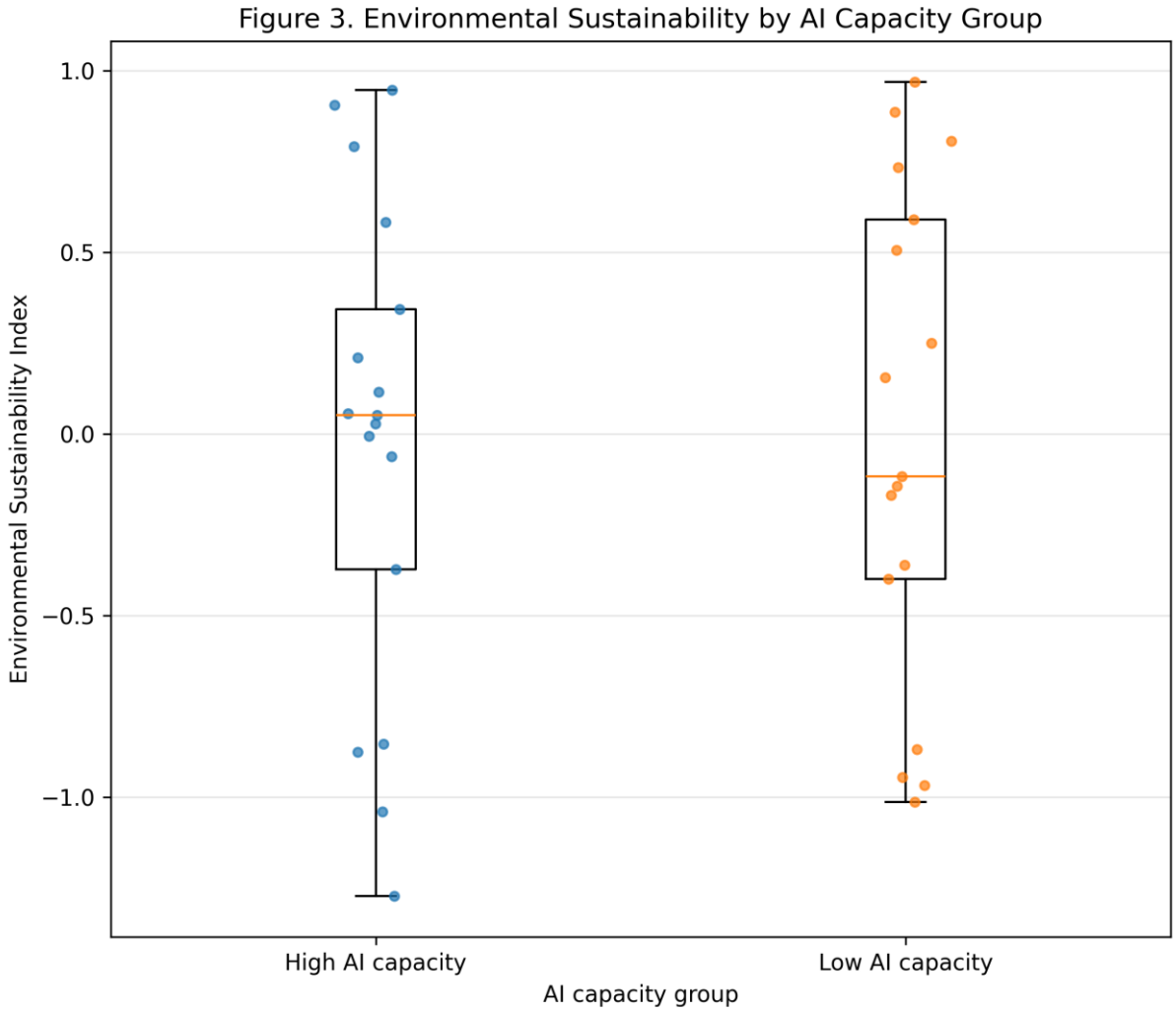


Figure 2. Distribution of the Environmental Sustainability Index among OECD countries with high and low AI capacity.

AI Development and the Pace of Decarbonization

The comparison of AI adoption growth and energy-system decarbonization included 24 countries with the required observations for both 2021 and 2025. The compound annual growth rate of AI adoption showed a negative but statistically nonsignificant association with the reduction in electricity carbon intensity ($\rho = -0.273$, $p = 0.197$).

AI–Decarbonization Gap values varied considerably across countries. Of the 24 countries examined, 15 recorded positive values and nine recorded negative values. Estonia had the highest positive gap (3.073), followed by Norway (1.743), Czechia (1.559), Austria (1.228) and Türkiye (1.207). The lowest values were observed in Denmark (−3.465), Finland (−2.876) and Portugal (−2.566) (Figure 4).

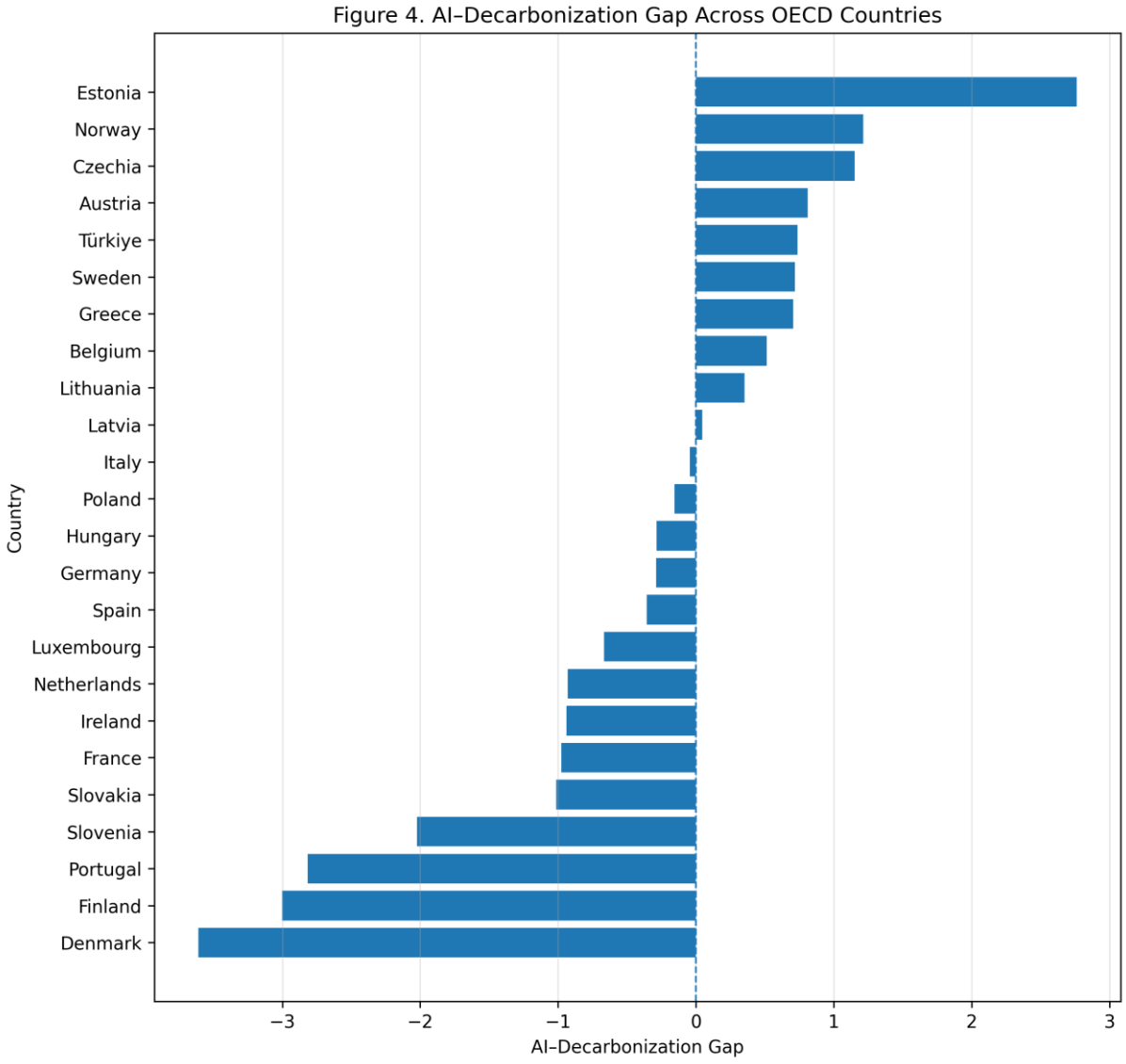


Figure 4. AI–Decarbonization Gap values across OECD countries, 2021–2025.

Positive AI–Decarbonization Gap values indicate that a country's relative growth in AI adoption outpaced its electricity-system decarbonization performance. Negative values indicate stronger relative progress in decarbonization than in AI adoption. However, no statistically significant positive association between these two transitions was found across the OECD countries examined.

