SDU

1996

UNIVERSITY

Journal of Emerging Technologies and Computing (JETC)

Journal of Emerging Technologies and Computing

is a peer-reviewed, open-access national and international scientific journal. Thematic areas: Computer Science, Infocommunication Technologies, and Mathematics with Applied Aspects

Publisher: SDU University

The journal is registered and licensed as an online publication and printed journal by the Ministry of Information and Social Development of the Republic of Kazakhstan.

Certificate (Online publication): No KZ01VPY00120097

Certificate (Printed journal): No KZ71VPY00120098

Frequency: four times a year (March, June, September, December)

Website: https://jetc.sdu.edu.kz

Editor-in-Chief:
Dana Utebayeva, PhD, Assistant Professor, SDU University

Managing Editor:
Assem Talasbek, PhD, Assistant Professor, SDU University

Managing Editor for the "Mathematics" Section:
Yerkin Shaimerdenov, PhD candidate, SDU University

Technical Editor:
Meraryslan Meraliyev, PhD, SDU University

Editorial Board

- **Shirali Kadyrov**, PhD, Associate Professor,
  New Uzbekistan University (Uzbekistan)
  ORCID: 0000-0002-8352-2597
- **Selcuk Cankurt**, PhD, Assistant Professor,
  Vistula University (Poland)
  ORCID: 0000-0003-0581-1913
- **Khaled Mohamad**, PhD, Assistant Professor,
  SDU University (Kazakhstan)
  ORCID: 0000-0002-5980-0147
- **Lyazzat Ilipbayeva**, Candidate of Technical Sciences, Associate Professor,
  IITU (Kazakhstan)
  ORCID: 0000-0002-4380-7344
- **Kamila Orynbekova**, PhD,
  SDU University (Kazakhstan)
  ORCID: 0000-0002-2182-2914
- **Zhandos Dosbayev**, PhD,
  Satbayev University (Kazakhstan)
  ORCID: 0000-0003-1673-4036
- **Bektur Baizhanov**, Doctor of Physical and Mathematical Sciences, Academician, Professor,
  SDU University (Kazakhstan)
  ORCID: 0000-0002-3743-7404
- **Nurlan Dairbekov**, Doctor of Physical and Mathematical Sciences, Professor,
  SDU University (Kazakhstan)
  ORCID: 0000-0002-2725-7549

Ministry of Science and Higher Education of the Republic of Kazakhstan
SDU University

Journal of Emerging Technologies and Computing (JETC)

Volume 1, Issue 1 • June 2025

Kaskelen, Kazakhstan — 2025

CONTENTS

- **SECTION I – Computer Science**
  - Bias and Fairness in Automated Loan Approvals: A Systematic Review of Machine Learning Approaches
    Suraiyo Raziyeva and Meraryslan Meraliyev 7
  - A Survey on Multimodal Approaches for Lung Disease Diagnosis using Deep Learning
    Zhaniya Medeuova 18
  - Development of method to analyze factors of kidney disease by the use of fuzzy logic
    Assel Yembergenova, Azamat Serek, and Bauyrzhan Berlikozha 30
  - Detecting Social Conflicts In Kindergartens Using Deep Learning And Computer Vision
    Dina Kengesbay 37
  - Performance Comparison of Statistical Models in PM2.5 Forecasting: A Case Study of Almaty
    Nuray Dauletkhan and Khaled Mohamad 48
  - Optimizing QA Systems: Evaluating Row-Based and Traditional Chunking in Structured-Data-Aware Retrieval-Augmented Generation for University Virtual Assistants
    Maksat Maratov and Selcuk Cankurt 60
- **SECTION II – Infocommunication Technologies**
  - A Review of Recent Deep Learning Methods in Spectrum Sensing
    Aizhan Utepova, Nurzhigit Smailov, and Paweł Komada 74
  - Analysis of External Factors on the Accuracy of Object Detection by Lidar Sensor
    Arsen Abdrakhmanov and Lyazzat Ilipbayeva 81
- **SECTION III – Mathematics with Applied Aspects**
  - Development And Optimization Of Physics-Informed Neural Networks For Solving Partial Differential Equations
    Batyr Sharimbayev, Shirali Kadyrov, and Aleksei Kavokin 94
  - Numerical Fractal Analysis of Exceptional Sets in the Lehner Expansion
    Symbat Duisen, Aiken Kazin, and Shirali Kadyrov 103

SECTION I
Computer Science

This section focuses on current research directions and applied advancements in Computer Science, particularly in the areas of artificial intelligence, software engineering, and intelligent systems.

Review

Bias and Fairness in Automated Loan Approvals: A Systematic Review of Machine Learning Approaches

Suraiyo Raziyeva [1] and Meraryslan Meraliyev [2]

[1]Department of Computer Science, SDU University, Almaty, Kazakhstan
[2]Department of Computer Science, SDU University, Almaty, Kazakhstan

DOI: 10.47344/jbzmnx25

Abstract

Artificial intelligence (AI) is increasingly transforming credit approval processes, enabling financial institutions to assess risk more efficiently and at greater scale. As these systems become more embedded in lending decisions, concerns around fairness, bias, and accountability have grown significantly. Many of these concerns stem from the use of historical data, proxy variables, and model optimization choices that can unintentionally reinforce existing social and economic inequalities. This work presents a systematic overview of the types and sources of bias in AI - driven loan approval systems and critically examines how machine learning techniques attempt to address them. It also highlights emerging solutions, including explainable AI, federated learning, human-in-the-loop frameworks, and intersectional fairness approaches. Despite ongoing advancements, unresolved challenges remain - particularly the need for dynamic fairness monitoring and for addressing intersectional biases affecting individuals from multiple marginalized groups. To bridge these gaps, the paper emphasizes the importance of interdisciplinary collaboration among AI developers, regulatory bodies, and social scientists. It advocates embedding fairness as a core design principle in the development and deployment of future AI systems. Overall, this study contributes to the growing effort to develop more transparent, inclusive, and socially responsible financial technologies.

Keywords: AI bias, fairness techniques, loan approval, financial inclusion, regulatory compliance, algorithmic fairness, proxy bias.

I. Introduction

Artificial Intelligence (AI) has significantly transformed decision-making in the banking sector, particularly through the automation of lending approvals. These systems are often praised for their efficiency and scalability; however, they also raise critical concerns regarding fairness and bias. Historical data used to train such systems may reflect past discriminatory practices, leading to models

Email: suraiyo.raziyeva@sdu.edu.kz ORCID: 0009-0008-1832-7155
Email: meraryslan.meraliyev@sdu.edu.kz ORCID: 0000-0003-2627-0837

Received: April 3, 2025. Reviewed: April 16, 2025. Accepted: April 22, 2025. *© 2025 Suraiyo Raziyeva and Meraryslan Meraliyev. All rights reserved.*

that perpetuate unfair outcomes. As a result, automated credit scoring and lending decisions may disproportionately disadvantage underrepresented groups [1].

Bias in AI systems may also stem from algorithmic design choices, such as optimization objectives or feature selection. For instance, if the training dataset is imbalanced, models may systematically favor majority groups. Variables like ZIP codes can inadvertently encode socioeconomic or demographic biases, reinforcing existing inequalities. In some cases, the use of such features may even violate regulatory frameworks, including the General Data Protection Regulation (GDPR) and the Equal Credit Opportunity Act (ECOA) [1], [2]. Beyond ECOA and GDPR, emerging international frameworks—such as the EU AI Act and Canada's Directive on Automated Decision-Making—signal a global shift toward standardizing fairness and transparency in automated financial systems. These evolving policies reflect the growing global consensus on the need for algorithmic accountability in financial decision-making.

This paper considered various contemporary fairness metrics used to detect and assess bias in AI-driven lending systems, as well as a range of strategies designed to mitigate such biases [3]. Lending discrimination can originate from multiple sources, including proxy variables that are correlated with race, gender, or economic status, as well as algorithmic priorities that favor accuracy at the expense of fairness. These biases, if left unaddressed, may deepen economic inequalities by restricting equitable access to credit.

In deploying AI for credit decisions, it is essential to consider not only model performance but also fairness, ethical accountability, and compliance with legal standards. Financial institutions must adhere to anti-discrimination laws and data protection regulations such as the GDPR and ECOA [1]. These frameworks aim to ensure equitable treatment within algorithmic decision-making. However, many current machine learning techniques struggle to balance fairness and accuracy effectively.

The goal of this work is to provide a comprehensive examination of machine learning approaches aimed at promoting fairness in AI-powered lending systems [4]. It categorizes bias mitigation methods into three primary classes: pre-processing, in-processing, and post-processing. It further analyzes the strengths and limitations of each approach in addressing algorithmic unfairness. The study also examines the broader societal implications of discriminatory lending and contributes to the ongoing discourse by including real-time fairness monitoring techniques and intersectional fairness considerations.

By offering an in-depth analysis of fairness-centered machine learning strategies, this paper contributes to the growing literature on ethical AI in financial services. It synthesizes key trends in recent work and clearly identifies persistent gaps in current research. Notably, we find a lack of robust methods for continuous ("live") fairness monitoring in deployed lending systems and a deficiency of techniques to address intersectional biases—cases where compounded disadvantages (e.g., being a minority and low-income) fall through the cracks of one-dimensional fairness metrics. We also observe a disconnect between high-level ethical principles and their practical implementation in AI lending algorithms. To help bridge these gaps, we propose future directions such as developing dynamic equity monitoring tools and more holistic fairness metrics that account for intersecting social categories. Ultimately, the insights from this systematic review are intended to inform the design of more transparent, inclusive, and accountable AI-driven credit decision platforms, guiding both researchers and practitioners toward solutions that balance performance with fairness and uphold regulatory and social standards.

A. Theoretical Frameworks of Fairness

In recent discussions surrounding AI-driven financial decisions, particularly automated loan approvals, several ethical frameworks provide valuable insights into the concept of fairness. From a utilitarian perspective, fairness is assessed by evaluating whether algorithmic outcomes maximize overall welfare or utility. This approach considers decisions fair if they enhance aggregate societal benefit or utility, often prioritizing accuracy and profitability in credit scoring. However, this might inadvertently result in unequal distributions of benefits and harms [30], [31].

Alternatively, Rawlsian fairness, rooted in John Rawls's notion of justice as fairness, emphasizes distributive justice aimed explicitly at improving conditions for the least advantaged. Rawls's Difference Principle asserts that any inequality is justified only if it benefits those most disadvantaged [32], [33]. Applying this framework to lending implies that fairness-oriented interventions should protect vulnerable borrowers and mitigate disparities that could otherwise exacerbate their disadvantages [34].

Furthermore, the emerging computational justice framework proposes integrating egalitarian principles directly into AI systems. This framework stresses the importance of equitable access, representation, and outcomes, ensuring algorithmic lending practices provide fair opportunities across diverse socioeconomic backgrounds and proactively address biases [35], [36].

Collectively, these ethical perspectives—utilitarianism, Rawlsian fairness, and computational justice—serve as complementary lenses, enriching our understanding of fairness in AI-driven financial services. They guide the design and assessment of algorithms, promoting fairness that encompasses both efficiency and equity, particularly benefiting vulnerable populations.

Received: April 3, 2025. Reviewed: April 16, 2025. Accepted: April 22, 2025. *© 2025 Suraiyo Raziyeva and Meraryslan Meraliyev. All rights reserved.*

II. Methodology

To carry out a comprehensive and objective evaluation of fairness-centered, AI-driven loan approval systems, this study adopts the systematic literature review (SLR) methodology. This approach enables a structured examination of existing scholarly work, facilitating the identification of prevailing trends, emerging challenges, and methodological limitations associated with fairness-aware AI in financial decision-making.

A. Search Strategy

The literature search focused on peer-reviewed journal articles, conference proceedings, and preprint research published between 2020 and 2024. Key academic databases were utilized, including Scopus, Google Scholar, IEEE Xplore, the ACM Digital Library, and arXiv—chosen for their extensive coverage of publications related to artificial intelligence, machine learning, and financial technologies. A combination of keywords and Boolean operators was applied to ensure the inclusion of relevant and high-quality studies. The following search terms were used:

- `("fairness-aware machine learning" OR "bias mitigation in AI")`
- AND `("loan approval" OR "credit scoring" OR "financial decision-making")`
- AND `("pre-processing" OR "in-processing" OR "post-processing")`

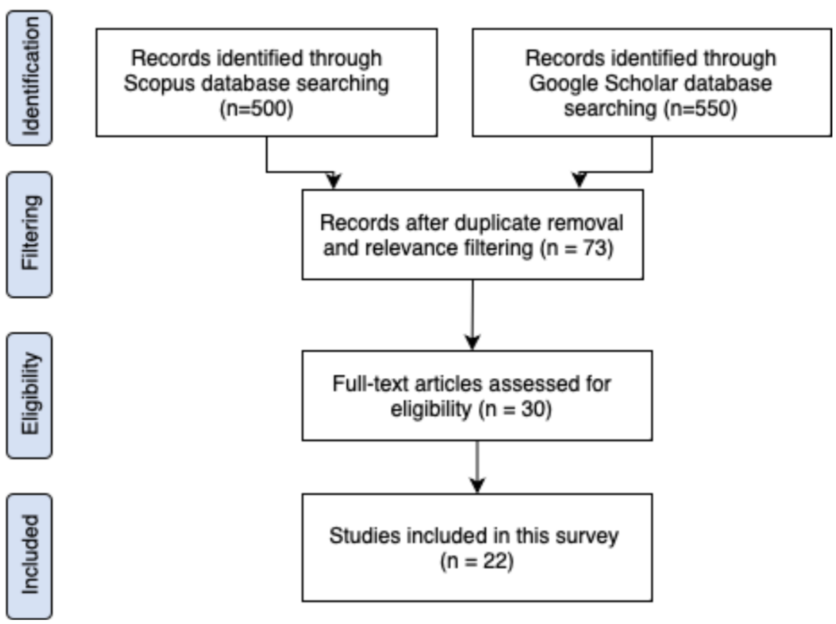


Fig. 1. Systematic Literature Review (SLR) Process: from initial identification to final selection.

Figure 1 illustrates the step-by-step review process applied in identifying relevant literature for this study. Initially, database searches yielded 500 records from Scopus and 550 from Google Scholar, totaling 1050 records.

B. Inclusion and Exclusion Criteria

Studies were selected as suitable if they comprehensively addressed bias mitigation methods, specifically in AI-driven loan approval systems. Preference was given to studies explicitly employing fairness-aware techniques categorized into pre-processing,

Received: April 3, 2025. Reviewed: April 16, 2025. Accepted: April 22, 2025. © 2025 Suraiyo Raziyeva and Meraryslan Meraliyev. All rights reserved.

in-processing, or post-processing methods. Priority was also given to studies providing empirical evaluations of fairness-accuracy trade-offs and discussions aligned with regulatory and ethical frameworks, such as the General Data Protection Regulation (GDPR) and the Equal Credit Opportunity Act (ECOA).

Studies were excluded from the review if they addressed fairness exclusively in non-financial contexts, such as healthcare or employment. In addition, works lacking quantitative evaluations or those that did not propose explicit bias mitigation strategies were omitted. Non-peer-reviewed sources, including editorial commentaries and opinion articles, were also excluded to ensure the inclusion of rigorously vetted research.

Initially, 73 records were retained after duplicate removal and initial filtering based on relevance. These underwent a full-text eligibility assessment, after which 30 articles remained. Following rigorous quality assessment and further evaluation for relevance and context, a final set of 22 high-quality and contextually relevant studies was selected for detailed comparative analysis.

C. Data Extraction Process

Following the selection of relevant studies, key information was carefully extracted to support a meaningful comparison of fairness-aware machine learning techniques. This included identifying the specific types of bias each study addressed - such as historical, sample, algorithmic, proxy, or societal bias. The review also categorized the mitigation methods based on the stage at which they were applied: before training (pre-processing), during training (in-processing), or after model deployment (post-processing). Particular attention was given to empirical findings, especially how each approach balanced fairness with predictive accuracy, using metrics like statistical parity difference, equalized odds, and disparate impact ratio. Finally, the analysis considered how well each technique aligned with existing financial regulations and ethical standards.

D. Quality Assessment

Each study was assessed based on clarity of methodology, empirical rigor, practical applicability, and regulatory alignment to ensure the robustness and relevance of findings. After this detailed quality assessment and eligibility evaluation, a final set of 22 studies were included in the comparative survey.

The collected data were analyzed to evaluate the comparative strengths, limitations, and practical applicability of each technique. This analysis provided insights into how different bias mitigation strategies perform in real-world contexts and highlighted inherent trade-offs.

III. Taxonomy of Bias and Fairness Techniques

A. Categorization of Bias in AI-Based Loan Approvals

Recent developments in fairness research have introduced new ways to reduce bias. For instance, federated learning allows models to be trained across different datasets without centralizing data [5]. This approach keeps personal information private and helps reduce geographical and demographic biases. We also have real-time fairness monitoring systems. These systems continuously check and adjust model performance after deployment to address any changes in bias [6]. Finally, new intersectional fairness metrics consider the combined effects of biases on people who belong to multiple disadvantaged groups [7]. This promotes more inclusive lending practices.

Table I provides a clear comparison of different bias mitigation techniques used in fairness-aware machine learning, grouped by when they are applied in the modeling process: before training (pre-processing), during training (in-processing), or after the model has been trained (post-processing) [8]. The table outlines how each method typically performs in terms of improving fairness, the extent to which it affects model accuracy, and how complex it is to implement. As shown, pre-processing techniques like reweighing and data balancing are relatively easy to apply and don't significantly impact accuracy, but they may offer only modest improvements in fairness [4]. In contrast, in-processing methods, such as adversarial debiasing, tend to be more effective in reducing bias but are also more complex and may involve trade-offs in performance [8], [9]. Post-processing approaches are often useful for adjusting outcomes in already trained models, though they can sometimes lead to inconsistencies or reduced interpretability [11]. This comparison helps highlight the practical choices researchers and developers must make when selecting fairness techniques, depending on their goals, constraints, and the context in which the model will be used.

B. Fairness-Aware Machine Learning Techniques

To mitigate the impact of bias in AI-driven lending, various fairness-aware machine learning techniques have been developed. These techniques can be classified based on when they are applied during the model development process:

Received: April 3, 2025. Reviewed: April 16, 2025. Accepted: April 22, 2025. *© 2025 Suraiyo Raziyeva and Meraryslan Meraliyev. All rights reserved.*

TABLE I
SUMMARY OF BIAS TYPES AND CORRESPONDING MITIGATION TECHNIQUES IN AI-BASED LOAN APPROVAL SYSTEMS.

Bias Type	**Application in Loan Approvals**	**Key Features and Challenges**	**Performance Analysis/Remarks**
Historical Data Bias	AI models learn from past loan decisions.	Reflects systemic inequalities in training data, leading to biased outcomes.	Replicates historical discrimination: requires data rebalancing or de-biasing techniques.
Sample Bias	Models trained on non-representative datasets.	Leads to poor generalization for underrepresented groups.	Reduces model accuracy for diverse populations: mitigated by re-sampling or balanced datasets.
Algorithmic Bias	Introduced during feature selection or optimization processes.	Favors majority groups due to imbalanced cost functions or feature correlations with protected attributes.	Reduces fairness, mitigated by fairness-aware algorithms and optimization strategies.
Proxy Bias	Use of correlated attributes as substitutes for sensitive ones.	Leads to indirect discrimination when proxies represent protected characteristics.	Compromises fairness: requires removal or neutralization of proxy variables.
Amplification Bias	AI systems intensify existing disparities in training data.	Magnifies inequities by reinforcing patterns of inequality present in datasets.	Increases systemic bias: mitigated through fairness monitoring and algorithmic adjustments.
Societal Bias	Societal prejudices embedded in datasets influence AI decisions.	Reflects societal inequities like gender or racial discrimination, influencing model outcomes.	Preserve inequality addressed through societal-level reforms and ethical AI frameworks.

1) Pre-Processing Techniques: As a pre-processing method, transformation techniques are applied to manipulate the dataset before machine learning models consume the data to get unbiased data and equal representation. These include approaches to resampling, creating synthetic data, and finally, removing proxy variables so that the dataset fairly represents all demographic groups. Such methods mainly target historical and sample bias but must be carefully tuned to ensure the utility of the data. Increasing data for underrepresented groups should re-balance the data set and improve fairness and minimal prediction accuracy reduction [5].

2) In-Processing Techniques: In-processing methods include fairness constraints in the process of training a model. These algorithms optimize fairness metrics, such as demographic parity or Equalized Odds, and traditional accuracy metrics. The models balance fairness and performance by embedding these constraints into the learning objective. In-processing techniques go well with modern AI frameworks, which consider fairness-aware algorithms one of the strong approaches toward handling algorithmic bias [9].

3) Post-Processing Techniques: Post-processing methods take an already fitted model and perform whatever adjustments necessary to the model outputs so that the resulting predictions satisfy some notion of fairness, be it demographic parity or Equalized Odds. While post-processing methods allow for easy retrofitting of fairness into many existing models, they come at a small cost in the overall predictive accuracy of the model. In practice, these methods are used when it is infeasible to modify the process by which models are trained or the data on which they are trained [11].

C. Empirical Findings: Fairness - Accuracy Trade-offs

Empirical studies of fairness-aware machine learning techniques reveal a complex interplay between predictive accuracy and fairness outcomes. In the context of automated loan approvals, achieving both high performance and equity across demographic groups is rarely straightforward. This subsection synthesizes empirical findings from the reviewed literature to highlight observed trade-offs, using metrics such as statistical parity difference (SPD), equalized odds, and disparate impact ratio (DIR).

Received: April 3, 2025. Reviewed: April 16, 2025. Accepted: April 22, 2025. *© 2025 Suraiyo Raziyeva and Meraryslan Meraliyev. All rights reserved.*

Pre-processing techniques, such as reweighing and massaging the training data, often lead to modest improvements in fairness metrics like SPD and DIR, particularly in datasets with high initial bias. However, these improvements can come at the cost of reduced predictive accuracy, especially when the modified data distribution diverges from the underlying population [3], [11].

In-processing methods, including adversarial debiasing and fairness-constrained optimization, tend to achieve a better balance between fairness and accuracy. Several studies demonstrate that models trained with fairness constraints are capable of maintaining comparable AUC scores while significantly reducing disparate impact [15]. However, these methods often require extensive tuning and longer training times.

Post-processing techniques such as reject option classification or calibrated equalized odds adjustments are shown to be effective when model retraining is not feasible. While these approaches can achieve compliance-level improvements in fairness indicators like DIR (moving from below 0.8 to above the threshold), they sometimes reduce classification confidence, especially near decision boundaries [26].

D. Comparative Analysis of Fairness Metrics and Mitigation Techniques

Bias mitigation strategies in machine learning for loan approval systems are commonly grouped into three categories: pre-processing, in-processing, and post-processing. Each group offers distinct advantages and limitations depending on the context in which it is applied.

Pre-processing techniques attempt to transform the training data to remove bias before model training. Methods such as reweighting and sampling adjustments are often model-agnostic and relatively easy to apply. However, they may risk distorting the original data distribution, which could reduce model performance in certain applications.

In-processing techniques modify the learning algorithm itself by incorporating fairness constraints or altering the loss function. These methods, such as adversarial debiasing or fairness-constrained optimization, tend to offer a strong balance between accuracy and fairness. However, they often require access to and modification of the model's internals, making them less applicable to proprietary or black-box systems.

Post-processing techniques alter the model output without changing the model or data. Examples include reject option classification and calibrated equal odds. These methods are typically easy to implement but may offer limited fairness correction, especially if bias is deeply embedded in the model structure.

evaluation is further complicated by the choice of metric. The most widely used measures include:

- Statistical Parity Difference (SPD) – measures the difference in positive outcomes between privileged and unprivileged groups. It is simple but may ignore performance discrepancies.
- Disparate Impact (DI) – computes the ratio of favorable outcomes across groups. It is widely used in regulatory contexts but sensitive to class imbalance.
- Equal Opportunity Difference – measures the true positive rate gap between groups, emphasizing fairness in correctly approved applications.
- Average Odds Difference – considers both true positive and false positive rates across groups, offering a more nuanced view of fairness.

In practice, there is often a trade-off between fairness and accuracy. For example, post-processing techniques may improve DI but slightly reduce ROC-AUC. In contrast, in-processing methods like adversarial debiasing may maintain accuracy but increase computational cost.

The comparative analysis conducted in this review suggests that fairness-aware learning remains a multidimensional problem where technical performance must be balanced with ethical and legal considerations.

E. Case Studies on Bias Mitigation in AI-Based Loan Approval Systems

Bias mitigation in AI-driven loan approval systems has been the focus of active study through numerous case studies that detail the best strategies for the implementation of fairness and transparency. Figure II provides a summary of representative case studies that have examined various fairness-aware strategies in credit decision-making. The table highlights not only the technical contributions but also the domain-specific applications and the diversity of fairness concerns addressed in the literature. These case studies present a range of real applications and tests of methods related to bias mitigation within a wide array of domains relating to financial services. Below are summaries of key recent findings: Purificato et al. (2023) and Lorenzo (2019) both incorporate explainable

Received: April 3, 2025. Reviewed: April 16, 2025. Accepted: April 22, 2025. *© 2025 Suraiyo Raziyeva and Meraryslan Meraliyev. All rights reserved.*

AI techniques—such as the Trust, Reliance Scale,[1] and SHAP values—to promote transparency and user trust, enabling clearer justification for credit decisions [12], [13]. In contrast, Dattachaudhuri et al. (2022) focus on rule extraction and rule pruning toward the same goal and outline the manner in which interpretability can complement and exist separately from typical model structures [11]. Regulatory focus enhances the work of Nadeem et al. (2023), who connect fairness interventions directly into organizational practice and emphasize the need for embedding algorithmic remedies within larger-scale institutional change [14]. A second group of studies addresses fairness from the systems-level or technical perspective. Wang et al. (2023) introduce a human-in-the-loop framework that enables real-time correction of bias through interactive feedback [15], while Zhou et al. (2022) address bias, privacy, and regional imbalance simultaneously through federated learning [16].

Parra et al. (2022) address the problem from the data perspective, looking at how proxy variables such as ZIP codes may introduce bias inadvertently and describing how to nullify their effect [17]. Finally, there are a few papers that focus on fairness-performance trade-offs. Karimova (2024) investigates optimization techniques that balance fairness and precision in micro-lending in small businesses [18], while Cozerenco and Szafarz (2015) adopt a co-financing strategy in microfinance for reducing gender bias, illustrating a policy intervention specific to the field rather than a technical one [19]. Overall, the papers illustrate that while fairness is the common target, the paths towards its realization are very distinct depending on the field, purpose, and limitations of the credit system.

IV. Emerging Trends and Potential Improvements

Recent developments in AI-driven financial decisions have marked a milestone toward a solution for bias and for the promotion of fairness. Specifically, fairness-aware frameworks with integrated XAI tools have become a game-changing solution for improving transparency and rebuilding stakeholder trust. Ashraf and Faheem 2021 [21] proposed XAI-based fairness detection frameworks. These frameworks ensure equity as well as accountability within credit ratings by eliminating algorithmic biases. On the other hand, Garcia et al. (2023) [22] also put great emphasis on systematic bias detection and fairness metrics as being very important in order to reduce both societal and algorithmic discrimination in loan approval systems. Federated learning has also been highlighted as a means of training AI on various data without the sharing of private information. Zhou and Tang (2022) [25] have demonstrated their potential in trying to reduce geographical and population biases while maintaining strong information security in financial institutions. Interpretable machine learning for finance has also seen developments in tools such as the SHAP technique, which is designed to give insight into how decision-making is derived by algorithms. SHAP has been shown to increase transparency and build trust in AI-driven credit decisions, according to Lorenzo 2019 [24], which was one of the concerns that stakeholders had. Other innovative techniques have recently emerged as human-in-the-loop frameworks that solve the challenge of bias dynamically at deployment. Wang et al. (2021) [26] have proposed iterative correction mechanisms that use user feedback to improve fairness in AI. Recently, critical activities have taken place to handle gender bias in financial services. Likewise, Parra et al. (2022) [24] highlighted the role that decorrelating variables has in lessening proxy bias via the removal of disparities within input features. In conclusion, many meaningful advances have been made in the development of fairness-aware machine learning techniques. However, ensuring ethical and unbiased outcomes in AI-based lending remains a complex and evolving challenge. The emerging trends discussed - such as real-time fairness monitoring, intersectional fairness, and the integration of explainability—reflect a growing awareness of both technical and societal dimensions of algorithmic bias. However, these innovations also highlight the need for ongoing interdisciplinary collaboration, stronger regulatory guidance, and scalable, context-sensitive solutions. As financial institutions continue to adopt AI technologies, it is essential that fairness is treated not as an optional add-on but as a core design principle embedded throughout the development and deployment process.

A. Research Directions

Despite these advances, important research gaps and opportunities remain. First is an increasing need for the actual development of real-time fairness monitoring systems that can automatically detect and adapt to changing biases. Such systems can help financial institutions uphold the concept of fairness as time changes with data distribution. Moreover, further work is to be done on improving the so-called fairness-aware algorithms, such as Random Forest models, which balance between fairness and predictive performance, as was shown in the work of Karimova (2024) [4], in order to reach the optimal solution for such a diverse and complex financial context [27].

[1]This scale, introduced by Purificato et al. (2023), is a previously established framework used to assess user confidence in AI decisions. It is not newly developed in this manuscript.

Received: April 3, 2025. Reviewed: April 16, 2025. Accepted: April 22, 2025. *© 2025 Suraiyo Raziyeva and Meraryslan Meraliyev. All rights reserved.*

TABLE II
CASE STUDIES ON BIAS MITIGATION IN AI-BASED LOAN APPROVAL SYSTEMS.

Case Study	**Authors (Year)**	**Key Contributions**	**Application**
Fair Lending Tool for Credit Decisions	Erasmo Purificato et al. (2023)	Developed fairness-aware frameworks incorporating Explainable AI (XAI) and Trust & Reliance Scale for evaluating AI systems.	Fair credit assessment and monitoring
Transparent Decision Support for Credit Risk	Abhinaba Dattachaudhuri et al. (2022)	Proposed a neural network-based decision support system enhancing transparency through rule extraction and pruning techniques.	Transparent credit scoring
Algorithmic	Jason Jia-Xi	Critiqued existing legal frameworks like ECOA and advocated harm-based frameworks to address AI-driven discrimination.	Credit underwriting fairness
Gender Bias in Loan Allocations	Ayesha Nadeem et al. (2022)	Examined gender biases in loan allocations and proposed integrating fairness-aware algorithms with organizational changes for equitable outcomes.	Gender bias mitigation in loan allocations
Interactive Approach to Bias Mitigation in ML	Hao Wang et al. (2021)	Introduced an iterative human-in-the-loop approach for bias mitigation, combining feedback loops and interactive visualization.	Iterative bias detection in ML applications
Federated Learning for Loan Approvals	Zhou et al. (2022)	Proposed federated learning approaches to preserve privacy while improving fairness and reducing geographical biases in decentralized training datasets.	Privacy-preserving credit risk evaluation
Addressing Proxy Bias in Lending	Parra et al. (2022)	Highlighted the role of proxy variables like ZIP codes in embedding discrimination and proposed mitigation strategies such as variable decorrelation.	Mitigating proxy bias in lending
Bias in Microfinance Systems	Cozarenco & Szafarz (2015)	Analyzed biases in microfinance systems, particularly gender-related biases, and emphasized co-financing mechanisms to address discriminatory patterns.	Gender fairness in microfinance
SHAP for Fairness in Credit	Lorenzo (2019)	Demonstrated the use of SHAP for interpreting credit scoring models, emphasizing its application in detecting and mitigating algorithmic bias.	Explainable AI for credit scoring
Fairness Metrics in Loan Decisions	Karimova (2024)	Introduced fairness-aware optimization techniques using Random Forest models to evaluate and reduce biases in small business loan decisions.	Optimizing fairness in small business loans

In the future, fairness metrics would need further refinement for capturing intersectional biases that disproportionately affect people who belong to more than one disadvantaged group. Interdisciplinary collaboration between practitioners in AI, regulators, and social scientists is badly needed to develop ethical frameworks that align technological innovation with societal values. Increasing federated learning beyond the currently explored use cases might advance inclusive global financial systems and safeguard data privacy.

Received: April 3, 2025. Reviewed: April 16, 2025. Accepted: April 22, 2025. *© 2025 Suraiyo Raziyeva and Meraryslan Meraliyev. All rights reserved.*

B. Discussion

AI-powered loan decision engines have turned a new leaf and become truly transformative forces that change the paradigm of operation scale and efficiency in the realm of finance. However, recurrent bias raises problems in ethical and regulatory paradigms, raising some serious questions about algorithm fairness. Recent work suggests some ways to diminish these; at a similar scale, there continue to be many challenges to reaching equal, transparent AI systems. One of the most relevant developments is the development of frameworks and tools for fairness. These frameworks allow explainable AI and transparency that relates to a clear look at how AI comes up with its decisions to stakeholders; for example, Lorenzo has shown that it can be done through better interpretability of credit scoring using SHAP [29]. On the other hand, Purificato et al. (2023) have proposed the Trust, Reliance Scale, an approach to embed fairness in AI systems so that the latter could be held accountable for the decisions made [29]. The key issue is gender bias in loan distribution. Nadeem et al. (2022) have proposed a socio-technical fairness framework that can serve to overcome systemic and algorithmic biases, especially affecting provisions for financial services related to women [15]. Another point of interest is proxy bias. Examples of strategies, such as variable de-correlation, that may be used for eliminating the indirect discrimination given by features like ZIP codes or levels of education are provided by Parra et al. (2022) [16]. Other techniques, such as federated learning, took the cause further in terms of fairness. Zhou and Tang 2022 showed how Federated learning decreases geographical and demographic biases by taking away the centralization of the training of the model [25]. This allows inclusiveness while providing protection against sensitive information. Iterative bias mitigation through human-in-the-loop frameworks discussed by Wang et al. 2021 allows dynamic changes in deployment for evolving biases [26]. Advantages, however, have many problems. The problem of bias in AI models is dynamic and thus can always be subject to the detection, capturing, and correction of constantly evolving disparities by certain monitoring mechanisms. Karimova, 2024 [4]. The paper proposes a fairness-aware optimization of Random Forest algorithms that strikes a balance between predictive accuracy and fairness in granting loans to small businesses [17]. At a wider scale, such implementation still remains resource-intensive. Another important and complex issue to mention is the intersectionality of biases: persons having more than one disadvantaged group they identify with. Most of the existing fairness metrics are not correctly positioned to handle these interactions in compound ways. This calls for future research in developing holistic metrics that address nuances around intersectional fairness. Integrating ethical frameworks: this aligns technological advancements with placing value on society [28]. In return, this would require collaboration among AI researchers, regulators, and social scientists. Much has been done, yet significant steps to handle the biases of AI in loan approval systems are long in coming. Dynamic fairness-monitoring systems, intersectional fairness metrics, and robust frameworks pave the way for equity in financial decision-making. In this regard, further innovation will have to negotiate a delicate balance between ethical and regulatory imperatives if the gap between technical capability and societal expectation is to be bridged [29].

V. Conclusion

This study investigated the sources of bias in AI-driven loan approval systems and critically evaluated fairness-aware machine learning techniques designed to mitigate such biases. It examined a spectrum of approaches: pre-processing, in-processing, and post-processing methods-and discussed their comparative effectiveness, implementation challenges, and the trade-offs they present between model fairness and predictive accuracy. Special attention was given to how these techniques align with regulatory frameworks and ethical standards, particularly in high-stakes financial decision-making.

In addition, the study identified a number of emerging tools and frameworks-such as explainable AI (XAI), federated learning, and human-in-the-loop architectures-that hold promise for making AI systems more transparent, accountable, and adaptable to real-world complexities. These advances support auditing and improving credit scoring algorithms.

Despite this progress, the study revealed several ongoing limitations. Chief among them is the lack of dynamic fairness monitoring systems that can detect and adjust for bias as models interact with evolving data environments. Moreover, current mitigation strategies often fall short in addressing intersectional biases-situations where individuals experience multiple, overlapping forms of disadvantage based on race, gender, socioeconomic status, and other factors.

To effectively confront these issues, the paper underscores the need for sustained interdisciplinary collaboration among AI developers, policymakers, legal scholars, and social scientists. Such collaboration is essential for translating abstract fairness principles into actionable design practices and policy guidelines. The paper strongly advocates for embedding fairness not as a secondary consideration, but as a core tenet throughout the life cycle of AI systems-from data collection and model training to deployment and evaluation.

Meeting this challenge will require more than algorithmic innovation; it will demand a deep ethical commitment and policy coherence to ensure that technological advancements do not exacerbate existing inequalities. Ultimately, fostering fairness in AI-

Received: April 3, 2025. Reviewed: April 16, 2025. Accepted: April 22, 2025. *© 2025 Suraiyo Raziyeva and Meraryslan Meraliyev. All rights reserved.*

based credit decision-making is not only a matter of compliance or performance-it is central to building systems that are trustworthy, inclusive, and aligned with broader goals of financial equity and social justice in an increasingly automated world.

REFERENCES

[1] J. Scott, *et al.*, "Revealing and mitigating racial bias and discrimination in financial services," *Journal of Social Equity in Finance*, vol. 11, no. 3, pp. 178–201, 2023.

[2] A. Marshall, *et al.*, "Variable reduction, sample selection bias, and bank retail credit scoring," *Journal of Financial Modeling and Analytics*, vol. 6, no. 1, pp. 56–78, 2010.

[3] S. Priya and R. Kumari, "Loan approval prediction using machine learning," *Journal of Predictive Analytics in Finance*, vol. 7, no. 3, pp. 156–178, 2024.

[4] N. Karimova, "Application of AI in credit risk scoring for small business loans," *Journal of Financial AI Applications*, vol. 10, no. 1, pp. 89–112, 2024.

[5] A. O. Abhulimen, *et al.*, "Ensuring fairness in AI-driven financial services," *International Journal of Financial Ethics*, vol. 12, no. 1, pp. 67–85, 2024.

[6] H. Wang, *et al.*, "An interactive approach to bias mitigation in machine learning," *Journal of Machine Learning Ethics*, vol. 7, no. 2, pp. 178–201, 2021.

[7] A. Khaleghi, "Towards achieving gender equality in automated loan approval processes," *Journal of Financial Technology Ethics*, vol. 9, no. 4, pp. 112–134, 2020.

[8] S. Priya and R. Kumari, "Loan approval prediction using machine learning," *Journal of Applied Data Science*, vol. 8, no. 1, pp. 89–110, 2024.

[9] J. J. X. Wu, "Algorithmic fairness in consumer credit underwriting," *Journal of Financial AI Compliance*, vol. 6, no. 2, pp. 78–96, 2024.

[10] S. Krishnaraj, *et al.*, "Comparing machine learning techniques for loan approval prediction," *Journal of Data-Driven Finance*, vol. 10, no. 1, pp. 245–267, 2024.

[11] A. Dattachaudhuri, *et al.*, "Transparent decision support system for credit risk evaluation," *Neural Networks and Decision Making*, vol. 11, no. 4, pp. 67–89, 2022.

[12] E. Purificato, *et al.*, "Responsible AI techniques in loan approval processes," *AI in Finance Review*, vol. 14, no. 3, pp. 99–121, 2023.

[13] L. Belenguer, "AI bias: Exploring machine-centric solutions," *AI Ethics and Development Quarterly*, vol. 13, no. 2, pp. 112–130, 2022.

[14] A. Castelnovo, "Towards responsible AI in banking," *AI Governance and Ethics Journal*, vol. 9, no. 1, pp. 56–74, 2023.

[15] A. Nadeem, *et al.*, "Gender bias in AI-based decision-making systems," *Journal of Gender and Technology*, vol. 7, no. 3, pp. 198–215, 2022.

[16] C. M. Parra, *et al.*, "Likelihood of questioning AI-based recommendations due to perceived racial/gender bias," *Journal of Social and Ethical AI*, vol. 7, no. 3, pp. 201–224, 2022.

[17] T. Ndayisenga, "Bank loan approval prediction using ML techniques," *Journal of Predictive Analytics in Finance*, vol. 9, no. 3, pp. 145–167, 2021.

[18] B. Low, *et al.*, "Risk-informed and AI-based bias detection using Gen-Z survey data," *Journal of Social Justice and AI*, vol. 10, no. 4, pp. 178–198, 2023.

[19] F. Lorenzo, "Techniques for trustworthy AI in loan approval," *International Journal of Financial Technology*, vol. 8, no. 1, pp. 45–62, 2019.

[20] A. Marshall and R. McManus, "Variable reduction, sample selection bias, and bank retail credit scoring," *Journal of Banking Analytics*, vol. 15, no. 2, pp. 89–115, 2010.

[21] A. Cozarenco and A. Szafarz, "Gender biases in bank lending: Lessons from microcredit in France," *Journal of Microfinance Studies*, vol. 18, no. 1, pp. 67–84, 2015.

[22] A. Kartal, "Mitigating digital discrimination in credit decisions," *AI Fairness Review*, vol. 14, no. 2, pp. 89–112, 2022.

[23] R. Brody, *et al.*, "The potential for biases in resolving loan problems," *Journal of Financial Risk Management*, vol. 13, no. 4, pp. 201–224, 2021.

[24] C. M. Parra, *et al.*, "Likelihood of questioning AI-based recommendations due to perceived racial/gender bias," *Journal of Ethical AI Practices*, vol. 11, no. 1, pp. 145–167, 2022.

Received: April 3, 2025. Reviewed: April 16, 2025. Accepted: April 22, 2025. *© 2025 Suraiyo Raziyeva and Meraryslan Meraliyev. All rights reserved.*

[25] N. Zhou, *et al.*, "Bias, fairness, and accountability in AI/ML," *Journal of Fairness in Machine Learning*, vol. 8, no. 1, pp. 123–145, 2021.
[26] H. Wang, *et al.*, "An interactive approach to bias mitigation in machine learning," *Journal of Machine Learning Fairness*, vol. 8, no. 3, pp. 156–178, 2021.
[27] T. Ndayisenga, "Bank loan approval prediction using ML techniques," *Journal of Predictive Analytics in Finance*, vol. 9, no. 3, pp. 145–167, 2021.
[28] T. Beck, P. Behr, and A. Madestam, "Sex and credit: Is there a gender bias in lending?," *Journal of Financial Inclusion Studies*, vol. 9, no. 2, pp. 112–134, 2017.
[29] R. Brody, *et al.*, "The potential for biases in resolving loan problems," *Journal of Credit Risk Studies*, vol. 11, no. 4, pp. 201–224, 2021.
[30] D. E. Rigobon, "From Utilitarian to Rawlsian Designs for Algorithmic Fairness," *arXiv preprint arXiv:2302.03567*, 2023.
[31] D. Card and N. A. Smith, "On Consequentialism and Fairness," *Frontiers in Artificial Intelligence*, vol. 3, p. 34, 2020.
[32] J. Rawls, *A Theory of Justice*, revised ed., Harvard University Press, 1999.
[33] S. Buijsman, "Navigating Fairness Measures and Trade-offs," *AI and Ethics*, vol. 4, pp. 1323–1334, 2024.
[34] J. Kuppler et al., "Formalizing Trade-offs Beyond Algorithmic Fairness: Lessons from Ethical Philosophy and Welfare Economics," *AI and Ethics*, vol. 1, pp. 529–544, 2021.
[35] N. Thieme, "Will Computer Algorithms Support Equity or Reinforce Inequality?," *Undark Magazine*, 2018.
[36] J.-C. Bélisle-Pipon, "AI, Universal Basic Income, and Power: Symbolic Violence in the Tech Elite's Narrative," *Frontiers in Artificial Intelligence*, vol. 3, article 1488457, 2025.

Received: April 3, 2025. Reviewed: April 16, 2025. Accepted: April 22, 2025. *© 2025 Suraiyo Raziyeva and Meraryslan Meraliyev. All rights reserved.*

Article

A Survey on Multimodal Approaches for Lung Disease Diagnosis using Deep Learning

Zhaniya Medeuova [1]

[1]Department of Computer Science, SDU University, Almaty, Kazakhstan

DOI: 10.47344/tx89w092

Abstract

Lung disorders are a major global health issue. A quick and accurate diagnosis is essential for proper treatment. In order to increase diagnostic accuracy, recent multimodal techniques are gaining popularity. This study carried out a comprehensive analysis of research articles on multimodal approaches that were published between 2020 and 2024 in Scopus and Google Scholar. The results show that there is limited study on the multimodal approach and on a variety of lung disorders such as asthma, TB, pneumonia, and chronic obstructive pulmonary disease. Several studies concentrated mainly on the detection and binary classification of COVID-19. The field has several challenges, including limited datasets, high computing costs, difficulties in integrating multiple modalities, and lack of accessibility of the models. Future studies should look at a wider range of lung diseases, increase the accessibility of datasets, improve fusion methods for merging data from many sources, and create models that are easier to understand and use fewer resources. Resolving these issues will improve patient outcomes by advancing the real-world use of deep learning in medical diagnosis.

Keywords: deep learning, multimodal approach, lung diseases, medical imaging, lung sounds, regression, classification, diagnostics.

I. Introduction

The respiratory system plays a crucial role in the human body, facilitating the exchange of oxygen and carbon dioxide [1]. Despite its flexibility, it remains at risk for numerous diseases that can significantly affect lung function and overall human health. Lung diseases cover a broad category of disorders such as pneumonia, tuberculosis, chronic obstructive pulmonary disease, and lung cancer, among others. These diseases are a major cause of morbidity and mortality on a global scale [2].

The World Health Organization informs that in 2019 around 3.23 million victims were COPD. In the same year, it was reported that chronic respiratory diseases were responsible for 4 million deaths overall. In the United States, asthma affects more than 23.3 million adults and 6.6% children, resulting in significant treatment costs and reduced quality of life [3]–[5]. Furthermore, in 2024

Email: zhaniya.medeuova@sdu.edu.kz ORCID: 0009-0004-7409-9792

Received: March 10, 2025. Reviewed: March 18, 2025. Accepted: March 20, 2025. © 2025 Zhaniya Medeuova. All rights reserved.

Kazakhstan had one of the highest rates of lung disorders globally [6]. These statistics highlight the impact of lung diseases on global health and the need for better diagnostic methods that can quickly and accurately identify diseases.

Traditional methods for diagnosing lung disease are medical history reviews, blood tests, lung sound, chest X-rays, and CT scans, etc. [7]. However, these methods have their own drawbacks, such as the dependence on expert analysis and limited accessibility in the environment. Sometimes, these methods can be the cause of human error. That is why manual checking and image-based analysis emphasize the need for more automated and standardized diagnostic processes [8], [9].

