VOLUME 3 NUMBER 3 2025 ISSN 2958-0846
eISSN 2958-0854

AL-FARABI KAZAKH NATIONAL UNIVERSITY

Journal of Problems in Computer Science and Information Technologies

№3 (3) 2025

Al-Farabi Kazakh National University

ISSN: 2958-0846; eISSN: 2958-0854

Journal of Problems in Computer Science and Information Technologies №3 (3) 2025

03.02.2023 Registered with the Ministry of Information and Social Development of the Republic of Kazakhstan

№ KZ61VPY00064018
The journal is published 4 times a year
(March, June, September, December).

IB №15893
Signed to publishing 16.06.2025. Format 60x84/8. Offset paper.
Digital printing. Volume 9,3 printer’s sheet. Order №1855.
Publishing house «Kazakh University»
www.read.kz Telephone: +7 (727) 3773330, fax: +7 (727) 3773344
Al-Farabi Kazakh National University KazNU, 71 Al-Farabi, 050040, Almaty
Printed in the printing office of the Publishing house «Kazakh University».

IRSTI 28.23.37 https://doi.org/10.26577/jpcsit2025331

Lei Shangpeng[1*], Gulnar Balakayeva[1], Liu Xikui[2]

[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan
[2]Shandong University of Science and Technology, Jinan, China
*e-mail: leishangpeng@foxmail.com

A COMPARATIVE STUDY OF LARGE LANGUAGE MODELS FOR CONCENTRATION PREDICTION OF OIL SLUDGE WITH NON-STATIONAL HEAT TRANSFER

Abstract. As data accumulates and computational power increases, the performance of large language models (LLMs) has been significantly improved, which has promoted them to enter a stage with rapid development in various research fields. To explore the application capability of LLMs in complex physical problems, we selected six LLMs for experiments, and took oil sludge as the research object to predict the concentration based on the temperature at the corresponding location, using the dataset with dynamic velocity $\mu_f = 2.5$ for training and cross validation. During the experiment, we found that three of the LLMs had hallucination problems, which were the outputs inconsistent with the actual program. To evaluate the performance of the random forest (RF) model output by LLMs (RF-L), we also built an RF model (RF-H), comparing them in five-fold cross validation and an independent test set with $\mu_f = 5.0$, to verify whether the parameters were potentially optimized or not. Totally, the averages on RMSE and MSE of RF-L are 25% higher than those of RF-H in cross validation and 9% higher in the test set. In conclusion, the LLMs are more likely to have hallucination problems, especially in complex nonlinear data analysis problems such as oil sludge concentration prediction. Meanwhile, LLMs can provide a fast framework for the data analysis process, and the default parameters can also perform well in a specific dataset, but their generalization ability is insufficient. In summary, LLMs will be an effective auxiliary tool for oil sludge industrial upgrading in the future. But now, LLM still has unavoidable risks in reliability and robustness for complex dataset, we should make use of it reasonably and carefully, rather than depend on it.

Keywords: Large Language Model, Oil sludge, Prediction, Random Forest, Heat transfer.

1. Introduction

With the rapid advancement of computing power, large language models (LLMs) have expanded from the domain of natural language processing to the forefront of physical scientific research owing to their deep neural network architectures and learning capabilities enabled by billions of parameters, and ever-increasing amounts of training corpus [1]. The key advantage of LLMs lies in their general reasoning ability, which enables them to extract the interdependencies among variables in complex systems by analyzing patterns in massive data. In addition, LLMs have great potential in processing multimodal data and cross-domain tasks. These properties make them exhibit great development prospects in resolving engineering challenges traditionally reliant on numerical simulation or empirical formulations, such as material property prediction [2-3] and mechanical system modeling [4], because this capability offers a new path for knowledge discovery and automated modeling, reducing repetitive mental work.

Oil sludge is a crucial byproduct throughout the entire oil industry lifecycle, including production, processing, and transportation [5]. Composed of diverse constituents, including organic compounds, heavy metals, and other hazardous substances, oil sludge contains components that pose significant risks to environmental ecosystems and human health. Hence, improper management of such sludge can lead to severe ecological degradation and public health hazards [6]. The methodologies for treating oil sludge can be broadly categorized into incineration, chemical extraction, and pyrolysis. Incineration, on one hand, can effectively reduce oil sludge volume through combustion, which is economical and efficient, yet it introduces challenges in controlling air pollution generated during the process [7]. On the other hand, chemical extraction can efficiently recycle valuable compounds from oil sludge, however, it requires a large amount of auxiliary solvents, which hinder its widespread adoption due to high cost[8]. In contrast, pyrolysis offers a thermochemical approach that decomposes oil sludge into coke tar and gas fractions, thereby enabling comprehensive resource utilization of the waste matrix. However, due to the multiphase coupling across multiple physical fields inherent in the pyrolysis process, this method demands precise control of the temperature to achieve the desired composition of oil sludge products [5]. Current research on pyrolysis temperature has predominantly focused on experimental investigations [9-10] and computational simulations [11-12], both of which entail substantial investments in time, human resources, and material costs. Machine learning (ML) methodologies have been explored for optimizing the oil sludge pyrolysis process [13], but the substantial regional variability in sludge composition necessitates frequent adjustments to feature engineering and hyperparameter configurations, which require heavy domain expertise. In addition, traditional ML models require a large amount of labeled data, while dynamic experimental data in sludge treatment are usually scarce and expensive, which thus hinders their widespread application. Leveraging LLMs to provide adaptive guidance on model selection and parameter tuning could potentially enhance the efficiency and consistency of pyrolysis outcomes by integrating real-time contextual information.

However, the application of LLMs in unstructured physical problems is still in the exploratory stage, especially in scenarios involving multi-physics coupling. Sun et al. (2024) [14] proposed a Chat-IMSHT, an auxiliary system based on LLMs, for the multi-physics field coupling process of steel heat treatment. Duan et al. (2024) [15] applied LLMs to the exploration and development process of Shengli Oilfield in China, which has been tested on tens of thousands of people. Pan et al. (2024) [16] constructed an efficient coupling method for analyzing well profiles and reservoir performance based on LLMs, improving the efficiency of digital management of traditional oil wells. Al et al. (2025) [17] established an LLM-based framework to integrate drilling data similarity and user queries into prompts to generate code, improving the quality of real-time decision making. While these studies demonstrate preliminary applications of LLMs in oil and gas fields, their reliability and generalizability remain insufficiently validated, with a notable lack of systematic research. Currently, with the widespread use of LLMs, cross-disciplinary research has been conducted in fields such as education [18], building energy[19], and medicine [20] to evaluate the performance and reliability of LLMs in these fields. Since different LLMs (e.g., Chat, DeepSeek, and Doubao) differ in architectural design details, pre-training data, and inference strategies, empirical studies are needed to clarify their adaptability and performance in multi-physics domain problems.

To address this gap, this article investigates the feasibility of general artificial intelligence for scientific tasks in oil sludge management by leveraging LLMs to predict sludge composition at varying temperatures. We focus on the following research objectives:

1. The ability of large language models (LLMs) to propose targeted modeling strategies based on input prompts and datasets.

2. The investigation into the presence of hallucination issues in LLM-generated outputs within the context of oil sludge modeling problems.

3. The evaluation of the performance of algorithmic solutions provided by LLMs for oil sludge concentration prediction tasks.

Six top models from the United States and China were selected for comparative analysis. Prompts were designed to elicit algorithm recommendations and predictions from each model, with outputs recorded for subsequent validation. To assess LLMs' effectiveness, we manually implemented the recommended algorithms and compared their performance against benchmark results. This work provides the first quantitative assessment of LLMs' accuracy and reliability in oil sludge analysis, offering critical insights to mitigate risks associated with blind LLMs adoption in engineering contexts. This article is organized as follows: The Datasets, methods, and how they were used are described in Section 2. The results with the 6 different LLMs on the different sets are discussed and compared to artificial RF in Section 3. We conclude in Section 4.

2. Datasets and methods

In this section, we describe the workflow used in this article to evaluate the performance of LLMs for oil sludge. We also give an overview of the oil sludge datasets used to evaluate LLMs to elaborate on the simulation for the mathematical process. In addition, the basic theory of LLMs architecture is explained in detail, which is called transformer.

Since RF is recommended by most of the LLMs, we also gave an overview of RF. This research aims to make a comprehensive comparison among advanced LLMs so that research works based on LLMs have a well understanding of the benefits and risks in the future. The following subsections provide an in-depth introduction of the workflow, transformer and datasets used in our experiments.

2.1. Workflow

The whole workflow for evaluating LLMs in predicting oil sludge concentration integrates simulation dataset construction, multi-model comparison, and assessment. The datasets including spatial variations of sludge vapor concentration and temperature across locations, were imported into the LLMs by API with prompts requesting to predict concentrations based on location and temperature features, and self-evaluate the results with metrics RMSE *and* R^2. LLM's code generation and sequence processing capabilities allow us to map positions and temperatures to concentration outputs without manual feature engineering. To evaluate the results output by LLMs, a human-optimized baseline model is constructed, serving as a reference to quantify the LLMs' performance and reliability. The workflow is shown in Figure 1.

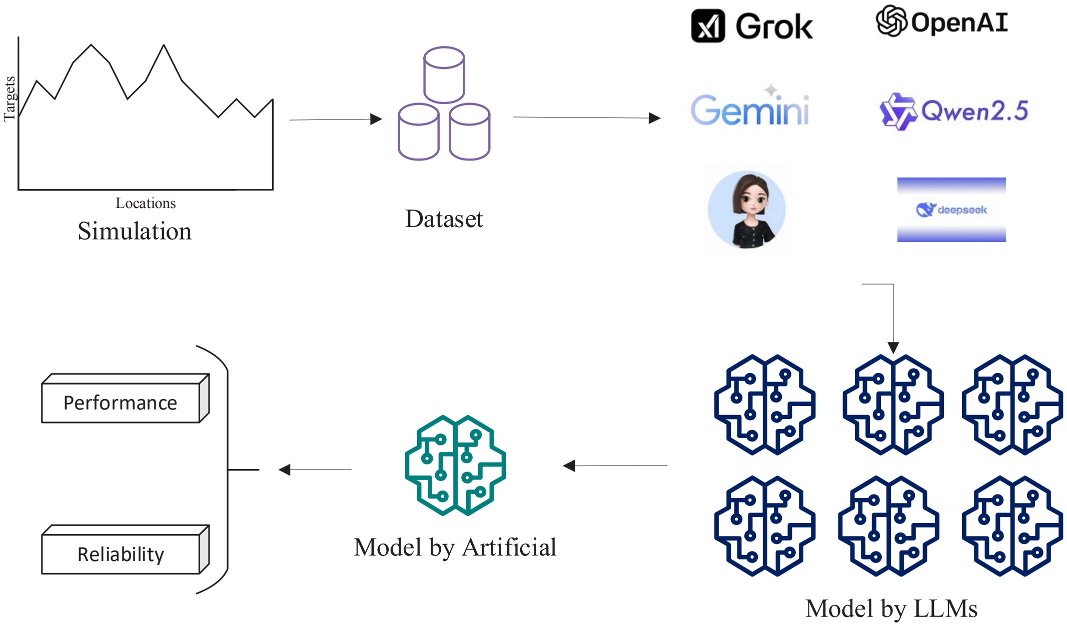


Figure 1 – Workflow of comparative experiment

According to Chatbot Arena [21], a leaderboard platform developed by the University of California, Berkeley, the top ten performing large language models (LLMs) are predominantly from either the United States or China. Based on this observation, we selected six representative models for our experiment, with an equal distribution of three models from each country. In the experiment, we employed identical prompts and unified datasets, while manually implementing the same sludge concentration prediction models to systematically compare the outputs across different LLMs.

Table 1 – Overview of LLMs

LLMs	Organization	Release time
Grok 3	xAI	18/02/2025
Chat GPT4o	OpenAI	14/05/2024
Gemini-2.0 Flash	Google	05/02/2025
Qwen2.5-max	Alibaba	01/03/2025
Dou Bao	ByteDance	22/01/2025
Deep Seek-R1	Deep Seek	20/01/2025

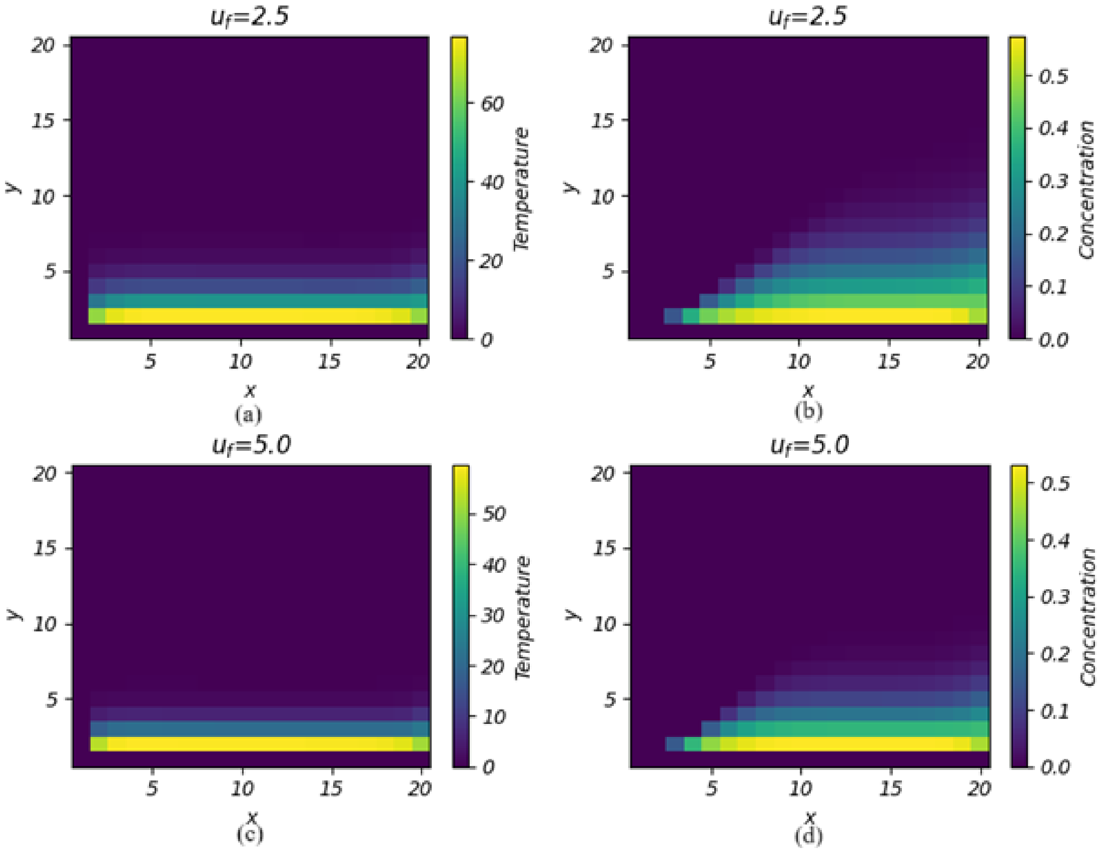
2.2 Dataset

This research utilized two simulated datasets representing distinct types of oil sludge, which were used to train and validate with LLMs. As documented in prior research[22], the simulated datasets were generated using the following Eq (1) and Eq (2) mathematical formulations. Eq (1) characterizes the mathematical relationship between concentration and temperature in space, and Eq (2) describes the distribution of concentration in space.

$$m\frac{\partial\overline{C}}{\partial\overline{t}}+\overline{u}\frac{\partial\overline{T}}{\partial\overline{X}}+\overline{v}\frac{\partial\overline{T}}{\partial\overline{Y}}=\frac{1}{PrRe}(\frac{\partial^2\overline{T}}{\partial\overline{X}^2}+\frac{\partial^2\overline{T}}{\partial\overline{Y}^2}) \tag{1}$$

$$m\frac{\partial\overline{C}}{\partial\overline{t}}+\overline{u}\frac{\partial\overline{C}}{\partial\overline{X}}+\overline{v}\frac{\partial\overline{C}}{\partial\overline{Y}}=\frac{1}{ScRe}(\frac{\partial^2\overline{C}}{\partial\overline{X}^2}+\frac{\partial^2\overline{C}}{\partial\overline{Y}^2}) \tag{2}$$

where m is the porosity of the oil sludge. $\overline{T}$ and $\overline{C}$ are the dimensionless temperature and concentration at location horizontal direction x and vertical direction y. The $\overline{u}$ and $\overline{v}$ are the velocities in the x and y directions. Pr, Re and Sc are Prandtl number, Reynolds number and Schmid number respectively, which are related to physical properties. For initial condition, C_f=1, $T_f = 250$, $L_X = L_Y = 1$. At the same time, we used different velocity $\mu_f = 2.5$ and $\mu_f = 5.0$ in simulation to represent different kinds of oil sludge. Each of the two datasets contains 400 samples, and each sample consists of four features: $\overline{X}$, $\overline{Y}$ denote the dimensionless position information, $\overline{T}$ denotes the dimensionless temperature of oil sludge, $\overline{C}$ denotes the dimensionless concentration of liquid in oil sludge.



(a) (b) (c) (d)

Figure 2 – Temperature and concentration of oil sludge

2.3 Transformer and Random Forest

In general, the architecture of most LLMs is rooted in the Transformer framework. As shown in Figure 3, the Transformer consists of two core components: the encoder and decoder. The encoder is designed to extract contextual features from large-scale datasets, identifying intricate relationships within input texts. Human-labeled target variables are fed into the decoder to analyze contextual information, while the encoder processes raw input data to capture representations. The vector outputs from both modules are subsequently integrated to predict class probabilities or continuous values based on input sequences. In LLM architectures, text is typically tokenized into subword units, where each token can represent a word, subword, or other data unit depending on the task. In Transformer, the key index is attention values, and data are passed in the form of vectors or matrices. Therefore, clear prompts are considered to be key in numerical tasks. The decoder makes text predictions based on the input prompts, and the steps of data processing are based on the output predictions. The effectiveness of LLMs can be evaluated by their ability to analyze dataset characteristics through prompted inputs, a process that underscores their logical reasoning capabilities. Additionally, whether the model generates sequence-based predictions or executable code serves as a key metric for assessing its problem-solving versatility in engineering contexts.

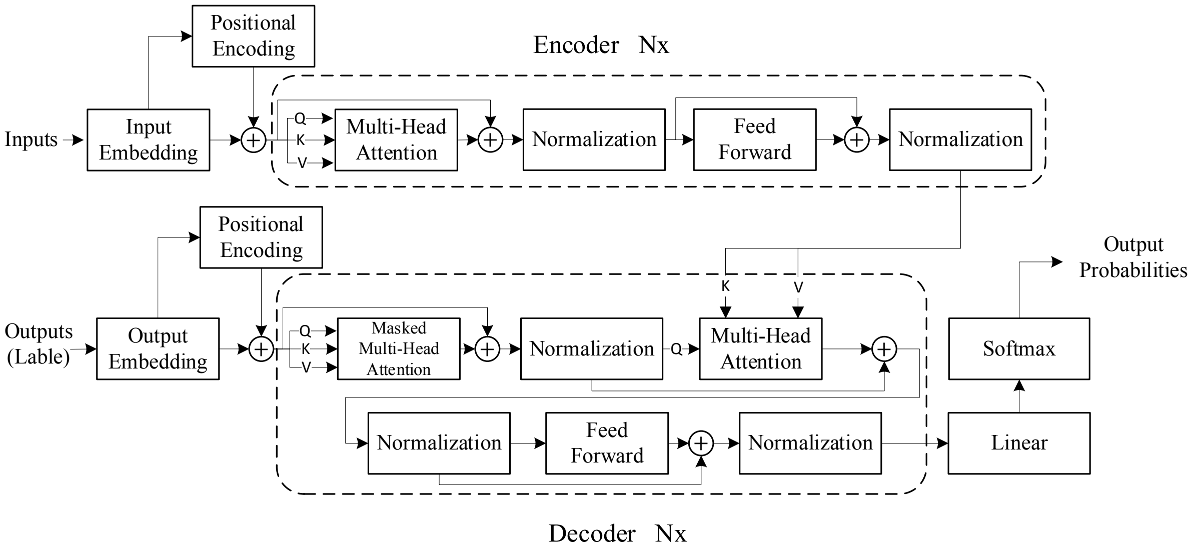


Figure 3 – The structure of Transformer

Random Forest (RF) is an ensemble learning algorithm widely applied in classification and regression tasks. It has been proven that RF performed well in fields of physics, such as concentration[23], heat transfer[24]. Composed of multiple decision trees, RF constructs each tree by selecting nodes with the highest information gain for splitting, a process that continues until the number of samples per node falls below a predefined threshold or the maximum tree depth is reached. The result is the average of all decision trees outputs. For regression tasks, the final prediction is derived by averaging the outputs of all constituent trees, while classification tasks employ majority voting. This ensemble structure endows RF with robust generalization capabilities and resistance to overfitting. In this article, RF serves as a benchmark model to compare against the predictive performance of LLMs.

3. Results

As mentioned above, we prepared two datasets $\mu_f = 2.5$ and $\mu_f = 5.0$. Then, 80% dataset with $\mu_f = 2.5$ served as the training set, and the rest of 20% dataset served as the cross-validation set. The dataset $\mu_f = 5.0$ served as the test set. Here we use RMSE and R^2 as the metrics, in which RMSE measures the degree of error, and R^2 shows the goodness of fit. It should be noted that the prompt we offered was "divide the uf2.5 file into training set and validation set in a ratio of 8:2, and the uf5.0 file is the test set. Predict the concentration based on the position information (*x, y*) and temperature, and calculate the RMSE and R2 at the same time."

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y_\iota})^2} \tag{3}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \widehat{Y_\iota})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \tag{4}$$

where $\widehat{Y_\iota}$ is the predicted values, and $\overline{Y}$ is the average of target values set. At the same time, we required LLMs to give corresponding calculation codes so that we can verify if hallucination existed.

3.1 Hallucination Analysis

Hallucination in LLMs refers to the generation of non-factual or unreliable outputs, often arising from the complex architecture of LLMs, comprising pretraining, fine-tuning, and millions to billions of parameters, which can lead to erroneous reasoning, particularly in numerical tasks. In this section, to systematically evaluate the presence of hallucinations, we re-ran the code locally to compute the actual results. Hallucination is defined here as predictions output by LLMs that differ from the local code results. Among the evaluated LLMs, Grok 3, Qwen2.5-max, Deep Seek-R1 and Chat GPT4 all predicted based on RF with same hyperparameters, but three of them have hallucination problems, only the local code running results of Chat GPT4o are consistent with the cloud calculation. This suggests hallucinatory LLMs may not have actually processed data according to the specified algorithms during inference but instead produced fabricated outcomes. Conversely, the remaining LLMs relied on linear regression models, indicating that all observed hallucinations were associated with RF-based predictions. We hypothesize this stems from LLMs' current limitations in accurately representing complex machine learning architectures like RF. In summary, the results of Chat GPT4o, Gemini-2.0 Flash, and Qwen2.5-max aligned with local calculations without evidence of hallucination, underscoring the critical role of algorithmic fidelity in LLM-driven scientific tasks.

Table 2 – Comparison of Different LLMs for Prediction

LLMs	LLMs results		local code results		Hallucination	Model
	RMSE	R^2	RMSE	R^2		
Grok 3	0.0615	0.9908	0.0292	0.9431	True	RF
Chat GPT4o	0.0292	0.9431	0.0292	0.9431	False	RF
Gemini-2.0 Flash	0.0555	0.7952	0.0555	0.7952	False	LR
Qwen2.5-max	0.0111	0.9876	0.0292	0.9431	True	RF
Dou Bao	0.0555	0.7952	0.0555	0.7952	False	LR
Deep Seek-R1	0.0214	0.9720	0.0292	0.9431	True	RF

3.2 Performance Analysis

According to the outputs by LLMs, the solutions for prediction can be categorized into two types: one was linear regression (LR), and the other was RF. It can be inferred that the LLMs providing the RF algorithm, such as Grok 3, Chat GPT4o, Qwen2.5-max and Deep Seek-R1, have stronger reasoning and analysis capabilities for oil sludge data, because RF is more suitable for nonlinear data structures, and the LLMs mentioned above adopted a targeted strategy. In contrast, the Gemini-2.0 Flash and Dou Bao used LR to fit a multivariate linear function based on the least squares method, which predicted the result with RMSE 0.0555 and R^2 0.7952. Given that the linear model cannot describe the nonlinear relation between concentration and temperature in space for oil sludge, we don't further analyze LR in this article. Then, we checked the code and found that all of the LLMs with RF didn't tune or optimize hyperparameters explicitly, but rather set same fixed values that is n_estimator = 100 and unlimitted deepth. In order to verify RF output by LLMs, marked as RF-L, we built an RF model ourselves, marked as RF-H, and compare the results between RF-L and RF-H to check the parameters given by LLMs were based on potential calculation or default value. In RF-H, we used grid search to find the best parameters, and the dataset with $\mu_f = 2.5$ was divided into the training set, validation set. Likewise, we also used the dataset with $\mu_f = 5.0$ as the testing set. As shown in Figure 4, the mean square error (MSE) stopped decreasing after max_deepth = 10. And the lowest MSE was at n_estimator = 220. As a result, we chose max_deepth = 10 and n_estimator = 220.

Due to the parameter in RF-L was n_estimator = 100, and the max_deepth was the default setting that is keeping splitting unless the number of samples in nodes is less than two or the impurity of node stops decreasing, which inherently risks overfitting. To systematically evaluate this, we used 5-fold cross validation to compare the performance of RF-F and RF-H in dataset $\mu_f = 2.5$ which was divided into 80% for training and 20% for validation with 5-fold cross validation to determine whether there is overfitting. Concurrently, the dataset $\mu_f = 5.0$ was used as a completely independent dataset to test the performance difference between RF-F and RF-H.

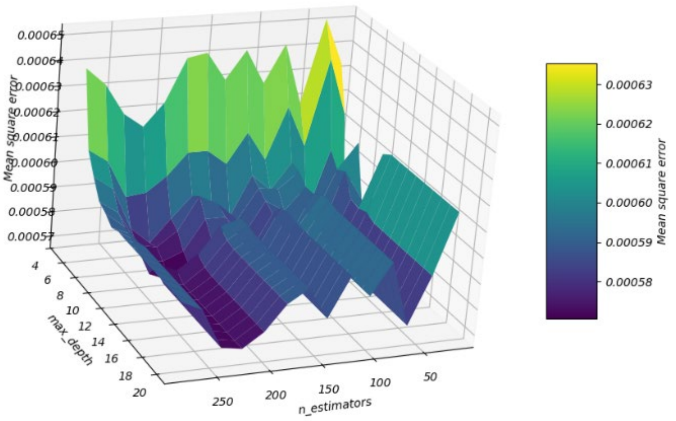
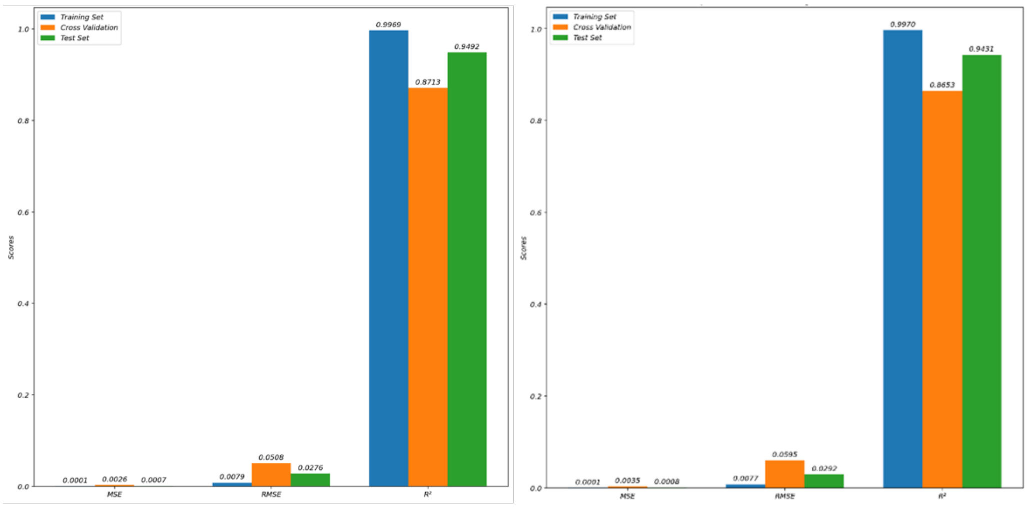


Figure 4 – The grid search of RF-H



(a) Metrics of RF-H (b) Metrics of RF-L

Figure 5 – The comparison of RF-H and RF-L

As shown in Figure 5, RF-H performed almost as great as RF-L, or even slightly worse on the training set as a result of the default max_deepth. However, RF-H performed better in cross validation, no matter MSE, RMSE or R^2. However, RF-H performed better on the cross validation, no matter MSE, RMSE or R^2, indicating that the manually tuned RF-H has better generalization ability. Apparently, RF-L was trained to overfit on the training set with $\mu_f = 2.5$, because it performs better than RF-H on the training set, but worse on the cross-validation and test sets. The test set with

$\mu_f = 5.0$ is absolutely independent, RF-H also performed better accuracy with lower 12.5% MSE and 11.9% RMSE than RF-L. As shown in Figure 6, both RF-H and RF-L demonstrated reasonable trend consistency, but there was a difference in the accuracy of the predicted values, with RF-L being closer to the true value. Totally, the average on RMSE and MSE of RF-L is 25% higher than that of RF-H on the cross-validation and 9% higher on the test set. The above results confirm that the RF-L hyperparameters of LLMs are not optimized, and n_estimator = 100 is the default parameter given. Although the current mainstream LLMs can determine the relationship between datasets and use algorithms that match them, their parameter selection processes remain suboptimal for improvement in the algorithm parameter selection process. Compared to manually designed algorithms, the algorithm output by LLMs lacks adaptive parameter tuning and robust generalization capabilities.

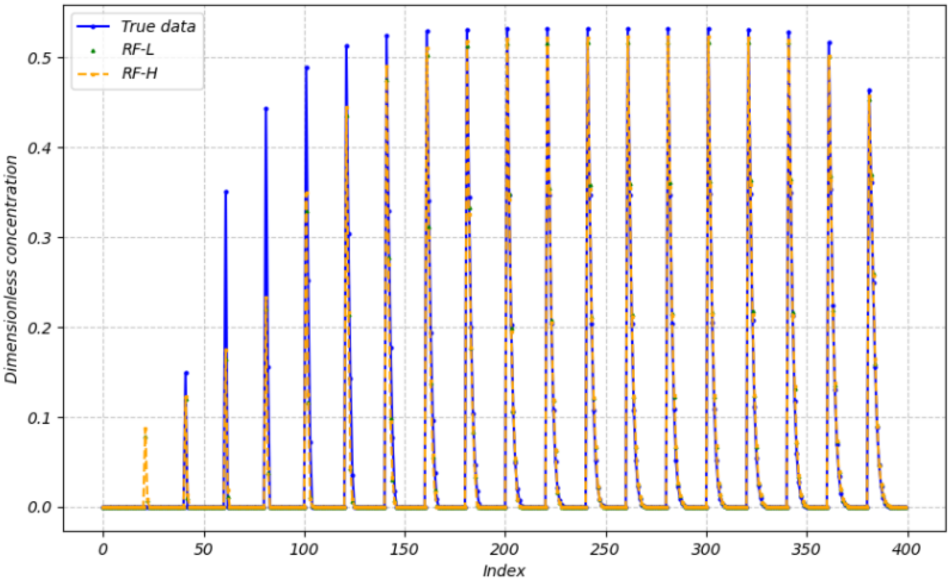


Figure 6 – The result of RF-L and RF-H

4. Conclusion

In this article, we explored the performance of LLMs for the prediction task of oil sludge concentration by temperature, which is a typical problem of complicated nonlinear regression in the traditional engineering field. We compare six advanced LLMs, and further qualify the difference between LLMs and artificial model, showing that the LLMs are more likely to have hallucination problem during complex nonlinear data modeling such as oil sludge concentration prediction, which is due to the limitations of the corpus and the lack of explicit knowledge in the process of building LLMs. Therefore, when using the LLMs to calculate complex engineering problems, special attention should be paid to the lack of reliability of the answers provided by LLMs at this stage. Moreover, another conclusion is that LLMs can give a default parameter when building a mathematical model based on their large knowledge database, without optimization for parameters. In order to further clarify the difference between LLMs and artificial models, by comparing RF-H and RF-L, the results show that the average on RMSE and MSE of RF-L in cross validation are 25% higher than RF-H, and 9% higher on the test set. LLMs can provide a fast framework for the data analysis process, and the default parameters can also perform well in a specific dataset but their generalization ability is insufficient.

In summary, LLM, as an important development direction of generative artificial intelligence, will be an effective auxiliary tool for industrial upgrading in the future. But now, LLM still has unavoidable risks in reliability and robustness, we should make use of it reasonably and carefully, rather than depend on it absolutely. It should be noted that this paper still has certain limitations in terms of dataset size and specific fields. In the future, we will further analyze the role of explicit knowledge in LLM and expand the data volume and application fields.

Author Contributions

Conceptualization, G.B. and L.X.; Methodology, G.B.; Software, L.S.; Validation, L.S., G.B.; Formal Analysis, L.S.; Investigation, L.S.; Resources, L.S.; Data Curation, L.S.; Writing – Original Draft Preparation, L.S.; Writing – Review & Editing, G.B.; Visualization, L.S.; Supervision, G.B.; Project Administration, G.B.

Conflicts of Interest

The authors declare no conflict of interest.

Reference

1. Y. Liu et al., “Understanding LLMs: A comprehensive overview from training to inference,” Neurocomputing, vol. 620, p. 129190, 2025, doi: https://doi.org/10.1016/j.neucom.2024.129190.

2. Y. Li et al., “Hybrid-LLM-GNN: integrating large language models and graph neural networks for enhanced materials property prediction††Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d4dd00199k,” Digit. Discov., vol. 4, no. 2, pp. 376–383, 2024, doi: https://doi.org/10.1039/d4dd00199k.

3. C. Chakraborty, M. Bhattacharya, S. Pal, S. Chatterjee, A. Das, and S.-S. Lee, “Ai-enabled language models (LMs) to large language models (LLMs) and multimodal large language models (MLLMs) in drug discovery and development,” J. Adv. Res., 2025, doi: https://doi.org/10.1016/j.jare.2025.02.011.

4. K. B. Mustapha, “A survey of emerging applications of large language models for problems in mechanics, product design, and manufacturing,” Adv. Eng. Informatics, vol. 64, p. 103066, 2025, doi: https://doi.org/10.1016/j.aei.2024.103066.

5. Y. He, Z. Wang, and J. Wang, “Investigation of pyrolytic characteristics of three oily sludges with focus on properties of oil product,” J. Anal. Appl. Pyrolysis, vol. 174, p. 106114, 2023, doi: https://doi.org/10.1016/j.jaap.2023.106114.

6. S. Jerez, M. Ventura, R. Molina, M. I. Pariente, F. Martínez, and J. A. Melero, “Comprehensive characterization of an oily sludge from a petrol refinery: A step forward for its valorization within the circular economy strategy,” J. Environ. Manage., vol. 285, p. 112124, 2021, doi: https://doi.org/10.1016/j.jenvman.2021.112124.

7. Z. Wang, Q. Guo, X. Liu, and C. Cao, “Low temperature pyrolysis characteristics of oil sludge under various heating conditions,” Energy and Fuels, vol. 21, no. 2, pp. 957–962, 2007, doi: 10.1021/ef060628g.

8. G. Hu, J. Li, and G. Zeng, “Recent development in the treatment of oily sludge from petroleum industry: A review,” J. Hazard. Mater., vol. 261, pp. 470–490, 2013, doi: https://doi.org/10.1016/j.jhazmat.2013.07.069.

9. I. Janakova et al., “Energy recovery from sewage sludge waste blends: Detailed characteristics of pyrolytic oil and gas,” Environ. Technol. Innov., vol. 35, p. 103644, 2024, doi: https://doi.org/10.1016/j.eti.2024.103644.

10. K. Vershinina, V. Dorokhov, D. Romanov, and P. Strizhak, “Oil sludge fuel mixtures with additives of fossil and biomass origin: Energy and operational parameters,” Energy, vol. 316, p. 134643, 2025, doi: https://doi.org/10.1016/j.energy.2025.134643.

11. H. Yu et al., “Pyrolysis characteristics of oil in oily sludge from experiments and simulation by model compounds,” J. Anal. Appl. Pyrolysis, vol. 183, p. 106738, 2024, doi: https://doi.org/10.1016/j.jaap.2024.106738.

12. X. Huang et al., “CPFD numerical study on tri-combustion characteristics of coal, biomass and oil sludge in a circulating fluidized bed boiler,” J. Energy Inst., vol. 113, p. 101550, 2024, doi: https://doi.org/10.1016/j.joei.2024.101550.

13. C. Lu, D. Li, B. Xi, G. Hu, and J. Li, “Machine learning-aided model for predicting oily sludge pyrolysis under various feedstock and operating conditions,” J. Hazard. Mater., vol. 489, p. 137654, 2025, doi: https://doi.org/10.1016/j.jhazmat.2025.137654.

14. Y. Sun et al., “Development of an intelligent design and simulation aid system for heat treatment processes based on LLM,” Mater. Des., vol. 248, p. 113506, 2024, doi: https://doi.org/10.1016/j.matdes.2024.113506.

15. M. C. Duan Hongjie Wang Zhen, Gong Xuchao, Jing Ruilin, Liu He, “Large Language Model of Oil and Gas Cognition Constructed and Applied in Shengli Oilfield.”

16. H. PAN et al., “Construction and preliminary application of large language model for reservoir performance analysis,” Pet. Explor. Dev., vol. 51, no. 5, pp. 1357–1366, 2024, doi: https://doi.org/10.1016/S1876-3804(25)60546-5.

17. S. T. Al Amin, K. Wu, A. Mathur, and C. Koritala, “LLM-In-The-Loop: A Framework for Enhancing Drilling Data Analytics,” Mar. 04, 2025. doi: 10.2118/223805-MS.

18. K. D. Wang, E. Burkholder, C. Wieman, S. Salehi, and N. Haber, “Examining the potential and pitfalls of ChatGPT in science and engineering problem-solving,” Front. Educ., vol. Volume 8-2023, 2024, [Online]. Available: https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2023.1330486

19. R. Kowalik and V. Vukašinović, “LARGE LANGUAGE MODELS AS TOOLS FOR PUBLIC BUILDING ENERGY MANAGEMENT : AN ASSESSMENT OF,” vol. 19, no. 3, 2023.

20. K. Giannakopoulos, A. Kavadella, A. Aaqel Salim, V. Stamatopoulos, and E. G. Kaklamanos, “Evaluation of the Performance of Generative AI Large Language Models ChatGPT, Google Bard, and Microsoft Bing Chat in Supporting Evidence-Based Dentistry: Comparative Mixed Methods Study,” J Med Internet Res, vol. 25, p. e51580, 2023, doi: 10.2196/51580.

21. L. L. M. Chatbot Arena, “Leaderboard: Community-driven Evaluation for Best LLM and AI chatbots,” 2024.

22. G. Balakayeva, G. Kalmenova, and C. Phillips, “Numerical modelling of the process of thermal treatment of oil slime,” Int. J. Oil, Gas Coal Technol., vol. 34, no. 2, pp. 157–172, 2023.