DISCUSSION AND CONCLUSION

The main finding of this study is that AI capacity and environmental sustainability do not necessarily develop in parallel across OECD countries. The absence of a statistically significant association between the AI Capacity Index and the Environmental Sustainability Index, together with the lack of significant environmental differences between countries with high and low AI capacity, suggests that technological

capacity alone is insufficient to explain a sustainable energy transition. Country profiles reinforce this interpretation. Some countries combine high AI capacity with high environmental sustainability, while others exhibit similarly high AI capacity alongside lower environmental performance.

This finding is relevant to the distinction between "AI for sustainability" and "the sustainability of AI" emphasized in the Green AI literature. Although AI can help address sustainability challenges, developing and operating AI systems requires substantial energy and computational resources. The International Energy Agency expects data-center electricity consumption to rise rapidly in the coming years, with AI applications contributing substantially to this growth. Assessments of national AI capacity should therefore consider the supporting energy infrastructure alongside adoption, publications, patents and investment.

The AI–Decarbonization Gap indicator developed in this study adds a temporal perspective to this relationship. Across the countries examined, AI adoption growth was not significantly associated with the pace of energy-system decarbonization. This finding suggests that digital transformation and energy transition do not automatically advance together. Supporting rapid AI adoption with low-carbon electricity generation and improvements in energy efficiency is therefore important for its environmental sustainability.

Several limitations should be acknowledged. AI indicators were not consistently available for all OECD countries in the same years, energy intensity data referred to an earlier year than the other indicators, and AI investment data covered a limited number of countries. In addition, the environmental indicators describe national energy systems rather than directly measuring AI-specific energy consumption. Future research could provide a more detailed assessment by incorporating direct measures such as data-center electricity consumption, AI-specific computing capacity and data-center carbon intensity.

Overall, the findings suggest that greater AI capacity does not automatically coincide with stronger environmental sustainability across OECD countries. Countries with high AI capacity exhibit different environmental profiles, and no statistically significant association was found between the two dimensions.

These results support closer coordination between AI, energy and climate policies. A sustainable AI ecosystem requires technological capacity to develop on a foundation of low-carbon, energy-efficient infrastructure. The AI Capacity Index, Environmental Sustainability Index and AI–Decarbonization Gap used in this study offer a comparative framework for examining digital transformation and energy transition together across OECD countries.

References

Dhiman R, Miteff S, Wang Y, Ma SC, Amirikas R, Fabian B, 2024. Artificial intelligence and sustainability—A review. Analytics, 3 (1): 140-164.

Ember, 2026. Yearly Electricity Data. Access address: https://ember-energy.org/data/yearly-electricity-data/; Date of access: 02.10.2026.

Huang L, Pal S, Mahalik MK, Gozgor G, 2026. AI vibrancy and renewable energy consumption: A cross-country analysis. Sustainable Development, 1-13.

International Energy Agency, 2025. Energy and AI. Access address: https://www.iea.org/reports/energy-and-ai; Date of access: 02.10.2026.

Karlilar S, Balcilar M, Emir F, 2023. Environmental sustainability in the OECD: The power of digitalization, green innovation, renewable energy and financial development. Telecommunications Policy, 47 (6): 102568.

Luo H, Zhang L, Sun Y, Zhang L, 2026. Artificial intelligence and environmental sustainability: A nonlinear analysis of the load capacity factor in OECD countries. Scientific Reports, 16: 27249.

Organisation for Economic Co-operation and Development, 2022. Measuring the environmental impacts of artificial intelligence compute and applications: The AI footprint. Access address: https://www.oecd.org/en/publications/measuring-the-environmental-impacts-of-artificial-intelligence-compute-and-applications_7babf571-en.html; Date of access: 02.10.2026.

Organisation for Economic Co-operation and Development, 2026. ICT access and usage by businesses. Access address: https://data-explorer.oecd.org/; Date of access: 02.10.2026.

Our World in Data, 2026. Energy intensity of economies. Access address: https://ourworldindata.org/grapher/energy-intensity-of-economies; Date of access: 02.10.2026.

Schwartz R, Dodge J, Smith NA, Etzioni O, 2020. Green AI. Communications of the ACM, 63 (12): 54-63.

Stanford Institute for Human-Centered Artificial Intelligence, 2025. The AI Index 2025 Annual Report. Access address: https://hai.stanford.edu/ai-index/2025-ai-index-report; Date of access: 02.10.2026.

Stanford Institute for Human-Centered Artificial Intelligence, 2026. Global AI Vibrancy Tool. Access address: https://hai.stanford.edu/ai-index/global-vibrancy-tool; Date of access: 02.10.2026.

Wang Q, Li Y, Li R, 2024. Ecological footprints, carbon emissions, and energy transitions: The impact of artificial intelligence (AI). Humanities and Social Sciences Communications, 11: 1043.

Zhong J, Zhong Y, Han M, Yang T, Zhang Q, 2024. The impact of AI on carbon emissions: Evidence from 66 countries. Applied Economics, 56 (25): 2975-2989.

Acknowledgment

The author would like to acknowledge the publicly available databases and institutional data sources, including OECD, Stanford HAI/CSET, Ember, the International Energy Agency, and Our World in Data, which provided the data used in this study.

Conflict of Interest

The author has declared that there is no conflict of interest.

Author Contributions

The author was responsible for the conceptualization of the study, data collection and integration, methodological design, statistical analysis, interpretation of the findings, visualization, and preparation of the manuscript.

Target-Dependent Source Utility in Transfer Learning for SPT-Based Shear-Wave Velocity Prediction

Engin Nacaroğlu[1], Ayşe Tuba Tuğrul[1]*, Emine Aylin Acarlar[1]

[1] Institute of Science, Civil Engineering, Pamukkale University, Denizli, Türkiye

*Corresponding author; e-mail: aysetubat22@gmail.com

Abstract

Published geotechnical datasets can extend the evidence available for predicting shear-wave velocity (Vs) from standard penetration test (SPT) measurements, but their value depends on how well each source represents a target region. Because geology and measurement practices vary among regions, pooling datasets may obscure target-specific contributions. This paper evaluates three published regional sources for SPT-based Vs prediction in Afyonkarahisar and Denizli, Türkiye, across three local-data budgets. Source-specific and pooled transfer models are compared with a target-only reference, alongside equal-weight and target-adaptive ensembles of source-specific predictions. The target regions are analyzed separately, with locked-test locations reserved for final evaluation. Model tuning and adaptive ensemble weights use validation data from target-training locations only. Location-macro mean absolute error (MAE) is the primary metric. In Afyonkarahisar, the Kathmandu Valley dataset yields the lowest source-specific MAE at all budgets, reducing error relative to the target-only reference by 9.13–16.02 m/s; both ensembles also improve average performance. In Denizli, source utility is mixed: the best individual source improves on the reference at the medium budget, but not at the low or full budget, while pooled transfer has higher error at each budget. Ensemble performance likewise varies with target and budget. These findings show that source utility depends on the target region and available local data; they do not establish a general regional ranking. Evaluating sources individually can inform whether to transfer from a single dataset, pool sources, or combine source-specific predictions.

Keywords: *Regional Transfer, Source Selection, Domain Adaptation, Shear-Wave Velocity, Standard Penetration Test*

INTRODUCTION

Shear-wave velocity (Vs) is widely used in geotechnical and seismic site characterization. Its empirical relationship with standard penetration test (SPT) measurements has been studied for estimation purposes (Akın et al., 2011). Cross-region prediction is complicated by differences in geology, testing practice, and data coverage; machine-learning performance may also change under distribution shift (Koh et al., 2021; Payan et al., 2025).

Recent geotechnical machine-learning studies have explored targeted data augmentation to address data scarcity (Nacaroglu et al., 2026), while transfer learning offers a complementary approach by reusing knowledge from source domains; however, its effectiveness for cross-regional SPT-based shear-wave velocity prediction remains insufficiently characterized.

Transfer learning can reduce dependence on local observations, but source-target mismatch can lead to negative transfer (Wang et al., 2019). Therefore, a source that is useful for one target need not be useful for another.

Comparing a pooled-source model with a target-only reference cannot show whether each source contributes similarly or whether one source drives the pooled result. Accordingly, this study asks whether the utility of an individual regional source is target-specific and whether that utility changes as local target data increase. For each target and budget, seven conditions are compared: three source-specific transfer models for Koreageotec, Quito, and SAFER; one pooled-source transfer model; a target-only reference; an equal-weight ensemble; and a target-adaptive ensemble. Afyonkarahisar and Denizli are analyzed separately at low, medium, and full training-location budgets, with locked-test locations held out for final evaluation and all tuning and adaptive weighting confined to target-training data. Location-macro mean absolute error, averaged across ten paired training-seed repetitions, is the primary outcome. The study therefore evaluates practical source-selection choices under a common location-based protocol; it does not propose a new SPT–Vs equation, claim a universally best source, or infer a causal reason for regional differences.