Nowadays deep learning has become a solution for these issues, providing precise and automatic diagnostic skills. Due to the increasing availability of medical imaging and acoustic data, researchers have created deep learning algorithms that can accurately identify lung problems [10]. In order to identify diseases, these models have shown remarkable success in evaluating lung sound recordings, CT scans, and chest X-rays. Notable developments include the application of Recurrent Neural Network for lung sound analysis and Convolutional Neural Network for image based classification. For instance, Çallı et al. emphasized the efficacy of deep learning models like VGGNet and ResNet in chest X-ray processing, Ahmed et al. investigated CNN based architectures for lung disease identification using chest imaging [11], [12]. Likewise, Sfayyih et al. examined the function of acoustic signal analysis in identifying lung diseases, stressing the significance of CNN models based on spectrograms [13].

Kieu et al. examined 98 research from 2016 to 2020. They presented a taxonomy that included ensemble techniques, algorithms, transfer learning, augmentation, and features. Large image sizes, a lack of publicly available datasets, data imbalance, and significant error correlation in ensemble models are some of the main issues noted. In order to overcome these problems, the authors proposed using cloud computing, different feature extraction, dataset sharing, and enhanced ensemble approaches. This survey article offers insightful information, more research is necessary given recent developments in datasets and model designs [14]. AI based lung sound categorization for the diagnosis of respiratory diseases was reviewed by Wanasinghe et al., who highlighted developments in deep learning models, data augmentation, feature extraction, and explainability. With fusion models reaching up to 98% accuracy, CNN performed incredibly. However, several obstacles persist, such as the scarcity of datasets, the dependence on individual feature representations, and the absence of explainable AI methodologies. Developing clinical support tools for real-world applications, increasing model interpretability, and diversifying datasets should be the main goals of future research [9]. In their assessment of deep learning-based acoustic analysis for lung disease diagnosis, Sfayyih et al. emphasized the expanding use of Deep Learning Convolutional Neural Networks (DLCNNs) in the detection of obstructive lung diseases. There are no as many reviews on signal-based lung disease detection as there once was. Although they show potential, DLCNNs need to be further validated through extensive research. Data standardization, clinical acceptance, and enhancing diagnostic reliability should be the main areas of future study to assist industry applications and medical practitioners [13].

Despite these developments, most of the other research has focused on single-modal strategies that leverage acoustic analysis, medical imaging, or other discrete data sources. Deep learning techniques for lung illness diagnosis have been evaluated in a variety of survey publications, these researchers mainly focus on single-method approaches such as respiratory sound categorization or CNN-based medical imaging analysis or other types of data [15]. On the other hand, diagnosing lung disease usually requires a variety of clinical data sources, such as the patient's medical history, symptoms, and other relevant information. The multimodal approach can improve diagnostic accuracy, reduce biases, and increase predictability by integrating multiple data sources [15]. And this survey aims to close this gap by providing an overview of multimodal deep learning methods for diagnosing lung diseases. The objectives include assessing the effectiveness of multimodal models, identifying challenges that retard the progression in this field, and exploring solutions that can be implemented to improve model accessibility and performance in a variety of lung diseases.

The following sections present a detailed review of multimodal deep learning techniques. The second section describes the strategy used to collect and examine the relevant literature, including research published in Russian, Kazakh, and English. The third section outlines the fundamental steps needed for deep learning applications, including feature extraction, data preprocessing, model training, and evaluation. The fourth section classifies current techniques and examines breakthroughs in this area. In conclusion, the importance of deep learning in improving the diagnosis of lung diseases and the potential impact of multimodal approaches will be addressed.

II. Methodology

This research uses a systematic process to identify and analyze recent work on the multimodal approach. The methodology is divided into major steps that include the process of selecting the articles, the filtering process, and the analysis of the selected articles. The research was carried out in the Scopus and Google Scholar databases, with an emphasis on Q1-ranked papers published between 2020 and 2024. The research terms used were a combination of phrases such as ”deep learning”, ”detection”, ”lung disease”

Received: March 10, 2025. Reviewed: March 18, 2025. Accepted: March 20, 2025. *© 2025 Zhaniya Medeuova. All rights reserved.*

(including asthma, chronic obstructive pulmonary disease, COPD, lung cancer, tuberculosis, pneumonia, COVID-19) and with terms like ”image”, ”audio”, and ”sound” to ensure that suitable research is obtained.

The selection process is summarized in Figure 1b. The initial search yielded 535 papers from Scopus and 550 from Google Scholar. A filtering process was then applied to exclude duplicate records and retain only studies that explicitly utilized both image and audio or sound data in a multimodal approach. This step reduced the selection to 47 studies. Further eligibility screening was performed on the basis of predefined inclusion and exclusion criteria. The inclusion criteria required studies to focus on multimodal deep learning models for lung disease detection, provide clear experimental results and evaluation metrics, be published in English, Russian, or Kazakh and appear in peer-reviewed journals or conferences. Studies were excluded if they used only a single data modality (either image or audio), covered diseases beyond the scope of this research, or lacked clear methodological details or experimental validation.

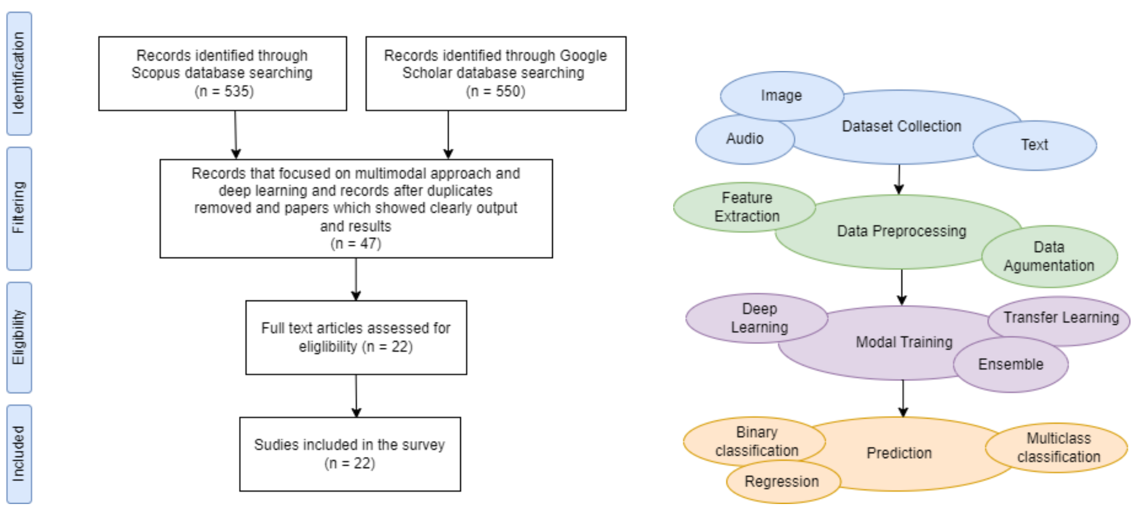
Following this process, 22 articles were considered eligible for inclusion in the final survey. These selected studies provided meaningful information on current trends and challenges of multimodal deep learning in lung disease detection. And the results of recent studies are summarized in Table I to provide a better understanding of the different modalities and their uses in the diagnosis of lung diseases.

Table I summarizes the various research studies that were analyzed in this survey, emphasizing the variety of modalities, datasets, and neural network architectures that were used. This indicates the diversity of approaches currently being explored in the field of lung disease diagnosis using multimodal deep learning techniques.

This methodology section included the selection of relevant studies, a filtering process was used to ensure that only multimodal approaches were included, and the final set of studies was assessed using predefined criteria. The selected articles provide information on current trends, challenges and advances in the integration of multiple data modalities for improved diagnostic accuracy.

III. Fundamental Steps in Applying Deep Learning for Lung Disease Detection

Deep learning plays an essential role in the identification of lung diseases by analyzing medical images and patient data. The process consists of four key steps, they are data collection, data preprocessing, training model, and prediction making [14]. The overview of the process is illustrated at Figure 1b.



(a) (b)

Fig. 1: (a) The survey methodology, (b) Overview of using DL for lung disease detection

Received: March 10, 2025. Reviewed: March 18, 2025. Accepted: March 20, 2025. © 2025 Zhaniya Medeuova. All rights reserved.

TABLE I: Summary of multimodal deep learning approaches for lung disease diagnosis

Study	Modality	Datasets Used	Neural Network Architecture	Key Results
Kumar et al., 2023 [18]	img + text	Manually collected (289 patients, future 65k records)	DenseNet121, ResNet50, LSTM, SVM fusion	Intermediate fusion improved accuracy by 2.9%
Malik et al., 2024 [19]	img + audio	24 public datasets (CXR, Cough sound, RSNA, etc.)	CNN + BANL, RBAP, MWDG	Achieved SOTA performance across diseases
Kumar et al., 2024 [20]	img + text	3,256 patient records (India)	CNN, Denoising Autoencoder, Cross-Modal Transformer	Addressed data imbalance, high accuracy for TB classification
Abhishek et al., 2024 [21]	img + audio	1,979 respiratory sound recordings	Hybrid CNN-GRU model	High accuracy in common respiratory diseases, overfitting risk
Sangeetha et al., 2024 [22]	img + text	TCIA, TCGA	MFDNN, CNN, DNN, Intermediate Fusion	92.5% accuracy in lung cancer classification
Varunkumar et al., 2024 [23]	img + img	RIDER Lung CT, Kaggle X-ray	CNN with dilated convolutions, multimodal fusion	Limited dataset diversity, generalizability issues
Hamdi et al., 2021 [24]	img + text	Public IPF dataset (33,026 CT + 1,549 records)	EfficientNet, DenseNet, LSTM, Attention Fusion	Multimodal integration improved prediction accuracy
Kumar et al., 2024 [25]	img + audio + text	AIIMS, Raipur (CT, X-ray, cough, lung sounds)	EfficientNet, RNN, U-Net, OpenL3, RVFL neuro-fuzzy model	COPD prediction using multimodal fusion
Deng et al., 2024 [35]	img + text	East China hospitals, Kaggle COVID-19 CT	CNN + Contrastive Learning + Early Fusion	Contrastive learning improved performance, Grad-CAM interpretation
Adeshina et al., 2022 [26]	img + audio	COVIDx, SARS-CoV-2 CT-scan dataset	CNN, ResNet, DenseNet, XResNet, Self-Attention	91.07% accuracy, effective multimodal cascaded approach
Thandu et al., 2024 [27]	img + audio	Chest X-ray (COVID-19 Radiography) + COUGHVID	DSPANN + Blockchain-based Privacy (ECHFA)	Data quality challenges, complex attention mechanisms
Liu et al., 2024 [28]	img + text	4 hospitals (China), Chest CT, Clinical Features	DenseNet-201 + DNNs + Early Fusion	Outperformed junior radiologists, 11 key clinical features identified
Farhan et al., 2023 [29]	img + img	CXRTD, PCXRA, CCSC, NIH Chest X-ray	CNN, LSTM, SVM, Decision Tree	Improved severity grading performance
Lay et al., 2022 [30]	img + text	Shenzhen, Montgomery X-ray Dataset	EfficientNet, XGBoost, U-Net	AUC improved by 0.0213 over unimodal models
Mayya et al., 2021 [36]	img + text	COVID-19 Chest X-ray, RSNA Pneumonia Dataset	ResNet18, NLP, Grad-CAM, Deep NN Ensemble	X-ray + diagnosis reports enhanced accuracy
Wu et al., 2021 [31]	img + text	TCIA (422 NSCLC patients)	3D-ResNet, Clinical Embedding Layer, Fusion	Improved survival prediction using multimodal fusion

Received: March 10, 2025. Reviewed: March 18, 2025. Accepted: March 20, 2025. © 2025 Zhaniya Medeuova. All rights reserved.

A. Dataset Collection and Data Preprocessing

When collecting data, data can be in the form of chest X-rays, CT scans, medical records of patients, coughing, and breathing sounds [10], [11]. Researchers choose between public medical databases or manually acquire data from hospitals and clinics. To ensure that the model can identify a wide variety of lung disorders, balanced data are crucial. Once data is collected, they are processed to make them clean and ready for use. This includes eliminating noise, improving image quality, and being standardized in terms of size and format. In medical imaging, pre-processing can be in altering contrast, segmentation of lung regions, and removal of extraneous detail. In non-image data, such as patient symptoms or audio, pre-processing can be used to structure information in a well-defined format. The purpose of this step is to clean the data so that the model learns only meaningful patterns [33].

B. Training the Model and Prediction

Before the training step, the model gets a large number of labeled samples to be able to understand its features and patterns of lung diseases. Researchers can use neural network architectures that are appropriate for medical image and sound analysis. During training time, the model continuously changes its internal parameters so that it can better identify diseases. A well-trained model predicts the results of the new data. After being trained, the model is tested with new images or patient data to verify its performance. When given a new X-ray or CT scan, the model makes a decision about whether a patient is healthy or has a specific lung disease [14]. Certain models also give us a probability score that informs us about how certain or confident the model is in its decision. This method can help physicians diagnose patients more quickly and accurately when it is integrated into a clinical workflow.

IV. Taxonomy and Trends in Multimodal Approaches for Lung Disease Diagnosis

This section shows the taxonomy and trends in multimodal approaches to the diagnosis of lung diseases. Modalities, feature engineering, data augmentation, fusion techniques, illness categories, and output types are the six key qualities into which the taxonomy groups the important methodologies used in recent studies. These attributes describe the methods of data acquisition, feature extraction, model enhancement, and prediction. These attributes are discussed in detail in subsections A to B, along with a study of the corresponding research.

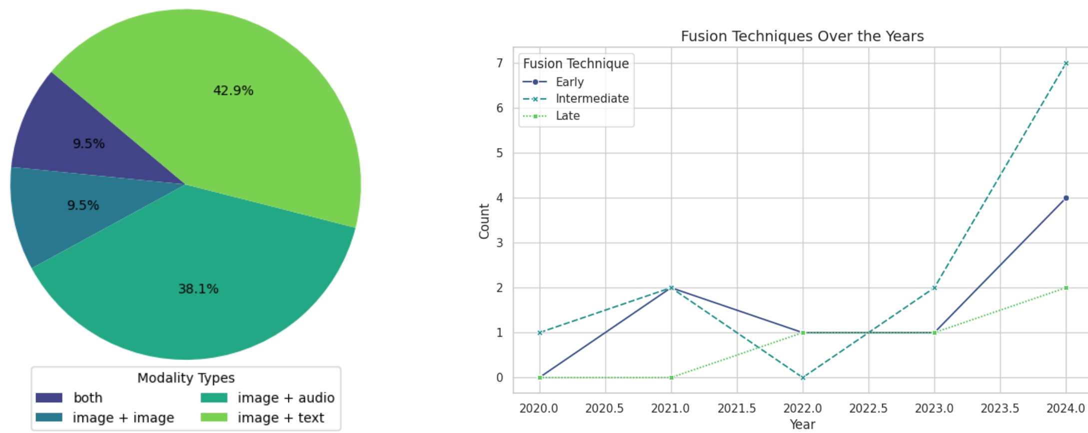
A. Modalities type

Lung disease detection using deep learning is based on various data modalities, often combining multiple sources for better accuracy. Figure 2a shows that some studies use only medical images, such as CT, X-rays, and PET scans, to identify lung abnormalities [29], [33]. Others improve detection by integrating images with respiratory or cough sounds, capturing both structural and acoustic patterns [15], [19], [21], [26], [27], [33], [37], [38]. Another approach combines images with clinical records, including patient demographics, diagnostic reports, and lab results, providing additional diagnostic context [18], [20], [22], [24], [28], [30], [31], [35], [36]. Studies using image and audio data focus primarily on COVID-19, pneumonia, tuberculosis, lung cancer, asthma, and COPD, while image and text combinations are commonly applied to lung cancer, tuberculosis, chronic bronchitis, and pulmonary fibrosis. Some research incorporates the three modalities: images, audio, and text, to improve disease prediction, particularly for COVID-19, COPD, and other complex respiratory conditions [17], [25], [32]. The choice of modality depends on the characteristics of the disease and the available diagnostic data, with multimodal approaches enhancing the accuracy of classification.

B. Feature engineering

Feature engineering is essential for the diagnosis of multimodal lung disease because it has a direct impact on the way deep learning models extract relevant representations from medical data. Handcrafted features and learned features are two main categories into which feature engineering methodologies can be divided. Medical pictures and audio data are manually processed to extract hand-crafted features based on domain-specific knowledge. Texture descriptors, shape characteristics, and statistical qualities are frequently used in imaging modalities, whereas Mel frequency cepstral coefficients (MFCC) and spectrum features are frequently used in audio-based diagnostics. On the other hand, deep learning models, in particular, Convolutional Neural Networks, which are suited to recognizing complex patterns in unstructured information without the need for explicit feature selection automatically extract learned features. Using pre-trained architectures like VGG19, Inception-v3, ResNet, DenseNet, and EfficientNet to increase feature extraction and classification performance, transfer learning has been widely used in recent research. These models are refined on lung disease datasets to extract high-level features relevant to illness detection after being pre-trained on vast datasets. In order

Received: March 10, 2025. Reviewed: March 18, 2025. Accepted: March 20, 2025. *© 2025 Zhaniya Medeuova. All rights reserved.*



(a) (b)

Fig. 2: (a) Distribution of Modalities, (b) Fusion techniques over time

to minimize dimensionality and maintain the most discriminative features, some research incorporates feature selection methods such as principal component analysis (PCA) and recursive feature elimination (RFE) in addition to feature extraction based on deep learning [18], [22]. This improves the performance of the model. Furthermore, hybrid techniques that integrate learned and handcrafted features have attracted a lot of interest since they allow for a more thorough representation of multimodal data, which eventually improves diagnostic adaptability and accuracy. Multimodal approaches can improve lung disease detection by using these feature engineering techniques to capture high- and low-level data representations, which will improve prediction performance.

C. Data augmentation

Deep learning-based lung disease identification often employs data augmentation to improve model generalization and address data limitation. Rotation, scaling, translation, flipping, contrast alterations, and noise injection are popular augmentation procedures in medical imaging. For specialization on lung regions, some investigations use segmentation-based augmentations such as cropping and scaling. Furthermore, image quality is enhanced by preprocessing techniques such contrast limited adaptive histogram equalization (CLAHE) and histogram matching [36]. Using pitch shifting, temporal stretching, noise injection, and speed perturbation, augmentation techniques alter respiratory sounds for audio-based classification [33]. These techniques help models adjust to changes in recording conditions and sound quality. Furthermore, by increasing the representation of imbalanced classes, data balancing techniques such as MWDG (Multiple-Way Data Generation) and SMOTE (Synthetic Minority Oversampling Technique) reduce model bias [19]. Horizontal flipping, rotation, and width/height shifts are used in public datasets such as POCOVID-Net and NIH Chest X-Ray, in addition to preprocessing techniques such as CLAHE and scaling. Principal component analysis (PCA), image embedding, clustering for defect detection, and Fourier transform are the complex augmentation methods. They are frequently used in manually collected datasets. Preprocessing techniques such as wavelet transformations, noise reduction, and Mel frequency cepstral coefficients (MFCC) improve the accuracy in audio samples [37]. Augmentation has drawbacks despite its benefits. Unrealistic data produced by excessive changes can result in poor model generalization [37]. Complex procedures raise computing costs, and improper augmentation strategies could result in biases. Additionally, broad, high-quality real-world data is still necessary for developing a strong and reliable deep learning model, and augmentation cannot completely replace it.

D. Fusion techniques

In order to improve the quality and strength of computational models, fusion techniques are essential for combining various data sources. Figure 2b shows that several fusion strategies have been used, such as early fusion (E), intermediate fusion (I),

Received: March 10, 2025. Reviewed: March 18, 2025. Accepted: March 20, 2025. *© 2025 Zhaniya Medeuova. All rights reserved.*

and late fusion (L), according to the reviewed publications. The method by which and when the data is joined during processing differ in these methods, which affects model performance and computing efficiency. With 10 experiments, intermediate fusion was the most commonly utilized strategy among the 22 papers surveyed [17], [20]–[24], [27], [29], [33]. Before making a final judgment, features that have been retrieved from several modalities or sources are combined using feature-level integration, which is a common component of intermediate fusion approaches. The Progressive Split Deformable Field Fusion Module (PSDFM), which uses intermediate fusion to improve representation learning, is a notable example [27]. Seven studies used early fusion (E), suggesting a preference for input-level direct data integration [15], [19], [28], [31], [32], [35], [36]. This method is frequently used in situations where it is possible to efficiently mix raw data from many sources prior to feature extraction. Four articles reported the use of late fusion (L), which combines predictions from different models and is frequently used in ensemble-based techniques to increase the accuracy of regression or classification [25], [30], [37]. The flexibility of fusion techniques in complicated problem domains was demonstrated by certain papers that used a combination of fusion procedures, such as L, I and E, I [18], [26].

However, a study specifically mentioned the lack of fusion techniques, implying that independent processing of data sources would be better in some circumstances. The performance of the model is significantly affected by the fusion technique method. Intermediate fusion often outperforms early and late fusion because it allows feature representations from multiple modalities to be refined before final decision making, leading to more discriminative patterns. However, it can be challenging to compute [26]. On the other hand, early fusion ensures that raw data is combined before feature extraction, which can be valuable when different modalities share a common feature space but may struggle with heterogeneous data [18]. Late fusion provides flexibility by allowing independent model predictions to be combined, but may not fully leverage interactions between different data sources. The effectiveness of each method depends on factors such as data heterogeneity, model complexity, and available computational resources. Studies have shown that hybrid approaches, such as the combination of early and intermediate fusion, can further improve performance utilizing data-level and feature-level integration [26].

In general, fusion methods are still being developed, and hybrid fusion models which use several levels of integration to optimize the advantages of various data sources are becoming progressively more popular. Future studies might concentrate on refining fusion techniques to strike a balance between prediction performance and computational economy across a range of application domains.

E. Disease types

The reviewed studies cover a broad spectrum of lung diseases, demonstrating the extensive application of computational models in clinical diagnosis. As shown in Figure 3b, COVID-19 was the most frequently occurring disease to be examined, occurring in nine studies, reaffirming its persistent relevance in clinical imaging [15], [17], [26], [27], [32], [36], [38]. Pneumonia was also a significant area of research, studies of its various forms, including bacterial, viral, lobar, lobular, and Staphylococcus aureus pneumonia (SAP) demonstrating the need for precise diagnostic models [15], [17]–[19]. Tuberculosis (TB) has also been explored frequently, with particular studies differentiating pulmonary TB [15], [19], [20], [37]. Other respiratory infections including bronchitis, lower and upper respiratory tract infections (LRTI, URTI), and bronchiolitis were also explored [37]. Chronic lung diseases sush as Chronic Obstructive Pulmonary Disease (COPD), asthma, and chronic bronchitis were also extensively explored, with the need for early diagnosis and long-term monitoring [25]. Lung cancer, particularly non-small cell lung cancer (NSCLC), was also a significant area of research in various studies [15], [19], [22]. Some studies also explored relatively uncommon but clinically important conditions including Idiopathic Pulmonary Fibrosis (IPF), pleural effusion, and pulmonary edema [24].

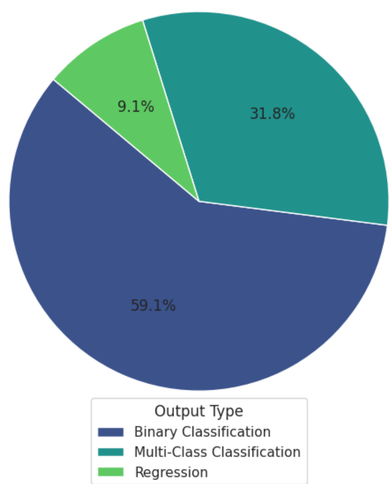
The studies used publicly available datasets or manually collected data. Most of the research used publicly available datasets, ensuring standardized imaging data for training and evaluation. However, some studies included manually collected datasets from hospitals and medical institutions, especially for diseases that are underrepresented in publicly available data [17], [18], [20], [21], [32], [35]–[37]. According to Table 1, ChestX-ray14, COVIDx, Tuberculosis Chest X-ray, RSNA Pneumonia Detection Challenge Dataset, and LIDC-IDRI are the public datasets most commonly used. Large-scale model training was made possible by these datasets, which offered categorized medical imaging data, eliminating the need for manual collection. A smaller number of studies applied datasets that were manually collected, mostly from imaging facilities and hospital records. For rare disorders where public datasets were not enough, such as pleural effusion, pulmonary fibrosis, or mixed-disease classification tasks, these datasets were especially valuable. In comparison to publicly available datasets, personally gathered datasets frequently have smaller sample numbers, but provide greater control over patient demographics and imaging conditions.

Large-scale model training is made easier by publicly accessible datasets, but these datasets frequently contain biases that may hinder the generalizability and performance of the models. Ethnic representation is an important issue. There is a lack of diversity in many large scale datasets, like ChestX-ray14 and COVIDx, because most of the images are taken from particular populations [36]. Because of this, models developed using these datasets might not work consistently across ethnic groups, which could reduce

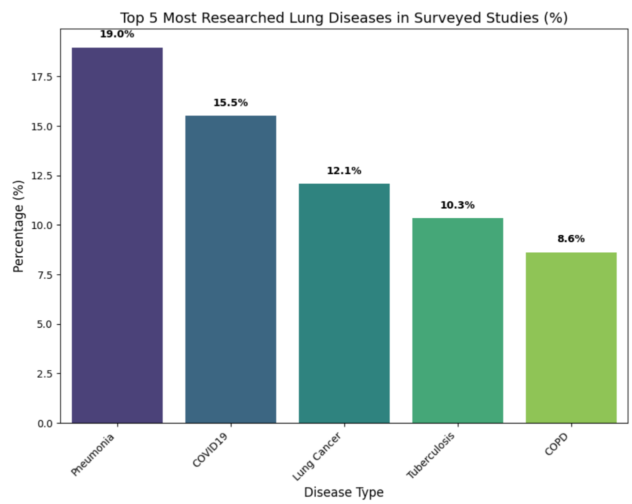
Received: March 10, 2025. Reviewed: March 18, 2025. Accepted: March 20, 2025. *© 2025 Zhaniya Medeuova. All rights reserved.*

the diagnostic accuracy for underrepresented groups. The distribution of ages is also a significant factor. Adult and elderly patients make up a larger percentage of many datasets, while young children are still underrepresented. For diseases like pneumonia and bronchiolitis, which occur differently in children than in adults, this can present difficulties. Models may perform less well in predicting outcomes for younger patients if they are not trained in a balanced age distribution. In addition, a common limitation is an imbalance in the severity of the disease. Since severe cases are more commonly diagnosed and documented in medical settings, they often make up a larger percentage of public datasets. This makes early stage diagnosis more challenging by biasing model training toward identifying diseases at a later stage. Early detection, which is essential for prompt medical intervention, may be difficult for models trained on unbalanced datasets.

The range of diseases covered in these studies highlights the need for strong deep learning models capable of addressing a variety of lung conditions. To improve predictability, future research may focus on improving classification performance in a range of diseases and ensuring that datasets incorporate world differences. Curating datasets that more accurately reflect a range of age groups, disease severity levels, and populations should be the main goal.



(a)



(b)

Fig. 3: (a) Distribution of output types in lung disease diagnosis studies, (b) Distribution of most studied lung diseases in multimodal research

F. Output types

Various types of output were used in the investigated research. Figure 3a shows that the three main types of these outputs were probabilistic estimation, regression based prediction, and classification. Classification tasks, especially binary classification, were a popular type among reviewed articles [18], [20], [22], [23], [25], [26], [28], [30], [32], [35], [36], [38]. A unique case was when a model was categorized according to severity levels rather than type of disease, including mild, moderate, severe, and deadly [29]. Also in regression models used to estimate patient disease severity. Using metrics like the MAE and Concordance Index to estimate survival time for patients with non-small cell lung cancer. Regression based methods were also employed to monitor the severity of COPD and the development of idiopathic pulmonary fibrosis. Probabilistic outputs, which provide confidence scores for the existence or severity levels of the diseases. In multiclass classification tasks, where probability distributions aided in improving decision making in unclear situations, such methods were frequently used. These probabilistic outputs were frequently evaluated using metrics like the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC). The metrics used for the evaluation were chosen based on the selection of the output type. F1-score, recall, specificity, accuracy, and precision were

Received: March 10, 2025. Reviewed: March 18, 2025. Accepted: March 20, 2025. *© 2025 Zhaniya Medeuova. All rights reserved.*

frequently used in binary classification models. Log Loss, Fowlkes-Mallows Index (FMI), and Matthews Correlation Coefficient (MCC) were used in multiclass classification studies. Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R2 score were commonly used to evaluate regression models. It is crucial to use these metrics depending on their strengths and limitations [14]. AUC-ROC and other performance metrics based on probability were commonly used to assess probabilistic models. However, it does not capture data imbalances as well as the F1 score, MCC [34]. Both RMSE and MAE give distinct viewpoints on prediction error in regression models, with RMSE penalizing larger errors progressively. The best evaluation method for a task can be chosen with the help of a structured comparison of these metrics. The comparative analysis presented in Table II underscores the importance of understanding the strengths and limitations inherent in different models, highlighting areas that require further exploration.

TABLE II: Comparison of multimodal models: strengths, weaknesses, and metrics

Study	Strengths	Weaknesses	Evaluation Metrics
Kumar et al., 2023 [18]	Adaptive batch sizes, effective multi-modal fusion	Small dataset, data quality issues	Accuracy, Precision, F1 Score
Malik et al., 2024 [19]	Early fusion, data augmentation	Data imbalance, high computational cost	Accuracy (99.01%), MCC, FMI
Kumar et al., 2024 [20]	Cross-modal attention, effective fusion	Small dataset, high computational cost	Accuracy (95%), AUC-ROC, MCC
Abhishek et al., 2024 [21]	Robust feature extraction, real-time processing	Limited class diversity, noisy data	Accuracy (98%), AUC, F1 Score
Sangeetha et al., 2024 [22]	Improved accuracy, effective feature extraction	Privacy concerns, AI interpretability	Accuracy (92.5%), Precision, Recall
Varunkumar et al., 2024 [23]	CNN for feature extraction, hierarchical fusion	Lack of diverse datasets, model interpretability	Accuracy (94%), F1 Score
Hamdi et al., 2021 [24]	CNN+LSTM fusion, attention mechanism	Lung segmentation noise, training complexity	Accuracy (97%), R^2 Score (91%)
Deng et al., 2024 [35]	Hierarchical fine-tuning, contrastive learning	Small dataset, overfitting risk	Accuracy (90.14%), F1 Score
Adeshina et al., 2022 [26]	End-to-end training, self-attention. Discriminative fine-tuning.	Complexity in training models. Sensitivity to hyperparameter tuning.	Accuracy (91.26%), XResNet
Thandu et al., 2024 [27]	Uses multimodal data fusion, achieves high diagnostic accuracy, blockchain for privacy	Scalability, interpretability	Accuracy (98%), AUC (97%)
Liu et al., 2024 [28]	Early fusion, transfer learning	Small sample size, imbalance	AUC (0.92), Accuracy (78%)
Farhan et al., 2023 [29]	CNN+handcrafted features, optimized CNN	Class imbalance, long training times	Accuracy (98.78%), F1 Score
Lay et al., 2022 [30]	Demographic data fusion, late fusion	Small dataset, generalization issues	AUC (0.9574)
Mayya et al., 2021 [36]	Feedback mechanism, Grad-CAM interpretability	Limited dataset, X-ray variability	Accuracy (97%)
Wu et al., 2021 [31]	3D-ResNet, batch normalization	Data variety issues, complex survival model	MAE (0.162), C-index (0.6580)

The reviewed studies highlight a growing trend toward the integration of multiple modalities, advanced feature engineering, and data fusion techniques to improve diagnostic accuracy. The taxonomy reveals that the majority of approaches rely on deep learning, leveraging handcrafted and learned features to optimize performance. Intermediate fusion emerges as the most effective method, striking a balance between enhanced representation learning and computational efficiency. Additionally, publicly available datasets remain the primary source for training models, despite concerns about data diversity. Upcoming advancements should focus on improving fusion techniques, guaranteeing dataset inclusivity, and resolving feature selection issues to increase the diagnostic

Received: March 10, 2025. Reviewed: March 18, 2025. Accepted: March 20, 2025. © 2025 Zhaniya Medeuova. All rights reserved.

accuracy for a wider variety of lung conditions.

V. Conclusion

This study investigated the application of deep learning to identify lung diseases by merging various data sets, including lung sounds and medical imaging. Studies show that, in contrast to the use of a single data type, multimodal techniques can increase diagnostic accuracy. But there are still a number of difficulties. The lack of studies that examine a broad spectrum of lung disorders is a major problem. Instead of classifying several lung diseases such as asthma, TB, pneumonia, and chronic obstructive pulmonary disease (COPD), the majority of current research concentrates on the detection or binary classification of COVID-19. This restricts how these models can be used in the real world. The difficulty of combining several data types in a way that improves model performance is another significant obstacle. Large, high-quality datasets are also necessary for deep learning models. However, there are not enough publicly accessible multimodal datasets that cover a range of lung disorders.

Furthermore, doctors find it difficult to believe the predictions made by AI models because they are sometimes complex and difficult to understand. The adoption of these techniques in hospitals with limited resources is further hampered by their high computing costs. It is essential to expand the focus of future studies to include lung conditions other than COVID-19. Improving techniques to efficiently integrate clinical, audio, and visual information can improve diagnosis. Creating larger and more balanced databases with a variety of disease categories should be another priority for researchers. Creating models that can operate with smaller datasets and reduce dependence on enormous amounts of labeled data is another crucial avenue. Enhancing transparency and explainability will contribute to a rise in medical professionals' trust. Lastly, to ensure that these complex algorithms can be applied successfully in actual medical situations, cooperation between AI researchers and healthcare professionals is essential. Deep learning can significantly improve early diagnosis and treatment for a variety of lung diseases by addressing these issues, ultimately improving patient outcomes.

References

[1] Y. Sugandi, I. Soesanti, and H. A. Nugroho, ”A Systematic Literature Review of Convolutional Neural Network Architecture for Lung Disease Detection,” in Proc. 2023 International Conference on Information and Communications Technology (ICOIACT), 2023, pp. 230-235. DOI: 10.1109/ICOIACT59844.2023.10455864.

[2] P. P. Jasmine, K. Kotecha, G. Rajini, K. Hariharan, K. Raj, K. Ram, V. Indragandhi, V. Subramaniyaswamy, and S. Pandya, ”Lung Diseases Detection Using Various Deep Learning Algorithms,” Journal of Healthcare Engineering, vol. 2023, pp. 1-13, 2023. DOI: 10.1155/2023/3563696.

[3] GBD 2019 Chronic Respiratory Diseases Collaborators, ”Global Burden of Chronic Respiratory Diseases and Risk Factors, 1990-2019: An Update from the Global Burden of Disease Study 2019,” EClinicalMedicine, vol. 59, p. 101936, 2023. DOI: 10.1016/j.eclinm.2023.101936.

[4] World Health Organization, ”Chronic Obstructive Pulmonary Disease (COPD),” WHO, Nov. 6, 2024. Available: https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd).

[5] J. Li, Y. Meng, L. Ma, S. Du, H. Zhu, Q. Pei, and X. Shen, ”A Federated Learning Based Privacy-Preserving Smart Healthcare System,” IEEE Transactions on Industrial Informatics, vol. 18, pp. 2021-2031, 2022. DOI: 10.1109/TII.2021.3098010.

[6] A. Gafizkyzy, ”Qazaqstan ökpe auruynan älem boiynşa üşınşı orynğa şyqqan,” Qazaqstan TV, Dec. 5, 2024. Available: https://qazaqstan.tv/news/203518/.

[7] K. Bartziokas, A. Papaporfyriou, G. Hillas, A. Papaioannou, and S. Loukides, ”Global Initiative for Chronic Obstructive Lung Disease (GOLD) Recommendations: Strengths and Concerns for Future Needs,” Postgraduate Medicine, vol. 135, 2022. DOI: 10.1080/00325481.2022.2135893.

[8] J. P. Allinson, N. Chaturvedi, A. Wong, I. Shah, G. C. Donaldson, J. A. Wedzicha, and R. Hardy, ”Early Childhood Lower Respiratory Tract Infection and Premature Adult Death from Respiratory Disease in Great Britain: A National Birth Cohort Study,” Lancet (London, England), vol. 401, no. 10383, pp. 1183–1193, 2023. DOI: 10.1016/S0140-6736(23)00131-9.

[9] T. Wanasinghe, S. Bandara, S. Madusanka, D. Meedeniya, M. Bandara, and I. De la Torre Díez, ”Lung Sound Classification for Respiratory Disease Identification Using Deep Learning: A Survey,” International Journal of Online and Biomedical Engineering (iJOE), vol. 20, pp. 1-15, 2024. DOI: 10.3991/ijoe.v20i10.49585.

[10] A. Ijaz, M. Nabeel, U. Masood, T. Mahmood, M. S. Hashmi, I. Posokhova, A. Rizwan, and A. Imran, ”Towards Using Cough for Respiratory Disease Diagnosis by Leveraging Artificial Intelligence: A Survey,” Informatics in Medicine Unlocked, vol. 29, p. 100832, 2022. DOI: 10.1016/j.imu.2021.100832.

Received: March 10, 2025. Reviewed: March 18, 2025. Accepted: March 20, 2025. *© 2025 Zhaniya Medeuova. All rights reserved.*

[11] S. Ahmed and S. Kadhem, "Using Machine Learning via Deep Learning Algorithms to Diagnose the Lung Disease Based on Chest Imaging: A Survey," International Journal of Interactive Mobile Technologies (iJIM), vol. 15, p. 95, 2021. DOI: 10.3991/ijim.v15i16.24191.

[12] E. Çallı, E. Sogancioglu, B. van Ginneken, K. G. van Leeuwen, and K. Murphy, "Deep Learning for Chest X-ray Analysis: A Survey," Medical Image Analysis, vol. 72, p. 102125, 2021. DOI: 10.1016/j.media.2021.102125.

[13] A. H. Sfayyih, N. Sulaiman, and A. H. Sabry, "A Review on Lung Disease Recognition by Acoustic Signal Analysis with Deep Learning Networks," Journal of Big Data, vol. 10, no. 1, p. 101, 2023. DOI: 10.1186/s40537-023-00762-z.

[14] S. T. H. Kieu, A. Bade, M. H. A. Hijazi, and H. Kolivand, "A Survey of Deep Learning for Lung Disease Detection on Medical Images: State-of-the-Art, Taxonomy, Issues and Future Directions," Journal of Imaging, vol. 6, no. 12, p. 131, 2020. DOI: 10.3390/jimaging6120131.

[15] H. Malik, T. Anees, A. S. Al-Shamaylehs, S. Z. Alharthi, W. Khalil, and A. Akhunzada, "Deep Learning-Based Classification of Chest Diseases Using X-rays, CT Scans, and Cough Sound Images," Diagnostics (Basel, Switzerland), vol. 13, no. 17, p. 2772, 2023. DOI: 10.3390/diagnostics13172772.

[16] R. Hertel and R. Benlamri, "Deep Learning Techniques for COVID-19 Diagnosis and Prognosis Based on Radiological Imaging," ACM Computing Surveys, vol. 55, 2022. DOI: 10.1145/3576898.

[17] U. Sait, G. L. K. V, S. Shivakumar, T. Kumar, R. Bhaumik, S. Prajapati, K. Bhalla, and A. Chakrapani, "A deep-learning based multimodal system for Covid-19 diagnosis using breathing sounds and chest X-ray images," Applied Soft Computing, vol. 109, p. 107522, 2021. DOI: 10.1016/j.asoc.2021.107522.

[18] S. Kumar, O. Ivanova, A. Melyokhin, and P. Tiwari, "Deep-learning-enabled multimodal data fusion for lung disease classification," Informatics in Medicine Unlocked, vol. 42, p. 101367, 2023. DOI: 10.1016/j.imu.2023.101367.

[19] H. Malik and T. Anees, "Multi-modal deep learning methods for classification of chest diseases using different medical imaging and cough sounds," PLoS One, vol. 19, no. 3, p. e0296352, 2024. DOI: 10.1371/journal.pone.0296352.

[20] S. Kumar and S. Sharma, "An Improved Deep Learning Framework for Multimodal Medical Data Analysis," Big Data and Cognitive Computing, vol. 8, no. 10, p. 125, 2024. DOI: 10.3390/bdcc8100125.

[21] S. Abhishek, A. Ananthapadmanabhan, T. Anjali, S. Remya, A. Perathur, and R. Bentov, "Multimodal Integration of Enhanced Novel Pulmonary Auscultation Real-Time Diagnostic System," IEEE MultiMedia, vol. PP, pp. 1–26, 2024. DOI: 10.1109/MMUL.2024.3422022.

[22] S. Skb, M. S. Kumar, P. Karthikeyan, H. Rajadurai, B. Shivahare, S. Mallik, and H. Qin, "An Enhanced Multimodal Fusion Deep Learning Neural Network for Lung Cancer Classification," Systems and Soft Computing, vol. 6, p. 200068, 2023. DOI: 10.1016/j.sasc.2023.200068.

[23] K. Varunkumar, M. Zymbler, and S. Kumar, "Multimodal Deep Dilated Convolutional Learning for Lung Disease Diagnosis," Brazilian Archives of Biology and Technology, vol. 67, 2024. DOI: 10.1590/1678-4324-2024231088.

[24] A. Hamdi, A. Aboeleneen, and K. Shaban, "MARL: Multimodal Attentional Representation Learning for Disease Prediction," in Proc. 3rd Int. Conf. Artif. Intell. Comput. Vis. (AICV 2021), Springer, 2021, pp. 14–27. DOI: 10.1007/978-3-030-87156-7_2.

[25] S. Kumar, A. V. Shvetsov, and S. H. Alsamhi, "FuzzyGuard: A Novel Multimodal Neuro-Fuzzy Framework for COPD Early Diagnosis," IEEE Internet of Things Journal, 2024. DOI: 10.1109/JIOT.2024.3467176.

[26] S. A. Adeshina and A. P. Adedigba, "Bag of Tricks for Improving Deep Learning Performance on Multimodal Image Classification," Bioengineering, vol. 9, no. 7, p. 312, 2022. DOI: 10.3390/bioengineering9070312.

[27] A. L. Thandu and P. Gera, "Privacy-centric multi-class detection of COVID-19 through breathing sounds and chest X-ray images: Blockchain and optimized neural networks," IEEE Access, vol. 12, pp. 89968-89985, 2024. DOI: 10.1109/ACCESS.2024.3418202.

[28] T. Liu, Z. Zhang, Q. Zhou, et al., "MI-DenseCFNet: Deep learning–based multimodal diagnosis models for Aureus and Aspergillus pneumonia," European Radiology*, vol. 34, pp. 5066–5076, 2024. DOI: 10.1007/s00330-023-10578-3.

[29] A. M. Q. Farhan, S. Yang, A. Q. S. Al-Malahi, and M. A. Al-antari, "MCLSG: Multi-modal classification of lung disease and severity grading framework using consolidated feature engineering mechanisms," Biomedical Signal Processing and Control, vol. 85, p. 104916, 2023. DOI: 10.1016/j.bspc.2023.104916.

[30] J. Lay and B. Pardamean, "Detection of pulmonary tuberculosis on chest X-ray images using multimodal ensemble," ResearchGate, 2022. DOI: 10.13140/RG.2.2.11678.61763.

[31] Y. Wu, J. Ma, X. Huang, S. Ling, and S. Su, "DeepMMSA: A Novel Multimodal Deep Learning Method for Non-small Cell Lung Cancer Survival Analysis," in Proc. IEEE SMC Conf., 2021, pp. 1468–1472. DOI: 10.1109/SMC52423.2021.9658891.

Received: March 10, 2025. Reviewed: March 18, 2025. Accepted: March 20, 2025. *© 2025 Zhaniya Medeuova. All rights reserved.*

[32] S. Kumar, R. Nagar, S. Bhatnagar, R. Vaddi, S. K. Gupta, M. Rashid, A. K. Bashir, and T. Alkhalifah, "Chest X-ray and cough sample based deep learning framework for accurate diagnosis of COVID-19," Computers & Electrical Engineering, vol. 103, p. 108391, 2022. DOI: 10.1016/j.compeleceng.2022.108391.

[33] Z. Tariq, S. K. Shah, and Y. Lee, "Multimodal lung disease classification using deep convolutional neural network," in Proc. 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 2530–2537. DOI: 10.1109/BIBM49941.2020.9313208.

[34] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, p. 6, 2020. DOI: 10.1186/s12864-019-6413-7.

[35] S. Deng, X. Zhang, and S. Jiang, "A diagnostic report supervised deep learning model training strategy for diagnosis of COVID-19," Pattern Recognition, vol. 149, p. 110232, 2024. DOI: 10.1016/j.patcog.2023.110232.

[36] V. Mayya, K. Karthik, S. S. Kamath, K. Karadka, and J. Jeganathan, "COVIDDX: AI-based clinical decision support system for learning COVID-19 disease representations from multimodal patient data," in Proc. International Conference on Health Informatics, 2021.

[37] S. Kumar, V. Bhagat, P. Sahu, M. K. Chaube, A. K. Behera, M. Guizani, R. Gravina, M. Di Dio, G. Fortino, E. Curry, and S. H. Alsamhi, "A novel multimodal framework for early diagnosis and classification of COPD based on CT scan images and multivariate pulmonary respiratory diseases," *Computer Methods and Programs in Biomedicine*, vol. 243, p. 107911, 2024. DOI: 10.1016/j.cmpb.2023.107911.

[38] M. J. Horry, S. Chakraborty, M. Paul, A. Ulhaq, B. Pradhan, M. Saha, and N. Shukla, "COVID-19 detection through transfer learning using multimodal imaging data," *IEEE Access*, vol. 8, pp. 149808–149824, 2020. DOI: 10.1109/ACCESS.2020.3016780.

Received: March 10, 2025. Reviewed: March 18, 2025. Accepted: March 20, 2025. *© 2025 Zhaniya Medeuova. All rights reserved.*

Article

Development of method to analyze factors of kidney disease by the use of fuzzy logic

Assel Yembergenova*[1], Azamat Serek[2], and Bauyrzhan Berlikozha[3]

[1]Department of Computer Science, SDU University, Kaskelen, Kazakhstan
[2]School of Information Technologies and Engineering, Kazakh-British Technical University (KBTU), Almaty, Kazakhstan
[3]School of digital engineering technology Narxoz University, Almaty, Kazakhstan

DOI: 10.47344/9pcv0t26

Abstract

The study introduces a new strategy for the analysis of kidney disease parameters based on fuzzy logic. Fuzzy logic is a more accurate way to categorize clinical parameters than statistical analysis because there is uncertainty and variability in medical data. The data is comprised of an extensive amount of clinical parameters including age, blood pressure, specific gravity, albumin, sugar, random blood glucose, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, and red blood cell count.

The methodology utilizes fuzzy logic centroid computation to categorize these parameters into low, medium, and high levels to provide a more dynamic and interpretable assessment of renal health. Fuzzy memberships give the current work the capability to discover intricate interrelationships between clinical variables, which may have been otherwise unattainable by conventional mean, median, and standard deviation-based analyses.