23. H. Ma, T. Peng, C. Zhang, C. Ji, Y. Li, and M. S. Nazir, “Developing an evolutionary deep learning framework with random forest feature selection and improved flow direction algorithm for NOx concentration prediction,” Eng. Appl. Artif. Intell., vol. 123, p. 106367, 2023, doi: https://doi.org/10.1016/j.engappai.2023.106367.

24. M. Rezaei, S. Bahramali Asadi Kelishami, and S. Sangin, “Iran’s comprehensive heat flow map generated by the Random Forest method and the Sequential Gaussian Simulation,” Geothermics, vol. 118, p. 102915, 2024, doi: https://doi.org/10.1016/j.geothermics.2024.102915.

Information about authors

Lei Shangpeng,– PhD student at Al-Farabi Kazakh National University, Almaty, Kazakhstan. ORCID iD: 0009-0002-7156-4631

Gulnar Balakayeva,– Professor, Dr. of mathematics and physics, Al-Farabi Kazakh National University, Almaty, Kazakhstan. ORCID iD:0000-0001-9440-2171

Liu Xikui – Professor, Dr. of mathematics, Shandong University of Science and Technology, Jinan, China, ORCID iD:0000-0002-4509-9468

Submission received: 14 March, 2025.
Revised: 10 July, 2025.
Accepted: 10 July, 2025.

IRSTI 20.23.27 https://doi.org/10.26577/jpcsit2025332

Aisulu Ataniyazova[1,2*], Timur Merembayev[2]

[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan
[2]Institute of Information and Computational Technologies, Almaty, Kazakhstan
[*]e-mail: aisulu.ataniyazova@gmail.com

SPATIOTEMPORAL ASSESSMENT OF SOIL SALINITY IN IRRIGATED AGRICULTURAL LANDS OF KAZAKHSTAN USING REMOTE SENSING

Abstract. Soil salinization poses a significant threat to agricultural productivity and environmental sustainability, particularly in arid and semi-arid regions. This study presents a comprehensive spatiotemporal analysis of soil salinity dynamics in irrigated lands of Alakol District, Zhetisu Region, Kazakhstan, using multi-temporal Sentinel-2 satellite imagery and the Normalized Difference Salinity Index (NDSI). The analysis covered the 2024 growing season, from March to November, with one cloud-free image selected for each month. NDSI values were calculated monthly and classified into four salinity categories: non-saline, slightly saline, moderately saline, and highly saline. Field sampling at 31 locations provided electrical conductivity (EC) data for validation, enabling comparison between surface reflectance-based salinity estimates and ground measurements. The results demonstrated pronounced seasonal trends: NDSI values were lowest in spring due to leaching by precipitation and early irrigation, gradually increasing through summer as evaporation concentrated salts at the surface, and fluctuating in autumn depending on rainfall and drainage conditions. Spatially, fields situated in topographic depressions or near Lake Alakol exhibited the highest salinity levels, whereas upland areas remained relatively unaffected. Notably, no fields exceeded the moderate salinity threshold, indicating that while salinization is present, it remains in early stages. The NDSI approach proved effective for surface salinity detection, capturing both temporal fluctuations and spatial heterogeneity. These findings underscore the utility of remote sensing for operational salinity monitoring and highlight the importance of continuous observation to inform timely land management interventions. This study offers actionable insights for sustainable agriculture, particularly in tailoring irrigation and drainage strategies to mitigate salinity risks across vulnerable farmlands in Central Asia.

Keywords: land degradation, soil salinity, electrical conductivity, remote sensing, satellite images, normalized difference salinity index, spatiotemporal dynamics.

1. Introduction

Soil salinization is a severe form of land degradation that threatens agricultural productivity and ecosystem health worldwide. Traditional methods of mapping soil salinity rely on extensive ground sampling and laboratory analysis, which are labor-intensive, costly, and impractical for large areas. In recent years, remote sensing satellites coupled with machine learning have emerged as efficient tools for assessing and mapping soil salinity across wide regions [1, 2]. Optical sensors (Landsat, Sentinel-2) and radar sensors (Sentinel-1) can detect spectral and backscatter signatures related to surface salt content, while ML algorithms can learn complex relationships between those signatures and ground-measured salinity.

Multi-spectral optical imagery and radar imagery are widely used to detect salinity-induced signals on the soil surface. Salt-affected soils often exhibit characteristic spectral signatures, such as high reflectance in visible and near-infrared bands or distinctive vegetation stress signals. Many studies derive spectral indices to enhance salinity detection. For example, researchers in the Great Hungarian Plain [3] (Eastern Europe) used Landsat 8 to compute vegetation and salinity indices (along with principal components and land surface temperature) as inputs to regression models. In arid regions of Abu Dhabi, indices like NDVI (Normalized Difference Vegetation Index) and BSI (Bare Soil Index) showed moderate correlation with soil electrical conductivity, and their combination improved salinity prediction models [4]. Such indices capture reduced vegetation vigor or exposed bright soils typical of saline areas. However, optical methods have limits: heavy vegetation can mask soil signals, and beyond the top ~5 cm of soil, optical reflectance is less sensitive to salt content [5]. To address this, some studies incorporate thermal infrared data (sensitive to soil moisture and salinity effects) or hyperspectral imagery for more diagnostic spectral features, though these data are less commonly available.

Central Asia has been a focal point for salinity research due to intensive irrigation and desertification. Mukhamediev et al. mapped soil salinity across Turkestan, Almaty, Zhambyl, and Kyzylorda using a fusion of Sentinel-1 SAR and Landsat optical data, combined with machine learning models [6]. Their approach – employing boosted regression trees (XGBoost/LightGBM) outperformed models using optical data alone, showing better agreement with ground EC measurements. Using explainable ML, they also optimized feature selection without reducing accuracy. Notably, their regional model outperformed a global salinity model, highlighting the importance of local calibration. Merembayev et al. studied salinity in arid irrigated farms using high-resolution radar textures and ML [7]. They highlighted strong spatial heterogeneity in salinity due to variable soil and irrigation conditions, which complicates mapping efforts. Their results showed that careful data partitioning and maintaining representative value distributions were key for model performance. LightGBM and Ridge regression achieved the best results (R^2 ~0.68). The authors suggest future work should explore deep learning and physics-based models to enhance accuracy.

Researchers in [8] assessed soil salinity changes under climate change in the Khorezm region of Uzbekistan – an area with extensive irrigation. Their analysis noted that over the last 40 years, soil salinity has increased due to rising temperatures and poor drainage, and using saline groundwater for irrigation has exacerbated secondary salinization. By integrating remote sensing with climate data, they linked periods of warming and reduced river inflows to spikes in soil salt levels, predicting that climate change will continue to aggravate salinity unless irrigation management improves. A broader scale analysis by [9] provides a striking forecast for Central Asia. Using an automated ML framework to analyze drivers of salinity across Central Asia and neighboring Xinjiang (western China), they found that meteorological factors (aridity, temperature) exert the strongest influence on soil salt content, often interacting with landscape position (e.g. low-lying basins). Their model projected that with extreme climate warming scenarios, average soil salt concentrations could rise by about +21% in Central Asia by 2100, and as much as +65% in Xinjiang. Areas around irrigation water sources and topographic low points are at highest risk of salinity escalation. This study's methodology – using ML to parse out interaction effects between climate, topography, and human factors – is an innovative approach to understand spatio-temporal dynamics. It provides a quantitative glimpse into the future, highlighting that without intervention, Central Asia's salinization will intensify under climate change.

A major advantage of satellite-based monitoring is the ability to track salinity changes over time – capturing both seasonal fluctuations and long-term trends. Historically, most remote sensing salinity studies focused on mapping spatial patterns at a single time, often neglecting the temporal dimension [10]. Recent work is beginning to fill this gap by leveraging multi-temporal image series and time-series analysis. In irrigated areas, soil salinity fluctuates seasonally–often rising during dry periods due to evaporation and decreasing after rain or irrigation. Dense time-series from Sentinel-2 have been used to detect such patterns. For instance, [11] identified higher salinity in the dry season in China's Ebinur Lake wetland using RF models. Similarly, a study in the Werigan–Kuqa Oasis [12] found salinity shifts linked to precipitation variability, with expansion during droughts and retraction in wetter years. Multi-date imagery reduces noise and improves model accuracy. Duan et al. proposed a "combined-temporal" approach using multiple Sentinel-2 images around the sampling date [13]. This stabilized spectral signatures and improved model performance ($R^2 = 0.72$, RMSE ≈ 0.87 dS/m). However, salinity signals are affected by vegetation phenology and cropping cycles, necessitating integration with seasonal land cover and crop type data.

Comparing salinity maps over years reveals degradation or improvement patterns. In Zaghouan, Tunisia, salinized areas expanded from 2000 to 2023, linked to reduced rainfall and land use change ($r \approx -0.85$ with precipitation), highlighting climate change as a key driver [13]. In contrast, the Xinjiang Oasis saw a salinity decline over 25 years, with non-salinized land increasing and severe salinity retreating due to improved irrigation and drainage. Evidence from Iran's Golestan Province also showed salinity reduction after drainage system installation, confirmed via satellite data. Meanwhile, in Central Asia's Kashgar region [14], time-series analyses revealed worsening salinity due to reclamation without drainage–mirroring issues in the Aral Sea basin.

A few studies directly compared different remote sensing approaches for mapping soil salinity under various environmental conditions. For instance, [15] compared Sentinel-2 vs Landsat-8 imagery for salinity mapping in a Mediterranean site, finding Sentinel-2's higher resolution gave it an edge in detecting fine-scale salinity patches. Conversely, another study in a Chinese wetland found Landsat-8's inclusion of a thermal band (absent in Sentinel-2) made it slightly superior for salinity estimation using a cubist model [16]. Such comparisons suggest that the "best" satellite platform may vary with context – Sentinel-2 excels with spatial detail and revisit frequency, whereas Landsat's thermal data can help in humid areas where evapotranspiration differences are key. In practice, many studies now use both (taking advantage of the combined 5-day revisit of Sentinel-2 and Landsat-8). On the ML side, comparisons like those by [17] in GEE have provided valuable guidance – they noted that while a CART model achieved the lowest error on training data, it tended to overfit extreme salinity, whereas RF provided more reliable generalization across landscapes. This hints that for operational mapping, a slightly less "accurate" but more stable model (RF) may be preferable to avoid speckled or noisy salinity maps.

Reliable ML modeling hinges on quality ground truth data, yet in many regions soil salinity sampling is sparse and infrequent due to the costs and effort required [18]. Small sample sizes can lead to unstable models. Most remote sensing methods primarily sense surface salt. This is problematic because harmful salinity can build up below the topsoil and escape detection until it surfaces [19, 20]. There is consensus that passive optical methods alone cannot fully capture subsoil salinity – thus, research is heading toward combining satellite data with geophysical surveys (EM induction, resistivity) or soil hydraulic models to infer salt distribution in the profile. Thus, despite existing advances, using remote sensing or machine learning in isolation often faces limitations – such as sensitivity to seasonal variability, limited model transferability, or instability in interpreting temporal signals. The present study is aimed at assessing the spatial and temporal dynamics of soil salinity within the selected territory in the period from March to November. Based on the interpretation of satellite images and data processing in the GIS environment, an analysis of salinity changes has been performed, which allows not only to identify seasonal trends, but also to justify the need for sustainable approaches to land management.

2. Materials and methods

The study was conducted in the agricultural lands of Alakol District, Zhetisu region, Kazakhstan, encompassing a semi-arid continental environment. The area lies approximately between 46°N and 81°E (near Lake Alakol's basin), characterized by hot, dry summers and cold winters . Annual precipitation is low (≈150–280 mm), with most rain falling in spring (April–May) and late autumn. This climate and intensive irrigation practices make the region prone to soil salinization, as evaporation often exceeds rainfall, leading to salt accumulation at the surface. Thirty-one sampling sites were selected across the district's farmlands to capture spatial variability. At each site, we collected surface soil samples and measured electrical conductivity in a 1:5 soil-water extract as a ground truth indicator of salinity. The EC values ranged from 0.07 (non-saline) up to 1.4 (highly saline) in the 1:5 extract. We also recorded terrain attributes at each site: elevation (from a DEM), local slope, and ambient surface temperature. Elevations spanned ~379–495 m above sea level (lower toward the lake plain, higher in uplands), and slopes were gentle (mostly <1° incline). Notably, the lowest-lying fields tended to have higher measured salinity, consistent with salt accumulation in topographic depressions. Table 1 summarizes the field data, including coordinates, electrical conductivity (EC), elevation, surface temperature, and slope values for each of the 31 sampling sites.

Table 1 – Field data

Field №	X	Y	EC (1:5), dS/m	DEM, m	Surface temp., °C	Slope
ALK1	81.263687	46.013337	0.16	387	42.35	0.8863131403923035
ALK2	81.263322	46.013021	0.2	387	42.35	0.8863131403923035
ALK3	81.262991	46.012058	0.44	387	42.35	0.8863131403923035
ALK4	81.252822	46.014737	0.15	386	42.78	0.4236890375614166
ALK5	81.252683	46.014185	0.48	386	42.78	0.4236890375614166
ALK6	81.258905	46.033323	0.35	380	39.97	0.41753801703453064
ALK7	81.262467	46.033464	0.52	380	39.16	0.41753801703453064
ALK8	81.235861	46.061892	0.37	396	39.55	0.5087683796882629
ALK9	81.236126	46.062183	0.38	396	39.55	0.5087683796882629
ALK10	81.227602	46.09355	0.19	395	41.56	0.5410060882568359
ALK11	81.226884	46.093285	0.13	395	41.56	0.5410060882568359
ALK12	81.20694	46.153272	0.34	369	40.9	0.272049218416214
ALK13	81.184885	46.143835	0.4	384	39.91	0.49335935711860657
ALK14	81.175125	46.125639	0.73	401	40.55	0.5325705409049988
ALK15	81.183753	46.120222	0.46	404	41.21	0.5124315023422241
ALK16	81.184296	46.120402	0.41	404	41.21	0.5124315023422241
ALK17	81.15495	46.164803	0.22	382	37.76	0.4547281265258789
ALK18	81.155732	46.165621	0.21	382	37.76	0.4547281265258789
ALK19	81.155817	46.165726	0.21	382	37.76	0.4547281265258789
ALK20	81.045543	46.194957	0.22	379	39.85	0.30169597268104553
ALK21	81.046017	46.194684	0.21	379	39.76	0.30169597268104553
ALK22	80.988318	46.183279	0.21	390	40.75	0.43475303053855896
ALK23	80.988355	46.183294	0.21	390	40.75	0.43475303053855896
ALK24	80.830143	46.244836	0.2	381	44.13	0.13844919204711914
ALK25	80.829652	46.212158	0.07	384	45.57	0.17097853124141693
ALK26	80.829596	46.213915	0.08	384	45.57	0.17097853124141693
ALK27	80.829592	46.213916	1.4	384	45.57	0.17097853124141693
ALK28	81.054566	46.069562	0.36	495	40.66	0.5086142420768738
ALK29	81.099052	46.036487	0.53	492	41.56	0.5497124791145325
ALK30	81.100584	46.036445	0.47	492	40.5	0.5497124791145325
ALK31	81.101486	46.036861	0.38	492	40.5	0.5497124791145325

Fig. 1 presents the collected soil samples, which were used for laboratory measurement of electrical conductivity (EC) as ground-truth data.

Fig. 2 shows the geographical location of the study area within Alakol District, Zhetisu region, including the spatial distribution of sampling sites.

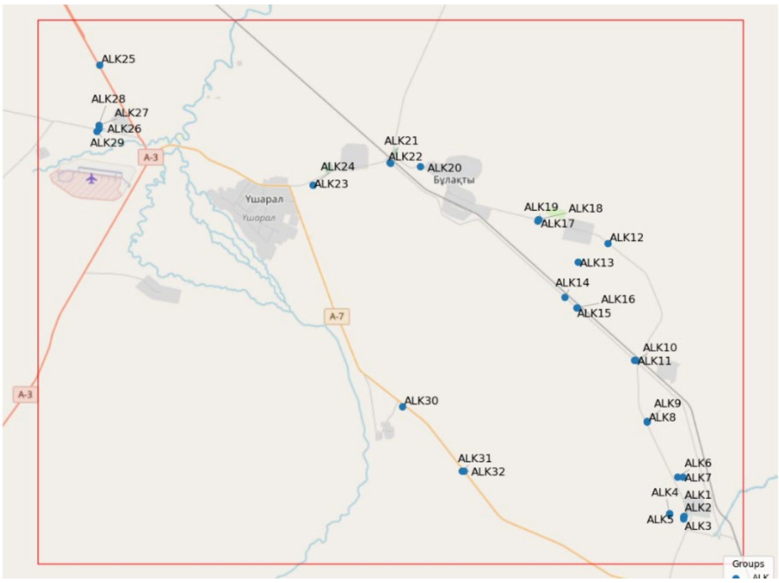


Figure 1 – Soil samples collected from agricultural fields

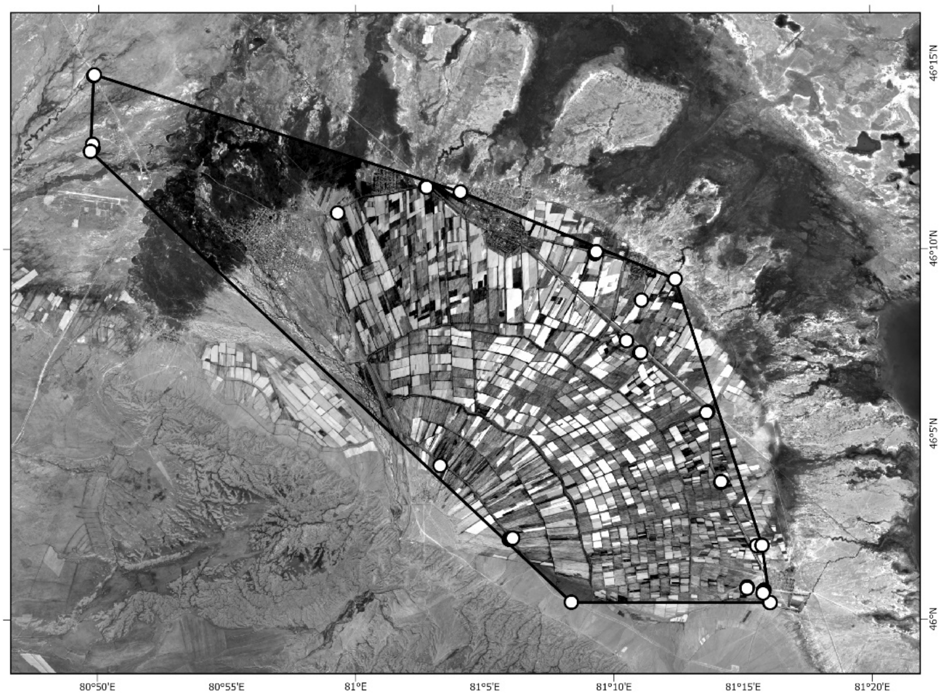


Figure 2 – Location of the study area and distribution of field sampling sites

To monitor salinity dynamics over the 2024 growing season, we acquired Sentinel-2 MultiSpectral Instrument (MSI) imagery for each month from March through November (one cloud-free scene per month). Sentinel-2 provides 13 spectral bands, including visible, near-infrared (NIR), and shortwave-infrared (SWIR) wavelengths, at spatial resolutions of 10–20 m, with a 5-day revisit frequency. We downloaded Level-2A surface reflectance products (which are already atmospherically corrected to bottom-of-atmosphere reflectance) covering the study area. Each monthly image was projected to the WGS 84 / UTM Zone 44N coordinate system and clipped to the boundaries of the target agricultural fields. Cloud masking was applied using the Sentinel-2 Scene Classification Layer (SCL) to remove cloud- and shadow-affected pixels. We ensured minimal cloud cover by selecting images on or near clear-sky dates for each month; if the primary monthly image had cloud contamination, an alternate cloud-free image from the same month was used. This preprocessing workflow yielded a time-series of nine cloud-free reflectance maps (March–November 2024) for the region. All images were co-registered to ensure that multi-date pixel-wise comparisons were spatially consistent, and radiometric consistency was maintained by using the atmospherically corrected reflectances (ensuring comparability across dates). We also extracted reflectance and index values at the 31 field sampling locations for each date to facilitate direct comparison with ground measurements of salinity. Fig. 3 illustrates the sequential workflow for mapping soil salinity using Sentinel-2 satellite imagery. The process comprises six main stages:

- High-resolution multispectral imagery from Sentinel-2 is acquired for the area of interest. These data provide the necessary spectral information to detect surface-level variations in soil properties, including salinity.
- Specific spectral bands (such as the Red, Green, and Near Infrared (NIR) bands) are selected based on their sensitivity to soil salinity and moisture content. These bands serve as the input for index-based salinity assessments.
- The selected images undergo preprocessing steps, including atmospheric correction, resampling, and cropping. The cropping operation ensures that the imagery conforms to the boundaries of the study area, facilitating localized analysis.

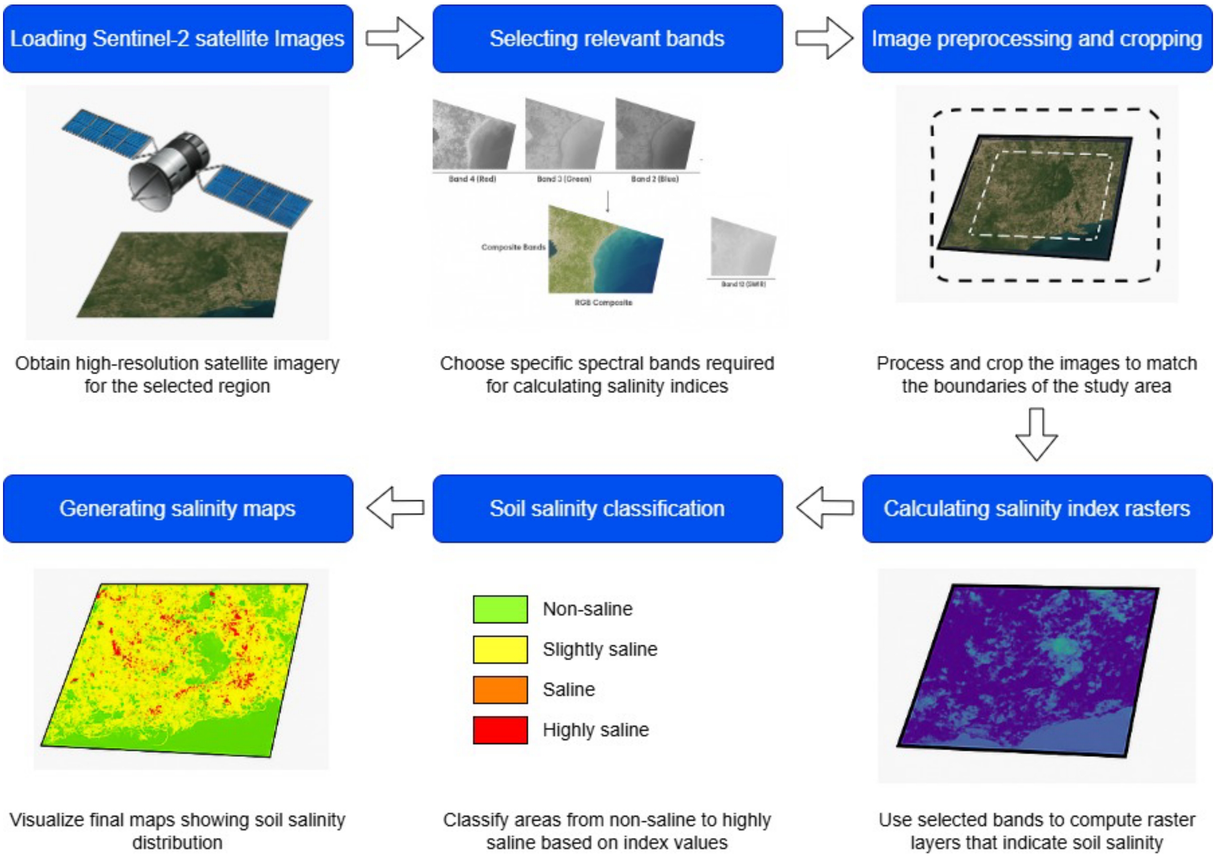

Loading Sentinel-2 satellite Images
Selecting relevant bands
Image preprocessing and cropping
Band 4 (Red)
Band 3 (Green)
Band 2 (Blue)
Composite Bands
RGB Composite
Band 12 (SWIR)
Obtain high-resolution satellite imagery for the selected region
Choose specific spectral bands required for calculating salinity indices
Process and crop the images to match the boundaries of the study area
Generating salinity maps
Soil salinity classification
Calculating salinity index rasters
🟩 Non-saline
🟨 Slightly saline
🟧 Saline
🟥 Highly saline
Visualize final maps showing soil salinity distribution
Classify areas from non-saline to highly saline based on index values
Use selected bands to compute raster layers that indicate soil salinity

Figure 3 – Stages of soil salinity assessment based on remote sensing

- Using the relevant bands, salinity indices are computed to generate raster layers. These layers indicate spatial variations in soil salinity levels based on spectral reflectance properties.

- The index values are classified into predefined salinity categories (non-saline, slightly saline, moderate saline, and highly saline). This classification supports the interpretation and evaluation of salinity severity across the landscape.
- The final classified raster outputs are visualized as salinity maps.

3. Results

The NDSI analysis from March to November 2024 shows clear temporal and spatial patterns of soil salinity across the study area. To interpret the index values, we categorised NDSI into four classes: non-saline, slightly saline, moderately saline and highly saline (Table 2).

The NDSI was calculated using the spectral bands Red (Band 4) and Near-Infrared (NIR, Band 8) from Sentinel-2 imagery [21]:

$$\mathrm{NDSI} = \frac{(\mathrm{Red}-\mathrm{NIR})}{(\mathrm{Red}+\mathrm{NIR})} \quad (1)$$

Across all nine months, the observed NDSI values ranged approximately from –0.58 to –0.03 (Table 3).

Table 2 – Soil salinity classification.

Categories	**EC$_{(1:5)}$**	**NDSI**
Non-saline	<0,16	<0,2
Slightly saline	0,16≤x<0,22	-0,2≤x<0
Moderately saline	0,22≤x<0,74	0≤x<0,2
Highly saline	≥0,74	≥0,2

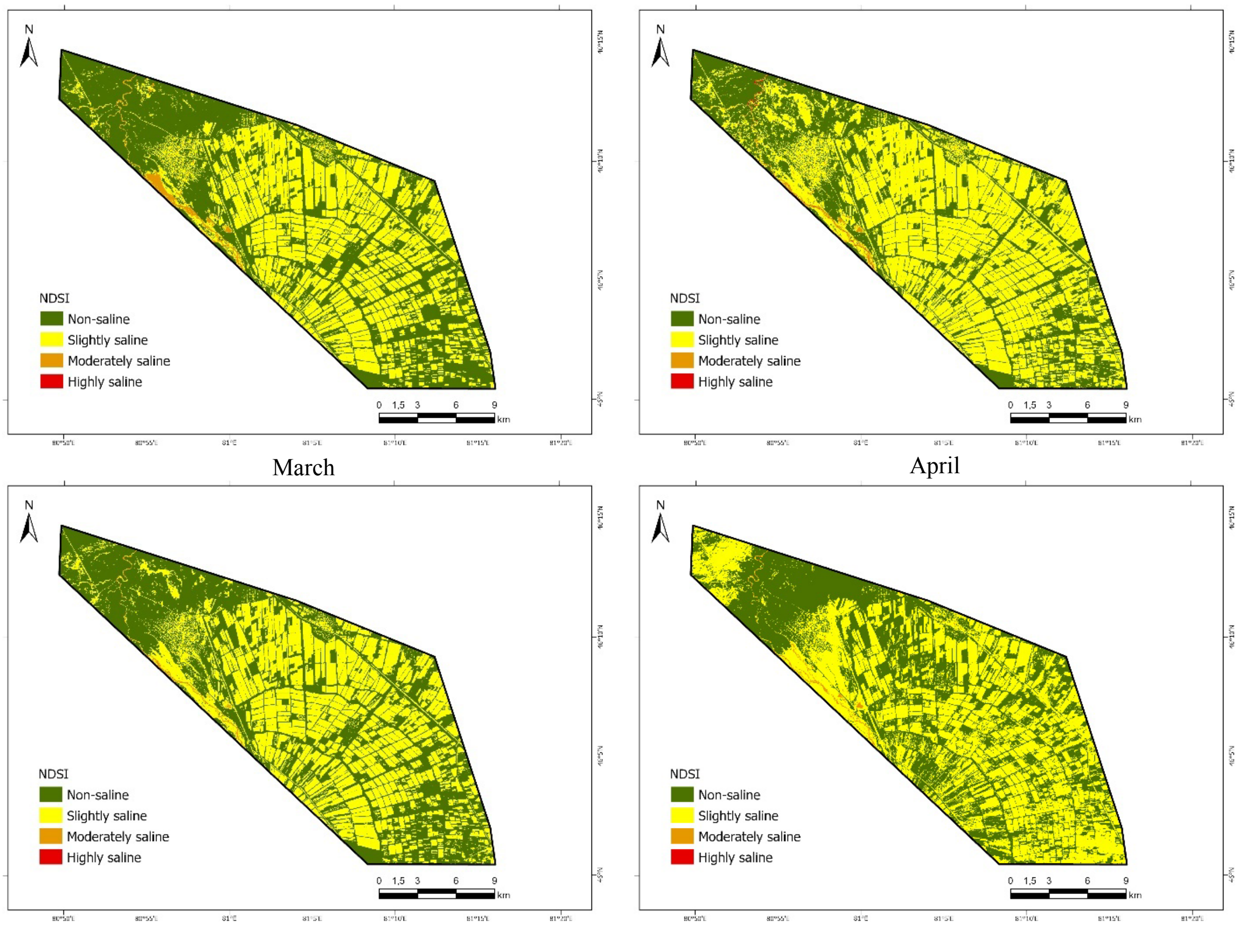
Table 3 – Normalized difference salinity index measurements.

Field №	**NDSI**								
	Mar	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**
ALK1	-0,171	-0,272	-0,295	-0,184	-0,130	-0,121	-0,144	-0,126	-0,112
ALK2	-0,145	-0,208	-0,247	-0,161	-0,107	-0,094	-0,102	-0,099	-0,113
ALK3	-0,112	-0,207	-0,223	-0,224	-0,171	-0,107	-0,114	-0,101	-0,103
ALK4	-0,117	-0,180	-0,093	-0,105	-0,178	-0,578	-0,532	-0,220	-0,075
ALK5	-0,128	-0,195	-0,150	-0,222	-0,282	-0,406	-0,493	-0,258	-0,117
ALK6	-0,128	-0,264	-0,364	-0,367	-0,448	-0,457	-0,509	-0,367	-0,111
ALK7	-0,098	-0,122	-0,103	-0,175	-0,252	-0,478	-0,373	-0,264	-0,141
ALK8	-0,132	-0,268	-0,309	-0,430	-0,481	-0,364	-0,286	-0,194	-0,116
ALK9	-0,101	-0,146	-0,171	-0,285	-0,292	-0,221	-0,253	-0,172	-0,090
ALK10	-0,169	-0,267	-0,289	-0,163	-0,112	-0,149	-0,220	-0,193	-0,053
ALK11	-0,081	-0,135	-0,158	-0,101	-0,521	-0,520	-0,297	-0,346	-0,128
ALK12	-0,090	-0,193	-0,232	-0,322	-0,308	-0,326	-0,316	-0,178	-0,087
ALK13	-0,104	-0,147	-0,197	-0,086	-0,304	-0,215	-0,106	-0,106	-0,104
ALK14	-0,083	-0,182	-0,163	-0,208	-0,198	-0,221	-0,129	-0,116	-0,033
ALK15	-0,138	-0,247	-0,280	-0,226	-0,126	-0,113	-0,150	-0,134	-0,078
ALK16	-0,183	-0,291	-0,284	-0,170	-0,121	-0,115	-0,140	-0,130	-0,128
ALK17	-0,110	-0,095	-0,242	-0,140	-0,116	-0,133	-0,177	-0,164	-0,138
ALK18	-0,078	-0,128	-0,100	-0,201	-0,403	-0,449	-0,420	-0,189	-0,057
ALK19	-0,078	-0,128	-0,100	-0,201	-0,403	-0,449	-0,420	-0,189	-0,057
ALK20	-0,079	-0,090	-0,160	-0,271	-0,398	-0,268	-0,231	-0,157	-0,063
ALK21	-0,064	-0,112	-0,135	-0,242	-0,427	-0,356	-0,162	-0,152	-0,109
ALK22	-0,201	-0,330	-0,391	-0,247	-0,168	-0,137	-0,137	-0,116	-0,185
ALK23	-0,117	-0,266	-0,485	-0,457	-0,189	-0,118	-0,125	-0,118	-0,156
ALK24	-0,162	-0,253	-0,283	-0,229	-0,217	-0,238	-0,246	-0,196	-0,184
ALK25	-0,203	-0,288	-0,295	-0,267	-0,214	-0,181	-0,176	-0,139	-0,053

Continuation of the table

Field №	NDSI								
	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov
ALK26	-0,185	-0,273	-0,232	-0,190	-0,143	-0,133	-0,145	-0,122	-0,120
ALK27	-0,185	-0,273	-0,232	-0,190	-0,143	-0,133	-0,145	-0,122	-0,120
ALK28	-0,107	-0,096	-0,066	-0,236	-0,253	-0,173	-0,133	-0,107	-0,136
ALK29	-0,088	-0,135	-0,186	-0,193	-0,140	-0,106	-0,109	-0,105	-0,124
ALK30	-0,068	-0,108	-0,142	-0,101	-0,072	-0,056	-0,095	-0,077	-0,081
ALK31	-0,155	-0,212	-0,215	-0,138	-0,109	-0,089	-0,096	-0,091	-0,097

Notably, almost none of the sampled locations exceeded an NDSI of 0, meaning moderate or high salinity levels were not reached in surface reflectance during 2024. Instead, most values fell in the non-saline or slightly saline categories. Figure 4, which displays monthly salinity maps, visually corroborates these findings by illustrating the expansion and contraction of saline areas over the seasons.

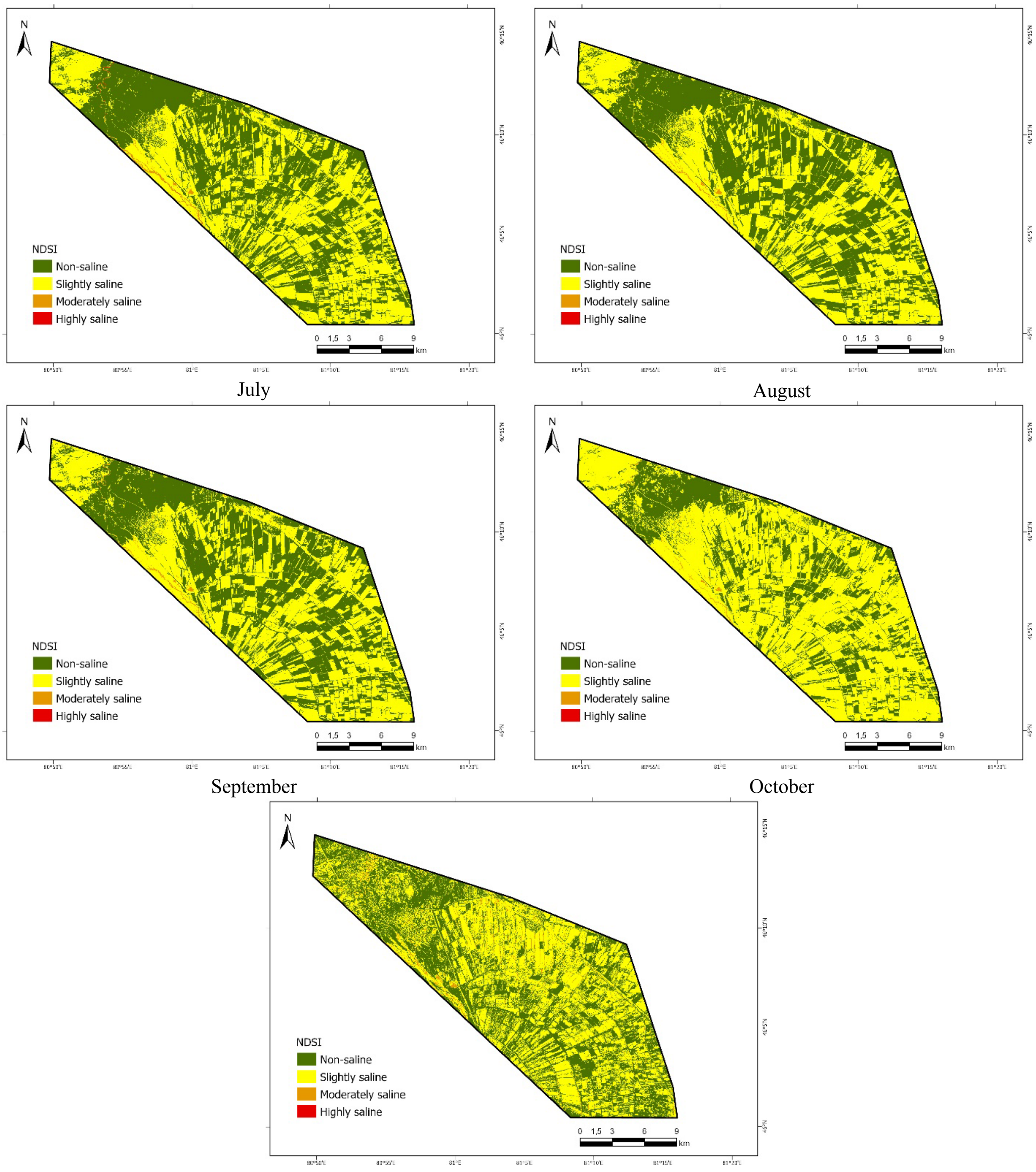


March

April

May

June



July

August

September

October

November

Figure 4 – Soil salinity distribution maps

- In early spring (March–April), soil salinity was generally low. March NDSI values averaged around –0.15 to –0.20 in many fields. By April, additional rainfall and early irrigation likely leached some surface salt, resulting in even more negative NDSI values in several fields. Many sites in April and May recorded NDSI below –0.2. This period represents the annual minimum for surface salinity; the land had been flushed by spring moisture, leaving little salt at the surface. For instance, ALK5 had an NDSI of –0.30 in May, compared to –0.16 in March – a drop indicating reduced salinity as spring progressed.

- Moving into summer, the trend reverses. By June, as temperatures rose and soils began drying, NDSI values in many fields started to increase, signaling a resurgence of salinity at the surface. The peak of summer (July–August) showed the most significant salinity levels. During July, a majority of fields shifted into the slightly saline category. By August, nearly all fields exhibited higher NDSI compared to spring: values commonly ranged from about –0.15 up to –0.08. A few fields even approached the threshold of moderate salinity – for example, one low-lying field reached an NDSI of –0.05 in August, the highest value observed. Although these values remained just below zero, they indicate that salts had considerably accumulated on the soil surface by late summer.
- In the autumn months (September–November), salinity levels exhibited slight improvements in some fields, while others remained high. September's NDSI values were still elevated (mostly in the –0.1 to –0.18 range), not significantly different from August in many cases. However, by October, a modest decrease in salinity is evident in a number of fields: for instance, fields that had NDSI around –0.10 in late summer dropped to approximately –0.13 to –0.15 in October. This suggests that as temperatures fell and crop water use declined, there was less evaporation to concentrate salts, and any early autumn rainfall may have begun to dissolve or move salts downward.
- By November, a few fields continued to show some of the highest salinity readings of the year (NDSI ≈ –0.05 to –0.08, remaining in the slightly saline class despite the season), especially those that are poorly drained. In other fields, November brought further slight decreases in NDSI (down to ~–0.18 to –0.20), nearly returning to springtime non-saline levels.