MATERIAL AND METHODS

Data and Preprocessing

The source data comprised three published geotechnical datasets representing South Korea, Quito, and the Kathmandu Valley: Koreageotec (Choi et al., 2026a, 2026b), Quito (Othman et al., 2023a, 2023b), and SAFER (Gilder et al., 2019; De Risi et al., 2021). After the common validity filter, the source datasets contained 17,450, 214, and 99 modeling rows, respectively. Thus, the pooled source set was strongly imbalanced toward Koreageotec; a balanced-source pooling condition was not evaluated. Target data for both Afyonkarahisar and Denizli were compiled from borehole-based soil investigation reports. The Denizli compilation is documented in 92 reports (Acarlar, 2026). The prediction task used depth and each dataset's supplied standardized SPT resistance as inputs and measured shear-wave velocity as the outcome.

Source Conditions And Location-Based Evaluation

For each target and local-data budget, we compared seven conditions: three source-specific transfer models based on Koreageotec, Quito, and SAFER; one pooled-source transfer model using all three sources; a target-only reference; an equal-weight ensemble; and a target-adaptive ensemble. Each transfer model was adapted with the same local target-training locations and compared with a reference trained only on the corresponding target data. The two ensembles combined predictions from the three source-specific models. All conditions used the same multilayer perceptron architecture (2-32-16-1, ReLU hidden layers, and a linear output) with the Adam solver. Transfer models were first pretrained on the selected source condition using a location-grouped source validation split; the resulting weights initialized all layers, and all layers were updated during target fine-tuning without freezing. Target hyperparameters were selected from eight candidates using three-fold location-grouped cross-validation within the target-training locations. The same model family was used across conditions, and the two target regions were analyzed separately.

All observations from a location were kept in the same data partition. Afyonkarahisar had eight locked-test locations and training budgets of 6, 15, and 30 locations; Denizli had six locked-test locations and budgets of 5, 12, and 23 locations. The 92 Denizli reports formed the source compilation; after modeling-data processing, observations were grouped into 29 distinct locations (six locked-test and 23 training locations). Within each target, the budget memberships were selected once in a deterministic nested order: the low set is contained in the medium set, which is contained in the full set, and this membership was reused across all ten paired training-seed repetitions. Therefore, budget comparisons change both the number and identity of training locations; the reported standard deviations capture training-seed variation only and not uncertainty from resampling alternative location subsets. The target-only model was standardized using target-training data, whereas transfer models retained standardization parameters derived from their source data. The comparison therefore reflects the complete transfer protocol, not the isolated effect of transferred weights. Locked-test locations were excluded from model fitting and selection. Figure 1 summarizes the evaluation design.

Model tuning used cross-validation grouped by location and restricted to target-training data. Here, out-of-fold (OOF) predictions mean predictions for target-training locations produced by models that did not train on the corresponding location. Adaptive ensemble weights and the OOF source ranking were estimated from these location-grouped OOF predictions only. The final comparison used the locked-test locations. Because model tuning and weight estimation used the same target-training pool, the weighting step was not independently nested within a separate validation procedure.

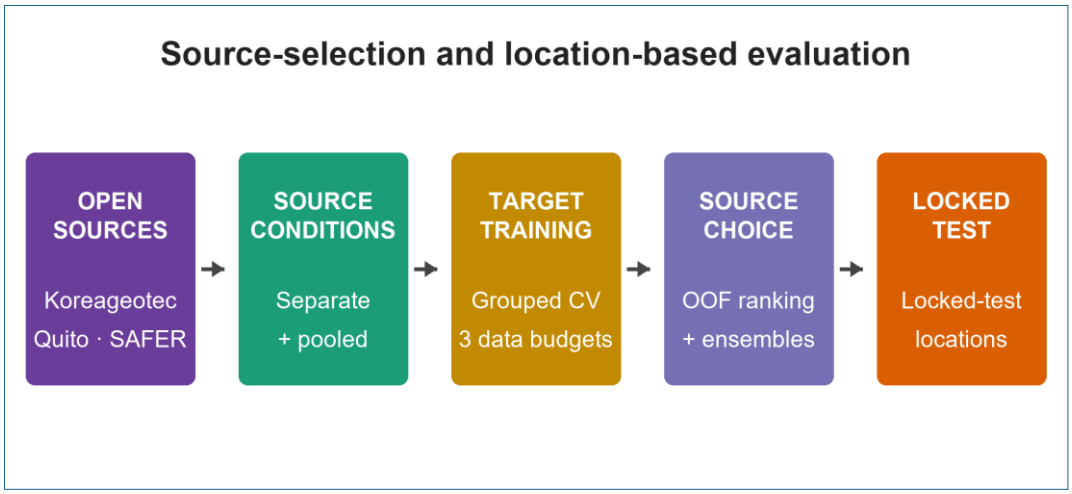


Figure 1. Source datasets are evaluated separately and pooled; fixed nested target-training budgets support model tuning and OOF source/ensemble assessment, while final evaluation uses locked-test locations.

Metrics

The primary metric was location-macro mean absolute error (MAE), calculated by averaging absolute errors within each locked-test location and then across locations. Source gain was calculated as target-only MAE minus the MAE of the evaluated source or ensemble; positive values indicate lower error than the reference. Negative-transfer rate is the proportion of paired training-seed repetitions in which a source condition had higher error than the target-only model. Means and standard deviations summarize

ten paired training-seed repetitions under one location-based split; the standard deviations are descriptive and are not confidence intervals.

RESULTS

Source-Specific And Pooled Performance

Table 1. Location-macro MAE (m/s; mean ± SD) by target and budget.

Target	Budget	Target-only	SAFER	Pooled	Equal ensemble	Adaptive ensemble
Afyonkarahisar	Low	105.01 ± 1.39	95.88 ± 3.66	102.90 ± 2.45	94.46 ± 2.28	95.23 ± 3.49
Afyonkarahisar	Medium	106.85 ± 3.84	90.83 ± 3.63	108.91 ± 1.77	93.95 ± 2.56	90.49 ± 2.99
Afyonkarahisar	Full	113.58 ± 2.11	104.08 ± 2.37	115.41 ± 1.16	109.90 ± 1.62	109.09 ± 1.80
Denizli	Low	52.75 ± 0.84	53.05 ± 0.59	65.14 ± 7.19	53.73 ± 2.38	54.53 ± 2.73
Denizli	Medium	60.29 ± 7.24	54.43 ± 2.29	73.61 ± 9.90	58.20 ± 4.66	58.93 ± 6.13
Denizli	Full	47.66 ± 2.48	49.57 ± 1.57	56.37 ± 5.86	52.69 ± 4.06	59.53 ± 8.52

Note. MAE = mean absolute error; SD = standard deviation. Values are mean ± SD across ten paired training-seed repetitions using one locked location-based test split.

SAFER is the lowest locked-test source-specific condition in every row of Table 1; this descriptive test result was not used to select a source during model training.

For Afyonkarahisar, SAFER had the lowest MAE among the source-specific conditions at every budget, with gains of 9.13–16.02 m/s over the target-only reference. Both ensembles also reduced average error, although adaptive weighting did not consistently outperform equal weighting.

For Denizli, the target-only reference had lower MAE at the low and full budgets. SAFER improved on it only at the medium budget, by 5.86 m/s; the pooled-source model had higher error than the reference at all three budgets. The medium-budget reference was also the least stable target-only condition (60.29 ± 7.24 m/s versus 52.75 ± 0.84 and 47.66 ± 2.48 m/s), so the isolated positive SAFER gain may partly reflect reference-model variability rather than a robust source benefit.

Afyonkarahisar target-only MAE increased from 105.01 to 113.58 m/s between the low and full budgets. Because the training-location subsets are fixed and nested, this pattern cannot be interpreted as a pure sample-size effect: it is also affected by which locations enter at each budget. On a fixed test set, a larger training budget did not guarantee lower error. Figure 2 summarizes source-specific and pooled gains across targets and budgets.