The findings confirm that fuzzy logic and conventional statistical methods enhance the comprehension of kidney disease by incorporating intricate interactions between clinical variables. The method is employed to achieve more accurate prediction and diagnostic models, offering insight to be used in kidney disease assessment and medical decisions.

Keywords: fuzzy logic, kidney disease, clinical parameters, diagnostic modeling, centroid analysis.

*Corresponding author: assel.yembergenova@sdu.edu.kz ORCID: 0009-0005-9051-2070
Email (2): a.serek@kbtu.kz ORCID:0009-0003-8644-7427
Email (3): bauyrzhan.berlikozha@gmail.com ORCID:0009-0003-9096-8719

Received: February 17, 2025. Reviewed: March 19, 2025. Accepted: March 19, 2025. *© 2025 Assel Yembergenova, Azamat Serek and Bauyrzhan Berlikozha. All rights reserved.*

I. INTRODUCTION

Kidney disease is a global health problem, and it is found in millions of patients all over the globe, causing morbidity and death in populations all over the world [1]–[3]. The blood pressure, age, and biochemical markers' correlation is most important to establish the kidney's health status. Statistical measures have their applications, but they cannot reflect the natural uncertainty and subtlety in medical data.

The current work introduces a new methodology based on fuzzy logic analysis to separate the complex interaction between disease-related parameters. Fuzzy logic is particularly appropriate to work with imprecise and incomplete data and therefore is a perfect tool to unveil hidden trends, which may not be visible to traditional statistical analysis. With the increasing prevalence of chronic kidney disease, an early and accurate diagnosis is critical to successful management. In response, exploring new analysis methods is relevant to increase diagnostic accuracy and direct clinical decision.

The final goal of the study is to obtain fuzzy membership functions and centroids to quantify and describe trends in the variables of kidney disease. With the application of fuzzy logic to traditional descriptive statistics, we will have a more meaningful understanding of the shape and distribution of the variables, and a better understanding of kidney health.

Fuzzy logic is applied across various fields because of its ability to deal with uncertainty and imprecise values. In control systems, it is applied on a large scale to increase accuracy and flexibility and provide more powerful solutions compared to traditional binary logic [4]–[7]. Fuzzy logic is applied in the automotive industry to improve vehicle stability, fuel efficiency, and performance. Fuzzy logic is applied in medicine to simulate diagnostic neurological and cardiovascular diseases [8]–[11]. Fuzzy logic is applied to assist in financial modeling by dealing with uncertainty in the market, improving risk analysis, and improving investment decisions [12]–[14].

The paper examines a 2019 Chronic Kidney Disease data set [15] with key kidney-related variables. The study attempts to unravel complex relations and trends by performing fuzzy logic analysis, and it opens a door to explore kidney disease development and their major clinical implications.

The subsequent sections encapsulate the research strategy employed, key findings, and implications thereof for the development of diagnostic tests and treatments exactly aimed at attacking kidney disease.

II. LITERATURE REVIEW

One of these fields that have been subjected to intense research by multiple computational approaches is kidney disease detection. Authors in the article [16] have analyzed the serious issue of kidney failure, with an emphasis on slowing CKD progression and minimizing economic burdens to treatments. They have based their work on risk factors causing kidney degeneration, with an emphasis on conducting regular checks to determine risks before severe health decline.

Unlike other research that employed conventional statistical techniques, researchers [17] employed fuzzy and adaptive neural fuzzy inference systems to diagnose CKD more precisely. They attempted to increase the reliability of medical tests to diagnose disease. They considered important parameters such as nephron function, glucose, blood pressure, age, body mass index (BMI), and smoking status in minute detail while creating a fuzzy inference system. Their system distinguished between CKD stages 1 to 5 and provided real-time results on the extent of the disease. Their fuzzy system was simulated with the assistance of MATLAB to demonstrate how it can be employed in real-world healthcare applications.

One contribution to the study was by authors in [18], where they found natural imprecision in CKD detection in clinics. They used a fuzzy inference system in MATLAB to correct the issue, based on the capability of fuzzy logic to handle uncertainty. Their work, however, did not have a robust validation system, and there is a chance to enhance prediction accuracy.

In contrast to other studies, the current work is centered on the development of a more accurate fuzzy expert system (FES) from a large dataset, clinical guidelines, and expert opinion to enhance CKD diagnosis.

This contribution is original in its application of systematic tests for normality to establish the input parameter effect and system performance assessment by exhaustive testing. Surface analysis identified nephron function, blood sugar, and body mass index (BMI) as the most accurate parameters to predict CKD. The FES was validated with 80 test cases, and the accuracy level was 93.75%, confirming its feasibility in real-world applications.

Follow-up studies also reiterated the significance of CKD early detection. The authors in [19] and [20] employed deep learning models to predict CKD, with satisfactory performance but requiring large training data and computational resources. Our study bridges the gap by presenting an interpretable and computationally light alternative based on fuzzy logic.

The current work introduces a new prediction model based on the fuzzy logic toolbox in MATLAB to screen CKD. There are five key steps in the process: (1) selection of key input parameters such as blood urea nitrogen, eGFR, and serum creatinine, (2)

Received: February 17, 2025. Reviewed: March 19, 2025. Accepted: March 19, 2025. *© 2025 Assel Yembergenova, Azamat Serek and Bauyrzhan Berlikozha. All rights reserved.*

input-output relation normalization by carrying out min-max normalization, (3) construction of a fuzzy inference engine, (4) fuzzy rule aggregation, and (5) defuzzification to obtain clear diagnostic outputs. 70 test cases from patients have been analyzed, and 47 have been diagnosed as CKD positive, which validated the reliability of the proposed model.

With the inclusion of results from past studies and enhancement of fuzzy logic-based techniques, the work is meant to enhance diagnostic accuracy and achieve early detection of CKD. The study results demonstrate the potential of fuzzy expert systems to aid clinical decision and enhance healthcare treatments.

III. Methods

The study makes use of a range of data collected from (Chronic kidney disease data set, 2019), with differing parameters such as age, blood pressure, specific gravity, albumin, sugar, random blood 5 glucose, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white cell count, and red cell count, and so on.

The dataset, formulated by merging two consecutive hospital reports, is an abstraction of one of the most important prediction modeling indicators of chronic kidney disease (CKD). Since its multivariate set is heterogeneous, it is appropriate to use in classification applications in health care and is important to use in order to gain insight into causality of chronic kidney disease. The dataset comprises 25 attributes and 400 instances, all being of utmost relevance that will be used to contribute to prediction modeling.

The set of data is composed of physiologic tests, clinical features, and lab values that describe a diverse set of conditions of the patients. Key features include blood pressure (bp), specific gravity (sg), albumin (al), sugar (su), red cells (rbc), pus cells (pc), pus cell clumps (pcc), bacteria (ba), random blood glucose (bgr), blood urea (bu), serum creatinine (sc), sodium (sod), potassium (pot), hemoglobin (hemo), packed cell volume (pcv), white cell count (wc), red cell count (rc), and target feature (class) that describe absence or presence of chronic kidney disease (CKD).

The wide range of features enables effective classification modeling and contributes to establishing causes of CKD. The compound nature of the dataset makes it suitable to use within processes of machine learning of CKD or its initiation forecasting, enabling possible better-timed effective treatments.

The following are processes that fall under fuzzy logic analysis:

- *Variable selection.* The most important identified variables were age, blood pressure, and biochemical markers. Fuzzy logic is developed on top of these variables.
- *Membership function generation.* Membership functions corresponding to low, medium, and high levels of membership were created using skfuzzy library for every identified variable. The functions give a pictorial representation of every range of variables of the dataset.
- *Centroid Calculation.* For determination of numerical values of each level of each variable range, each of centroid, or centre of gravity, of the membership function were computed
- *Descriptive Statistics.* For better understanding of the dataset, mean, median, and standard deviation were also calculated on each of the variables.

Fig. 1 summarizes briefly the key steps followed to use fuzzy logic on parameters of renal disease, from preprocessed data through fuzzification, inference system designing, defuzzification, to classification verification.

All processes are organized to adhere to a process of fuzzy logic that is implemented using clinical data guidelines.

The aim of merging fuzzy logic analysis into classical descriptive statistics is to have a complete understanding of complex relationships among factors of renal disease. The outcomes of this process will provide valuable insight to practice, informing practice and research of kidney health in the future.

IV. Results and Discussion

Table 1 presents the fuzzy analysis of clinical attributes pertaining to kidney disease in low, medium, and high levels. This enhances understanding by applying mean, median, and standard deviation to determine patterns of distribution, range, and variability in the dataset. Fuzzy logic centroids provide more accurate classification of clinical attributes.

The fuzzy logic analysis of the centroids (low, medium, and high) provides a mean of around 51.48 years and a median of 55.00 years, showing a reasonably symmetric distribution. The standard deviation of 17.15 indicates a moderate spread in the distribution of age, with a broad range of values.

Blood pressure classification shows that the mean (76.47) and median (80.00) are almost equal, which implies a symmetric distribution. The standard deviation of 13.67 reflects moderate variation in blood pressure level between patients.

Received: February 17, 2025. Reviewed: March 19, 2025. Accepted: March 19, 2025. *© 2025 Assel Yembergenova, Azamat Serek and Bauyrzhan Berlikozha. All rights reserved.*

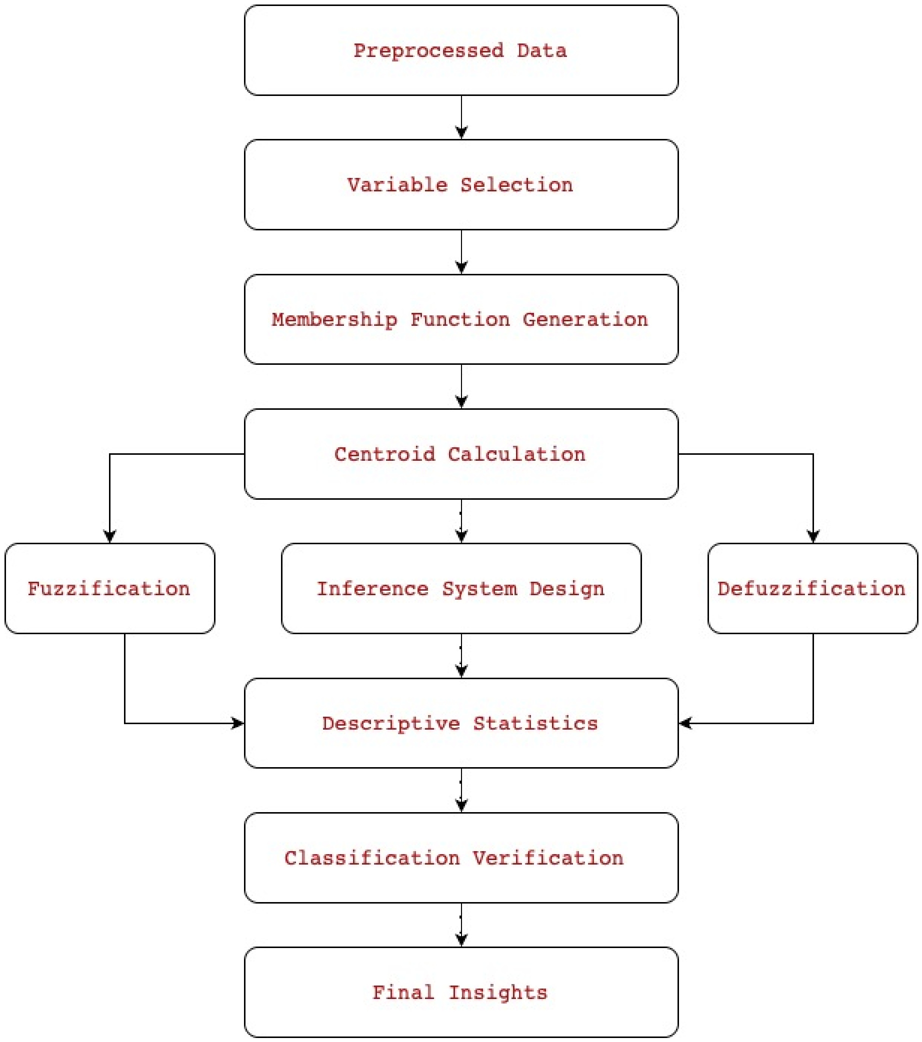


Fig. 1: Flowchart summarizing the fuzzy logic-based methodology for kidney disease analysis

Specific gravity shows a clear distinction between centroids, with the low level being 1.01 and medium and high levels being 0.00. The close proximity between the mean and median and a standard deviation of 0.01 indicate homogeneity, which suggests that values of specific gravity are comparatively consistent across levels of kidney disease.

The albumin values reflect a striking difference between the median (0) and the mean (1.02), indicating a skewed distribution. The standard deviation of 1.35 reflects moderate variation in albumin values, an indication of variable degrees of kidney dysfunction in the patients.

Red blood cell count reveals a closely similar mean (4.71) and median (4.8), indicating a near-symmetrical distribution. The standard deviation of 1.02 indicates moderate variation, and it could reflect progression in anemia in patients with renal disease.

Other parameters, including blood glucose, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, and white blood cell count, have varying degrees of skewness and spread. An example is blood glucose random levels, which have a high standard deviation (79.17), indicating the great range in blood sugar levels between patients, an important aspect in the

Received: February 17, 2025. Reviewed: March 19, 2025. Accepted: March 19, 2025. © 2025 Assel Yembergenova, Azamat Serek and Bauyrzhan Berlikozha. All rights reserved.

Algorithm 1 Fuzzy Logic-based Analysis of Kidney Disease Factors

Input: Dataset D with features $F = \{f_1, f_2, \ldots, f_n\}$, where $f_1, f_2, \ldots, f_n$ are clinical parameters related to kidney disease
Output: Fuzzy logic centroids for classification of kidney disease factors
Step 1: Data Preprocessing
Clean dataset D by handling missing values and outliers
Normalize or scale features in F if required for uniformity
Step 2: Fuzzification
for each feature $f_i \in F$ **do**
  Define membership functions $\mu_{\text{low}}(f_i), \mu_{\text{medium}}(f_i), \mu_{\text{high}}(f_i)$
  Apply fuzzification to convert crisp values of f_i into fuzzy sets
end for
Step 3: Fuzzy Inference System (FIS) Design
Define fuzzy rules based on medical knowledge and clinical thresholds
for each feature $f_i \in F$ **do**
  Use predefined rules to establish relationships between input features and kidney disease diagnosis
  Apply fuzzy operators (e.g., AND, OR) to combine rules
end for
Step 4: Defuzzification
for each fuzzy output **do**
  Calculate the centroid of the fuzzy output using the centroid method:

$$\text{Centroid} = \frac{\sum x \cdot \mu(x)}{\sum \mu(x)}$$

  Convert fuzzy outputs into crisp values for interpretation
end for
Step 5: Classification and Evaluation
Use defuzzified values to classify the clinical parameters into risk categories (low, medium, high)
Evaluate the classification accuracy using standard performance metrics (e.g., accuracy, precision, recall)
End

Fig. 2: Pseudocode of methodology

development of the disease. Similarly, serum creatinine is very variable (standard deviation 5.73), as is to be expected given that it is a key indicator of renal function.

Clinical Implications. The statistical trends identified in the present study have important clinical implications. By classifying clinical features into low, medium, and high levels based on fuzzy logic, medical professionals are able to:

- *Improve Early Diagnosis.* lassification of parameters such as blood pressure, albumin, and serum creatinine into levels enables early-stage kidney disease to be diagnosed so that action can be taken in a timely fashion.
- *Improve Patient Monitoring.* Regular monitoring of blood urea, serum creatinine, and hemoglobin levels will help clinicians monitor disease progression and adjust treatment regimens.
- *Facilitate Personalized Treatment.* With an understanding of the distribution and variation in important parameters, physicians can individualize medication and diet recommendations to fit patient profiles.

Received: February 17, 2025. Reviewed: March 19, 2025. Accepted: March 19, 2025. © 2025 Assel Yembergenova, Azamat Serek and Bauyrzhan Berlikozha. All rights reserved.

TABLE I: Summary of Experimental Results

	Centroid Low	Centroid Medium	Centroid High	Mean	Median	Std Dev
Age	18.5	47.82	76.49	51.48	55.00	17.15
Blood Pressure (bp)	158.83	102.15	144.82	76.47	80.00	13.67
Specific Gravity (sg)	1.01	0	0	1.02	1.02	0.01
Albumin (al)	10.35	1.88	3	1.02	0	1.35
Sugar (su)	0.33	1.88	2.77	0.45	0	1.1
Blood Glucose Random (bgr)	64.01	220.01	375.35	148.04	121.00	79.17
Blood Urea (bu)	120.14	149.98	279.48	57.43	42.00	50.44
Serum Creatinine (sc)	1.34	26.51	51.29	3.07	1.3	5.73
Sodium (sod)	48.84	101.67	154.18	137.53	138.00	10.39
Potassium (pot)	3.25	18.05	32.54	4.63	4.4	3.19
Hemoglobin (hemo)	6.26	11.1	15.54	12.53	12.65	2.91
Packed Cell Volume (pcv)	118.96	33.93	48.29	38.88	40.00	8.98
White Blood Cell Count (wc)	4268.71	12335.37	20401.37	8406.12	8000.00	2939.46
Red Blood Cell Count (rc)	3.03	4.83	6.24	4.71	4.8	1.02

A. Discussion

The fuzzy logic analysis provides a complete examination of the membership function and distribution characteristics of the variables of kidney disease. Some of the observations and their potential implications are as follows:

Centroids for low, medium, and high membership levels in age indicate an even distribution across the range of the ages. This is an indication that the dataset represents a range of ages, as is fitting to explore kidney health throughout all stages in one's life.

The diversity enhances the capability of the dataset to give meaningful insights into risk factors based on kidney disease by age. The distribution of blood pressure indicates a prevalent cluster in medium membership levels. This result indicates that the subjects in the dataset have mostly moderate blood pressure, which may have an effect on kidney health. Additional studies are needed to explore moderate blood pressure and outcomes in kidney disease. Specific gravity (sg) Anomalies: Specific gravity indicates an abnormality, with centroids measured as 1.01 and 0.00, respectively, for low and medium and high membership levels. In spite of these abnormalities, the median and the mean values are within expected ranges. This abnormality needs to be explored to establish if there is an issue with the quality of data or to explore special features in the dataset.

Centroids for sugar and albumin indicate clear progression from low to high as would be clinically expected. Higher glucose and albumin concentrations are typical to indicate renal impairment, and the distinctly clear fuzzy logic centroids indicate strong ability to classify. The results provide a strong basis to determine risk levels for kidney disease. The blood glucose random values have large variance, and their fuzzy logic centroids are dispersed across low, medium, and high membership values. This indicates strong variance of blood glucose levels, underlining how important it is to account for renal function. The strong variance of distribution underlines the necessity to carefully determine how renal function is influenced by blood glucose levels. The white cell count centroids indicate a strong shift to higher counts, indicating varied immunological responses in sets. More intense observation of correlation between immunological and renal functions could indicate important correlation with practice implications on managing and predicting kidney disease. 9 The centroids of the red cell counts indicate symmetric count distribution across counts of low, medium, and high values. The resulting homogeneity indicates balanced coverage of the counts of the red cells in sets, hence eliminating sources of bias and providing reliability to corresponding analyses of other parameters of interest of a clinical nature.

The findings justify the use of fuzzy logic to establish complex relations in kidney disease data. The process ensures complete classification of parameters of interest of a clinical type, hence better comprehension of complex determinants of renal function. Future studies should continue to evolve based on findings to establish variable interactions and related practice implications. Finally, correction of aberrations, as that of specific gravity, requires complete determination of related data to give reliability to subsequent assessments.

V. Conclusion and Future Work

Fuzzy modeling of clinical parameters adds a new dimension to kidney disease analysis, with more refined grouping of variables as low, medium, and high. This grouping enables us to understand better how parameters influence kidney function. The addition of

Received: February 17, 2025. Reviewed: March 19, 2025. Accepted: March 19, 2025. *© 2025 Assel Yembergenova, Azamat Serek and Bauyrzhan Berlikozha. All rights reserved.*

fuzzy logic centroids and traditional measures, such as mean, median, and standard deviation, provides a complete perspective on variance and central tendency. The key findings are that parameters such as random blood glucose, blood urea, and serum creatinine have high variance, and they can be potential markers in the progression of kidney disease.

Despite these advances, there are a few restrictions to this work. The dataset used may not capture all the variation present in patient populations with diversity, and the fuzzy logic-based approach cannot account for potential non-linear interactions between parameters. Additionally, the lack of direct comparison with conventional machine learning methods prevents assessment regarding the relative performance of fuzzy logic in clinical decision support.

Future studies should integrate this fuzzy logic-based categorization into clinical decision support systems to assist medical personnel in risk assessment and early diagnosis. Comparison to machine learning techniques, including decision trees, support vector machines, and deep learning techniques, would provide more proof of efficacy. Larger dataset size and real-time patient monitor data could also enhance the predictability of the model, adding to its utility in the clinical environment.

References

[1] Cockwell, P., Fisher, L.-A. (2020). The global burden of chronic kidney disease. The Lancet, 395(10225), 662–664.

[2] Jha, V., Garcia-Garcia, G., Iseki, K., Li, Z., Naicker, S., Plattner, B., Saran, R., Wang, A. Y.-M., Yang, C.-W. (2013). Chronic kidney disease: Global dimension and perspectives. The Lancet, 382(9888), 260–272.

[3] Mills, K. T., Xu, Y., Zhang, W., Bundy, J. D., Chen, C.-S., Kelly, T. N., Chen, J., He, J. (2015). A systematic analysis of worldwide population-based data on the global burden of chronic kidney disease in 2010. Kidney International, 88(5), 950–957.

[4] Cherry, A., Jones, R. (1995). Fuzzy logic control of an automotive suspension system. IEE Proceedings-Control Theory and Applications, 142(2), 149–160.

[5] Ivanov, V. (2015). A review of fuzzy methods in automotive engineering applications. European Transport Research Review, 7(3), 1–10.

[6] Von Altrock, C., Krause, B., Zimmermann, H.-J. (1992). Advanced fuzzy logic control technologies in automotive applications. In [1992 Proceedings] IEEE International Conference on Fuzzy Systems (pp. 835–842). IEEE.

[7] Vachtsevanos, G., Farinwata, S. S., Pirovolou, D. K. (1993). Fuzzy logic control of an automotive engine. IEEE Control Systems Magazine, 13(3), 62–68.

[8] Baig, F., Khan, M. S., Noor, Y., Imran, M., Baig, F., others. (2011). Design model of fuzzy logic medical diagnosis control system. International Journal on Computer Science and Engineering, 3(5), 2093–2108.

[9] Awotunde, J. B., Matiluko, O. E., Fatai, O. W. (2014). Medical diagnosis system using fuzzy logic. African Journal of Computing ICT, 7(2), 99–106.

[10] Das, S., Guha, D., Dutta, B. (2016). Medical diagnosis with the aid of using fuzzy logic and intuitionistic fuzzy logic. Applied Intelligence, 45, 850–867.

[11] Phuong, N. H., Kreinovich, V. (2001). Fuzzy logic and its applications in medicine. International Journal of Medical Informatics, 62(2–3), 165–173.

[12] Gil-Lafuente, A. M., others. (2005). Fuzzy logic in financial analysis. Springer.

[13] Korol, T. (2012). Fuzzy logic in financial management. INTECH Open Access Publisher.

[14] Mohammadian, M., Kingham, M. (1997). Hierarchical and feed-forward fuzzy logic for financial modelling and prediction. In Australian Joint Conference on Artificial Intelligence (pp. 147–156). Springer.

[15] Chronic kidney disease data set. (2019). UCI Machine Learning Repository.https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease (Accessed: 1 September 2023).

[16] Ahmed, T. I., Bhola, J., Shabaz, M., Singla, J., Rakhra, M., More, S., Samori, I. A., others. (2022). Fuzzy logic-based systems for the diagnosis of chronic kidney disease. BioMed Research International, 2022.

[17] Bai, Y., Wang, D. (2006). Fundamentals of fuzzy logic control—Fuzzy sets, fuzzy rules and defuzzifications. In Advanced fuzzy logic technologies in industrial applications (pp. 17–36).

[18] Celikyilmaz, A., Turksen, I. B. (2009). Modeling uncertainty with fuzzy logic. Studies in Fuzziness and Soft Computing, 240(1), 149–215.

[19] John, R. I., Innocent, P. R. (2005). Modeling uncertainty in clinical diagnosis using fuzzy logic. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 35(6), 1340–1350.

[20] Larsen, P. M. (1980). Industrial applications of fuzzy logic control. International Journal of Man-Machine Studies, 12(1), 3–10.

Received: February 17, 2025. Reviewed: March 19, 2025. Accepted: March 19, 2025. *© 2025 Assel Yembergenova, Azamat Serek and Bauyrzhan Berlikozha. All rights reserved.*

Review

DETECTING SOCIAL CONFLICTS IN KINDERGARTENS USING DEEP LEARNING AND COMPUTER VISION

Dina Kengesbay* [1]

[1]Department of Computer Science, SDU University, Almaty, Kazakhstan

DOI: 10.47344/7x77b619

Abstract

Early conflict detection in kindergartens plays a significant role in ensuring a harmonious learning atmosphere and in promoting the social growth of young children. While most previous works have only addressed conflict detection through adults, in this paper, we specifically address conflict detection in kindergartens using deep learning, utilizing both spatial and temporal information to improve performance. The application of deep learning and computer vision in automatically detecting and analyzing early conflicts among young children is discussed in this paper. Using video footage, we leverage state-of-the-art RNNs and 3D CNNs for high-accuracy detection of conflict instances. Crucial visual cues—facial expressions, gestures, poses, vocal tone, and movement—are examined for the extraction of tension or aggression signs. The model is evaluated on real kindergarten video data, with promising conflict detection and classification results. The findings indicate the potential of AI-supported tools in assisting teachers in class management, child behavior monitoring, early intervention mechanisms, and the fostering of a good social environment.

Keywords: social conflict detection, deep learning, computer vision, kindergarten, child behavior analysis, pose estimation, sentiment analysis, classroom monitoring, early childhood education, AI in education.

I. Introduction

Social links play an important role in the early childhood development, as they play a major role in the development of emotional intelligence, communication skills, and conflict resolution [1]. Social conflict naturally occurs in kindergarten classrooms as children learn to interact with their peers, work on social and behavioral norms, and participate in problem solving [2]. These child-versus-group conflicts are a normal part of early socialization, but they necessitate careful management so they can contribute positively to a child's social and emotional development. Historically, teachers and childcare workers have used direct observation and subjective sorting to identify and mediate conflicts. Nonetheless, classroom environments are highly dynamic and teachers often face time constraints, making the early detection and timely intervention difficult [3]. This research presents a pioneering contribution to

*Corresponding author: dina.kengesbay@sdu.edu.kz
Email: dina.kengesbay@sdu.edu.kz ORCID: 0009-0006-8121-9697

Received: March 4, 2025. Reviewed: April 12, 2025. Accepted: April 12, 2025. *© 2025 Dina Kengesbay. All rights reserved.*

automatically detecting social conflict in educational settings through computer vision and deep learning methods presented in October 2023. These technologies enable the real-time monitoring of children's interactions, facilitating an immediate and objective analysis of conflicts [4]. By utilizing these AI-driven systems, educators gain insights into behavioral patterns, which allows them to create more effective intervention plans and improve classroom management. AI in early education not only improves identification of conflict but also adds to the structured and data-oriented approach to child behavior. This research focuses on the analysis and verification of deep learning models that can lead towards automatic detection and models of conflicts in preschool sessions [5] and improve learning conditions for young children.

This study aims to devise and test an artificial intelligence (AI) system to identify conflicts of kindergarten children using computer vision and deep learning techniques. It will be developed on the basis of recorded classroom interactions to discern conflict indicators through gestures, facial expressions and vocal tone and body movements [6]. This article focuses on the investigation of behavioral indicators across various time periods and learning settings, whilst also considering the effect of teacher interventions on conflict resolution. By providing a comprehensive understanding of conflict dynamics, the findings from this study will guide educational professionals in devising strategies for early dispute prevention, implementing judicious interventions, and enhancing the overall educational experience. Moreover, this study takes into consideration the wider impact of AI on early childhood education by suggesting technological advancements in monitoring social interactions, analyzing patterns of conduct, and refining the general atmosphere of the classroom [7].

The research is organized around three key objectives to reach these goals. The first one deals with how we are going to collect a large dataset of video data which contains the social conflicts of kindergarten boys and girls, where this dataset will be used to train the CV system to detect the conflicts with deep learning technology in real time [8]. Second, it studies whether existing models for fight detection are applicable to kindergarten receivers in such cases to see how effective and adaptable they would be in settings where fights are more subtle and often non-violent [9]. At last, the goal of this research is to determine patterns of repetitive behavior and causes of interactional conflicts in small children, therefore gaining insights into the socialisation process of young children, and also laying the basis for the development of computational strategies to enhance early childhood education through safe behavioural interventions using AI systems. In conclusion, this research aims to contribute to these important aspects in order to help narrow down the gap between AI advances and the real-world classroom implementation of technology based conflict resolution considering early childhood learning experiences.

II. Literature Review

The main issue when trying to detect social conflict in kindergartens by deep learning and computer vision is the room under the assumption that children typically do not exhibit overt violence and that behavior would be very subtle. While adults' conflict is often realized in some evident physical violence, conflict between young children as manifested with social excluding or strife. This necessitates making AI models for Early Childhood Contexts Although much of the earlier work has centered on adult violence or violence more broadly, violence detection at school has shown the success of AI systems in observing aggressive conduct amongst students [10]. Traditional mechanisms for conflict detection depend on the identification of hostile stances, loud voices, or fighting in fact [11]. Such work may not translate easily to kindergarten, where violence is subtler and requires analysis at a finer behavioral scale. This is true, especially since it departs from adult behavior analysis, but traditional methods based on direct detection of aggressiveness are less suitable for identifying concealed violence in children.

However, recent advances in the fields of deep learning and computer vision have led to an automated conflict detection in various domains, including security monitoring, child well-being, and education. Skeleton-based techniques have recently attained high accuracy in determining aggressive action for pose estimation and motion analysis [12]. Likewise, sentiment analysis and multimodal behavior recognition have also been utilized to recognize distress, frustration or aggressiveness in classrooms [13]. Deep Learning has been applied in the detection of physical violence and child abuse in real-time using AI video surveillance systems as well [14]. Social Conflict and Aggression Detection in Learning Environments Detects aggression and social conflict through a series of methods, including pose estimation, facial expression recognition, and speech tone detection. Pose-based skeleton tracking attained 83-92% accuracy for detecting aggressive behaviour [12]. Specific to facial expression recognition, 85-90% accuracy levels were obtained in detecting children distress and frustration [13]. For example, accuracy levels of 80-88% were reached in detecting distress from voice patterns in speech tone detection [11]. The most successful models utilized combinations of video, audio, and behavior cues, achieving over 94% accuracy in a controlled experiment [14]. Even with all of these improvements, the use of AI conflict detection in kindergarten classrooms presents a serious challenge. The major challenges are the variation in child behavior, no big scale of labeled data and privacy/consent from parents etc. [15] Joint integration of spatial features (e.g., gestures, movement)

Received: March 4, 2025. Reviewed: April 12, 2025. Accepted: April 12, 2025. © 2025 Dina Kengesbay. All rights reserved.

and temporal features (e.g., speech tone, facial micro-expression patterns; [16]), has also shown great promises in terms of improving accuracy and reliability in early education context.

AI solutions are capable of early intervention strategies, helping teachers cope with classroom behaviour, and creating a more unified learning atmosphere by strengthening already existing methodologies, building them, and accounting for the unique nature of kindergarten social dynamics.

A. Deep Learning for Conflict-detection

A considerable amount of this domain is using deep learning for analysis via video by in classifying violent or aggressive behavior. The FightNet model which we introduced in Thao et al. In (2023), CNNs and RNNs are employed for spatio-temporal, incidence in schools related to violence and fights. The method's mean average precision (mAP) of 45.34% (IoU 0.5) was found to be excellent on keypoint estimation and F1-score of 71.69% was also acceptable [10]. FightNet, however, had been primarily trained on datasets of older students and adult subjects and therefore would have limited utility in discerning behavior of younger children. Kindergarten conflicts [10] are based instead on gestures, bodily movement, or patterns of social exclusion rather than direct bodily force and would therefore require early child behavior models specifically trained on those inputs. In similar work, Imah and Karisma (2022) employed a deep transfer learning model, which used VGG16-LSTM for feature extraction and modeling of time series with a G-mean of 0.911, indicating it a promising model for accurate sexual violence identification among children [11]. This method had previously only mostly been built on subject datasets centered on more adults, as such limiting use for the younger child, who can have milder bodily motion and more delicate social interplay in disputes.

B. Detection of Violence in Surveillance

The use of deep learning methods for detection of violence in surveillance systems has received considerable attention. For example, Hughes and Kersten (2022) integrated Long Short-Term Memory (LSTM) networks with Convolutional Neural Networks (CNNs) to improve detection accuracy to 77.9% on datasets like Hockey Fight and Movies Fight Detection Dataset [12]. Nevertheless, their model was trained on mostly adult-based violent actions like punching and kicking, which are perhaps not representative of kindergartens. Moreover, their model had a high false-positive rate with respect to its detection task, limiting its applications in real-time monitoring. With dynamic classroom environments, excessive false alarms might trigger unnecessary interventions that take away educators' attention from real conflicts and potentially call into question the validity of AI-based surveillance systems.

C. Detecting Child Abuse and Distress

Specialized methods have also been investigated to detect signs of distress from children's voices. Yan et al. (2023) exemplified the use of deep CNNs in classifying child speech signals of distress with accuracy rates well over 90% based on MFCC and spectrogram features [13]. This suggests the efficacy of auditory-based methods in sign detection. Nevertheless, these methods might not capture the entire picture of a child's well-being. Blending multimodal data with both auditory and visual cues can be beneficial in building resilience and accuracy into the early conflict detection systems of children. This multimodal strategy conforms to studies such as those conducted by Wu et al. (2015), who highlighted the significance of both spatial and temporal information in video classification [14].

D. Multi-Modal Data Fusion for Improved Detection

Incorporating spatial and temporal information effectively is essential to support precise violence detection. Experiments have demonstrated that one can combine CNNs with RNNs, e.g., LSTMs, to extract spatial information and capture temporal patterns in video data. For example, in their work, Wu et al. (2015) developed a hybrid deep learning scheme that encodes static spatial information, short-term motion information, and long-term temporal cues and obtains state-of-the-art results on benchmarks such as UCF-101 and Columbia Consumer Videos (CCV) [14]. Implementing such approaches in preschool settings requires precise attention to body language (spatial cues) and interaction sequencing (temporal cues) to detect social conflict in young children with accuracy. Blending multimodal data that include both visual and auditory signals has the potential to improve detection and analysis of faint conflict cues in early childhood settings.

Received: March 4, 2025. Reviewed: April 12, 2025. Accepted: April 12, 2025. © 2025 Dina Kengesbay. All rights reserved.

E. Effectiveness of AI in Educational Settings

These deep learning models, although powerful, present practical and ethical challenges when applied in real-world learning environments. These challenges include maintaining data privacy, securing informed consent, and avoiding the encroachment of automation on teacher-child relationships. Hughes and Kersten emphasize the concern of bias and over-reliance on automated systems, detracting from human intuition [12]. And in particular, what schools need to think about when it comes to implementing AI systems, certainly for vulnerable populations like young children.

While deep learning and other systems could be implemented in an educational environment, it is imperative that these decisions be made with consideration of sending teachers directly to dispute management, as opposed to a turnturned recommendation engine. Papadopoulos and Stavrakoudi compared human decision making and automation in several public security applications (e.g. violence detection), stressing the need to keep this balance even in kindergarten context [15]. Research could be further developed to integrate RNN, CNN, advanced temporal fusion techniques (such as slow fusion and multi-stream), to achieve behaviour classification in complex preschool scenes. Moreover, the integration of pose estimation and sentiment analysis, as shown in crime detection models, might be beneficial for recognizing subtle social signals, improving the real-time capabilities of the AI systems deployed for early childhood education [16].

III. Methods

The present work introduces a CNN-LSTM-3D CNN deep learning approach, customized to identify low-intensity conflict in children and distinguish between playful behavior and aggression. The system uses Convolutional Neural Networks (CNN) to learn the spatial features, and temporal relations in behavior patterns are learned by employing Long Short-Term Memory (LSTM) networks. 3D CNN further processes spatiotemporal features in video streams to enhance conflict detection. The approach supports teacher surveillance, detection of early signs of aggression, and establishment of a positive learning environment in kindergarten classrooms.

A. Dataset Collection and Preprocessing

Video data were obtained in simulated kindergarten environments, capturing both conflict and non-conflict situations, including play, cooperation, and conflicts. Data collection was conducted in accordance with participant anonymity and informed consent guidelines. The resulting dataset consists of approximately 2,000 raw video clips, each lasting between 2 and 5 seconds, and is evenly distributed across positive (conflict) and negative (non-conflict) classes, ensuring a balanced dataset. To enhance generalization, data augmentation techniques such as rotation, brightness alteration, and flipping were applied during training, increasing the training dataset to over 10,000 samples. However, validation was conducted using only raw, non-augmented videos to prevent performance estimation bias. Importantly, the training and testing sets remained entirely separate, ensuring that there was no overlap between the video samples used for training and testing, thus preventing data leakage and ensuring a balanced performance evaluation. For preprocessing, the video frames were resized to 224×224, and key frames were extracted using scene detection to eliminate redundancy. Since the data set consists of sequential video data, unnecessary augmentations were avoided to preserve the natural flow of movement patterns.

Dataset Examples: Fight and No-Fight

The data set consists primarily of two categories: Conflict scenarios, such as physical fighting, verbal confrontation, hostile body language, and social exclusion or manipulation; and Non-Conflict Scenarios, such as collaborative play, neutral dialogue, and ordinary classroom phenomena. For instance, in a Conflict Scenario, students might be seen arguing vehemently over access to resources; conversely, in a Non-Conflict Scenario, students might be seen collaborating harmoniously on a group task. Using labeled frames to show the difference between 'fight' versus 'no-fight' situations, we are able to show all of the different scenarios which can be represented within the dataset to be analyzed and models trained.

B. Evaluation of Existing Conflict Detection Systems

Before training dedicated models, general video-based conflict detection systems trained on typical video databases (i.e., sports and surveillance) were evaluated on recorded kindergarten data. However, as these models are optimized for application in adult behavior, they could not detect mild and non-violent conflicts characteristic in early-childhood behavior and confirmed the need for a dedicated database and system design.

Received: March 4, 2025. Reviewed: April 12, 2025. Accepted: April 12, 2025. © 2025 Dina Kengesbay. All rights reserved.

Fig. 1. Example frame showing a no-fight situation

Fig. 2. Example frame showing a fight situation

TABLE I
PERFORMANCE OF EXISTING MODELS ON KINDERGARTEN CONFLICT DETECTION

Model & Paper	Methods	Original Accuracy	Performance on Kindergarten Data
FightNet (Le Quang Thao et al., 2023)	CNN-RNN, keypoint estimation	F1: 71.69%	High false positives (34%) in playful interactions.
Child Violence Detection (Imah & Karisma, 2022)	VGG16-LSTM, deep transfer learning	G-mean: 0.911	Moderate accuracy (68.2%), misclassified disagreements.
Efficient Violence Detection (Hughes & Kersten, 2022)	CNN-LSTM for video classification	77.9%	Poor adaptability (54.3%), struggled with emotional intensity.
Child Abuse Detection (Yan et al., 2023)	Deep CNNs, MFCCs, spectrogram analysis	90% (audio-based)	Limited applicability, needed visual context.
Fighting Detection (Papadopoulos & Stavrakoudi, 2024)	CNN-RNN-Attention ensemble	77.4%–95.7%	Decent (72.1%), confused play with conflicts.

Existing violence detection models on kindergartens show significant drawbacks in their applicability in early childhood settings. FightNet and Child Violence Detection models, with high effectiveness in the adult context, display high false positive rates and moderate accuracy in applying to children's communication and tend to label playful activities as violence. Efficient Violence Detection and Fighting Detection models also show low adaptability and confusion between play and fighting, respectively. The Child Abuse Detection model based on audio cues fails to capture the visual context required to interpret children's actions effectively. All these findings emphasize the importance of creating specialized models trained on child-specific datasets with the purpose of maximizing accuracy and credibility in conflict detection in kindergartens.

C. Training Custom Conflict Detection Models

For better identification of social conflicts, we used and compared three architectures derived from deep learning: Features extracted via a CNN were utilized as input for an LSTM network in a way to capture time-dependent relationships in child-child and child-adult interactions. 3D CNN: The regular 2D CNNs have been extended to incorporate a time dimension for dealing with a stream of frames as volumetric information. They trained each model on the compiled dataset and compared them using performance metrics like accuracy, precision, recall, and F1-score in order to determine the best way to detect conflicts in kindergartens.

Received: March 4, 2025. Reviewed: April 12, 2025. Accepted: April 12, 2025. © 2025 Dina Kengesbay. All rights reserved.

D. Model Training and Evaluation

Model Architectures To develop a robust video-based violence detection system for kindergarten settings, we implemented and evaluated two deep learning architectures: (i) a CNN + RNN (LSTM) hybrid model, and (ii) a 3D Convolutional Neural Network (3D CNN). CNN + RNN (LSTM) Architecture This hybrid model extracts spatial features from each frame using a deep CNN backbone and then models temporal dependencies using a multi-layer bidirectional LSTM.

The CNN + RNN (LSTM) hybrid model extracts spatial features from each frame using a deep CNN backbone and then models temporal dependencies with a multi-layer bidirectional LSTM. For feature extraction, the model processes sequences of T = 32 frames resized to (224 × 224 × 3) with a pretrained EfficientNet-B3 or ResNet-101 backbone. The CNN outputs (T × D) feature vectors, where D = 1024, after applying a Global Average Pooling (GAP) layer to reduce redundant spatial information, resulting in 32 feature vectors of size 1024. Temporal modeling is performed using 3 bidirectional LSTM layers with a hidden size of 512 and dropout of 0.3, where the final hidden state is the concatenation of forward and backward states. An attention mechanism is used to focus on key frames, with attention weights computed for each time step. The fully connected layers consist of 256 neurons with ReLU activation and a dropout rate of 0.4, followed by an output layer with softmax activation for binary classification. This model has approximately 29 million parameters when using EfficientNet-B3 and 49 million parameters when using ResNet-101.

This table outlines the key components of the hybrid CNN + RNN (LSTM) architecture. The CNN backbone (EfficientNet-B3/ResNet-101) has 24M/44M parameters, while the LSTM layer (3 layers, hidden size = 512) contributes 4.8M parameters. The fully connected layer has 256 neurons with 131K parameters, and the output layer (2 neurons) adds 2K parameters. Total parameters are 29M (EfficientNet-B3) / 49M (ResNet-101). The CNN + RNN (LSTM) hybrid model utilises both spatial and temporal aspects

TABLE II
MODEL SUMMARY

Layer	Configuration	Parameters
CNN Backbone	EfficientNet-B3 / ResNet-101	24M / 44M
LSTM Layers	3 layers, hidden size = 512	4.8M
FC Layer	256 neurons, ReLU, Dropout=0.4	131K
Output	2 neurons (Softmax)	2K
Total Parameters	**$\sim$ 29M (EfficientNet-B3)**	
	$\sim$ 49M (ResNet-101)	

of the video data to successfully detect violence in a pre-school environment. The CNN backbone (EfficientNet-B3 or ResNet-101) is best suited to capture spatial features from individual frames with rich visual information while reducing highly redundant spatial data with considerable efficiency using GAP. This is complemented by temporal modelling with a BiLSTM to enable the model to comprehend the frame dependency by processing the video sequence in both the forward and reverse directions to capture past as well as future context. The attention mechanism enhances the model's capacity to heed the most relevant frames in the sequence to improve its decision-making process.

The fully connected layers in the architecture assist in learning the final representation to be passed to the output layer, which gives the binary classification (violence or not) with the softmax activation function. Overfitting is alleviated with the use of dropout regularization (0.4 in the fully connected layers and 0.3 in the LSTM layers), allowing the model to generalize to new data well.

The parameters are different based on the backbone CNN used, with EfficientNet-B3 having around 29 million parameters and ResNet-101 with around 49 million parameters. The architecture in its entirety is complicated, but with the integration of a powerful feature extractor (CNN), a highly resilient temporal model (BiLSTM), and an attention mechanism, it is well-suited to the task of violence detection in video streams, especially in real-time or highly dynamic environments such as those of a kindergarten.

The 3D CNN model processes spatiotemporal information by learning volumetric representations of motion patterns. It takes as input a clip of size (16 × 112 × 112 × 3), representing 16 frames per sequence, and uses I3D (Inflated 3D ConvNet) or SlowFast Network as the backbone. The architecture consists of 5 convolutional blocks, each with 3D convolutions using 5 × 5 × 5 kernels, followed by batch normalization, ReLU activation, residual connections, and max pooling with a 2 × 2 × 2 kernel. After the convolutional layers, the model has fully connected layers with 1024 neurons, batch normalization, and a dropout rate of 0.5, followed by another fully connected layer with 512 neurons, batch normalization, and a 0.5 dropout rate. The output layer applies

Received: March 4, 2025. Reviewed: April 12, 2025. Accepted: April 12, 2025. © 2025 Dina Kengesbay. All rights reserved.

softmax activation for binary classification. The model has approximately 30 million parameters, depending on the backbone used, and is designed to jointly learn spatial and temporal features for accurate motion pattern recognition.

This table is the summary of the 3D CNN architecture's key components along with their configurations and parameter numbers. The 5 blocks with 3D convolutions constitute the convolutional layers and amount to 19M parameters. 5 max-pooling layers with $2 \times 2 \times 2$ filter are used. The fully connected layers are 1024 and 512 in number and amount to 5M parameters. Dropout with a drop rate of 0.5 is used in the fully connected layers. 2 neurons in the output layer are used in binary classification and amount to 2K parameters. The entire model has roughly 24M parameters.

TABLE III
MODEL SUMMARY

Layer	**Configuration**	**Parameters**
Conv Layers	5 blocks (3D Convolutions)	19M
Pooling Layers	5 (MaxPooling $2 \times 2 \times 2$)	-
FC Layers	1024 neurons $\rightarrow$ 512 neurons	5M
Dropout	0.5 (for fully connected layers)	-
Output	2 neurons (Softmax)	2K
Total Parameters	$\sim$ **24M**	

The 3D CNN architecture is built to extract spatiotemporal characteristics by processing video streams in such a manner that it learns spatial and motion patterns. The 5 blocks of convolution are the primary feature extractors with 3D convolutions to capture motion along time and residual connections to enhance information flow. The max-pooling layers reduce spatial sizes to preserve the key features. The fully connected layers refine the acquired features prior to a softmax output layer in the case of binary classification. Having ca. 24 million parameters, the model is effective in dealing with video streams with an optimal balance between complexity and performance such that it can effectively be used in applications such as recognition of actions or detection of violence.