To assess the validity of the NDSI satellite indicator, a correlation analysis was performed using the Pearson correlation coefficient between EC and NDSI [22]:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - y)^2}} \qquad (2)$$

where x_i – EC value for i field,

y_i – NDSI value for i field,

$\bar{x}$ – average EC value,

$\bar{y}$ – average NDSI value.

To estimate how much the observed value deviates from the expected value, if the null hypothesis is correct, we will calculate the t-statistics:

$$t = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}} \qquad (3)$$

After calculating t, this value will be compared with the Student's distribution (t-distribution) with df = n-2:

$$p = 2 \cdot (1 - T_{CDF}(|t|, df)) \qquad (4)$$

T_{CDF} – cumulative distribution function,

df = n-2

The calculation results are shown in Table 4. There is a weak but positive relationship between the EC and NDSI values, which gives the right to use NDSI as a reliable indicator of salinity for monitoring large areas without the need for continuous sampling. Non-simultaneous measurements are a limitation for calculations, since the EC is taken at one moment, and the NDSI changes monthly.

Table 4 – NDSI satellite indicator validity assessment.

Month	r	t	p-value	Interpretation
March	–0.0047	–0.025	0.980	No connection
April	–0.0260	–0.140	0.889	No connection
May	+0.1128	+0.611	0.546	Weak positive connection
June	+0.0494	+0.266	0.792	No connection
July	+0.1889	+1.036	0.309	Weak positive connection
August	+0.1775	+0.972	0.339	Weak positive connection
September	+0.1563	+0.852	0.401	Weak positive connection
October	+0.1673	+0.914	0.368	Weak positive connection
November	+0.0147	+0.079	0.938	No connection

Superimposed on these temporal trends are distinct spatial patterns in salinity. Certain fields consistently showed higher salinity than others, underlining the role of site-specific factors. In particular, fields situated at lower elevations or poorly drained positions were much more prone to salinisation. For example, ALK27 (located in a topographic depression near the lake plain) had NDSI values of approximately –0.28 in May (virtually non-saline after spring rains) but rose to around –0.06 by November, categorising it as one of the most saline fields by year's end. In contrast, ALK28, which lies on higher ground, ranged from about –0.25 (May) to –0.12 (August), never exceeding the slightly saline range and ending the season near –0.18 in November. This comparison illustrates that the low-lying field accumulated and retained far more salt over the season than the upland field. Such patterns were typical: nearly all the lowest-lying fields had the highest salinity readings, whereas fields at higher elevations or with better natural drainage remained relatively less affected. This observation aligns with the ground truth data, which showed that the lowest elevation sites had the highest soil EC measurements.

Series of maps in Fig. 4 highlights these spatial differences – the same areas (notably, the northern and central parts of the district closer to Lake Alakol) repeatedly show up as saline-hued zones in summer and autumn, whereas southern and higher-elevation plots stay blue (non-saline) throughout.

4. Discussion

The spatio-temporal patterns observed in this study reflect the interplay of climatic, hydrological, and land-use factors characteristic of semi-arid irrigated environments. During spring, precipitation and irrigation water dilute and leach salts from the topsoil, whereas in summer, high evaporation rates draw moisture up, causing dissolved salts to crystallise at the surface [23, 24]. Our results confirm this cycle – the pronounced increase in salinity from May to August indicates evaporative salt concentration under hot, dry conditions. By late autumn, the slight reductions in NDSI in some fields suggest that cooler temperatures and occasional rain may have partially re-dissolved surface salts. However, the fact that many fields remained more saline in November than in March implies that seasonal flushing was incomplete. In practice, this means salts can carry over into the next year, leading to a gradual buildup if not managed. Thus, even though salinity may appear to recede each spring, the summer accumulations pose a recurring stress that can contribute to long-term soil degradation if proper remediation is not in place.

Topography and water flow emerge as critical drivers of the spatial salinity patterns [25]. Fields in depressions or near the lake plain consistently showed higher salinity, which is consistent with water pooling and evaporating in these low-lying areas, leaving behind salt deposits. In contrast, fields on slight rises or with better drainage had lower NDSI values, as excess water (and salt) could more easily percolate away. This observation corresponds with well-known behaviour of salts accumulating in landscape low points. Similar findings have been reported in other Central Asian studies [8, 9] – for instance, a regional analysis noted that areas around irrigation water sources and topographic low points are at highest risk of salinity escalation. Our field-scale heterogeneity (where adjacent fields had quite different salinity levels) also echoes the work of [7], who found strong spatial variability in salinity due to differences in soil properties and irrigation practices even within a small area. Such comparisons highlight that local factors (micro-relief, irrigation scheduling, soil texture, etc.) can cause significant divergence in salinity outcomes, underlining the importance of site-specific management strategies. Regarding the efficacy of the NDSI approach, our use of a simple spectral index proved effective in capturing surface salt dynamics.

The temporal trends in NDSI aligned with expected seasonal salinity changes and qualitatively matched ground EC data–fields with higher EC generally showed less negative NDSI values. This confirms NDSI's usefulness as a rapid, cost-effective tool for surface salinity monitoring. Its sensitivity enabled detection of subtle monthly variations and emerging salinity hotspots. However, NDSI reflects only surface conditions and can be influenced by vegetation or soil moisture. In areas with dense crop cover, salinity may be underestimated due to spectral masking. Moreover, similar reflectance can result from dry soil or carbonates, reducing specificity. In this study, the predominantly bare fields enhanced the reliability of NDSI, though its effectiveness may decline in heavily vegetated areas.

Another limitation of NDSI is its inability to detect subsurface salinity–salts beneath a thin surface layer may go unnoticed until they re-emerge. To address this, integrating radar data (e.g., Sentinel-1) can improve detection, as radar is sensitive to surface roughness and moisture and is unaffected by vegetation cover. Studies in Kazakhstan confirm that combining optical and radar imagery enhances salinity mapping accuracy [6, 7]. Thermal infrared data could also help by revealing moisture and evaporation patterns linked to salinity. Additionally, machine learning models that integrate multiple indices and auxiliary data (e.g., terrain, climate) can improve prediction. While NDSI was effective for surface monitoring in this study, a multi-source, multi-index approach would offer a more comprehensive assessment.

5. Conclusion

This study provided a detailed spatio-temporal assessment of soil salinity in irrigated agricultural lands of Alakol District, Kazakhstan, using time-series Sentinel-2 imagery. By tracking the NDSI over the 2024 growing season (March to November), we identified clear seasonal patterns: salinity was lowest in spring after winter and early rains, increased markedly in summer due to evaporation and irrigation practices, and persisted into autumn to varying degrees across the landscape. Spatial analysis further revealed that salinity issues are concentrated in specific areas – notably, low-lying fields near the lake basin experienced the greatest salt accumulation, whereas upland fields were relatively less affected. Importantly, the salinity levels observed (as indicated by NDSI) remained in the slight to moderate range, with no extreme salinity outbreaks during the study period. This suggests that while salinisation is a concern, it may still be at a manageable stage if addressed promptly. The findings underscore the importance of monitoring soil salinity over time. A one-off measurement provides only a snapshot; in contrast, the temporal approach adopted here captures the dynamic nature of salinity, revealing when peaks occur and when remediation would be most needed.

Although the statistical correlation between the Normalized Difference Salinity Index (NDSI) and field-measured electrical conductivity (EC) was modest, the index effectively captured distinct seasonal and spatial patterns of soil salinity. Given the limitations in ground data frequency and temporal alignment, NDSI values should not be interpreted in absolute terms. Instead, they should be regarded as qualitative indicators of salinity variation, capable of supporting spatiotemporal monitoring and the identification of emerging salinity hotspots. Future research should aim to improve ground validation protocols through more frequent and temporally aligned EC sampling and explore the integration of NDSI with complementary data sources–such as radar imagery, soil moisture metrics, and topographic parameters – using hybrid or machine learning-based models to enhance the accuracy and robustness of salinity assessments.

Funding

This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan grant number AP23488281.

Author Contributions

Conceptualization, A.A. and M.T.; Methodology, A.A. and M.T.; Software, A.A. and M.T.; Validation, A.A. and M.T.; Formal Analysis, A.A.; Investigation, A.A. and M.T.; Resources, A.A. and M.T.; Data Curation, A.A. and M.T.; Writing – Original Draft Preparation, A.A.; Writing – Review & Editing, A.A. and M.T.; Visualization, A.A.; Supervision, M.T.; Project Administration, M.T.; Funding Acquisition, M.T.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. S.K. Sarkar, R. Rudra, A. Reza, P. Das and A. Islam, “Coupling of machine learning and remote sensing for soil salinity mapping in coastal area of Bangladesh,” *Scientific Reports*, vol. 13, 2023. doi: 10.1038/s41598-023-44132-4.

2. R. Taghizadeh-Mehrjardi, “Sentinel-2 and Ensemble ML for Soil Salinity Mapping in Iran,” *Geoderma*, vol. 409, 2022.

3. G. Sahbeni, “Soil salinity mapping using Landsat 8 OLI data and regression modeling in the Great Hungarian Plain.” *Applied Sciences*, vol. 3, no. 587, 2021. doi: 10.1007/s42452-021-04587-4.

4. A.S. Alqasemi, M. Ibrahim, A.Q. Fadhil and G. Kaplan, “Detection and modeling of soil salinity variations in arid lands using remote sensing data,” *Open Geosciences*, vol. 13, no. 1, 2021, pp. 443-453. doi: 10.1515/geo-2020-0244.

5. F. Wang et al., “Advancements and perspective in the quantitative assessment of soil salinity utilizing remote sensing and machine learning algorithms: A review,” *Remote Sens*., vol. 16, no. 24, p. 4812, 2024, doi: 10.3390/rs16244812.

6. R. I. Mukhamediev et al., “Soil salinity estimation for South Kazakhstan based on SAR Sentinel-1 and Landsat-8,9 OLI data with machine learning models,” *Remote Sens*., vol. 15, no. 17, p. 4269, 2023, doi: 10.3390/rs15174269.

7. Y. Amirgaliyev, R. Mukhamediev, T. Merembayev et al., “Remote sensing and machine learning algorithms to predict soil salinity in southern Kazakhstan,” *Discover Sustainability*, vol. 5, p. 363, 2024, doi: 10.1007/s43621-024-00594-8.

8. M. Khamidov, J. Ishchanov, A. Hamidov, C. Donmez, and K. Djumaboev, “Assessment of soil salinity changes under the climate change in the Khorezm region, Uzbekistan,” Int. J. Environ. Res. *Public Health*, vol. 19, no. 14, p. 8794, 2022, doi: 10.3390/ijerph19148794.

9. L. Wang et al., “An automated framework for interaction analysis of driving factors on soil salinization in Central Asia and Western China,” *Remote Sens*., vol. 17, no. 6, p. 987, 2025, doi: 10.3390/rs17060987.

10.Y. Zhang et al., “Mapping the soil salinity distribution and analyzing its spatial and temporal changes in Bachu County, Xinjiang, based on Google Earth Engine and machine learning,” *Agriculture*, vol. 14, no. 4, p. 630, 2024, doi: 10.3390/agriculture14040630.

11. S. Bandak, S. A. Movahedi-Naeini, S. Mehri, and A. Lotfata, “A longitudinal analysis of soil salinity changes using remotely sensed imageries,” *Sci. Rep*., vol. 14, no. 1, p. 10383, May 2024, doi: 10.1038/s41598-024-60033-6.

12. S. Ma, B. He, B. Xie, X. Ge, and L. Han, “Investigation of the spatial and temporal variation of soil salinity using Google Earth Engine: a case study at Werigan-Kuqa Oasis, West China,” *Sci. Rep*., vol. 13, no. 1, p. 2754, Feb. 2023, doi: 10.1038/s41598-023-27760-8.

13. C. Duan, Y. Zhang, C. Hu, H. Chen, and P. Liu, “Soil salinity inversion by combining multi-temporal Sentinel-2 images near the sampling period in coastal salinized farmland,” *Front. Environ. Sci*., vol. 13, p. 1533419, 2025, doi: 10.3389/fenvs.2025.1533419.

14. N. Erkin et al., “Method for predicting soil salinity concentrations in croplands based on machine learning and remote sensing techniques,” *J. Appl. Remote Sens*., vol. 13, no. 3, pp. 034520–034520, 2019, doi: 10.1117/1.JRS.13.034520.

15. E. Davis, C. Wang, and K. Dow, “Comparing Sentinel-2 MSI and Landsat 8 OLI in soil salinity detection: A case study of agricultural lands in coastal North Carolina,” *International Journal of Remote Sensing*, vol. 40, no. 16, pp. 6134–6153, 2019, doi: 10.1080/01431161.2019.1587205.

16. J. Wang et al., “Machine learning-based detection of soil salinity in an arid desert region, Northwest China: A comparison between Landsat-8 OLI and Sentinel-2 MSI,” *Sci. Total Environ*., vol. 707, p. 136092, 2020, doi: 10.1016/j.scitotenv.2019.136092.

17. S. Aksoy et al., “Assessing the performance of machine learning algorithms for soil salinity mapping in Google Earth Engine platform using Sentinel-2A and Landsat-8 OLI data,” *Adv. Space Res.,* vol. 69, no. 2, pp. 1072–1086, 2022, doi: 10.1016/j.asr.2021.10.024.

18. J. W. Sirpa-Poma et al., “Towards the improvement of soil salinity mapping in a data-scarce context using Sentinel-2 images in machine-learning models,” *Sensors*, vol. 23, no. 23, p. 9328, Nov. 2023, doi: 10.3390/s23239328.

19. A. Bannari and Z. M. Al-Ali, “Assessing climate change impact on soil salinity dynamics between 1987–2017 in arid landscape using Landsat TM, ETM+ and OLI data,” *Remote Sens.*, vol. 12, no. 17, p. 2794, 2020.

20. V. Habibi, H. Ahmadi, M. Jafari, and A. Moeini, “Mapping soil salinity using a combined spectral and topographical indices with artificial neural network,” *PLoS One*, vol. 16, no. 5, 2021.

21. K.-A. Nguyen, Y.-A. Liou, P. Tran, H. P. Phung, and T.-H. Nguyen, “Soil salinity assessment by using near-infrared channel and Vegetation Soil Salinity Index derived from Landsat 8 OLI data: a case study in the Tra Vinh Province, Mekong Delta, Vietnam,” *Prog. Earth Planet. Sci.*, vol. 7, 2020, doi: 10.1186/s40645-019-0311-0.

22. J. Ding, S. Yang, Q. Shi, Y. Wei, and F. Wang, “Using apparent electrical conductivity as indicator for investigating potential spatial variation of soil salinity across seven oases along Tarim River in Southern Xinjiang, China,” *Remote Sens.,* vol. 12, p. 2601, 2020, doi: 10.3390/rs12162601.

23. N. E. Silvero et al., “Soil property maps with satellite images at multiple scales and its impact on management and classification,” *Geoderma*, vol. 397, p. 115089, 2021.

24. A. S. Abuzaid et al., “Multi-indicator and geospatial based approaches for assessing variation of land quality in arid agroecosystems,” *Sustainability*, vol. 14, p. 5840, 2022.

25. A. Okasha et al., “Effects of irrigation method and water flow rate on irrigation performance, soil salinity, yield, and water productivity of cauliflower,” *Agriculture*, vol. 12, pp. 1–18, 2022, doi: 10.3390/agriculture12081164.

Information about authors

Aisulu Ataniyazova – PhD student at Al-Farabi Kazakh National University, Faculty of Information Technologies (Almaty, Kazakhstan, aisulu.ataniyazova@gmail.com). Her research interests include machine learning and remote sensing. ORCID iD: 0000-0003-1122-6614.

Timur Merembayev – PhD, Research Assistant at the Institute of Information and Computational Technologies (Almaty, Kazakhstan, merembaevt@gmail.com). His current research covers various topics in a mathematical simulation of physical processes, machine learning, and geoscience problems. ORCID iD: 0000-0001-8185-235X.

Submission received: 18 April, 2025.
Revised: 28 May, 2025.
Accepted: 28 May, 2025.

IRSTI 05.01.07 https://doi.org/10.26577/jpcsit2025333

Fakhriddin Nuraliev[1*], Begdulla Sultanov[2], Nurkadem Kaldybayev[1]

[1]Tashkent University of Information Technology Named after Muhammad Al-Khwarizmi,
Tashkent, Uzbekistan
[2]Independent Researcher, Nukus, Uzbekistan
*e-mail: nuraliev2001@mail.ru

MODELING OF THE DEFORMED STATE OF MESH PLATES USING COMPLEX CONFIGURATION

Abstract: In this article, the deformed state processes of mesh plates with complex configurations are modeled mathematically. Specifically, a computational algorithm comprising of the R-function methods of V.L. Rvachev (RFM) and the Bubnov-Galerkin method is applied. The mathematical model describes the behavior of mesh plates under external loads by representing equilibrium equations in a Cartesian coordinate system. The solution structures are built using constructive RFM approaches, and discretization is carried out with the Bubnov-Galerkin technique. Computational experiments are conducted to determine the deformation characteristics of mesh plates with intricate geometries. The proposed approach significantly reduces the computational complexity and increases the accuracy of results when compared to conventional analytical methods. Furthermore, the algorithm enables numerical analysis of rhomboidal and hexagonal plate configurations under different boundary conditions. These results may be utilized in designing lightweight yet structurally efficient components in aerospace, civil, and mechanical engineering.

Keywords: mathematical modeling, stress-strain state, mesh plates with complex configurations, R-function method (RFM), Bubnov-Galerkin method, computational mechanics.

1. Introduction

Mesh shells and plates are common structural forms in different technological domains as well as the construction sector. In this sense, there has been a recent surge in interest in these structures both domestically in our country and elsewhere. The high industrial production of the primary structural components and a notable rise in the walls of their ready structures, the objectivity of the products, and the potential for broad unification of them not only for individual structures but also for buildings with different objects, loads, and operation patterns are the benefits of building structures. The theory of mesh systems was developed with contributions from well-known scientists including L.N. Lubo, B.A. Mironkov, G.I. Pshenichnov, V.K. Kabulov, T.Sh. Shirinkulov, T.Buriev, and K.S.Abdurashitov, among others. However, the configurations of these mesh plates have classical forms in many of the investigated engineering practice issues where the stress-strain condition of the plates is being studied. However, in engineering practice, complicated mesh plate designs present a problem for designers and design engineers to solve mathematically.

This makes the development of a computational algorithm and the software that goes along with it, as well as the computation of mesh plates with intricate configurations, highly pertinent.

2. Materials and methods

In the development of the theory and methods for calculating lattice shells and plates, significant contributions were made by such well-known scientists as G.I. Pshenichnikov, V.V. Kuznetsov, L.N. Lubo, B.A. Mironov, V.I. Volchenko, R.I.Khisamov, A.L. Filin, I.G. Tagiev, A.R.Rzhanitsyn, B.S. Volikov, V.V. Bolotin and others [1-11].

The work [6] presents a new, fairly accurate method for calculating the processes and stability of lattice cylindrical shells, which allows taking into account the main features of the structure under consideration. The results obtained earlier by the author, valid for shells with a square mesh, are generalized to the case of a shell with a rhombic lattice.

The theory of thin elastic lattice shells and plates, as well as its applications in technology, are presented in [7]. The main attention is paid to the issues of the theory of single-layer structures, mesh, which are rod systems.

Based on the hypotheses of the technical theory of thin-walled rods by V.Z. Vlasov, a system of differential equations of thin-walled composite rods with variable stiffness characteristics with elastically compliant shear connections was obtained by the displacement method [8]. It is shown that in this problem, the force method leads to a system of integro-differential equations if the shear forces in the seams are taken as the main unknowns. The solution of the problem in the case of static variability of stiffness characteristics is carried out by the perturbation method using Green's functions.

In [9], the effect of rigid gussets in nodes and the frequency and shape of oscillations of trusses with different numbers of panels are considered. Loads and masses are taken at the nodes. The effect of changing the stiffness of rods on the displacement of truss nodes is investigated.

In [10], theoretical studies of the strength of a two-support beam with a defect arbitrarily located along the length and height are presented. The beam is loaded with a concentrated force. Calculation formulas for determining the forces in the transverse ties are derived.

The author of the work [11] proposes to present an expression for the differential operator and derivatives under limiting and boundary conditions in finite differences for calculating a statically indeterminate beam of an asymmetric cross-section made of a material with different modulus. The corresponding system of algebraic equations is solved by the Seidel method, modified by the authors due to the presence of limiting conditions. An example of calculation is given.

In this research [12], vibration analysis of a thin elastic circular Aluminium plate has been investigated both experimentally and computationally. First six natural frequencies of thin Aluminium circular plate having diameter 220mm were calculated by performing modal analysis on ANSYS. In order to perform computational study, multizone quad/tri method was applied on the thin elastic circular plate to generate the quadrilateral mesh and simply supported boundary condition was applied. Mesh independence study was performed by varying the size of the mesh through sizing option and examining the natural frequency. To determine fundamental natural frequency experimentally, forced vibrations were induced into the plate with the help of system comprises of DC motor, steel wire and cam with the eccentricity of 8mm.

In this work [13], the conventional method of modeling and analyzing cold plates involves the use of empirical correlations to predict pressure drop for standard fin geometries used for various designs. A more reliable approach to pressure drop prediction is to perform a detailed CFD analysis of the cold plate design. Conventional CFD analysis of fluid flow through cold plates requires filling in details of the cold plates (passages and fin structures) as well as fluid volumes in the cold plates. For complex fin patterns produced by additive manufacturing, this can be very computationally expensive. A new analytical method for CFD analysis of the cold plate structure with only the captured fluid volume was developed, which significantly reduced the time required to mesh and solve complex fin structures in cold plate fluid passages.

Additive manufacturing (AM) technologies offer design freedom to create complex metal structures with significantly increased area-to-volume ratios with minimal time and cost. This allows creating a new generation of two-phase exchangers, thermosyphons and heat pipes. The pile design has been identified as the main parameter affecting the thermal performance of heat pipes and steam chambers. Traditionally, these rods were constructed by sintering metal powder or welding wire mesh welded to the inner wall of the heat pipe. This research paper [14] investigates the use of Laser Powder Bed Fusion (L-PBF) to construct AM microstructured rods that cannot be produced using conventional methods. Four pore configurations are produced using L-PBF: (i) body-centered cubic (BCC), (ii) face-centered cubic (FCC), (iii) regular cubic (SC), and (iv) Voronoi. The porosity of each configuration was varied by varying the strut thickness from 0.25 mm to 0.35 mm. The capillary performance of these rods was tested through a rate of rise experiment in which the sample is immersed vertically in a well of ethanol and the mass rate is recorded using an accurate laboratory scale.

In addition to the classical methods for analyzing the stress-strain behavior of lattice shells and plates, recent studies have explored the nonlinear deformation processes of anisotropic plates under complex fields. For instance, in [26], the authors developed a mathematical model based on the Hamilton-Ostrogradsky variational principle and Kirchhoff-Lyaw hypothesis to study the thermo-electro-magneto-elastic behavior of plates with intricate shapes. Their analysis integrated electromagnetic and thermal effects using Cauchy relations, Hooke's law, and Maxwell’s equations. Furthermore, a related investigation [27] addressed the independent solution of the magnetoelastic problem for thin compound-shaped isotropic plates. The study applied a variational approach to derive plate motion equations under electromagnetic fields, contributing to the theoretical base of compound geometries. These approaches reflect an increasing trend toward incorporating multi-physical interactions in plate modeling tasks.

From the review of studies it follows that among the considered mesh plates the classical configurations have practical significance. Even their algorithm has not been developed when the mesh plate configurations have complex shapes.

3. Results

Equilibrium equations for a mesh plate element in a Cartesian coordinate system relatively have the form [1-3]

$$\frac{\partial Q_1}{\partial x_1}+\frac{\partial Q_2}{\partial y}+Z=0 \tag{1}$$

where

$$Q_1=\frac{\partial H_2}{\partial y}-\frac{\partial M_1}{\partial x},\, Q_2=\frac{\partial H_1}{\partial x}-\frac{\partial M_2}{\partial y}$$

Here Z is the external load; the parameters are bending and torque; lateral forces.

Bending and torque moments, when the number of rods is equal to n (general case), are determined by the following formulas:

$$\left.\begin{array}{c} M_1=-(D_{11}\chi_1+D_{12}*\chi_2+2D_{16}*\tau) \\ M_2=-(D_{21}\chi_1+D_{12}*\chi_2+2D_{16}*\tau) \\ H_1=D_{61}*^{(i)}\chi_1+D_{62}*^{(i)}\chi_2+D_{66}*^{(i)},i=1{,}2 \end{array}\right\} \tag{2}$$

in which

$$\left.\begin{array}{c} D_{11}^*=D_{11}+K_{11},D_{12}^*=D_{12}-K_{12},D_{16}^*=D_{16}-\frac{1}{2}K_{16}, \\ D_{21}^*=D_{k1}-K_{11},D_{22}^*=D_{22}+K_{22},D_{26}^*=D_{26}+\frac{1}{2}K_{26}, \\ D_{16}^{*(i)}=D_{61}+K_{61}^{(i)},D_{26}^{*(i)}=D_{62}-K_{62}^{(i)}, \\ D_{66}^{*(i)}=D_{66}-K_{66}^{(i)},i=1{,}2, \end{array}\right\} \tag{3}$$

$$\left.\begin{array}{c} D_{11}=\sum_{i=1}^{n}I_iC_i^4,D_{12}=\sum_{i=1}^{n}I_iS_i^2C_i^2,D_{66}=2D_{12} \\ D_{12}=\sum_{i=1}^{n}I_iS_iC_i^3,D_{22}=\sum_{i=1}^{n}I_iS_i^4,D_{26}=\sum_{i=1}^{n}I_iS_i^3C_i,D_{ij}=D_{ij}, \\ K_{61}^{(1)}=K_{62}^{(1)}=\sum_{i=1}^{n}C_iS_iC_i^3,K_{66}^{(1)}=\sum_{i=1}^{n}C_ic_i^2\,cos\,2\,\phi_i, \\ K_{61}^{(2)}=K_{62}^{(2)}=\sum_{i=1}^{n}C_iS_i^2C_i,K_{66}^{(2)}=\sum_{i=1}^{n}C_iS_i^2\,cos\,2\,\phi_i, \\ K_i=E_iF_i/a_i,I=F_iJ_{1i}/a_i,C_i=G_iJ_{3i/a_i}. \end{array}\right\} \tag{4}$$

Let's characterize the parameters provided in (4). *I*I_i, C_i show the relationship between the rods' matching stiffness properties and the separation between their axes.; a_i – distance between the axes of adjacent rods; F_i – rod area; J_{1i}, J_{3i} – main, central moments of inertia; n – number of mesh bar families; φ_i – angle between axis α and the axis of the rod (measured from the axis α in the direction of the axis β); $S_i = \sin\varphi_i, C_i = \cos\varphi_i$,

$$x_1 = -\frac{\partial^2 W}{\partial x^2}, x_1 = -\frac{\partial^2 W}{\partial x^2}, \tau = -\frac{\partial^2 W}{\partial x \partial y}$$

Substituting (2) taking into account (3) into (1), we obtain the equilibrium equation of mesh plates with respect to deflection – W:

$$\begin{gathered}
(D_{11} + K_{11})\frac{\partial^2 W}{\partial x^2} + 2(D_{61} + D_{16})\frac{\partial^4 W}{\partial x^3\,\partial y} + (3D_{66} + K_{16})\frac{\partial^4 W}{\partial x^2\,\partial y^2} \\
+2(D_{62} + K_{16})\frac{\partial^4 W}{\partial x\,\partial y^3} + (D_{22} + K_{11})\frac{\partial^4 W}{\partial y^4} \\
+[2\frac{\partial}{\partial x}(D_{11} + K_{11})\frac{\partial}{\partial y}(2D_{61} - K_{16})]\frac{\partial^3 W}{\partial x^3} \\
+[3\frac{\partial}{\partial x}(2D_{61} - K_{16}) + \frac{\partial}{\partial y}(3D_{66} - K_1^0)]\frac{\partial^3 W}{\partial x^3\,\partial y} \\
+\frac{\partial}{\partial y}(3D_{66} + K_1^0) + 3\frac{\partial}{\partial x}(2D_{62} + K_{16})]\frac{\partial^3 W}{\partial x\,\partial y^2} \\
+[\frac{\partial}{\partial y}(2D_{62} + K_{16}) + 2\frac{\partial}{\partial y}(D_{22} + K_{11})]\frac{\partial^3 W}{\partial y^3} \\
+[\frac{\partial^2}{\partial x^2}(D_{11} + K_{11})\frac{\partial}{\partial x\,\partial y}(2D_{61} - K_{16}) + \frac{\partial^2}{\partial y^2}(2D_{61} + K_{16}) \\
+\frac{\partial}{\partial x\,\partial y}(2D_{66} + K_1^0) + \frac{\partial}{\partial y^2}(2D_{62} + K_{16})]\frac{\partial^2 W}{\partial \mathrm{x}} \\
+[\frac{\partial^2}{\partial x^2}(D_{12} - K_{11}) + \frac{\partial^2}{\partial x\,\partial y}(2D_{62} + K_{16}) + \frac{\partial^2}{\partial y^2}(D_{22} + K_{11})]\frac{\partial^2 W}{\partial y^2} \\
-Z = 0
\end{gathered} \tag{5}$$

where

$$K_1^{(0)} = \sum_{i=1}^{h}(1 - 6j_i^2 c_i^2)c_i\,, K_2^{(0)} = \sum_{i=1}^{h} C_i\, cos^2\, 2\, \phi_i \tag{6}$$

Substituting (4) into (3), then the resulting relations into (2), we have:

$$\left.\begin{aligned}
M_1 &= \sum_{i=1}^{4} c_i \nabla_i\,(I_i c_i \nabla_i + C_i s_i \nabla_i)w, \\
M_2 &= \sum_{i=1}^{4} s_i \nabla_i\,(I_i s_i \nabla_i + C_i c_i \nabla_i)w, \\
H_1 &= -\sum_{i=1}^{4} c_i \nabla_i\,(I_i c_i \nabla_i - C_i c_i \nabla_i)w, \\
H_2 &= -\sum_{i=1}^{4} s_i \nabla_i\,(I_i c_i \nabla_i - C_i c_i \nabla_i)w.
\end{aligned}\right\} \tag{7}$$

$$\left.\begin{aligned} Q_1 &= -\sum_{i=1}^{4} \nabla_i^2 \,(I_i c_i \nabla_i + C_i s_i \Delta_i)w, \\ Q_1 &= -\sum_{i=1}^{4} \nabla_i^2 \,(I_i s_i \nabla_i + C_i c_i \Delta_i)w, \end{aligned}\right\} \quad (8)$$

By virtue of (7), equation (5) takes the form

$$\sum_{i=1}^{4} \nabla_i^2 \,(I_i \nabla_i^2 + C_i S_i^2)w - z = 0 \quad (9)$$

Equation (5) is solved under the condition

$$M_n \delta \frac{\partial W}{\partial n}\bigg| = 0,\, R_n \delta W|_G = 0 \quad (10)$$

where n – outer normal direction; (□-variation operation sign; $M_n -$ normal bending moment,

$$\begin{gathered} M_n = M_1 \cos^2\alpha + \\ +(H_1 + H_2)\cos\alpha\sin\alpha + M_2 \sin 2\alpha \end{gathered} \quad (11)$$

$R_n -$ normal ground reaction, having the form

$$\begin{gathered} R_n = Q_n + \frac{\partial M_n \tau}{\partial S}, \\ Q_n = Q_1 \cos\alpha + Q_2 \sin\alpha\,, \\ M_n = H_1(\cos^2\alpha - \sin^2\alpha) + \\ +(M_2 - M_1)\cos\alpha\sin\alpha = \\ = H_1 \cos 2\alpha + \frac{1}{2}(M_2 - M_1)\sin 2\alpha, \end{gathered} \quad (12)$$

here (□-angle between outer normal and axis x; W – plate deflection; W – plate deflection; S – the length of the arc for the boundary of the Γ region of the plate

For the purpose of computing mesh plates with complex shapes, this section presents a computational technique that builds a joint combination of R-function approaches by V.L. Rvachev and Bubnov Galerkin. The procedure consists of two steps:

Constructing coordinate sequences, also known as solution structures;

- building equations for solving problems (discretization by spatial variables) using the Bubnov-Galerkin method;
- approximating double integral values;
– solving equations for solving problems;
- determining necessary parameter values;
– registering calculation outcomes.

Depending on how the mesh plate's edges are secured, differentiable equations (1.3.5) for the mesh plate's equilibrium can be solved given suitable boundary conditions based on the number of families of rods. Condition (10), in a generalized form, determines these boundary conditions.

Generally speaking, we use the form to represent the coordinate sequences that match criterion (5).

$$W = \sum_{i=1}^{n} C_i \phi_i \quad (13)$$

where C_i – unknown coefficients to be determined; $\varphi_i -$ basis system of coordinate functions, meeting the boundary criteria, which were created using V.L. Rvachev's R-function approach.

If we replace (13) with (5), we get the following system of algebraic equations by applying the Bubnov-Galerkin techniques;

$$AC=f \quad (14)$$

Where

$$\begin{gathered} A = \{a_{ij}\}, dim(A) = n \times n, \\ C = \{c_i\}, f = \{f_i\}, dim(C) = \\ = n \times 1, dim(f) = n \times 1, \end{gathered}$$

Here in general elements a_{ij} and f_i for equations (5) have a look

$$a_{ij} = \iint_{\Omega} \tilde{A} W_i W_j d\Omega \quad (15)$$

$$f_j = \iint_{\Omega} Z W_j d\Omega \quad (16)$$

Here

$$\begin{aligned}\tilde{A}W_i := \quad & (D_{11}+K_{11})\frac{\partial^4 W}{\partial x^4}+2(D_{61}+D_{16})\frac{\partial^4 W}{\partial x^3\,\partial y}+(3D_{66}+K_{16})\frac{\partial^4 W}{\partial x^2\,\partial y^2}\\ & +2(D_{62}+K_{16})\frac{\partial^4 W}{\partial x\,\partial y^3}+(D_{22}+K_{11})\frac{\partial^4 W}{\partial y^4}\\ & +[2\frac{\partial}{\partial x}(D_{11}+K_{11})-2\frac{\partial}{\partial y}(2D_{61}-K_{16})]\frac{\partial^3 W}{\partial x^3}\\ & +[3\frac{\partial}{\partial x}(2D_{61}-K_{16})+\frac{\partial}{\partial y}(3D_{66}-K_{16}^{0})]\frac{\partial^3 W}{\partial x^2\,\partial y}\\ & +[3\frac{\partial}{\partial y}(2D_{62}+K_{16})+\frac{\partial}{\partial x}(2D_{62}+K_{16})]\frac{\partial^3 W}{\partial x\,\partial y^2}\\ & +[\frac{\partial}{\partial y}(D_{22}+K_{11})]\frac{\partial^3 W}{\partial y^3}\\ & +[\frac{\partial^2}{\partial x^2}(D_{11}+K_{11})+\frac{\partial^2}{\partial x\,\partial y}(2D_{61}-K_{16})+\frac{\partial^2}{\partial y^2}(2D_{61}+K_{16})\\ & +\frac{\partial^2}{\partial x\,\partial y}(D_{66}+K_{16}^{0})+\frac{\partial^2}{\partial y^2}(2D_{62}+K_{16})]\frac{\partial^2 W}{\partial x^2}\\ & +[\frac{\partial^2}{\partial x^2}(D_{12}-K_{11})+\frac{\partial^2}{\partial x\,\partial y}(2D_{62}+K_{16})+\frac{\partial^2}{\partial y^2}(D_{22}+K_{11})]\frac{\partial^2 W}{\partial y^2}\end{aligned} \qquad (17)$$

Equations (14) and (15) respectively assume the following form if we examine specific instances of equations (5).

$$\begin{aligned}a_{ij}=\iint_{\Omega}(D_1\frac{\partial^4 W_i}{\partial x^4}+D_3\frac{\partial^4 W_i}{\partial x^2\partial y^2}\\ +D_2\frac{\partial^4 W_i}{\partial y^4})_i W_j d\Omega\end{aligned} \qquad (18)$$

$$f_j=\iint_{\Omega} aW_j d\Omega \qquad (19)$$

Here a, depending on the consideration of equations (15) or special cases, respectively, takes the following values:

$$\text{a:=}\ \frac{Z}{I}\ ;\text{a:=}\ \frac{a_3}{EJ_1}Z\ ;\ \text{a:=}\ \frac{a_4}{EJ_1}Z\ ;$$

In this case, the expressions for a_{ij} in equations (15) or others similar in form to expressions (18), difference is, expressions here D_1, D_2, D_3 will differ.

Equation (16) is solved by the Gaussian elimination method. Unknown coefficients C$_i$ are determined. Then substituting the values C$_i$ into (13), we will find solution W. Afterwards, the transverse forces of the plate and the values of its bending and torque moments are calculated using the structural formulas. Further, using the known values of bending (M_1, M_2) and torque (M_{12}) moments and shear forces Q_1, Q_2; force values are calculated (N_i^*, S_i^*) and moments M_i^* rod.

We will see that the precise Gauss formula is used to determine the value of double integrals.

Let's now examine how structural formulas are actually constructed. Anisotropic plate structural formulas are utilized for this purpose; the lone exception being when the coefficient D_{ij} is changed into D_{ij}^* in case of mesh plates.

when the edges of the plate are rigidly clamped, the structure of the solution has the form [4-5]:

$$W=\omega^2\Phi \qquad (20)$$

The simply supported boundary condition is now examined. In this instance, the solution's structure takes the form

$$W(x) = \omega\Phi_1 - \frac{\omega^2}{2(A_1-\omega)}\left(A_1(2D_1\Phi_1 + \Phi_1 D_2\omega) + 2A_2T_1\Phi_1 - \frac{1}{\rho}A_3\Phi_1\right) + \omega^3\Phi_2 \tag{21}$$

We will look through the shifted boundary condition. For example, when Γ_1 – area Γ_1 – rigidly clamped plate, and the rest $\Gamma_1 = \Gamma - \Gamma_1$ free, hen the solution structure according to [4-5] has the form

$$\begin{gathered} W \;=\; \omega_1^2\,\Phi_1 + \frac{\omega_1^2\,\omega_2^2}{2\left(\omega_1^2+\omega_2^2\right)}\times \\ \times\,\{\omega_2^2\,\Phi_2 - \{\frac{\omega_2}{3}(S_1^{*(2)}[1 - \frac{\omega_2^2}{2S_1^*}\,B_2^{*(2)}] + B_3^{*(2)} - \frac{2A_6^*}{S_1^*}\,KB_2^{*(2)})\} - \frac{1}{S_2}\,B_2^{*(2)}\}\,(W_1^2\,\Phi_1) \end{gathered} \tag{22}$$

here Φ_1, Φ_2 – undefined components of structural formulas, which are usually presented in the form [4-5]:

$$\Phi_s = \textstyle\sum_{i=1}^{m_s} C_{is}\psi_i^s,\ \text{s=1,2;} \tag{23}$$

where $\{\psi_i^s\}$ – selected sequence of polynomials, such as power polynomial, Chebyshev polynomial, trigonometric polynomial, etc.; ω_1, ω_2 – correspondingly normalized boundary equations Γ_1 and Γ_2 areas Γ_1 and Γ_2.