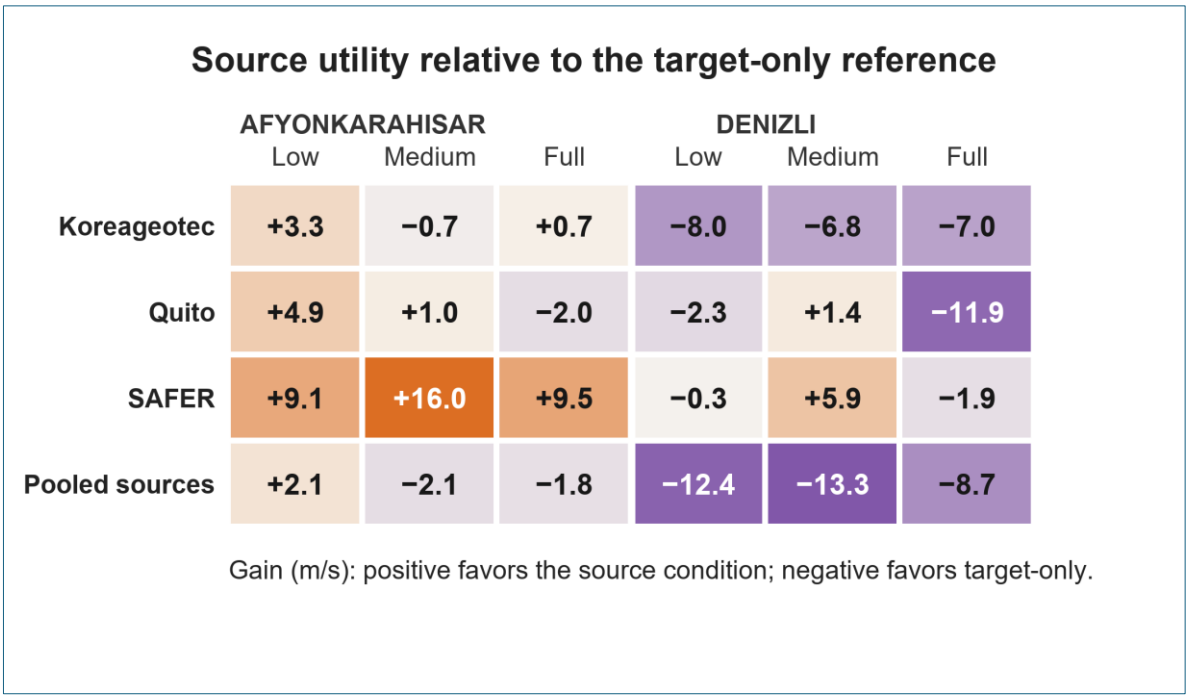


Figure 2. Heatmap of mean gain over the target-only reference for each source condition, target, and budget. Positive values favor the listed source condition; negative values favor the target-only reference. Values are averaged across ten paired training-seed repetitions.

Source Ranking And Negative Transfer

Table 2 reports the OOF-selected source and its locked-test gain at each budget, together with pooled gain and source-specific negative-transfer rates.

Table 2. OOF-selected source, locked-test gain, pooled gain, and negative-transfer rate by target and budget.

Target	Budget	OOF-selected source / locked-test gain (m/s)	Pooled gain (m/s)	SAFER negative-transfer rate	Pooled negative-transfer rate
Afyonkarahisar	Low	SAFER / 9.13	2.12	0.00	0.30
Afyonkarahisar	Medium	SAFER / 16.02	−2.06	0.00	0.70
Afyonkarahisar	Full	Koreageotec / 0.65	−1.83	0.00	0.90
Denizli	Low	SAFER / −0.30	−12.39	0.50	1.00
Denizli	Medium	Quito / 1.36	−13.32	0.20	0.80
Denizli	Full	Quito / −11.86	−8.70	0.70	1.00

Note. OOF = out-of-fold; MAE = mean absolute error. Gain is target-only MAE minus source-condition MAE (m/s), so positive values favor transfer. Negative-transfer rate is the proportion of paired training-seed repetitions with higher MAE than the target-only model.

The OOF-selected source varied by target and budget. At the full budget, OOF selection favored Koreageotec for Afyonkarahisar and Quito for Denizli, whereas SAFER had the lowest descriptive

locked-test source MAE in both regions. The table makes this distinction explicit: OOF selection uses target-training predictions, while the reported gain is calculated on locked-test locations.

Ensemble Analysis

Table 3 reports ensemble gains relative to the target-only reference and the adaptive weights estimated from target-training predictions.

Table 3. Ensemble gains relative to the target-only reference.

Target	Budget	Equal ensemble gain (m/s)	Adaptive ensemble gain (m/s)	Adaptive weights: Koreageotec / Quito / SAFER
Afyonkarahisar	Low	10.55	9.78	0.051 / 0.000 / 0.949
Afyonkarahisar	Medium	12.90	16.36	0.201 / 0.343 / 0.455
Afyonkarahisar	Full	3.68	4.49	0.443 / 0.184 / 0.373
Denizli	Low	−0.99	−1.78	0.488 / 0.000 / 0.512
Denizli	Medium	2.09	1.36	0.000 / 1.000 / 0.000
Denizli	Full	−5.03	−11.86	0.000 / 1.000 / 0.000

Note. Ensemble gain is target-only MAE minus ensemble MAE (m/s), so positive values favor the ensemble. Gains are averaged across ten paired training-seed repetitions; adaptive weights are listed as Koreageotec / Quito / SAFER and sum to one.

Adaptive weights varied across targets and budgets. At Denizli's medium and full budgets, the training predictions assigned all ensemble weight to Quito, but this ensemble did not achieve the lowest locked-test MAE at either budget. These weights describe a training-data preference, not guaranteed test-set performance. Figure 3 shows the weights; Table 2 reports the same OOF-versus-locked-test distinction across all three budgets.

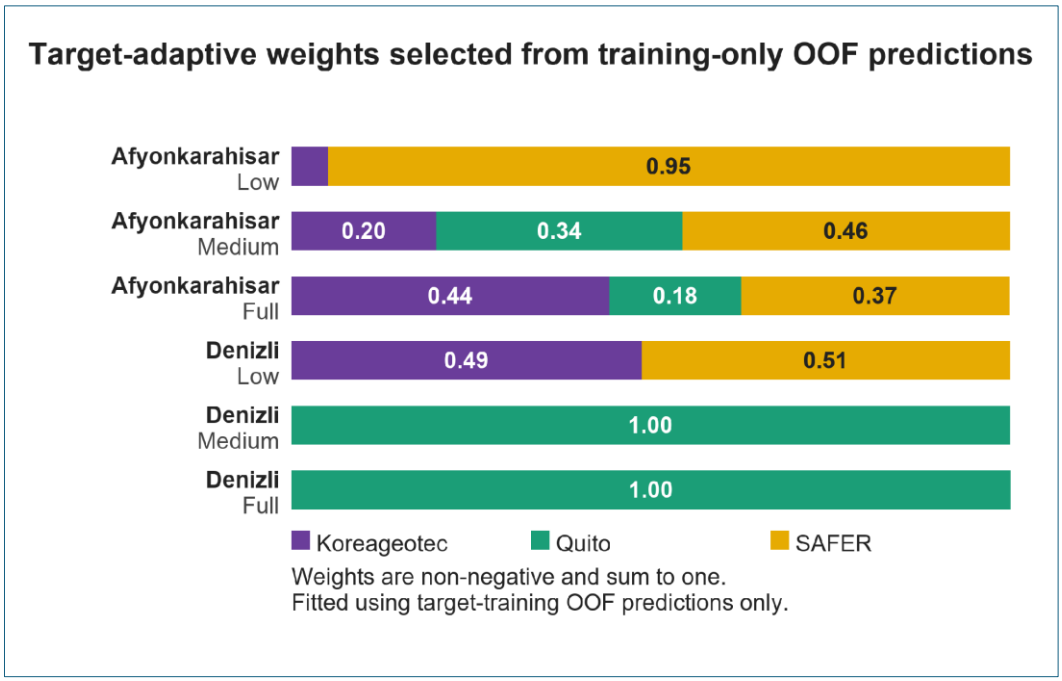


Figure 3. Target-adaptive ensemble weights by target and budget (non-negative and summing to one), estimated from location-grouped OOF predictions for target-training locations only.

DISCUSSION AND CONCLUSION

Source utility varied by target region and local-data budget. Transfer from SAFER, the Kathmandu Valley source, performed best among the individual sources for Afyonkarahisar. In Denizli, no individual source improved on the target-only reference at every budget, and pooling increased error at all three budgets. However, the medium-budget Denizli result should be interpreted cautiously because the target-only reference was comparatively variable at that budget; the positive SAFER gain may partly reflect baseline instability rather than a stable source advantage.

The pooled-source result is also difficult to separate from source-size imbalance: the filtered pooled training rows comprised 17,450 Koreageotec, 214 Quito, and 99 SAFER observations, so Koreageotec contributed about 98.2% of the pooled rows. This may allow the largest source to dominate the pooled representation and can help explain why pooled Denizli performance more closely followed the large-source condition. Balanced-source pooling by equal row counts or source weights was not evaluated and should be treated as a follow-up sensitivity analysis, not as a conclusion of this study. Interpretation is limited by one test split per target, few locked-test locations, differences in measurement practices, the fixed nested budget membership, and ensemble weights estimated from the same target-training pool used for model tuning. The standard deviations across the training-seed repetitions are descriptive, not confidence intervals. The results support evaluating candidate sources separately before pooling or combining their predictions; broader regional conclusions require additional independent locations and external validation.