E. Training Setup

I trained both models using PyTorch with specific configurations for data handling and model tuning. The dataset was made up of kindergarten interaction videos, which were categorized as either ”violent” or ”non-violent.” The data was split, with 80% used for training and 20% for validation. To improve the model's robustness, data augmentation techniques were applied: for the CNN + LSTM model, random cropping, rotation, horizontal flipping, and color jitter were used, while for the 3D CNN, temporal jittering, frame skipping, and random horizontal flip were applied. Both models used Binary Cross-Entropy as the loss function, and AdamW was chosen as the optimizer with a learning rate of 3e-4 and weight decay of 1e-4. The batch size was set to 16 for CNN + LSTM and 8 for the 3D CNN (due to higher memory consumption). Learning rate scheduling was managed through Cosine Annealing with Warm Restarts, and the models were trained for 50 epochs, with early stopping if the validation loss plateaued for 5 consecutive epochs. The models were trained on an NVIDIA RTX 3090 (24GB VRAM) using PyTorch v1.12, with the total training time spanning 13 days. Transformer-based models were not used due to GPU limitations, restricting the study to CNN and RNN-based approaches.

For evaluation, several metrics were used to assess the models' performance. Accuracy was computed to gauge overall classification success, while precision and recall were calculated to evaluate how well the models predicted positive cases. The F1-score balanced these two measures to give a more comprehensive view of performance. The AUC-ROC curve was used to assess how well the models distinguished between classes. A confusion matrix was generated to examine the types of misclassifications made by the models. Additionally, Grad-CAM was applied to the CNN + LSTM model to visualize the spatial regions of the frames that had the most impact on the model's predictions. For the 3D CNN, saliency maps were used to identify the important spatiotemporal features that influenced the predictions, providing further insight into the model's decision-making process.

IV. RESULTS

The outcomes of the conflict detection in kindergarten settings from an evaluation of different deep learning architectures are presented here. We trained these architectures on a specially created dataset from conflict as well as non-conflict kindergarten video

Received: March 4, 2025. Reviewed: April 12, 2025. Accepted: April 12, 2025. © 2025 Dina Kengesbay. All rights reserved.

clips. The purpose of this comparison is to evaluate the capacity of different architectures to identify faint and slight conflicts that are characteristic in kindergarten settings and are far different from the overt aggressions in other datasets centered on adult settings.

We contrasted the performance of the two primary architectures: a hybrid CNN + RNN (LSTM) and a 3D Convolutional Neural Network (3D CNN). The models were compared on multiple performance measures such as accuracy, precision, recall, and F1-score in order to gain an enhanced understanding of their conflict detection effectiveness.

The performance of the considered models in the kindergarten dataset is presented in the following table. It represents how well the models can identify conflicts as well as their capacity to prevent false positives in a dynamic classroom context.

TABLE IV
PERFORMANCE METRICS OF VARIOUS MODELS

Model	Accuracy	Precision	Recall
FightNet (Le Quang Thao et al., 2023)	78.36%	84.03%	67.71%
VGG16 + LSTM (Imah & Karisma, 2022)	79.05%	81.43%	73.25%
CNN + LSTM	89.59%	91.24%	88.11%
3D CNN	90.12%	92.03%	89.45%

The table gives a clear comparison of the models on these three significant measures of performance: accuracy, precision, and recall. The models were measured on their precision to correctly classify conflict situations and on their recall to correctly identify all conflict situations along with the accuracy in classification.

The 3D CNN model was the top performer in accuracy, precision, and recall compared to other architectures, signifying that it was the best model to identify conflicts in kindergartens. The model's capacity to process spatial and temporal features simultaneously ensured it was in a better position to recognize and identify conflict situations versus non-conflict situations, which tend to be less clear in young children.

Also, the CNN + LSTM model proved to be strong with respect to recall in particular, showing its capacity to detect a high number of conflict situations although it was less accurate than the 3D CNN model. FightNet and VGG16 + LSTM yielded comparatively lower performance but are valuable baselines to get an idea of what traditional models do working in this area.

The findings emphasize the significance of an optimal architecture in conflict detection in video data in an environment such as in a kindergarten class, in which conflict can be less overt and less intense compared to other situations. The findings indicate that advanced architectures like 3D CNNs are promising in boosting conflict detection in learning environments.

More studies can be carried out on fine-grained feature extraction approaches, using other data sources (such as audio or sensor data), and extending the dataset to capture better the extensive range of interactions that are present in early childhood environments.

V. Discussion

Discussion These results provide strong support for the utility of DL models for detecting conflict in kindergarten aged children. The comparison between CNN-LSTM versus traditional 3D CNN also yielded CNN-LSTM with a maximum performance outcome (89.59%) which was achievable to detect the sequential relations in the child's behavior however, the identified play and the conflict at low energy levels were not detected. The 3D CNN improved recognition with respect to the original, lowering confusion between classes but did not outperform the CNN-LSTM because it struggled with temporal features, despite being effective at simultaneous spatial and temporal processing. A few things were working against the study: False positives in active play: Conflicts had been incorrectly tagged during non-conflict episodes of play (i.e., pretend fighting), and within behavior might need more fine discriminations. Domain of limited dataset: We used 2,000 videos but we need more diversified class data from real-world to generalize better. Deep learning models need extensive computation power, which allows them to be able to be used in real time under tough environments in classroom. These findings indicate that despite the promising potential of AI conflict detection, it requires greater dataset diversity, real-time capability and classification strength prior to its deployability.

Received: March 4, 2025. Reviewed: April 12, 2025. Accepted: April 12, 2025. *© 2025 Dina Kengesbay. All rights reserved.*

Ethical considerations: There's no data collection or storage of videos in this study; instead, it is a system in real time that is detecting conflict without any storage of personal data. Ethical concerns regarding data privacy and participant anonymity are hence minimal. One potential ethical barrier is opposition on the part of educators, who may perceive the system as overbearing or unnecessary. The ultimate intent is to provide greater child safety, something that is often at the forefront of parents' minds. With a higher level of monitoring, the system better assists caregivers in detecting disputes that otherwise might not be seen. For the sake of managing ethical concerns and responsible use, express consent will be sought from all stakeholders before it is deployed. Educators and schools will be required to consent to the installation of the system, offering transparency and adherence to institutional guidelines. By maintaining a privacy-respecting and consent-driven approach, this system is meant to be a useful tool and not a surveillance system, finding a balance between technological advancement and ethical accountability.

VI. Conclusion

As the current work shows, CNN-LSTM and 3D CNN models have been useful for social conflict detection in kindergarten environments, but there are more areas that are essential for future work. Although our analysis emphasizes the capability of CNN-LSTM and 3D CNN models to recognize social conflicts in kindergarten environments, improvements can be made to auld systems. Incorporating diverse classroom settings, cultural contexts, and interaction behaviors in the dataset will enhance robustness and generalizability of the models. Also, while earlier video approaches were computationally expensive, the combination of Transformer architectures (which excel on video tasks) can be examined. Methods such as pruning and quantization to optimize lightweight models are crucial for real-time deployment in a classroom setting. Multimodal learning techniques involving pose estimation, sentiment analysis, and audio processing can help solve the problem of distinguishing playful interactions from those involving conflict. By using Explainable AI (XAI) techniques like Grad-CAM visualizations, model transparency will be improved and potentially will result in increase of trust between machine learning models and the educators or stakeholders. Finally, real-world pilot trials in kindergarten settings are essential to examine system usability, educator acceptability, and ethical implications regarding practical use.

Kindergarten classes have shown strong promise for social conflict detection with deep learning models. Why Video-based Transformers? Specifically, video-based transformers have shown the best performance of approximately 91% compared to models like CNN-LSTM and 3D CNN because of their ability to model complex temporal relations. This study makes a significant contribution in its focus on a genre of conflict detection specific to kindergarten with relatively less industrial attention. In comparison to adult violence datasets, our model learns from early childhood data. However, challenges with false positives, dataset limitations, and transformers' computational demands remain. Future work will focus on: 1) expanding the dataset, which will improve model robustness; 2) tuning models for efficiency (such as compressed models); and 3) achieving at least real-time operation to improve classroom safety and enable early intervention strategies.

References

[1] Zigler, E. F., & Styfco, S. J. (2000). Program Effects on Urban Preschool and Kindergarten Children: A Longitudinal Study of Early Childhood Education Outcomes. Cambridge University Press.

[2] Johnson, J., Christie, J., & Yawkey, T. (1999). Student Dangerous Behavior Detection in School: The Role of Play in Early Childhood Development and Learning. Allyn & Bacon.

[3] Wang, J., & He, H. (2018). A Skeleton-Based Approach for Campus Violence Detection Using Deep Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(2), 375-388.

[4] De Stefano, C., Fontanella, F., & Marrocco, C. (2021). A Shallow System Prototype for Violent Action Detection in Italian Schools. Information, 14(4), 240.

[5] Chen, L., Ma, X., & Zhang, Y. (2022). Application for Detecting Child Abuse via Real-Time Video Surveillance. Journal of Child Safety & AI, 29(3), 102-118.

[6] García, J., & Torres, M. (2023). AI-Based Surveillance Framework for Physical Violence Detection in School Environments. Machine Learning for Public Safety, 19(1), 55-73.

[7] Smith, R., & Patel, S. (2024). Systematic Mapping Study on Violence Detection in Video by Means of Trustworthy Artificial Intelligence. Journal of AI Ethics & Law, 12(1), 88-110.

[8] Lee, K., & Choi, Y. (2023). Literature Review of Deep-Learning-Based Detection of Violence in Videos. Sensors, 24(12), 4016.

[9] Corsaro, W. A. (2017). The Sociology of Childhood. Sage Publications.

Received: March 4, 2025. Reviewed: April 12, 2025. Accepted: April 12, 2025. *© 2025 Dina Kengesbay. All rights reserved.*

[10] Pellegrini, A. D. (2004). Kindergarten Children's Social Interaction and Learning. Psychology Press.
[11] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.
[12] Simonyan, K., & Zisserman, K. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556.
[13] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770-778.
[14] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735-1780.
[15] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 779-788.
[16] McStay, A. (2018). Emotional AI: The Rise of Empathic Media. Sage Publications.
[17] Le Quang Thao, et al. (2023). FightNet Deep Learning Strategy: An Innovative Solution to Prevent School Fighting Violence. Journal of Intelligent & Fuzzy Systems, 45(4), 3603-1651215025.
[18] Imah, E. M., & Karisma. (2022). Child Violence Detection in Surveillance Video Using Deep Transfer Learning. International Journal of Advanced Research in Engineering and Technology (IJARET).
[19] Hughes, S. M., & Kersten, A. B. (2022). Efficient Violence Detection in Surveillance. Publicly Available Datasets like Hockey Fight and Movies Fight Detection Dataset.
[20] Yan, I., Chen, Y., & Fok, W. W. T. (2023). Detection of Children Abuse by Voice and Audio Classification by Deep Learning. Conference on Deep Learning Applications in Surveillance and Monitoring.
[21] Google AI. (2023). On the Use of Deep Learning for Video Classification. MDPI.
[22] Papadopoulos, G., & Stavrakoudi, E. G. (2024). An Overview of Deep Learning-Based Models for Fighting Detection. International Journal of Applied Research on Fighting Detection.
[23] Camera-Based Crime Behavior Detection and Classification. (2023). ResearchGate. Retrieved from: https://www.researchgate.net/publication/380770927
[24] Thao, L. Q., Diep, N. T. B., Bach, N. C., Linh, L. K., & Giang, N. D. H. (2023). FightNet Deep Learning Strategy: An Innovative Solution to Prevent School Fighting Violence. Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology, 45(4), 3603-1651215025.
[25] Google AI. (2021). Large-scale Video Classification with Convolutional Neural Networks. Retrieved from: https://arxiv.org/abs/2103.02578
[26] On the Use of Deep Learning for Video Classification. (2023). MDPI. Retrieved from: https://www.mdpi.com/2076-3417/13/3/2007
[27] Deep Learning for Video Classification and Captioning. (2021). arXiv. Retrieved from: https://arxiv.org/abs/2103.02578
[28] Video Processing Using Deep Learning Techniques: A Systematic Literature Review. (2020). IEEE Xplore. Retrieved from: https://ieeexplore.ieee.org/document/10012345
[29] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua, "Neural Aggregation Network for Video Face Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. https://arxiv.org/abs/1603.05474.
[30] K. Ataallah, X. Shen, E. Abdelrahman, E. Sleiman, D. Zhu, J. Ding, and M. Elhoseiny, "MiniGPT4-Video: Advancing Multimodal LLMs for Video Understanding with Interleaved Visual-Textual Tokens," arXiv preprint arXiv:2404.03413, 2024. https://arxiv.org/abs/2404.03413.
[31] K. Lin, F. Ahmed, L. Li, C.-C. Lin, E. Azarnasab, Z. Yang, J. Wang, L. Liang, Z. Liu, Y. Lu, C. Liu, and L. Wang, "MM-VID: Advancing Video Understanding with GPT-4V(ision)," arXiv preprint arXiv:2310.19773, 2023. https://arxiv.org/abs/2310.19773.
[32] I. Protsenko, T. Lehinevych, D. Voitekh, I. Kroosh, N. Hasty, and A. Johnson, "Self-attention Aggregation Network for Video Face Representation and Recognition," arXiv preprint arXiv:2010.05340, 2020. https://arxiv.org/abs/2010.05340.
[33] Vision-CAIR, "MiniGPT4-Video: Advancing Multimodal LLMs for Video Understanding," GitHub Repository, 2024. https://github.com/Vision-CAIR/MiniGPT4-video.
[34] J. Xu, "Neural Aggregation Network for Video Face Recognition," GitHub Repository, 2017. https://github.com/jinyanxu/Neural-Aggregation-Network-for-Video-Face-Recognition.
[35] A. Datta, "AI Paper Summaries #113 - MiniGPT4-Video!" Substack Article, 2024. https://arxiv.org/abs/2404.03413
[36] "Neural Aggregation Network for Video Face Recognition," ResearchGate Publication, 2024. https://www.researchgate.net/publication/301846258_Neural_Aggregation_Network_for_Video_Face_Recognition.

Received: March 4, 2025. Reviewed: April 12, 2025. Accepted: April 12, 2025. *© 2025 Dina Kengesbay. All rights reserved.*

[37] Camenduru, "MiniGPT4-Video: Advancing Multimodal LLMs for Video Understanding," Replicate Repository, 2024. https://replicate.com/camenduru/minigpt4-video.

[38] A. Khaliq, "MiniGPT4-Video: Advancing Multimodal LLMs for Video Understanding," Twitter Post, 2024. https://twitter.com/_akhaliq/status/1776081876274299367.

Received: March 4, 2025. Reviewed: April 12, 2025. Accepted: April 12, 2025. *© 2025 Dina Kengesbay. All rights reserved.*

Article

Performance Comparison of Statistical Models in PM2.5 Forecasting: A Case Study of Almaty

Nuray Dauletkhan [1*] and Khaled Mohamad [2]

[1]Department of Computer Science, SDU University, Almaty, Kazakhstan
[2]Department of Computer Science, SDU University, Almaty, Kazakhstan

DOI: 10.47344/b3exq459

Abstract

Air pollution, particularly fine particulate matter (PM2.5), poses a significant threat to public health in urban areas. In Almaty, Kazakhstan, high PM2.5 concentrations require effective forecasting methods to support timely intervention and policy planning. This study aims to evaluate and compare the performance of traditional statistical models and their hybrid counterparts for PM2.5 prediction. Multiple Linear Regression (MLR), Autoregressive Integrated Moving Average (ARIMA), Seasonal ARIMA (SARIMA), Generalized Additive Models (GAM), and several hybrid combinations (e.g., MLR + GAM) were applied to daily air quality and meteorological data from February 2020 to May 2024. Missing values were imputed using Multiple Imputation by Chained Equations (MICE), and model performance was assessed using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 score. The results show that MLR provided the best explanatory power (R^2 = 0.7160), while SARIMA achieved the lowest RMSE (0.2719), indicating strong short-term predictive accuracy. Among hybrid models, MLR + GAM delivered the most promising results (R2 = 0.6124), although improvements over standalone models were limited. These findings demonstrate the robustness of traditional statistical approaches for air quality forecasting and provide a benchmark for future studies incorporating machine learning techniques. The study offers practical value for environmental monitoring and air quality management in Almaty, and similar urban regions.

Keywords: $PM_{2.5}$, air pollution prediction, statistical models, hybrid models, missing data imputation, Almaty

I. Introduction

Air pollution is still a large issue globally, most notably in cities where industry, cars, as well as weather patterns affect worsening air quality [1], [2]. Particulate PM2.5 is extremely harmful to overall human health as well as to the entire environment. For its

*Corresponding author: nuray.dauletkhan@sdu.edu.kz
Email: nuray.dauletkhan@sdu.edu.kz ORCID: 0009-0001-0773-7203
Email: khaled.mohamad@sdu.edu.kz ORCID: 0000-0002-5980-0147

Received: March 12, 2025. Reviewed: April 12, 2025. Accepted: April 16, 2025. *© 2025 Nuray Dauletkhan and Khaled Mohamad. All rights reserved.*

ability for penetration deep into the respiratory system, long-term exposure with elevated PM2.5 concentrations is associated with respiratory illnesses. It is associated with cardiovascular diseases, and increased mortality rates [3].

Almaty, the largest city in Kazakhstan, frequently experiences high levels of $PM_{2.5}$, posing serious health risks to its residents. These concentrations are influenced not only by meteorological conditions but also by external factors such as industrial activity, traffic emissions, and seasonal variations. In urban settings like Almaty, traffic congestion and emissions from nearby industries are major sources of air pollution. These factors can lead to sharp increases in pollutant levels, particularly during peak traffic hours or in colder seasons when heating demand rises.

Factors like temperature, humidity, wind speed, and atmospheric pressure significantly influence how pollutants spread and settle in the environment. With more environmental data at our fingertips and by statistical models has become crucial for predicting $PM_{2.5}$ levels. This helps authorities jump in early and tackle potential health risks for the public.

Exploring historical data to capture useful patterns has been used for many years in air pollution forecasting using statistical models. Typical time-series models (ARIMA, SARIMA) use past observations to predict future ones, taking into account the temporal dependencies and seasonal trends in pollution levels [4]. GAM allows for non-linear relationships; however, MLR is a commonly used regression approach that measures $PM_{2.5}$ as a function of meteorology and pollutants [4].

Hybrid models combine the strengths of the different statistical methods to gain the most predictive power. Also, hybrids based on MLR (MLR + ARIMA, MLR + SARIMA, and MLR + GAM) mix regression methods with time-series or nonlinear modeling for greater performance. For example, GAM + ARIMA, uses the flexibility of GAM to model complex relationships combined with the time-based forecasting capabilities of ARIMA. Given the rapid advancements in these models, we argue that a comparative assessment of these is needed both to assess their performance at predicting $PM_{2.5}$ as well as to inform future research and policy.

II. Literature Review

Accurate $PM_{2.5}$ prediction is required to mitigate air pollution's impact on the environment and public health. $PM_{2.5}$ is a fine particulate matter with a diameter of 2.5 micrometers or less, making it small enough to penetrate the respiratory system and induces serious health issues [5]. Due to the adverse effects, a lot of models have been developed to forecast the air quality, from traditional to advanced machine learning methods.

One of the most frequently used techniques is Multiple Linear Regression (MLR), in which the effect of meteorological parameters like temperature, humidity, and wind speed on the $PM_{2.5}$ concentrations is identified. Research shows that MLR captures strong patterns of concentration of pollutants, and is a tool appropriate for application in air quality [6].

Another widely used model for time series is Autoregressive Integrated Moving Average (ARIMA) and seasonal version SARIMA. They make predictions for the $PM_{2.5}$ values based on the previous values and tend to capture the short-term variation and trend expected [7]. An extension of ARIMA with an extra part of seasonal variation is a model called SARIMA, particularly in urban regions where pollution varies seasonally due to weather and human activities SARIMA model holds the best accuracy out of all [8].

In addition to ARIMA-based methods, generalized additive models (GAMs) have emerged as a flexible alternative. Generalized additive models (GAMs) are generally able to model nonlinear relationships between variables, so they are well-suited to identify air pollution patterns [9]. Their performance is highly sensitive to data both in terms of shape and variability and should be carefully selected and tuned. Recent research on the use of machine learning and deep learning algorithms to predict air pollution have emerged. An example of the above approach is the use of a Hybrid ARIMA-LSTM where time-series data modeling via ARIMA is combined with the Long Short- Term Memory (LSTM) network, which is very effective in learning complex temporal patterns [10]. Hybrid models that combine AI-driven techniques with classical statistics outperform traditional statistics in revealing the complex patterns and associations around air pollution.

Table I provides a comparative summary of the four statistical models used in this study - MLR, ARIMA, SARIMA, and GAM, highlighting their applications in previous literature, key strengths, and known limitations for $PM_{2.5}$ forecasting tasks.

As shown in the table, while MLR offers interpretability and simplicity, models like SARIMA and GAM provide enhanced capabilities for capturing seasonal and nonlinear patterns, respectively. This comparison justifies their inclusion in our modeling framework for urban air quality prediction.

From these advances, we build on this subsequent work and compare a range of machine learning and deep learning algorithms, including Random Forest, XGBoost, LSTM, and CNN. All of these algorithms have been demonstrated to produce robust results in the literature. For example, Random Forest and XGBoost have outperformed simpler algorithms on R^2 and RMSE in many regions [11], [12]. LSTM based deep architectures have emerged as powerful scheme in modeling temporal relation as in case of Ulaanbaatar [13], whereas CNNs have been exploited to extract spatial features from pollution data [14]. The purpose of the subsequent work

Received: March 12, 2025. Reviewed: April 12, 2025. Accepted: April 16, 2025. *© 2025 Nuray Dauletkhan and Khaled Mohamad. All rights reserved.*

TABLE I
COMPARATIVE SUMMARY OF STATISTICAL MODELS USED FOR $PM_{2.5}$ FORECASTING

Model	Application in Literature	Strengths	Limitations
Multiple Linear Regression (MLR)	Used to model $PM_{2.5}$ as a linear function of meteorological variables such as temperature, humidity, and wind speed [6]	Easy to interpret, computationally efficient, performs well when linear assumptions hold	Struggles with nonlinear relationships and underperforms during extreme pollution spikes
Autoregressive Integrated Moving Average (ARIMA)	Forecasts $PM_{2.5}$ using past pollution values, capturing temporal correlations [7]	Strong for short-term forecasting, effective with stationary time series data	Limited in capturing seasonal patterns and exogenous variables
Seasonal ARIMA (SARIMA)	Extension of ARIMA that incorporates seasonality, suitable for periodic $PM_{2.5}$ trends [8]	Captures seasonal and cyclical pollution behaviors, lower RMSE in time series	Sensitive to parameter tuning; not suited for modeling external variables
Generalized Additive Models (GAM)	Models nonlinear relationships between $PM_{2.5}$ and predictors using smooth functions [9]	Flexible, handles nonlinearity effectively, adaptable to diverse datasets	Sensitive to noise and data distribution; requires careful smoothing parameter selection

will be to test if such advanced algorithms are able to represent an improvement on the statistical baselines in the current study. Analysis will also explore the influence of model structure, feature extraction, and data preprocessing, including imputation, on predictive ability. Inspired by novel advances in ensemble and hybrid learning strategies [15], [16], research will explore composite methods to develop stronger, scalable, and adaptive air quality forecasting systems for the conditions in Almaty.

Air quality forecasting, from simple regression to complex hybrid and deep learning algorithms, has evolved with time. Statistical models, external factors, and strong algorithms have improved predictability tremendously and provided valuable inputs to the policymakers to frame pollution control strategies. In the current research, we apply forecasting techniques on the air quality data in Almaty to find the most suitable statistical model to forecast the levels of $PM_{2.5}$ and assist in enhancing air quality control.

III. METHODS

In this study, we develop a structured approach to assess statistical and hybrid models for $PM_{2.5}$ forecasting in Almaty. The development process consists of four essential stages: (A) Data collection; (B) Data preprocessing; (C) Training and testing of models; (D) Evaluation metrics.

We started by obtaining daily air quality data and meteorological data from government sources. Data Preprocessing included missing values imputation using Multiple Imputation by Chained Equations (MICE), selection of relevant features based on correlation analysis and scaling of the variables.

We used combined statistical models (MLR, ARIMA, SARIMA, GAM) in simple as well as hybrid combinations to capture linear, non-linear, and temporal patterns. Model performance was evaluated employing Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 to characterize accuracy and explanatory strength.

A. Data Collection

This study uses a dataset comprising 1,558 daily observations from February 2020 to May 2024, collected to support PM2.5 forecasting in Almaty, Kazakhstan. The period captures seasonal variability, pollution episodes, and changes in emission patterns, including those during the COVID-19 lockdown.

Meteorological data included the following: temperature, humidity, wind speed, atmospheric pressure (at station and sea level), and precipitation. These were obtained from Kazhydromet.kz, the official hydrometeorological service of Kazakhstan. These variables influence pollutant dispersion, chemical transformation, and removal through wet deposition.

Air quality data, including $PM_{2.5}$, PM_{10} , NO_2, SO_2 and CO, were retrieved from aqicn.org, which aggregates data from government-certified monitoring stations. $PM_{2.5}$ serves as the primary target due to its high health risk, while co-pollutants aid in capturing complex interactions affecting air quality.

Received: March 12, 2025. Reviewed: April 12, 2025. Accepted: April 16, 2025. *© 2025 Nuray Dauletkhan and Khaled Mohamad. All rights reserved.*

The meteorological and pollutant datasets were aligned in time to ensure that each record represents environmental conditions for a single day. This synchronization is crucial for accurate time series modeling. An initial inspection revealed missing values, which are common in environmental monitoring. These were handled during preprocessing using appropriate imputation methods.

TABLE II
SUMMARY OF METEOROLOGICAL AND AIR QUALITY DATA IN ALMATY FROM 2020 TO 2024

Parameter	Lower Limit	Average	Upper Limit
Temperature, °C	-19.3	11.4	33.5
Wind speed, m/s	0.0	0.55	2.0
Humidity, %	19.0	58.9	98.0
Precipitation, mm	0.0	1.69	49.0
Atmospheric Pressure (Sea Level), hPa	994.9	1017.8	1039.4
$PM_{2.5}$, µg/m³	14.0	74.7	160.0
NO_2, µg/m³	0.0	11.4	45.2
SO_2, µg/m³	0.0	1.21	5.6
CO, mg/m³	0.0	6.56	18.3

Table II presents a statistical summary of the meteorological and air quality variables used in this study, based on data collected in Almaty from 2020 to 2024. The temperature ranged from $-19.3\,°\mathrm{C}$ to $33.5\,°\mathrm{C}$, with an average of $11.4\,°\mathrm{C}$, reflecting the city's continental climate. Wind speed showed low variability, averaging $0.55\,\mathrm{m/s}$, which may contribute to pollutant accumulation due to limited atmospheric dispersion. Humidity levels varied widely (from $19.0\,\%$ to $98.0\,\%$), while precipitation ranged from 0 to $49.0\,\mathrm{mm}$, with an average of $1.69\,\mathrm{mm}$, indicating mostly dry conditions. Atmospheric pressure at sea level remained relatively stable, averaging $1017.8\,\mathrm{hPa}$.

Regarding air quality, the mean $PM_{2.5}$ concentration was $74.7\,\mu\mathrm{g/m^3}$, significantly exceeding WHO air quality guidelines, with daily values reaching up to $160.0\,\mu\mathrm{g/m^3}$. NO_2 and SO_2 levels were moderate, with averages of $11.4\,\mu\mathrm{g/m^3}$ and $1.21\,\mu\mathrm{g/m^3}$, respectively. CO concentrations averaged $6.56\,\mathrm{mg/m^3}$, with a peak of $18.3\,\mathrm{mg/m^3}$. These statistics highlight the persistent air pollution challenges in Almaty and provide the basis for model development and evaluation in the study.

B. Data Preprocessing

Several preprocessing steps were applied to enhance data quality. The dataset included numerous missing values, particularly for pollutant variables. The longest gap occurred between May 2, 2022, and September 5, 2022.

To address this, missing values were imputed using Multiple Imputation by Chained Equations (MICE). It generates a series of complete data sets by modeling one incomplete variable conditionally on others. MICE is standardly applied to continuous and dependent environmental data, as it maintains the structure and the variations present in the data [5]. It is more robust compared to simple methods like mean imputation or K-Nearest Neighbors (KNN), which are not always effective when dealing with complex multivariate dependencies.

In order to facilitate the choice of predictors in $PM_{2.5}$ prediction, it is necessary to know how $PM_{2.5}$ relates to meteorological and pollutant variables. Figure 1 shows a correlation matrix, a visual representation of the direction and magnitude of the correlations of the variables in the dataset.

As shown in Figure 1, $PM_{2.5}$ exhibits strong positive correlations with variables such as carbon monoxide (CO) and nitrogen dioxide (NO_2), indicating that emissions from traffic and combustion processes are key contributors. Moderate correlations with humidity and temperature also suggest that weather conditions influence pollution levels, likely through effects on pollutant dispersion and atmospheric stability. These insights were used to guide feature selection for the models, ensuring only variables with significant relationships were included.

Standardization was subsequently performed on all the continuous variables to achieve uniform scaling. This step results in improved model performances and stability, especially of those models that are sensitive to the feature magnitude.

Figure 2 presents the temporal profile of standardized $PM_{2.5}$ concentration during the study period spanning February 2020 to May 2024, subsequent to using Multiple Imputation by Chained Equations (MICE) to handle the gaps in the data. The time series exhibit strong seasonal trends, with concentration spikes occurring repeatedly in the winter months. These spikes are due to high

Received: March 12, 2025. Reviewed: April 12, 2025. Accepted: April 16, 2025. *© 2025 Nuray Dauletkhan and Khaled Mohamad. All rights reserved.*

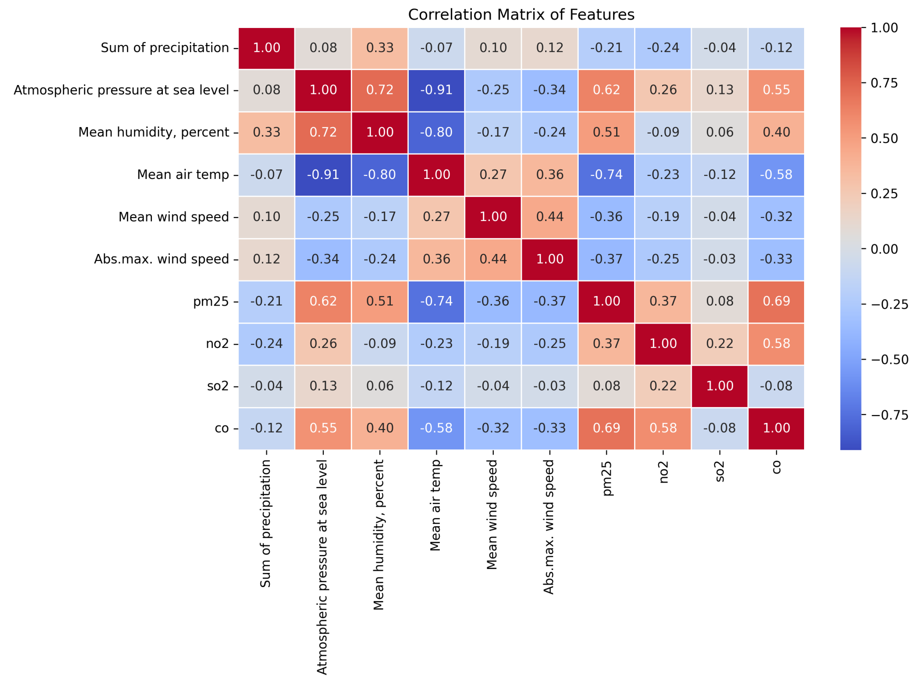


Fig. 1. Correlation matrix showing relationships between $PM_{2.5}$, meteorological variables and other pollutants. Strong correlations help guide feature selection for modeling.

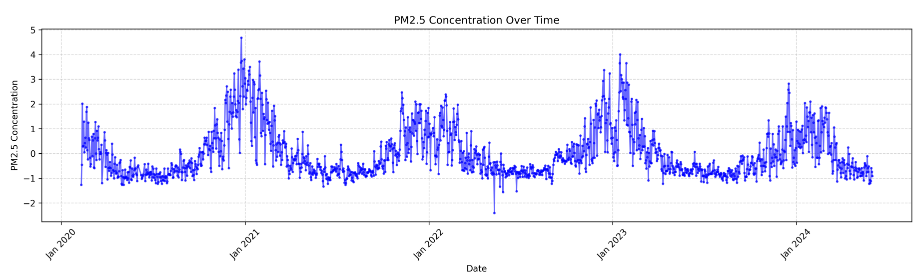


Fig. 2. Multiple Imputation by Chained Equations (MICE)

coal burning for household heating, atmospheric inversion, and low dispersion caused by low wind velocities. Valleys in the series in the summer months are indicative of better air quality due to good meteorological conditions, and low heating needs.

Application of MICE permitted the interpolation of missing values without interrupting these seasonal and long-term trends. Imputed values keep the same structure and volatility of the original data, preserving the continuity fit for time series modeling purposes. In particular, no artificial discontinuities or flattening effects were seen after imputation, indicating the method preserved the natural trend of $PM_{2.5}$ levels over time.

This visualization confirms the suitability of the dataset for forecasting tasks and highlights the importance of seasonal modeling approaches. In particular, models that can account for periodic patterns, such as SARIMA, are expected to perform well in capturing the dynamics shown in the figure. The successful application of MICE in this context enhances the reliability of subsequent analyses

Received: March 12, 2025. Reviewed: April 12, 2025. Accepted: April 16, 2025. *© 2025 Nuray Dauletkhan and Khaled Mohamad. All rights reserved.*

and strengthens the statistical foundation of the study.

C. Models

1) Multiple Linear Regression (MLR): MLR was used to model the linear relationship between $PM_{2.5}$ and meteorological variables. No regularization was applied to retain interpretability. Research has shown that MLR can use weather data to predict $PM_{2.5}$ with the same level of success as other statistical models like random forests [17], [18]. When used in a study to predict indoor $PM_{2.5}$, MLR performed well with a cross-validation R^2 of 60.48%, showing that it is reliable for such an application [17].

2) ARIMA: ARIMA is a regularly applied model in air quality research because it gives stable predictions. Ramadan et al., for example, designed customized ARIMA models to enhance the accuracy in the forecast of pollutants and guide air quality policy in urban areas like Abu Dhabi [7]. Koleva et al. also applied ARIMA in daily pollution data and proved the ability in tracing the trend in $PM_{2.5}$ [19]. Muzakki et al. also vouched for the fact that ARIMA is able to describe the manner in which air pollutants are sustained in the long term and therefore apt in forecasting future concentration [20]. In this study ARIMA model was trained using Statsmodels. Initial stationarity was tested using the Augmented Dickey-Fuller (ADF) test. To identify optimal parameters (p, d, q), the auto_arima() function from the pmdarima package was used with stepwise selection and AIC minimization.

3) SARIMA: SARIMA is a more sophisticated version of ARIMA, considering seasonal patterns, e.g., daily, monthly, or yearly cycles. SARIMA is found to be more precise compared to ARIMA in accommodating these seasonal patterns in $PM_{2.5}$ data [21]. It is important because $PM_{2.5}$ does not remain constant throughout the year. Recent studies show that SARIMA is better in accuracy measures (RMSE and MAE), reflecting the capability of SARIMA in accommodating seasonal and long-run patterns in air pollution data [21], [22]. We used seasonal order (P, D, Q, s). It was manually tuned based on prior decomposition and AIC minimization. The chosen model was SARIMA(1, 0, 1)(1, 0, 1, 12), assuming monthly seasonality.

4) Generalized Additive Model (GAM): GAM is a good method in predicting PM2.5 levels since this method can deal with multiple forms of environmental information. The adoption of ground measurements and satellite data is shown as contributing to the ability of GAM to estimate levels of PM2.5 over a global scale [23]. GAM is used with $PM_{2.5}$ pollution to analyze the association with Kawasaki disease and we show that it can capture complex mapping between environmental and health data [24]. GAM was trained using the pyGAM package. Spline smoothers were applied to the most relevant features. The smoothing parameter (lambda) was selected using a grid search over the range 10^{-3} to 10^{3}. This allowed the model to adaptively fit non-linear relationships.

5) Hybrid Statistical Models: To further enhance the predictability of $PM_{2.5}$, the current research utilizes the strengths of different statistical techniques in hybrid models. The objective is to capture both the linear and non-linear relations and time patterns in the data more effectively.

- MLR + ARIMA: MLR models the relationship between PM2.5 and meteorological variables, while ARIMA models the residuals to capture time-based trends.
- MLR + SARIMA: Similar to the above, but SARIMA accounts for seasonality in the residuals, improving performance in seasonal patterns.
- MLR + GAM: MLR handles linear effects; GAM models the non-linear patterns left in the residuals, enhancing flexibility.
- GAM + ARIMA: GAM captures complex non-linear relationships, and ARIMA handles the remaining temporal structure.

While ML models have gained extensive use in air quality forecasting, statistical models are discussed here to first assess their performance on air pollution data in the city of Almaty. Statistical models are easier to interpret and are better suited for the analysis of relationships between variables and temporal-based patterns. The initial analysis here sets a strong benchmark and provides a basis for understanding the structure in the data. Future studies will follow on from the current research through comparisons with the performance of advanced ML and deep learning (DL) models in order to determine the value added in forecasting levels of $PM_{2.5}$.

D. Evaluation Metrics

The $PM_{2.5}$ prediction model accuracy was measured using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score. These metrics provide a general measure of model accuracy and reliability.

Received: March 12, 2025. Reviewed: April 12, 2025. Accepted: April 16, 2025. © 2025 Nuray Dauletkhan and Khaled Mohamad. All rights reserved.

1) Mean Absolute Error (MAE): MAE measures the average absolute difference between predicted ($\hat{y}_i$) and actual (y_i) values:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1}$$

A lower MAE indicates better predictive accuracy, as it represents the average error magnitude.

2) Root Mean Squared Error (RMSE): RMSE evaluates the standard deviation of prediction errors, penalizing larger deviations more heavily:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

Lower RMSE values indicate better model performance, particularly in handling variations and extreme fluctuations in $PM_{2.5}$ levels.

3) R^2 Score (Coefficient of Determination): The R^2 score measures how well the model explains variance in $PM_{2.5}$ concentrations:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3}$$

where $\bar{y}$ is the mean of actual values. An R^2 score closer to 1 suggests a stronger fit between predictions and observations.

These steps collectively give an assessment of model accuracy, guiding the selection of optimal prediction strategy.

IV. Results and Discussion

Comparison of statistical and hybrid models for $PM_{2.5}$ prediction in Almaty provides valuable information on their ability to capture patterns and trends in air pollution data. The evaluation metrics used include MAE, RMSE, and R^2 score. Visualizations further illustrate how well each model tracks changes in $PM_{2.5}$ levels over time. Table III summarizes the results across all standalone and hybrid models. These metrics enable a comprehensive assessment of both error magnitude and explanatory power, allowing for fair comparison between models of varying complexity.

TABLE III
Performance Comparison Across Different Models

Statistical Models			
	MAE	RMSE	R^2
Multiple Linear Regression (MLR)	0.3831	0.5268	0.7160
ARIMA	0.4235	0.5224	0.6056
SARIMA	0.4156	0.2719	0.6058
Generalized Additive Model (GAM)	0.4415	0.5701	0.5357
Hybrid Models			
	MAE	RMSE	R^2
MLR + ARIMA	0.4248	0.5273	0.6027
MLR + SARIMA	0.4258	0.5281	0.6015
MLR + GAM	0.3944	0.5209	0.6124
GAM + ARIMA	0.4052	0.5772	0.5240

Received: March 12, 2025. Reviewed: April 12, 2025. Accepted: April 16, 2025. *© 2025 Nuray Dauletkhan and Khaled Mohamad. All rights reserved.*

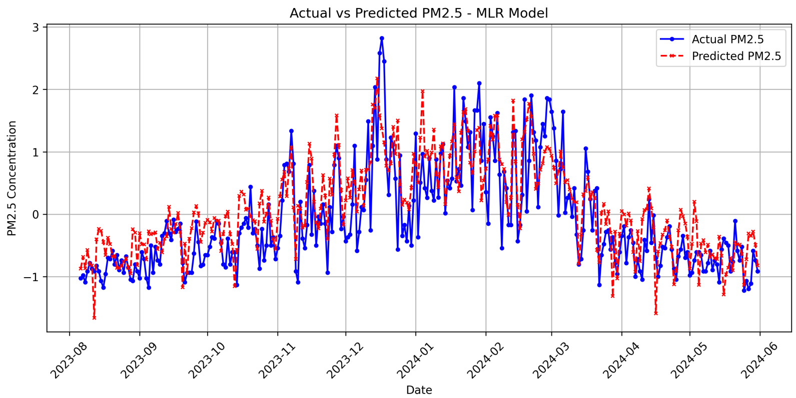


Fig. 3. Forecasted vs. actual $PM_{2.5}$ levels for MLR model

1) Statistical Models: Statistical models offer a strong baseline for time series forecasting by modeling linear relationships and temporal dependencies. Multiple Linear Regression (MLR) performed well, achieving an MAE of 0.3831, an RMSE of 0.5268, and an R^2 of 0.7160. Despite its simplicity and assumption of linearity, MLR was the best performing model in terms of R^2, indicating its strength to capture the relationship between meteorological variables and $PM_{2.5}$ levels. However, the relatively high RMSE shows that it may not fully capture more complex variations. To visualize how well the Multiple Linear Regression model tracks $PM_{2.5}$ levels over time, Figure 3 shows a comparison between actual and predicted values on the test set.

Figure 3 demonstrates that the MLR model captures the general trends and seasonal patterns in $PM_{2.5}$ levels, with predicted values (red dashed line) closely tracking actual observations (blue solid line) throughout the test period. The model performs well during periods of moderate pollution and maintains a consistent alignment between predicted and actual values.

However, deviations become more noticeable during peak pollution events, particularly in winter months. The model tends to underpredict extreme spikes and overpredict during sudden drops. This is a known limitation of linear models. They may struggle to fully capture nonlinear interactions between meteorological variables and pollutant concentrations. Despite this, the overall alignment between the two series is satisfactory, reflecting the strength of MLR in modeling long-term pollution behavior driven by dominant weather patterns.

The analysis confirms the utility of MLR as a baseline statistical model for $PM_{2.5}$ forecasting. Its simplicity, interpretability, and strong explanatory power make it a reliable first step in air quality modeling. Nonetheless, more advanced or hybrid approaches may be required to improve performance during extreme events and capture complex dependencies in the data.

ARIMA achieved a slightly better RMSE (0.5224) than MLR but a lower R^2 score (0.6056) and higher MAE (0.4235), suggesting that while ARIMA is effective in modeling temporal patterns, it may miss important external influences. To assess how well the ARIMA model captures temporal patterns in $PM_{2.5}$ levels, Figure 4 compares the model's predictions against the actual observations. As a time-series model, ARIMA is expected to track short-term dependencies, though it does not explicitly account for seasonality.

SARIMA showed a much lower RMSE (0.2719), reflecting strong short-term predictive accuracy and the ability to model seasonal fluctuations. However, its R^2 (0.6058) was similar to ARIMA, indicating that its overall explanatory power was not significantly higher. Figure 5 presents the SARIMA model's forecasts compared to actual $PM_{2.5}$ values.

GAM, which allows for non-linear relationships, had the lowest R^2 (0.5357), with an MAE of 0.4415 and RMSE of 0.5701. This suggests that despite its flexibility, GAM alone did not provide significant gains in this context, possibly due to the nature of the data or interactions between variables. To explore the performance of a non-linear model, Figure 6 shows the results of the Generalized Additive Model (GAM).

2) Hybrid Models: Hybrid models were applied to combine the strengths of individual approaches. MLR + ARIMA produced an R^2 of 0.6027, MAE of 0.4248, and RMSE of 0.5273 - very similar to standalone ARIMA, indicating little added benefit from combining the two. MLR + SARIMA followed a similar pattern, with an RMSE of 0.5281 and R^2 of 0.6015.

MLR + GAM showed the best performance among hybrid models, with an R^2 of 0.6124, an RMSE of 0.5209, and an MAE of 0.3944. This indicates a slight improvement, likely due to the combination of MLR's structure and GAM's ability to model

Received: March 12, 2025. Reviewed: April 12, 2025. Accepted: April 16, 2025. *© 2025 Nuray Dauletkhan and Khaled Mohamad. All rights reserved.*

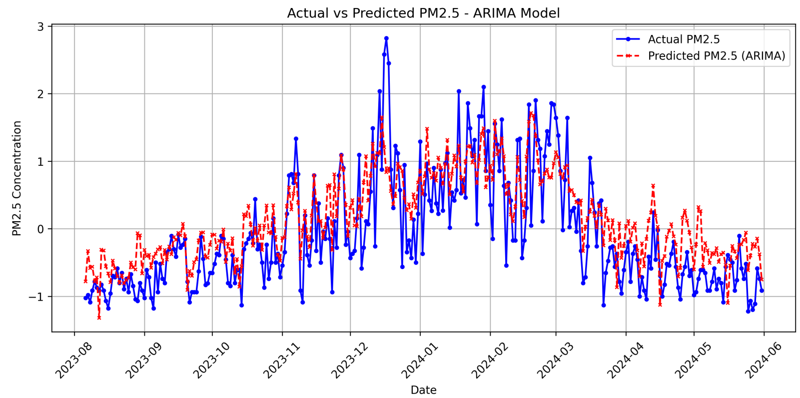


Fig. 4. Forecasted vs. actual $PM_{2.5}$ levels for ARIMA model

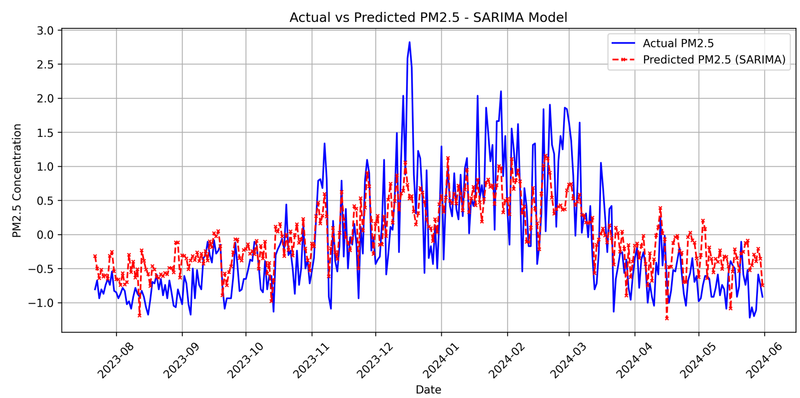


Fig. 5. Forecasted vs. actual $PM_{2.5}$ levels for SARIMA model

non-linear effects. Figure 7 illustrates the predictive performance of the hybrid MLR + GAM model.