4. Discussion

Calculate mesh plates with a rhomboidal form is the focus of this section. Here the plate arrangements with their complex shapes are considered [7–10].

Assuming that the edges of the hexagonal plate (Fig. 1) are simply supported and firmly clamped, we can compute the surface area of the plate.

The problem comes down to integrating equation (15), i.e.

$$D_1\frac{\partial^4 W}{\partial x^4} + D_3\frac{\partial^4 W}{\partial x^2\partial y^2} + D_2\frac{\partial^4 W}{\partial y^4} = \frac{z}{I} \tag{24}$$

in case

$$W|\Gamma = 0, M_4|\Gamma = 0 \tag{25}$$

The geometry equation of the region in this case has the form:

$$\begin{gathered} \omega = \omega_1 \wedge \omega_2 \wedge_0 \omega_3 \\ \omega_1 = \frac{\left(\frac{3}{4} - y^2\right)}{\sqrt{3}}, \\ \omega_2 = f_1 \wedge_0 f_3, \omega_3 = \\ = f_4 \wedge_0 f_5, f_1 = \frac{\left(\sqrt{3} - \sqrt{3}x - y\right)}{2}, \\ f_3 = \left(\sqrt{3} + \sqrt{3}x - y\right)\sqrt{2}, \\ f_4 = \frac{\frac{\sqrt{3}}{2}\left(\sqrt{3} + \sqrt{3}x + y\right)}{2}, \\ f_5 = \left(\sqrt{3} - \sqrt{3}x + y\right)/2 \end{gathered} \tag{26}$$

Take note that the formulas for bending moments for traditional orthotropic elastic plates differ slightly from the expressions for moments in form used here. Consequently, we describe the moments of mesh plates in standard form and create structural formulas for expressing them.

For this purpose, consider the relation:

$$\begin{gathered} M_1 = 2c^2\left[(Ic^2 + Cs^2)\frac{\partial^2 w}{\partial x^2} + (I - C)s^2\frac{\partial^2 w}{\partial y^2}\right], \\ M_2 = 2s^2\left[(I - C)c^2\frac{\partial^2 w}{\partial x^2} + (Is^2 + Cc^2)\frac{\partial^2 w}{\partial y^2}\right], \\ H_1 = -2c^2(2Is^2 + C\cos 2\phi)\frac{\partial^2 w}{\partial x\partial y}, \\ H_2 = 2s^2(C\cos 2\phi - 2Ic^2)\frac{\partial^2 w}{\partial x\partial y} \end{gathered} \tag{27}$$

We will rewrite the relation (27) in form

$$M_1 = D_{1c}\left[\frac{\partial^2 w}{\partial x^2} + v_{2c}\frac{\partial^2 w}{\partial y^2}\right],$$
$$M_2 = D_{2c}\left[\frac{\partial^2 w}{\partial y^2} + v_{1c}\frac{\partial^2 w}{\partial x^2}\right], \qquad (28)$$
$$H_1 = \alpha_1\frac{\partial^2 w}{\partial x\partial y},$$
$$H_2 = \alpha_2\frac{\partial^2 w}{\partial x\partial y}.$$

where

$$D_{1c} = 2c^2(Ic^2 + Cs^2),$$

$$v_{2c} = \frac{(I - C)s^2}{Ic^2 + Cs^2},$$

$$D_{2c} = 2s^2(Is^2 + Cc^2),$$

$$v_{1c} = \frac{(I - C)c^2}{Is^2 + Cc^2},$$

$$C = GJ_3/a, I = EJ_1/a,$$

$$C = cos\,\phi\,, S = sin\,\phi,$$

$$\alpha_1 = -2c^2(2Is^2 + C\,cos\,2\,\phi)\frac{\partial^2 w}{\partial x\partial y},$$

$$\alpha_2 = 2s^2(C\,cos\,2\,\phi - 2Ic^2)\frac{\partial^2 w}{\partial x\partial y}$$

Rhomboidal plate structure $\varphi = \frac{\pi}{4}$. Thus s=c и $D_{1c} = D_{2c}$; $v_{1c} = v_{2c}$, therefore, we consider and $J_3 = 0$ we can rewrite these formulas in the form

$$M_1 = D_c\left(\frac{\partial^2 w}{\partial x^2} + v_c\frac{\partial^2 w}{\partial y^2}\right),$$

$$M_2 = D_c\left(\frac{\partial^2 w}{\partial y^2} + v_c\frac{\partial^2 w}{\partial x^2}\right), \qquad (29)$$

$$H_1 = H_2 = -4Ic^2s^2$$

where $D_C = D_C = D_{2C}, v_c = v_{1c} = v_{2c}$.

According to (29) we will have this

$$M_n = D_c\left(\frac{\partial^2 w}{\partial n^2} + v_c\frac{\partial^2 w}{\partial \tau^2}\right) \qquad (30)$$

so

$M_n = M_1\cos^2\alpha + M_2\sin^2\alpha + 2M_{12}\sin\alpha\cos\alpha$.

Here n – outer normal, □ – tangent.

Considering the remarks above, relations of the form determine the boundary condition of a simply supported plate with a rhomboidal structure.

$$W|r = 0, \left(\frac{\partial^2 w}{\partial n^2} + v_c\frac{\partial^2 w}{\partial \tau^2}\right)\Big|_{\&_r} = 0 \qquad (31)$$

Consequently, the structure of this boundary condition's solution in this instance is shown as

$$W = \omega\Phi -$$
$$-\frac{\bar{\omega}^2}{2}[\Phi(D_2\bar{\omega} + v_cT_2\bar{\omega}) + 2D_1\Phi] \qquad (32)$$

We note that if the stiffness of the rods is not taken into account $J_3 = 0$, then C=0, ν c =1 and (3.2.7) takes the form

$$W = \omega\Phi -$$
$$= \frac{\bar{\omega}^2}{2}[\Phi(D_2\bar{\omega} + T_2\bar{\omega}) + 2D_1\Phi] \qquad (33)$$

Let us now examine the rhombic structure's hexagonal plate, as depicted in Figure 1. Let a firmly clamped rhomboidal mesh plate form the image's border in Figure 1. Then, the structure of the answer to this issue takes the following form:

$$W = \omega^2\Phi \qquad (34)$$

Here ω is found out with the formula (26) and Φ is given as (23).

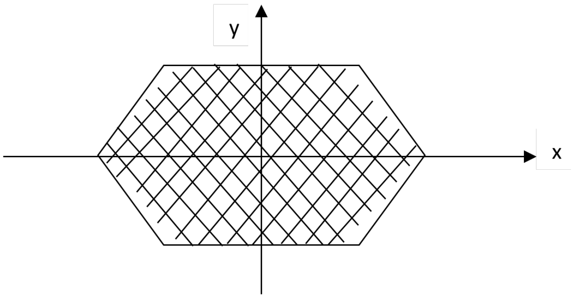


Figure 1 – Complex poligon

x	W n=3 c=20
-1	0
-0,9	0,00010152
-0,8	0,00071714
-0,7	0,00205652
-0,6	0,00400341
-0,5	0,00626503
-0,4	0,00851907
-0,3	0,01049807
-0,2	0,01201645
-0,1	0,01296362
0	0,01328472
0,1	0,01296362
0,2	0,01201645
0,3	0,01049807
0,4	0,00851907
0,5	0,00626503
0,6	0,00400341
0,7	0,00205652
0,8	0,00071714
0,9	0,00010152
1	0

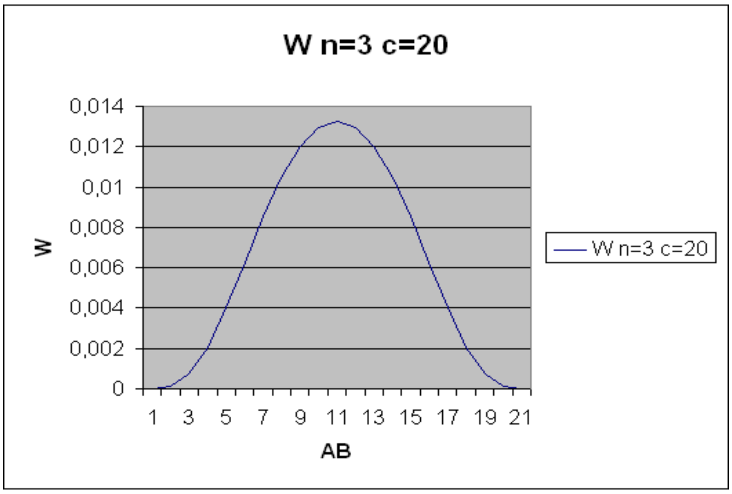


Figure 2 – The figure shows numerical and graphical results for one of the sections1.

Below (Fig. 2) the computational experiment's findings are displayed.

5. Conclusions

Summarize the key findings and their significance. Clearly state the main conclusions drawn from the study and their implications. Avoid introducing new data or extensive discussions not previously covered.

Author Contributions

Conceptualization, F.N.; Methodology, F.N. and B.S.; Software, N.K.; Formal Analysis, F.N., B.S. and N.K.; Investigation, B.S.; Resources, F.N., B.S. and N.K.; Data Curation, B.S. and N.K.; Writing – Original Draft Preparation, F.N. and B.S.; Writing – Review & Editing, F.N.; Visualization, B.S. and N.K.; Supervision, F.N.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. G. I. Pshenichnov, *Theory of Thin Elastic Mesh Shells and Plates*. Moscow, Russia: Nauka, 1982, 352 pp. (in Russian).
2. G. I. Pshenichnov, *Calculation of Mesh Shells. Research on the Theory of Structures*. Moscow, Russia, 1976. (in Russian; publisher not specified).
3. G. I. Pshenichnov, *Theory of Thin Elastic Mesh Shells*, Doctor of Science dissertation, Moscow, USSR, 1968. (in Russian).
4. V. L. Rvachev, *Theory of R-Functions and Their Applications*. Kyiv, Ukraine: Naukova Dumka, 1982, 552 pp. (in Russian).
5. V. L. Rvachev and L. Kurpa, *The R-Function Method in Problems of Bending Plates with Complex Shapes*. Kyiv, Ukraine: Naukova Dumka, 1988. (in Russian).
6. L. N. Lubo and B. A. Mironkov, *Plates of Regular Spatial Structure*. Moscow, Russia: Stroyizdat, 1976, 105 pp. (in Russian).
7. S. G. Lechnicky, *Anisotropic Plates*. Leningrad, USSR: Gostekhizdat, 1947, 355 pp. (in Russian).
8. G. I. Pshenichnov, “On the calculation of mesh cylindrical shells with an arbitrary lattice,” *Bulgarian Academy of Sciences*., 1962. (in Russian).
9. G. I. Pshenichnov, *Analysis of Cylindrical Grid Shells*. Moscow, USSR: Izdat. AN SSSR (Publishing House of the Academy of Sciences of the USSR), 1961. (in Russian).
10. G. I. Pshenichnov, *Free and Forced Vibrations of a Thin Elastic Cylindrical Shell of Open Profile*. Moscow, USSR: AN SSSR, 1967. (in Russian).
11. G. I. Pshenichnov, *Free and Forced Axisymmetric Vibrations of Thin Elastic Shells of Revolution*. Baku, Azerbaijan, 1996. (in Russian; publisher not specified).
12. M. Yasir, Z. I. Muhammad, and R. Ali, “Experimental and computational study of simply supported thin elastic circular plate under forced vibrations,” in *Proc. 2021 Int. Bhurban Conf. Appl. Sci. Technol. (IBCAST)*, Islamabad, Pakistan, Jan. 12–16, 2021, doi: 10.1109/IBCAST51254.2021.9393228.

13. U. Girish, “CFD analysis of extracted fluid volume for predicting pressure drop in additively manufactured cold plates with complex fin patterns,” in *Proc. 2023 22nd IEEE Intersoc. Conf. Thermal Thermomech. Phenomena Electron. Syst. (ITherm)*, Orlando, FL, USA, May 30–Jun. 2, 2023, doi: 10.1109/ITherm55368.2023.10177632.

14. E. Ahmed, D. Jason, and K. Roger, “Development of additive manufactured metal wick structures for two-phase heat transfer applications,” in *Proc. 2022 21st IEEE Intersoc. Conf. Thermal Thermomech. Phenomena Electron. Syst. (ITherm)*, San Diego, CA, USA, May 31–Jun. 3, 2022, doi: 10.1109/iTherm54085.2022.9899556.

15. M. Golub and O. Doroshenko, “Analysis of eigenfrequencies of a circular interface delamination in elastic media based on the boundary integral equation method,” *Mathematics*, vol. 10, no. 1, Art. no. 38, 2022, doi: 10.3390/math10010038.

16. M. Kolev, M. Koleva, and L. Vulkov, “An unconditional positivity-preserving difference scheme for models of cancer migration and invasion,” *Mathematics*, vol. 10, no. 1, Art. no. 131, 2022, doi: 10.3390/math10010131.

17. P. Di Barba, L. Fattorusso, and M. Versaci, “Electrostatic-elastic MEMS with fringing field: A problem of global existence,” *Mathematics*, vol. 10, no. 1, Art. no. 54, 2021, doi: 10.3390/math10010054.

18. F. Nuraliev, S. Safarov, M. Artikbayev, and O. Sh. Abdirozikov, “Calculation results of the task of geometric nonlinear deformation of electro-magneto-elastic thin plates in a complex configuration,” in *Proc. 2022 Int. Conf. Inf. Sci. Commun. Technol. (ICISCT)*, Tashkent, Uzbekistan, Sep. 28–30, 2022, doi: 10.1109/ICISCT55600.2022.10146920.

19. F. Nuraliev, S. Safarov, and M. Artikbayev, “Solving the problem of geometrical nonlinear deformation of electro-magnetic thin plate with complex configuration and analysis of results,” in *Proc. 2021 Int. Conf. Inf. Sci. Commun. Technol. (ICISCT)*, Tashkent, Uzbekistan, Nov. 3–5, 2021, doi: 10.1109/ICISCT52966.2021.9670282.

20. F. M. Nuraliev, B. Sh. Aytmuratov, Sh. Sh. Safarov, and M. A. Artikbaev, “Mathematical modeling of geometrical non-linear processes of electromagnetic spring thin plate layer configuration,” *Problems of Computational and Applied Mathematics*, vol. 1, no. 38, pp. 90–109, 2022.

21. F. M. Nuraliev, B. Sh. Aytmuratov, and M. A. Artikbaev, “Mathematical model and computational algorithm of vibration processes of thin magnetoelastic plates with complex form,” in *Proc. 2021 Int. Conf. Inf. Sci. Commun. Technol. (ICISCT)*, Tashkent, Uzbekistan, Nov. 3–5, 2021, doi: 10.1109/ICISCT52966.2021.9670313.

22. F. Nuraliev, S. Safarov, and M. Artikbayev, “A computational algorithm for calculating the effect of the electromagnetic fields to thin complex configured plates,” in *Proc. 2020 Int. Conf. Inf. Sci. Commun. Technol. (ICISCT)*, Tashkent, Uzbekistan, Nov. 4–6, 2020, doi: 10.1109/ICISCT50599.2020.9351447.

23. F. Nuraliev, M. Artikbayev, and M. Uralbekova, “Analysis of the results of solving the magnetoelastic problem of thin anisotropic plates with complex form,” in *Proc. Int. Conf. Modern Trends in Developing the Mathematics in the Era of Digital Transformation*, Almaty, Kazakhstan, Mar. 14, 2024, pp. 114–119.

24. F. Nuraliev, B. Sultanov, and K. Qasimova, “Modeling of processes of deformation state of mesh plates with complex configuration,” in *Proc. Int. Conf. Modern Trends in Developing the Mathematics in the Era of Digital Transformation*, Almaty, Kazakhstan, Mar. 14, 2024, pp. 104–113.

25. B. Sultanov, “Justification of the reliability of numerical results of the approximate solution to the calculation of mesh plates,” in *Proc. Republican Sci. Tech. Conf. “Current State and Ways of Development of Information Technologies”*, Tashkent, Uzbekistan, Sep. 23–25, 2008, pp. 226–229.

26. F. Nuraliev, N. Tojiev, and B. Tahirov, “Mathematical model and computational analysis of the nonlinear stress–strain processes in anisotropic thermo–electro–magnetoelastic plates with complex shapes,” *Development of Science*, vol. 1, no. 3, pp. 18–23, 2024.

27. F. Nuraliev, M. Mirzaakhmedov, N. Tojiyev, and O. Abdullayev, “Independent solution of the magnetoelastics problem of thin compound-shaped isotropic plates,” *Modern Problems of Applied Mathematics and Information Technology – Al-Khwarizmi*, no. 1, 2024.

Information about authors

Prof. Dr. Faxriddin Murodullaevich Nuraliev is a Doctor of Technical Sciences in the specialty 05.01.07 – Mathematical Modeling, Numerical Methods and Software Systems (PhD, 2000; DSc, 2016). He currently serves as the Head of the Department of Television and Media Technologies at the Tashkent University of Information Technologies (TUIT), Uzbekistan. Prof. Nuraliev is a distinguished specialist in mathematical and geometric modeling and has made substantial contributions to the field. Email: nuraliev2001@mail.ru , ORCID iD: 0000-0002-0574-9278

Begdulla Sultanov is an independent researcher specializing in mathematical modeling. He collaborates on research projects in applied mathematics and computational modeling.

Nurkadem Kaldybaev, a citizen of Kazakhstan, is a doctoral student under the academic supervision of Prof. Dr. Nuraliev. He conducts research in the field of 05.01.07 – Mathematical Modeling, Numerical Methods and Software Systems, maintaining close academic cooperation across national borders.

Submission received: 31 May, 2025.
Revised: 10 July, 2025.
Accepted: 10 July, 2025.

IRSTI 49.03.03 https://doi.org/10.26577/jpcsit2025334

Amirkhan Temirbayev

Cluster of Engineering and High Technologies, Al-Farabi Kazakh National University, Almaty, Kazakhstan
e-mail: amirkhan@kaznu.kz

PHASE-BASED INTERFEROMETRIC METHOD FOR PRECISE DISPLACEMENT ESTIMATION: THEORY AND COMPUTATIONAL POTENTIAL

Abstract. This paper presents the theoretical foundation and computational modeling of a novel phase-based interferometric method for precise displacement estimation in environmental monitoring applications. The method leverages the phase difference between two coherent radio signals transmitted over a wireless forward link, enabling sub-millimeter resolution without relying on reflected signals or embedded sensors. Unlike radar interferometry and distributed fiber optic systems, the proposed technique operates entirely in a forward-link architecture, making it more scalable, energy-efficient, and suitable for low-infrastructure deployments. Special attention is given to the computational procedures required for real-time signal interpretation, including instantaneous phase extraction using the Hilbert transform, phase unwrapping algorithms, and noise mitigation via digital filters. Simulation results confirm that the method is theoretically robust and computationally tractable, offering a practical path toward implementation using lightweight embedded platforms such as software-defined radios (SDRs) with GPS-disciplined oscillators. The results also demonstrate how design parameters such as carrier frequency and dual-tone spacing – affect the sensitivity and resolution of displacement estimates. This study lies at the intersection of applied computer science, signal processing, and geospatial engineering. It provides both a mathematical and algorithmic foundation for future systems aimed at distributed, real-time sensing in civil infrastructure and geohazard management.

Keywords: Wireless monitoring, signal processing, intelligent systems, phase-based interferometry, displacement monitoring, Hilbert transform.

1. Introduction

Accurate detection of small-scale ground or structural displacements is critical in applications such as landslide forecasting, infrastructure safety, and environmental monitoring. Traditional approaches rely on mechanical inclinometers, differential GPS, or optical surveying, which often suffer from limited spatial resolution, high costs, or low temporal granularity. These limitations motivate the development of lightweight, scalable, and computationally efficient alternatives.

Recent advancements in Distributed Fiber Optic Sensing (DFOS) have significantly expanded the applicability of fiber-based technologies in structural and geotechnical monitoring. The study [1] explored the potential application of fiber optic sensors in monitoring key-point displacements by leveraging their sensitivity to optical parameters and spectral changes. The results demonstrated that fiber optic sensors could accurately measure circumferential strain within the elastic range and reliably reflect key-point displacement trends through the linear relationship model. The survey [2] is a comprehensive collection of recently published research articles on Structural Health Monitoring (SHM) campaigns performed by means of Distributed Optical Fiber Sensors (DOFS). Authors show the large scale of using DOFS which are cutting-edge strain, temperature and vibration monitoring tools with a large potential pool, namely their minimal intrusiveness, accuracy, ease of deployment and more. Its most state-of-the-art feature, though, is the ability to perform measurements with very small spatial resolutions (as small as 0.63 mm). The review [3] highlights the latest progress in distributed optical fiber sensors with an emphasis on energy applications such as energy infrastructure monitoring, power generation system monitoring, oil and gas pipeline monitoring, and geothermal process monitoring. This review aims to clarify challenges and limitations of distributed optical fiber sensors with the goal of providing a pathway to push the limits in distributed optical fiber sensing for practical applications. Studies [4], [5] explore the possibilities of distributed fiber-optic sensors (DFOS) in the field of geotechnics for detecting soil deformation. The authors compare the results of the new inclinometer with a traditional inclinometer. The works show that the measurement results are consistent with the results of a traditional inclinometer, which indicates its reliability and practicality. The review [6] explores various sensor technologies in structural health monitoring (SHM), such as piezoelectric, fibre optic, force, MEMS devices, GPS, LVDT, electromechanical impedance techniques, Doppler effect, and piezoceramic sensors, focusing on advancements from 2019 to 2024. This study also shows increasing of articles from 15 in 2019 to 359 in 2023. Velocity sensors also play a crucial role in SHM by capturing the movement of structures and providing valuable data for predicting potential damage [7]-[9]. These sensors utilise two main sensing methods: electromechanical impedance techniques and the Doppler effect, each offering unique capabilities in measuring structural velocity. The paper [10] summarizes the use of GPS technology for structural health monitoring.

While InSAR and ground-based interferometric radar (GB-InSAR) have demonstrated sub-centimeter resolution through analysis of reflected signals [11]-[12], these systems rely on coherent backscattering from natural or artificial surfaces, which can be influenced by vegetation, weather, and line-of-sight constraints. In contrast, the method proposed in this study is fundamentally different: it does not rely on reflected signals. Instead, it utilizes the direct transmission of two closely spaced coherent signals from a fixed transmitter to a fixed receiver. The receiver performs phase-difference analysis to estimate relative displacement between the two endpoints. This enables continuous, low-power, and infrastructure-independent monitoring, even in environments where radar reflection is unreliable or unfeasible.

This paper focuses on laying the theoretical and computational groundwork for such a system. Rather than describing a specific hardware implementation, we develop the analytical model, examine its sensitivity, and explore algorithms necessary to convert raw phase data into actionable displacement information. In doing so, we establish a framework that merges physical modeling with digital signal processing – an essential step toward future embedded and intelligent monitoring systems.

2. Theory of phase-based sensing

Let a narrowband radio signal of wavelength λ be transmitted from a source to a receiver. If the distance $L(t)$) between them changes due to motion of the transmitter (e.g., ground sliding), the phase of the received signal will change accordingly:

$$\phi(t) = \frac{2\pi}{\lambda} \cdot L(t) \qquad (1)$$

Taking the difference over time yields the displacement-sensitive expression:

$$\Delta L = \frac{\lambda}{2\pi} \cdot \Delta\phi \qquad (2)$$

This relationship allows us to infer displacement from phase measurements alone, assuming the system maintains phase coherence. By transmitting two closely spaced frequencies f_1 and f_2, we obtain two phase trajectories $\phi_1(t)$, $\phi_2(t)$, and analyze their difference:

$$\Delta\phi(t) = \phi_2(t) - \phi_1(t) = \frac{2\pi\cdot\Delta f}{c} \cdot L(t) \qquad (3)$$

where $\Delta f = f_2 - f_1$ is small and known, and c is the speed of light. This technique suppresses common-mode noise and enhances resolution through interferometric gain.

The effectiveness of phase-based displacement sensing strongly depends on three key parameters:

1. Carrier wavelength λ
2. Phase resolution $\delta\phi$ of the measurement system
3. Frequency stability of the system’s oscillators

The minimum detectable displacement δL is determined from the phase-displacement relationship:

$$\delta L = \frac{\lambda}{2\pi} \cdot \delta\phi \qquad (4)$$

While the physical basis of the method is straightforward, practical implementation requires robust signal processing techniques to extract displacement information from real-world phase measurements. Three key computational procedures are critical to the system’s reliability:

1. Instantaneous Phase Extraction:

To retrieve the phase of the received signal in real time, the analytic representation of the signal is obtained using the Hilbert transform. Given a received time-domain signal $x(t)$, the complex analytic signal $z(t)$ is constructed as:

$$z(t) = x(t) + j \cdot \mathcal{H}\{x(t)\} \qquad (5)$$

where $\mathcal{H}\{\cdot\}$ denotes the Hilbert transform. The instantaneous phase $\phi(t)$ is then computed as:

$$\phi(t) = arg(z(t)) \qquad (6)$$

This approach allows for real-time demodulation of the carrier signal and precise tracking of phase evolution.

2. Phase Unwrapping:

Since phase is inherently modulo-2π, it is necessary to perform phase unwrapping to obtain a continuous phase trajectory over time. Numerical algorithms are applied to detect discontinuities greater than π and correct them by adding or subtracting 2π accordingly.

3. Error Mitigation Under Noise:

In practical environments, phase measurements are affected by thermal noise, multipath interference, and oscillator jitter. To mitigate these effects:

- Averaging and smoothing filters (e.g., moving average, Kalman filters) are applied to suppress high-frequency noise;
- Differential phase analysis between dual-frequency signals is used to cancel common-mode phase drift;
- Thresholding and signal quality estimation are used to discard outlier measurements.

These computational techniques are lightweight and well-suited for real-time implementation on embedded systems, such as microcontrollers or FPGA-equipped SDR platforms.

3. Results

Let us consider a signal frequency f=433 MHz, which corresponds to:

$$\lambda = \frac{c}{f} = \frac{3 \cdot 10^8\, m/s}{433 \cdot 10^6 Hz} \approx 0.692 \text{ m}$$

If the receiver can resolve a phase difference of 1° (i.e., $\delta\phi$=π/180), then:

$$\delta L \approx \frac{0.692}{2\pi} \cdot \frac{\pi}{180} \approx 0.61 \text{ mm}$$

With more advanced DSP (e.g., Hilbert transform + filtering), systems can achieve sub-degree phase resolution, making it possible to detect displacements on the order of 0.1 mm or better.

If we use two coherent signals with a small frequency separation Δf, the phase difference between them increases linearly with distance:

$$\Delta\phi(t) = \frac{2\pi \cdot \Delta f}{c} \cdot L(t) \qquad (7)$$

Taking derivative over time:

$$\frac{dL}{dt} = \frac{c}{2\pi \cdot \Delta f} \cdot \frac{d(\Delta\phi)}{dt} \qquad (8)$$

This magnifies displacement changes into faster phase dynamics – effectively amplifying sensitivity.

To validate the theoretical framework proposed in this study, we present a series of simulated results that illustrate the key functional relationships between phase variation and physical displacement. These results demonstrate the high sensitivity and analytical predictability of the phase-based sensing method under realistic conditions. The Figure 1 illustrates the linear relationship between phase difference (in degrees) and the corresponding displacement (in millimeters). The result confirms the core equation $\Delta L = \frac{\lambda}{2\pi} \cdot \Delta\phi$, where small phase shifts directly correlate with measurable displacement. The wavelength corresponds to 433 MHz ($\lambda \approx 0.692$ m), and results are shown for phase shifts ranging from 0° to 180°.

Figure 2 illustrates the evolution of instantaneous phase over time under a condition of uniform ground motion (constant velocity of 0.5 mm/s). As expected, the phase increases linearly, confirming that the proposed method can track displacement in real time through continuous phase accumulation.

To analyze how operating frequency influences system performance, Figure 3 presents the dependency of displacement sensitivity (in mm per degree of phase shift) on the carrier frequency. The trend reveals that lower frequencies provide greater displacement sensitivity, supporting the selection of UHF bands for high-resolution sensing.

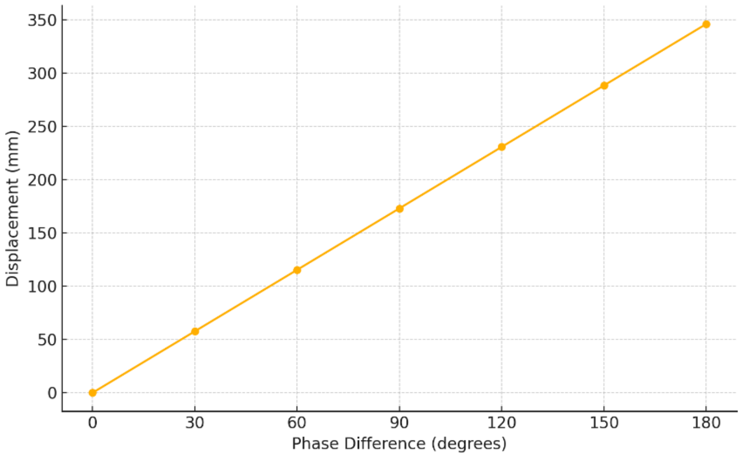


Figure 1 – Linear dependence of displacement on phase difference at 433 MHz.

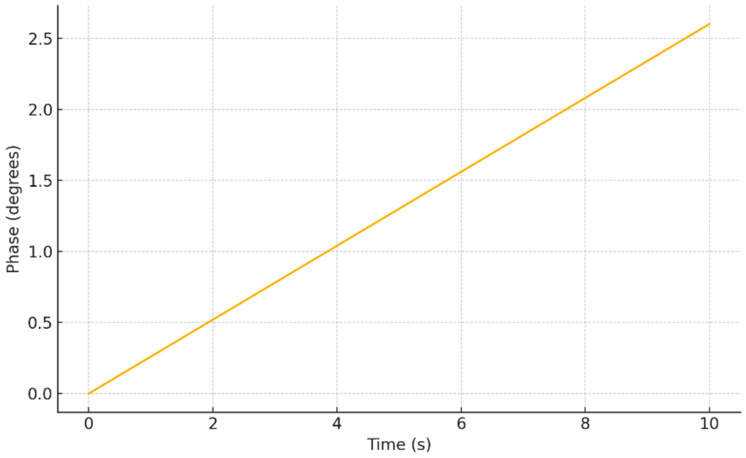


Figure 2 – Temporal evolution of phase in response to uniform displacement.

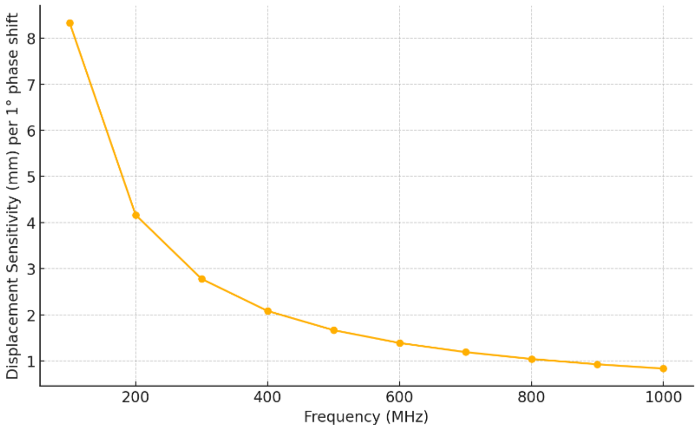


Figure 3 – Influence of carrier frequency on displacement sensitivity.

Finally, Figure 4 simulates the dual-frequency approach, where the phase difference between two closely spaced coherent signals (433 MHz and 434 MHz) is monitored over time during uniform displacement. The linearly increasing phase separation illustrates the enhanced resolution afforded by interferometric phase amplification, making this technique particularly robust in low-SNR or noisy environments.

Together, these simulation results demonstrate the feasibility, responsiveness, and computational tractability of the proposed phase-based method for precision displacement estimation in wireless settings. In the following discussion, we explore how these results compare to other displacement sensing technologies and outline the potential for embedded implementations using low-power SDR platforms.

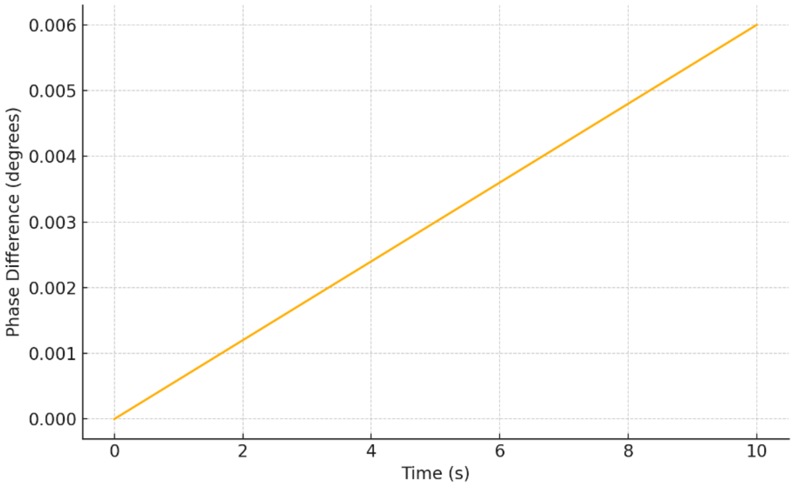


Figure 4 – Accumulation of dual-frequency phase difference during uniform motion.

4. Discussion

The results presented in the previous section confirm that phase-based interferometric sensing offers a theoretically grounded and computationally tractable method for precise displacement monitoring. In this section, we compare this approach with other commonly used geotechnical and structural monitoring techniques, while outlining its practical implications and unique advantages.

Comparison with Existing Technologies.

Distributed Fiber Optic Sensing (DFOS) technologies are widely regarded for their high spatial resolution and immunity to electromagnetic interference [1]-[5], [13]. However, their dependence on continuous optical fiber infrastructure imposes high installation costs and maintenance complexity, especially in remote or inaccessible terrains. Differential GPS (DGPS), on the other hand, provides centimeter-level accuracy [6], [14], [15], but suffers from performance degradation in environments with obstructed satellite visibility and typically requires expensive base station infrastructure. Mechanical inclinometers, while cost-effective and easy to deploy [16], lack real-time capabilities and are limited in resolution, especially for detecting slow or micro-scale deformations.

Novelty of the Proposed Method.

Unlike radar interferometry and DFOS, which rely on backscattered or embedded signals, the proposed method does not require any reflected wave or physical contact with the monitored medium. Instead, it utilizes a wireless forward-link between a transmitter and receiver, with displacement inferred solely from changes in phase difference. This makes the system inherently non-intrusive, infrastructure-light, and more deployable in challenging environments such as mountainous slopes, rural infrastructure, or temporary installations.

Computational and Energy Efficiency.

The method benefits from its low computational overhead: phase difference extraction can be performed in real time using signal demodulation and Hilbert transform techniques. Moreover, because the system can operate in a duty-cycled mode (e.g., one-minute transmission every five minutes), it is compatible with energy-autonomous deployments powered by compact solar panels or batteries. The use of GPS-disciplined oscillators (GPSDO) ensures that phase coherence is maintained across sessions, even in low-power states.

Limitations and Future Directions.

Despite its promise, the method does face certain limitations. The range is constrained by the power of the transmitter and antenna design, typically below 2 km in line-of-sight conditions without amplification. Environmental noise and multipath effects may require careful site-specific calibration and filtering. Future work should focus on multi-node configurations, real-time signal quality assessment, and adaptive algorithms to compensate for noise-induced phase anomalies.

Overall, the phase-based displacement sensing method proposed here provides a compelling alternative to existing technologies, particularly for use in distributed, low-cost, and real-time monitoring systems. It complements – not replaces – other methods, and is especially suited for lightweight deployments where traditional infrastructure is unavailable or impractical.

5. Conclusions

This study has presented a theoretical and computational framework for a novel phase-based interferometric method for displacement sensing. Unlike conventional techniques such as radar interferometry, differential GPS, or distributed fiber optic sensing, the proposed approach does not rely on reflected signals or physical embedding of sensors. Instead, it leverages the phase difference between two coherent radio signals transmitted through a wireless forward link, with displacement estimated based on analytical relationships between phase and path length.

Through a series of simulations, we have demonstrated the method's sensitivity, resolution, and scalability. The results confirm that sub-millimeter displacement can be detected using modest hardware configurations operating in the UHF band. Additionally, computational procedures such as instantaneous phase extraction and phase difference analysis can be executed in real time, making the system suitable for embedded and low-power applications.

The proposed method complements existing displacement monitoring technologies by offering a cost-effective, lightweight, and scalable solution for geohazard detection, structural health monitoring, and remote sensing scenarios. Its low energy footprint and wireless nature make it particularly attractive for deployment in difficult-to-access locations where conventional technologies are impractical.

Future work will focus on prototyping, real-world validation, and the development of a software-defined radio (SDR) implementation for field experiments. Moreover, expanding the system into a multi-node sensing network may open new possibilities for spatially distributed interferometric sensing in environmental and civil monitoring contexts.

Funding

This research has been funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP27510563).

Conflicts of Interest

The author declares no conflict of interest.

References

1. J. Wang, Z. Xiong, S. Li, H. Lu, M. Sun, Z. Li, and H. Chen, "Research on displacement monitoring of key points in caverns based on distributed fiber optic sensing technology," *Sensors*, vol. 25, no. 8, Art. no. 2619, 2025, doi: 10.3390/s25082619.
2. M. F. Bado and J. R. Casas, "A review of recent distributed optical fiber sensors applications for civil engineering structural health monitoring," *Sensors*, vol. 21, no. 5, Art. no. 1818, 2021, doi: 10.3390/s21051818.
3. P. Lu, N. Lalam, M. Badar, B. Liu, B. T. Chorpening, M. P. Buric, and P. R. Ohodnicki, "Distributed optical fiber sensing: Review and perspective," *Appl. Phys. Rev.*, vol. 6, no. 4, Art. no. 041302, 2019, doi: 10.1063/1.5113955.
4. R. C. S. B. Allil, L. A. C. Lima, A. S. Allil, and M. M. Werneck, "FBG-based inclinometer for landslide monitoring in tailings dams," *IEEE Sens. J.*, vol. 21, no. 15, pp. 16670–16680, 2021, doi: 10.1109/JSEN.2021.3081025.
5. E. Damiano, M. de Cristofaro, E. Molitierno, and L. Olivares, "DFOS-based inclinometers: Challenges and potentialities in monitoring slow landslides," *Procedia Struct. Integr.*, vol. 64, pp. 1628–1635, 2024, doi: 10.1016/j.prostr.2024.09.418.

6. A. Sivasuriyan, D. S. Vijayan, P. Devarajan, A. Stefańska, S. Dixit, A. Podlasek, W. Sitek, and E. Koda, "Emerging trends in the integration of smart sensor technologies in structural health monitoring: A contemporary perspective," *Sensors*, vol. 24, no. 24, Art. no. 8161, 2024, doi: 10.3390/s24248161.

7. W. K. Chiu, W. H. Ong, T. Kuen, and F. Courtney, "Large structures monitoring using unmanned aerial vehicles," *Procedia Eng.*, vol. 188, pp. 415–423, 2017, doi: 10.1016/j.proeng.2017.04.503.