References

Acarlar, E. A. (2026). Denizli için deterministik sismik tehlike analizi ve yer tepki analizleri [Master's thesis, Pamukkale University]. Thesis No. 998139.

Akın, M. K., Kramer, S. L., & Topal, T. (2011). Empirical correlations of shear wave velocity (Vs) and penetration resistance (SPT-N) for different soils in an earthquake-prone area (Erbaa-Turkey). Engineering Geology, 119, 1–17. https://doi.org/10.1016/j.enggeo.2011.01.007

Choi, I., Heo, G., Ahn, J.-K., Park, S.-C., Jang, Y.-G., & Kwak, D. (2026a). A curated geotechnical database from over 220,000 boring logs including VS, SPT-N, stratigraphy, and geologic and topographic attributes for South Korea. Scientific Data, 13, advance online publication. https://doi.org/10.1038/s41597-026-07719-w

Choi, I., Heo, G., Ahn, J.-K., Park, S.-C., Jang, Y.-G., & Kwak, D. (2026b). A curated geotechnical database from over 220,000 boring logs including VS, SPT-N, stratigraphy, and geologic and topographic attributes for South Korea (Version 1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.19777659

De Risi, R., De Luca, F., Gilder, C. E. L., Pokhrel, R. M., & Vardanega, P. J. (2021). The SAFER geodatabase for the Kathmandu valley: Bayesian kriging for data-scarce regions. Earthquake Spectra, 37(2), 1108–1126. https://doi.org/10.1177/8755293020970977

Gilder, C. E. L., Pokhrel, R. M., & Vardanega, P. J. (2019). The SAFER Borehole Database (SAFER/GEO-591_v1.1) [Data set]. University of Bristol Research Data. https://doi.org/10.5523/bris.3gjcvx51lnpuv269xsa1yrb0rw

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B. A., Haque, I. S., Beery, S. M., Leskovec, J., Kundaje, A., . . . Liang, P. (2021). WILDS: A benchmark of in-the-wild distribution shifts. Proceedings of Machine Learning Research, 139, 5637–5664. https://proceedings.mlr.press/v139/koh21a.html

Nacaroglu, E., Tugrul, A. T., & Yagcioglu, B. (2026). A tabular data augmentation framework based on error-focused XAI-supported weighting strategy: Application to soil liquefaction classification. Applied Sciences, 16(1), 330. https://doi.org/10.3390/app16010330

Othman, M. M., Hen-Jones, R., Zapata, C., Jiménez, E., De Luca, F., Holcombe, E. A., & Vardanega, P. J. (2023a). Geotechnical variability of the soils of Quito, Ecuador: A geodatabase study. Bulletin of Engineering Geology and the Environment, 82, Article 433. https://doi.org/10.1007/s10064-023-03438-3

Othman, M., Hen-Jones, R., Zapata, C., Jiménez, E., De Luca, F., Holcombe, E. A., & Vardanega, P. J. (2023b). Quito Geotechnical Database v1.1 [Data set]. University of Bristol Research Data. https://doi.org/10.5523/bris.ys9nfsd66knb21h3twx2rfin6

Payan, M., Asadi, P., Jamaldar, A., Salimi, M., Zanganeh Ranjbar, P., Jahed Armaghani, D., He, X., & Sheng, D. (2025). Artificial intelligence-based predictive models for shear wave velocity of soils: A comprehensive review. Engineering Applications of Artificial Intelligence, 155, 111095. https://doi.org/10.1016/j.engappai.2025.111095

Wang, Z., Dai, Z., Póczos, B., & Carbonell, J. (2019). Characterizing and avoiding negative transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11293–11302). https://doi.org/10.1109/CVPR.2019.01155

Acknowledgment

This research was supported by the TÜBİTAK BİDEB 2211-A National PhD Scholarship Program. The authors also acknowledge the published datasets and their original data providers.

Conflict of Interest

The authors have declared that there is no conflict of interest.

Author Contributions

E.N.: study design, analysis, methodological review, interpretation, and manuscript preparation. A.T.T.: data review, methodological review, and manuscript revision.E.A.A.: data organization, result checking, and manuscript revision. All authors reviewed and approved the final version.

Spatiotemporal Validation of Machine-Learning Models for Weekly Precipitation pH Estimation

Ayse Tuba Tugrul[1*], Özkan Günalp[2], Duygu Erişken[3], Olaf Yunus Laitinen Imanov[4], Rana İrem Turhan[5], Tuğba Ceren Turhan[6]

[1] Institute of Science, Civil Engineering, Pamukkale University, Denizli, Türkiye
[2] Institute of Health Sciences, Medical Biology, Ege University, İzmir, Türkiye
[3] Faculty of Science, Mathematics, Trakya University, Edirne, Türkiye
[4] Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland
[5] MSc Artificial Intelligence, School of Computer Science, University of Sheffield, Sheffield, United Kingdom
[6] Institute of Information Technology, Faculty of Computer Science, Information Technology and Energy, Riga Technical University, Riga, Latvia

*Corresponding author e-mail: aysetubat22@gmail.com

Abstract

Acid deposition reaches Earth's surface through rain, snow, and fog, or when gases and particles settle directly from the air. Weekly pH indicates sample acidity. We tested whether air-pollution and weather data from the same collection period can estimate sample pH, and whether estimates remain accurate in a later year or at stations not used for training. The outcome was weekly pH measured by the National Atmospheric Deposition Program/National Trends Network (NADP/NTN). Inputs were station location, elevation, and collection month; U.S. Environmental Protection Agency Air Quality System (EPA AQS) daily measurements of SO_2, NO_2, PM_{10}, and $PM_{2.5}$; and hourly ERA5 weather data on precipitation, temperature, humidity, pressure, cloud cover, and wind. EPA and ERA5 records were summarized over the sample's collection dates. The model estimated sample pH without using its measured chemistry as an input. The 2010–2014 primary analysis included 11,712 samples from 131 stations. Each sample had all four EPA pollutant matches within 100 km and at least 90% of expected ERA5 hourly records. We compared a monthly-median baseline, Ridge regression, random forest, and histogram-based gradient boosting using three tests: randomly splitting samples into training (80%) and test (20%) sets; training on 2010–2013 and testing samples starting in 2014; and holding out each station in turn. Random-forest MAE (average absolute pH error) was 0.303, 0.312, and 0.342 pH units across these tests. Gradient-boosting errors were 0.304, 0.307, and 0.350. Errors were similar in the first two tests and higher at stations kept out of training.

Keywords: *Precipitation pH, Acid Deposition, Temporal Validation, Station-Held-Out Validation*

INTRODUCTION

Acid deposition transfers acidic compounds from the atmosphere to Earth's surface. It reaches the surface as wet deposition, when rain, snow, or fog carries dissolved compounds, or as dry deposition of gases and particles. In the atmosphere, sulfur dioxide (SO_2) and nitrogen oxides (NO_x) can form sulfuric and nitric acids that contribute to both pathways (U.S. EPA, 2026b). Deposited acidity can affect soils, surface waters, vegetation, and built surfaces. This study examines wet deposition only: the pH of a

collected precipitation sample indicates that sample's acidity, but does not measure dry deposition or the total amount of acid deposited. Rain is naturally slightly acidic (about pH 5.6) because carbon dioxide dissolves in water (U.S. EPA, 2026b). We use 5.6 only as a descriptive reference for examining lower-pH observations, not as a class boundary for the model. Because pH is logarithmic, lower values mean greater hydrogen-ion concentration; accordingly, we model measured weekly pH as a continuous value.

Recent machine-learning studies connect precipitation acidity to environmental data, but differ in the outcome they model and the geographic and time scales they cover. Huang et al. (2024) used XGBoost, a tree-ensemble classifier, to classify acid-deposition outcomes across the Guangdong-Hong Kong-Macao Greater Bay Area during 2006–2021 from 19 weather, air-pollution, and land-surface variables; Shapley additive explanations (SHAP) summarized how those variables contributed to the model output. Wei and Huang (2024) estimated daily rainwater pH from 2008–2020 records at one Taipei station using a deep neural network and a long short-term memory model; they also used backward trajectories from the Hybrid Single Particle Lagrangian Integrated Trajectory (HYSPLIT) model to describe air-mass paths reaching the station, including during severe events.

These studies address related questions, but their scores cannot be compared directly with ours because they use different outcomes, locations, time scales, and evaluation designs. We ask whether air-pollution and weather measurements from the same collection week can estimate the measured pH of a precipitation sample. If so, such estimates could provide information for weeks when pH was not measured. We do not test filling in missing pH values or replacing laboratory measurements. Because the inputs come from the sample's collection week, this is an estimate of an already-collected sample, not a forecast or early-warning system. A random split can place weeks from the same station in both the training and test sets. It therefore does not show how well a model works in a later year or at a station it did not see during training; we evaluate both settings separately (Roberts et al., 2017; Boser, 2024).