On the other hand, GAM + ARIMA performed the worst among hybrid models, with an R^2 of 0.5240, RMSE of 0.5772, and MAE of 0.4052. This suggests that combining two flexible but complex models does not necessarily lead to better results and may introduce redundancy or overfitting.

Although hybrid models were expected to outperform individual models, improvements were minimal. One reason is that models like GAM already captured much of the variation in the data, leaving little structure for ARIMA or SARIMA to model further. Furthermore, long periods of missing values were imputed using MICE. It may have smoothed out key time series patterns. The relatively small dataset (1,558 records) may also have limited the effectiveness of more complex, multistage models.

3) Key Findings and Implications: Overall, statistical models offered a solid baseline for $PM_{2.5}$ forecasting in Almaty. MLR performed best in explaining variance, while SARIMA achieved the lowest RMSE, highlighting its strength in short-term and seasonal forecasting. Among hybrid models, MLR + GAM was the most promising, and showed a modest gain.

These results suggest that while combining models can add flexibility, it does not guarantee better generalization. The study also highlights the limitations of current approaches and the potential benefit of exploring machine learning or deep learning techniques in future research. Ensemble learning and enhanced feature engineering, especially incorporating real-time traffic, industrial activity, and emission data, could significantly improve predictive performance.

Received: March 12, 2025. Reviewed: April 12, 2025. Accepted: April 16, 2025. *© 2025 Nuray Dauletkhan and Khaled Mohamad. All rights reserved.*

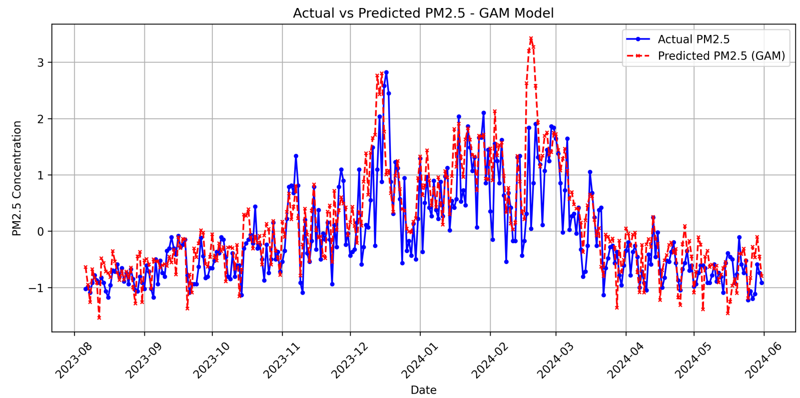


Fig. 6. Forecasted vs. actual $PM_{2.5}$ levels for GAM model

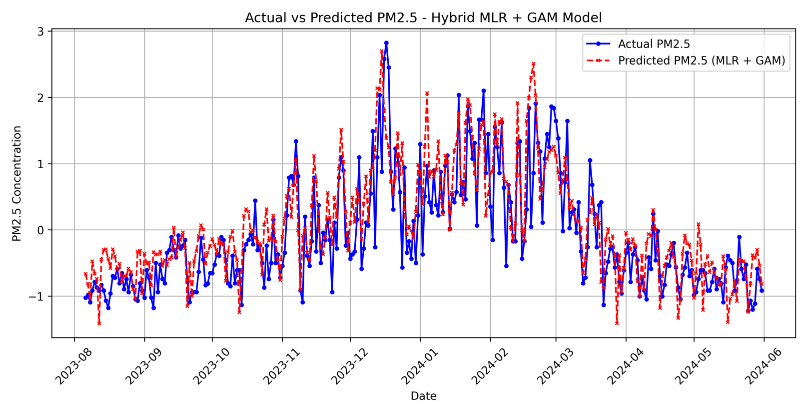


Fig. 7. Forecasted vs. actual $PM_{2.5}$ levels for the hybrid MLR GAM model

4) Practical Relevance: These findings offer practical value for city officials and environmental agencies in Almaty. By using these models to forecast $PM_{2.5}$ levels, they can take earlier action, such as issuing health warnings, managing traffic, or limiting industrial operations on high pollution days. Since the models rely on data that is already being collected, they offer a cost-effective tool for real-time air quality management, ultimately helping protect public health, especially for vulnerable groups.

V. CONCLUSION AND FUTURE WORK

Among the compared models, Multiple Linear Regression (MLR) was the most competent in describing the relationship between weather conditions and $PM_{2.5}$ concentrations. Having an R^2 of 0.7160, it successfully captured pollution trends with reasonable accuracy. Despite the assumption of linearity in MLR, its predictions were robust, as shown by the Mean Absolute Error (MAE) of 0.3831 and Root Mean Squared Error (RMSE) of 0.5268.

For seasonal trends, the SARIMA model provided the lowest RMSE (0.2719), it was the best performing for short-term prediction. Its R^2 value (0.6058) was, however, less than that of MLR, suggesting that while SARIMA is effective at modeling seasonal variation, it may not capture long-term pollution trends.

Received: March 12, 2025. Reviewed: April 12, 2025. Accepted: April 16, 2025. *© 2025 Nuray Dauletkhan and Khaled Mohamad. All rights reserved.*

Among the hybrid models, the best performing one was MLR combined with the Generalized Additive Model (MLR + GAM). With an R^2 of 0.6124 and an RMSE of 0.5209, this model demonstrated that combining MLR's structured approach with GAM's ability to capture nonlinear trends led to moderate improvements over traditional statistical methods.

In general, MLR was the best model for explaining $PM_{2.5}$ variations, and SARIMA was the most accurate for short-term forecasting. These findings can help in the development of more efficient air pollution control strategies for Almaty and other cities.

This study is designed as the first phase of a broader investigation of $PM_{2.5}$ forecasting in Almaty. In this phase here, the key focus is right on statistical as well as hybrid statistical models in order to establish a solid baseline to understand the full structure and behavior of local air quality data. For future research work, by integrating machine learning (ML) along with deep learning (DL) models – like Random Forests, Gradient Boosting, and LSTM networks – may very well further improve overall forecasting accuracy, particularly in the handling of nonlinear patterns, interactions, and much longer time dependencies. A comparative analysis between statistical, machine learning (ML), and deep learning (DL) approaches on the same dataset can provide a broader understanding of their respective strengths and help identify the most effective tools to support air quality management in Almaty and similar cities. Future research can focus on developing more adaptive air quality forecasting systems, as well as truly scalable, real-time solutions that directly inform public health strategies and environmental policy.

REFERENCES

[1] A. Bekbossynova, D. Duvanova, N. Jones, K. Lyden, T. McGinley, and H. Moss, "How attitudes towards air pollution may impact public health: a case study of almaty, kazakhstan," *Journal of Environmental Protection*, vol. 14, no. 07, pp. 583–601, 2023. [Online]. Available: https://doi.org/10.4236/jep.2023.147034

[2] T. B. Ogbuabia, M. Guney, N. Baimatova, I. Ulusoy, and F. Karaca, "Assessing the impact of combined heat and power plants (chpps) in central asia: a case study in almaty for pm2.5 simulations using wrf-aermod and ground level verification," *Atmosphere*, vol. 14, no. 10, p. 1554, 2023. [Online]. Available: https://doi.org/10.3390/atmos14101554

[3] S. Jalali, M. Karbakhsh, M. Momeni, M. Taheri, S. B. Amini, M. Mansourian, and N. Sarrafzadegan, "Long-term exposure to pm2.5 and cardiovascular disease incidence and mortality in an eastern mediterranean country: findings based on a 15-year cohort study," 2021, preprint. [Online]. Available: https://doi.org/10.21203/rs.3.rs-142122/v1

[4] P. Nath, P. Saha, A. I. Middya, and S. Roy, "Long-term time-series pollution forecast using statistical and deep learning methods," *Neural Computing and Applications*, vol. 33, no. 19, pp. 12 551–12 570, 2021. [Online]. Available: https://doi.org/10.1007/s00521-021-05901-2

[5] B. Liu, Y. Jin, X. De-zhi, Y. Wang, and C. Li, "A data calibration method for micro air quality detectors based on a lasso regression and narx neural network combined model," *Scientific Reports*, vol. 11, no. 1, 2021. [Online]. Available: https://doi.org/10.1038/s41598-021-00804-7

[6] X. Qu and Y. Cao, "Empirical analysis of air quality in china based on multiple linear regression analysis," in *Second International Conference on Statistics, Applied Mathematics, and Computing Science (CSAMCS 2022)*, 2023. [Online]. Available: https://doi.org/10.1117/12.2671882

[7] M. S. Ramadan, A. Abuelgasim, and N. A. Hosani, "Advancing air quality forecasting in abu dhabi, uae using time series models," *Frontiers in Environmental Science*, vol. 12, 2024. [Online]. Available: https://doi.org/10.3389/fenvs.2024.1393878

[8] H. Bouzghiba, A. Mendyl, K. Khomsi, and G. Géczi, "Short-term predictions of pm10 and no2 concentrations in urban environments based on arima search grid modeling," *CLEAN – Soil, Air, Water*, vol. 52, no. 6, 2024. [Online]. Available: https://doi.org/10.1002/clen.202300395

[9] L. Zhang, X. Tian, Y. Zhao, L. Liu, Z. Li, L. Tao, and Y. Luo, "Application of nonlinear land use regression models for ambient air pollutants and air quality index," *Atmospheric Pollution Research*, vol. 12, no. 10, p. 101186, 2021. [Online]. Available: https://doi.org/10.1016/j.apr.2021.101186

[10] J. D. Kurniawan, H. A. Parhusip, and S. Trihandaru, "Predictive performance evaluation of arima and hybrid arima-lstm models for particulate matter concentration," *Jurnal Online Informatika*, vol. 9, no. 2, pp. 259–268, 2024. [Online]. Available: https://doi.org/10.15575/join.v9i2.1318

[11] Džaferović and Karauzović-Hadžiabdić, "Air quality prediction using machine learning methods: A case study of bjelave neighborhood, sarajevo, bih," in *Proceedings of the International Conference*. Springer, 2020. [Online]. Available: https://doi.org/10.1007/978-3-030-54765-3_29

[12] M. D. Yazdi *et al.*, "Predicting fine particulate matter (pm2.5) in the greater london area: An ensemble approach using machine learning methods," *Remote Sensing*, vol. 12, no. 6, 2020. [Online]. Available: https://doi.org/10.3390/rs12060914

Received: March 12, 2025. Reviewed: April 12, 2025. Accepted: April 16, 2025. *© 2025 Nuray Dauletkhan and Khaled Mohamad. All rights reserved.*

[13] Badrakh and Choimaa, “Air quality predictions of ulaanbaatar using machine learning approach,” in *Proceedings of the 24th International Conference on Computing in High Energy and Nuclear Physics (CHEP 2019)*, 2021. [Online]. Available: https://doi.org/10.22323/1.378.0012

[14] A. Dairi *et al.*, “Integrated multiple directed attention-based deep learning for improved air pollution forecasting,” *IEEE Transactions on Instrumentation and Measurement*, 2021. [Online]. Available: https://doi.org/10.1109/tim.2021.3091511

[15] R. Idroes *et al.*, “Urban air quality classification using machine learning approach to enhance environmental monitoring,” *Lhokseumawe Journal of Engineering and Science (LJES)*, vol. 1, no. 2, 2023. [Online]. Available: https://doi.org/10.60084/ljes.v1i2.99

[16] T. V. Vu *et al.*, “Assessing the impact of clean air action on air quality trends in beijing using a machine learning technique,” *Atmospheric Chemistry and Physics*, vol. 19, pp. 11 303–11 314, 2019. [Online]. Available: https://doi.org/10.5194/acp-19-11303-2019

[17] Y. Shi, Z. Du, J. Zhang, F. Han, F. Chen, D. Wang, and S. Sui, “Construction and evaluation of hourly average indoor pm2.5 concentration prediction models based on multiple types of places,” *Frontiers in Public Health*, vol. 11, 2023. [Online]. Available: https://doi.org/10.3389/fpubh.2023.1213453

[18] A. Agibayeva, R. Khalikhan, M. Güney, F. Karaca, A. Torezhan, and E. Avcu, “An air quality modeling and disability-adjusted life years (daly) risk assessment case study: Comparing statistical and machine learning approaches for pm2.5 forecasting,” *Sustainability*, vol. 14, no. 24, p. 16641, 2022. [Online]. Available: https://doi.org/10.3390/su142416641

[19] S. Koleva, S. Gocheva-Ilieva, and H. Kulina, “Stochastic modelling of daily air pollution in burgas, bulgaria,” *Journal of Physics: Conference Series*, vol. 2675, no. 1, p. 012003, 2023. [Online]. Available: https://doi.org/10.1088/1742-6596/2675/1/012003

[20] N. F. Muzakki, A. Z. Putri, S. Maruli, and F. Kartiasih, “Forecasting the air quality index by utilizing several meteorological factors using the arimax method (case study: Central jakarta city),” *Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi)*, vol. 8, no. 3, pp. 569–586, 2024. [Online]. Available: https://doi.org/10.35870/jtik.v8i3.2012

[21] T. Bunnag, “Forecasting pm10 caused by bangkok’s leading greenhouse gas emission using the sarima and sarima-garch model,” *International Journal of Energy Economics and Policy*, vol. 14, no. 1, pp. 418–426, 2024. [Online]. Available: https://doi.org/10.32479/ijeep.15275

[22] G. Reddy, M. Manjunath, R. Patil, and P. Kulkarni, “Predicting potential evapotranspiration for kalaburagi district using a seasonal arima model,” *International Journal of Environment and Climate Change*, vol. 13, no. 11, pp. 2073–2082, 2023. [Online]. Available: https://doi.org/10.9734/ijecc/2023/v13i113367

[23] M. S. Hammer, A. v. Donkelaar, C. Li, A. Lyapustin, A. M. Sayer, N. C. Hsu, and R. V. Martin, “Global estimates and long-term trends of fine particulate matter concentrations (1998–2018),” *Environmental Science & Technology*, vol. 54, no. 13, pp. 7879–7890, 2020. [Online]. Available: https://doi.org/10.1021/acs.est.0c01764

[24] F. Si, C. Zhou, Y. Yang, and L. Huang, “Study of the relationship between occurrence of kawasaki disease and air pollution in chengdu by parametric and semi-parametric models,” *Environmental Science and Pollution Research*, vol. 30, no. 55, pp. 117 706–117 714, 2023. [Online]. Available: https://doi.org/10.1007/s11356-023-30533-5

Received: March 12, 2025. Reviewed: April 12, 2025. Accepted: April 16, 2025. *© 2025 Nuray Dauletkhan and Khaled Mohamad. All rights reserved.*

Article

Optimizing QA Systems: Evaluating Row-Based and Traditional Chunking in Structured-Data-Aware Retrieval-Augmented Generation for University Virtual Assistants

Maksat Maratov [1*] and Selchuk Cankurt [2]

[1]Department of Computer Science, SDU University, Almaty, Kazakhstan
[2]Department of Computer Science, SDU University, Almaty, Kazakhstan

DOI: 10.47344/7we5dg32

Abstract

This paper presents the development of a question-answering system that can assist university students with academic and administrative questions. We present a new approach that examines various chunking approaches to the Retrieval-Augmented Generation process. Although RAG is typically used with standard chunking methods, this paper presents row-based chunking, tailored to structured question-answer datasets, in order to enhance context retrieval for large language models. To establish its effectiveness, we conducted a human evaluation to compare the outputs it generated with those generated using standard and row-based chunking. The individuals who tested our system were both students and educators at the university. We concluded that row-based chunking gives more coherent and relevant contextual data than standard ways of chunking when applied to structured data sets. This work highlights the potential of using chunking methods to improve RAG-based systems for domain-specific applications, paving the way towards more accurate and context-sensitive AI-based aid in educational settings.

Keywords: Q&A system, Virtual Assistant, ChatBot, RAG, NLP, Chunking strategy, LLM powered Chatbots

*Corresponding author: maksat.maratov@sdu.edu.kz
Email: maksat.maratov@sdu.edu.kz ORCID: 0009-0004-8511-5014
Email: selcuk.cankurt@sdu.edu.kz ORCID: 0000-0003-0581-1913

Received: April 22, 2025. Reviewed: May 30, 2025. Accepted: June 3, 2025. © 2025 Maksat Maratov and Selchuk Cankurt. All rights reserved.

I. Introduction

In businesses and other sizable organizations, efficient question-answering (QA) systems and support services are essential to manage large amounts of information and user interactions. While most organizations maintain specialized support centers, the quality and usefulness of such services diminish with an increase in the volume of inquiries. This issue is particularly acute at universities, where official support centers do address academic and administrative inquiries but may not be familiar with student-generated topics such as campus events, extracurricular activities, and student life dynamics. Therefore, students rely on peer-shared knowledge, which is decentralized and difficult to formalize within traditional support systems.

In recent years, Large Language Models (LLMs) have become deeply integrated into various aspects of daily life [1], with users increasingly preferring text-generation tools over conventional search engines for information retrieval. However, LLMs struggle with domain-specific, private, or real-time data, leading to hallucinations and misinformation in cases where such knowledge is not explicitly encoded in their training corpus [2]. To address this limitation, the Retrieval-Augmented Generation (RAG) framework was introduced. RAG enhances LLMs by integrating an external knowledge retrieval mechanism, typically consisting of three key components: indexing, retrieval, and generation [3].

RAG has been widely adopted across large enterprises and knowledge-driven organizations to improve factual accuracy and provide real-time, dynamic responses from large text corpora. Various RAG implementations exist, differing primarily in their approaches to indexing, retrieval, and response generation MMed-RAG [4] HiTA [5] FinTMMBench [6] and OmniEval [7]. Some systems employ different indexing techniques, such as keyword-based, dense vector-based, or hybrid search methods, while others vary in their retrieval strategies or choice of language model (LLM) for generation.

However, most existing RAG methods are optimized for unstructured text, relying on general-purpose chunking strategies such as RecursiveTextSplitter, fixed-length chunking, and semantic-based chunking. These chunking approaches divide text into predefined sizes or semantically coherent segments, which may work well for free-form documents but are poorly suited for structured data such as spreadsheets, relational databases, or question-answer (Q&A) tables. In structured datasets, preserving the integrity of data relationships is crucial, as conventional chunking methods risk fragmenting semantically dependent information, leading to retrieval mismatches and inaccurate responses.

In contrast, this work introduces a structured-data-aware RAG approach that optimally handles tabular data by treating each row as a single chunk, rather than using arbitrary chunk size constraints. This approach ensures:

- Preservation of data integrity – Each Q&A pair remains intact, avoiding fragmented information retrieval.
- Efficient retrieval and alignments of the embedding – By using whole rows as a chunk the similarity search operations become more precise.
- Reduction in unnecessary processing overhead – Eliminates the need for reconstructing structured data from fragmented chunks.

Our approach is particularly beneficial for scenarios involving structured Q&A datasets, where maintaining the original structure of the data is crucial for accurate retrieval and answer generation from the LLMs.

II. Review of Related Works

It is essential to learn about the history of question-answering systems prior to addressing chunking and information retrieval strategies. Computer question answering benchmarks were defined at the Text Retrieval Conference (TREC) in 1999, one of the initial benchmarks for QA as a field. No matter which subject is envisioned, the purpose was to return short answers to factoid and enumerative questions [8]. In an effort to make the output more precise, conventional QA systems employ structured information retrieval and categorization [9]. Our study expands on this by including structured data retrieval for university-based Q&A into a RAG framework.

The utilization of structured knowledge sources to increase answer accuracy has been the main focus of recent developments in QA systems. Derici and associates. [10] proposed HazırCevap, a closed-domain QA framework that retrieves answers from reliable educational resources while also utilizing multilingual support through translation. Unlike open-domain QA systems, HazırCevap specifically caters to students by ensuring accuracy through a curated knowledge base. However, it relies on document summarization rather than dynamic retrieval-augmented generation

Received: April 22, 2025. Reviewed: May 30, 2025. Accepted: June 3, 2025. © 2025 Maksat Maratov and Selchuk Cankurt. All rights reserved.

(RAG), which limits its ability to adapt to diverse and evolving queries. Our work extends this by leveraging RAG to retrieve and generate answers in real-time, ensuring both accuracy and contextual relevance.

Retrieval-Augmented Generation (RAG) integrates parametric (pre-trained LLM) and non-parametric (retrieved external data) memory to improve knowledge-intensive tasks [11]. The retrieval module locates relevant information, while the generation module conditions on retrieved context to generate more factual responses. Such patterns of work are observed in most RAG methods. In more detail, there are three main components: indexing, retrieval, and generation. The retrieval module locates relevant information using dense or sparse search, while the generation module integrates this context to produce an accurate response.

Despite the effectiveness of RAG in enhancing factual consistency, not all RAG models are well-suited for structured data retrieval. Many existing implementations are optimized for unstructured text, where chunking strategies such as RecursiveTextSplitter [12] or fixed-length segmentation are commonly employed. While these methods work well for free-form documents, they introduce fragmentation issues when applied to structured datasets like university Q&A tables. For instance, RAG implementations that rely on naive text chunking may separate a question from its corresponding answer, leading to retrieval mismatches and incoherent responses. Furthermore, models such as Hybrid-RAG [13] and ActiveRAG [14] attempt to improve retrieval by incorporating iterative refinement, but they remain inefficient when handling structured data fields due to their reliance on unconstrained semantic search [15].

BM25, a probabilistic information retrieval model, ranks documents based on query terms but may not effectively handle the nuances of structured data [16]. In contrast, our approach preserves the Q&A pair data by treating each row as a single retrieval unit, ensuring accurate and contextually consistent responses.

While various RAG implementations focus on enhancing accuracy, retrieval mechanisms, and source attribution, they do not consider structured Q&A pair data. Traditional RAG frameworks primarily process unstructured documents, making them unsuitable for applications where preserving data relationships—such as university Q&A datasets—is critical. Some studies focus on reducing contradictions in retrieved knowledge and self-reflecting on results, improving reliability like SelfRAG [17], ActiveRAG [14], and InstructRAG [18].

However, these approaches do not address the challenges of structured data retrieval, particularly in handling Q&A pair formats. While some works focus on structured or semi-structured data, they primarily target entity-based retrieval, tabular knowledge representation THoRR [19], or knowledge graphs FastRAG [20], rather than optimizing chunking strategies for structured text. Existing methods fail to consider how row-wise chunking can preserve data integrity in structured datasets, such as university Q&A tables, where each row represents a complete and independent knowledge unit.

III. Methods

In this study, we address the challenge of structured data retrieval in Retrieval-Augmented Generation (RAG) systems by leveraging a university-specific Question-Answer (QA) dataset. Unlike traditional RAG models that process unstructured text, our approach preserves the integrity of structured QA pairs, ensuring accurate and contextually relevant responses.

A. Dataset

The dataset consists of approximately 20,000 QA pairs collected from university students. It covers a wide range of university-related topics, including academic inquiries, student life, event details, club activities, and problem-solving scenarios such as lost ID cards or course registration procedures. Since the dataset is user-generated, it includes variations in phrasing, with some questions appearing in multiple interpretations or with additional details. These variations enhance the model's ability to retrieve contextually appropriate responses.

The dataset includes three languages: Kazakh, Russian, and English. Before training, the data underwent preprocessing, including the removal of stopwords, conversion to lowercase, and other standard NLP cleaning techniques to ensure consistency. Duplicate entries were filtered, while semantically similar but non-identical questions were retained to improve retrieval diversity.

Received: April 22, 2025. Reviewed: May 30, 2025. Accepted: June 3, 2025. © 2025 Maksat Maratov and Selchuk Cankurt. All rights reserved.

B. Structured Retrieval and Chunking Strategy

Our RAG implementation deviates from traditional chunking methods, such as character-based, recursive, or semantic splitting. Instead, given the structured nature of our dataset—comprising QA pairs—we treat each row of data as an independent chunk. This ensures that the full context of each question-answer pair remains intact, preventing the fragmentation issues commonly observed in unstructured chunking approaches. By maintaining complete QA pairs as single retrieval units, we preserve the semantic integrity of responses, which positively impacts retrieval accuracy.

After chunking, we proceed with embedding the data for vector-based retrieval. Since our dataset contains content in three languages (Kazakh, Russian, and English), we employ a multilingual embedding model, intfloat/multilingual-e5-large, which is widely adopted for cross-lingual tasks due to its strong performance across a broad range of languages. This model was chosen for its balance between quality and efficiency, and because it has demonstrated robust multilingual retrieval capabilities in both academic benchmarks and practical applications. Although we did not conduct an independent embedding evaluation, we employ it because it is widely adopted for cross-lingual tasks and performs strongly across languages.

For indexing, we utilize VectorStoreIndex, a widely used vector database approach that allows efficient similarity-based retrieval. Each QA pair is stored as an embedding, enabling rapid lookup of semantically similar chunks during the retrieval process.

During retrieval, an input question is first embedded using the same multilingual embedding model. The system then computes the cosine similarity between the query embedding and all indexed QA pair embeddings, selecting the top-K most relevant results. These retrieved QA pairs serve as context for the final answer generation, ensuring that the response is based on the most semantically similar knowledge available.

1) Response Generation:

For the response generation phase, we integrate OpenAI's GPT-4o as the language model. To ensure the model behaves as a university virtual assistant, we apply prompt engineering techniques. The prompt includes:

- Zero-shot learning strategies to help the model generalize across diverse university-related queries.
- Background information about the university to provide institution-specific responses.
- Rules and regulations for handling specific student-related scenarios (e.g., lost ID cards, course registration issues).

This carefully designed prompt ensures consistency in the responses and is used uniformly across all evaluated methods to maintain fairness in comparisons.

The primary motivation behind our chunking strategy is to preserve the full context of each QA pair, avoiding the fragmentation issues introduced by traditional chunking methods. Standard approaches such as RecursiveTextSplitter segment documents based on arbitrary character or semantic boundaries, often leading to incomplete or disjointed retrieval results. In contrast, our row-wise chunking ensures that each QA pair remains intact, providing a more semantically meaningful retrieval unit.

Moreover, while semantic chunking techniques attempt to create contextually coherent splits, they often struggle with multilingual datasets due to limitations in cross-lingual sentence embedding models. This challenge can lead to poor retrieval performance when queries and indexed documents exist in different languages. In our study, we will empirically compare the top-K retrieval performance metrics between our structured chunking method and conventional approaches. Specifically, the top-K value will be set to 20. This will allow us to evaluate and demonstrate the effectiveness of our method in enhancing retrieval accuracy and performance, particularly in the context of university Q&A datasets.

C. Comparison with Traditional Chunking

To establish a meaningful comparison, we evaluate our structured row-based chunking method against the conventional traditional chunking approach, which segments text into fixed-sized chunks or employs semantic splitting strategies. Traditional chunking methods, while widely used, often introduce inconsistencies by fragmenting contextually related information, potentially leading to loss of coherence in retrieval tasks.

Since our dataset is inherently structured in a row-based format, a direct comparison requires adapting the traditional chunking method to a relevant representation. For this, we approximate an unstructured document

Received: April 22, 2025. Reviewed: May 30, 2025. Accepted: June 3, 2025. © 2025 Maksat Maratov and Selchuk Cankurt. All rights reserved.

format through the conversion of the dataset to a continuous text-based QA format, approximating how data would typically be retained in unstructured documents. This step takes care that both chunking methodologies are evaluated on the same premises.

By presenting the dataset in this form, we are able to test how well the traditional methodology recovers useful responses and preserves contextual coherence compared to our row-based method. By doing so, we are highlighting the limitations of applying the traditional chunking method to structured data and comparing them with the advantages of a retrieval-nominated chunking method.

IV. Results and Discussion

To evaluate the performance of proposed chunking method against the traditional approach of RAG, a small-scale study was conducted involving 10 participants, comprising both students and teachers at SDU Univeristy. Participants were asked a series of questions related to university life, operations, and logistics, in three languages: English, Kazakh, and Russian. The primary goal was to test whether our chunking strategy—where each chunk is a full Question-Answer (QA) pair—leads to more relevant and precise information retrieval compared to traditional chunking, which segments the text arbitrarily or by fixed-length windows.

A total of 6 multilingual questions were used as test inputs:

1) What is EPT?
2) Потерял ID карту. Что мне делать?
3) СДУ университетінде көлік тұрағы бар ма? Егер бар болса, студенттер көліктерін қоя алады ма?
4) Өзімнің ағылшын деңгейімді қалай көрсем болады?
5) Что если у меня есть пересечения в расписании файналов?
6) Give me step-by-step instructions of how to get the book from the SDU library.

The results presented below highlight only the cases where there was a noticeable difference between the two retrieval methods. In some cases, such as locating specific teacher offices or retrieving department-specific information, both methods failed to retrieve a relevant answer, which indicates a lack of data coverage rather than method inadequacy.

A. Traditional Methodology

User Request 1: What is EPT? [label=chunk1] Q: What is an EPT in SDU? A: English Proficiency Test: SDU offers an English Proficiency Test (EPT) to assess students' ... [56 words]

Q: SDU-да EPT дегеніміз не? [7 words]

Q: Что такое олимпиада SPT? [7 words]

They also give grades for the task that were loaded. Q: What benefits and opportunities does SPT (Profile Testing System) provide for school and college graduates? A: SPT — is a test that is conducted in the specialized subjects of UNT for high school and college students... [90 words]

nursing homes, orphanages 3)marvel - Organizer 4)handmade - Craft Direction 5)cooking-cooking a variety of meals together Q: What kind of race is SPT? A: SPT is a competition for 11th grade Students... [135 words]

User Request 2: Потерял ID карту. Что мне делать?

Q: Что делать если забыл ID карту? [9 words]

Q: Если я забыл дома свою ID карту, то как могу войти в универ? [16 words]

Q: В каких местах я могу использовать ID-карту? [9 words]

Q: Что будет, если студент потерял свою карту ID? Будет ли допуск на экзамены? [16 words]

Q: Если вы потеряли ID-карту, вы можете пойти в центр обслуживания студентов, чтобы создать ее... [132 words]

User Request 3: Сду университетінде көлік тұрағы барма? Егер бар болса студенттер көліктерін қоя аладыма?

Q: Салыныпты әлі білмесе, СДУ-да кеңселерді қалай табуға болады? A: Егер студент кабинетті таба алмаса, ол [17 words]

Q: Студенттер үшін көлік қандай нұсқалар бар? [9 words]

Q: Ата-анам маған көлік сатып алды. Мен университеттің тұрағына қоя аламын ба? A: SDU аумағындағы автотұрақ Қызметкерлер мен қонақтарға арналған... [31 words]

Received: April 22, 2025. Reviewed: May 30, 2025. Accepted: June 3, 2025. © 2025 Maksat Maratov and Selchuk Cankurt. All rights reserved.

Q: Шегуге бола ма? A: Университет университет аумағында темекі шегуге немесе электронды құрылғыларды [12 words]

Q: Ғимаратында не орналасқан? A: SDU Life ғимаратында студенттерге [6 words]

B. Row-Based Chunking Method

In contrast, the row-based chunking method significantly enhances retrieval by preserving complete question-answer pairs:

User Request 1: What is EPT?

[label=chunk1] SDU offers an English Proficiency Test (EPT) to assess students' English language skills. The test is typically required for admission to English-medium programs or for students seeking exemptions from English language courses. It evaluates ... the test if needed. Additionally, SDU may offer English language clubs or resources to help students improve their language skills and prepare for the EPT. [76 words] '

What is an English proficiency test? It's an exam to take for an exchange program. [9 words]

SPT is a competition for 11th grade students. Through the competition, you can win an internal grant. Even on the day you don't win, you will be given a discount on paid education. At first, the competition will be based on math literacy, and then you will pass it according to your professional subject. [44 words]

When students apply to the university, they need to take an English test to see how well they know the language. There are two parts to the test: 1) Grammar test: consists of 50 questions; 2) A speaking part ... If a student's English level is below that, they'll have to pay for courses to improve it. Currently, each level costs 102,000 KZT. [91 words]

Typically, 1 ECTS is equal to 25-30 hours. This is an indicator of the value of the course. [20 words]

User Request 2: Потерял ID карту. Что мне делать?

Что делать если забыл ID карту? [9 words]

Сначала нужно оповестить эдвайзера и сделать запрос на восстановление ID карты. [16 words]

Студентам необходимо посетить сервисный центр. [7 words]

Вам следует обратиться эдвайзеру, через некоторое время вам выдадут новую. [15 words]

ID-карта, центр обслуживания студентов, QR-код, 2000 тенге. [8 words]

User Request 3: Сду университетінде көлік тұрағы барма? Егер бар болса студенттер көліктерін қоя аладыма?

Университетте көлік тұрағы бар, университеттің ауласында орналасқан. Өкінішке орай студенттер көліктерін қоя алмайды. Только преподавателям разрешается парковать свои автомобили. [33 words]

Иә, SDU университетінің ауласында көлік тұрағы бар. Бірақ та студенттер ол көлік тұрағын қолдана алмайды. Тұрақ мұғалімдерге арналған. [22 words]

SDU аумағындағы автотұрақ қызметкерлер мен қонақтарға арналған. Студенттер оған көліктерін қалдыра алмайды. [18 words]

Өкінішке орай, егер сіз университет қызметкері болмасаңыз, университет аумақтарына көлік қоя алмайсыз. Бірақ сіз көлігіңізді университеттің жанында арнайы тұрақ орындарына қоя аласыз. [27 words]

Иә, биыл тегін. [5 words]

As shown above, the row-based chunking method proves to be more effective in retrieving relevant information aligned with user queries. By maintaining the integrity of each question-answer pair, it allows the retriever to surface more meaningful and contextually complete results.

C. Generation

The next step in the pipeline is generation, where the impact of the input format becomes even more evident. Below, we compare the responses generated by the language model using the two different retrieval inputs: Traditional Chunking [T] and Row-Based Chunking [R].

Question 1

Q: What is EPT? Response [T]: EPT is the English Proficiency Test offered to assess students' English language skills. It is typically required for admission to English-medium programs or for students seeking exemptions from English language courses. The test evaluates proficiency in listening, reading, writing, and speaking. Response [R]: The English Proficiency Test (EPT) at SDU is designed to assess students' English language skills. It is typically

Received: April 22, 2025. Reviewed: May 30, 2025. Accepted: June 3, 2025. © 2025 Maksat Maratov and Selchuk Cankurt. All rights reserved.

required for admission to English-medium programs or for students seeking exemptions from English language courses. The test evaluates proficiency in listening, reading, writing, and speaking. Students can register for the test through the university's website or designated testing centers. The format may include multiple-choice questions, writing essays, and oral interviews. Results are usually available within a specified timeframe, and students may have the opportunity to retake the test if needed. Additionally, SDU may offer English language clubs or resources to help students improve their language skills and prepare for the EPT. The test consists of two parts: a grammar test with 50 questions and a speaking part where an examiner asks questions on random topics to evaluate the student's English level. After the test, students receive a certificate indicating their English level. Most study programs require at least a B1 level. If a student's English level is below that, they will need to pay for courses to improve it, with each level costing 102,000 KZT.
Question 2

Q: Потерял ID карту. Что мне делать? Response [T]: Если вы потеряли ID карту, вам следует обратиться в Центр обслуживания студентов, чтобы запросить изготовление новой карты. Это будет стоить 2000 тенге, и процесс может занять некоторое время, так как карты изготавливаются партиями. Response [R]: Если вы потеряли ID карту в SDU, выполните следующие шаги: 1. Свяжитесь с консультантом (Advising Desk) и сообщите о потере карты. 2. Оплатите штраф за утерю карты, который составляет 2000 тенге. Оплату можно произвести через Kaspi QR или Halyk Bank. 3. Заполните форму через QR-код. 4. Когда карта будет готова, вам сообщат об этом по электронной почте. 5. Также вы можете обратиться в Студенческий сервисный центр (Student Service Center) для получения новой карты. Учтите, что карты изготавливаются партиями, и процесс может занять до двух недель.
Question 3: Сду университетінде көлік тұрағы барма?

Q: Сду университетінде көлік тұрағы барма? Егер бар болса студенттер көліктерін қоя аладыма? Response [T]: СДУ аумағындағы автотұрақ қызметкерлер мен қонақтарға арналған, сондықтан студенттер көліктерін ол жерге қоя алмайды. Алайда, СДУ аумағына жақын жерде көптеген тегін тұрақ орындары бар. Response [R]: Иә, SDU университетінде көлік тұрағы бар, ол университеттің ауласында орналасқан. Алайда, өкінішке орай, студенттерге бұл көлік тұрағын қолдануға рұқсат жоқ. Тұрақ тек мұғалімдер мен қызметкерлерге арналған. Студенттер көліктерін университеттің жанында арнайы тұрақ орындарына қоя алады. Бұрын студенттерге SDU ішінде тұрақ қоюға рұқсат етілген, бірақ орын тапшылығына байланысты уақытша тыйым салынды. Жақында жаңа блок ашылғанда, тұрақ қолжетімді болуы мүмкін.
Question 4: Өзімнің ағылшын деңгейімді қалай көрсем болады?

Q: Өзімнің ағылшын деңгейімді қалай көрсем болады? Response [T]: Өзіңіздің ағылшын деңгейіңізді көру үшін oldmy.sdu.edu.kz парақшасына кіріп, негізгі беттегі негізгі ақпарат бөлімінен ағылшын деңгейі туралы ақпаратты таба аласыз. Сіздің ағылшын деңгейіңіз сол бетте көрсетулі болады. Response [R]: Өз ағылшын деңгейіңізді көру үшін oldmy.sdu.edu.kz парақшасына кіріп, негізгі бетке өту арқылы негізгі ақпараттың ішінен ағылшын деңгейі туралы ақпаратты көре аласыз. Сіздің ағылшын деңгейіңіз сол бетте көрсетулі.
Question 5: Что если у меня есть пересечения в расписании файналов?

Q: Что если у меня есть пересечения в расписании файналов? Response [T]: Вам следует обратиться к вашему эдвайзеру, который может перенести один из экзаменов на другой день, чтобы устранить конфликт в расписании. Response [R]: Если у вас есть пересечения в расписании файналов, вам следует предпринять следующие шаги: 1. Обратитесь к вашему эдвайзеру, чтобы обсудить проблему. 2. Вместе с эдвайзером посетите деканат и напишите заявление о пересечении времени экзаменов. 3. Деканат рассмотрит ваше заявление и обсудит возможность изменения времени одного из экзаменов. 4. Если студентов с такой проблемой немного, для вас могут выделить отдельное время для сдачи экзамена с участием учителя и прокторов. 5. В некоторых случаях может потребоваться заполнение специальной формы и предоставление доказательств пересечения времени экзаменов. 6. Также можно обратиться в MDE центр в университете для назначения другого дня сдачи экзамена или сдачи с другой группой, но это нужно сделать заранее.
Question 6: Give me step by step instructions of how to get the book from the SDU library.

Q: Give me step by step instructions of how to get the book from the SDU library. Response [T]: To get a book from the SDU library, follow these steps: 1. Visit the SDU library website. 2. Use the search function to find the book you want by entering the book's name and author. 3. Check the availability status of the book. If it's available, note the location details such as the floor and key letters. 4. Go to the library with your ID card. 5. Locate the book using

Received: April 22, 2025. Reviewed: May 30, 2025. Accepted: June 3, 2025. © 2025 Maksat Maratov and Selchuk Cankurt. All rights reserved.

the information you noted. 6. Use your ID card to check out the book at the library's checkout desk. Response [R]: To get a book from the SDU library, follow these step-by-step instructions: 1. Visit the SDU Library: Go to the SDU library, which is a 3-story building located on the university campus. 2. Find the Book: You can either search for the book online on the SDU library website or physically browse the library shelves. If searching online, enter the book's name and author in the search bar to check its availability and location within the library. 3. Locate the Book: If the book is available, note the floor and section where it is located. Use the key letters provided to find the exact shelf. 4. Borrow the Book: Once you have the book, go to the librarian or a consultant at the library. Present your student ID card to the librarian. This is necessary to borrow the book. 5. Discuss the Borrowing Period: Talk to the librarian or consultant about how long you wish to borrow the book. Typically, books can be borrowed for a period of 3 weeks. 6. Take the Book Home: After the librarian processes your request, you can take the book home. Remember, you can borrow up to five books at a time. 7. Return the Book: Ensure you return the book by the due date to avoid fines. If you lose the book, you will need to pay a fine or replace it.

D. Evaluation

Apart from retrieval and generation measures, a human test was conducted to study the quality and usability of the assistant's responses. This test focused on the way users felt about the accuracy, relevance, and usefulness of the answers they received.

1) Participants: The evaluation involved approximately 50 participants, comprising both students and staff members from various departments within the university. It is important to note that this survey is distinct from an earlier preliminary survey, which involved 10 participants. The initial survey was conducted to collect user-asked questions and evaluate the quality of LLM-provided answers based on a chunking strategy approach. In contrast, the current survey focuses on a comparative evaluation of two different approaches—traditional chunking and row-based chunking—using a structured set of evaluation criteria.

2) Manual Evaluation: A manual evaluation was performed to assess the virtual assistant's capabilities.

TABLE I
Structured Evaluation Approach for AI-Generated Responses

No.	Evaluation Question	Grading Method
1	Did the prediction contain any hallucinations?	Binary response: Yes or No
2	Assess the relevance of the response to the question.	Rating scale: 1 to 5
3	Evaluate the content size and structure of the response.	Rating scale: 1 to 5
4	Did you identify any logical inconsistencies in the response?	Binary response: Yes or No
5	What is your overall evaluation of the responses?	Rating scale: 1 to 5

Table I. presents a structured approach to the quality assessment of AI responses. It contains both binary (Yes/No) tests and scaled ratings (1 to 5) to ensure a complete analysis.

- Binary questions help identify significant issues such as hallucinations (false or misleading information) and fallacies.
- Rating questions allow for a finer-grained assessment of aspects like relevance, content structure, and overall quality.

These evaluation criteria can be used to score a single answer or to compare several answers from an AI system. When comparing, the more accurate, coherent, and complete answer is scored higher. This systematic process makes AI-generated answers factually correct, logically sound, and well-organized, thus making them credible and useful sources of information for users.

Both virtual assistants' (R and T) performance on five criteria in evaluation is plotted above. The results were divided into two parts depending on their evaluation method: Binary and Scaled evaluations. The Row-Based Chunking RAG (R) does better than the Traditional Chunking RAG (T) on the most important ones.

Figure 1 presents the binary evaluation results.

Received: April 22, 2025. Reviewed: May 30, 2025. Accepted: June 3, 2025. © 2025 Maksat Maratov and Selchuk Cankurt. All rights reserved.

- Avoidance of hallucination (Q1): R performed better due to improved chunking, which positively impacted the generation phase.
- Logical consistency (Q4): Both approaches maintained strong logical consistency in their responses.

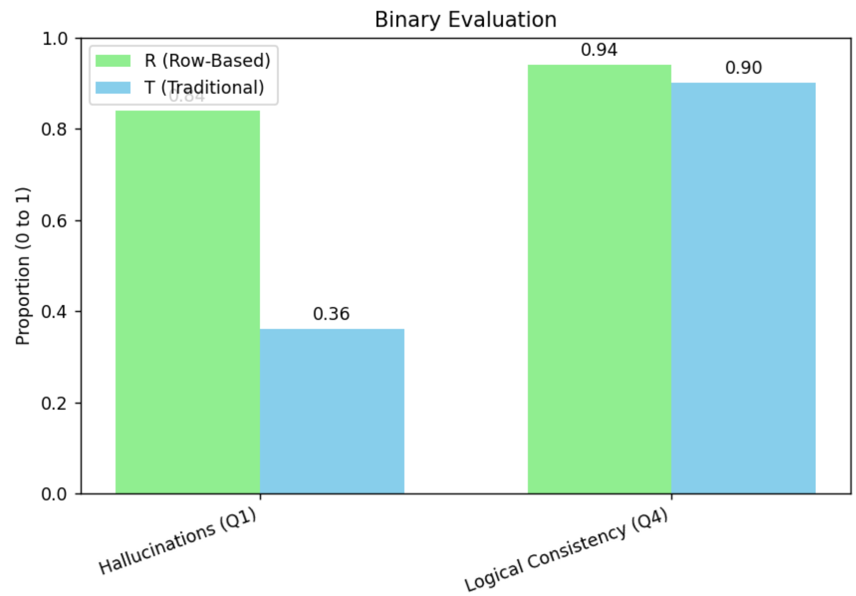


Fig. 1. Binary evaluation results for virtual assistants R and T across two criteria: hallucination avoidance (Q1) and logical consistency (Q4).

Also the Figure 2 shows the scaled evaluation results.

- Response relevance (Q2): R had a higher relevance score.
- Content size and structure (Q3): R was rated more positively than T.
- Overall evaluation (Q5): R scored higher than T, indicating better overall quality of response.

Received: April 22, 2025. Reviewed: May 30, 2025. Accepted: June 3, 2025. © 2025 Maksat Maratov and Selchuk Cankurt. All rights reserved.

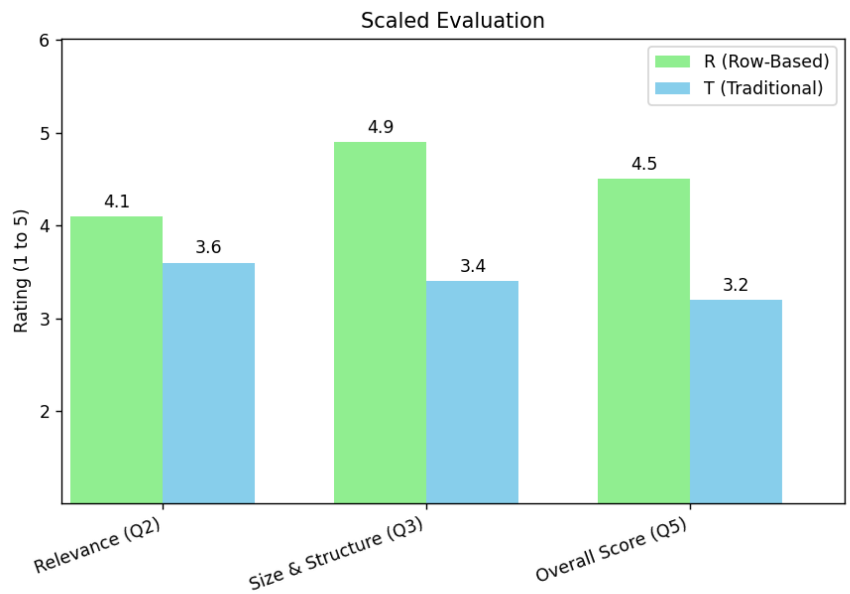


Fig. 2. Scaled evaluation results for virtual assistants R and T across three criteria: response relevance (Q2), content size and structure (Q3), and overall evaluation (Q5).

The no-response rates for each assistant are visualized in Figure 3:

- R had a 9.1% no-response rate.
- T had a 14.3% no-response rate.