8. B. W. Isah, H. Mohamad, and N. R. Ahmad, "Rock stiffness measurements fibre Bragg grating sensor (FBGs) and the effect of cyanoacrylate and epoxy resin as adhesive materials," *Ain Shams Eng. J.*, vol. 12, no. 2, pp. 1677–1691, 2021, doi: 10.1016/j.asej.2020.09.007.

9. G. Morgenthal, J. F. Eick, S. Rau, and J. Taraben, "Wireless sensor networks composed of standard microcomputers and smartphones for applications in structural health monitoring," *Sensors*, vol. 19, no. 9, Art. no. 2070, 2019, doi: 10.3390/s19092070.

10. S. B. Im, S. Hurlebaus, and Y. J. Kang, "Summary review of GPS technology for structural health monitoring," *J. Struct. Eng.*, vol. 139, no. 10, pp. 1653–1664, 2013, doi: 10.1061/(ASCE)ST.1943-541X.0000475.

11. D. Massonnet and K. L. Feigl, "Radar interferometry and its application to changes in the Earth's surface," *Rev. Geophys.*, vol. 36, no. 4, pp. 441–500, 1998, doi: 10.1029/97RG03139.

12. Q. Lin, S. Li, and W. Yu, "Review on phase synchronization methods for spaceborne multistatic synthetic aperture radar," *Sensors*, vol. 24, no. 10, Art. no. 3122, 2024, doi: 10.3390/s24103122.

13. Z. Sun, X. Wang, T. Han, H. Huang, J. Ding, L. Wang, and Z. Wu, "Pipeline deformation monitoring based on long-gauge fiber-optic sensing systems: Methods, experiments, and engineering applications," *Measurement*, vol. 248, Art. no. 116911, 2025, doi: 10.1016/j.measurement.2025.116911.

14. K. Kim, J. Choi, J. Chung, G. Koo, I. H. Bae, and H. Sohn, "Structural displacement estimation through multi-rate fusion of accelerometer and RTK-GPS displacement and velocity measurements," *Measurement*, vol. 130, pp. 223–235, 2018, doi: 10.1016/j.measurement.2018.07.090.

15. G. E. Vázquez B., J. R. Gaxiola-Camacho, R. A. Bennett, G. M. Guzman-Acevedo, and I. E. Gaxiola-Camacho, "Structural evaluation of dynamic and semi-static displacements of the Juarez Bridge using GPS technology," *Measurement*, vol. 110, pp. 146–153, 2017, doi: 10.1016/j.measurement.2017.06.026.

16. B. H. van Duren, M. Al Ashqar, J. N. Lamb, H. G. Pandit, and C. Brew, "A novel mechanical inclinometer device to measure acetabular cup inclination in total hip arthroplasty," *J. Med. Eng. Technol.*, vol. 44, no. 8, pp. 481–488, 2020, doi: 10.1080/03091902.2020.1825846.

Information about authors

Amirkhan Temirbayev was born in Tashkent, Uzbekistan, in April 1986. He received the Ph.D. degree in Physics from the Al Farabi KazNU, in 2012, Kazakhstan. Currently he is a General Director of Cluster of Engineering and High Technologies at Al-Farabi KazNU. Dr. Amirkhan A. Temirbayev is involved in the design of a KazNU's nanosatellites. Repeatedly passed scientific training in leading centers in Germany, Japan, Holland. His research interests include nanosatellites, subsystems for CubeSats, STEM education, Information and Communication Technologies. ORCID iD: 0000-0001-6759-2774.

Submission received: 24 May, 2025.
Revised: 4 August, 2025.
Accepted: 4 August, 2025.

IRSTI 28.23.29 https://doi.org/10.26577/jpcsit2025335

Zhansaya Abildaeva[1], **Raissa Uskenbayeva**[1], **Nurbek Konyrbaev**[2]*, **Gulzhanat Beketova**[3,4], **Valery Lakhno**[5], **Alona Desiatko**[6]

[1]Kazakh National Technical Research University named after K. I. Satbayeva, Almaty, Kazakhstan
[2]Institute of Engineering and Technology, Korkyt Ata Kyzylorda University, Kyzylorda, Kazakhstan
[3]Almaty University of Energy and Communications G. Daukeeva, Almaty, Kazakhstan
[4] Academy of Logistics and Transport, Almaty, Kazakhstan, Almaty, Kazakhstan
[5] National University of Life and Environmental Sciences of Ukraine, Ukraine
[6] State University of Trade and Economics, Ukraine
*e-mail: nurbek@korkyt.kz

OPTIMIZATION OF MARKETING STRATEGIES IN THE AGRO-INDUSTRIAL COMPLEX OF KAZAKHSTAN BASED ON A HYBRID METHOD

Abstract. In the development of the agro-industrial complex (AIC) of the Republic of Kazakhstan, one of the key aspects is the improvement of information and consulting activities of enterprises and companies in the agricultural sector, in particular, aimed at increasing the efficiency of production and marketing of agricultural products. A significant aspect is also the development of agromarketing in the AIC, which will ultimately improve market mechanisms and increase the competitiveness of products of local agricultural producers. The study proposes a hybrid method for solving the problem of multi-criteria optimization of marketing strategies in the AIC. The method combines the use of the NSGA-II algorithm and machine learning based on K-means to analyze the results. The quality of the solution was assessed using the hypervolume parameters and visualization of the found optimal strategies using the Pareto front. The theoretical and practical significance of the study is confirmed by the possibility of adapting the proposed hybrid method for enterprises of the AIC of Kazakhstan, taking into account existing regional restrictions.

Keywords: agro-industrial complex (AIC), digital marketing, agromarketing, strategy, target functions, optimization, hybrid method, machine learning, models, agricultural products.

1. Introduction

As part of the strategic development of the agro-industrial complex (AIC) of the Republic of Kazakhstan (hereinafter RK), one of the priority areas is the improvement of information and consulting work with enterprises and companies in the agricultural sector. This process is aimed at increasing production efficiency, improving the marketing of agricultural products and creating favorable conditions for sustainable growth of the industry. The main areas in this area are: dissemination of information on the latest technologies and developments introduced in the agricultural sector, the creation of regional consulting centers, as well as the involvement of foreign specialists, which will help improve the qualifications of local specialists. An essential element of this process is the analysis of the needs of agricultural market participants in order to optimize offers and improve interaction between various segments of the industry [1], [2].

An essential factor for the successful development of the agro-industrial complex of the Republic of Kazakhstan is agromarketing, which contributes to the improvement of market mechanisms and increased competitiveness of local producers. Marketing in the agro-industrial complex of Kazakhstan has a number of specific features due to the structural features of the economy of the Republic of Kazakhstan, geographical and demographic conditions, as well as the level of digitalization of the agricultural sector. Unlike marketing in traditional consumer segments, the promotion of agricultural products in the conditions of Kazakhstan requires taking into account many factors, such as significant spatial dispersion of producers and consumers, limited infrastructure and seasonality of demand [16], [20].

The specific features of the agricultural sector of Kazakhstan are largely related to climatic and natural conditions that affect the stability of production cycles and, as a consequence, the variability of supply. This, in turn, requires the implementation of adaptive marketing strategies that are able to flexibly respond to changes in the volume and structure of agricultural production. An essential aspect is the need to develop marketing solutions that would effectively interact with changing market conditions and taking into account the instability in agricultural production [1], [13].

Another characteristic feature is the prevalence of B2B communications over B2C models. A significant share of agricultural products in Kazakhstan is sold through wholesale channels and processing enterprises, which necessitates the development of marketing strategies aimed not only at the end consumer, but also at the corporate segment. This requires a different logic of interaction, as well as the use of more complex methods of analytics and demand forecasting. In this regard, marketing strategies should include flexible mechanisms that would take into account the interests and needs of both business and the end consumer.

The third significant feature is the insufficient level of digitalization of rural areas of the Republic of Kazakhstan, which limits the possibilities of using standard digital communication channels, such as social networks and targeted advertising. This requires the development of hybrid marketing models that would combine online and offline activities, as well as use traditional means of informing consumers. However, with the growing availability of the Internet and mobile devices in rural areas, new opportunities are opening up for the use of machine learning (ML) and data mining (DMP) technologies. These methods can be effectively used in areas such as demand forecasting, market segmentation and personalization of offers for agricultural products, which helps to increase the effectiveness of marketing campaigns.

In addition, it is necessary to take into account the regional specifics of demand in Kazakhstan, which depends on ethno cultural preferences, income levels of the population and logistics capabilities both within the country and in neighboring states. Regional differences can significantly affect the demand for agricultural products, which requires the creation of more differentiated and adapted marketing strategies.

Based on the analysis of the current state of agricultural production infrastructure in Kazakhstan, conducted in works [1], [3], [4], it can be concluded that the existing mechanisms aimed at supporting marketing and agricultural production are insufficient to achieve effective results. Despite this, a number of studies, such as works [1], [14], propose specific steps to create and optimize marketing systems that are aimed at improving technological links and logistics in the process of movement of agricultural goods. These proposals are an important step towards improving the state of the agricultural market in Kazakhstan, increasing the competitiveness of local producers and ensuring sustainable development of the agro-industrial complex. Digitalization of the agro-industrial complex of Kazakhstan has given rise to the problem of optimal distribution of marketing budgets between various channels for promoting agricultural products. And traditional methods based on heuristics or single-criterion optimization [5], [6], [7] do not take into account the multiplicity of target indicators, such as advertising effectiveness [8], audience reach [9], and the cost of an advertising campaign [9], leading, as a rule, to suboptimal solutions. In this paper, the multi-criteria evolutionary algorithm NSGA-II (Non-dominated Sorting Genetic Algorithm II) [10], [11] is proposed to solve the problem of Pareto-optimal distribution of marketing resources in the agro-industrial complex. The relevance of the study is due to the need to develop adaptive strategies that can take into account non-linear dependencies between budgetary distributions of items and key metrics of digital marketing efficiency in the agro-industrial complex. Unlike classical methods, where scalar optimization dominates [12], the hybrid method proposed in the article will simultaneously maximize efficiency and coverage while minimizing costs, meeting the real requirements of agro-industrial enterprises of the agro-industrial complex of Kazakhstan, which usually operate under limited budgets [17], [18], [19].

2. Materials and methods

The aim of the study is to develop and implement digital marketing solutions based on machine learning, focused on the specifics of the agro-industrial complex of Kazakhstan.

2.1. The main material of the article

The reform of the agro-industrial complex of the Republic of Kazakhstan in recent decades, despite the complex challenges, has also prompted the sector to adapt to market conditions. In particular, the optimization of digital marketing and advertising processes has become an important tool for agricultural producers. In the market conditions, not only efficient production is becoming increasingly important, but also attention to the needs of end consumers of agricultural products, which is relevant for the agriculture of Kazakhstan, where the production of food, raw materials and food directly affects the socio-economic well-being of rural areas and the security of the state as a whole. Modern agricultural organizations in Kazakhstan have a great influence on the development of rural settlements, often becoming the main employers and economic centers for the local population [15]. It is important to note that the success of agricultural production today largely depends on how well the marketing processes are organized, on the focus on the consumer and on the construction of an effective system of commodity circulation. Otherwise, even successful production may face economic difficulties and a deterioration in the social situation in rural areas. The task of not only state policy, but also agricultural entrepreneurs is to create conditions for the effective operation and development of agricultural enterprises focused on modern market tools and requirements. It is important to invest in the education of agricultural leaders, farmers and managers of agricultural enterprises who will be able to effectively manage marketing and production processes. In this regard, the introduction of digital solutions in marketing for effective communication with end consumers is becoming especially important. Given the dynamically changing food market, agricultural enterprises in Kazakhstan should develop flexible and low-cost marketing mechanisms focused on consumer needs and innovative approaches to product promotion. It is essential that marketing strategies be developed taking into account the specifics of local markets, using modern technologies and tools, including digital platforms, to maximize the availability and attractiveness of products to consumers. Thus, marketers and agricultural producers should focus on studying market needs, reducing costs, improving product availability and an effective advertising strategy. At the same time, it is necessary to study how government protectionist measures can be used to support domestic producers. An important aspect is the creation of an effective marketing system based on a project approach and process management, which will help strengthen the competitiveness of the agricultural sector of Kazakhstan in the domestic and international markets.

Mathematically, the problem includes a vector of target functions reflecting the effectiveness, coverage and cost of marketing campaigns, as well as a system of constraints taking into account budget and industry specifications. Formally, the problem is reduced to finding a set of Pareto-optimal budget allocations, for which a modified NSGA-II with adapted crossover and mutation operators is used. The approach is validated on synthetic data simulating real conditions of agro-industrial marketing, with subsequent assessment of the quality of solutions through hypervolume (Hypervolume indicator) and visualization of the 3D Pareto front; real data on enterprises of the Republic of Kazakhstan in the agro-industrial complex were also used.

Let's consider the problem of distributing the marketing budget betweenchannels – digital, TV, Radio, Print, Events (exhibitions, presentations, etc.). We assume that for each channeland segmentthe following parameters are set: efficiency, audience reach, price. The problem of distributing the marketing budget between $n = 5$ channels – digital (x_1), TV (x_2), Radio (x_2), Seal (x_4), Events (exhibitions, presentations, etc.) (x_5). We assume that for each channel (i) and segment $(s \in [0,1])$ the following parameters are set: efficiency $e_{s,i} \in [0,1]$, audience reach $c_{s,i} \in [0,1]$, price $p_i > 0$. We will also define the efficiency and coverage matrices: $E_s = [e_{s.1}\ e_{s.2}\ e_{s.3}\ e_{s.4}\ e_{s.5}]$, $S_s = [s_{s.1}\ s_{s.2}\ s_{s.3}\ s_{s.4}\ s_{s.5}]$. And the cost vector $P = [p_1\ p_2\ p_3\ p_4\ p_5]^T$.

Let's describe the objective functions. It is required to maximize the overall efficiency $f_1(X,s) = \sum_{i=1}^{5} x_i \cdot e_{s,i}$ and overall audience reach $f_2(X,s) = \sum_{i=1}^{5} x_i \cdot c_{s,i}$. And accordingly, minimize the cost of implementing a certain marketing strategy and advertising $f_3(X) = \sum_{i=1}^{5} x_i \cdot p_i \rightarrow max.$

In the problem of multicriterial optimization we are looking for Pareto-optimal solutions: $\underset{X\in F}{max}\left(f_1(X), f_2(X)\right), \underset{X\in F}{\min} f_3(X)$, Where F – set of feasible solutions.

And we will formulate the corresponding restrictions. On the general budget $\sum_{i=1}^{5} x_i < 2,0.$ Min/Max Channel Shares $0,1 \le x_1 \le 0,9$, $0,1 \le x_2 \le 0,9$, $0,1 \le x_3 \le 0,9$, $0,1 \le x_4 \le 0,5$, $0,1 \le x_5 \le 0,2.$ And combined restrictions $x_1 + x_2 \ge 0,2$, $x_1 \ge 0,2.$

As a solution method, we use NSGA-II at the first stage to search for the Pareto front. The quality assessment (Hypervolume Indicator) is performed as follows

$$HV = Volume\left(\bigcup_{X\in P} \left|f_1(X), r_1\right| \times \left|f_2(X), r_2\right| \times \left|r_3, f_3(X)\right|\right),$$

where P – Pareto front, r – reference point. To test the proposed hybrid method, a computational experiment was conducted. The initial data for the two segments was taken as follows based on data from several enterprises in the agro-industrial complex of Kazakhstan:

$$E_0 = [0.95, 0.75, 0.85, 0.60, 0.90],$$
$$C_0 = [0.85, 0.65, 0.80, 0.50, 0.75],$$
$$E_1 = [0.85, 0.85, 0.75, 0.65, 0.80],$$
$$C_0 = [0.75, 0.75, 0.70, 0.55, 0.70].$$

And the cost $P = [1.1, 0.8, 1.0, 0.6, 0.9]$.

The solutions found must satisfy the Pareto-optimal solution, i.e.

$$X^* = \arg\max_{X\in F} (f_1, f_2), \arg\min_{X\in F} f_3.$$

2.2. Methodology details.

In this study, the NSGA-II algorithm was applied with the following parameters: population size = 100, number of generations = 200, crossover probability = 0.9, mutation probability = 0.1. Initial solutions were seeded using uniform distribution of the budget across channels. Constraints were handled by penalizing infeasible solutions exceeding budget or violating channel share bounds.

For the hypervolume indicator, normalization of all objectives was performed, and the reference point was set to (0,0, max cost). K-means clustering was applied after normalization of solutions, with the optimal number of clusters determined by the elbow method (k=3). Standard scaling was used to balance efficiency, coverage, and cost in clustering.

K-means clustering was applied on normalized solutions using k-means++ initialization. The number of clusters (k=3) was chosen based on the elbow method and validated with silhouette score (0.67) and Davies–Bouldin index (0.54), indicating robust separation. Experiments used both synthetic and real aggregated data. The synthetic dataset simulates budget allocations across 5 channels (Digital, TV, Radio, Print, Events) with varying effectiveness, coverage, and costs. Real aggregated data were collected from 10 enterprises of Kazakhstan's agro-industrial complex for 2018–2022, including budget allocations, efficiency (ROI, %), coverage (thousand people), and costs (K USD). Only anonymized aggregated statistics are reported for confidentiality.

3. Results

The results of the computational experiments are presented in Table 1 (synthetic data) and in Figures 1, 2.

Table 1 – Results of modeling the analysis of the effectiveness of various media channels for promoting agricultural products on the RK market.

Cluster	Avg Efficiency	Avg Coverage	Avg Cost	Avg Score	Key Channels
0	1.260	1.110	1.290	0.690	TV, Print, Radio
1	0.680	0.600	0.750	0.610	Radio, Print
2	1.560	1.390	1.730	0.660	Digital, TV

The conducted research was aimed at analyzing the effectiveness of various media channels (Digital, TV, Radio, Print, Events) to study their impact on the indicators of effectiveness (Efficiency), coverage (Coverage) and cost (Cost) for various segments (Segment) and clusters (Cluster). As a result of data processing, the following key patterns were identified. Cluster 0 demonstrates high values of the Efficiency and Coverage indicators, which indicates the high effectiveness of media strategies focused on this segment. The greatest contribution is made by TV and Print channels.

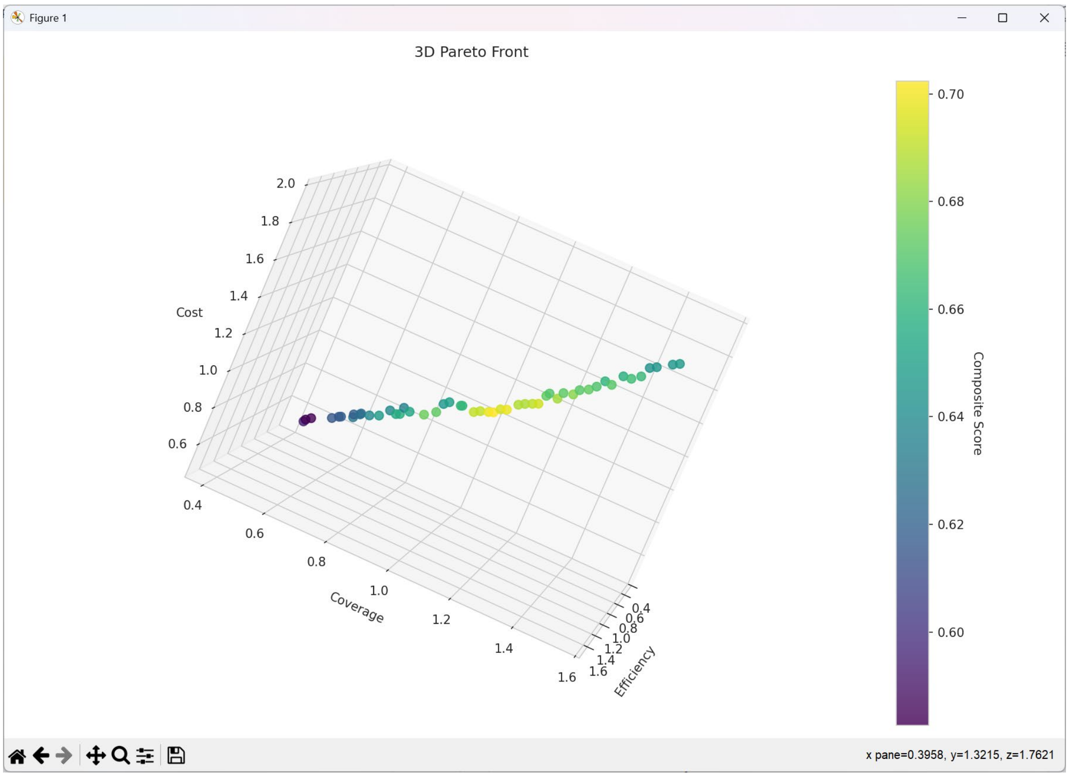


Figure 1 – 3D Pareto front in 3D format, formed as a result of applying the hybrid method. Axes: Efficiency (ROI, %), Coverage (audience ×10³), Cost (K USD).

The results of the study allowed us to conclude that the effectiveness of media strategies significantly depends on the choice of channels and their adaptation to specific segments and clusters. The greatest effectiveness is achieved with the combined use of Digital and TV, especially in cluster 2. The data obtained can be used to optimize media planning and increase the ROI of marketing campaigns of agricultural enterprises in Kazakhstan.

Figure 2 – 2D Pareto frontier reflecting the trade-off between the effectiveness of marketing strategies and the reach of the target audience. Axes: Efficiency (ROI, %), Coverage (audience ×10³). Marker size represents cost (K USD).

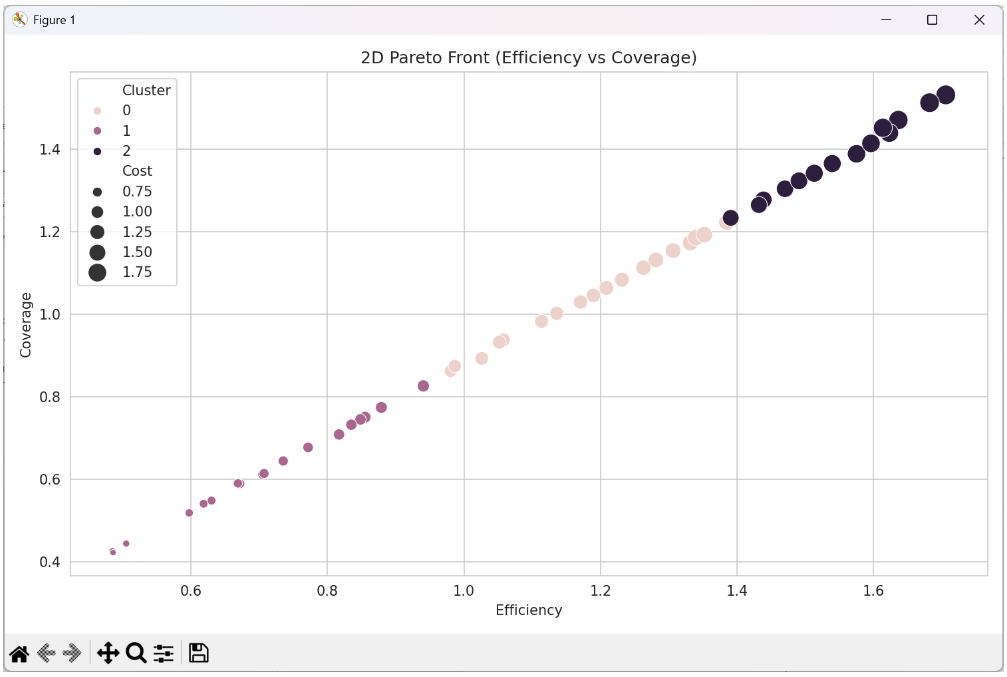


Figure 2 – 2D Pareto frontier reflecting the trade-off between the effectiveness of marketing strategies and the reach of the target audience. Axes: Efficiency (ROI, %), Coverage (audience ×10³). Marker size represents cost (K USD).

The size of the markers on the graph is proportional to the cost of implementing the strategy, providing a clear representation of the relationship between the key parameters – efficiency and coverage. Analysis of the distribution of solutions shows that the strategies of cluster 0 are characterized by relatively high efficiency with moderate coverage, while cluster 1 combines solutions with increased coverage, but lower efficiency. Cluster 2 includes strategies with intermediate values of both indicators. The upper right part of the graph shows the solutions demonstrating the best values of both efficiency and coverage, but their implementation is associated with higher costs, which is evident from the increased size of the corresponding markers.

4. Discussion

Such strategies can be recommended for use in conditions of sufficient budget for enterprises in the agro-industrial complex. On the contrary, solutions with smaller marker sizes, corresponding to lower costs, can be preferable for agro-industrial enterprises with limited resources.

It should be noted that the novelty of the model lies in the development of a comprehensive method combining evolutionary algorithms and machine learning for marketing optimization tasks in the agro-industrial complex. And the practical significance is confirmed by the application of the obtained solutions for planning advertising campaigns taking into account the regional characteristics of agro-industrial enterprises in Kazakhstan and resource constraints. The proposed approach expands the arsenal of decision-making methods in the conditions of multi-criteria and uncertainty characteristic of agro-industrial marketing. Further research can be aimed at integrating predictive models of channel efficiency and taking into account dynamic changes in market conditions.

4.1. Reproducibility.

To ensure reproducibility, the random seed for NSGA-II and K-means was fixed (seed=42). The code used for optimization and clustering is available from the authors upon request. The synthetic dataset can be regenerated following the described procedure, while aggregated real data statistics are provided in the Data description section.

4.2. Correlation analysis.

A correlation matrix was constructed to support the observed relationships between channels.

Pearson's correlation coefficients are shown in Table 2. Results indicate a strong positive correlation between Digital and Radio (0.88), a moderate positive correlation between Digital and TV (0.65), and a negative correlation between Print and other channels (around -0.40).

Table 2 – Correlation matrix of media channels

	Digital	TV	Radio	Print	Events
Digital	1.00	0.65	0.88	-0.42	0.21
TV	0.65	1.00	0.55	-0.36	0.30
Radio	0.88	0.55	1.00	-0.41	0.25
Print	-0.42	-0.36	-0.41	1.00	-0.28
Events	0.21	0.30	0.25	-0.28	1.00

5. Conclusions

It is shown that the developed hybrid method combining the NSGA-II algorithm and K-means clustering provides an effective solution to the problem of multi-criteria optimization of marketing budgets under conditions of limited resources typical for a number of enterprises in the agro-industrial complex of Kazakhstan. The obtained Pareto-optimal solutions demonstrate a compromise between the criteria of efficiency, coverage and cost, which is confirmed by the analysis of the hypervolume and cluster structure of the decision front. The greatest efficiency is demonstrated by strategies with a combination of digital and television channels (cluster 2), while radio and print media are appropriate for segments with a moderate budget (cluster 1). Correlation analysis revealed a strong relationship between digital channels and radio (0.88), as well as a negative correlation of print media with other channels. The approach proposed in the article allows adapting marketing strategies to the dynamic conditions of the agro-industrial complex market in Kazakhstan, minimizing costs while maximizing key metrics.

Note: Avg Score is a normalized composite indicator combining efficiency, coverage, and cost with equal weights (0.33 each).

Author Contributions

Conceptualization, – Zh.A. and R.K.; Methodology – Zh.A, ; Software, N.K.; Formal Analysis, G.B.; Investigation, X.X.; Resources, V.L, A.D.; Data Curation, X.X.; Writing – Original Draft Preparation, .L, A.D.; Writing – Review & Editing, Zh.A.; Visualization, Zh.A. and R.K.; Supervision, Zh.A. and R.K.; Project Administration, Zh.A.; Funding Acquisition, Zh.A.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. G. T. Sultanova, *Problems and Prospects of the Agro-Industrial Complex Development in Kazakhstan*, Doctoral dissertation, 2020.

2. A. Jumabayeva, A. Sankussov, and T. Satbayeva, "Ensuring food security in Kazakhstan's market of fish products," *Brazilian Journal of Food Technology*, vol. 26, e2023079, 2023, doi: 10.1590/1981-6723.07923. ResearchGate

3. B. Khadys, D. Sikhimbayeva, and A. Bozhkarauly, "State regulation of the development of the agro-industrial complex of the Republic of Kazakhstan," *Journal of Advanced Research in Law and Economics*, vol. 9, no. 1(31), pp. 127–138, 2018, doi: 10.14505/jarle.v9.1(31).15. econpapers.repec.org

4. A. K. Otesheva, M. O. Myrzagaliyeva, K. I. Yesken, and Z. S. Atauly, "The role of agriculture in ensuring the security of the national economy," *Problemy Agorrynka*, no. 2, pp. 34–40, 2019.

5. V. Zakshevsky, S. Kh, D. Nekrasova, and A. Tuluganova, "Complex development strategy of agribusiness of the Orenburg region," *IOP Conf. Ser.: Earth Environ. Sci.*, vol. 274, 012007, 2019, doi: 10.1088/1755-1315/274/1/012007. MDPI

6. O. Andrianova, O. Kirillova, and J. Kirillova, "Basic principles of marketing as a management system of an agricultural enterprise," in *SGEM 2019 Conf. Proc.*, vol. 19, iss. 5.3, pp. 863–868, 2019.

7. R. Bilovol and A. Chaikina, "Persistence of competitiveness of enterprise marketing strategy in the digital economy," *Baltic Journal of Economic Studies*, vol. 2, no. 5, pp. 16–21, 2016.

8. I. F. Gorlov et al., “Marketing management in the agro-industrial complex in the context of digitalization,” in *Digital Economy and the New Labor Market: Jobs, Competences and Innovative HR Technologies*, 2020, pp. 219–226, doi: 10.1007/978-3-030-29586-8_27.

9. O. Kuzyk, “Information technologies in the formation of the marketing strategy of agricultural enterprises,” *Universal Journal of Agricultural Research*, vol. 11, no. 2, pp. 217–229, 2023, doi: 10.13189/ujar.2023.110201. MDPI

10. K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, “A fast and elitist multiobjective genetic algorithm: NSGA-II,” *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002, doi: 10.1109/4235.996017.

11. A. B. N. Djami, J. Mbozo’o, D. Ntamack, and B. A. Pofa, “Multi-objective optimization of cutting parameters in hard turning using the NSGA-II algorithm,” *Operations Research Forum*, vol. 5, 86, 2024, doi: 10.1007/s43069-024-00364-2. ResearchGate

12. A. Ahmed, A. Das, and A. J. Smola, “Scalable hierarchical multitask learning algorithms for conversion optimization in display advertising,” in *Proc. 7th ACM Int. Conf. on Web Search and Data Mining (WSDM ’14)*, 2014, pp. 153–162, doi: 10.1145/2556195.2556264. research.google.com

13. Y. Yu. Blinova, “Methodology of marketing research of the agricultural market,” *Marketing in Russia and Abroad*, no. 4, pp. 30–38, 2015.

14. N. V. Pershukov and O. S. Karnadud, “Marketing communications management in the enterprises of the agro-industrial complex,” *Vestnik ASTU*, no. 3, pp. 85–92, 2019.

15. K. V. Ponomarenko and I. V. Zatonskaya, “Marketing communications management at food industry enterprises,” in *Proc. Int. Sci.-Prac. Conf. Problems and Prospects for the Development of the Industry*, 2016, pp. 252–256.

16. A. Garrido, “Mathematical programming models applied to the study of water markets in Spain’s agricultural sector,” *Annals of Operations Research*, vol. 94, pp. 105–123, 2000, doi: 10.1023/A:1018979010078. hrpub.org

17. S. Kozlovskiy, V. Khadzynov, A. Lavrov, V. Skaydan, O. Ivanyuta, and I. Varshavska, *Economic-Mathematical Modeling of the Competitiveness of the Agricultural Sector*, Monograph, 2019.

18. O. I. Laburtseva, N. I. Golik, O. V. Ryabchenko, and S. V. Kolisnichenko, “Strategic communications in the development of the agricultural sector,” *Research in Applied Economics*, vol. 39, no. 5, 2021.

19. S. V. Kovalchuk, O. V. Udovenko, L. E. Antonova, and N. I. Kosenko, “On the problems of implementing Internet marketing in the industry,” *Proc. Int. Sci.-Prac. Conf. Innovatsiyni Mekhanizmy Upravlinnya*, 2019.

20. T. Hung-Yi, “Explanation of an explainable AI framework for efficient employee hiring using feature importance,” in *Proc. IEEE ICITEICS 2024*, pp. 1–5, 2024.

Information about authors

Zh.T. Abildayeva – first author, 3rd year doctoral student of the “Software engineering” program, Department of Software Engineering, Kazakh National Technical Research University named after K. I. Satbayeva, Almaty, 050013, Kazakhstan. https://orcid.org/0000-0002-2637-0443

R.K. Uskenbayeva – Corresponding author, Professor, Vice-Rector of the Kazakh National Technical Research University named after K. I. Satbayeva, Almaty, 050013, Kazakhstan. https://orcid.org/0000-0002-8499-2101

N.B. Konyrbayev – Corresponding Author, Professor of the Department of Computer Science, PhD, Institute of Engineering and Technology, Korkyt Ata Kyzylorda University, Aiteke Bi Street. 29a, Kyzylorda 120014, Kazakhstan. https://orcid.org/0000-0002-8788-4149

G.S. Beketova – author, PhD, Almaty University of Energy and Communications named after Gumarbeka Daukeeva, department – "IT-engineering and artificial intelligence", position – associate professor. https://orcid.org/0000-0001-7160-1514

Valerii Lakhno – Professor of the Department of Computer Systems, Networks and Cybersecurity, National University of Life and Environmental Sciences of Ukraine, Ukraine.email: lva964@nubip.edu.ua. https://orcid.org/0000-0001-9695-4543

A. Desyatko – PhD, Associate Professor at the Department of Software Engineering and Cyber Security State University of Trade and Economics. Her research focuses on developing models to understand and analyze the dynamics and interactions within economic and financial systems. Among her other areas of research are cloud technologies, information systems, cybersecurity, Product IT, Project IT, software architecture. She can be contacted on this email address, desyatko@gmail.com https://orcid.org/0000-0002-2284-3418

Submission received: 7 July, 2025.
Revised: 22 August, 2025.
Accepted: 22 August, 2025.

IRSTI 20.19.27 https://doi.org/10.26577/jpcsit2025336

Assel Ospan* , Kanat Auyesbay ,
Talshyn Sarsembayeva , Aman Mussa
Al Farabi Kazakh National University, Almaty, Kazakhstan
*e-mail: asselyaospan@gmail.com

APPLICATION OF RULE-BASED METHOD FOR AUTOMATIC EXTRACTION OF TAGS FROM COLUMN-STYLE PDF-DOCUMENTS

Abstract. This study presents a rule-based hybrid pipeline for the automated extraction of structured metadata from PDF versions of Kazakh-language newspaper articles, focusing on the national newspaper Egemen Qazaqstan. The primary goal is to support the development of a machine-readable knowledge base for future use in training large language models (LLMs) and building an AI-powered assistant for data journalism in Kazakhstan. The pipeline integrates three open-source parsers – pdfminer.six, PyMuPDF, and pdfplumber – to extract key elements such as title, author, date, abstract, text, journal name, and category. To evaluate extraction quality, we compared the results of the automated parser against manually annotated reference files across three real-world issues of the newspaper. The evaluation employed three complementary metrics: Precision, Textual Semantic Similarity (TSS), and Holistic Precision, which jointly assess both exact and semantic matches. The experimental results show that three structured tags – date, journal, and category – achieved perfect Holistic Precision (1.00), while the remaining tags obtained high scores (author 0.93, abstract 0.98, text 0.91, title 0.85), yielding a macro-average Holistic Precision of 0.95. The validated pipeline was then applied to the full corpus of 2,140 newspaper PDFs published between 2017 and March 2025, successfully converting 159,135 articles into structured JSON format. This enriched corpus serves as a foundational knowledge base for Kazakh-language AI systems in journalism and media analysis.

Keywords: PDF parsing, Rule-based extraction, Metadata extraction, Document structure recognition, Text mining, Low-resource language processing, Knowledge base for LLMs.

1. Introduction

Currently, the development of an AI assistant for data journalism in the Kazakh language is gaining particular relevance due to the rapid advancement of large language models (LLMs), such as OpenAI GPT-4 [1], Google Gemini [2], Meta LLaMA [3], and others. One of the key factors determining the effectiveness of such models is the availability of a high-quality corpus in the target language. However, for the Kazakh language, there is a significant shortage of digital and structured texts, especially in the domain of official and analytical journalism. Most previously published materials exist only as scanned or digital PDF versions of newspapers, lacking HTML markup and accessible APIs, which significantly complicates the task of automatic data extraction and annotation.

In this context, it is particularly important to enrich the model's knowledge base with only verified information from trusted sources. Newspaper articles that have undergone editorial and journalistic reviews serve as a reliable foundation for building such corpora. To choose an optimal source, our team consulted professional journalists. Based on their recommendation, archival issues of the newspaper Egemen Qazaqstan [4] were selected as the base corpus – a leading national publication known for its credibility and stylistic quality.

The use of PDF versions of newspapers is due to the fact that the official Egemen Qazaqstan website provides access to archives only from 2017 onward. A significant portion of the materials from 2017 to 2020 was published mainly in print or as PDF files, without HTML versions or machine-readable markup. This limits the ability to automatically collect and structure texts using standard web tools.

As a result, there is a need to apply specialized methods for extracting text and metadata from PDF documents with column layouts, embedded illustrations, and Kazakh fonts. Developing an effective hybrid pipeline to process such materials becomes a key step in building a high-quality corpus suitable for training Kazakh-language LLMs aimed at analysis and generation tasks in data journalism.

In scientific literature, various approaches to text extraction from PDFs are proposed, primarily divided into rule-based and learning-based methods.

Rule-based methods rely on predefined rules using regular expressions, keywords, positional and typographic cues to identify meaningful text elements. Their main advantages include interpretability, flexibility in adapting to specific documents, and no need for training data. In our prior work, we have repeatedly relied on rule-based and hybrid pipelines to structure data from unstructured sources. In [5], authors introduced TableProcessor, which used rule-driven parsers and schema mapping to interpret statistical tables from the Bureau of National Statistics (Word/Excel) and convert them into JSON suitable for GIS and knowledge-base ingest. In [6], it was proposed an ontology-guided, iterative method that combines deterministic rules with neural NER to classify cells and expand a domain ontology, yielding RDF triples from semi-structured tables. Building on the same design principles, [7] presented Qurma, a modular system that extracts tables from HTML/PDF/images and applies rule-based normalization and semantic interpretation within a Clean Architecture for scalability and domain portability. These experiences directly inform the present newspaper pipeline – specifically our use of regular expressions, keyword lexicons, and positional/typographic cues – while preserving interpretability and eliminating the need for training data. A successful rule-based approach is presented in [8], where Alamoudi et al. extract metadata from PDF books using PDFBox and logical rules to identify key elements such as book title, author name, and ISBN; their rule set encodes deterministic patterns and constraints to maximize accuracy, yielding overall accuracies of 94.62% (training) and 90.27% (test) on their dataset, which underscores the effectiveness of rule-based designs for document-specific metadata extraction. Similarly, [9] describes a method for citation metadata extraction using Hidden Markov Models, classifying token sequences based on reference structure patterns. In the PDFBoT tool [10], the authors propose a system for high-precision extraction of the main text from academic PDFs using HTML replication, formatting analysis, and syntactic cues. The method achieves high accuracy while preserving sentence and paragraph structure, removing auxiliary elements like tables and images.