Each record is one valid weekly precipitation sample from an NADP/NTN station. The value to estimate is the pH measured in that sample. Inputs come from three sources. NADP/NTN station records and collection dates provide location, elevation, and month. EPA AQS provides daily surface measurements of SO_2, NO_2, PM_{10}, and $PM_{2.5}$; we summarize each pollutant over the sample's collection dates using its mean, maximum, and standard deviation. ERA5 provides hourly precipitation, temperature, dew point, relative humidity, pressure, cloud cover, and wind; we summarize these weather records over the same dates. The model uses these inputs to estimate the sample's pH. Chemical measurements from that sample are excluded from the model inputs. We evaluate performance on randomly held-out weeks, samples beginning in 2014, and stations kept out of training. We also examine how far EPA monitors are from the sampled stations and how that distance affects sample coverage and model error.

MATERIAL AND METHODS

Material

The outcome was measured pH in weekly wet-deposition samples collected by NADP/NTN; the network samples rain and other precipitation, not dry deposition (NADP, 2021). Each valid station and collection interval (dateOn to dateOff) formed one station-week row, with pH as the response (y). Predictors (X) came from EPA AQS surface monitors and ERA5 hourly gridded weather estimates

(Hersbach et al., 2020). Chemistry measured in the target precipitation sample was excluded from X because it comes from the sample whose pH is being estimated.

Methods

The Collection Of The Data

Valid NADP/NTN samples with official station coordinates and collection intervals in 2010–2014 formed the analysis cohort. This was the verified common-coverage window in the assembled target, EPA AQS, and ERA5 data; it provided 2010–2013 for model fitting and a later year, 2014, for temporal evaluation. EPA AQS and ERA5 records were matched to each sample's exact dateOn–dateOff interval.

EPA AQS supplied SO_2 (parameter 42401), NO_2 (42602), PM_{10} (81102), and $PM_{2.5}$ (88101). Candidate daily records required a nonnegative arithmetic mean, positive observation percentage, valid date, reported unit, and coordinates. For duplicate records at a site on one day, we retained one record in this order: 24-hour, 1-hour, 8-hour, 3-hour, then other durations; ties favored higher observation percentage, lower POC, lower method code, and lower source-row number. For each pollutant and station-year, the nearest AQS site with valid selected records was chosen by great-circle distance and used for all weekly windows in that year. If it had no record within a window, we did not substitute a farther site. Each pollutant window was summarized by mean, maximum, and standard deviation. Gas concentrations remained in ppb and particle concentrations in the EPA-reported $\mu g/m^3$ units. The sample-level eligibility rule required the recorded selected duration to be 1 hour for SO_2/NO_2 and 24 hours or 24-hour block average for $PM_{10}/PM_{2.5}$; EPA provenance fields were retained (U.S. EPA, 2026a).

ERA5 hourly fields distributed through the Open-Meteo Historical Weather API were matched to the nearest 0.25-degree grid cell for each NADP station. We retained timestamps within the sample interval and required at least 90% of expected hourly records. For each window, precipitation was summed and hours with precipitation above zero were counted; temperature was summarized by its mean, minimum, maximum, and standard deviation. Dew point, relative humidity, surface pressure, cloud cover, and wind speed were averaged. Hourly wind speed and direction were converted to u and v components before their period means were calculated; direction angles were not averaged (Open-Meteo, 2026).

Open-Meteo and ERA5 data-use note. The hourly weather values were retrieved through the Open-Meteo Historical Weather API for non-commercial academic research under the provider's free-API terms. Open-Meteo identifies the API data as available under the Creative Commons Attribution 4.0 International licence (CC BY 4.0), which requires appropriate credit, a link to the licence, and an indication of modifications. We used model=era5, nearest 0.25-degree grid-cell selection, and elevation=nan, which disables statistical downscaling; the hourly values were then aggregated over each NADP collection interval. This use does not imply endorsement by Open-Meteo, ECMWF, or Copernicus. Sources: https://open-meteo.com/en/terms; https://open-meteo.com/en/pricing; https://creativecommons.org/licenses/by/4.0/. ERA5 data were produced by ECMWF for the Copernicus Climate Change Service (C3S) and accessed through the Open-Meteo Historical Weather API (Hersbach et al., 2023).

Statistical Analysis

F0 contained latitude, longitude, elevation, and cyclic month terms. F1 contained the four EPA pollutants' window means, maxima, and standard deviations. F2 contained the ERA5 window summaries described above. F3 combined F1 and F2; F4 combined all three groups. Every feature-group comparison used the same complete-case station-weeks.

The monthly-median reference predicted the training-set pH median for the same calendar month, falling back to the overall training median if needed. Ridge used median imputation, standardization, and alpha=10. Random forest used 60 trees, min_samples_leaf=3, random_state=42, and n_jobs=1; other parameters followed the scikit-learn defaults. Histogram-based gradient boosting used 180 iterations, learning_rate=0.05, max_leaf_nodes=31, l2_regularization=1.0, and random_state=42. These settings were fixed for the reported run; no hyperparameter search was conducted, and the 2014 test was not used to select or configure the models. Median imputation was fitted on training data only; scaling was used for Ridge.

S0 randomly assigned 80% of station-week records to training and 20% to testing using seed 42; because a station can occur in both sets, this is a within-network reference rather than a test at unseen stations. S1 assigned records by collection start year: samples beginning in 2010–2013 were used for training and those beginning in 2014 for testing. Eleven training samples began in 2013 and ended in early 2014; they were retained, so S1 is a start-year holdout rather than a fully time-disjoint cutoff. In S2 leave-one-station-out validation, all records from one station were withheld for testing while the model was trained on the remaining stations, repeating this process for each station. In practical terms, S0 tests other weeks from a network that may include the same stations in training, S1 tests samples whose collection starts in 2014, and S2 tests stations whose pH records were entirely absent from model fitting. Mean absolute error (MAE), the average absolute difference between predicted and measured pH, was the primary metric. Root mean square error (RMSE) gives greater weight to larger errors. The coefficient of determination (R^2) compares model squared error with prediction from the test-set mean and can be negative when the model is worse than that reference; signed bias shows whether predictions tend to be high or low. Station macro-MAE averages absolute error within each station and then across stations, giving each station equal weight. For S0 and S1, 95% confidence intervals were calculated by resampling stations rather than individual weeks, reflecting repeated observations within sites. Transfer loss was the matching S1 or S2 MAE minus the S0 MAE, so a positive value means greater error than the random-row reference.

RESULTS

The primary 100 km cohort comprised 11,712 weekly observations from 131 stations, or 26.9% of valid NADP/NTN pH targets. A sample was retained only when all four pollutant-monitor matches met the distance limit and at least 90% of expected ERA5 hourly records were available, so the results describe this matched subset rather than every valid NADP/NTN sample. For each retained station-week, the models used location and season, same-window EPA pollutant measurements, and ERA5 weather summaries to estimate a numeric pH for the already-collected precipitation sample. This benchmark asks whether such retrospective estimates remain accurate for samples starting in a later year and for stations absent from training, and may inform estimates where pH is unavailable; it does

not test missing-data recovery or replace measured pH. The 50 km and 250 km sensitivity cohorts comprised 2,920 observations from 41 stations and 21,796 observations from 236 stations, respectively; cohort coverage is summarized in Table 1. Annual sample counts in the primary cohort were 2,483, 2,150, 2,286, 2,484, and 2,309 from 2010 through 2014, and the median pH was below the descriptive 5.6 reference in each year (Figure 1).

Table 1. Data sources and analysis cohorts

Item	Specification / verified count
Target	NADP/NTN weekly wet-only pH; 2010–2014
EPA AQS predictors	Surface SO_2, NO_2, PM_{10}, and $PM_{2.5}$ observations; pollutant-specific concentration units
ERA5 predictors	Hourly weather via Open-Meteo; nearest 0.25° grid cell; ≥90% window coverage
Primary 100 km cohort	11,712 station-weeks; 131 stations; 26.9% of valid NADP targets
50 km sensitivity cohort	2,920 station-weeks; 41 stations; 6.7% of valid NADP targets
250 km sensitivity cohort	21,796 station-weeks; 236 stations; 50.1% of valid NADP targets

EPA distance limits apply to the four pollutant-monitor matches, not to the ERA5 grid cell; cohorts differ in sample composition. Abbreviations: NADP, National Atmospheric Deposition Program; NTN, National Trends Network; EPA, U.S. Environmental Protection Agency; AQS, Air Quality System; ERA5, fifth-generation global atmospheric reanalysis; PM_{10} and $PM_{2.5}$, particulate matter with aerodynamic diameters of 10 µm or less and 2.5 µm or less, respectively. ppb, parts per billion; µg m^{-3}, micrograms per cubic metre.