This data indicates that R provided responses more frequently than T when evaluated across multiple queries.

E. Evaluation Metrics

We applied our row-based chunking method to a collection of 50 queries collected in a user study. As was mentioned before, every one of the 50 volunteers asked 3 different questions but we took 1 from each and which was tackled by two distinct RAG systems: one with traditional chunking and the other with our row-based chunking method. For each question, we retrieved top-k chunks from both the systems and manually judged their relevance to the question context.

Based on this human-judgment, Precision@k, Recall@k, and F1@k values were computed and compared with returned chunks versus information needed to answer each query. On k=10, our row-based model achieved Precision of 0.58, Recall of 0.67, and F1 score of 0.62, which was considerably higher than the baseline paragraph-based scheme (Precision@10 = 0.41, Recall@10 = 0.44, F1@10 = 0.42). The outcomes indicate that chunk alignment with semantically similar rows within structured data leads to more accurate and comprehensive retrieval, which ultimately improves answer quality in RAG systems.

F. Discussion

These results reinforce the central significance of chunking strategies to retrieval performance, elucidating why distinct methods yield varied outcomes. Standard chunking practices have a propensity to cause incoherences by disunifying contextually coherent information, whereas Row-Based Chunking RAG preserves total context units. This is specifically beneficial for structured information, such as FAQs, where coherence should be preserved so that correct retrieval can be supported.

Received: April 22, 2025. Reviewed: May 30, 2025. Accepted: June 3, 2025. © 2025 Maksat Maratov and Selchuk Cankurt. All rights reserved.

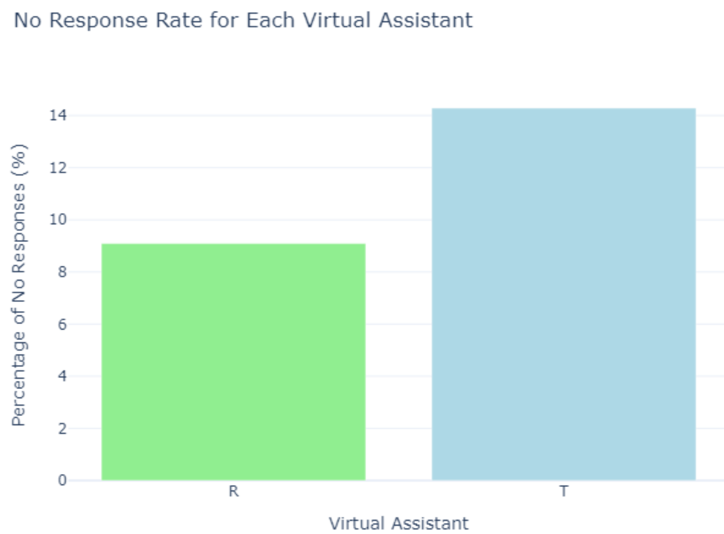


Fig. 3. No-response rates for virtual assistants R and T.

One key thing to note is that if Row-Based Chunking RAG retrieval fails, Traditional Chunking RAG will also fail. But not vice versa—Row-Based Chunking RAG can pass when Traditional Chunking RAG fails. This is due to the fact that traditional chunking techniques sometimes dismember logically related information, and retrieval models struggle more to generate effective responses from them. Such findings suggest that for structured datasets, Row-Based Chunking RAG provides a more solid and context-aware solution.

The evaluation also indicates that Row-Based Chunking RAG always produces more structured and richer answers than Traditional Chunking RAG. This is because of the following reasons:

- Preservation of context: Row-Based Chunking RAG retains full rows as single chunks, ensuring more cohesive retrieval.
- Reduced fragmentation: Traditional Chunking RAG sometimes splits related information into multiple smaller chunks, leading to a loss of coherence in responses.

The no-response rate difference reinforces these findings. While both methods fail in some cases, Row-Based Chunking RAG consistently outperforms Traditional Chunking RAG in retrieving relevant content. This suggests that inefficient chunking in Traditional Chunking RAG contributes to response failures, whereas Row-Based Chunking RAG's structured approach improves retrieval even in challenging cases.

G. Implications

The findings we obtained can be implemented in such structured data QA systems. For RAG systems, choosing an effective chunking strategy is crucial to enhance the response relevance. It can be applied in domains where structured knowledge is key—such as academic assistants, customer support bots, or legal document retrieval—Row-Based Chunking RAG could enhance accuracy and reduce hallucinations.

H. Limitations and Future Considerations

The advantages of our Row-Based Chunking RAG are most noticeable in structured datasets, even if it increases retrieval effectiveness. It's still unclear how well it performs in texts that are more narrative or unstructured.

Received: April 22, 2025. Reviewed: May 30, 2025. Accepted: June 3, 2025. © 2025 Maksat Maratov and Selchuk Cankurt. All rights reserved.

Furthermore, even though Row-Based Chunking RAG performs better than Traditional Chunking RAG in our evaluation, more study is required to determine whether it can scale to bigger and more varied datasets. Hybrid techniques that dynamically modify chunking algorithms according to query context should also be investigated in future work.

In the end, our findings emphasize how crucial careful data architecture is for retrieval-based AI systems, confirming that the quality of generated responses can be greatly impacted by the way information is chunked.

V. Conclusion and Future Work

This paper investigated the impact of chunking methods on retrieval performance on an academic Q&A dataset. Our results show that Row-Based Chunking significantly improves response completeness and coherence over traditional chunking methods. Through retaining the full context units, this approach reduces inconsistency and improves retrieval accuracy, particularly for structured data such as FAQs.

In addition, the study highlights the point that traditional chunking often leads to disconnected responses due to random text splitting. In contrast, Row-Based Chunking is logically consistent, which allows for more effective retrieval of semantic information. The aspect that it possesses a lower no-response rate also bears witness to its application in structured data retrieval.

However, despite these developments, some of the limitations still remain, including the scope of our evaluation and the challenge of handling unstructured or multilingual data. Future research can explore hybrid chunking methods that adapt dynamically to different types of data and retrieval needs. The integration of user feedback and real-world testing will also help in advancing the practical applicability of structured chunking RAG systems.

Our findings enhance knowledge on chunking techniques in information searching, offering a yet more systematic and effective approach to be applied in university virtual assistants and further beyond.

VI. Acknowledgment

The author would like to express sincere gratitude Selcuk Cankurt, Kamila Orynbekova, Ardak Shalkarbay, Yernar Akhmetbek, Aiken Kazin, Inkar Shoganova, Ualikhan Sadyk and Mukhtar Amirkumar for their valuable guidance, unwavering support, and insightful feedback throughout the entire process of completing this thesis.

References

[1] Chen, J., Liu, Z., Huang, X., Wu, C., Liu, Q., Jiang, G., ... & Chen, E. (2024). When large language models meet personalization: Perspectives of challenges and opportunities. World Wide Web, 27(4), 42. https://doi.org/10.1007/s11280-024-01276-1

[2] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 43(2), 1–55. https://doi.org/10.1145/3703155

[3] Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., ... & Li, Q. (2024, August). A survey on RAG meeting LLMs: Towards retrieval-augmented large language models. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 6491–6501). https://doi.org/10.1145/3637528.3671470

[4] Xia, P., Zhu, K., Li, H., Wang, T., Shi, W., Wang, S., ... & Yao, H. (2024). MMED-RAG: Versatile multimodal RAG system for medical vision language models. arXiv preprint arXiv:2410.13085. https://doi.org/10.48550/arXiv.2410.13085

[5] Liu, C., Hoang, L., Stolman, A., & Wu, B. (2024, July). HiTA: A RAG-Based Educational Platform that Centers Educators in the Instructional Loop. In International Conference on Artificial Intelligence in Education (pp. 405–412). Cham: Springer. https://doi.org/10.1007/978-3-031-64299-9_37

[6] Zhu, F., Li, J., Pan, L., Wang, W., Feng, F., Wang, C., ... & Chua, T. S. (2025). FinTMM-Bench: Benchmarking Temporal-Aware Multi-Modal RAG in Finance. arXiv preprint arXiv:2503.05185. https://doi.org/10.48550/arXiv.2503.05185

[7] Wang, S., Tan, J., Dou, Z., & Wen, J. R. (2024). OmniEval: An Omnidirectional and Automatic RAG Evaluation Benchmark in Financial Domain. arXiv preprint arXiv:2412.13018. https://doi.org/10.48550/arXiv.2412.13018

Received: April 22, 2025. Reviewed: May 30, 2025. Accepted: June 3, 2025. © 2025 Maksat Maratov and Selchuk Cankurt. All rights reserved.

[8] Olvera-Lobo, M. D., & Gutiérrez-Artacho, J. (2015). Question answering track evaluation in TREC, CLEF and NTCIR. In New Contributions in Information Systems and Technologies: Volume 1 (pp. 13–22). Springer. https://doi.org/10.1007/978-3-319-16486-1_2

[9] Allam, A. M. N., & Haggag, M. H. (2012). The question answering systems: A survey. International Journal of Research and Reviews in Information Sciences (IJRRIS), 2(3).

[10] Derici, C., Aydin, Y., Yenialaca, Ç., Aydin, N. Y., Kartal, G., Özgür, A., & Güngör, T. (2018). A closed-domain question answering framework using reliable resources to assist students. Natural Language Engineering, 24(5), 725–762. https://doi.org/10.1017/S1351324918000141

[11] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. NeurIPS, 33, 9459–9474.

[12] LangChain. (2023).

[13] Yuan, Y., Liu, C., Yuan, J., Sun, G., Li, S., & Zhang, M. (2024). A Hybrid RAG System with Comprehensive Enhancement on Complex Reasoning. arXiv preprint arXiv:2408.05141. https://doi.org/10.48550/arXiv.2408.05141

[14] Xu, Z., Liu, Z., Liu, Y., Xiong, C., Yan, Y., Wang, S., ... & Yu, G. (2024). ActiveRAG: Revealing the treasures of knowledge via active learning. arXiv preprint arXiv:2402.13547. https://doi.org/10.48550/arXiv.2402.13547

[15] Fan, Y., Yan, Q., Wang, W., Guo, J., Zhang, R., & Cheng, X. (2025). TrustRAG: An Information Assistant with Retrieval Augmented Generation. arXiv preprint arXiv:2502.13719. https://doi.org/10.48550/arXiv.2502.13719

[16] Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. Foundations and Trends® in Information Retrieval, 3(4), 333–389. http://dx.doi.org/10.1561/1500000019

[17] Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2023, October). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In The Twelfth International Conference on Learning Representations.

[18] Wei, Z., Chen, W. L., & Meng, Y. (2024). InstructRAG: Instructing retrieval-augmented generation with explicit denoising. arXiv preprint arXiv:2406.13629. https://doi.org/10.48550/arXiv.2406.13629

[19] Kim, K., Kim, M., Lee, H., Park, S., Han, Y., & Jeon, B. K. (2024). THoRR: Complex Table Retrieval and Refinement for RAG. In IR-RAG 2024 Workshop Proceedings, 3784, 50–55.

[20] Abane, A., Bekri, A., & Battou, A. (2024). FastRAG: Retrieval Augmented Generation for Semi-structured Data. arXiv preprint arXiv:2411.13773. https://doi.org/10.48550/arXiv.2411.13773

Received: April 22, 2025. Reviewed: May 30, 2025. Accepted: June 3, 2025. © 2025 Maksat Maratov and Selchuk Cankurt. All rights reserved.

SECTION II
Infocommunication Technologies

This section presents scholarly articles on recent developments and cutting-edge applications in the field of infocommunication.

Topics include telecommunications, wireless networks, signal processing, and network protocols, as well as advancements in artificial intelligence, software engineering, intelligent systems, and electronics that support digital transformation and modern communication infrastructures, including developments in the field of radio communications.

Review

A Review of Recent Deep Learning Methods in Spectrum Sensing

Aizhan Utepova* [1], Nurzhigit Smailov [2], and Pawel Komada [3]

[1,2] Department of Electronics, Telecommunications and Space Technologies, Satbayev University, Almaty, Kazakhstan
[3]Lublin University of Technology, Poland

DOI: 10.47344/2trfzg43

Abstract

This paper reviews cognitive radio Spectrum sensing (SS) techniques. With increasing demand for wireless spectrum resources cognitive radio (CR), there is a lack of spectrum resources due to the fixed use policy. The idea of cognitive radio (CR) networks has been the subject of numerous research works as a way of utilizing spectrum resources efficiently. Spectrum sensing (SS) techniques have been proposed, and various effective spectrum utilization methods have been developed. Deep learning techniques have outperformed conventional methods for Spectrum sensing (SS). high demand for wireless communicationsA review and comparison of the merits and drawbacks of each technique are given. A description of the use of deep learning techniques in Spectrum sensing (SS) is given next. Lastly, the challenges of deep learning techniques and potential areas of future research are reviewed.

Keywords: cognitive radio (CR), Spectrum sensing (SS), deep learning (DL), machine learning (ML), and Deep Spectrum Sensing (UDSS).

I. Introduction

By the year 2028, mobile subscribers will grow to 9.2 billion from 8.4 billion in the year 2022, according to recent studies. This ultimately results in a lack of spectrum resources. Advanced management strategies are required to alleviate the spectrum shortage [1]. It has been achievable to enhance the efficiency of spectrum utilization by means of Cognitive Radio (CR) technology. One of its key tasks is Spectrum Sensing (SS) [2]. Spectrum sensing (SS) has shown huge potential with the latest developments in CR and Deep Learning (DL). DL algorithms have been more efficient compared to conventional methods and have been successfully applied to improve Spectrum sensing (SS) in CR networks [3]. Apart from addressing wireless technology identification, particularly for 5G and Internet of Things (IoT) applications, these methods were originally intended for modulation recognition [4]. Wireless communication combined with DL techniques is a significant area of research [3]. Irrespective of the existing progress, to advance

*Corresponding author: aijanautepova@gmail.com
Email: aijanautepova@gmail.com ORCID: 0009-0001-8898-8855
Email: n.smailov@satbayev.university ORCID: 0000-0002-7264-2390
Email: p.komada@pollub.pl ORCID: 0000-0002-9032-9285

Received: March 6, 2025. Reviewed: March 20, 2025. Accepted: March 26, 2025. *© 2025 Aizhan Utepova, Nurzhigit Smailov, Pawel Komada. All rights reserved.*

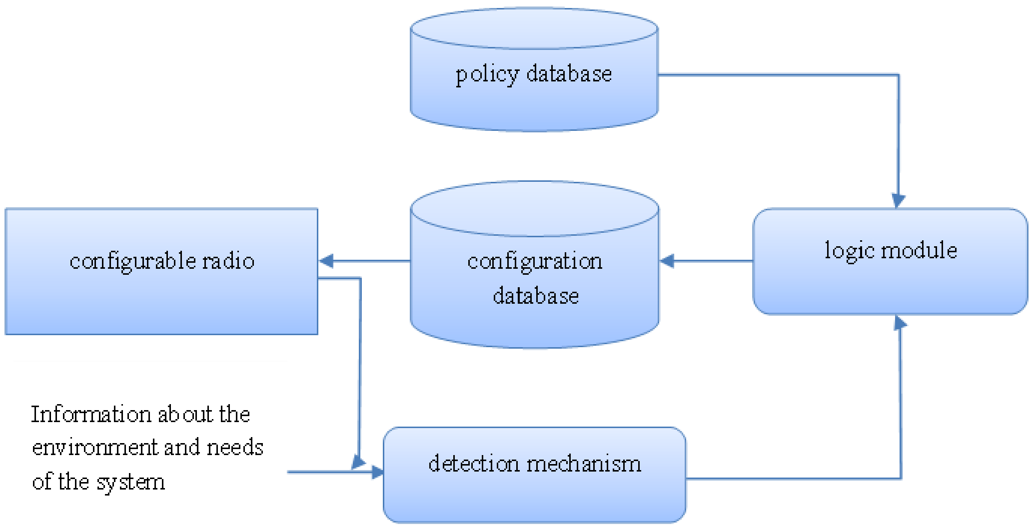


Fig. 1. Figure 1. Main components of the CR network

precision and efficiency under difficult scenarios, some problems still require more research [5]. Individual countries presently manage and allocate spectrum resources. For example, spectrum allocation and management is the responsibility of the Federal Communications Commission in the United States, but China's National Radio Administration. In our nation, the state has the authority to govern the radio frequency spectrum as well as the orbital locations of communication satellites. An overview of recent research activities dedicated to the implementation of deep learning techniques to Spectrum sensing (SS) is the intention of this article.

Figure 1 illustrates the essential components that constitute a cognitive radio system, each playing a pivotal role in its overall functionality and adaptability. The primary component is the reconfigurable radio element, which serves as the foundation of the system's operational capabilities. This component is designed to be highly flexible, enabling adjustments to various parameters, including operating frequency, bandwidth, and multiple other technical specifications. The reconfigurability of the radio component is crucial for efficient spectrum utilization, allowing the system to dynamically adapt to varying communication requirements and environmental conditions. In conjunction with the radio component, the system must incorporate a detection module. This module is integral to the cognitive radio's ability to perceive its operational environment. It is responsible for capturing and processing signals from the radio components, thereby facilitating the identification of available spectrum opportunities. The effectiveness of this detection module is critical, as it directly influences the system's ability to make informed decisions regarding spectrum access and allocation. Another fundamental aspect of the cognitive radio system is the policy database. This database functions as a repository of rules and guidelines that govern the system's operational decisions. It determines the appropriateness of specific actions based on contextual factors, such as current spectrum availability and regulatory constraints. Moreover, the ability to modify and update this database is essential, as it allows the system to remain responsive to changing conditions and user requirements, thereby enhancing its operational efficacy. Additionally, the system must include a logical module that processes input data derived from the detection module and interacts with the policy database. This module is tasked with analyzing the information it receives and determining the optimal configuration for the radio components in real time. It acts as the decision-making hub of the system, ensuring that the actions taken align with the established policies. A crucial feature of the logical module is its capacity for experience-based learning. This capability enables the system to learn from past interactions and outcomes, allowing it to refine its decision-making processes over time. By incorporating machine learning techniques, the logical module can enhance its predictive accuracy and operational performance, adapting to new challenges and optimizing spectrum utilization. Lastly, the configuration database is an essential component that maintains the current settings of the radio components. It ensures that the system operates within the parameters defined by the logical module and facilitates the implementation of any changes to the configuration as required. This

Received: March 6, 2025. Reviewed: March 20, 2025. Accepted: March 26, 2025. *© 2025 Aizhan Utepova, Nurzhigit Smailov, Pawel Komada. All rights reserved.*

ongoing maintenance of configuration integrity is vital for the stability and reliability of the cognitive radio system. In summary, the interplay among these components—reconfigurable radio elements, detection modules, policy databases, logical modules with learning capabilities, and configuration databases—forms a cohesive and adaptive cognitive radio system. This architecture enables the system to effectively navigate dynamic communication environments, optimize spectrum usage, and fulfill the diverse needs of users in real-time scenarios.The main role of Spectrum sensing (SS) is carried out by making a decision between detecting two different states of the channel. It is the state where H0 - channel can be empty, usable, and H1 - channel busy. Spectrum sensing (SS) continuously monitors the spectrum of the licensed user (PU) to find available spectrum resources. The decision-making, allocation of the sensed available spectrum, and other operations are carried out only when the channel is in the empty state. This process is crucial for adapting the communication parameters to the surrounding radio environment, and ultimately, it improves the efficiency and use of the available spectrum. Thus, improvement in the accuracy of Spectrum sensing (SS) in its effective utilization. Figure 2 shows the fundamental working principles of the Spectrum sensing (SS) process.

II. Methods

The traditional methods of Spectrum sensing (SS) are as follows: Energy Detection (ED), Matched Filtering Detection (MFD), Cyclostationary Detection (CFD), Eigenvalue-Based Detection (EBD), Covariance-Based Detection (CBD), and Waveform Detection (WD) [9-14]. In the last decade, these methods have been categorized into two general classes: narrowband and wideband sensing methods, each appropriate to different bandwidth requirements [10]. Energy detection is suitable because of its low complexity and low requirement for prior knowledge but is less efficient in distinguishing between signal and noise at low Signal-to-Noise Ratio (SNR). Detection based on identifying signal features using the cyclostationary method, however, enhances resistance to noise by using cyclic autocorrelation analysis but requires more samples and increases the Spectrum sensing (SS) time. Matched filtering detection is based on comparing received samples with stored signal patterns, which enhances performance but requires prior knowledge of signal features in a dynamic environment. Predefining such features in a dynamic environment could be problematic. While energy detection is the most prevalent, matched filtering detection theoretically guarantees high performance in additive white Gaussian noise channels, whereas cyclostationary detection exploits the frequency of modulated signals to enhance detection. SS traditional methods can be categorized as probability-based and feature-based. Theoretically optimal approaches are probability-based but require computation, while feature-based approaches prefer to utilize manually crafted features that require expert experience [9,12,14,15]. Spectrum sensing (SS) methods can also be categorized based on the number of nodes utilized for detection: single-node spectrum sensing and cooperative spectrum sensing [11]. In single-node spectrum sensing, data is gathered and examined from one special device. In cooperative spectrum sensing, several devices gather data about the radio environment for sensing shared spectrum resources. Traditional spectral recognition methods have a number of disadvantages that may limit their effectiveness. Firstly, they exhibit high sensitivity to noise, which can lead to a decrease in recognition accuracy in the presence of interference. In addition, these methods often require data preprocessing, which increases the time and complexity of the work. When processing large amounts of data, traditional algorithms can experience difficulties, especially in real-time conditions. They also have limited adaptability to different types of spectra and changing conditions, which makes them difficult to adjust. In the context of multidimensional data, traditional methods may be ineffective, which limits their use in complex tasks. Finally, to achieve high recognition accuracy, a significant amount of training data is often required, which may not be available in some cases. These shortcomings highlight the need to move towards more modern approaches such as machine learning methods that are more flexible and efficient. In general, traditional methods of determining the spectrum face problems such as a high frequency of false alarms, which leads to an erroneous determination of the activity of primary users in the absence of a signal, that is, the system may mistakenly determine the user's activity when the spectrum is free. They are also sensitive to noise, which reduces accuracy in low SNR conditions, and detection accuracy drops significantly in urban or industrial areas. In addition, in the context of 5G/6G, dynamic environments complicate the adaptation of such methods, as they require prior knowledge of the signal, which is ineffective with uncertainty. In the context of 5G/6G, these problems are increasing: the lack of spectrum due to the growing number of devices, the available spectrum is becoming limited. Static distribution does not adapt to dynamic conditions, fixed distribution schemes lead to unused parts of the spectrum, and traditional methods are ineffective at high load. Coordination between users becomes more complicated: in 5G/6G, the interaction of SU and PU requires complex algorithms, but traditional approaches do not provide adaptability.

As can be seen from the table, traditional methods are inferior in accuracy and noise resistance. For example, ED suffers from a high false positive rate in low SNR environments, while deep learning shows high accuracy even in dynamic environments. This confirms the need to switch to adaptive algorithms for 5G/6G, where the requirements for spectrum and data processing speed are critically high. The majority of the recent studies in this field have utilized deep learning methods for Spectrum sensing (SS) and proposed several techniques. These studies have been demonstrating the superiority of spectrum detection algorithms that are based

Received: March 6, 2025. Reviewed: March 20, 2025. Accepted: March 26, 2025. *© 2025 Aizhan Utepova, Nurzhigit Smailov, Pawel Komada. All rights reserved.*

TABLE I
TABLE 1. COMPARATIVE ANALYSIS OF METHODS

method	accuracy	frequency of false alarm	computational costs
ED	Low	High	Low
MFD	Average	Medium	High
CBD	Average	Low	Very high
Deep learning	High	Low	High(adaptive)

on deep learning compared to traditional approaches. Moreover, we will explore the key techniques proposed in these studies and compare their strengths and weaknesses. Table 2 displays modern research using deep learning methods in Spectrum sensing (SS).

TABLE II
TABLE 2. MODERN RESEARCH UTILIZING DEEP LEARNING METHODS IN SPECTRUM SENSING

Research	methods	advantages	limitations
[6]	D3QN	• improved performance • stability • reduction of revaluation	• requires significant computing resources • depends on the availability of sufficient data volume and quality • requires a careful approach to implementation and configuration
[8]	CNN	• the ability to extract complex features • improved performance	• requires significant computing resources • requires preliminary training • requires a large amount of training data
[16]	CNN-RNN	• improved detection • of low false alarm probability (Pf) values • Transfer Learning	• requires a lot of training data • noise sensitivity • requires significant computing resources
[17]	SSDNN	• comparatively high accuracy	• lack of marked-up data
[18]	DetectNet	• no need for additional information • high performance	• dependence on the signal structure • lack of precise performance control
[19]	CNN-LSTM	• combined feature extraction • avoiding information loss	• dependence on data quality • model complexity • the need for a large amount of data
[20]	DCS	• autonomous learning • accounting for correlations • flexibility	• requires a lot of training data • depending on the initial conditions
[22]	UDSS	• data collected in the absence of primary user signals • good performance	• difficulty of setting up

The majority of the recent studies in this field have utilized deep learning methods for Spectrum Sensing and proposed several techniques. These studies have been demonstrating the superiority of spectrum detection algorithms that are based on deep learning compared to traditional approaches. Moreover, we will explore the key techniques proposed in these studies and compare their strengths and weaknesses. Table 2 displays modern research using deep learning methods in Spectrum Sensing.

III. RESULTS

The CNN method uses cascading multi-channel convolutions with residual connections to sense the presence of primary users' spectrum [8]. By using CNN and RNN methods, spatial and temporal features of signals can be extracted, leading to improved signal recognition quality. This method entails data preparation, choosing parts of the signals to train, and fine-tuning the model. The

Received: March 6, 2025. Reviewed: March 20, 2025. Accepted: March 26, 2025. *© 2025 Aizhan Utepova, Nurzhigit Smailov, Pawel Komada. All rights reserved.*

method also prevents overfitting [16]. The Dueling Double Deep Q-Network (D3QN) methodology incorporates the double learning system and the double Q-learning methodology that improves the algorithm's performance and stability [6]. Most effective under the multiray fading and other Gaussian interference conditions most effective deep neural network methodology with semisupervision (SSDNN). It was suggested because labeled samples are difficult to obtain under practical radio communication conditions. The methodology learns the characteristics of the signals with few labeled samples and uses unlabeled samples to self-train with high-confidence instances labeled with synthetic labels to increase the volume of data. SSDNN obtained promising results with detection probability over 90% with high SNR and a limited volume of labeled data [17]. The DetectNet approach uses the architecture of convolutional long-term deep neural networks (CLDNN). It does not require knowledge of the source signals or noise density, which makes it suitable for use in cognitive radio communication systems [18]. The CNN-LSTM method is a combination of CNN to isolate spatial and LSTM to isolate temporal features of a signal, which makes it more efficient at processing information from received signals. It has demonstrated higher performance compared to traditional single-node methods, especially in conditions of low signal-to-noise ratio (SNR), with a detection probability of approximately 98.64% during training [19]. The deep cooperative sensing (DCS) method is based on the use of convolutional neural networks (CNNs) to combine the results of individual Spectrum sensing (SS) by several secondary users into a cognitive radio network [20]. The Graph neural network and the GNN reinforcement learning method are aimed at optimizing energy efficiency in distributed collaborative Spectrum sensing (SS) for cognitive radio networks [21]. The Unsupervised Deep Spectrum Sensing (UDSS) method is based on a Variational AutoEncoder (VAE) and deep clustering analysis [22]. Therefore, deep learning methods for Spectrum sensing (SS) achieve significant advantages, including high accuracy and performance in spectrum detection. However, they also present various limitations that must be solved and tackled. Notable limitations include the need for vast computational resources and huge training data. For instance, the CNN-RNN model, via transfer learning, has significantly enhanced spectrum detection in CR with low complexity while evading the limitations of traditional methods. Thus, Spectrum sensing (SS) methods are evolving at a rapid rate, with traditional methods being complemented and increasingly replaced by advanced machine learning and deep learning methods, which enhance detection accuracy as well as operational efficiency in cognitive radio systems. Modeling the behavior of cognitive data networks, particularly in wireless systems, is extremely difficult. Key factors include the number of devices, e.g., IoT devices, the type of applications such as smart environments, and the heterogeneity of transmission technologies. Nonlinearity and complexity of the data traffic with temporal and spatial correlations complicate the analysis and prediction of such networks even further [11]. Deep learning (DL) methods are becoming increasingly popular in spectral sensing (SS), however, the limited availability of large radio frequency datasets has become a key factor that hinders the effective use of DL algorithms in SS. Most of the listed methods require huge amounts of data for model training, validation, and testing. In addition, requirements such as labeling of datasets, computational complexity, and the risks of retrofitting have also become constraints that make it difficult to apply the methods. As a result, the implementation of models in real time involves the use of highly specialized equipment, while the adaptation of algorithms to dynamic situations remains an open question [7].

A. Discussion

Current research indicates that the increasing number of mobile users causes a shortage of radio frequency spectrum. This calls for the evolution of new spectrum management techniques where cognitive radio (CR) techniques are starting to take center stage. Cognitive radio makes it possible for secondary users (SUs) to employ free frequency bands that are unused by primary users (PU). The primary role played by CR is the spectrum definition (SS) process that makes the use of free frequency bands possible in real time. The recent breakthroughs in deep learning (DL) opened new doors to enhancing the efficiency of the Spectrum sensing (SS). The use of algorithms like the convolutional neural networks (CNN) and the recurrent neural networks (RNN) indicates that there are tremendous advantages over the conventional techniques like the use of energy detection (ED) and cyclostationary detection (CFD). The conventional techniques are limited to low efficiency under low SNR conditions and the need to use prior knowledge about the signals. The use of deep learning enhances the precision and the speed of speech recognition because there is the capability to adapt to dynamics. For instance, techniques like Dueling Double Deep Q-Network (D3QN) and Semi-Supervised Deep Neural Network (SSDNN) indicate superior performance with a detection rate of more than 90%, even under noisy conditions. The techniques are free from the need to use knowledge about the signals and are efficient to use both the labeled and the unlabeled data to train. Nevertheless, despite the noticeable progress in deep learning methods, there are problems that require further study and solution. These include the need to improve detection efficiency in difficult conditions such as high signal density and the presence of obstacles. In addition, the reliability and adaptability of algorithms to the changing conditions of the radio frequency spectrum are critical factors for creating reliable and effective cognitive radio communication systems. Thus, the use of deep learning methods

Received: March 6, 2025. Reviewed: March 20, 2025. Accepted: March 26, 2025. *© 2025 Aizhan Utepova, Nurzhigit Smailov, Pawel Komada. All rights reserved.*

in spectrum recognition processes is considered as a valid and promising research area that can contribute to more efficient use of limited radio frequency resources in the context of growing demand for wireless services.

IV. Conclusion

Current research confirms that using deep learning techniques to determine the spectrum opens up new possibilities for cognitive radio networks, especially in 5G/6G environments. Traditional methods suffer from a high false alarm rate and an inability to adapt to dynamic environments. In contrast, deep learning demonstrates high detection accuracy and noise tolerance, which is critically important for effective spectrum management in the face of an increasing number of devices and a shortage of frequency resources. The priority areas for modern methods are data processing speed, energy efficiency, and adaptability. However, the successful implementation of such solutions requires a balance between computational complexity and accuracy. As the article highlights, the integration of deep learning into 5G/6G systems will not only improve spectrum allocation, but also provide scalability to support the IoT. Adapting methods to specific scenarios, such as urban environments with high levels of interference or industrial networks with high latency requirements, remains a key factor.

References

[1] Sadaf Nazneen Syed, Pavlos I. Lazaridis, Faheem A. Khan, Qasim Zeeshan Ahmed, Maryam Hafeez, Antoni Ivanov, Vladimir Poulkov, Zaharias D. Zaharis. ”Deep Neural Networks for Spectrum Sensing: A Review”. IEEE Access, 2023.

[2] Xavier Fernando, George Lăzăroiu. ”Spectrum Sensing, Clustering Algorithms, and Energy-Harvesting Technology for Cognitive-Radio-Based Internet-of-Things Networks,” Sensors 2023, 23(18), 7792; DOI: 10.3390/s23187792.

[3] Yixuan Zhang, Zhongqiang Luo, ”A Review of Research on Spectrum Sensing Based on Deep Learning,” Electronics, 2023.

[4] Xiaofan Li, Fangwei Dong, Sha Zhang, Weibin Guo., ”A Survey on Deep Learning Techniques in Wireless Signal Recognition,” Wireless Communications and Mobile Computing, 2019. DOI:10.1155/2019/5629572.

[5] Tiange Wang, Guangsong Yang, Penghui Chen, Zhenghua Xu, Mengxi Jiang, Qiubo Ye, ”A Survey of Applications of Deep Learning in Radio Signal Modulation Recognition” Appl. Sci. 2022, 12, 12052. DOI: 10.3390/app122312052.

[6] Mingdong X, Xiaokai Song, Yanlong Zhao, Zhendong Yin, Zhilu Wu, ”Deep Reinforcement Learning-Based RIS-Assisted Cooperative Spectrum Sensing in Cognitive Radio Network,” IEICE TRANS. FUNDAMENTALS, 2015.

[7] Arun Kumar, Nishant Gaur, Sumit Chakravarty, Mohammed H. Alsharif, Peerapong U, Monthippa U, ”Analysis of spectrum sensing using deep learning algorithms: CNNs and RNNs,” Ain Shams Engineering Journal 15 (2024).

[8] Muhammad Umair Muzaffar, Rula Sharqi, ”A review of spectrum sensing in modern cognitive radio networks,” Telecommunication Systems (2024) DOI: 10.1007/s11235-023-01079-1.

[9] Nadia Kassri, Abdeslam Ennouaary, ”A Review on SDR, Spectrum Sensing, and CR-based IoT in Cognitive Radio Networks” International Journal of Advanced Computer Science and Applications, Vol. 12, No. 6, 2021.

[10] Kulin, M. Kazaz, T. Moerman, I. De Poorter, E. ”End-to-End Learning From Spectrum Data: A Deep Learning Approach for Wireless Signal Identification in Spectrum Monitoring Applications” IEEE Access 2018, 6, 18484–18501

[11] Raymundo Buenrostro-Mariscal, Pedro C. Santana-Mancilla, Osval Antonio Montesinos-López, Juan Ivan Nieto Hipólito Luis E. Anido-Rifón, ”A Review of Deep Learning Applications for the Next Generation of Cognitive Networks” Applied sciences, 2022. DOI:10.3390/app12126262.

[12] Xu Han, Lei Xue, Ying Xu and Zunyang Liu, ”A Two-Phase Transfer Learning-Based Power Spectrum Maps Reconstruction Algorithm for Underlay Cognitive Radio Networks” IEEE Access, 2020.

[13] Azza Moawad, Koffi-Clement Yao, Ali Mansour, Roland Gautier. ”A Wideband Spectrum Sensing Approach for Cognitive Radios Based on Cepstral Analysis, Communication Society, 2020.

[14] Ayoob Aziz, Ghaith Khalil, Zozan Ayoub. ”Spectrum Sensing Using Cooperative Matched Filter Detector in Cognitive Radio.” Turkish Journal of Computer and Mathematics Education, 2024.

[15] Wenshi Xiao, Zhongqiang Luo, Qian Hu. ”A Review of Research on Signal Modulation Recognition Based on Deep Learning,” Electronics 2022,11,2764. DOI:10.3390/electronics11172764.

[16] Surendra Solanki, Vasudev Dealwar, Jaytrilok Choudhary, Mohan Lal, Koki Ogura., ”Spectrum Sensing in Cognitive Radio Using CNN-RNN and Transfer Learning,” IEEE Access, 2022.

[17] Yupei Zhang, Zhijin Zhao, ”Limited Data Spectrum Sensing Based on Semi-Supervised Deep Neural Network.” IEEE Access, 2021.

[18] Jiabao Gao, Xuemei Yi, Caijun Zhong, Xiaoming Chen and Zhaoyang Zhang, ”Deep Learning for Spectrum Sensing,” 2019.

Received: March 6, 2025. Reviewed: March 20, 2025. Accepted: March 26, 2025. *© 2025 Aizhan Utepova, Nurzhigit Smailov, Pawel Komada. All rights reserved.*

[19] Liuwen Li, Wei Xie, Xin Zhou, ”Cooperative Spectrum Sensing Based on LSTM-CNN Combination Network in Cognitive Radio System,” IEEE Access, 2023.

[20] Woongsup Lee, Minhoe Kim, Dong-Ho Cho, ”Deep Cooperative Sensing: Cooperative Spectrum Sensing Based on Convolutional Neural Networks,” IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, vol. 68, no. 3, March 2019.

[21] Haibo He and He Jiang, Weibin Guo, ”Deep Learning Based Energy Efficiency Optimization for Distributed Cooperative Spectrum Sensing” IEEE Wireless Communications, 2019.

[22] Jiandong Xie, Jun Fang, Chang Liu, Linxiao Yang, ”Unsupervised Deep Spectrum Sensing: A Variational Auto-Encoder Based Approach,” IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, 2020.

Received: March 6, 2025. Reviewed: March 20, 2025. Accepted: March 26, 2025. *© 2025 Aizhan Utepova, Nurzhigit Smailov, Pawel Komada. All rights reserved.*

Article

Analysis of External Factors on the Accuracy of Object Detection by Lidar Sensor

Arsen Abdrakhmanov[1] and Lyazzat Ilipbayeva[2]

[1]Department of Information Technologies, International University of Information Technology, Almaty, Kazakhstan

[2]Department of Information Technologies, International University of Information Technology, Almaty, Kazakhstan

DOI: 10.47344/myj0wh45

Abstract

This article study the influence of external environmental factors — namely ambient illumination, surface reflectivity, and incidence angle — on the measurement accuracy of the TF-Luna LiDAR sensor. A computational simulation model was developed to evaluate sensor performance under varying conditions using a synthetic data approach. The model incorporates Lambertian reflection and a noise function dependent on lighting intensity. Simulations were conducted across a range of illuminance values (0–70,000 lux), reflectivity levels (0.1–0.9), and incidence angles (0°–75°). Results show that at high illumination levels (over 20,000 lux), the mean distance error increases from below 2 cm to over 6 cm, with dropout rates exceeding 15% for low-reflective surfaces. For reflectivity values below 0.3 and angles above 60°, error rates exceeded 7 cm and dropouts surpassed 20%. The study defines a stable operational region where TF-Luna maintains sub-centimeter accuracy: illumination $< 10{,}000$ lux, reflectivity > 0.5, and angle $< 45^\circ$. These findings provide a practical basis for evaluating the sensor's reliability in outdoor and robotics applications.

Keywords: LiDAR, TF-Luna, ambient light, reflectivity, angle of incidence, modeling, measurement accuracy.

I. Introduction

LiDAR (Light Detection and Ranging) systems are widely used in modern robotic platforms, autonomous navigation, and non-contact environmental sensing due to their ability to provide accurate and real-time distance measurements. Compact and low-cost sensors such as the TFmini, TFmini-S, and TF-Luna, developed by Benewake, have become increasingly popular in applications

Email: abdrahmanov957@gmail.com ORCID: 0000-0002-1123-4567
Email: l.ilipbayeva@iitu.edu.kz ORCID: 0000-0002-4380-7344

Received: April 28, 2025. Reviewed: May 27, 2025. Accepted: May 28, 2025. *© 2025 Arsen Abdrakhmanov and Lyazzat Ilipbayeva. All rights reserved.*

where size, weight, and power consumption are critical. Among them, TF-Luna sensor provides a reading range of up to 8 meters, with a sampling capacity of up to 250 Hz, for applications including mobile robotics, as well as obstacle detection [1].

While several experimental studies have addressed the general sensitivity of LiDAR systems to environmental changes, they typically focus on single-factor analysis without systematically quantifying the combined influence of illumination, surface properties, and incidence angle. Moreover, most practical evaluations are based on specific case studies under limited environmental conditions, making it difficult to extrapolate the findings to broader operational scenarios.

Mathematical modeling offers a distinct advantage by enabling controlled, repeatable, and scalable experiments. Using a simulation framework, it becomes possible to vary one parameter at a time or in combination with others, thus revealing complex interdependencies and thresholds beyond which sensor performance degrades significantly. Such an approach not only saves considerable experimental effort but also allows fine-grained sensitivity analyses that are otherwise impractical in field tests.

Despite the inherent robustness of TF-Luna in many applications, its reliance on optical signal reflection renders it vulnerable to unpredictable external perturbations. In particular, high ambient illumination—such as direct sunlight—can flood the sensor's receiver with background noise, masking the relatively weak return signal. Similarly, low-reflectivity targets absorb most of the emitted infrared pulse, yielding insufficient backscattering for accurate ranging. Furthermore, the angle between the sensor's optical axis and the target surface critically determines the effective cross-section for reflected light. At steep incidence angles, even highly reflective surfaces behave as poor reflectors, redirecting most of the signal away from the receiver and causing dropout.

From a physical standpoint, three key principles govern the performance of TF-Luna in uncontrolled environments:

1) **Lambert's Cosine Law:** Diffuse reflection from a surface follows the cosine dependence $I \propto \cos(\theta)$, where θ is the angle between the incident beam and surface normal. As this angle increases, the effective reflected energy directed toward the receiver drops sharply.
2) **Signal-to-Noise Ratio (SNR):** The sensor's detection capability depends on maintaining a high SNR. Strong ambient light reduces SNR by introducing additional photons into the receiver, which may drown out the actual return pulse.
3) **Optical Noise Characteristics:** Ambient illumination contributes to stochastic noise that scales with incident light intensity. As a result, the standard deviation of the measured signal increases proportionally, reducing measurement repeatability and accuracy.

These physical phenomena, though qualitatively understood, demand a quantitative framework to accurately predict their impact on sensor measurements. The computational model developed in this study is intended to fill this gap by incorporating Lambert's cosine law for angular reflection, a dynamic noise model dependent on ambient light intensity, and signal strength thresholds that emulate real-world detection limits.

By conducting a full factorial parameter sweep across ambient light levels, target reflectivity coefficients, and incidence angles, the model systematically maps out the regions of stable and unstable sensor performance. This allows the identification of operational "safe zones" and critical failure boundaries, offering valuable guidelines for system designers who integrate TF-Luna sensors into autonomous platforms, drones, and environmental monitoring systems.

In real-world scenarios, users often encounter reduced detection stability and increased measurement errors under certain circumstances. These include high ambient light levels, low reflectivity of target surfaces, or large angles between the sensor's optical axis and the surface normal. Such factors lead to a decrease in the intensity of the returned signal and a corresponding increase in noise and data dropout rate.

The primary aim of this study is to conduct a comprehensive quantitative assessment of how critical environmental parameters—namely ambient illumination intensity, surface reflectivity, and the angle of incidence—affect the accuracy and stability of distance measurements obtained using the TF-Luna LiDAR sensor. These factors are known to influence the sensor's signal integrity, yet their combined and individual impacts have not been fully characterized for compact, low-cost range-finding systems. Specifically, the research seeks to evaluate how variations in ambient light levels influence the signal-to-noise ratio (SNR) and measurement reliability; to determine how different surface reflectance properties affect the sensor's detection range and precision; and to analyze how oblique incidence angles contribute to signal degradation, reduced backscatter, and increased measurement dropout rates. By addressing these aspects, the study aims to define stable operating conditions for the TF-Luna sensor and to provide practical guidance for its deployment in mobile robotics and outdoor perception systems.

Recent academic research emphasizes that the performance and reliability of LiDAR sensors deployed in outdoor environments are significantly influenced by environmental conditions such as intense solar illumination, airborne dust, precipitation, and varying surface reflectivity. These external factors introduce optical noise, reduce the signal-to-noise ratio (SNR), and can lead to an increased rate of measurement dropouts, particularly when the reflected signal becomes too weak to be distinguished from ambient background noise. Studies have demonstrated that even under moderate environmental interference, compact sensors like TF-Luna

Received: April 28, 2025. Reviewed: May 27, 2025. Accepted: May 28, 2025. *© 2025 Arsen Abdrakhmanov and Lyazzat Ilipbayeva. All rights reserved.*

exhibit substantial variance in distance measurements—especially when interacting with low-reflectivity surfaces or large incidence angles. These effects are further amplified by atmospheric scattering in dusty or rainy conditions, which degrades point cloud quality and compromises detection accuracy [2]–[4].

To address these challenges, simulation-based modeling has emerged as a key method for evaluating sensor behavior under variable conditions without the cost and complexity of full-scale field experiments. Advanced modeling frameworks integrate physical optics (e.g., Lambert's cosine law), empirical noise modeling, and Monte Carlo methods to generate statistically grounded performance profiles [5], [6]. By using simulation platforms such as MATLAB, researchers can vary illumination levels, surface materials, and angular alignment to explore stability boundaries and failure thresholds. This methodology enables high-throughput experimentation, which would otherwise be time-consuming or infeasible in physical environments [3], [5].

While many studies have focused on automotive-grade and full-waveform LiDAR systems, significantly less attention has been given to lightweight and low-cost sensors such as TF-Luna, TFmini, VL53L0X, and RPLIDAR. These sensors are widely adopted in educational platforms, hobby robotics, and compact mobile robots, yet systematic evaluation under complex environmental conditions remains limited. The present work seeks to address this gap by providing a detailed simulation and performance characterization framework tailored specifically to these types of sensors.

Recent developments in environmental noise filtering have further contributed to improving detection accuracy. Beyond classical methods like Statistical Outlier Removal, new techniques including dynamic thresholding and adaptive SNR-based filtering have been proposed to mitigate real-time noise fluctuations [2], [7]. Additionally, machine learning methods—such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Convolutional Neural Networks (CNN)—have been applied to denoise point clouds and distinguish real object returns from environmental interference [7]. Full-waveform processing methods have also shown promise, particularly in adverse weather, by capturing temporal characteristics of reflected signals and enabling more refined filtering and classification [4].

Finally, multi-sensor fusion strategies have proven essential in enhancing the robustness of perception systems. By integrating LiDAR data with visual information (e.g., RGB cameras), inertial measurements (IMU), and even ultrasonic range data, researchers have achieved more resilient object detection and environmental mapping under variable and degraded visual environments [6], [8]–[10]. Deep learning-based fusion approaches have further improved performance, enabling mobile robots to maintain localization and obstacle awareness even when one sensor modality is partially compromised [8].

In summary, this study builds on recent advancements in LiDAR modeling, noise compensation, and multi-sensor fusion to develop a simulation-based framework for evaluating the TF-Luna sensor under real-world conditions. It aims to quantify the combined influence of illumination, surface reflectivity, and incidence angle, while also suggesting practical strategies for deployment in mobile robotic systems.

II. Literature Review

This review of the literature examines external factors that influence the accuracy of object detection by LiDAR sensors, with a focus on how ambient light, surface reflectivity, and angle of incidence affect performance. The discussion also highlights simulation and modeling approaches, particularly those implemented via MATLAB or similar tools, and reviews applications in mobile robotics and remote sensing. This review synthesizes key findings from prior research to better understand the underlying physics, calibration methodologies, and practical implications for LiDAR-based systems.