In contrast, learning-based methods (based on machine learning and neural network architectures) are used to automatically train models to extract data from documents with complex or unstable structures. The authors of [11] proposed an end-to-end neural architecture for image-based sequence recognition, which has been successfully applied to scene text recognition tasks. In the research work [12], an OCR-free Document Understanding Transformer was developed, which uses a transformer architecture to understand document structure without relying on traditional OCR. The authors of [13] introduced TrOCR, a transformer-based optical character recognition model pre-trained on large-scale corpora, enabling high accuracy across a wide range of document types. Additionally, for learning-based methods, researchers created DocLayNet [14], a large annotated dataset for document layout segmentation, which is used to train models for automatic document structure analysis.

Given the limited availability of annotated data, the specific layout of newspapers (columns, fonts, illustrations), and the need for fine-tuned customization for a single source (Egemen Qazaqstan), a rule-based approach proves to be the most rational and effective solution for automatic extraction of structured tags from Kazakh-language PDF documents.

Based on this, we formulate the following research question: how can rule-based methods be adapted for the automatic extraction of semantically significant tags from Kazakh-language newspaper PDFs, considering their complex visual structure and formatting features?

The goal of this study is to develop and implement a hybrid pipeline based on the integration of rule-based tools – pdfminer.six, PyMuPDF, and pdfplumber – aimed at extracting structured elements (url, title, author, date, abstract, text, journal, category) from PDF versions of Kazakh-language newspaper articles.

Why were these particular tools chosen? To justify this, we refer to the study [15], where the authors conducted a comprehensive comparative analysis of ten popular PDF parsing tools across six types of documents: financial reports, manuals, scientific articles, laws and regulations, patents, and government tenders. The results showed that rule-based tools such as pdfminer.six [16], PyMuPDF [17], and pdfplumber [18] perform well in extracting text from standard documents, while deep learning models such as Nougat [19] and Table Transformer (TATR) [20] significantly outperform them in more complex cases, especially when dealing with mathematical formulas or non-standard tables.

The structure of this paper is as follows: Section II presents the experimental algorithm and data extraction methods from the PDF corpus, including a detailed description of the rule-based pipeline; Section III provides quantitative and qualitative results of the selected parsers; Section IV discusses their strengths and weaknesses in the context of Kazakh-language newspapers; Section V concludes the study and outlines directions for further development of the hybrid solution.

The outcome of this work is the construction of a machine-readable JSON-format corpus suitable for subsequent training of large language models and the development of an intelligent assistant in the field of Kazakh-language data journalism.

2. Materials and methods

For the experimental study, a corpus of PDF documents from the Egemen Qazaqstan newspaper was used. The processing was carried out using the tools PyMuPDF, pdfminer.six, and pdfplumber, in accordance with established computational standards. Based on empirical observations, a set of heuristic rules was developed for the automatic extraction of semantically meaningful tags such as title, author, date, abstract, text, category, and journal.

2.1. Description of dataset

For the purposes of this study, a specialized corpus was compiled, consisting of all PDF issues of the Egemen Qazaqstan newspaper from January 2017 to March 2025 inclusive. A total of 2,140 PDF documents were collected, each representing a full issue of the national newspaper.

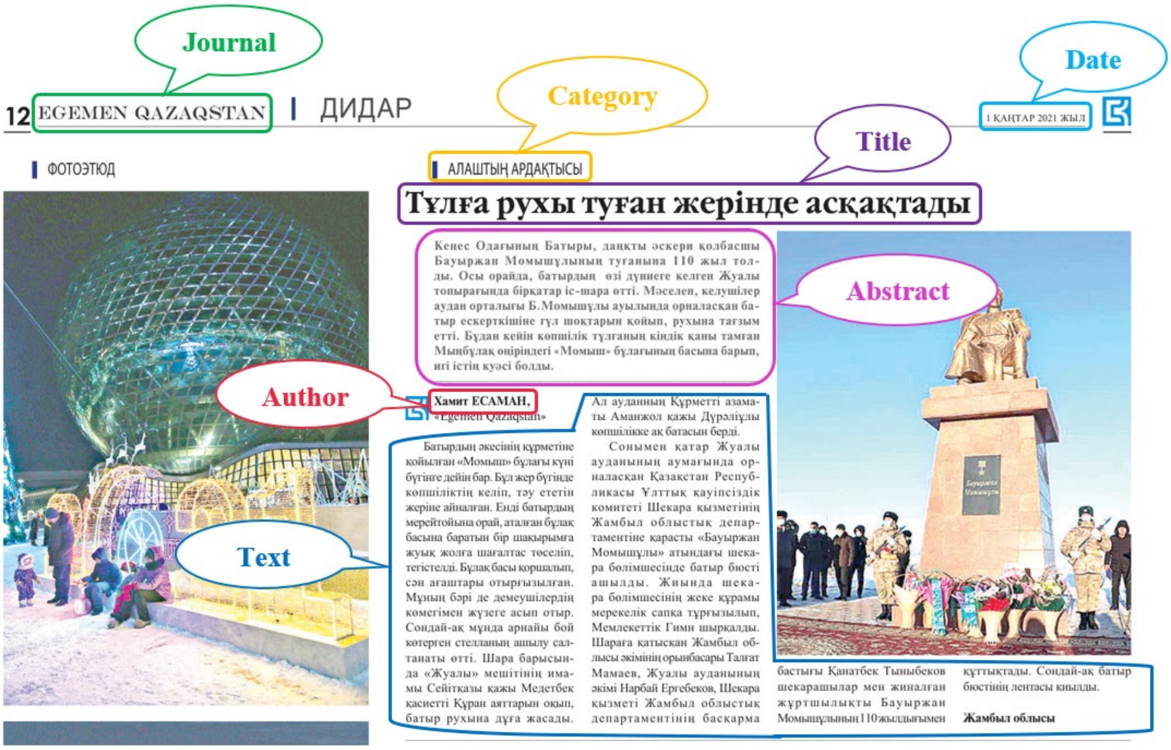
Each PDF file contains an average of 16-18 pages, formatted in a multi-column format with illustrations and multilingual inserts. On average, there are about 4.6 full articles per page, so the total number of text units in the corpus is 159,135 articles. These materials cover a wide range of topics – with a total of 208 unique categories. The primary categories include: Economics, Politics, Society, Culture, Education, History, Literature, Religion, Health, Spirituality, Regions, Agriculture, Technology, and Sports.

2.2. Rule-based methodology

This section outlines the rule-based methodology developed for the automatic extraction of semantically meaningful tags and metadata from the PDF versions of the collected newspapers. During the preprocessing stage, text cleaning and normalization methods were applied, including removal of OCR artifacts, merging of hyphenated line breaks, and Unicode normalization.

The core of the approach is a set of heuristic rules that leverage visual, lexical, and structural features to identify key tags such as: title, author, date, category, text, abstract, and journal. Figure 1 visually highlights all the main components corresponding to the target tags, which are subject to automatic extraction using the rule-based method.

The heuristic rules for extracting each tag were manually developed based on the analysis of the structure and visual characteristics of the PDF documents. All applied heuristics and filters are summarized in Table 1. The proposed rule set is tailored for the automatic annotation of newspapers with column-based layout and can serve as a foundation for building an annotated corpus or for subsequent training of machine learning models.



12 EGEMEN QAZAQSTAN | ДИДАР

1 ҚАҢТАР 2021 ЖЫЛ

ФОТОЭТЮД

АЛАШТЫҢ АРДАҚТЫСЫ

Тұлға рухы туған жерінде асқақтады

Кеңес Одағының Батыры, даңқты әскери қолбасшы Бауыржан Момышұлының туғанына 110 жыл толды. Осы орайда, батырдың өзі дүниеге келген Жуалы топырағында бірқатар іс-шара өтті. Мәселен, келушілер аудан орталығы Б.Момышұлы ауылында орналасқан батыр ескерткішіне гүл шоқтарын қойып, рухына тағзым етті. Бұдан кейін көпшілік тұлғаның кіндік қаны тамған Мыңбұлақ өңіріндегі «Момыш» бұлағының басына барып, игі істің куәсі болды.

Хамит ЕСАМАН,
«Egemen Qazaqstan»

Батырдың әкесінің құрметіне қойылған «Момыш» бұлағы күні бүгінге дейін бар. Бұл жер бүгінде көпшіліктің келіп, тәу ететін жеріне айналған. Енді батырдың мерейтойына орай, аталған бұлақ басына баратын бір шақырымға жуық жолға шағалтас төселіп, тегістелді. Бұлақ басы қоршалып, сән ағаштары отырғызылған. Мұның бәрі де демеушілердің көмегімен жүзеге асып отыр. Сондай-ақ мұнда арнайы бой көтерген стелланың ашылу салтанаты өтті. Шара барысында «Жуалы» мешітінің иманы Сейітқазы қажы Медетбек қасиетті Құран аяттарын оқып, батыр рухына дұға жасады. Ал ауданның Құрметті азаматы Аманжол қажы Дүрәліұлы көпшілікке ақ батасын берді.

Сонымен қатар Жуалы ауданының аумағында орналасқан Қазақстан Республикасы Ұлттық қауіпсіздік комитеті Шекара қызметінің Жамбыл облыстық департаментіне қарасты «Бауыржан Момышұлы» атындағы шекара бөлімшесінде батыр бюсті ашылды. Жиында шекара бөлімшесінің жеке құрамы мерекелік сапқа тұрғызылып, Мемлекеттік Гимн шырқалды. Шараға қатысқан Жамбыл облысы әкімінің орынбасары Талғат Мамаев, Жуалы ауданының әкімі Нарбай Ергебеков, Шекара қызметі Жамбыл облыстық департаментінің басқарма бастығы Қанатбек Тыныбеков шекарашылар мен жиналған жұртшылықты Бауыржан Момышұлының 110 жылдығымен құттықтады. Сондай-ақ батыр бюстінің лентасы қиылды.

Жамбыл облысы

Figure 1 – Example of a newspaper article from the PDF issue
of Egemen Qazaqstan dated January 1, 2021, featuring column-based layout.

Table 1 – Heuristic extraction rules for newspaper tags.

№	Tag	Rules	Rules description
Rule 1	title	if 10 < len(text) < 180 and block_index == 0 and span["size"] > 15	The title is identified as the first text block on the page that contains between 10 and 180 characters and is printed in a font size larger than 15 pt.
Rule 2	author	if re.search(r"[А-ЯӘӨҮҰІ][а-яәөүұі]+ [А-ЯӘӨҮҰІ][а-яәөүұі]+", text)	The author is identified by the presence of a full name in the format "First Last," starting with capital letters and matching a predefined regular expression.
Rule 3	date	if re.search(r"\d{1,2} \D+ 20\d{2}", text)	The date is extracted as a string containing a day, month name, and four-digit year (e.g., "15 шілде 2023"), conforming to the "day month year" pattern.
Rule 4	category	if any(keyword in text.upper() for keyword in ["Экономика", "Саясат", "Қоғам", …, "Спорт"] and span["size"] < 11)	The category is recognized by the presence of an uppercase keyword from a predefined list of 208 unique topics.
Rule 5	text	cleaned_text = unicodedata.normalize('NFC', re.sub(r"\s+", " ", re.sub(r"(\w+)-\s*\n\s*(\w+)", r"\1\2", re.sub(r"cid:\d+", "", text.replace('\n', ' ').replace('\r', ''))))).strip()) if isinstance(text, str) else ""	The article text is merged from multiple columns into a single string, with hyphenated line breaks removed, extra spaces and line breaks cleaned, and Unicode normalization applied for text standardization.
Rule 6	abstract	abstract = text.split(".")[0] if "." in text else text[:200]	The abstract is extracted as the first sentence (up to the first period); if no period is found, the first 200 characters are used.
Rule 7	journal	if block_index < 3 and span["size"] > 10 and text.isupper() and len(text.split()) <= 4:	The journal name is identified as an uppercase string located in one of the top three text blocks on the page, consisting of no more than four words and printed in a font larger than 10 points.

The rules presented in Table 1 will be further integrated into a hybrid pipeline that leverages PyMuPDF for analyzing the visual structure of the document, pdfminer.six for extracting the main text, and pdfplumber for refining block positions when necessary.

2.2.1 Pipeline architecture

Before designing the architecture of the hybrid pipeline, a comparative evaluation of three popular Python libraries – PyMuPDF, pdfplumber, and pdfminer.six – was conducted to assess their effectiveness in extracting various tags from newspaper-style PDF documents (Table 2).

The results indicate that none of the libraries alone ensures high accuracy across all tag categories: for instance, pdfminer.six performs better at extracting main text and abstracts, PyMuPDF shows the best performance for title and date tags, while pdfplumber demonstrates low robustness when dealing with the multi-layered structure of newspaper layouts. These differences stem from the internal architecture of the libraries: PyMuPDF relies on visual block positioning, pdfminer.six focuses on linear text extraction, and pdfplumber applies character-level parsing.

Based on this comparative analysis, a well-grounded decision was made to integrate all three tools into a single rule-based pipeline, leveraging the strengths of each library to achieve optimal accuracy for individual tag extraction tasks. The proposed pipeline implements a rule-based method for automatic extraction of semantically labeled tags from Kazakh-language newspaper PDF documents (see Figure 2).

Table 2 – Comparison of Tag Extraction Accuracy Using PyMuPDF, pdfplumber, and pdfminer.six.

Tag	PyMuPDF (%)	Pdfplumber (%)	pdfminer.six (%)
Title	90	50	60
Author	70	30	60
Abstract	60	40	80
Text	80	60	90
Date	90	40	40
Category	75	30	60
Journal	80	50	50

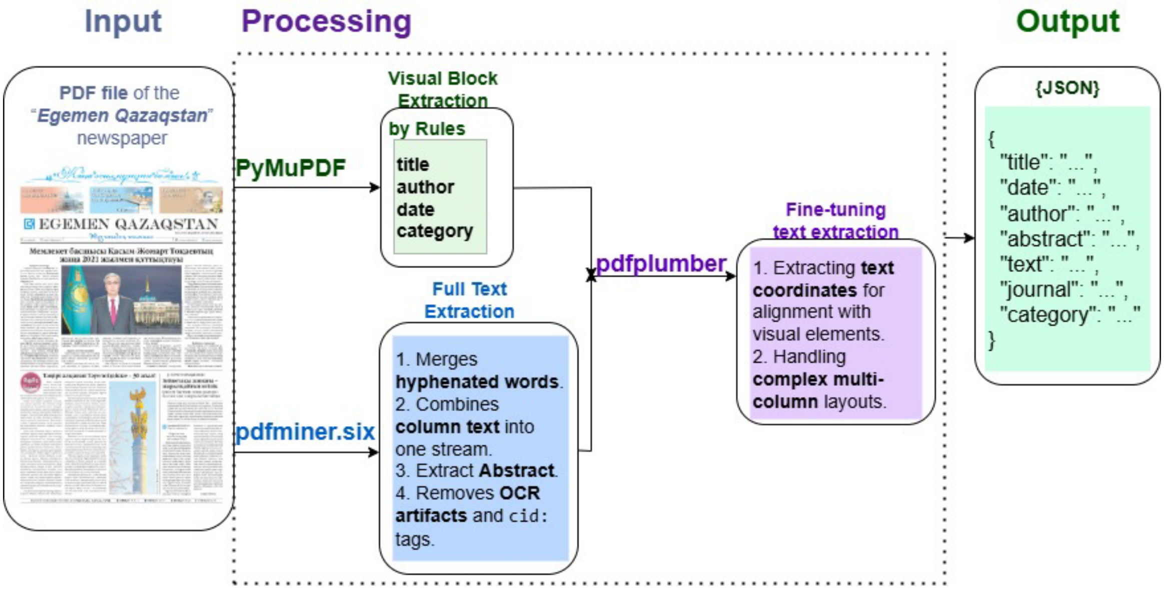

Input
Processing
Output
PDF file of the "Egemen Qazaqstan" newspaper
EGEMEN QAZAQSTAN
PyMuPDF
pdfminer.six
Visual Block Extraction by Rules
title
author
date
category
Full Text Extraction
1. Merges hyphenated words.
2. Combines column text into one stream.
3. Extract Abstract.
4. Removes OCR artifacts and cid: tags.
pdfplumber
Fine-tuning text extraction
1. Extracting text coordinates for alignment with visual elements.
2. Handling complex multi-column layouts.
{JSON}
{
"title": "...",
"date": "...",
"author": "...",
"abstract": "...",
"text": "...",
"journal": "...",
"category": "..."
}

Figure 2 – Hybrid PDF Tagging Architecture using PyMuPDF, pdfminer.six and pdfplumber.

The input to the system is a PDF file representing a print issue of the Egemen Qazaqstan newspaper, featuring a multi-column layout with both textual and graphical elements. The first stage employs the PyMuPDF library to extract visual text blocks by analyzing coordinates, font sizes, and spatial positioning. Based on a set of heuristics, this module identifies and extracts key tags such as title, author, date, and category. To retrieve the full textual content of the articles, the pipeline utilizes pdfminer.six, which disregards visual layout constraints and merges columnar content into a unified text string. At this stage, hyphenated word breaks are resolved, Unicode normalization is applied, and OCR artifacts (e.g., “cid:123”) are removed.

The resulting content is then used to generate the abstract and to extract the journal name based on the top blocks of the page. Additionally, pdfplumber is used for geometry refinement, coordinate verification, and correction of extracted tags in cases with complex visual layouts. This enhances tag-level precision through improved structural parsing. The output is a machine-readable JSON structure containing all extracted tags (url, title, author, date, abstract, text, journal, category), suitable for knowledge base enrichment, language model training, and intelligent Kazakh-language text analysis systems.

2.2.2 Algorithm

The developed algorithm formalizes this pipeline into a sequential metadata extraction procedure: Rules 1–4 are applied to visual blocks via PyMuPDF, followed by Rules 5–6 for text and abstract generation using pdfminer.six, and Rule 7 for journal identification. Pdfplumber serves as an auxiliary module for final verification and correction. This hybrid rule-based pipeline demonstrates the effective integration of three PDF processing tools, ensuring automated and interpretable extraction of textual information from Kazakh-language newspaper documents.

Algorithm: Rule-based Metadata Extraction from PDF Newspapers

```
Input: PDF file D of Egemen Qazaqstan newspaper
Output: JSON object J with fields: url, title, author, date, abstract, text, journal, category
function EXTRACT_METADATA(D)
    V ← extract_visual_blocks(D) using PyMuPDF
    INITIALIZE title, author, date, category
    for each text span T in V do
        T ← clean_and_normalize_text(T)
        if title == "" and apply Rule 1 on T then
            title ← T
        else if author == "" and apply Rule 2 on T then
            author ← T
        else if date == "" and apply Rule 3 on T then
            date ← T
        else if category == "" and apply Rule 4 on T then
            category ← T
    end for
    text_raw ← extract_text_pdfminer(D)
    text ← normalize_hyphens_and_columns(text_raw) ⟵ apply Rule 5
    abstract ← extract_abstract(text) ⟵ apply Rule 6
    for each top span T in first 3 blocks of V do
        if journal == "" and apply Rule 7 on T then
            journal ← T
    end for
    J ← {
        "title": title,
        "author": author,
        "date": date,
        "abstract": abstract,
        "text": text,
        "journal": journal,
        "category": category
    }
    J ← refine_with_pdfplumber(J, D)
    return J
end function
```

2.2.3 Algorithmic complexity

Let P denote the number of pages, B=|V| the number of visual blocks, S the number of text spans, C the total number of characters, and R=7 the number of heuristic rules. Enumerating blocks across all pages is O(B), while establishing per-page reading order can require O(B log B) in the worst case. Cleaning and normalizing the text for all spans is linear in the total character count, O(C). The single pass that applies Rules 1–4 examines up to S spans until all four fields are filled, and per-span costs are constant for metadata checks tied to span/block attributes (title), O(|T|) for non-pathological regular expressions (author, date), and O(|T|) for uppercasing combined with average O(1) hash-set membership over a fixed 208-label category, which aggregates to O(S+C) in the worst case. Full-text extraction followed by column stitching, de-hyphenation, and whitespace repair is also linear, O(C). Abstract derivation is O(C) in the worst case but typically sublinear due to early termination at the first sentence. Journal identification is restricted by design to the first three blocks and therefore runs in O(1) on average, although it can degrade to O(B) if fallbacks expand the search region. The optional geometry refinement applies local checks and can be bounded by O(B).

Consequently, the algorithm runs in $O(B\log B+S+C)$ per issue in the worst case, which in practice tends toward near-linear O(B+C) thanks to page-zone filtering and early stopping (e.g., title and journal found early; abstract stops at the first period). Viewed per tag, title selection is O(1) per span via metadata with an O(B) worst case if page scans are required; author and date are O(C) due to regex evaluation; category is O(C) using the constant-size lexicon; text and abstract are O(C); and journal is O(1) on average with an O(B) worst case.

2.3. Evaluation Methodology

Evaluation of the performance of rule-based information extraction (IE) systems, unlike machine learning models, does not require a training phase; however, it demands meticulous verification of the correctness of the extracted data. The most common approaches include manual inspection (see Formulas 1–3), semantic similarity measurement between extracted and reference text fragments (see Formula 4), and expert evaluation using a checklist (see Formulas 5–7). These methods are widely applied in scenarios where labeled datasets are unavailable.

Table 3 presents three quality assessment methods for evaluating the performance of rule-based information extraction systems.

Table 3 – Methods and formulas for evaluating the quality of rule-based information extraction (IE) systems.

№	Evaluation method	Formula	Description
1	Manual inspection	For each entity *E={w₁,w₂,...,wₖ},* where *wᵢ* are whitespace-separated words, count $\|E_{true} \cap E_{pred}\|$, then compute $Precision = \frac{\|E_{true} \cap E_{pred}\|}{E_{pred}}$ (1), $Recall = \frac{\|E_{true} \cap E_{pred}\|}{E_{true}}$ (2), $holistic\ F1 = 2 * \frac{Precision*Recall}{Precision+Recall}$ (3)	It is used as a baseline quality metric for rule-based approaches in the absence of training data and allows for direct comparison with the gold-standard annotation [21].
2	Textual semantic similarity (TSS)	Modified cosine similarity measure: $Sim(T_1,T_2) = \frac{\overrightarrow{T_1} * \overrightarrow{T_2}}{\|\|\overrightarrow{T_1}\|\| * \|\|\overrightarrow{T_2}\|\|}$ (4), where $T_1, T_2 \in R^n$ – vector representations of texts, $Sim(T_1,T_2) \in [0,1]$, tags are considered correct if *Sim≥0.85*	It evaluates the degree of semantic similarity between the reference and extracted text based on vector representations. A modified cosine similarity measure with a threshold ≥ 0.85 is applied [22].
3	Expert checklist evaluation	1. Exploration: $H_e = \frac{T_e - U_e}{T_e}$ (5), where H_e – homogeneity of vocabulary, T_e – total number of words in the selected entity texts, T_e – the number of unique words in the same texts. 2. Term frequency analysis: $TF.IDF_{t,e} = \frac{f_{t,e}}{\sum_{t' \in e} f_{t',e}} * log\left(\frac{\|E\|}{\|E_t\|}\right)$ (6),	The REST-tool is designed to optimize resources in information extraction (IE) tasks by analyzing whether rule-based approaches can be effectively applied to each entity or whether machine learning (ML) methods should be used, involving multiple expert reviewers [23].

Continuation of the table

№	Evaluation method	Formula	Description
		where $f_{t,e}$ – the number of occurrences of term t in the selections of entity e, |E| – total number of entities, $\|E_t\|$ – the number of entities containing the term t. 3. Expert Evaluation: $Precision = \frac{TP}{TP+FP}$ (7), where TP – true positive, FP – false positive. Then do checklist – assessment of the applicability of the rules: *TH: Text Highlights ≥ 25%,* *LH: Linguistic Homogeneity ≥ 10%,* *ER: Entity Recall ≥ 75%,* *EP: Entity Precision ≥ 75%,* If all 4 criteria are met, the rules apply, otherwise – ML.	

In [21], the authors present a system for structured information extraction from scientific texts that combines manual annotation with automated entity extraction. They introduce the *holistic F1* metric, which accounts for both exact matches and semantic similarity between extracted and gold-standard fragments. This methodology emphasizes comprehensive expert evaluation, making it especially applicable to rule-based systems without training data.

The TSS method evaluates the degree of semantic alignment between reference and extracted text using vector representations, applying a modified cosine similarity threshold of ≥ 0.85 [22]. The study in [22] demonstrates that transformer models pre-trained and fine-tuned on clinical data achieve a high correlation with expert judgments (Pearson r ≈ 0.89–0.91), confirming the effectiveness of this approach for measuring semantic similarity in clinical texts.

In [23], the authors developed a rule-based Evaluation and Support Tool (REST) to optimize resource allocation in information extraction tasks. Their workflow began with fully manual expert annotations, followed by the introduction of metrics to assess rule adequacy, and the use of expert checklists to determine whether rule-based methods should be applied to specific entity types.

After reviewing these three effective evaluation metrics (see Table 3), the authors of this study selected two – Precision and TSS. We compute TSS with the Sentence-Transformers library [24] using the multilingual model sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 [25]. The model maps sentences to 384-dimensional embeddings, then it is used the model's Hugging Face AutoTokenizer (BERT-style) with max_seq_length = 128 and apply means pooling over token embeddings to obtain sentence vectors. Similarity is measured by cosine similarity between the reference and extracted text. We count a pair as semantically equivalent when TSS ≥ 0.85. This threshold was selected on a held-out, manually labeled set by maximizing F1 and was robust in the 0.80–0.88 range. The model is multilingual (≈50 languages, including Kazakh), so no additional language adaptation was required; we only apply Unicode NFC normalization prior to encoding [24, 25]. We then integrated these two metrics into a unified Holistic Precision formula for a more comprehensive assessment of the extracted tags (see formulas 8 and 9):

$$holistic\ Precision_{tag} = \\ = \alpha * Precision_{tag} + (1-\alpha) * TSS_{tag} \quad (8)$$

$$\alpha \in [0, 1] \quad (9)$$

where $Precision_{tag}$ – proportion of exactly matching extracted values, TSS_{tag} – average cosine similarity between reference and extracted text, α – weighting coefficient (in our case 0.5 for equal contribution of both metrics).

Both metrics are normalized to [0,1] and capture complementary error modes: Precision penalizes string-level deviations – critical for structured fields (e.g., date, category, journal) – consistent with standard IE/NER evaluation where scoring is performed at the entity level [26]. TSS rewards meaning-preserving paraphrases – critical for free-text fields – so that variants of title, abstract, or full text that differ lexically yet preserve meaning are not unduly penalized. These practices reflect long-standing IE/NER evaluation conventions and explain why the two signals are complementary.

In the absence of domain-specific cost asymmetries, we set α=0.5 as a neutral prior for three reasons: (i) it treats the two components symmetrically; (ii) it avoids degeneration to a single metric at the extremes $\alpha \in [0,1]$; and (iii) it is invariant under affine rescalings of either component. This choice aligns with the classical balanced setting in IR/IE, where the Precision measure corresponds to equal weighting ($\beta = 1 \Rightarrow \alpha = \frac{1}{2}$), and with the REST rationale to balance exact and semantic signals in rule-based IE evaluation [23].

Holistic Precision provides a more flexible and realistic estimate of extraction quality, particularly useful when analyzing rule-based systems where semantically correct but not form-identical extractions are possible.

2.4. Rights, licensing, and intended use

Copyright in the original newspaper content remains with the rights holders. No full article texts or images are redistributed by this work.

Published metadata is licensed under a CC BY-NC 4.0 (Creative Commons Attribution–NonCommercial 4.0) license. Any reuse must be related to this article/dataset, remain non-commercial, and comply with applicable law and publisher policies. We honor copyright holder takedown requests; requests must include the publication date, author, and title so that relevant records can be removed.

The dataset and code are intended for academic research and benchmarking of information-extraction methods for low-resource languages and to support development of a non-commercial AI assistant for journalists. Use of the original article texts for model training lies outside the scope of this release and requires appropriate permissions.

3. Results

As part of the experimental evaluation of the proposed rule-based approach, we manually annotated a sample of 113 articles obtained from three issues of the Egemen Qazaqstan newspaper. The annotation covered key semantic tags, including title, author, abstract, text, date, journal, and category. To assess the extraction quality, we applied our proposed integrated metric – holistic Precision (see Formula 8) – which takes into account not only the exact match (Precision), but also the TSS between the reference and extracted fragments. This dataset was used as a test sample for the experiment. All training and testing files are available in the GitHub repository at https://github.com/AsselOspan/Rule_based_PDF2JSON. Figure 3 presents an example of the original print layout of the newspaper and the corresponding gold-standard JSON file created through manual annotation.

To evaluate the quality of data extraction, a comparative analysis was conducted between the results obtained using the rule-based parsing approach and manually annotated gold-standard files. The comparison was performed across each of the seven key tags: title, author, abstract, text, date, journal, and category. During the experiment, the values of Precision, TSS (Textual Semantic Similarity), and the final metric Holistic Precision – which integrates both indicators – were calculated. The results of the experiment based on three selected PDF files are presented in Tables 4–7.

EGEMEN QAZAQSTAN

Атажұртқа алғашқы көш

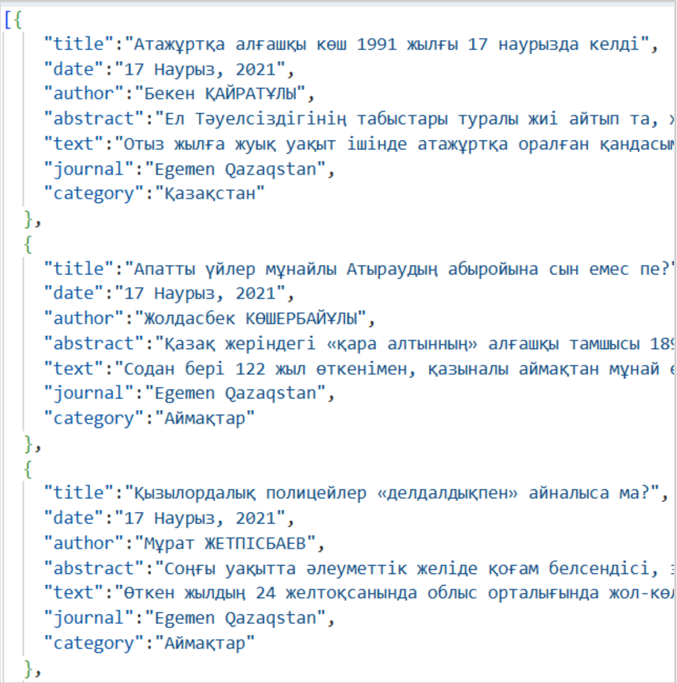
1991 жылғы 17 наурызда келді

Апатты үйлер

мұнайлы Атыраудың абыройына сын емес пе?

Қызылордалық полицейлер «делдалдықпен» айналыса ма?

(a)



```
[{
    "title":"Атажұртқа алғашқы көш 1991 жылғы 17 наурызда келді",
    "date":"17 Наурыз, 2021",
    "author":"Бекен ҚАЙРАТҰЛЫ",
    "abstract":"Ел Тәуелсіздігінің табыстары туралы жиі айтып та,
    "text":"Отыз жылға жуық уақыт ішінде атажұртқа оралған қандасы
    "journal":"Egemen Qazaqstan",
    "category":"Қазақстан"
  },
  {
    "title":"Апатты үйлер мұнайлы Атыраудың абыройына сын емес пе?"
    "date":"17 Наурыз, 2021",
    "author":"Жолдасбек КӨШЕРБАЙҰЛЫ",
    "abstract":"Қазақ жеріндегі «қара алтынның» алғашқы тамшысы 18
    "text":"Содан бері 122 жыл өткенімен, қазыналы аймақтан мұнай
    "journal":"Egemen Qazaqstan",
    "category":"Аймақтар"
  },
  {
    "title":"Қызылордалық полицейлер «делдалдықпен» айналыса ма?",
    "date":"17 Наурыз, 2021",
    "author":"Мұрат ЖЕТПІСБАЕВ",
    "abstract":"Соңғы уақытта әлеуметтік желіде қоғам белсендісі,
    "text":"Өткен жылдың 24 желтоқсанында облыс орталығында жол-кө
    "journal":"Egemen Qazaqstan",
    "category":"Аймақтар"
  },
```

(b)

Figure 3 – PDF presentation of the Egemen Qazaqstan newspaper in JSON format: (a) original printed page layout; (b) corresponding JSON representation.

Table 4 – Results of the experiment for the newspaper "Egemen Qazaqstan" from January 23, 2020 on three quality indicators.

Tag	Precision	TSS	Holistic Precision
title	0,68	0,87	0,77
author	0,74	0,87	0,80
date	1,00	1,00	1,00
abstract	1,00	1,00	1,00
text	0,88	0,93	0,91
journal	1,00	1,00	1,00
category	1,00	1,00	1,00

Table 5 – Results of the experiment for the newspaper “Egemen Qazaqstan” from January 1, 2021 on three quality indicators

Tag	Precision	TSS	Holistic Precision
title	0,80	0,93	0,86
author	1,00	1,00	1,00
date	1,00	1,00	1,00
abstract	0,98	0,96	0,97
text	0,94	0,93	0,93
journal	1,00	1,00	1,00
category	1,00	1,00	1,00

Table 6 – Results of the experiment for the newspaper "Egemen Qazaqstan" from May 20, 2021 on three quality indicators

Tag	Precision	TSS	Holistic Precision
title	0,87	0,93	0,90
author	1,00	1,00	1,00
date	1,00	1,00	1,00
abstract	0,97	0,96	0,96
text	0,87	0,92	0,90
journal	1,00	1,00	1,00
category	1,00	1,00	1,00

Table 7 – Average results of the experiment

Tag	Precision	TSS	Holistic Precision
title	0,78	0,91	0,85
author	0,91	0,96	0,93
date	1,00	1,00	1,00
abstract	0,98	0,97	0,98
text	0,90	0,93	0,91
journal	1,00	1,00	1,00
category	1,00	1,00	1,00
Average	**0,94**	**0,97**	**0,95**

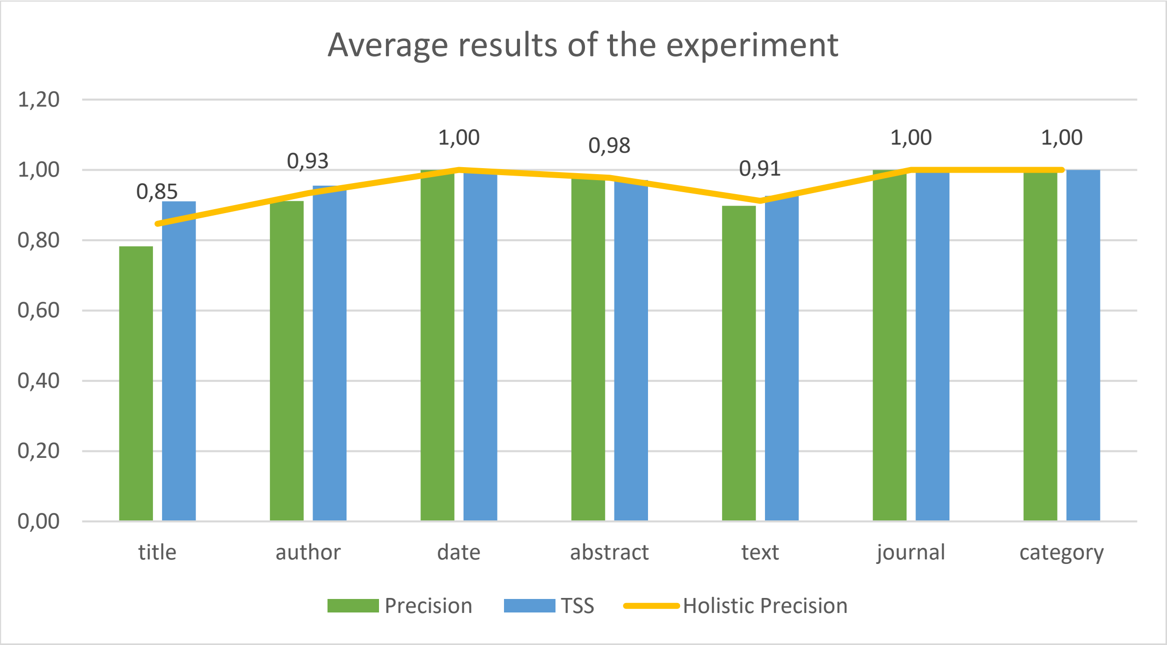


Figure 4 – Graph of averaged results of the experiment on assessing the quality of semantic tag extraction using the rule-based method.

As shown in Figure 4, the TSS and Holistic Precision metrics for most tags fall within the 0.95–1.00 range, indicating high accuracy of the proposed rule-based approach. Slightly lower Precision values were observed for the title tag, which can be attributed to the variability of headlines and minor stylistic or syntactic differences.

The results demonstrate excellent extraction accuracy for formalized tags such as date, journal, and category, where the integrated Holistic Precision metric reached a perfect score of 1.000. High semantic similarity was also observed for less formal fields such as title, abstract, and text, with Holistic Precision values ranging from 0.85 to 0.98.

Following the experimental evaluation and validation of the method's effectiveness, the approach was scaled to the entire PDF corpus of the Egemen Qazaqstan newspaper from 2017 to March 2025. As a result of automated extraction and structuring, 2,140 PDF files were processed, and 159,135 articles were successfully converted into JSON format (see Table 8).

Table 8 – Number of Egemen Qazaqstan PDF issues from 2017 to March 2025 and the number of articles successfully converted to JSON format.

Year	Number of PDF newspapers	Number of articles in JSON
2017	272	17 147
2018	285	16 000
2019	265	22 647
2020	280	21 630
2021	301	22 365
2022	337	28 855
2024	340	26 491
2025	60	4 000
Total number	**2 140**	**159 135**

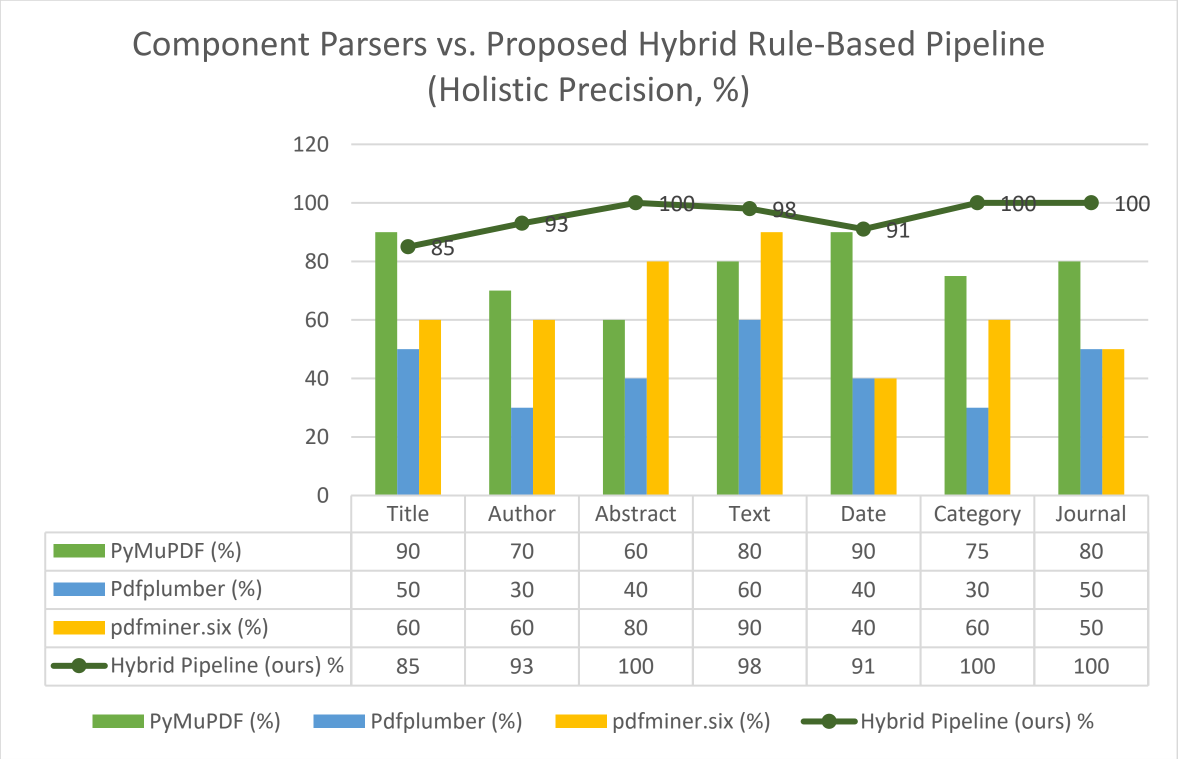
The resulting collection of structured materials will serve as a knowledge base for the development of an intelligent AI assistant tailored to data journalism tasks.