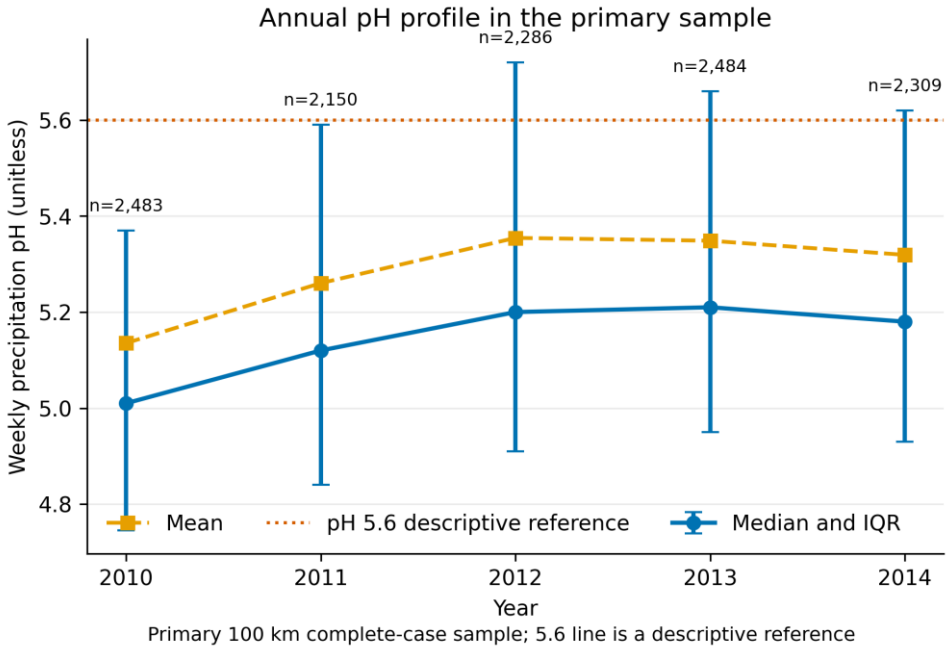


Figure 1. Annual weekly precipitation pH in the primary 100 km cohort; bars show the median and interquartile range (IQR), dashed markers show the mean, and the 5.6 line is a descriptive reference.

Table 2 summarizes full-context performance. The two tree-based models had random-row mean absolute errors of about 0.30 pH units; their predictions differed from measured pH by about 0.30 units

on average. On the 2014 holdout, histogram-based gradient boosting and random forest errors were 0.307 and 0.312; pooled leave-one-station-out errors were 0.350 and 0.342, respectively.

Table 2. Full-context model performance

Model	Split (n; stations)	MAE	RMSE	R^2	Bias
Monthly-median baseline	S0 (2,342; 125)	0.426	0.576	0.020	-0.126
Monthly-median baseline	S1 (2,309; 95)	0.407	0.565	-0.069	-0.179
Ridge regression	S0 (2,342; 125)	0.383	0.495	0.278	+0.009
Ridge regression	S1 (2,309; 95)	0.366	0.498	0.168	-0.055
Random forest	S0 (2,342; 125)	0.303	0.409	0.506	+0.009
Random forest	S1 (2,309; 95)	0.312	0.433	0.372	-0.059
Histogram-based gradient boosting	S0 (2,342; 125)	0.304	0.409	0.506	+0.000
Histogram-based gradient boosting	S1 (2,309; 95)	0.307	0.424	0.398	-0.066
Monthly-median baseline	S2 (11,712; 131)	0.436	0.590	—	-0.133
Ridge regression	S2 (11,712; 131)	0.390	0.508	—	-0.002
Random forest	S2 (11,712; 131)	0.342	0.469	—	-0.016
Histogram-based gradient boosting	S2 (11,712; 131)	0.350	0.471	—	-0.015

MAE, RMSE, and bias are expressed in pH units; bias is predicted minus observed. MAE, mean absolute error; RMSE, root mean square error; R^2, coefficient of determination. S2 pooled errors combine station-held-out predictions; S2 R^2 is not reported.

Figure 2 compares S0 and S1 mean absolute error; station-block bootstrap intervals show uncertainty across stations, while pooled S2 metrics are reported in Table 2.

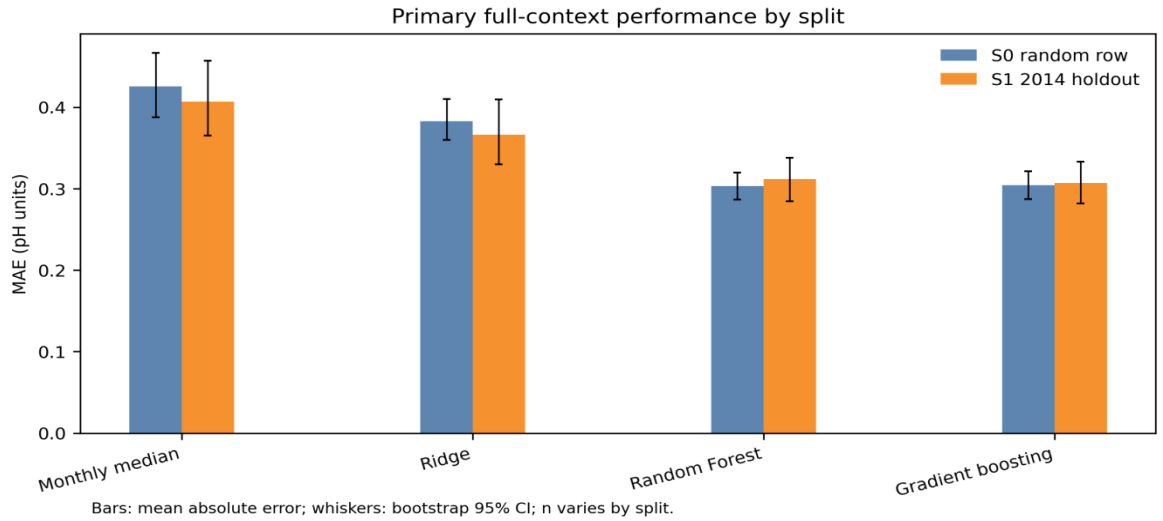


Figure 2. Full-context mean absolute error for the S0 random-row reference and S1 2014 holdout; whiskers show station-block bootstrap 95% confidence intervals (CIs).

Compared with each model's S0 reference, S1 MAE was higher by 0.008 for random forest and 0.003 for histogram-based gradient boosting; pooled S2 MAE was higher by 0.039 and 0.046 pH units, respectively (Table 2). These are descriptive contrasts across different test samples and designs, not isolated estimates of a pure temporal or spatial effect.

When each station was weighted equally, S2 station macro-MAE was 0.438 for the monthly-median reference, 0.406 for Ridge regression, 0.346 for random forest, and 0.352 for histogram-based gradient boosting.

On the 2014 test set, the combined pollutant-weather feature group (F3) had lower MAE than either group alone for both tree models. The full feature set (F4) yielded the lowest MAE (0.312 for random forest and 0.307 for histogram-based gradient boosting), reducing error relative to the location-and-season group (F0) by 0.015 and 0.005 pH units, respectively.

For the 2014 test, random forest and histogram-based gradient boosting MAE decreased from 0.329 and 0.333 at 50 km to 0.310 and 0.305 at 250 km; the S0 pattern was not monotonic (Figure 3). The 2014 test cohort also grew from 636 observations at 50 km to 4,128 at 250 km, so these comparisons reflect both the matching distance and which station-weeks were retained.

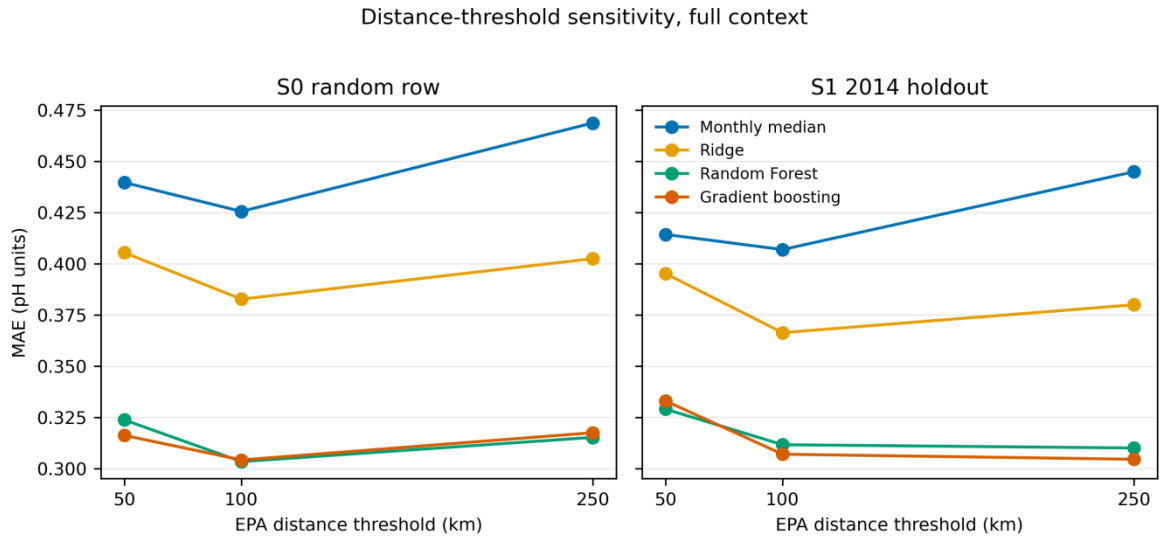


Figure 3. Full-context mean absolute error under EPA pollutant-monitor distance limits of 50, 100, and 250 km; panels show the S0 and S1 tests.