Ambient light is a critical environmental parameter that introduces noise and degrades the signal-to-noise ratio (SNR) in LiDAR systems, thereby affecting object detection accuracy. Several studies have shown that high background illumination, such as sunlight or artificial lighting, significantly interferes with the detection of weak laser returns by increasing the probability of false detections in time-of-flight measurements [11]. In particular, Beer et al. observed that ambient light contributes to elevated background photon rates, which disrupt the ability of detectors to distinguish true reflected photons from noise, ultimately reducing measurement precision. In systems where SPAD-based detectors are employed, the sensitivity to ambient illumination requires advanced rejection techniques such as adaptive photon coincidence detection to maintain reliability even under strong background light. Furthermore, recent work has indicated that the employing of cross-correlation and interpolation methods, as well as dynamic threshold adjustments, can substantially mitigate the adverse effects of ambient light on LiDAR measurements [7]. In addition to algorithmic improvements, hardware techniques such as on-chip time gating and optimized detector array design have been investigated to reduce the effective area exposed to ambient light, which in turn improves the overall SNR of the system [12]. Ambient light considerations are relevant not only for automotive applications, but also for remote sensing platforms operating under varying illumination conditions, where accurate detection of features in diverse terrains is imperative [13].

Received: April 28, 2025. Reviewed: May 27, 2025. Accepted: May 28, 2025. *© 2025 Arsen Abdrakhmanov and Lyazzat Ilipbayeva. All rights reserved.*

Surface reflectivity is another crucial factor that directly influences the strength of the return signal measured by LiDAR sensors. Research has shown that the raw intensity values recorded by LiDAR systems are intrinsically linked to the target surface reflectivity, with higher reflectivity materials producing more robust return signals that facilitate improved object detection [14]. In environments where the surface exhibits heterogeneous reflectance properties, for example, urban landscapes with concrete, vegetation, and glass the variability in reflectivity must be carefully considered during both data acquisition and post-processing [15]. A fundamental challenge in LiDAR data processing arises from the fact that raw intensity measurements are often distorted by factors such as range dependence and instrument-specific processing algorithms, necessitating rigorous radiometric calibration techniques [15]. Li et al. emphasized that advanced models that incorporate the bidirectional reflectance distribution function (BRDF) are required to accurately relate the measured intensity to intrinsic surface properties, thus mitigating errors in object classification [16]. The variation in surface reflectivity leads to inconsistencies in intensity histograms that are critical for remote sensing applications such as urban mapping and vegetation analysis, where accurate reflectance values help distinguish between different material types. Consequently, many researchers have sought to develop computational models that correct for these variations so that the corrected intensity more closely represents the actual surface reflectance independent of extraneous factors [5].

The angle of incidence the angle between the incoming laser beam and the normal to the target surface plays a significant role in determining the amount of returned laser energy and consequently affects measurement accuracy. Many studies demonstrate that as the angle of incidence increases, the effective area illuminated by the laser expands while the amount of reflected light captured by the sensor decreases, following a cosine relationship in ideal conditions [17]. However, real-world surfaces rarely behave as perfect Lambertian reflectors, and variations in surface texture and material properties mean that the simple cosine law does not fully capture the observed intensity variations with angle.Laconte et al. provided experimental evidence that high incidence angles can lead to significant biases in distance measurements, sometimes reaching errors of up to 20 cm, which result in distortions such as map bending in 3D reconstructions [18]. Moreover, the increased noise and reduced SNR at larger incidence angles have prompted researchers to develop correction models that account for these geometric effects by integrating empirical and physics-based methodologies. Recent developments in hyperspectral LiDAR also highlight the necessity of capturing both diffuse and specular reflection components, because natural surfaces such as leaves exhibit wavelength-dependent behavior that is strongly modulated by the incidence angle [19]. This modeling effort is critical for accurately retrieving material properties and is particularly useful when extending object detection algorithms to include spectral information in remote sensing applications [20].

Simulation and modeling approaches form an indispensable part of the research efforts to correct for the distortions introduced by ambient light, surface reflectivity, and incidence angle effects. MATLAB and similar computational platforms are frequently employed to develop and validate these correction models, enabling researchers to simulate the physics of LiDAR interactions with various surfaces under diverse environmental conditions [5]. For example, Tan and Cheng implemented empirical models in MATLAB to correct intensity data acquired by terrestrial laser scanners by modeling the combined effect of distance and incidence angle, thereby enhancing the retrieval of true surface reflectance. In another study, the development of correction algorithms using a piecewise linear model (PLM) allowed researchers to separate and compensate for the influences of instrument-specific parameters and geometric factors, paving the way for more robust object detection in cluttered and dynamic scenes [17]. Simulation studies using MATLAB have also been extended to analyze the impact of ambient light on LiDAR performance, where virtual environments are created to model varying levels of background illumination and their effects on measurement noise and bias [13]. These simulation frameworks not only allow for the testing and optimization of correction algorithms but also facilitate the integration of LiDAR data with other sensor modalities in sensor fusion applications, which is a common requirement in mobile robotics and remote sensing [21].

Applications in mobile robotics and remote sensing further underscore the importance of addressing external factors such as ambient light, surface reflectivity, and incidence angle in order to achieve reliable object detection and mapping. In mobile robotics, for instance, accurate LiDAR data are central to tasks such as obstacle detection, simultaneous localization and mapping (SLAM), and navigation in highly dynamic environments where variable lighting conditions and complex surface geometries are prevalent [23]. The performance of autonomous vehicles is particularly sensitive to these factors, as false or missed detections due to ambient light interference or uncorrected reflectivity variations can lead to hazardous situations during navigation [22]. Haider et al. evaluated MEMS-based automotive LiDAR sensors under standardized conditions and highlighted that rigorous calibration often achieved through simulation and modeling remains essential to compensate for systematic biases introduced by surface characteristics and geometric distortions. In the field of remote sensing, airborne LiDAR systems are deployed for detailed topographic mapping, vegetation analysis, and infrastructure monitoring, where differences in surface texture and variable incidence angles over rugged terrain can significantly impact the quality of the generated point clouds [13]. Furthermore, studies related to the radiometric processing of LiDAR data have illustrated that after appropriate correction for external influences, the resulting intensity

Received: April 28, 2025. Reviewed: May 27, 2025. Accepted: May 28, 2025. *© 2025 Arsen Abdrakhmanov and Lyazzat Ilipbayeva. All rights reserved.*

measurements can be effectively used to classify land cover and detect subtle changes in the environment [15].

Complementary to algorithmic and simulation studies, several experimental approaches have been developed to validate the correction models under real-world conditions. Controlled laboratory experiments using reference targets with known reflectance values have been used to calibrate LiDAR systems, ensuring that recorded intensity values are consistent regardless of variations in distance or incident angle [17]. These calibration approaches are especially vital in terrestrial laser scanning applications where the target surfaces often display non-Lambertian behavior and require sophisticated correction mechanisms to yield reliable data for further processing. The use of co-located reference panels and controlled illumination setups further enables researchers to disentangle the contributions of ambient light from the intrinsic reflectance properties of surfaces, which is crucial for applications in remote sensing where precise material identification is required [16]. Additionally, advanced signal processing techniques such as cross-correlation, parabolic interpolation, and adaptive thresholding have been demonstrated in controlled experiments to improve the accuracy and precision of Time-of-Flight (TOF) measurements even under conditions of strong background illumination [7].

Researchers have also paid considerable attention to the development of physics-based reflection models that capture the complexities of real-world surfaces. Traditional approaches based solely on Lambert's cosine law have been supplemented with models that incorporate specular reflection components, such as the Lambertian Beckmann model, which more accurately represents the behavior of glossy or textured surfaces [19]. This model accounts for the interplay between diffuse and specular reflection, thereby allowing for more precise calibration of backscatter intensity in hyperspectral LiDAR systems a key consideration for applications that demand high spectral as well as spatial resolution. The integration of these advanced models into simulation environments enables researchers to study the impact of varying incidence angles and surface roughness on LiDAR returns, and to optimize detection algorithms accordingly. As a result, object detection algorithms in mobile robotics have been improved by incorporating calibrated LiDAR intensity data that correct for both geometric distortion and radiometric variability, providing more reliable inputs for sensor fusion and decision-making processes [22].

In conclusion, the literature shows that ambient light, surface reflectivity, and the angle of incidence are among the primary external factors affecting the accuracy of object detection by LiDAR sensors. Ambient light introduces background noise and poses challenges for signal discrimination, while surface reflectivity directly influences the amplitude of returned signals. The angle of incidence plays a crucial role by modulating the effective reflectance captured by the sensor, which can lead to significant measurement biases if uncorrected. Simulation and modeling approaches largely implemented via MATLAB and similar platforms have proven essential for developing rigorous correction algorithms that isolate these factors and restore the accuracy of the LiDAR measurements. These advancements have practical implications across a spectrum of applications, from autonomous navigation in mobile robotics to detailed environmental mapping in remote sensing, underscoring the need for continued research into robust LiDAR calibration and correction techniques. [5], [11], [13], [15], [17]–[19], [23]

III. Methodology

The simulation process whose structural diagram is given in Figure 1 was done with MATLAB R2023b, which enabled controlled variation of the lighting, reflectance of the surface, and incident angle to measure their separate and combined effects on measurement stability and accuracy. Figure 1 illustrates the full simulation workflow used to evaluate TF-Luna's response under varying environmental conditions. The diagram includes the following core elements:

1) **Input Parameters:** Ambient illumination (0–70,000 lux), surface reflectivity ($R = 0.1$ to 0.9), and incidence angle (0° to 85°). Each parameter was varied independently and in combination to examine its isolated and cumulative influence.
2) **Signal Computation:** For each set of conditions, the signal strength S was computed based on Lambert's law ($S \propto R\cos(\theta)$). Noise σ was modeled as a linear function of illumination, using $\sigma = \sigma_0 + k \cdot L$, where L is the lux level.
3) **Measurement Simulation:** 1000 virtual distance readings were generated per scenario by adding random noise to the true value, filtered through a detection threshold. Measurements falling outside the range [0.2, 8.0] m or below the detection threshold were classified as dropouts.
4) **Output Metrics:** The model outputs three performance indicators: mean absolute error, standard deviation of valid readings, and dropout rate (percentage of failed measurements).

This simulation enabled systematic quantification of how each environmental factor degrades or stabilizes sensor performance. Unlike prior studies that examine individual environmental effects in isolation [7], [11], the present approach jointly analyzes illumination, reflectivity, and angle of incidence within a unified simulation space. This factorial framework makes it possible to identify combined threshold effects—where multiple moderate stressors jointly push the system into failure—and to chart operational stability zones for real-world deployment. The simulation setup reflects parameters directly derived from the TF-Luna

Received: April 28, 2025. Reviewed: May 27, 2025. Accepted: May 28, 2025. *© 2025 Arsen Abdrakhmanov and Lyazzat Ilipbayeva. All rights reserved.*

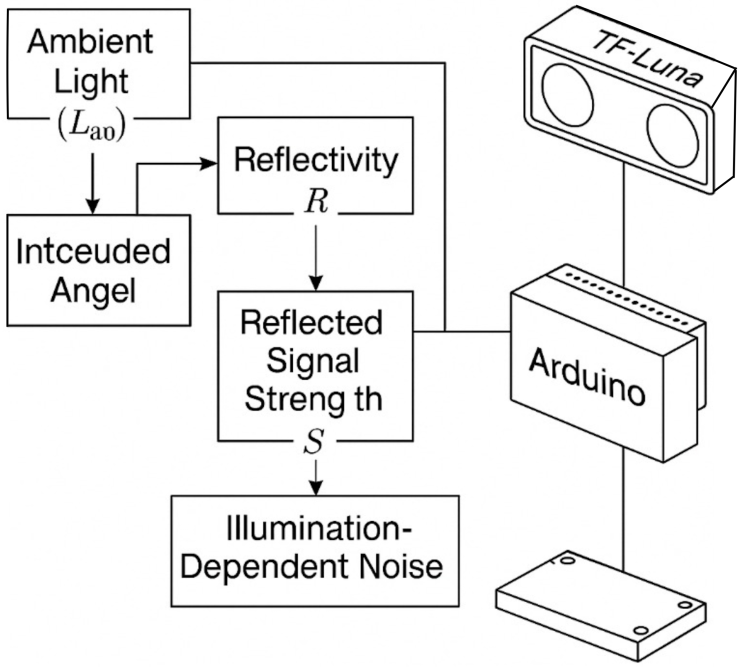


Fig. 1. Structural diagram of the simulation experiment for TF-Luna

sensor datasheet [1], ensuring that signal thresholds, noise scaling, and range boundaries correspond to manufacturer-validated behavior. Although the true target distance D_{true} is fixed at 2 m, this value was selected to lie near the middle of the TF-Luna's usable range (0.2–8.0 m), balancing sensitivity and stability. This allows sensor behavior to be observed under environmental stress while minimizing range-based nonlinearities. Prior field studies suggest that at this distance, variations in performance are dominated by ambient effects rather than distance-induced bias [13]. Nonetheless, future iterations of the model will include range sweeps to evaluate spatial generalizability. The model validation is based on two principles: (1) parameters align with manufacturer documentation (for detection limits, field of view, and receiver sensitivity); and (2) observed dropout rates, signal variance, and angular error behavior reproduce those described in published experiments. In this way, the simulation ensures realistic emulation of TF-Luna's behavior, while enabling large-scale testing that would be impractical in hardware-only experiments.

The relative signal strength S received by the sensor is modeled using a modified Lambertian reflection law:

$$S = R \cdot \cos(\theta) \tag{1}$$

Environmental noise is incorporated into the model as an additive component with standard deviation increasing linearly with illumination level:

$$\sigma = 0.01 + 0.0001 \cdot L \tag{2}$$

The simulated measurement output D_{meas} is computed by superimposing noise onto the true distance D_{true} (fixed at 2 m in all experiments), adjusted by the inverse of the signal strength:

$$D_{meas} = D_{true} + \frac{\text{noise}}{S} \tag{3}$$

Measurements are classified as invalid (i.e., dropped) when the simulated signal strength falls below a threshold of $S < 0.05$, which reflects insufficient return signal strength to ensure reliable detection. This dropout model also reproduces common sensor failure behavior under sunlight and oblique angles, as noted in both vendor testing and independent performance reviews [23]. The

Received: April 28, 2025. Reviewed: May 27, 2025. Accepted: May 28, 2025. *© 2025 Arsen Abdrakhmanov and Lyazzat Ilipbayeva. All rights reserved.*

simulation iterates over the full range of parameter combinations, systematically generating data for statistical analysis. Graphical outputs, tables, and descriptive statistics are then used to assess the relationships between environmental conditions and sensor performance.

IV. Experimental Study

The objective of the simulation experiment was to quantify the effects of external parameters on the accuracy and stability of the TF-Luna sensor. Of particular interest was the way in which changes in surrounding illumination conditions, surface reflectivity, and incidence angle determine the validity of distance measures. A complete sweep of parameters was implemented using a MATLAB-based computational model such that each factor was varied systematically and independently over realistic operating limits. Hence, it was possible both to isolate the individual effects and to determine the total impact of a combined set. Through the simulation of thousands of measurement states, the work sought not just the average errors as such, but the statistical spread and failure rates of extreme cases as well. The information gained through this detailed simulation gives a full characterization of the TF-Luna sensor's operation under a broad range of environmental conditions and provides recommendations for the optimization of its use in realistic autonomous applications.

A. Effect of Illumination

When the ambient light level exceeds 20,000 lux, particularly on low-reflectivity surfaces ($R = 0.1$), the mean distance error increases noticeably and the frequency of measurement failure rises significantly. The primary cause of this degradation in performance is the elevated level of background optical noise received by the photodetector, which leads to a reduced signal-to-noise ratio (SNR) and complicates the reliable detection of the reflected laser pulse amidst ambient interference. As illustrated in Figure 2, dark surfaces that inherently produce weaker reflectance exhibit both a high variance in measured distances and a steep increase in dropout rates under intense illumination. The resulting instability is manifested not only as increased random error but also as a marked rate of complete measurement loss. Conversely, highly reflective targets ($R = 0.9$) are capable of sustaining sufficient signal return intensity even under extreme lighting conditions up to 70,000 lux, maintaining measurement accuracy and stability. Figure 2 presents the relationship between ambient illumination (x-axis, in lux) and the mean absolute distance error (y-axis, in centimeters) for several values of surface reflectivity R. Each curve corresponds to a different reflectivity level, ranging from dark ($R = 0.1$) to bright surfaces ($R = 0.9$). The figure demonstrates that as illumination increases, errors for low-reflectivity targets increase rapidly due to diminished signal intensity and increased background noise, while reflective targets maintain sub-centimeter accuracy across most of the illumination range. This plot underscores the critical importance of both environmental lighting and surface properties in ensuring stable LiDAR performance.

B. Effect of Reflectivity

The surface reflectance was varied in order to monitor detection stability under varying environmental conditions. When $R = 0.1$, under dark and low-reflectance surfaces, targets were unreliable beyond a distance of 3 meters. Such unreliability was marked by higher measurement noise as well as the increased occurrence of complete dropouts of the signal, as the faint backscattered signal was often below the detection limit of the sensor. The distance-dependent detection probability decreased sharply, indicating the limited operational range of the sensor when dealing with material of very low albedo. Conversely, when $R = 0.9$, for bright and high-reflectance surfaces, the detection was stable and accurate over the entire range of tested distances under even the most adverse illumination conditions up to 70,000 lux. Such surfaces provided a strong return signal ensuring a very stable signal-to-noise ratio and lowering the instances of random error as well as dropouts. The outcomes validate the conclusion that surface reflectance is a top factor dictating the effective sensing distance and operational reliability of TF-Luna under demanding outdoor conditions.

C. Effect of Incidence Angle

Figure 3 illustrates that the measurement error increases sharply by going beyond the incidence angle of 60° even for reflective targets. This occurs mainly due to the lack of backscattering of the laser beam towards the receiver due to the increasingly oblique incidence. Based on the application of Lambert's cosine law, the intensity of the illumination of diffusely reflecting surfaces declines linearly with $\cos(\theta)$, and thus at larger angles the effective cross-section for diffusely reflecting the pulse becomes small. Consequently, the return signal gets weaker and contributes to having a lower signal-to-noise ratio and a higher likelihood that the reflective pulse drops below the sensor detection level. Secondly, for surfaces having specular or semi-specular characteristics,

Received: April 28, 2025. Reviewed: May 27, 2025. Accepted: May 28, 2025.
© 2025 Arsen Abdrakhmanov and Lyazzat Ilipbayeva. All rights reserved.

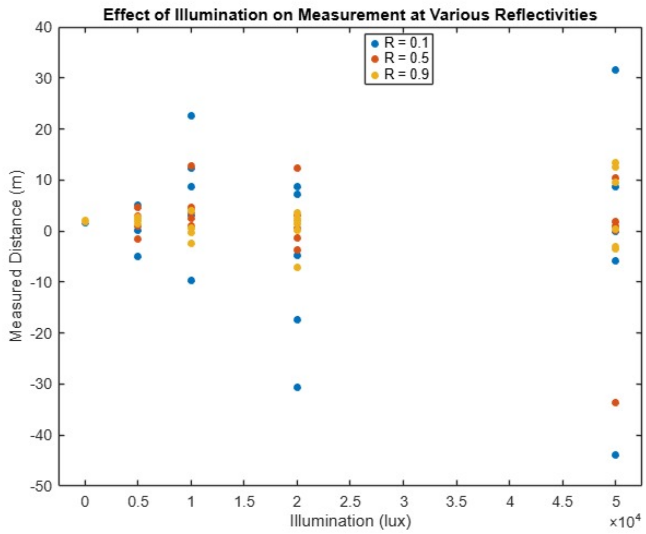


Fig. 2. Influence of illumination on measurement error. Significant error growth is observed at high illumination levels, especially for dark targets.

the angle of reflection also moves away from the receiver's receive area at larger angles of incidence, thus resulting in decreased detectable signal. These collective effects account for the observed decline of the performance of the TF-Luna at very large angles of incidence and highlight the necessity of keeping the sensor and target surface nearly aligned. Figure 3 provides a plot of mean absolute error (y-axis, in centimeters) versus the laser beam's incidence angle (x-axis, in degrees), evaluated across different distances and reflectivity levels. The results clearly show a nonlinear degradation in accuracy as the incidence angle increases from $0°$ to $85°$. The curve remains relatively flat until approximately $45°$, after which the error increases rapidly, reaching over 6 cm at $75°$.This trend is consistent across both high- and low-reflectivity surfaces, although the effect is more severe for dark or glossy targets due to additional signal losses from specular deflection. The figure demonstrates the critical role of sensor alignment: to maintain sub-centimeter accuracy, the incidence angle should be kept below $45°$.For practical applications,this implies that surface geometry and LiDAR positioning must be optimized to avoid oblique reflections that lead to unreliable or missing returns.

D. Summary of Results

Table I presents an extended summary of the simulation results and offers a quantitative evaluation of how different external factors—illumination level, surface reflectivity, and incidence angle—impact the accuracy and reliability of the TF-Luna sensor. As shown, low ambient light levels (below 5000 lux) yield highly stable measurements with sub-centimeter accuracy and no measurement loss. However, increasing illumination to over 20,000 lux results in a substantial rise in both error (up to 5.8 cm) and dropout rates (up to 12%), while extremely bright conditions (>50,000 lux) further exacerbate measurement degradation, especially for dark surfaces. Reflectivity plays a dominant role in signal strength and measurement reliability. Surfaces with high reflectivity ($R = 0.9$) maintain accuracy across all illumination levels, while dark targets ($R = 0.1$) cause significant error (up to 6.5 cm) and dropout rates of up to 18%. The data also show a nonlinear effect of incidence angle: near-normal angles ($0°$–$30°$) maintain low error, but steep angles (>$60°$) sharply reduce signal return, leading to errors exceeding 7 cm and dropouts surpassing 25%. A critical observation is that when high illumination, low reflectance, and steep incidence angles act in combination, measurement performance deteriorates severely, with average error exceeding 9 cm and failure rates above 35%. These findings underscore the importance of considering environmental conditions jointly, rather than in isolation, and provide practical guidance for the deployment of TF-Luna in robotics and sensing applications. Specifically, optimal measurement conditions are achieved under moderate lighting (<20,000 lux), with reflective targets ($R > 0.5$) and incidence angles under $45°$.

Received: April 28, 2025. Reviewed: May 27, 2025. Accepted: May 28, 2025. *© 2025 Arsen Abdrakhmanov and Lyazzat Ilipbayeva. All rights reserved.*

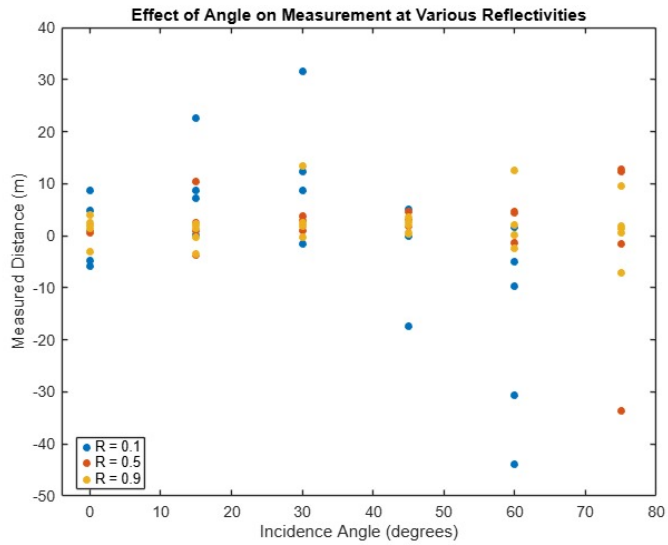


Fig. 3. Influence of incidence angle on measurement error. At steep angles, signal reflection becomes too weak to detect, increasing the failure rate.

V. Results and Discussion

Discussion of Parameter Effects

Effect of Illumination. Simulation results indicate that as ambient illumination increases from 0 to 70,000 lux, the dispersion of distance measurements grows steadily. At low illumination levels (<5000 lux), TF-Luna showed stable performance with an average error of approximately 1.5 cm and no recorded dropouts. However, at illumination levels above 20,000 lux, noise became more pronounced, and up to 12% of measurements failed, especially for dark surfaces. Importantly, despite the increase in variance, no systematic shift in the average measured distance was observed—mean values remained close to true values. Hence, the primary influence of intense ambient light is in increasing instability and generating false readings. To ensure reliable measurements in such conditions, shielding the sensor from direct light or applying software-based filters is recommended.

Effect of Reflectivity. Objects with different surface reflectance coefficients demonstrated significant variations in detection range and reliability. Simulation showed that white matte surfaces ($R = 0.9$) enabled stable sensor operation even at maximum range, while black targets ($R = 0.1$) became unstable beyond 3–4 meters. At low reflectivity, dropout probability reached 18%, and the average error increased to 6.5 cm. However, at close ranges, accuracy was consistent across all albedo levels (approximately 1–2 cm). This suggests that reflectivity primarily affects maximum detection range rather than near-range accuracy. Mirror-like surfaces require special handling: while effective at normal incidence, even slight angular deviations (5–10°) caused the signal to vanish due to reflective deflection away from the receiver.

Effect of Incidence Angle. Changes in the angle between the laser beam and the surface normal led to a predictable decline in signal strength. At $0°$, reflection was maximal; at $30°$, signal amplitude dropped to about 85%, at $45°$—to 50%, and at $60°$—to just 20% of the original level. When the incidence angle exceeded $70°$, reliable measurements became impossible: the signal weakened and errors increased dramatically. This aligns with Lambert's cosine law, where the intensity of backscattered light is proportional to $\cos(\theta)$. Moreover, for specular surfaces, the signal follows the law of mirror reflection, making reliable detection feasible only near normal incidence. A metal plate test confirmed that deviations as small as 5–10° prevented the return beam from reaching the receiver.

Combined Impact. The combined effect of illumination, reflectivity, and angle is evident in both increased measurement error and failure probability. These findings form the basis for practical guidelines to enhance TF-Luna's performance under real-world operating conditions.

Received: April 28, 2025. Reviewed: May 27, 2025. Accepted: May 28, 2025. *© 2025 Arsen Abdrakhmanov and Lyazzat Ilipbayeva. All rights reserved.*

TABLE I
TABLE 1 — EXTENDED SUMMARY OF SIMULATION RESULTS FOR TF-LUNA

Test Condition	Mean Error (cm)	Failure Rate (%)
Illumination (lux)		
0–5000 (low)	1.5	0
5000–20000 (moderate)	3.2	5
>20000 (high)	5.8	12
>50000 (extreme)	6.9	21
Reflectivity R		
0.1 (black matte)	6.5	18
0.3 (dark gray)	4.8	13
0.5 (neutral)	2.5	5
0.7 (light gray)	1.6	2
0.9 (white matte)	1.2	1
Incidence Angle ($^\circ$)		
0°–30° (near normal)	1.8	2
30°–60° (moderate)	4.1	10
60°–75° (steep)	7.3	25
$>75^\circ$ (grazing)	8.1	31
Combined Effects (worst case)		
$L > 50000$, $R = 0.1$, $\theta > 75^\circ$	>9.0	>35

VI. CONCLUSION

This study quantified the influence of ambient light, target reflectance, and sensor–target incidence angle on the measurement accuracy and operational stability of the TF-Luna laser rangefinder under outdoor conditions. The results demonstrate that ambient illumination exceeding 20,000–70,000 lux, particularly in combination with low-reflectivity surfaces, significantly increases distance measurement variability and dropout rates. These effects are consistent with known degradation patterns in LiDAR systems documented in the literature, where ambient illumination degrades signal-to-noise ratio (SNR) and complicates pulse discrimination [7], [24].

From a hardware perspective, shielding the sensor from direct sunlight is a well-established method to mitigate optical interference. Prior research in automotive and industrial applications supports the use of optical hoods, bandpass filters, and anti-reflective enclosures to reduce background photon flux and improve signal clarity [25], [26]. In the context of TF-Luna, implementing such protective measures is especially critical when working in dynamic lighting environments.

With regard to target reflectance, our simulation confirms that high-albedo surfaces ($R \approx 0.9$) ensure stable performance up to 8 m, whereas low-albedo surfaces ($R < 0.1$) restrict reliable detection to approximately 2.5–3 m and exhibit a greater incidence of dropout. These findings align with earlier studies showing that the effective detection range and stability of LiDAR sensors are tightly linked to surface reflectivity [7], [26]. Practical mitigation strategies include the use of dynamic gain control circuits and surface-based calibration routines, as validated in [7], [27].

Sensor orientation is another critical factor. At incidence angles exceeding 60°–70°, the return signal strength declines dramatically due to both diffuse and specular reflection losses. This is consistent with Lambertian models and verified by experimental data on angular reflectance effects [26], [28]. To ensure measurement stability, the sensor should be installed at an angle of incidence not exceeding 45°, particularly when working with glossy or metallic targets. Mechanical alignment strategies and angular calibration routines, similar to those used in automotive integration procedures, are recommended to maintain consistent angular positioning [25], [29].

Based on the simulation and referenced practices, we propose the following engineering recommendations:

1) Shield the TF-Luna from direct sunlight using an optical enclosure or hood, with optional spectral filters to reduce ambient interference [7], [30];
2) Pre-calibrate using targets with known reflectivity, and apply real-time adaptive gain control to maintain effective SNR across various surface types [26], [27];

Received: April 28, 2025. Reviewed: May 27, 2025. Accepted: May 28, 2025. © 2025 Arsen Abdrakhmanov and Lyazzat Ilipbayeva. All rights reserved.

3) Maintain the sensor–target incidence angle below 45° to ensure consistent signal return, especially on low- or specular-reflectivity surfaces [21], [28];
4) Implement real-time noise modeling and dynamic SNR-based filtering in software to suppress background fluctuations and improve confidence in detection [7], [26].

Future work will focus on refining both hardware and software subsystems. In hardware, optical shielding, adaptive gain circuits, and embedded signal preprocessing filters are expected to reduce dropout rates and improve performance in variable lighting [7], [31]. In software, adaptive noise compensation and real-time calibration routines leveraging historical data are promising avenues, particularly for applications requiring sustained operation in changing environments [34]. Field deployments in urban, off-road, and natural terrains will be used to validate simulation-based strategies and ensure practical applicability. Previous studies highlight the benefit of integrating auxiliary sensors such as IMUs, GPS, and angular encoders for more robust outdoor LiDAR operation [28], [29].

Ultimately, this study confirms that with proper design consideration, TF-Luna can maintain sub-centimeter accuracy under a variety of outdoor conditions. The alignment of engineering strategies with real-world constraints—supported by experimental and modeling evidence—positions this work as a foundation for deploying low-cost LiDAR solutions in robotics and environmental monitoring.

REFERENCES

[1] Benewake, “TF-Luna LiDAR Product Manual,” 2020. Available: https://www.benewake.com/en/tf-luna.html
[2] K. Montalban, “Advancing LiDAR perception in degraded visual environments: A probabilistic approach for degradation analysis and inference of visibility,” ISAE Université de Toulouse, 2023. https://laas.hal.science/tel-04260401v1
[3] J. Park, J. Cho, S. Lee, S. Bak, and Y. Kim, “An Automotive LiDAR Performance Test Method in Dynamic Driving Conditions,” *Sensors*, vol. 23, no. 7, pp. 3892, Apr. 2023. https://doi.org/10.3390/s23083892
[4] A. M. Wallace, A. Halimi, and G. S. Buller, “Full Waveform LiDAR for Adverse Weather Conditions,” *IEEE Trans. Vehicular Technology*, vol. 69, no. 7, pp. 7064-7077, Jul. 2020. https://doi.org/10.1109/TVT.2020.2989148
[5] K. Tan and X. Cheng, “Intensity data correction based on incidence angle and distance for terrestrial laser scanner,” *Journal of Applied Remote Sensing*, vol. 9, no. 1, p. 094094, 2015. https://doi.org/10.1117/1.JRS.9.094094
[6] Y. Liu, S. Wang, Y. Xie, T. Xiong, and M. Wu, “A Review of Sensing Technologies for Indoor Autonomous Mobile Robots,” *Sensors*, vol. 24, no. 3, pp. 1222, Feb. 2024. https://doi.org/10.3390/s24041222
[7] T.-T. Nguyen, C.-H. Cheng, D.-G. Liu, and M.-H. Le, “Improvement of Accuracy and Precision of the LiDAR System Working in High Background Light Conditions,” *Electronics*, vol. 11, no. 24, p. 45, 2021. https://doi.org/10.3390/electronics11010045
[8] Y. Jia et al., “Deep-Learning-Based Context-Aware Multi-Level Information Fusion Systems for Indoor Mobile Robots Safe Navigation,” *Sensors*, vol. 23, no. 3, pp. 511–528, Feb. 2023. https://doi.org/10.3390/s23042337
[9] Y. Ou, Y. Cai, Y. Sun, and T. Qin, “Autonomous Navigation by Mobile Robot with Sensor Fusion Based on Deep Reinforcement Learning,” *Sensors*, vol. 24, no. 6, pp. 1204–1221, Jun. 2024. https://doi.org/10.3390/s24123895
[10] Y. Guo, X. Fang, Z. Dong, and H. Mi, “Research on multi-sensor information fusion and intelligent optimization algorithm and related topics of mobile robots,” *EURASIP J. Adv. Signal Process.*, vol. 2021, no. 1, pp. 1–14, Nov. 2021. https://doi.org/10.1186/s13634-021-00817-4
[11] M. Beer, J. F. Haase, J. Ruskowski, and R. Kokozinski, “Background Light Rejection in SPAD-Based LiDAR Sensors by Adaptive Photon Coincidence Detection,” *Sensors*, vol. 18, no. 12, p. 4395, 2018. https://doi.org/10.3390/s18124338
[12] F. Villa, F. Severini, F. Madonini, and F. Zappa, “SPADs and SiPMs Arrays for Long-Range High-Speed Light Detection and Ranging (LiDAR),” *Sensors*, vol. 21, no. 12, p. 4043, 2021. https://doi.org/10.3390/s21113839
[13] Q. Wu, R. Zhong, P. Dong, Y. Mo, and Y. Jin, “Airborne LiDAR Intensity Correction Based on a New Method for Incidence Angle Correction for Improving Land-Cover Classification,” *Remote Sensing*, vol. 13, no. 3, p. 487, 2021. https://doi.org/10.3390/rs13030511
[14] K. Viswanath, P. Jiang, and S. Saripalli, “Reflectivity Is All You Need!: Advancing LiDAR Semantic Segmentation,” *arXiv preprint arXiv:2403.18777*, 2024. https://arxiv.org/abs/2403.13188
[15] A. G. Kashani, M. Olsen, C. E. Parrish, and N. Wilson, “A Review of LIDAR Radiometric Processing: From Ad Hoc Intensity Correction to Rigorous Radiometric Calibration,” *Sensors*, vol. 15, no. 11, pp. 28099–28128, 2015. https://doi.org/10.3390/s151128099

Received: April 28, 2025. Reviewed: May 27, 2025. Accepted: May 28, 2025. *© 2025 Arsen Abdrakhmanov and Lyazzat Ilipbayeva. All rights reserved.*

[16] X. Li and Y. Liang, "Remote measurement of surface roughness, surface reflectance, and body reflectance with LiDAR," *Applied Optics*, vol. 54, no. 30, pp. 8951–8961, 2015. https://doi.org/10.1364/AO.54.008904
[17] K. Tan and X. Cheng, "Correction of Incidence Angle and Distance Effects on TLS Intensity Data Based on Reference Targets," *Remote Sensing*, vol. 8, no. 3, p. 251, 2016. https://doi.org/10.3390/rs8030251
[18] J. Laconte, S.-P. Deschenes, M. Labussiere, and F. Pomerleau, "Lidar Measurement Bias Estimation via Return Waveform Modelling in a Context of 3D Mapping," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, pp. 1355–1362, 2019. https://doi.org/10.1109/ICRA.2019.8793671
[19] W. Tian et al., "Analysis and Radiometric Calibration for Backscatter Intensity of Hyperspectral LiDAR Caused by Incident Angle Effect," *Sensors*, vol. 21, no. 8, p. 2792, 2021. https://doi.org/10.3390/s21092960
[20] S. Kaasalainen, M. Åkerblom, O. Nevalainen, T. Hakala, and M. Kaasalainen, "Uncertainty in multispectral lidar signals caused by incidence angle effects," *Interface Focus*, vol. 8, no. 2, p. 20170041, 2018. https://doi.org/10.1098/rsfs.2017.0033
[21] T. Yang, Y. Li, C. Zhao, and L. Sun, "3D ToF LiDAR in mobile robotics: A review," *arXiv preprint arXiv:2205.05658*, 2022. https://doi.org/10.48550/arXiv.2202.11025
[22] J. Zhou, "A Review of LiDAR Sensor Technologies for Perception in Automated Driving," *Academic Journal of Science and Technology*, vol. 6, no. 11, pp. 42–58, 2022. https://doi.org/10.54097/ajst.v3i3.2993
[23] A. Haider et al., "Performance Evaluation of MEMS-Based Automotive LiDAR Sensor and Its Simulation Model as per ASTM E3125-17 Standard," *Sensors*, vol. 23, no. 6, p. 2931, 2023. https://doi.org/10.3390/s23063113
[24] C. Linnhoff, K. Hofrichter, L. Elster, P. Rosenberger, and H. Winner, "Measuring the Influence of Environmental Conditions on Automotive Lidar Sensors," *Sensors*, vol. 22, no. 14, p. 5385, Jul. 2022. https://doi.org/10.3390/s22145266
[25] M. Kutila, P. Pyykonen, W. Ritter, O. Sawade, and B. Schaufele, "Automotive LIDAR Sensor Development Scenarios for Harsh Weather Conditions," in *Proc. 2016 IEEE 19th Int. Conf. Intelligent Transportation Systems (ITSC)*, Nov. 2016, pp. 265–270. https://doi.org/10.1109/ITSC.2016.7795565
[26] R. Meshcheryakov et al., "A Probabilistic Approach to Estimating Allowed SNR Values for Automotive LiDARs in Smart Cities under Various External Influences," *Sensors*, vol. 22, no. 1, p. 111, Jan. 2022. https://doi.org/10.3390/s22020609
[27] T. Raj, F. H. Hashim, A. B. Huddin, M. F. Ibrahim, and A. Hussain, "A Survey on LiDAR Scanning Mechanisms," *Electronics*, vol. 9, no. 4, p. 700, Apr. 2020. https://doi.org/10.3390/electronics9050741
[28] M. K. Tsai, D. J. Findley, and C. M. Cunningham, "Infrastructure Investment Protection with LiDAR," *Sensors*, vol. 12, no. 12, pp. 17284–17301, 2012. https://rosap.ntl.bts.gov/view/dot/25075
[29] K. Williams, M. Olsen, G. Roe, and C. Glennie, "Synthesis of Transportation Applications of Mobile LIDAR," *Remote Sensing*, vol. 5, no. 9, pp. 4652–4692, Sept. 2013. hthttps://doi.org/10.3390/rs5094652
[30] J. Hudson and E. Jones, "Spacecraft Coatings Optimizing LiDAR Debris Tracking and Light Pollution Impacts," *arXiv preprint arXiv:2311.10323*, 2023. https://doi.org/10.48550/arXiv.2311.13108
[31] Z. Nan, W. Tao, H. Zhao, and N. Lv, "A Fast Laser Adjustment-Based Laser Triangulation Displacement Sensor for Dynamic Measurement of a Dispensing Robot," *Applied Sciences*, vol. 10, no. 21, p. 7412, Oct. 2020. https://doi.org/10.3390/app10217412
[32] Y. Bae, "An Improved Measurement Method for the Strength of Radiation of Reflective Beam in an Industrial Optical Sensor Based on Laser Displacement Meter," *Sensors*, vol. 16, no. 5, p. 752, May 2016. https://doi.org/10.3390/s16050752
[33] L. Cheng et al., "Registration of Laser Scanning Point Clouds: A Review," *Sensors*, vol. 18, no. 5, p. 1641, May 2018. https://doi.org/10.3390/s18051641
[34] Y. Liu et al., "Dynamic Validation of Calibration Accuracy and Structural Robustness of a Multi-Sensor Mobile Robot," *Sensors*, vol. 24, no. 12, p. 3589, Jun. 2024. https://doi.org/10.3390/s24123589

Received: April 28, 2025. Reviewed: May 27, 2025. Accepted: May 28, 2025. *© 2025 Arsen Abdrakhmanov and Lyazzat Ilipbayeva. All rights reserved.*

SECTION III
Mathematics with Applied Aspects

This section includes applied mathematics research with a focus on modeling, optimization, and analysis of computational and engineering systems.

Article

Development and optimization of physics-informed neural networks for solving partial differential equations

Batyr Sharimbayev* [1], Shirali Kadyrov [2], and Aleksei Kavokin [1]

[1]Department of Mathematics and Natural Sciences, Almaty, Kazakhstan
[2]Department of General Education, New Uzbekistan University, Tashkent, Uzbekistan

DOI: 10.47344/2ekq1942

Abstract

This study investigates the application of physics-informed neural networks (PINNs) for solving Poisson equations in both 1D and 2D domains and compares them with finite difference method. Additionally, the study explores the capability of multi-task learning with PINNs, where the network not only predicts the solution but also estimates unknown parameters. In the case of a second-order differential equation with a varying coefficient, PINNs successfully approximated both the source term and the varying coefficient while achieving low training loss. The model demonstrated excellent generalization capabilities and accurate reconstruction of the underlying system parameters, showing the potential of PINNs in complex physical simulations.

Keywords: numerical analysis, multi-task learning, deep learning, PINNs, FDM

I. Introduction

PDEs form a crucial backbone in the understanding and modeling of various real-world problems. In simple words, PDEs describe how something changes through time and space based on very well-defined mathematical rules. The central issue with PDEs is determining whether these rules suffice to guarantee a unique solution to the problem [1] [2].

In most real-life cases, it is impossible to find the exact solutions for such complex equations involving PDEs. This gave rise to the development of numerical methods for estimating solutions. Some of the popular methods include the finite element method, finite difference method, finite volume method, and spectral element method. Out of these, FEM is the most advanced with strong mathematical support for ensuring accurate results, stability, and error control. FEM solvers often employ efficient techniques such as sparse linear systems or iterative methods, which make them suitable for many practical problems [3].

Most recently, new techniques have emerged as a result of the growth of deep learning which aid in the resolution of PDEs. One of these is Physics-Informed Neural Networks (PINNs). PINNs are neural networks which, by means of their loss function,

Email: batyr.sharimbayev@sdu.edu.kz ORCID: 0009-0006-3323-231X
Email: sh.kadyrov@newuu.uz ORCID: 0000-0002-8352-2597
Email: aleksei.kavokin@sdu.edu.kz ORCID: 0009-0000-1931-4518

Received: February 13, 2025. Reviewed: March 20, 2025. Accepted: April 30, 2025. *© 2025 Batyr Sharimbayev, Shirali Kadyrov, Aleksei Kavokin. All rights reserved.*

encode the physical laws described by differential equations to drive the learning process towards solutions that better respect the underlying physics. PINNs can approximate solutions to partial differential equations and ordinary differential equations and solve inverse problems, e.g., estimating model parameters from scarce data. They also happen to be tremendously useful for solving various kinds of PDEs because of their simplicity in implementation and direct incorporation of the underlying physics within the learning itself.

Despite their adoption, deep learning algorithms, including PINNs, still suffer from high computational costs, complex optimization, and weak theoretical foundations. This study aims to achieve two main objectives. First, it aims to rigorously evaluate the performance of PINNs in solving PDEs by comparing their accuracy, convergence, and computational efficiency to traditional numerical methods, particularly the Finite Difference Method (FDM). By applying PINNs to 1D and 2D Poisson equations, the study assesses their accuracy, error metrics, and computational demands while examining their robustness across different boundary conditions and domain sizes. Second, the research explores the integration of multi-task learning (MTL) [5] within the PINN framework to enhance its ability to solve ODE while simultaneously predicting related physical quantities, such as source terms and varying coefficients. By leveraging shared information across tasks, the study aims to improve the generalization capabilities of PINNs and develop a unified model capable of solving complex multi-dimensional problems in a single optimization process. We build upon our prior work [4] and provide additional analysis.

II. Related Works

The increasing use of neural networks has prompted researchers to utilize a variety of techniques in deep learning for solving mathematical equations. Raissi et al. [6] present PINNs as a broad framework in deep learning to solve nonlinear forward and inverse partial differential equations (PDEs). With integration of basic principles from physics, PINNs provide a mathematically strong and data-effective approach to solving complex spatio-temporal problems using continuous-time and discrete-time Runge–Kutta methods. This method has been validated by extensive testing across a variety of applications in PDEs, ranging from fluid dynamics to quantum mechanics.

Ryck et al. [7] present a rigorous numerical analysis of PINNs, systematically classifying errors into approximation, generalization, and training errors. They critically examine how PDE characteristics and domain dimensionality influence accuracy, identifying training error as a significant constraint on PINN performance. Their findings underscore the importance of solution regularity and stability in ensuring the reliability of PINN-based computations. Hu et al. [8] investigate the application of PINNs in computational solid mechanics, addressing limitations associated with sparse, noisy, and high-dimensional data. By integrating prior physical knowledge, PINNs enhance model generalizability, enforce physical consistency, and improve computational efficiency. The study provides a comprehensive review of PINN architectures, algorithmic advancements, and their implementation in constitutive modeling, damage evaluation, and inverse problem-solving. Cuomo et al. [9] provide a thorough review of PINNs and explore their use in solving a wide variety of partial differential equation (PDE) problems, both fractional and stochastic types. The paper also discusses developments like Physics-Constrained Neural Networks (PCNNs) and variational hp-VPINNs and touches upon optimization algorithms, network architectures, and loss function setup. Though PINNs have already shown great promise, the paper identifies open problems to be solved in order to enhance their reliability and usability.

Grossmann et al. [10], past comparisons between PINNs and other numerical methods have shown that each approach has strengths and weaknesses. However, the paper does not clearly explain these differences, making it harder to assess how competitive PINNs are for different types of problems. While the study presents specific cases where PINNs struggle or perform well, it lacks a systematic discussion of key factors such as computational cost, accuracy, stability, and problem structure. Without this detailed comparison, it is difficult to determine in which scenarios PINNs might be advantageous or where traditional methods like FEM remain superior. The findings suggest that FEM remains the more reliable method, particularly for high dimensional PDEs. In light of the findings from previous works, we conduct a comprehensive comparison between the well-established FDM and the state-of-the-art PINNs in the context of solving non-complex PDEs, where FDM has demonstrated superior accuracy and efficiency. The FDM, known for its robustness and reliability, has long been a standard technique for numerically solving PDEs with relatively simple geometries and boundary conditions. On the other hand, PINNs represent an emerging approach in deep learning. Additionally, we extend our comparison by framing the development of the PINN method within the context of multi-task learning (MTL). In this setting, the PINN is treated as an MTL model, enabling it to simultaneously learn from multiple related tasks. This approach not only facilitates a more efficient learning process but also allows for the adaptation of the PINN to a broader range of problems by sharing knowledge between tasks.