4. Discussion

The goal of this study was to determine how rule-based methods can be adapted for the automatic extraction of semantically meaningful tags from PDF issues of Kazakh-language newspapers–despite their complex visual structure and formatting–and whether high-quality results can be achieved. Although there exist widely used PDF parsing methods [16–20], our experiments showed that they underperformed on Kazakh texts and struggled with multi-column layouts. We therefore targeted these challenges specifically, since building a knowledge base for downstream NLP inevitably involves such complex files. The methods we designed are detailed in the Methodology section, with custom heuristic rules as the key to improving extraction quality.

We evaluated extraction quality with three complementary metrics – Precision, TSS, and their integrated Holistic Precision – each addressing a distinct need in Kazakh-language processing. Precision quantifies exact matches between extracted and reference tags and is crucial for highly formalized elements (e.g., date, category, journal), where correctness can be measured without contextual interpretation. TSS is especially important for Kazakh's rich morphology and agglutination: if strict Precision alone were used, many partially matching yet semantically valid outputs (e.g., for title, abstract and text) would be incorrectly deemed wrong. Holistic Precision combines both signals to provide a balanced view of system effectiveness, which is particularly relevant for rule-based systems in low-resource settings where large annotated datasets are unavailable.

A comparative analysis of tag-level results across the three component parsers – PyMuPDF, pdfplumber, and pdfminer.six – shows that no single component dominates all tags. PyMuPDF performs best for title, date, category, and journal thanks to its use of layout cues, while pdfminer.six leads on abstract and text due to its linear text analysis that is more resilient to multi-level layouts; pdfplumber is the weakest on most tags. Our Hybrid Rule-Based Pipeline, which integrates all three modules, achieves the highest overall performance, with Holistic Precision (α = 0.5) reaching 91–100% on six of the seven tags (and 85% on title). Figure 5 summarizes these outcomes by reporting per-tag Holistic Precision for the three single-component variants (bars) and the proposed hybrid pipeline (line).



	Title	Author	Abstract	Text	Date	Category	Journal
🟩 PyMuPDF (%)	90	70	60	80	90	75	80
🟦 Pdfplumber (%)	50	30	40	60	40	30	50
🟨 pdfminer.six (%)	60	60	80	90	40	60	50
●― Hybrid Pipeline (ours) %	85	93	100	98	91	100	100

Figure 5 – Results reporting per-tag Holistic Precision for the three parsers (bars) – PyMuPDF, pdfplumber, and pdfminer.six – and the proposed hybrid rule-based pipeline (line).

These results align with prior findings [21, 23] on the advantages of rule-based approaches for documents with predictable visual structure in low-resource settings. The Egemen Qazaqstan newspaper, with its consistent multi-column layout and recurring placement of key elements, is a suitable testbed for such heuristics.

Nonetheless, potential bias introduced by heuristic rules should be considered. Because the pipeline relies on hand-crafted heuristics over visual, lexical, and structural cues (Rules 1–7; Fig. 1; Table 1), systematic biases may arise. The title rule (first block, large font, 10–180 chars) can over-split short briefs or miss multi-line/small-caps headlines; the author regex (two capitalized tokens) may under-detect three-part names, initials, hyphenated surnames, or mixed-script forms; the date pattern (dd Month yyyy) can miss numeric formats or variant month spellings; the category rule (uppercase keyword list of 208 topics, small font) may bias counts toward frequent uppercase rubrics and under-represent rare/lowercase labels; the text merger (de-hyphenation, column stitching, NFC) can over-join captions or sidebars; the abstract heuristic (first sentence / first 200 chars) may be non-representative for quotes or very short items; and the journal rule (uppercase in top blocks) can confuse mastheads or slogans with the outlet name. Multilingual inserts and dense layouts amplify these risks. In evaluation terms, such effects tend to increase Precision errors on structured tags (boundary/format mismatches), while TSS may mask minor string differences in free-text fields. We mitigate these risks via Unicode normalization (NFC), near-duplicate removal at the (title, date) level, lexicon expansion, and layout refinement with pdfplumber, but residual skew is possible across issues and years and should be considered when interpreting per-tag results.

Some limitations follow directly from these observations. For example, relatively lower Precision for the title tag (avg. 0.78) indicates sensitivity to stylistic and linguistic variation. Newspaper headlines often feature creative phrasing, non-standard punctuation, or unusual syntax, making fixed-template extraction challenging. Although the final Holistic Precision for title reaches 0.85 due to compensation via semantic similarity, this highlights opportunities for improvement – e.g., integrating lightweight ML components to capture more flexible patterns.

Scaling the method to the full Egemen Qazaqstan corpus (2017–March 2025) confirmed robustness and practicality: 2,140 issues processed and 159,135 articles converted to JSON demonstrate suitability for building large machine-readable corpora and knowledge bases. Such resources are vital for training Kazakh LLMs and developing AI assistant for data journalism.

Finally, several operational constraints remain. Manual rule design requires occasional maintenance as layout/editorial policies evolve. Despite combining three libraries (PyMuPDF, pdfminer.six, pdfplumber), throughput may lag behind optimized ML-based pipelines in large-scale deployments. The current version also lacks automatic segmentation for long multi-page articles and pre-processing for scanned PDFs with heavy OCR noise.

Future work includes expanding the tag set (e.g., subtitles, image captions), adding active learning to partially automate rule configuration, and developing a hybrid architecture that couples rule-based logic with transformer-based models for more flexible post-processing. Evaluating the pipeline on additional multilingual news corpora will further illuminate its generalizability.

In summary, the proposed rule-based method achieves high accuracy and strong interpretability on Kazakh-language newspapers, offering a reliable foundation for structured corpora, AI applications, and knowledge-base enrichment in low-resource settings.

5. Conclusions

In this study, a hybrid rule-based pipeline was developed and experimentally validated for the automatic extraction of semantically annotated tags from Kazakh-language newspaper PDF documents. The system combines the strengths of three libraries – PyMuPDF (visual layout analysis), pdfminer.six (linear text extraction), and pdfplumber (structure refinement) – and utilizes a set of manually crafted heuristic rules to identify key metadata fields: title, author, date, abstract, text, journal, and category.

Experimental evaluation on a sample of 113 manually annotated articles demonstrated high values for Precision, TSS, and the integrated Holistic Precision metric (averaging 0.95), confirming the reliability and accuracy of the proposed method. Particularly strong results were achieved for the date, journal, and category tags (1.00), as well as high semantic similarity for the title and abstract fields. Following this validation, the method was scaled to the entire Egemen Qazaqstan corpus from 2017 to March 2025, successfully extracting and structuring 159,135 articles from 2,140 PDF files.

The research goal was achieved: an effective and scalable pipeline was developed for structured data extraction from Kazakh-language PDFs. The study also provided a clear answer to the research question – rule-based methods, when adapted to the visual and linguistic characteristics of newspaper layouts, proved highly suitable for automatic tag extraction under conditions of limited annotated data.

The proposed approach lays a strong foundation for building a machine-readable Kazakh-language corpus, which is strategically important for training large language models (LLMs) and developing an AI assistant focused on data journalism tasks.

Acknowledgments

The authors express their sincere gratitude to the doctoral students of the Faculty of Journalism for their invaluable assistance in collecting and organizing the archival PDF documents of the Egemen Qazaqstan newspaper from 2017 to March 2025.

Funding

This research was funded by the Science Committee of the Ministry of Science and Higher

Education of the Republic of Kazakhstan grant number BR24993001 “Creation of a large language model (LLM) to maintain the implementation of Kazakh language and increase the technological progress”.

Author Contributions

Conceptualization, A.O. and T.S.; Methodology, A.O.; Software, A.M.; Validation, A.O., A.M. and K.A.; Formal Analysis, T.S.; Investigation, K.A.; Resources, K.A.; Data Curation, K.A.; Writing – Original Draft Preparation, A.O.; Writing – Review & Editing, T.S. and K.A.; Visualization, A.M.; Supervision, K.A.; Project Administration, T.S.; Funding Acquisition, T.S.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, et al., “Language models are few-shot learners,” arXiv preprint arXiv:2005.14165, May 2020. [Online]. Available: https://arxiv.org/abs/2005.14165

2. A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, et al., “PaLM: Scaling language modeling with pathways,” arXiv preprint arXiv:2204.02311, Apr. 2022. [Online]. Available: https://arxiv.org/abs/2204.02311

3. H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, et al., “LLaMA 2: Open foundation and fine-tuned chat models,” arXiv preprint arXiv:2307.09288, Jul. 2023. [Online]. Available: https://arxiv.org/abs/2307.09288

4. V. Barakhnin, M. Mansurova, I. Grigorieva, O. Kozhemyakina, and A. Ospan, “TableProcessor: The tool for the analysis and the interpretation of web tables to create the geo knowledge base of Kazakhstan,” in Artificial Intelligence in Models, Methods and Applications, AIES 2022, Studies in Systems, Decision and Control, vol. 457, Cham, Switzerland: Springer, 2023. [Online]. Available: https://doi.org/10.1007/978-3-031-22938-1_15

5. M. Mansurova, V. Barakhnin, A. Ospan, and R. Titkov, “Ontology-driven semantic analysis of tabular data: An iterative approach with advanced entity recognition,” Appl. Sci., vol. 13, no. 19, pp. 10918, 2023. https://doi.org/10.3390/app131910918

6. A. B. Nugumanova, K. S. Apayev, Y. M. Baiburin, M. Mansurova, and A. G. Ospan, “QURMA: A table extraction pipeline for knowledge base population,” J. Math., Mech. Comput. Sci., vol. 114, no. 2, 2022. https://doi.org/10.26577/JMMCS.2022.v114.i2.08

7. “Egemen Qazaqstan – Official Republican Socio-Political Newspaper of Kazakhstan.” Accessed: Jul. 21, 2025. [Online]. Available: https://egemen.kz/

8. A. Alamoudi, A. Alomari, S. Alwarthan, and A. Rahman, “A rule-based information extraction approach for extracting metadata from PDF books,” ICIC Express Lett., Part B: Appl., vol. 12, no. 2, pp. 121–132, Feb. 2021. [Online]. Available: https://www.researchgate.net/publication/347948003

9. E. Hetzner, “A simple method for citation metadata extraction using hidden Markov models,” in Proc. 8th ACM/IEEE-CS Joint Conf. Digital Libraries, JCDL '08, pp. 280–284, 2008. https://doi.org/10.1145/1378889.1378937

10.C. Yu, C. Zhang, and J. Wang, “Extracting body text from academic PDF documents for text mining,” arXiv preprint arXiv:2010.12647, Oct. 2020.

11.B. Shi, X. Bai, and C. Yao, “An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 11, pp. 2298–2304, Nov. 2016.

12.G. Kim, T. Hong, M. Yim, J. Nam, J. Park, and J. Yim, et al., “OCR-free document understanding transformer,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Cham, Switzerland: Springer, 2022, pp. 498–517.

13.M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, et al., “TrOCR: Transformer-based optical character recognition with pre-trained models,” in Proc. AAAI Conf. Artif. Intell., vol. 37, pp. 13094–13102, 2023.

14.B. Pfitzmann, C. Auer, M. Dolfi, A. S. Nassar, and P. Staar, “DocLayNet: A large human-annotated dataset for document-layout segmentation,” in Proc. 28th ACM SIGKDD Conf. Knowl. Discov. Data Mining, KDD '22, pp. 3743–3751, 2022. https://doi.org/10.1145/3534678.3539043

15.N. S. Adhikari and S. Agarwal, “A comparative study of PDF parsing tools across diverse document categories,” arXiv preprint arXiv:2410.09871, Oct. 2024. [Online]. Available: https://arxiv.org/abs/2410.09871

16.pdfminer.six, “PDF parser for Python.” [Online]. Available: https://github.com/pdfminer/pdfminer.six

17.PyMuPDF, “Python bindings for the MuPDF library.” [Online]. Available: https://pypi.org/project/PyMuPDF

18.pdfplumber, “Tool for extracting text, tables, and metadata from PDFs.” [Online]. Available: https://github.com/jsvine/pdfplumber

19.L. Blecher, G. Cucurull, T. Scialom, and R. Stojnic, “Nougat: Neural optical understanding for academic documents,” arXiv preprint arXiv:2308.13418, Aug. 2023. [Online]. Available: https://arxiv.org/abs/2308.13418

20.Microsoft, “Table Transformer (TATR): Transformer-based table detection model.” [Online]. Available: https://github.com/microsoft/table-transformer

21.J. Dagdelen, A. Dunn, S. Lee, et al., “Structured information extraction from scientific text with large language models,” Nat. Commun., vol. 15, p. 1418, 2024. https://doi.org/10.1038/s41467-024-45563-x

22.X. Yang, X. He, H. Zhang, Y. Ma, J. Bian, and Y. Wu, “Measurement of semantic textual similarity in clinical texts: Comparison of transformer-based models,” JMIR Med. Inf., vol. 8, no. 11, p. e19735, 2020. https://doi.org/10.2196/19735

23.G. Bazin, X. Tannier, F. Adda, A. Cohen, A. Redjdal, and E. Kempf, “Development of the user-friendly decision aid Rule-based Evaluation and Support Tool (REST) for optimizing the resources of an information extraction task,” arXiv preprint arXiv:2506.13177, Jun. 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2506.13177

24.N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” *Proceedings of EMNLP-IJCNLP*, 2019.

25.Sentence-Transformers. *Model card: sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2* (multilingual, 384-dim, BERT backbone, default max_seq_length 128). Hugging Face, accessed 22.08.2025.

26.Manning, C. D. (2010). *Information extraction & named entity recognition* (CS224N lecture slides). Stanford University. Available at: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1106/handouts/InfoExtract-cs224n-2010-1up.pdf

Information about authors

Assel Ospan is a senior lecturer at the Department of Artificial Intelligence and Big Data, al-Farabi Kazakh National University (Almaty, Kazakhstan, assel.ospan@kaznu.edu.kz). Her research focuses on the development of large language models for the Kazakh language, intelligent information extraction, and knowledge base construction. She actively participates in national AI research initiatives and has authored several publications on NLP and data journalism. ORCID iD: 0000-0002-1860-6997.

Kanat Auyesbay is the Dean of the Faculty of Journalism at Al-Farabi Kazakh National University (Almaty, Kazakhstan, kanat.auyesbay@kaznu.edu.kz). He is a journalist-educator who bridges the fields of media and higher education. Dr. Auesbay holds a Candidate of Philological Sciences degree (equivalent to PhD) and has extensive experience in both media production and academic leadership. As a recipient of the Bolashak International Scholarship, Kanat Auesbay completed a research and teaching internship at the University of East Anglia, UK (Norwich, 2013–2014). He served as Chairman of the State Attestation Commission at the Faculty of Journalism and Political Science of L.N. Gumilyov Eurasian National University (2023–2024). Since 2018, he has been a corresponding member of the Kazakhstan Academy of Pedagogical Sciences and a member of the Educational-Methodical Association under the Republican Educational-Methodical Council (ROӘK) for Journalism and Information (2019–2021). He has also served on the expert commission for training specialists abroad under the Bolashak program and has supervised and reviewed numerous theses and doctoral dissertations in media studies. ORCID iD: 0009-0001-3529-9888

Talshyn Sarsembayeva is a a senior lecturer at the Department of Artificial Intelligence and Big Data, al-Farabi Kazakh National University (Almaty, Kazakhstan, talshyn.sagdatbek@kaznu.edu.kz). Her work focuses on the integration of artificial intelligence and data processing tools in journalistic practice. She has contributed to projects involving the structuring of large-scale media archives and the development of AI-assisted systems for Kazakh-language content. ORCID iD: 0000-0001-7668-2640.

Aman Mussa is a research assistant at the Department of Artificial Intelligence and Big Data, al-Farabi Kazakh National University (Almaty, Kazakhstan, mussa.aman0519@gmail.com). He is engaged in the development of rule-based and hybrid NLP pipelines, with a focus on Kazakh-language PDF processing. His work supports large-scale knowledge base generation for intelligent assistants in data journalism. ORCID iD: 0009-0001-9972-7677.

Submission received: 28 July, 2025.
Revised: 24 September, 2025.
Accepted: 24 September, 2025.

IRSTI 28.23.01 https://doi.org/10.26577/jpcsit2025337

Aizhan Nurzhanova[1], Miras Mussabek[2], Gokhan Ince[3], Mas Rina Mustaffa[4], Ainur Zhumadillayeva[1*]

[1]L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
[2]Astana IT University, Astana, Kazakhstan
[3]Istanbul Technical University, Istanbul, Turkey
[4]Universiti Putra Malaysia, Serdang, Malaysia
*e-mail: zhumadillayeva_ak@enu.kz

DETECTION OF MENTAL DISORDERS BASED ON THE ANALYSIS OF EMOTION, FACIAL EXPRESSIONS AND FACIAL MOVEMENTS IN A VIDEO STREAM

Abstract. Traditional emotion recognition systems typically use publicly available models without regard to differences in future emotions. In this paper, we find the possibility of violations based on the analysis of facial expressions and movements in a video stream taking into account features. A personalized approach to emotion recognition is proposed, implemented using machine and deep learning algorithms. The system uses the MediaPipe FaceMesh to extract 468 facial keypoints and analyze expressions associated with conditions such as anxiety, depression, post-traumatic stress disorder(PTSD), mania, or fatigue. Experiments have confirmed that personalized models provide higher accuracy compared to robustness. We propose a video-based framework for predicting mental disorders by analyzing temporal facial dynamics, micro-expression volatility, and asymmetry using RGB streams. Our system identifies disorder-specific biomarkers like delayed emotional reactivity and slow blinks, enabling real-time mental health detection.

Keywords: emotion recognition, video models, individual differences, personalized models, deep learning, affective computing.

1. Introduction

Emotions and facial expressions play an important role in human communication, influencing interpersonal interactions, cognitive processes, and decision-making. Automated emotion recognition systems have gained great importance in affective computing and human-computer interaction systems. However, traditional models generalize emotional expressions without taking into account individual differences. The variability of facial expressions, head movements, and micro-emotions requires the development of personalized recognition models that are sensitive to the characteristics of a particular person.

In light of the growing importance of human-computer interaction, research in the field of psycho-emotional state recognition is becoming especially relevant. Automatic emotion recognition attracts the attention of both researchers and developers, which makes this task significant for fundamental research, applied developments in the field of information technology and neural interfaces.

Mental health is an integral part of a person's overall condition, determining their emotional well-being, cognitive abilities, and quality of life. In recent decades, there has been a steady increase in the number of people suffering from mental disorders, such as depression, anxiety disorders, bipolar disorder, emotional burnout, and chronic fatigue. These conditions often develop gradually and may remain unnoticed until a critical phase. Therefore, early detection of signs of psychoemotional disorders is of particular importance.

One of the key indicators of a person's internal state is emotion. Emotions reflect a person's reaction to external and internal stimuli, and form the basis of motivation and behavior. They are accompanied by physiological changes, behavioral reactions, and, above all, are expressed through facial expressions – movements of the facial muscles that form facial expressions. Facial expressions, in turn, are a universal and intuitive communication channel: they allow you to convey emotional states even without words.

Of particular importance are microexpressions – short-term, involuntary facial expressions that last less than 0.5 seconds. These expressions arise in response to real emotions, even if a person tries to hide them. Scientific research has proven that microexpressions are a highly informative feature that can indicate internal conflicts, stress, anxiety or suppressed emotions. Changes in the nature of emotional manifestations – their frequency, intensity, symmetry and diversity – can be a sign of disorders in the psycho-emotional sphere. In recent years, the rapid development of computer vision, machine and deep learning methods has made it possible to automate the process of analyzing emotions. Modern systems are able to accurately detect a face on video, track key points (for example, using technologies such as MediaPipe FaceMesh), extract facial features and classify emotional states. This has created the prerequisites for building intelligent systems for detecting mental disorders that can assess a person's emotional reactions in real time. In modern society, there is an increase in the number of psycho-emotional disorders, such as depression, anxiety, bipolar disorder, chronic fatigue and other forms of mental disorders. These conditions often go unnoticed in the early stages due to the subjectivity of self-diagnosis, social stigma, and limited access to mental health professionals. In this regard, there is an increasing need to develop non-invasive, automated, and objective methods for the early detection of signs of mental disorders.

One of the promising areas is the analysis of a human face video stream to assess facial expressions, microexpressions, and emotional reactions. The face reflects a wide range of psycho-emotional states, and its movements are a powerful indicator of an individual's internal state. Modern computer vision and machine learning technologies make it possible to track the smallest changes in facial expressions and analyze them in real time, which opens up opportunities for continuous monitoring of the emotional state without the need for direct intervention.

Analysis of non-verbal signs, including facial expressions, microexpressions, and motor patterns, opens up new opportunities in the early detection of psychoneurological and mental disorders, such as depression, anxiety, autism, post-traumatic stress disorder, and schizophrenia.

The use of algorithms based on facial recognition features such as expression asymmetry, blink rate, muscle tension, lip and eye movement dynamics, in combination with emotion classification, provides a deeper understanding of a person’s internal psycho-emotional state. Such systems can be effectively used in telemedicine, psychotherapy, education, as well as for monitoring the condition of employees in responsible areas of activity.

Today’s algorithms provide high accuracy and can serve as the basis for scalable, non-invasive, and personalized mental health support systems. Among all approaches, deep neural networks can be noted, where deep learning algorithms, in particular convolutional neural networks (CNN), demonstrate high efficiency in the task of facial expression recognition (FER), significantly surpassing traditional methods in the accuracy and stability of results [1],[2].

The paper discusses the application of video analysis and deep learning methods to assess a person’s mental state in order to detect signs of mental disorders. The approaches are based on the synthesis of computer vision, affective computing, and machine learning methods.

A person’s mental health can manifest itself in various behavioral features or physiological changes. For example, micro-movements of the eyes and pupil dilation can be markers of internal stress. Deep gaze features extracted by autoencoders are effectively classified using Random Forest, achieving 100% accuracy [3]. Pupil diameter is tracked in the video stream and complements facial expression analysis to improve accuracy [4].

Body posture and gestures also reflect the psycho-emotional state. Using Pose Net, it is possible to analyze posture and body movements associated with anxiety (e.g. fidgeting, closed postures). Accuracy is F1: 0.98–0.99 [5]. Head movements (nodding, tilting) and body movements are analyzed together with facial expressions and gaze to obtain a comprehensive picture [6].

Video data allows non-invasive extraction of physiological signals: remote photoplethysmography (rPPG) is used to estimate HR and HRV. Random Forest classifiers achieve 99% accuracy [1,7]. Hemoglobin (HC) changes are estimated by skin color. The analysis is based on bit planes and ML models [8].

Facial expressions are one of the most informative channels for identifying emotional states, including anxiety. Active shape models (ASM) capture key facial landmarks (eyes, lips, eyebrows) and generate feature vectors classified by

SVM to determine anxiety and stress [9]. Deep neural networks trained on specialized datasets allow accurate recognition of anxious expressions and are integrated into web applications for real-time analysis [10]. Multi-task learning with attention mechanism combines facial expression with physiological characteristics (HR, HHR), achieving an accuracy of up to 94.33% [11].

We propose a video-centric framework to predict mental disorders by analyzing temporal dynamics, micro-expression volatility, and facial asymmetry metrics derived exclusively from RGB video streams. Leveraging the MediaPipe FaceMesh tool, our system quantifies disorder-specific biomarkers–such as delayed emotional reactivity in depression and slow blinks with yawning in fatigue. Our framework enables real-time mental health identification through optimized edge deployment, processing video streams directly on consumer-grade devices.

2. Materials and methods

In this study, we utilized a real-time emotion detection system with the help of the MediaPipe Face Mesh model developed by Google. We analyzed individual differences in emotional expression and recognition using 468 facial landmarks of the human face. The eyes, eyebrows, nose, mouth, and jawline were among the important face traits that the system tracked. Geometric features such as ratio of the human face, and distances between each keypoints were extracted using these landmarks.

In face detection side MediaPipe Face Mesh's utilize BlazeFace, a neural network architecture tailored for mobile GPU inference [21]. BlazeFace uses a simplified feature extractor made especially for faces, but it is based on the Single Shot MultiBox Detector (SSD). It uses a unique backbone that is similar to but different from MobileNetV1/V2 with custom residual structures known as BlazeBlocks and double BlazeBlock, which contains 5×5 depthwise separable convolutions instead of the typical 3×3 kernels. These design components aid in expanding the receptive field with little overhead and are computationally efficient. The model processes inputs at 128×128 resolution and uses an anchor-based detection scheme that stops at an 8×8 spatial resolution to minimize latency, replacing non-maximum suppression with a regression-weighted blending strategy to enhance stability over time.

The face landmark side generates a dense 3D mesh with 468 vertices, each of which is treated as an independent landmark [22]. These points are placed to capture perceptually significant facial regions. The model is trained using a combination of synthetic 3D renderings and annotated 2D landmarks from real-world mobile images, as well as special noise modeling and lighting variability techniques.

For our experiments we implemented the rule-based approach to identify different human mental disorders, including anxiety, depression, panic disorder, bipolar disorder and fatigue based on the relative positions and movements of facial features (see Table 1). It takes position coordinates of face landmarks and calculates the differences and distances between them. All positions are relative, so the distance between face and recording device will not affect the results.

Table 1 – Emotion detection criteria based on facial landmark features.

Condition	Facial Signs	Detection Logic
Anxiety	Rapid blinking, lip compression [12],[13]	Blink rate: EAR < 0.2 for 3+ frames = blink. >20/min = anxiety.
		Lip compression: Vertical distance between [61] and [291] < 0.05 (normalized).
Depression	Reduced smiling, flat affect [14],[15]	Smile absence: Mouth corner distance (61-291) < 0.2 (neutral) for >80% of frames.
		Slow blinks: EAR < 0.2 for >0.5s.
Hypervigilance (Post-traumatic stress disorder)	Sudden eye widening, mouth slackening [16],[17]	Eye widening: EAR > 0.3 (baseline: 0.25) + mouth openness (13-14 distance > 0.15).
Mania (Bipolar disorder)	Excessive smiling, eyebrow raises [18],[19]	Smile intensity: Mouth corner distance > 0.4 for >50% of video.
		Eyebrow volatility: Rapid AU2 (brow raise) spikes.
Fatigue (Sleep Deprivation)	Slow blinks, yawning [20]	Slow blinks: EAR < 0.2 for >0.5s.
		Yawning: Jaw (17-181 distance) > 0.3 for 2+ sec.

We manually annotated 100 video samples from human subjects, labeling each with one of five mental states. Each sample consisted of a short video sequence in which facial behavior was observed and classified using established psychological markers. The system's predictions, based on facial landmark thresholds, were then compared to these ground truth labels to evaluate classification performance. Experiments conducted on the machine with these specifications: NVIDIA GeForce RTX 3080 Laptop GPU and 16GB 3200MHz RAM.

For the real-time our system processed video frames at 30 frames per second using a standard webcam(see Figure 1). Each frame was converted to RGB format and passed through the MediaPipe Face Mesh pipeline to extract facial landmarks. They helped us to build prediction statements for human mental disorders. Facial signs conditions are written in python code as well as implementation of MediaPipe Face Mesh.

In comparison with usual similar systems, our framework employs a multi-stage adaptive architecture that integrates personalized baseline profiling, dynamic threshold adjustment (see Figure 2). By that we can address critical gaps in handling individual variability for each person. Also the system uses its lightweight CNNs for real-time detection for possibility to implement it in edge devices, where other models can use multiple models, separate for detection and disorder classification (see Figure 3).

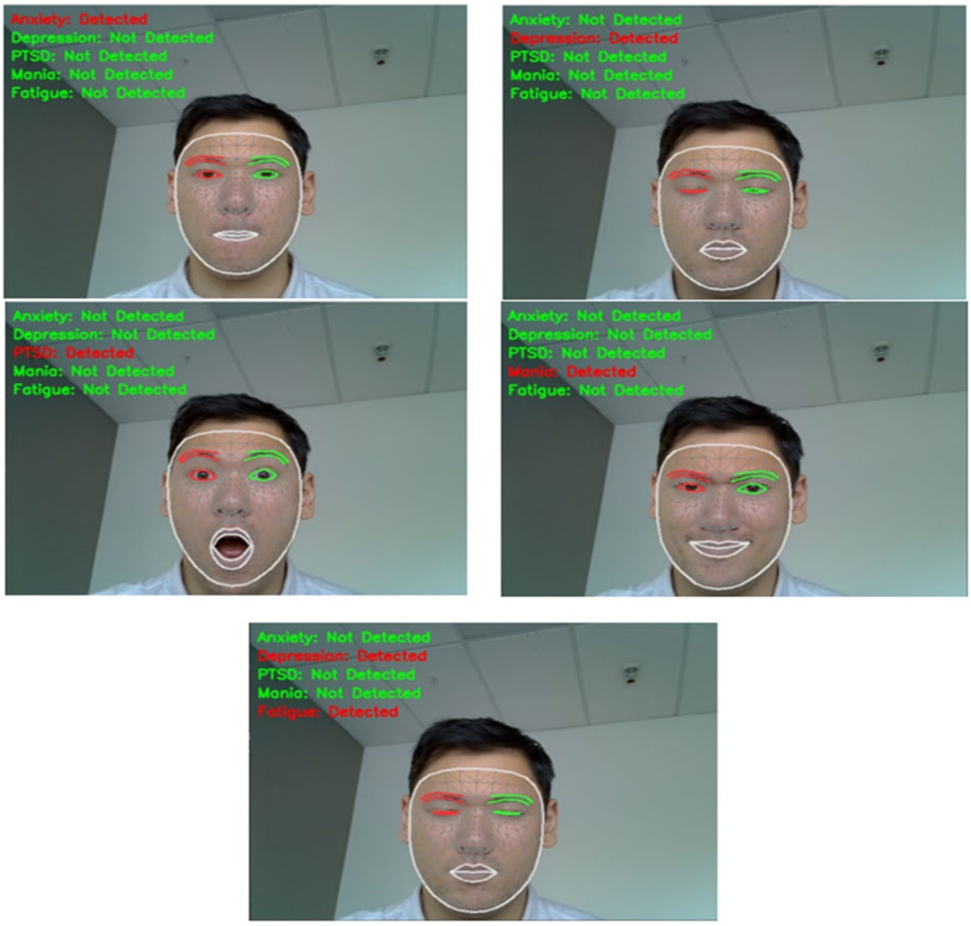


Figure 1 – Emotion recognition examples based
on facial expression analysis with one member of our team.

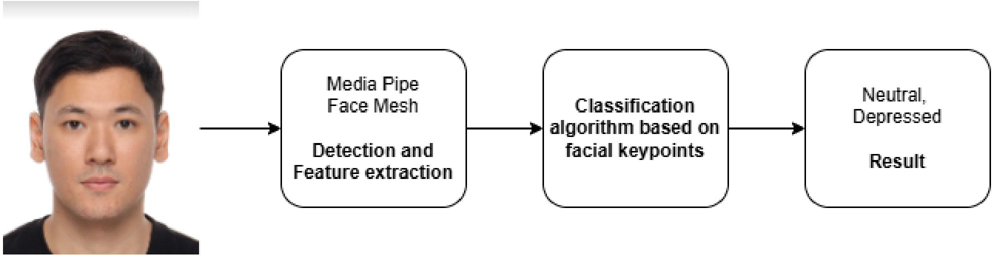

Media Pipe Face Mesh
Detection and Feature extraction
Classification algorithm based on facial keypoints
Neutral, Depressed
Result

Figure 2 – Proposed mental disorder detection system architecture.

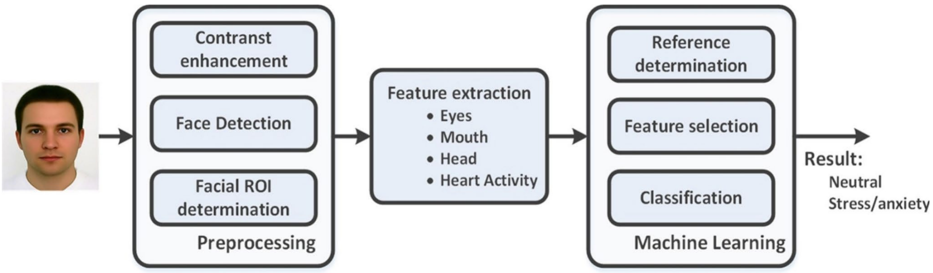


Figure 3 – Emotion recognition system example in [12].

To assess the performance of our rule-based system, we used standard classification metrics like precision, recall, F1-score and accuracy. These metrics enable us to quantify how well the system identifies each mental state, particularly in the presence of class overlap or imbalanced predictions.

Precision measures the reliability of the system when it predicts a specific class – in other words, how many of those predictions were actually correct. For example, when the system predicts that someone is experiencing mania, precision tells us how often that prediction is accurate (1):

$$\text{Precision} = \frac{\text{True Positives+False Positives}}{\text{True Positives}} \quad (1)$$

Recall quantifies the system's ability to detect all actual instances of a given class – for example, how many true cases of depression the system was able to identify (2):

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives+False Negatives}} \quad (2)$$

The F1-score strikes a balance between precision and recall, making it an especially useful metric when dealing with uneven class distributions or overlapping symptoms. It is the harmonic mean of precision and recall, yielding a single score that reflects both the system's accuracy and sensitivity (3).

$$\text{F1 score} = 2 * \frac{\text{Precision}\times\text{Recall}}{\text{Precision+Recall}} \quad (3)$$

Accuracy represents the overall effectiveness of the rule-based classification system (4):

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}} \quad (4)$$

3. Results

Anxiety is identified through heightened blink rates and lip compression [12,13]. The algorithm calculates blink frequency using the Eye Aspect Ratio (EAR), where an EAR below 0.2 for three consecutive frames registers as a blink. A sustained rate exceeding 20 blinks per minute signals potential anxiety. At the same time, lip compression measured by the vertical distance between landmarks 61, which is upper lip, and 291, which is lower lip, is flagged when normalized to less than 0.05. MediaPipe Face Mesh tracks these metrics via eyelid landmarks, for example indices 362, 385 are for the left eye and 33, 160 for the right eye, and lip landmarks.

According to [15], depression is associated with reduced smiling and prolonged blinks. Our algorithm monitors the distance between mouth corners, the landmarks 61 and 291, where a value below 0.2 for over 80% of frames suggests diminished positive affect. Slow blinks, defined as eyelid closure, with EAR smaller than 0.2 lasting longer than 0.5 seconds, are tracked using the same eye landmarks as anxiety. MediaPipe's precision in capturing subtle lip and eyelid movements allows for continuous assessment of emotional withdrawal.

Hypervigilance in Post-traumatic stress disorder(PTSD), is detected through sudden eye widening and mouth slackening [16,17]. We fully used MediaPipe's capability to track subtle changes in eye and inner lip landmarks for real-time detection. An EAR exceeding 0.3, signals exaggerated eye openness. Meanwhile inner lip distance, which are the landmarks 13 and 14, surpassing 0.15 reflects mouth tension release.

Mania is characterized by excessive smiling and erratic eyebrow movements and most of the time is a Bipolar disorder [18,19]. In our algorithm a mouth corner distance, which are landmarks from 61 to 291, exceeding 0.4 for more than half of observed frames indicates intense, sustained smiling. Also eyebrow volatility, which is measured by rapid displacement of brow landmarks 336 and 296 for the left, 66 and 105 for the right brow, reflects heightened arousal. We map these dynamics through upper lip curvature, 37 and 267, and brow motion.

Figure 4 illustrates temporal trajectories of selected facial landmarks**.** For example, eyebrow landmark movements (66, 105, 296, 336) display pronounced volatility in mania compared to stable patterns in neutral or depressive states. Such temporal dynamics highlight the relevance of personalized thresholding.

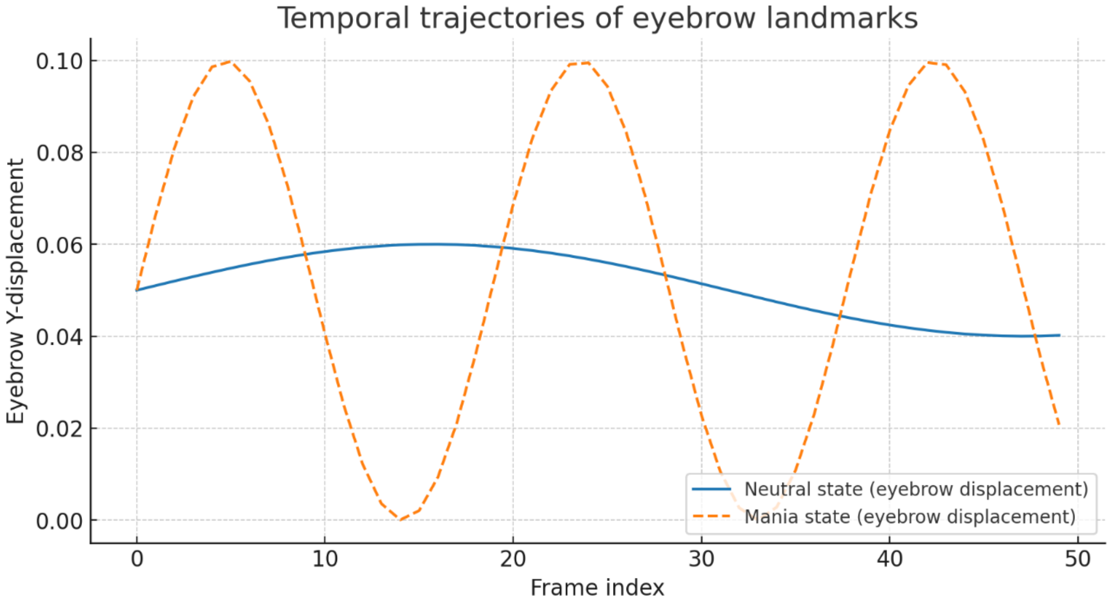


Figure 4 – Temporal trajectories of selected facial landmarks

Fatigue measured in our algorithm through prolonged blinks and yawning [20]. Slow blinks, EAR smaller than 0.2 for more than 0.5 seconds, are paired with jaw dropping. It is measured by the distance between chin landmark 17 and jawline point 181 exceeding 0.3 for two seconds. We used MediaPipe's jaw and eyelid tracking to ensure robust detection across varied scenarios like stretching or screen interaction.