Among 2014 test observations with measured pH below 5.6 (n = 1,709), MAE was 0.219 pH units for histogram-based gradient boosting and 0.223 for random forest (Figure 4; U.S. EPA, 2026b). All four models had positive mean prediction bias: predicted pH tended to be higher than measured pH, indicating less acidity than was observed in this descriptive subset.

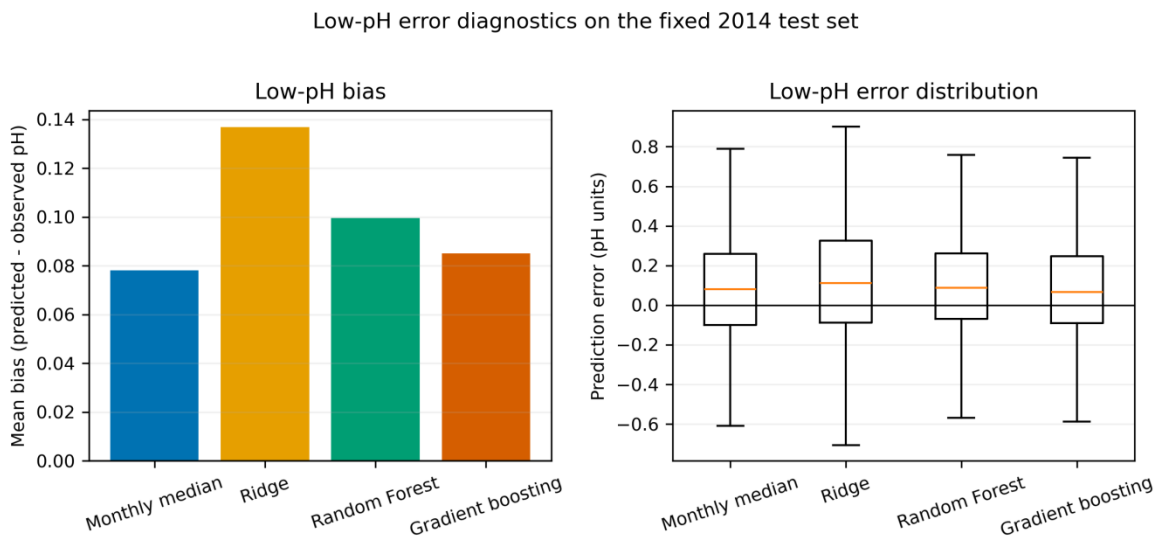


Figure 4. Prediction bias and error distribution for 2014 test samples with observed pH below 5.6; observed pH below 5.6 identifies a descriptive subset only.

DISCUSSION AND CONCLUSION

The two tree models showed similar errors in S0 and S1, but higher pooled error under S2 (Table 2; Figure 2). S0 can benefit from station-specific patterns shared by training and test records, whereas S2 withholds every record from the test station during fitting. The higher S2 error is consistent with weaker performance when a station has no pH records in training. These MAE differences are descriptive contrasts among different test samples, not isolated causal effects of time or station novelty; changes in station-week composition may also contribute. In addition, 11 training intervals began in 2013 and extended into 2014. Since they were assigned by start year, S1 should be read as a 2014 start-year holdout, not as a test with every input timestamp strictly after the training period.

This study complements earlier acid-rain machine-learning work without ranking scores across unlike tasks. It adds a continuous weekly pH benchmark across multiple NADP/NTN stations and evaluates within-network, later-year, and station-held-out performance. Its contribution is the matched weekly comparison of these validation settings, rather than a claim that one model universally outperforms prior studies.

On the same 2014 test observations, combining pollutant and weather summaries reduced error relative to either group alone, and adding location and season produced the lowest MAE. This pattern is consistent with complementary predictive information in the measured groups, but it does not identify causal mechanisms and may not hold outside this cohort.

Among 2014 observations with measured pH below 5.6, all four models had positive mean prediction bias (Figure 4): their predicted pH was higher, and therefore indicated less acidity, than the measured pH on average. This direction matters if regression estimates are used to screen acidic conditions, because average error can conceal underestimation at the acidic end. The 5.6 value defines only a descriptive subset here; no classifier, threshold-detection sensitivity, or false-negative rate was evaluated, so these results do not establish alert performance.

The 250-km sensitivity cohort included more station-weeks and stations than the 50-km cohort, and tree-model MAE was lower in the 2014 test. Because both monitor distance and retained sample composition changed, these results do not show that farther monitors improve estimation. Findings apply to the 2010–2014, 100-km matched cohort. In S2, other and potentially nearby stations remained in training, so the results describe station holdout within the sampled network, not regional extrapolation.

Overall, this study provides a matched weekly benchmark for retrospective pH estimation from pollutant and weather observations. On the screened cohort, tree-model MAE was about 0.31 pH units in the random-row and 2014 start-year tests and 0.34–0.35 in pooled station-held-out evaluation. The results show why performance should be judged under the intended transfer setting: random-row accuracy alone gives a less cautious estimate for a station absent from model fitting. The benchmark can inform retrospective estimates in monitoring networks with comparable weekly sampling and data coverage, but it does not validate replacing direct pH measurements or deploying the models in a new region without external testing.

References

Boser A, 2024. Validating spatio-temporal environmental machine learning models: Simpson's paradox and data splits. Environmental Research Communications, 6 (3): 031003. https://doi.org/10.1088/2515-7620/ad2e44.

Hersbach H, Bell B, Berrisford P, et al., 2020. The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society, 146 (730): 1999–2049. https://doi.org/10.1002/qj.3803.

Huang Z, Fu J, Liu B, Zhao X, Zhang Y, Wang X, 2024. Acid rain prediction in the Guangdong-Hong Kong-Macao Greater Bay Area using an explainable machine learning framework. Atmospheric Pollution Research, 15 (9): 102201. https://doi.org/10.1016/j.apr.2024.102201.

NADP, 2021. National Trends Network weekly metadata. National Atmospheric Deposition Program. Access address: https://nadp.slh.wisc.edu/wp-content/uploads/2021/10/ntn-weekly-Meta.pdf; Date of access: 10.16.2026.

Open-Meteo, 2026. Historical Weather API documentation, Terms of Use, and Pricing. Access addresses: https://open-meteo.com/en/docs/historical-weather-api; https://open-meteo.com/en/terms; https://open-meteo.com/en/pricing; CC BY 4.0: https://creativecommons.org/licenses/by/4.0/; accessed October 16, 2026.

Hersbach H, Bell B, Berrisford P, et al., 2023. ERA5 hourly data on single levels from 1940 to present [data set]. ECMWF. https://doi.org/10.24381/cds.adbb2d47.

Roberts DR, Bahn V, Ciuti S, et al., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40 (8): 913–929. https://doi.org/10.1111/ecog.02881.

U.S. EPA, 2026a. Air Quality System (AQS) Data Mart [internet database]. Available at: https://aqs.epa.gov/aqsweb/ and https://www.epa.gov/aqs; accessed October 16, 2026.

U.S. EPA, 2026b. What is Acid Rain? Access address: https://www.epa.gov/acidrain/what-acid-rain; Date of access: 10.16.2026.

Wei CC, Huang R, 2024. A hybrid approach of air mass trajectory modeling and machine learning for

acid rain estimation. Water, 16 (23): 3429. https://doi.org/10.3390/w16233429.

Conflict of Interest

The authors have declared that there is no conflict of interest**.**

Author Contributions

A.T.T. led conceptualization, methodology, data curation, formal analysis, visualization, investigation, validation and writing of the original draft. Ö.G. contributed to methodology, supervision, and manuscript review and editing. D.E. and O.Y.L.I. contributed to methodology, data curation, formal analysis, visualization, investigation, validation. R.I.T. and T.C.T. contributed to methodology, data curation, formal analysis, investigation, validation, and manuscript review and editing. All authors approved the final manuscript.