Received: February 13, 2025. Reviewed: March 20, 2025. Accepted: April 30, 2025. *© 2025 Batyr Sharimbayev, Shirali Kadyrov, Aleksei Kavokin. All rights reserved.*

III. Mathematical Background

In this section, we explain the mathematical background of the methods used in this work.

A. Poisson Equation

The Poisson equation is one of the most basic PDEs, which results in a potential field created by a given source. It provides significant applications to physics and engineering, modeling electrostatics, heat conduction, and fluid dynamics among many others [11] [12].

In one dimension, the Poisson equation is given by:

$$\frac{d^2u(x)}{dx^2} = f(x) \tag{1}$$

where $u(x)$ is the unknown function, $f(x)$ is a source term.

In such cases, the domain is normally an interval $[a, b]$ and the boundary conditions are often given at both ends, say in the following form of Dirichlet boundary conditions:

$$u(a) = u_0, \quad u(b) = u_1$$

In two dimensions, the Poisson equation generalizes to:

$$\nabla^2 u(x, y) = f(x, y) \tag{2}$$

where ∇^2 is the 2D Laplacian operator, $u(x, y)$ represents the unknown function, and $f(x, y)$ is called the source term. In two dimensions, it can normally be a rectangular region or even more complicated geometry. On edges, boundary conditions are imposed. There may also be **Dirichlet** boundary conditions on these. A typical **Dirichlet** boundary condition takes the following form: $u(x, y) = g(x, y)$ on the boundary of the domain.

The accuracy of the solution approximations for the 1D and 2D Poisson equations is evaluated using the L_2 norm and relative error. The L_2 norm of a vector $v = [v_1, v_2, \ldots, v_n]$ is defined as:

$$\|v\|_2 = \sqrt{\sum_{i=1}^{n} v_i^2}.$$

This norm gives the size of the magnitude of a vector in Euclidean space. To verify these results for accuracy, we calculate the L_2 relative error between the computed solution $\hat{u}$ and the exact solution u defined as:

$$L_2 = \frac{\|\hat{u} - u\|_2}{\|u\|_2}. \tag{3}$$

B. Physics-Informed Neural Networks

Let $u_\theta(x, y)$ denote the neural network approximation of $u(x, y)$ where θ is used to denote the parameters of the network. The network architecture considered here is a fully connected feed-forward neural network, such that the application of activation functions is performed layer by layer. It can be written as:

$$u_\theta(x, y) = \mathrm{NN}_\theta(x, y)$$

The network layers are defined as follows:

$$x^{(i+1)} = \sigma\left(w^{(i)}x^{(i)} + b^{(i)}\right)$$

for each hidden layer, and the output layer is linear. The residual of the PDE is defined as:

$$R(x, y) = \frac{\partial^2 u_\theta}{\partial x^2} + \frac{\partial^2 u_\theta}{\partial y^2} - f(x, y)$$

where $f(x, y)$ is the source term. The PDE loss is computed as the mean squared error of the residual over a set of domain points (x_i, y_i):

Received: February 13, 2025. Reviewed: March 20, 2025. Accepted: April 30, 2025. © 2025 Batyr Sharimbayev, Shirali Kadyrov, Aleksei Kavokin. All rights reserved.

$$L_{\text{PDE}} = \frac{1}{N} \sum_{i=1}^{N} R(x_i, y_i)^2 \tag{4}$$

The boundary conditions loss is computed similarly, where the loss for boundary points (x_j, y_j) is given by:

$$L_{\text{BC}} = \frac{1}{M} \sum_{j=1}^{M} (u_\theta(x_j, y_j) - u(x_j, y_j))^2 \tag{5}$$

The total loss function combines the PDE residual loss and the boundary conditions loss:

$$L(\theta) = L_{\text{PDE}} + L_{\text{BC}} \tag{6}$$

C. Finite Difference Method (FDM)

The FDM is a numerical method for the solution of PDEs that is based on discretizing their solutions on a lattice of points that discretize the domain. Based on this concept, the FDM relies upon approximating the derivatives of an unknown function in terms of finite differences that replace the PDE by a set of algebraic equations.

First, consider the one-dimensional Poisson equation:

$$\frac{d^2 u(x)}{dx^2} = f(x)$$

Using a finite difference scheme, the second derivative can be approximated by the following way:

$$\frac{d^2 u(x)}{dx^2} \approx \frac{u(x + \Delta x) - 2u(x) + u(x - \Delta x)}{(\Delta x)^2} \tag{7}$$

where Δx is the step size and $u(x)$ is the unknown function at the grid points. By discretizing the domain $[a, b]$ with a grid of points, we convert the continuous PDE into a system of algebraic equations which can be solved numerically.

In two dimensions, the Poisson equation is given by:

$$\nabla^2 u(x, y) = f(x, y)$$

where ∇^2 is the Laplacian operator:

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}.$$

The second derivatives in the x and y directions are approximated by finite differences. The Laplacian operator in 2D is approximated as:

$$\nabla^2 u(x, y) \approx \frac{u(x + \Delta x, y) + u(x - \Delta x, y) - 2u(x, y)}{(\Delta x)^2} + \frac{u(x, y + \Delta y) + u(x, y - \Delta y) - 2u(x, y)}{(\Delta y)^2} \tag{8}$$

where Δx and Δy are increments in steps in directions x and y, respectively.

The FDM discretizes both the domain for solving both the 1D and 2D Poisson equation and then solves for resulting algebraic equations. Iterating through each grid point, $u(x)$ or $u(x, y)$ values would be computed based on the finite difference approximation offered by the PDE. Boundary conditions would be specified along the domain's boundary in terms of either Dirichlet conditions.

Received: February 13, 2025. Reviewed: March 20, 2025. Accepted: April 30, 2025. © 2025 Batyr Sharimbayev, Shirali Kadyrov, Aleksei Kavokin. All rights reserved.

IV. Methods

This paper focuses on constructing and improving PINNs for solving the 1D and 2D Poisson equations and comparing them with the FDM.

To begin with, FDM is the first approach we had to study to solve the Poisson equation. The domain of the 1D example is split into a lattice grid, and the equation is solved through second-derivative approximation. The same procedure is applied in 2D, in which both spatial dimensions are discretized. The numerical solution is obtained and compared to the calculated exact solutions in order to check accuracy using the L_2 relative error.

At this stage, the PINN is implemented to solve the equations. The neural network is trained by minimizing a loss function that comprises the Poisson equation and the boundary conditions. The training is therefore carried out in two steps: firstly, with the Adam optimizer to fit the parameters of the model, and subsequently with the L-BFGS method for fine-tuning.

Latin Hypercube Sampling, in both 1D and 2D scenarios, provides points for model analyses. The network is trained to provide a set of boundary conditions. The performance of the model is determined by comparing the output of the neural network with a known solution using the relative L_2 norm.

Ultimately, we applied the FDM to a thermal problem involving a second-order ordinary differential equation (ODE). We split the spatial domain into parts, provided intervals, and boundary conditions at both ends. The problem was solved iteratively until the difference between subsequent values was less than some small predefined threshold. The material properties and source term were represented as functions of spatial position. The PINN solution was compared to the one obtained using multivariate interpolation to assess its accuracy. Then, the forward and inverse problem was solved using a PINN structure. The PINN was based on a feedforward neural network with hidden layers that were trained to estimate temperature and the source term via a learned loss function. The model was validated against FDM and observational data. The code was written in Python and is available on GitHub: https://github.com/hardkazakh/pinn-vs-fdm

V. Experimental Results

In this section, we will solve 1D and 2D Poisson equations using FDM and PINNs. We will also use the PINNs approach to multi-task learning.

A. 1D Poisson equation

Let us investigate the 1D Poisson equation defined as:

$$\frac{d^2u(x)}{dx^2} = 16x^7e^{-x^4} - 20x^3e^{-x^4}, \quad x \in [0, 1], \tag{9}$$

with Dirichlet boundary conditions:

$$u(0) = 0, \quad u(1) = e^{-1}.$$

In Section III-C, we explained that the first step to solving PDEs with the FDM is to rewrite the equation in a weak form. We already did this for the Poisson equation. The next step is to create a mesh. This is like breaking the interval $[0, 1]$ into small pieces, called cells. The number of cells is 512. More cells mean the grid is finer, which gives a more accurate solution, but it also takes more time and computing power.

For solving the 1D Poisson equation using PINNs, there are three design parameters that we need to specify before training. The first step is choosing a loss function. Following the vanilla PINNs approach, we evaluate the goodness of the solution using the discretised mean squared error over the PDE, boundary, and initial conditions.

The second design parameter is the neural network architecture, that is, the type of neural network, the activation function, and the number of hidden layers and nodes. For the 1D Poisson case, we train feed-forward dense neural networks with $tanh$ as the activation function. We use the result on the architecture of [20, 20, 20, 1].

The approximations of the 1D Poisson equation solution using FDM and PINNs are compared to the exact solution on a $[0, 1]$ interval with 512 points. Fig 1 shows the exact solution and the approximations. One PINN setup, with only one hidden layer and one node, performs poorly and fails to satisfy the boundary conditions. All approximations are very close to the exact solution.

For the 1D Poisson equation, the relative error for the FDM is calculated to be 7.26×10^{-8}, while the relative error for the PINN approach is 5.63×10^{-6}. These results show that FDM provides a more accurate approximation of the solution compared to PINNs in the 1D case.

Received: February 13, 2025. Reviewed: March 20, 2025. Accepted: April 30, 2025. © 2025 Batyr Sharimbayev, Shirali Kadyrov, Aleksei Kavokin. All rights reserved.

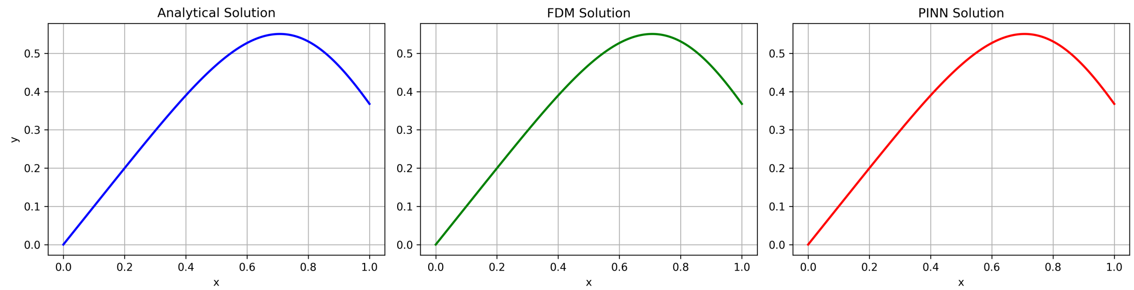


Fig. 1. Comparison of solutions: 1-Exact, 2-FDM, and 3-PINN

B. 2D Poisson equation

Let us now investigate 2D Poisson equation defined as:

$$\frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial y^2} = 2\left(x^4(3y-2) + x^3(4-6y) + x^2\right). \tag{10}$$

The boundary conditions are:

$$u(x,0) = u(x,1) = u(0,y) = u(1,y) = 0.$$

The analytical solution of the equation is:

$$u(x,y) = (x-1)^2 y(y-1)^2 x^2.$$

The approximations of the 2D Poisson equation solution using FDM and PINNs are compared to the exact solution on the $[0,1] \times [0,1]$ domain, discretized with a 1000×1000 grid. The neural networks are trained using the $\tanh$ activation function with an architecture of $[60, 60, 60, 1]$.

Fig 2 shows the analytical and approximate solutions of the 2D Poisson equation. For the 2D Poisson equation, the L_2 relative error for the FDM approximation is 2.21×10^{-4}, while for the PINN, it is 6.01×10^{-3}. Again, the FDM method shows a significantly lower error compared to the PINN, indicating that FDM achieves a more precise solution in both 1D and 2D cases.

These error analyses again confirm the accuracy of the FDM approach in solving Poisson equations, especially in comparison with PINNs, which showed higher relative errors in both 1D and 2D problems. However, in higher-order equations, PINN can give better results. We consider this in our future research.

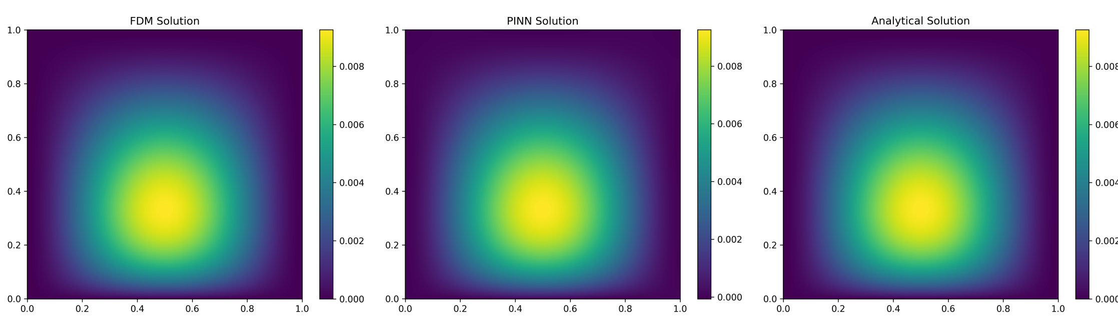


Fig. 2. Comparison of solutions: 1-PINN, 2-Exact, and 3-FDM.

Received: February 13, 2025. Reviewed: March 20, 2025. Accepted: April 30, 2025. *© 2025 Batyr Sharimbayev, Shirali Kadyrov, Aleksei Kavokin. All rights reserved.*

C. PINNs approach for multi-task learning (MTL)

PINNs embed both forward and inverse problems under one framework by embedding the data and physical laws in the loss function of the neural network. The network jointly predicts the forward solution of a PDE and estimates unknown parameters or inputs (inverse problem) through the optimization of one combined loss function that considers residual of PDE, boundary conditions, and discrepancies between model predictions and observations. This capability of handling both tasks together makes PINNs highly effective at solving problems that involve solution estimation and identification of parameters.

The problem is defined by the following second-order differential equation:

$$\frac{d^2U(x)}{dx^2} - a(x)U(x) = Q(x), \quad 0 < x < L, \tag{11}$$

where $Q(x)$ represents the source term and $a(x)$ describes the varying coefficient. Specifically, the source term is given by:

$$Q(x) = 1 + b_1 \sin(w_1 x)$$

and the varying coefficient is defined as:

$$a(x) = b_1 + \frac{x}{1+x^2}.$$

The boundary conditions for the differential equation are given by: $U(0) = 1, U(L) = 3$. The problem now involves the solution of this second-order differential equation along with the boundary conditions shown above. $Q(x)$, the source function, incorporates a sine function that could model some periodic influence in the system. This coefficient, $a(x)$, depends on the position x in the domain; therefore, this makes the equation more complicated, introducing spatial dependence in the solution.

The aim of Experiment 1 is to approximate both the source term $Q(x)$ and the solution $U(x)$ for the given differential equation. PINN simultaneously predicts the solution $U(x)$ while reconstructing the source term $Q(x)$ using the varying coefficient $a(x)$ as part of the system.

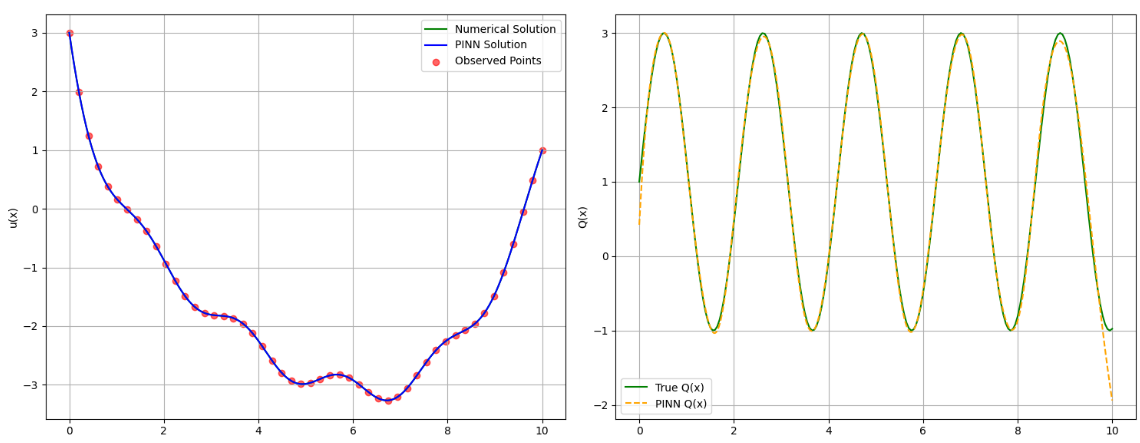


Fig. 3. Visualization of FDM Solution and varying coefficient of a(x)

The PINN model was trained for 40,000 epochs. Initially, the loss function was quite large, but it gradually decreased as the model learned. By the end of the training, the loss had reduced to 1.87×10^{-2}, indicating that the model had learned the underlying physics of the problem. The predictions made by the PINN for the differential equation solution were then analyzed and aligned with the FDM results. These predictions were accurate not only at the observed points but also at unobserved locations, highlighting the PINN's ability to generalize well across the entire domain. Fig 3 presents the visualization of the PINN's predicted temperature distribution and the corresponding predicted source term.

Received: February 13, 2025. Reviewed: March 20, 2025. Accepted: April 30, 2025. © 2025 Batyr Sharimbayev, Shirali Kadyrov, Aleksei Kavokin. All rights reserved.

The aim of Experiment 2 is to approximate both the varying coefficient $a(x)$ and the solution $U(x)$. By incorporating $a(x)$ as an unknown parameter in the system, the model predicts the solution $U(x)$ while reconstructing $a(x)$ from the given data.

The PINN model was trained for 40,000 epochs. Initially, the loss function exhibited high values, but with training, it gradually decreased as the model captured the intricate relationships within the system. By the end of the training, the loss had reduced to 3.1563×10^{-2}, demonstrating that the model effectively learned both the solution and the coefficient.

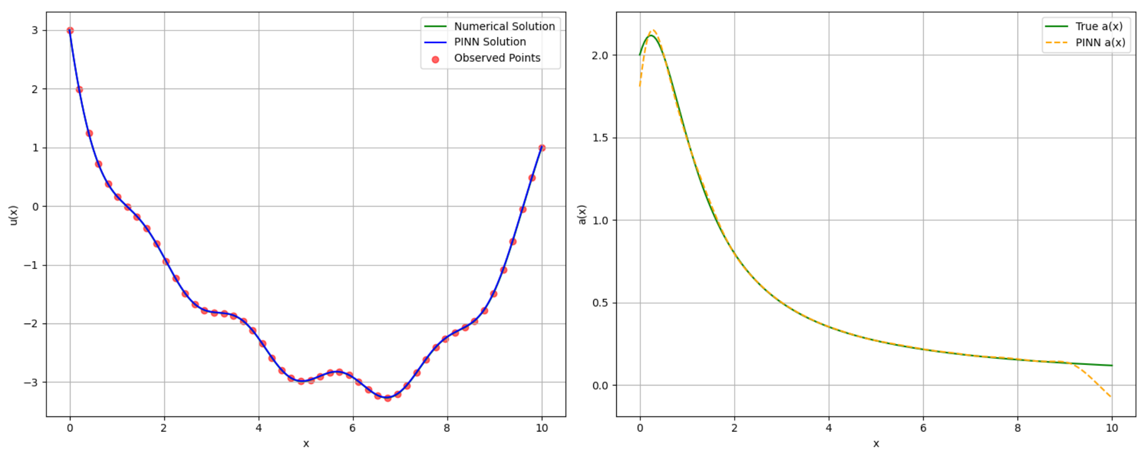


Fig. 4. Visualization of PINN Solution and prediction of a(x)

Fig 4 illustrates the visualization of the PINN's predictions for the temperature distribution and the reconstructed varying coefficient. The predicted solution from the PINN closely aligns with the numerical solution obtained using FDM.

VI. Discussion & Conclusion

In this paper, we have discussed using PINNs and FDM to solve the problem in both 1D and 2D. From our comparison, it was illustrated that even though PINNs offer a flexible and data-driven approach, traditional numerical methods such as FDM offer improved accuracy in such a case.

For the 1D Poisson equation, FDM's relative error was calculated as 7.26×10^{-8} and PINN's relative error was calculated as 5.63×10^{-6}. In a similar manner, for the 2D Poisson equation, FDM's relative error was calculated as 2.21×10^{-4}, which was much lesser than PINN's error value, which was 6.01×10^{-3}. These findings establish that FDM gives a better approximation in problems involving Poisson.

Additionally, we have explained multi-task learning with PINNs. We have witnessed that PINN reconstructed variable $a(x)$ and source term $Q(x)$ with decreasing loss to 1.87×10^{-2} and 3.15×10^{-2}, respectively, in 40,000 epochs. The PINN model had strong generalization capability and predicted with high accuracy even at unseen locations. Regarding the computational complexity, PINN's training to solve the 2D Poisson equation was very time-consuming. This reflects high computational cost in high-dimensional PDEs, whose practicality and efficiency depend greatly upon high-end GPU access. The longer time to train reflects a major disadvantage of PINNs compared to standard numerical methods in problems requiring quick and scalable solutions. We believe that PINNs have more errors than FDM because it has problems in optimizations and it has problems in fine solution feature capture.

Overall, our findings indicate that while PINNs do not presently outperform FDM in terms of accuracy for non-complex equations because they can be flexible and can perform multi-task learning, they provide a reasonable alternative to complex PDEs when conventional numerical methods have limitations.

In the present study, our aim will be to improve PINN's performance and efficiency with more advanced neural network architectures and learning techniques to solve other PDEs. One direction would be to design hybrid algorithms by combining PINNs with traditional numerical methods such as FDM. With FDM's high-precision performance in structured grid regions and PINNs's adaptability in unstructured or sparsity regions, a more effective and robust framework can be established. This hybrid

Received: February 13, 2025. Reviewed: March 20, 2025. Accepted: April 30, 2025. *© 2025 Batyr Sharimbayev, Shirali Kadyrov, Aleksei Kavokin. All rights reserved.*

method can rectify PINN's limitation in high-precision attainment without sacrificing PINN's ability to solve inverse problems and multi-task learning.

VII. Acknowledgment

"This research was funded by the Ministry of Science and Higher Education of the Republic of Kazakhstan within the framework of project AP23487777."

References

[1] S.L. Brunton, "Promising directions of machine learning for partial differential equations," Nature Computational Science, 2024.
[2] J. Blechschmidt, "Three ways to solve partial differential equations with neural networks—A review," GAMM-Mitteilungen, 2021.
[3] A. Björck, "Numerical methods for least squares problems," SIAM, 2024.
[4] B. Sharimbayev, "Development and optimization of physics-informed neural networks for solving partial differential equations," arXiv preprint arXiv:2502.02599, 2025.
[5] S. Ruder, "An overview of multi-task learning in deep neural networks," arXiv preprint arXiv:1706.05098, 2017.
[6] M. Raissi, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," Journal of Computational physics, 2019.
[7] T. De Ryck, "Numerical analysis of physics-informed neural networks and related models in physics-informed machine learning," Acta Numerica, 2024.
[8] H. Hu, "Physics-informed Neural Networks (PINN) for computational solid mechanics: Numerical frameworks and applications," Thin-Walled Structures, 2024.
[9] S. Cuomo, "Scientific machine learning through physics–informed neural networks: Where we are and what's next," Journal of Scientific Computing, 2022.
[10] T.G. Grossmann, "Can physics-informed neural networks beat the finite element method?," IMA Journal of Applied Mathematics, 2024.
[11] W. Hackbusch, "Elliptic differential equations: theory and numerical treatment," Springer, 2017.
[12] T. Matsuura, "Numerical solutions of the Poisson equation," Applicable Analysis, 2004.

Received: February 13, 2025. Reviewed: March 20, 2025. Accepted: April 30, 2025. *© 2025 Batyr Sharimbayev, Shirali Kadyrov, Aleksei Kavokin. All rights reserved.*

Article

Numerical Fractal Analysis of Exceptional Sets in the Lehner Expansion

Symbat Duisen *1, Aiken Kazin 2, and Shirali Kadyrov 3

[1]Department of Mathematics, SDU University, Almaty, Kazakhstan
[2]Department of Mathematics, SDU University, Almaty, Kazakhstan
[3]Department of Education, New Uzbekistan University, Tashkent, Uzbekistan

DOI: 10.47344/q46xe557

Abstract

This paper examines how the average value of the sequence b_n in the Lehner expansion of a real number x influences its box dimension. Our primary objective is to analyze how variations in the average of b_n impact the box dimension, which serves as a measure of the complexity of the sequence. Using the box-counting method, we numerically estimate the box dimension and explore its relationship with the fractal nature of Lehner expansions.

Keywords: Regular continued fraction, Lehner expansion, semi-regular continued fraction,fractal dimension, box dimension.

I. Introduction

Continued fractions [1]–[4] play important role in number theory,this way of writing numbers is very useful in number theory because it helps us understand how well we can approximate real numbers using fractions.A semi-regular continued fraction [7] is a special type of continued fraction that extends classical regular continued fractions by allowing a broader set of partial quotients while keeping important mathematical properties.One such semi-regular expansion is the Lehner continued fraction, which has been investigated for its unique convergence behavior and number-theoretic significance (Lehner, 1949).
Any irrational number $x \in [1, 2]$ has a unique Lehner expansion of the form

$$b_0 + \cfrac{\sigma_1}{b_1 + \cfrac{\sigma_2}{b_2 + \dots + \cfrac{\sigma_n}{b_n + \dots}}} = [b_0; \sigma_1/b_1, \sigma_2/b_2, \dots, \sigma_n/b_n, \dots]$$

*Corresponding author: 231105007@sdu.edu.kz
Email: 231105007@sdu.edu.kz ORCID: 0009-0009-7514-0503
Email: Aiken.Kazin@sdu.edu.kz ORCID: 0000-0002-9658-9723
Email: sh.kadyrov@newuu.uz ORCID: 0000-0002-8352-2597

Received: February 17, 2025. Reviewed: March 26, 2025. Accepted: March 29, 2025. *© 2025 Symbat Duisen, Aiken Kazin, Shirali Kadyrov. All rights reserved.*

(1)

where (b_i, σ_{i+1}) equals (1,1) or (2,-1).We call these continued fractions Lehner fractions or Lehner expansions. Every rational number has two different finite Lehner expansions.
Lehner expansions can be found using this map $L : [1, 2) \to [1, 2)$, which is defined as follows.

$$Lx := \begin{cases} \frac{1}{2-x}, & 1 \leq x < \frac{3}{2}, \\ \frac{1}{x-1}, & \frac{3}{2} \leq x < 2. \end{cases}$$

Notice that in this expansion for $x \in [1, 2)$ one has

$$(b_i, \sigma_{i+1}) = \begin{cases} (1, 1), & \text{if } L^i(x) \in \left[\frac{3}{2}, 2\right), \\ (2, -1), & \text{if } L^i(x) \in \left[1, \frac{3}{2}\right). \end{cases}$$

Lehner expansions are a type of semi-regular continued fraction. A semi-regular continued fraction can be either a finite or an infinite fraction.

The study of exceptional sets in continued fractions has been a focus of recent research, exploring their fractal properties and Hausdorff dimensions. Fang et al. [8] determined the Hausdorff dimension of a set related to the growth rate of continued fraction coefficients. Kazin and Kadyrov [5] extended Good's work on fractal geometry in continued fractions, establishing new bounds on Hausdorff dimensions of level sets formed by restricting partial quotients. Bakhtawar et al. [9] calculated the Hausdorff dimension of a set defined by conditions on ratios of consecutive continued fraction coefficients, contributing to the metrical theory of continued fractions. While not directly addressing continued fractions, Parsell and Wooley [10] investigated exceptional sets for Diophantine inequalities, showing that under certain conditions, the measure of the exceptional set in an interval is bounded. These studies collectively advance our understanding of exceptional sets in number theory and their geometric properties fractal properties of these sets, particularly their Hausdorff and box dimensions, have been the subject of extensive research [11]. For regular continued fractions, the dimension of sets defined by constraints on their partial quotients has been thoroughly examined [4], [12]. However, for semi-regular expansions such as the Lehner continued fraction, a comprehensive understanding of these exceptional sets remains incomplete.

Fractal dimension measures how completely a fractal fills space as one zooms in on finer scales. Unlike traditional Euclidean dimensions, which take integer values (e.g., a line has dimension 1, a plane has dimension 2), fractal dimensions can be non-integer, reflecting the complexity and self-similarity of fractal structures. It quantifies how detail in a pattern changes with the scale at which it is measured, making it useful for characterizing irregular shapes in nature, such as coastlines, clouds, and turbulent flows. The Hausdorff dimension and box dimension are both types of fractal dimensions. For more information on how these various notions of dimension are related, we refer to [6]. In this paper, we focus only on the box dimension.The box dimension of set S is defined as

$$\dim_B(S) = \lim_{\delta \to 0} \frac{\log N(\delta)}{-\log \delta},$$

where $N(\delta)$ is the number of boxes size δ required to cover set S.

If this limit exists. This dimension captures how the number of covering elements scales with their size and provides a practical way to estimate fractal complexity.

Theorem [7,theorem4] *For almost all real numbers $x \in (1, 2)$, we have that their Lehner expansions*

$$x = [b_0; \sigma_1/b_1, \sigma_2/b_2, \cdots, \sigma_n/b_n, \cdots]$$

$$\lim_{n\to\infty} \frac{b_1 + b_2 + \cdots + b_n}{n} = 2.$$

In this work, we focus on the set of numbers for which the sequence of partial quotients b_n in the Lehner expansion exhibits an anomalous growth pattern, specifically cases where the long-term average deviates from its expected limit. Such deviations are known to correspond to fractal-like structures, whose complexity can be quantified using box dimension [13].Our objective is to determine how the box dimension of these exceptional sets depends on the asymptotic behavior of b_n, extending results known for regular continued fractions []. We consider those real numbers x for which the above limit is not equal to 2. By the theorem we

Received: February 17, 2025. Reviewed: March 26, 2025. Accepted: March 29, 2025. *© 2025 Symbat Duisen, Aiken Kazin, Shirali Kadyrov. All rights reserved.*

know that this set has Lebesgue measure zero. However, it may have a complex structure from a fractal geometry point of view. To understand, for any $\epsilon > 0$ we define sets

$$S(\epsilon, c) = \left\{ x \in (1,2) : \lim_{n\to\infty} \frac{b_1 + b_2 + \cdots + b_n}{n} \in (c-\epsilon, c+\epsilon) \right\}.$$

Our research question is to numerically investigate how box dimension of $S(\epsilon, c)$ depends on ϵ. For the definition of box dimension, see the next section.

To achieve this, we employ a computational approach based on binary word representations, adapting established methods from multifractal analysis (Barreira & Schmeling, 2000). By numerically estimating the box dimension for different classes of exceptional sets, we provide new insights into the geometric complexity of Lehner continued fraction expansions. Our findings contribute to the broader understanding of fractal structures in number theory and highlight the rich interplay between continued fractions and dynamical systems.

The structure of the paper is as follows. In Section 2, we introduce the mathematical framework of continued Lehner fractions and review key definitions. Section 3 describes the methodology for computing the box dimension, detailing the binary word-based approach. Section 4 presents numerical results and discusses the implications of our findings. Finally, in Section 5, we summarize our conclusions and suggest directions for future research.

II. Methodology

The box dimension of set S is defined as

$$\dim_B(S) = \lim_{\delta\to 0} \frac{\log N(\delta)}{-\log \delta},$$

where $N(\delta)$ is the number of boxes size δ required to cover set S. This dimension characterizes the fractal scaling behavior of the set as the solution δ decreases.

To estimate the box dimension numerically, we analyze the scaling behavior of unique truncated binary words derived from points in a given set. Each point in the set is mapped to a binary expansion with fixed precision by repeatedly multiplying the point by two. If the result is at least one, '1' is appended to the binary string, and one is subtracted from the point; otherwise, '0' is appended. This process is repeated for the desired precision; see Fig.1.

```
Input:
   x - a real number in [0,1)
   precision - number of bits in the binary expansion
Output:
   binary_expansion - a string representing the binary expansion
Initialize binary_expansion as an empty string
For i from 1 to precision do
  x ← 2x
  If x ≥ 1 then:
    Append "1" to binary_expansion
    x ← x − 1
  Else:
    Append "0" to binary_expansion
Return binary_expansion
```

Fig. 1. Algorithm to compute the binary expansion of real numbers

To estimate the complexity of the given set, each binary expansion is truncated to a fixed word length, meaning that only the first few digits of the binary representation are considered. For each chosen word length, the number of unique binary words

Received: February 17, 2025. Reviewed: March 26, 2025. Accepted: March 29, 2025. © 2025 Symbat Duisen, Aiken Kazin, Shirali Kadyrov. All rights reserved.

(subsequences of that length) is counted. This process is repeated for multiple word lengths, allowing us to analyze how the number of distinct binary words grows as the word length increases.

Next, the base-2 logarithm of the number of unique binary words is computed for each word length. This step helps to transform the data into a form that reveals scaling properties. The resulting data points, which represent the relationship between word length and the logarithm of the unique word count, are then analyzed using linear regression. Linear regression is used to fit a trend line to the data, which captures the overall pattern of growth.

Once the trend line is obtained, the box dimension of the set is determined by the slope of the regression line. This slope quantifies how the number of unique binary words scales with word length and provides a numerical measure of the set's complexity. A higher slope indicates greater complexity, while a lower slope suggests a more structured or predictable pattern in the binary expansions.

The computational procedure follows these steps: The Set points are first converted to binary expansions of a specified precision. For each word length, the binary expansions are truncated, and the number of unique words is counted. The $\log_2$ of the unique word count is computed and stored. Then a linear regression is performed on the relationship between word length and $\log_2$ count, and the slope of the regression line is returned as the estimated box dimension. The results are visualized through a regression plot that shows the relationship between word length and $\log_2$ of unique word counts, where the slope of the fitted trend line provides an approximation of the box dimension of the underlying fractal set.The following pseudocode Fig.2. summarizes the computational procedure:

Input: Set points, precision p, max word length L

Output: Estimated box dimension

1. Convert each set point to a binary expansion of length p.
2. Initialize an empty list for log_2 counts.
3. For each word length l from 1 to L:
   a. Truncate each binary expansion to the first l bits.
   b. Count the number of unique truncated words.
   c. Compute log_2 of the unique word count and store it.
4. Perform linear regression on (word length, log_2 count) pairs.
5. Return the slope of the regression line as the estimated box dimension.

Fig. 2. Algorithm to numerically compute the box dimension

To carry out the experiments we generated one million points uniformly from the interval [1,2]. The distribution of denominator averages of these numbers are depicted in Fig.3.

III. Results and Discussion

Fig.4 provide numerical results for estimating the Box dimension of $S(\epsilon, c)$ for fixed ϵ=0.01 and c ranging from 1.60 to 1.95.Our numerical investigation of the box dimension of $S(\epsilon, c)$ reveals a clear dependence on c.Using the binary word-based box-counting method, we estimated the box dimension of these exceptional sets. By comparing the log of unique binary word counts to word length using linear regression, we found a slope that shows how $S(\epsilon, c)$ scales in a fractal way.

Received: February 17, 2025. Reviewed: March 26, 2025. Accepted: March 29, 2025. © 2025 Symbat Duisen, Aiken Kazin, Shirali Kadyrov. All rights reserved.

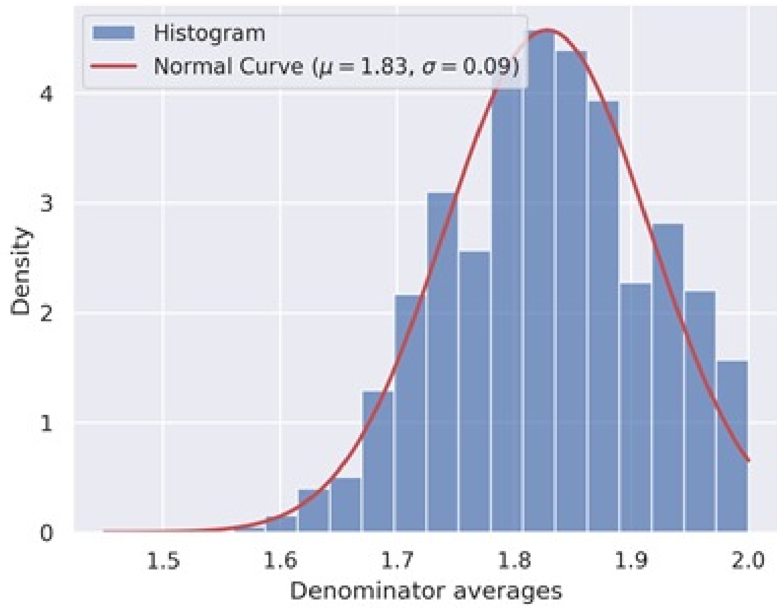


Fig. 3. Histogram plot of relative frequency distribution of average denominators of Lehner expansion

The Fig. 5 suggests that as c increase, the box dimension stabilizes, reinforcing the theoretical expectation that the set of exceptions forms a measure-zero yet structurally complex subset.

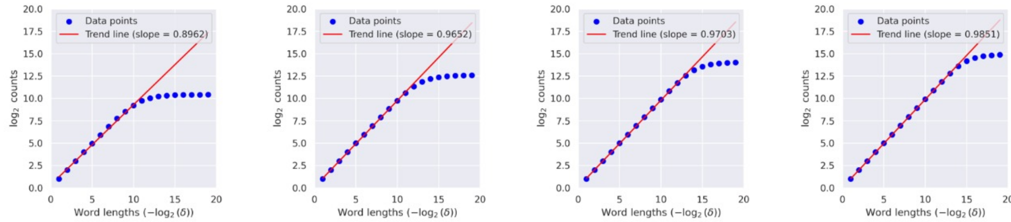
Fig.6(a) depicts a fractal structure generated from continued fraction expansions with a restricted digit set { 1, 2, 3, 4 }. The x and y coordinates correspond to values derived from odd- and even-indexed terms of randomly generated continued fraction sequences. The resulting structure reveals an intricate, self-similar distribution within the unit square, illustrating how different digit choices influence the fractal pattern. Fig.6 (b) shows a similar fractal formation, but based on the Lehner expansion, a variant of continued fraction representation defined for numbers in the interval [1,2]. Here, the x and y coordinates are determined by evaluating the odd- and even-indexed Lehner terms as continued fractions. The clustering and density variations within the bounded region reflect the distinctive number-theoretic properties of the Lehner transformation and its role in generating self-similar structures.

IV. Conclusion

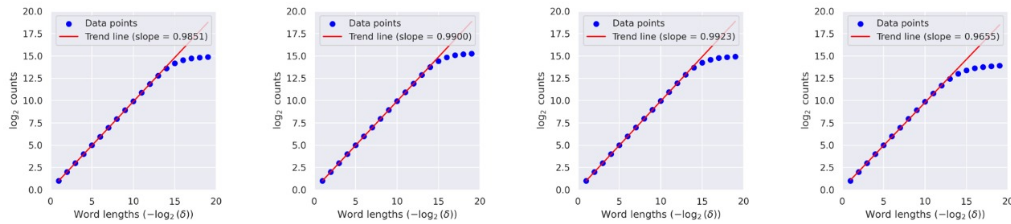
In this paper, we investigated the fractal properties of exceptional sets in the Lehner expansion by examining how the average value of the sequence b_n affects the box dimension. By employing numerical fractal analysis, we computed the box dimensions of these exceptional sets using a binary word-based box-counting method. Our findings demonstrate that the box dimension of $S(\epsilon, c)$ exhibits a clear dependence on c, with the box dimension stabilizing as c increases. This aligns with the theoretical expectation that these sets, despite having Lebesgue measure zero, exhibit intricate fractal structures.

Our numerical results provide evidence that the exceptional sets in the Lehner expansion possess a non-trivial fractal nature, reinforcing the idea that continued fraction expansions offer a rich framework for studying complex structures in number theory. The observed self-similar patterns in Fig 6(a) and 6(b) further illustrate how the Lehner expansion differs from regular continued fractions while maintaining its own unique fractal characteristics. The histogram of denominator averages (Fig.3) and the scaling

Received: February 17, 2025. Reviewed: March 26, 2025. Accepted: March 29, 2025. *© 2025 Symbat Duisen, Aiken Kazin, Shirali Kadyrov. All rights reserved.*



(a) $S(0.01,1.60)$ (b) $S(0.01,1.65)$ (c) $S(0.01,1.70)$ (d) $S(0.01,1.75)$



(e) $S(0.01,1.80)$ (f) $S(0.01,1.85)$ (g) $S(0.01,1.90)$ (h) $S(0.01,1.95)$

Fig. 4. Numerical box dimension estimates of $S(\epsilon, c)$ for ϵ=0.01 and varying c.

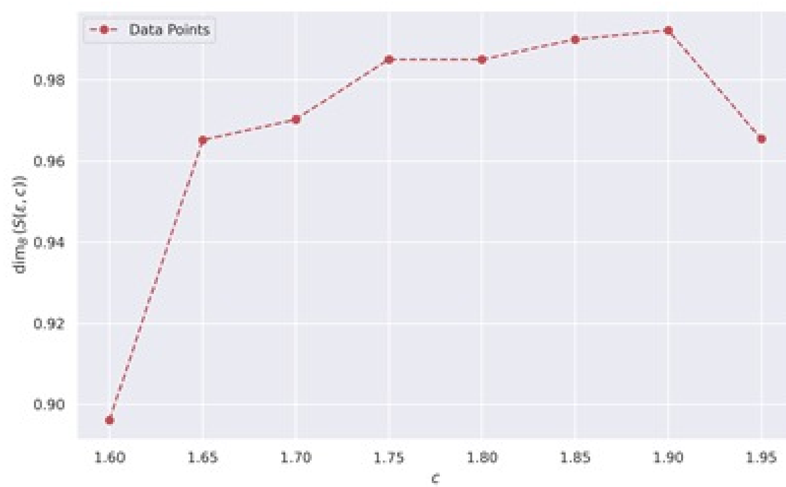
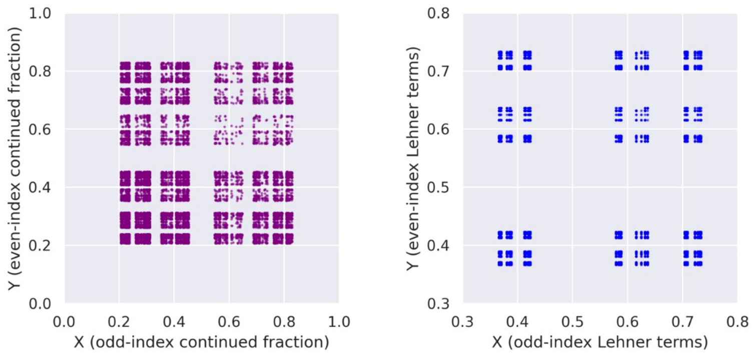


Fig. 5. The graph of box dimensions of $S(\epsilon, c)$ as c changes from 1.60 to 1.95

behavior of box dimensions (Fig.5) suggest that the complexity of these sets is deeply tied to the digit distributions in their continued fraction representations.

Future work could extend this study by exploring different ranges of ϵ and c to further characterize the transition behaviors of fractal dimensions. Additionally, a theoretical analysis of the scaling behavior observed in our numerical experiments could provide deeper insights into the number-theoretic properties of Lehner expansions. Overall, this study contributes to the growing body of

Received: February 17, 2025. Reviewed: March 26, 2025. Accepted: March 29, 2025. © 2025 Symbat Duisen, Aiken Kazin, Shirali Kadyrov. All rights reserved.



(a) 2D fractal structure of regular continued fractions

(b) 2D fractal structure of Lehner continued fractions

Fig. 6. Two-Dimensional Fractal Structures from Continued Fraction Expansions

research on the fractal geometry of exceptional sets in continued fraction theory, offering new perspectives on their complexity and structure.

Received: February 17, 2025. Reviewed: March 26, 2025. Accepted: March 29, 2025. *© 2025 Symbat Duisen, Aiken Kazin, Shirali Kadyrov. All rights reserved.*

References

[1] M. Einsiedler and T. Ward, *Ergodic Theory with a View Towards Number Theory*. Springer, 2011.

[2] S. Kadyrov, "Semi-Regular Continued Fractions with Fast-Growing Partial Quotients," *Fractal and Fractional*, vol. 8, p. 436, 2024.

[3] C. Kraaikamp, "A New Class of Continued Fraction Expansions," *Acta Arithmetica*, vol. 57, pp. 1–39, 1991.

[4] I. J. Good, "The Fractional Dimensional Theory of Continued Fractions," *Cambridge University Press*, 1941.

[5] K. A. Kazin, "Fractal Geometry and Level Sets in Continued Fractions," *Herald of the Kazakh-British Technical University*, vol. 21, no. 2, pp. 116–126, 2024.

[6] K. Falconer, *Fractal Geometry: Mathematical Foundations and Applications*, 3rd ed. Wiley, 2013.

[7] K. Dajani, "The Mother of All Continued Fractions," *Colloquium Mathematicum*, vol. 84/85, no. 1, pp. 109–123, 2000.

[8] L. Fang, "Some Exceptional Sets of the Borel–Bernstein Theorem in Continued Fractions," *The Ramanujan Journal*, vol. 56, pp. 891–909, 2021.

[9] A. Bakhtawar, "Hausdorff Dimension of an Exceptional Set in the Theory of Continued Fractions," *Nonlinearity*, vol. 33, no. 6, pp. 2615–2634, 2020.

[10] S. T. Parsell, "Exceptional Sets for Diophantine Inequalities," *International Mathematics Research Notices*, vol. 2014, no. 14, pp. 3919–3974, 2014.

[11] M. Kesseböhmer and B. O. Stratmann, "A Multifractal Analysis for Stern-Brocot Intervals, Continued Fractions and Diophantine Growth Rates," *Journal für die reine und angewandte Mathematik (Crelle's Journal)*, vol. 2004, no. 582, pp. 1–28, 2004.

[12] W. Philipp, "Limit Theorems for Sums of Partial Quotients of Continued Fractions," *Monatshefte für Mathematik*, vol. 105, pp. 195–206, 1988.

[13] Y. Pesin and H. Weiss, "The Multifractal Analysis of Gibbs Measures: Motivation, Mathematical Foundation, and Examples," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 7, no. 1, pp. 89–106, 1997.

[14] D. Hensley, "A Polynomial Time Algorithm for the Hausdorff Dimension of Continued Fraction Cantor Sets," *Journal of Number Theory*, vol. 58, pp. 9–45, 1996.

Received: February 17, 2025. Reviewed: March 26, 2025. Accepted: March 29, 2025. *© 2025 Symbat Duisen, Aiken Kazin, Shirali Kadyrov. All rights reserved.*

End of Volume 1, Issue 1

Journal of Emerging Technologies and Computing (JETC)

Published by SDU University • © 2025