In Table 2 we show the confusion matrix, which summarizes the classification results across the five mental states. Each row represents the actual label, and each column represents the predicted label assigned by the system. Diagonal entries represent correctly classified samples, whereas off-diagonal entries indicate misclassification. The model was highly accurate in detecting mania and depression, with few misclassifications. However, there was a significant overlap between fatigue, anxiety, and PTSD, owing to the visual similarity of certain facial cues such as droopy eyelids and decreased facial tension across these states.

Table 2 – Confusion matrix illustrating the classification performance of the rule-based system across five mental state categories.

Actual \ Predicted	Anxiety	Depression	PTSD	Mania	Fatigue
Anxiety	14	2	1	0	3
Depression	1	16	0	0	3
PTSD	2	1	13	1	3
Mania	0	0	1	18	1
Fatigue	3	2	2	1	12

We calculated the precision, recall, and F1-scores for each of the five mental state categories using the confusion matrix(see Table 3). Mania performed the best in all metrics, with precision and recall both at 0.90, followed by depression and PTSD. Fatigue had the lowest scores, owing to visual similarities with other classes, which resulted in a higher false positive rate. The system's overall accuracy was 73%, which demonstrates the reliability of the threshold-based approach.

Table 3 – Performance metrics of the system across five mental state categories.

Metric	Anxiety	Depression	PTSD	Mania	Fatigue
Precision	0.700	0.762	0.765	0.900	0.545
Recall	0.700	0.800	0.650	0.900	0.600
F1-score	0.700	0.780	0.703	0.900	0.571
Accuracy	**0.730**				

All information taken from a person's facial behaviour can predict some mental disorders. Nonetheless, similar patterns in each case may arise during different human moods. For example for Hypervigilance conditions, the same patterns can happen in the moments of surprise or concentration, underscoring the need for corroborating behavioral data to distinguish pathology from transient reactions. It risks misinterpreting naturally expressive individuals, emphasizing the importance of longitudinal tracking to differentiate episodic mental disorder actions from temperamental exuberance.

The experimental results demonstrate that personalized video-based emotion recognition models can be used in real-world applications, with good camera conditions, where a person can be detected with the Mediapipe library. This result stems from the models' ability to account for individual variations in facial dynamics, such as unique microexpression patterns and temporal differences in emotional expression transitions (Figure 1). However, the analysis revealed challenges in distinguishing between visually similar states, such as fatigue and depression, because of overlapping facial conditions like drooping eyelids and reduced smile intensity.

MediaPipe's facial landmark tracking proved effective in capturing facial features like jaw tension, eyelid closure duration, and inner lip compression. They are all associated with mental states. For instance, we can differentiate landmark-based metrics with intentional smiles from spontaneous ones. To address limitations for individual persons, we incorporated adaptive thresholds that adjusted classification criteria based on user-specific baselines. For example, resting lip distance and blink frequency were calibrated during initial sessions, reducing false positives in natural settings.

4. Discussion

In this study, we conducted research aimed at detecting mental disorders by analyzing videos that assess emotions, facial expressions, and facial movements. The proposed approach, based on computer vision technologies, demonstrates strong potential for the early diagnosis of mental disorders. The developed technology can be useful in healthcare practice for pre-identifying signs of mental distress before a patient visits a physician, thereby accelerating diagnosis and improving treatment outcomes. In this study, only possible mental disorders were considered, but in the future, this method may be extended to diagnose neurological conditions, such as Bell's palsy, which is characterized by facial asymmetry.

A promising direction for the further development of this technology is its application in the diagnosis of neurological disorders. In particular, video stream analysis may be beneficial in detecting Parkinson's disease, Huntington's disease, bulbar and pseudobulbar syndromes, multiple sclerosis, facial neuritis (including Bell's palsy), and Tourette's syndrome. Including such conditions would significantly increase the diagnostic value of the proposed approach and expand the possibilities for early detection of diseases manifesting through facial and motor patterns.

The difficulty in distinguishing between fatigue and depression highlights one of the key limitations of unimodal systems. Although both conditions are associated with reduced ocular activity and slowed facial expressions, depression is often characterized by prolonged microexpressions of sadness. The integration of speech analysis or a combination of head pose tracking with galvanic skin response (GSR) data may allow for a more accurate differentiation between cognitive fatigue and emotional withdrawal.

Thresholds for indicators such as lip compression, gaze deviation, or smile intensity may vary significantly across ethnicities, ages, and cultural backgrounds. For instance, norms for lip distance are influenced by facial structure, while eye contact conventions are determined by cultural norms. It is well known that individuals with darker skin tones often experience higher false-positive stress detections due to improper calibration of infrared cameras. This can lead to entrenched diagnostic bias. To reduce such risks, participatory design principles should be applied: involving diverse user groups in threshold calibration, using synthetic datasets to supplement underrepresented facial structures, and including cultural consultants to adapt models to regional norms. For example, increasing the acceptable threshold for gaze deviation in cultures where avoiding eye contact is perceived as a sign of respect rather than an attempt at deception.

While the current version of the system applies rule-based thresholds, including metrics like eye aspect ratio, lip distance, and blink rate, based on existing empirical studies, we acknowledge the need for a more rigorous scientific foundation. In future studies, we will perform systematic threshold validation across larger and more diverse samples using statistical analyses, expert clinical input, and data-driven optimization techniques. This will ensure that the selected facial parameters have robust diagnostic relevance rather than relying solely on heuristic definitions.

Another important aspect not yet addressed in this study is the comorbidity of mental disorders. Many individuals exhibit overlapping symptoms from multiple conditions–such as depression co-occurring with anxiety, or PTSD alongside fatigue–which complicates classification based on isolated features. Since our current framework is designed to detect single-condition markers, it may fail to capture the interaction between multiple symptom domains.

To address this, we plan to introduce multilabel classification models capable of identifying multiple co-occurring conditions simultaneously. Additionally, we aim to adopt temporal modeling techniques (e.g., LSTMs, Transformers) that account for behavioral patterns over time–patterns which often reflect comorbid states more reliably than static snapshots. These enhancements will be supported by expanding the dataset to include clinically verified cases with annotated comorbid symptoms.

Improving the accuracy and reliability of the system requires refinement of the algorithms and expansion of the training data. Introducing more complex analysis conditions and training the model on larger datasets with annotated mental disorders will allow much of the diagnostic process to be delegated to artificial intelligence. This will not only enhance recognition accuracy but also improve model robustness to environmental factors such as lighting, camera angle, background, and more.

Nevertheless, this study has a number of significant limitations related to the dataset used. First, the dataset is proprietary and private, created solely for experimental validation. It contains video recordings of only one subject, captured under controlled conditions. As such, the dataset lacks diversity in facial expressions, facial structure, and behavioral responses. This limited scope restricts the generalizability of results and prevents broader validation of model robustness.

In future research, we plan to address these limitations. First, we will develop a larger dataset involving a diverse group of participants, including both clinical cases (e.g., depression, anxiety, bipolar disorder) and healthy individuals. This will allow us to balance class representation and improve the generalizability of the model.

Second, we plan to ensure demographic diversity across factors such as gender, age, ethnicity, and cultural background. This is essential for creating a robust and ethically sound model that does not discriminate based on appearance or cultural behavior.

Third, in future work, we intend to shift from a unimodal approach to a multimodal system, integrating visual cues with speech characteristics, physiological data (e.g., heart rate, EEG, GSR), and audio signals. This approach will enhance the diagnostic accuracy, particularly in complex cases where behavioral cues may be absent or masked.

Furthermore, we aim to implement an open and reproducible methodology, including the publication of anonymized data (with participant consent), label definitions, feature descriptions, and transparent evaluation protocols. This will support scientific verification and comparability with other research efforts.

Thus, the future development of our system is focused on creating a scalable, ethically responsible, and personalized tool for the early diagnosis of mental and neurological disorders in real-world conditions.

5. Conclusions

This research demonstrates the feasibility of using video-centric affective computing as a non-invasive, scalable tool for mental health assessment. By analyzing temporal facial dynamics and microexpressions extracted from RGB video using MediaPipe FaceMesh, the system identifies disorder-specific facial markers such as delayed emotional reactivity (in depression) and microexpression volatility (in anxiety). The real-time capability of the framework, coupled with personalized adaptation that adjusts thresholds to individual facial baselines, highlights its potential for practical deployment across neurodiverse and demographically varied populations.

However, the study also reveals important limitations. The model was developed and tested on a private, single-subject dataset without class balance or demographic variation. Rule-based labeling was applied without clinical verification, which limits generalizability and diagnostic reliability. In addition, the current unimodal design restricts the system's ability to resolve overlapping emotional states–such as distinguishing between fatigue and depression–both of which may exhibit similar facial attenuation.

To overcome these challenges, future work will focus on developing a large, demographically diverse dataset with clinically annotated cases, including both isolated and comorbid mental health conditions. The system will evolve toward a multimodal architecture by integrating facial analysis with speech prosody, physiological signals (e.g., heart rate variability, EEG, GSR), and temporal modeling techniques (e.g., LSTMs, Transformers) to better capture dynamic and overlapping behavioral cues.

Ethical deployment is central to this work. As systems like this infer highly sensitive mental states, safeguards must include transparency at multiple levels: how decisions are made (e.g., which facial features contribute), how data is stored and used, and how predictions are interpreted. In high-stakes applications such as workplace monitoring or insurance assessment, dynamic consent mechanisms and user-facing explainability are critical to ensuring trust and preventing misuse.

Cultural and neurodivergent inclusion must remain a design priority. Participatory design with underrepresented groups, systematic bias audits, federated learning methods, and region-specific model calibration will help reduce risks of stigmatization and increase fairness and model robustness across diverse real-world contexts.

In conclusion, while this work serves as a foundational proof of concept, its continued development–through theoretical grounding, multimodal expansion, and ethical reinforcement–will be essential for building a clinically reliable and socially responsible system for early mental health and neurological disorder detection.

Funding

This research was funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan, Grant No. AP26100559 “Development of computational methods for determining and generating positive and negative emotions from photo and video data for monitoring of the educational process”.

Author Contributions

Conceptualization: A.N., M.M. and A.Z.; Methodology: A.N., M.M.; Software: M.M.; Validation: G.I., M.R. M. and A.Z.; Formal analysis: A.N., M.M.; Investigation: A.N., M.M.; Resources: A.N., M.M.; Data Curation: M. M.; Writing – Original Draft Preparation: A.N., M.M.; Writing – Review & Editing: A.N., M.M.; Visualization: M.M.; Supervision: G.I., M.R.M. and A.Z.; Project administration: A.Z.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. A. Devarapalli and J. Gonda, "Investigation into facial expression recognition methods: a review," *Indonesian Journal of Electrical Engineering and Computer Science*, 2023. doi: 10.11591/ijeecs.v31.i3.pp1754-1762.

2. D. Uneza and S. Gupta, "Facial Expression Analysis: Unveiling the Emotions Through Computer Vision," in *Proc. Int. Conf. Recent Innovations in Technology and Optimization (ICRITO)*, 2024, pp. 1–5. doi: 10.1109/icrito61523.2024.10522418.

3. N.S. Harshit, N. K. Sahu, and H. R. Lone, "Eyes Speak Louder: Harnessing Deep Features From Low-Cost Camera Video for Anxiety Detection," in *Proc. ACM Conf.*, 2024, pp. 23–28. doi: 10.1145/3662009.3662021.

4. A. Tiwari, B. Matejek, and D. Haehn, "Non-Invasive Stress Monitoring From Video," in *Proc. IEEE Int. Symp. Biomedical Imaging (ISBI)*, 2024, pp. 1–5. doi: 10.1109/isbi56570.2024.10635725.

5. Y. Amirgaliyev, I. Krak, I. Bukenova, B. Kazangapova, and G. Bukenov, "Determining the psycho-emotional state of the observed based on the analysis of video observations," *Eastern-European Journal of Enterprise Technologies*, 2024. doi: 10.15587/1729-4061.2024.296500.

6. A. Kargarandehkordi and P.S.W. Vecilla, "Computer Vision Estimation of Stress and Anxiety Using a Gamified Mobile-based Ecological Momentary Assessment and Deep Learning: Research Protocol," *medRxiv*, 2023. doi: 10.1101/2023.04.28.23289168.

7. M. Khomidov, D. Lee, C. Kim, and J.-H. Lee, "The Real-Time Image Sequences-Based Stress Assessment Vision System for Mental Health," *Electronics*, vol. 13, no. 11, p. 2180, 2024. doi: 10.3390/electronics13112180.

8. K. Lee, P. Zhang, and S. Wu, "System and method for camera-based stress determination," 2019.

9. M. Penev, A. Manolova, and O. L. Boumbarov, "Active Shape Models with 2D profiles for Stress/Anxiety recognition from face images," in *Proc. Int. Conf. Communications*, 2014, vol. 1, pp. 108–112. [Online]. Available: http://e-university.tu-sofia.bg/e-publ/files/1697_CEMA14_Martin_Agata.pdf

10. H. Chandika, B. Soumya, B.N.E. Reddy, and B.M.S. SaiManideep, "Real-Time Stress Detection and Analysis using Facial Emotion Recognition," *Int. J. Adv. Res. Comput. Commun. Eng.*, 2024. doi: 10.17148/ijarcce.2024.13324.

11. J. Xu, C. Song, Z. Yue, and S. Ding, "Facial Video-Based Non-Contact Stress Recognition Utilizing Multi-Task Learning With Peak Attention," *IEEE J. Biomed. Health Inform.*, pp. 1–12, 2024. doi: 10.1109/jbhi.2024.3412103.

12. G. Giannakakis et al., "Stress and anxiety detection using facial cues from videos," *Biomed. Signal Process. Control*, vol. 31, pp. 89–101, 2017. doi: 10.1016/j.bspc.2016.06.020.

13. H. Chandika, B. Soumya, B. N. E. Reddy, and B. M. S. SaiManideep, "Real-Time Stress Detection and Analysis using Facial Emotion Recognition," *Int. J. Adv. Res. Comput. Commun. Eng.*, 2024. doi: 10.17148/ijarcce.2024.13324.

14. G. Orrù et al., "Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review," *Neurosci. Biobehav. Rev.*, vol. 36, no. 4, pp. 1140–1152, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22305994

15. J. Singh and G. Goyal, "Decoding depressive disorder using computer vision," *Multimedia Tools Appl.*, vol. 80, pp. 8189–8212, 2021. doi: 10.1007/s11042-020-10128-9.

16. W.R. Ringwald et al., "Day-to-day dynamics of facial emotion expressions in posttraumatic stress disorder," 2024. doi: 10.31234/osf.io/6fqg9.

17. L.R. Enders, H. Roy, T. Rohaly, A. Jeter, and J. Villarreal, "Impacts of Posttraumatic Stress Disorder on Eye-Movement during Visual Search in an Open Virtual Environment under High and Low Stress Conditions," *J. Vis.*, vol. 23, no. 9, p. 5691, 2023. doi: 10.1167/jov.23.9.5691.

18. J. Gruber et al., "Associations between hypomania proneness and attentional bias to happy, but not angry or fearful, faces in emerging adults," *Cognition & Emotion*, vol. 35, no. 1, pp. 207–213, 2021. doi: 10.1080/02699931.2020.1810638.

19. J. Wang, Y. Song, H. Li, Z. Leng, M. Li, and H. Chen, "Impaired Facial Emotion Recognition in Individuals with Bipolar Disorder," *Asian J. Psychiatry*, vol. 102, p. 104250, 2024. doi: 10.1016/j.ajp.2024.104250.

20. R. Schleicher, N. Galley, S. Briest, and L. Galley, "Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired?," *Ergonomics*, vol. 51, no. 7, pp. 982–1010, 2008. doi: 10.1080/00140130701817062.

21. V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran, and M. Grundmann, "BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs," *arXiv preprint*, arXiv:1907.05047, 2019. doi: 10.48550/arXiv.1907.05047.

22. Y. Kartynnik, A. Ablavatski, I. Grishchenko, and M. Grundmann, "Real-time Facial Surface Geometry from Monocular Video on Mobile GPUs," *arXiv preprint*, arXiv:1907.06724, 2019. doi: 10.48550/arXiv.1907.06724.

Information about authors

Aizhan Nurzhanova is a 1st year doctoral student in the Department of Computer and Software Engineering at L.N. Gumilyov Eurasian National University (Astana, Kazakhstan, nuraizhan87@mail.ru, +77028307620). Her research interests include video-based emotion recognition, facial expression analysis, and machine learning applications in mental health. ORCID iD: 0009-0006-9871-9823.

Miras Mussabek is a 1st year doctoral student in the Department of Computer Engineering at Astana IT University (Astana, Kazakhstan, miras.k@astanait.edu.kz, +77071771011). His research interests include video-based detection, recognition. ORCID iD: 0009-0009-2353-3524.

Dr. Gokhan Ince is an Associate Professor in the Computer Engineering Department, Faculty of Computer and Informatics Engineering at Istanbul Technical University (Istanbul, Turkey, gokhan.ince@itu.edu.tr, +90 (212) 285 69 86 ext: 6986). He has extensive experience in signal processing, affective computing, and human–robot interaction. ORCID iD: 0000-0002-0034-030X.

Dr. Mas Rina Mustaffa is an Associate Professor at the Faculty of Computer Science, University Putra Malaysia (Serdang, Malaysia, MasRina@ump.edu.my). Her research interests include pattern recognition, emotion detection, and deep learning for intelligent systems. ORCID iD: 0000-0001-5088-2871.

Dr. Ainur Zhumadillayeva is an Associate Professor in the Faculty of Information Systems, Department of Computer and Software Engineering at L.N. Gumilyov Eurasian National University (Astana, Kazakhstan, zhumadillayeva_ak@enu.kz, +77025295999). Her research focuses on machine learning, data mining, and educational technologies. ORCID iD: 0000-0003-1042-0415.

Submission received: 16 April, 2025.
Revised: 30 August, 2025.
Accepted: 30 August, 2025.

IRSTI 06.52.35 https://doi.org/10.26577/jpcsit2025338

Ali Rakhimzhanov[1*], Jelena.V. Caiko [2]

[1]Kazakh-British Technical University, Almaty, Kazakhstan
[2]Riga Technical University, Riga, Latvia
*e-mail: ali.rakhimzhanov11@gmail.com

FORECAST OF HOUSING PRICES IN ALMATY USING MACHINE LEARNING ALGORITHMS

Abstract. Precise prediction of housing values is an important task for various stakeholders involved in the housing market, including investors, builders, and city planners. In this research, supervised machine learning models are used to predict the price of apartments in Almaty, Kazakhstan, which is a dynamic urban market in Central Asia. With an openly available dataset of apartments for sale, Linear Regression, Lasso Regression, Random Forest, and XGBoost models are implemented and tested. The data is scaled and encoded with scalable pipelines, and models are evaluated with regards to Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R² Score. The best performing model amongst those tested was Random Forest Regressor with an R² of 0.9158, followed by XGBoost with 0.8438. Feature importance visualization identifies district, area, and construction year as primary influencing factors. The research supports that ensembling machine learning models are efficient and scalable predictors for housing forecasts and suggests future improvements with time-series and geospatial features.

Keywords: Housing price prediction; Machine learning algorithms; Regression models; Random Forest; Real estate market analysis; Urban economy.

1. Introduction

Housing markets worldwide have witnessed unprecedented deviations through economic, social, and political drivers in recent years. Almaty, Kazakhstan, and urban hubs of emerging economies around the world, have not escaped these trends. As the largest city and economic capital of the country, Almaty, with its high urbanization growth rate, has witnessed heightened demand for housing, infrastructural growth, and an uptrend for housing prices. Official numbers have Almaty place consistently at or near the head of the list of highest property prices per square meter in Kazakhstan[1], but pricing patterns remain geographically localized and data-dispersed. Such elements inject uncertainty and inconsistency into housing price valuation, posing a most sensitive challenge to stakeholders along the property spectrum–including buyers and sellers, as well as developers, banks, and policymakers.

Traditional housing price appraisal methods depend on statistical methods or human judgment, both of which suffer from inherent drawbacks. Linear regression models, for example, are unable to identify nonlinear relationships and interactions between features that exist in housing data. Further, these approaches are prone to human bias and tend to fail to cope with fast-evolving market conditions. On the other hand, machine learning (ML) algorithms have proven to be high-performance tools with the capability to learn intricate patterns based on large databases and make precise forecasts. The fact that ML models can perform prediction automatically and improve with time makes them the best fit for real estate prediction [3][2].

Around the world, several studies have established machine learning's capability for accurate price prediction of real estate. Decision Trees, Random Forest, Gradient Boosting Machines (i.e., XGBoost), and Artificial Neural Networks have been proven to outshine traditional methods, particularly for large, heterogenic datasets. For instance, studies done in China [4], India[5], and the United States [6] have established ML's potential for enhancing house price estimation. The majority of these studies, albeit, are based on developed real estate markets, where data access and quality facilitate such analyses. Conversely, there is a clear void of scholarly literature focusing on Central Asia and, more specifically, Kazakhstan, where there is underutilization of real estate data for model analyses. It is this gap that creates an imperative for regionally based studies utilizing contemporary data science approaches for emerging markets.

This paper seeks to close that gap by examining machine learning algorithm applicability and effectiveness for residential housing price prediction in Almaty. We gather and preprocess actual estate listings from open data sources, extract features that are applicable, and use various regression-based ML models to fit and validate price forecasts. Models considered include Linear Regression, Lasso Regression, Random Forest, and XGBoost. Model performance was evaluated using three widely accepted metrics: the coefficient of determination (R^2, Eq. 1), Root Mean Squared Error (RMSE, Eq. 2), and Mean Absolute Error (MAE, Eq. 3). These metrics provide complementary perspectives on predictive accuracy, penalizing systematic bias, large deviations, and overall fit respectively.

The contributions of this study are three: First, we propose an extensive ML-based framework specific to Almaty's housing market, based on available public datasets and sophisticated algorithms. Second, we compare several different models to arrive at a suitable method for housing price prediction under conditions specific to Almaty. Third, we make actionable contributions to understanding what drives housing price dynamics in Almaty, which can help stakeholders make informed decisions based on data.

The rest of this paper is outlined below: Section 2 contains a literature review of machine learning and housing price forecasting. Section 3 outlines the data, feature engineering, and data preprocessing steps. Section 4 states the methodology, including model and training procedures. Section 5 presents results and evaluation. Section 6 concludes and offers future research guidelines and recommendations.

2. Literature review

Increases in data science adoption within real estate analytics have resulted in dramatic improvements to housing price forecasting. Historically, valuation was controlled by hedonic methodologies, which employed linear regression to link attributes of housing (size, location, number of bedrooms, etc.) to price. Though successful under restricted circumstances, these models fail to represent nonlinearities and high-level interactions between features, which are common under dynamic housing conditions. As a result, ML methods have become favored because of high-dimensional modeling, automatic discovery of patterns, and improved predictive performance under various market conditions.

The theoretical foundation of these approaches lies in the hedonic pricing model [7], which has long been used to capture structural and locational effects in property values. However, subsequent research demonstrated that linear frameworks struggle with multicollinearity and fail to incorporate spatial dependencies effectively [8].

Among the first to criticize traditional approaches was [2] who presented the shortcomings of a hedonic price model and promoted more adaptive, data-driven methods, particularly for differentiated urban contexts. A later review by [3] reinforced this development, illustrating that Random Forest, XGBoost, and Artificial Neural Networks outperformed linear models substantially through international case studies.

Other early explorations of machine learning in housing price prediction, such as [9], also demonstrated the potential of structured data integration (e.g., energy efficiency, accessibility), showing that nontraditional features could strongly influence property valuations.

Specifically, [4] proposed an XGBoost-driven housing price prediction model for urban China. It identified that the algorithm performed well in dealing with both high-dimensional inputs and missing data and surpassed Decision Trees and Ridge Regression with better RMSE and R^2 scores. Analogously, [5] surveyed more than a hundred research papers and concluded that ensemble learning methods, namely Random Forest and boosting methods, generalized well on actual housing datasets. Notably, Random Forest achieved the lowest RMSE (Eq. 2), confirming its robustness across error measures

Researchers have pursued hybrid methods to increase prediction accuracy. As an example, [10] employed a stacked model with Linear Regression, Random Forest, and XGBoost to achieve high precision for a Turkish housing market dataset. Another creative solution was proposed by [11] who integrated their prediction model into an MLOps pipeline, which supports auto-deployment and perpetual optimization within live real estate platforms.

In spite of this quick progress, research in emerging markets, and especially Central Asia, is still limited. A major contribution to this field is provided by [12], who implemented Naive Bayes, Decision Trees, and AdaBoost on housing market data in Kazakhstan. Their research shows that machine learning can identify insightful market patterns and price drivers with only limited data available locally. This regional lack of research emphasizes how crucial and innovative it is to have localized ML models for cities such as Almaty.

In addition, deep learning–based models have been tested in international contexts (Cheng & Wang, 2018) [13], though their higher computational demand and sensitivity to feature scaling make them less commonly adopted in smaller-scale or emerging market studies compared to ensemble methods.

More recent studies are moving towards spatiotemporal modeling as well. [14] investigated applying geographically and temporally weighted machine learning, and demonstrated that including neighborhood-level and seasonal dynamics greatly enhances Sydney's forecasting. In a similar vein, [15] used Explainable AI (XAI) methods to model affordable housing dynamics with land value and zoning data, with a focus on interpretability for urban planning.

Recent research also considered feature engineering and multiobjective optimization. [16] employed evolutionary algorithms to hybridize ML with optimization methodologies and demonstrated that models with domain-specific objectives perform well in actual deployments. The relevance of hyperparameter fine-tuning, cross-validation, and importance of features has been reinforced through most recent literature, and ensembling models have remained at or near the top of performance and stability rankings. Furthermore, [17] suggested a hybrid model of TLBO and XGBoost with inherent uncertainty estimations to provide confidence-scored predictions for construction and real estate evaluations. Further, research as presented by [18][19] identifies the trend towards universal AI frameworks for construction and real estate.

3. Data Description and Preprocessing

3.1 Dataset Overview

We obtained data for this research through Kaggle [20], and it contains 11,883 residential apartments for sale in Almaty, Kazakhstan. Each record is a unique listing and includes rich features that specify the physical attributes, address, and price of the selling apartment. The dataset contains a variety of different types of apartments, including those found within Soviet-era structures to those newly developed high-rise buildings.

Table 1 – Features for Modeling

Feature Name	Original Data Type	Description / Unit	Role
price	Numerical	Total apartment sale price (in KZT)	Target variable
area	Numerical	Apartment floor area (in square meters)	Input feature
no_of_rooms	Numerical	Number of rooms	Input feature
floor	Numerical	Floor level	Input feature
year_of_construction	Numerical	Year building was constructed	Input feature
district	Categorical	Administrative district of Almaty	Input feature
structure_type	Categorical	Type of building construction (e.g., Brick)	Input feature
quality	Ordinal	Subjective quality score (Very Poor to Excellent)	Input feature
Id	Nominal	Unique listing identifier	Dropped
price_per_sqm	Derived (Numerical)	Price divided by area (KZT per m²)	EDA-only

3.2 Preprocessing

3.2.1 Data Cleaning

During initial data inspection, there were no missing or null values found in important fields. Duplicates were checked based on the Id field and deleted as required. All the numeric fields,price, area, and year of construction,were within desirable limits, reflecting good data consistency.

3.2.2 Feature Decoding

The original data had several categorical features encoded numerically. These were interpreted as follows:

- districts: 0–7 → Almalinsky, Auezovsky, Bostandyk, etc.
- structure_type: 0–3 → Panel, Brick, Monolithic, Other
- quality: 0–4 → Very Poor to Excellent

3.2.3 Feature Engineering

A derived variable, price per square meter, was defined by the formula:

price_per_sqm = price / area

The engineered feature price_per_sqm was calculated only for exploratory analysis purposes and was not included in the set of input features used for model training or testing.

Using this variable as a predictor would create target leakage, since it is a transformation of the target (price).

Instead, we used price_per_sqm only for visualizations in EDA (e.g., price distribution plots by district) to help understand pricing patterns and variability.

3.2.4 Summary Statistics

As indicated by Table 2, Almaty's average apartment costs around 55.7 million KZT and measures around 67.9 square meters. Price and area, as expected, both show wide variation, reflecting a highly diverse market. Construction years span from 1932 to 2023, showing both old Soviet structures and new developments.

Table 2 – Descriptive Statistics of Main Apartment Features in Almaty

Feature	Mean	Std. Dev.	Min	25%	Median	75%	Max
Price (KZT)	55,745,160	507,514,200	3,500,000	27,900,000	35,000,000	47,000,000	5.7B+
Area (m²)	67.9	42.9	8.0	44.0	59.5	80.0	600+
Rooms	2.24	1.01	1	1	2	3	5+
Floor	5.13	4.11	1	2	5	7	24
Year Built	2001	20.7	1932	1982	2008	2020	2023

3.3 Statistical Tests and Feature Diagnostics

- Normality Tests:

In addition to visual inspection of price and area histograms, we conducted statistical tests to examine distributional assumptions. Both Shapiro–Wilk and Kolmogorov–Smirnov tests rejected the null hypothesis of normality for apartment prices ($p < 0.001$) and living area ($p < 0.01$), confirming skewness and heavy tails.

- Multicollinearity:

Variance Inflation Factor (VIF) analysis indicated no severe multicollinearity (VIF < 5 across predictors). However, strong correlations were found between area, number of rooms, and price, which may explain instability in linear models.

- Feature Importance Beyond Gain:

SHAP (SHapley Additive exPlanations) values were used to interpret feature contributions beyond the XGBoost gain metric. SHAP confirmed that district, area, and quality were dominant predictors, but also revealed nonlinear effects (e.g., diminishing returns for very large apartments).

3.3 Exploratory Visualizations

To better understand data distribution and detect potential modeling issues, the following visualizations were created:

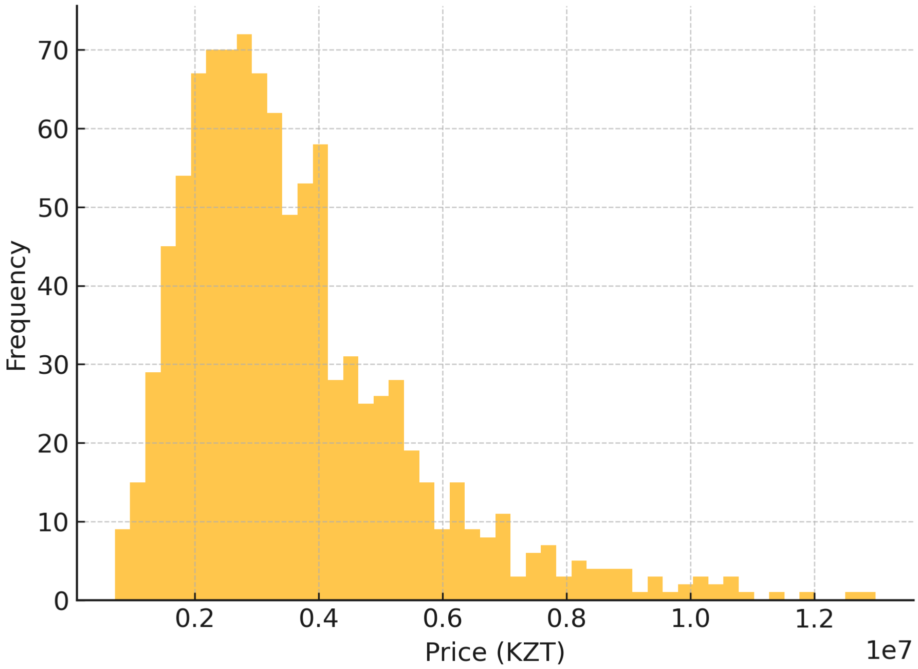


Figure 1 – Distribution of apartment prices in Almaty.

Price Distribution: Skewed to the right; most apartments are priced between 10 and 50 million KZT. Some listings exceed 500 million KZT, creating a long tail,indicating luxury segments that modeling needs to account for.

Area Distribution: Apartments range from compact 8 m² units to over 200 m². Most listings fall between 40–70 m², aligning with typical urban layouts.

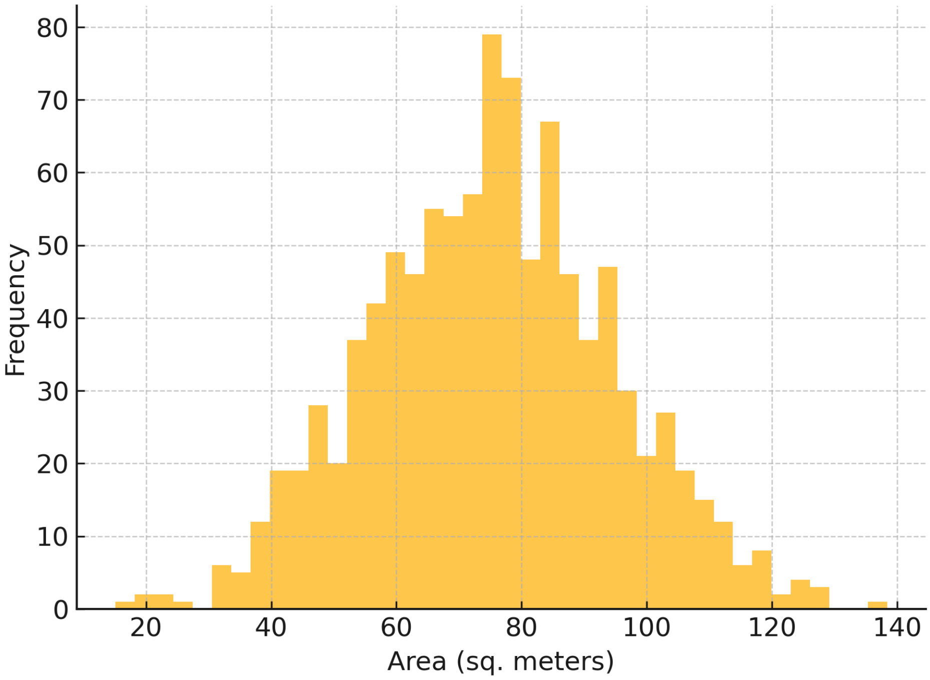


Figure 2 – Distribution of apartment areas (square meters).

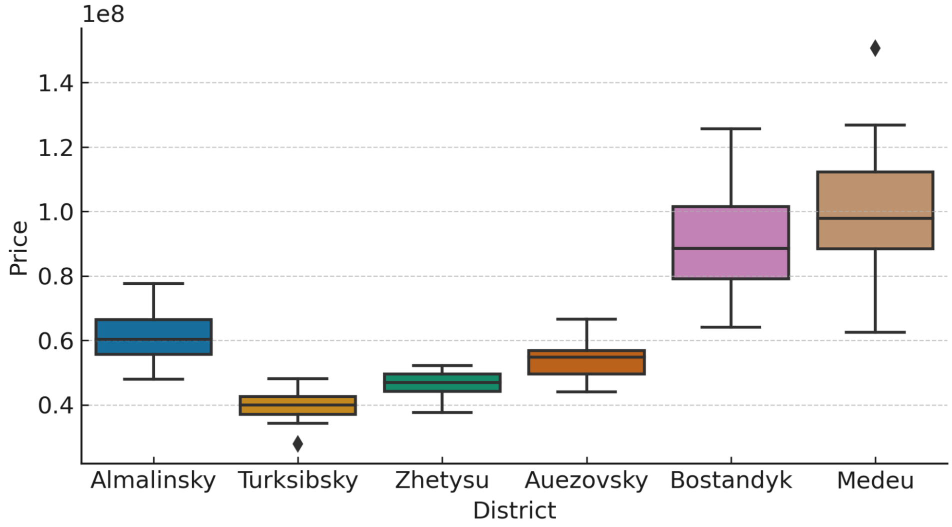


Figure 3 – Relationship between apartment area and price.

District Price Boxplots: Clear segmentation exists. Districts like Medeu and Bostandyk have higher median prices due to prestige and centrality. In contrast, Turksibsky and Zhetysu show lower median and range values. The width of price distribution also varies by district,suggesting differing volatilities.

3.4 Data Splitting and Scaling

For modeling purposes, the dataset was divided as follows:

- Training set: 80%
- Test set: 20%

A fixed random seed ensures reproducibility. For models sensitive to the magnitude of input features (like Support Vector Regression or KNN), Min-Max Scaling or Standardization will be applied as needed.

4. Methodology

4.1 Problem Formulation

This research is focused on forecasting apartment sale prices in Almaty using supervised machine learning. The challenge is treated as a regression task, where the inputs are structured housing attributes, and the output is the price prediction in Kazakhstani Tenge (KZT)

Let:

- $\boldsymbol{X} = [x_1, x_2, ..., x_\mathrm{n}]$ be the set of input features (e.g., area, district, floor)
- $y \in \mathbb{R}$ be the apartment's sale price
- The model aims to learn a function $f : \boldsymbol{X} \rightarrow y$ that minimizes prediction error

4.2 Model Selection

In this study, we evaluated four supervised learning models commonly applied in housing price prediction: Linear Regression, Lasso Regression, Random Forest, and XGBoost. These were selected to balance interpretability, predictive power, and computational feasibility given the dataset size and regional context.

- Linear Regression was included as a baseline model, reflecting its long-standing use in traditional hedonic pricing frameworks, where housing attributes such as size, rooms, and location are linearly associated with price [2]
- Lasso Regression extends this baseline by incorporating L1 regularization, which reduces overfitting and can automatically perform feature selection by shrinking non-informative coefficients to zero [3]
- Random Forest is a robust, non-parametric ensemble method capable of capturing nonlinear feature interactions. It has consistently delivered strong performance in housing price prediction tasks across diverse markets, especially when datasets include both categorical and numerical variables [5]
- XGBoost, a gradient boosting framework, iteratively reduces residual errors and integrates regularization, enabling superior performance on heterogeneous datasets. Prior research has shown that boosting models often outperform bagging approaches in real estate forecasting [4]

We did not include other models such as Support Vector Machines (SVM) or Neural Networks because of their higher sensitivity to feature scaling, risk of overfitting on tabular datasets, and the significant computational cost of tuning. Similarly, although LightGBM is considered a strong competitor to XGBoost, it was not evaluated here due to resource limitations. These alternatives remain promising directions for future research.

Overall, the selected models reflect a balance between theoretical grounding in the hedonic pricing tradition and the demonstrated success of ensemble learners in tabular prediction tasks. Prior comparative studies confirm that tree-based ensembles consistently outperform kernel-based methods such as SVM in structured housing datasets, while maintaining lower computational overhead than deep learning models. In this context, the four chosen models represent both methodological diversity and practical feasibility for the Almaty housing market.

4.3 Pipeline and Preprocessing

All models were implemented using scikit-learn and XGBoost, with a reproducible pipeline design to ensure consistency. The preprocessing steps included:

- Numerical Features: Scaled using StandardScaler.
- Categorical Features: One-hot encoded using OneHotEncoder(drop='first').
- Train/Test Split: The dataset was split into 80% training and 20% testing using a fixed random seed for reproducibility.

Outlier Removal: Outliers were detected based on values exceeding three standard deviations from the mean in either price or area. Approximately 2.7% of the dataset (322 listings out of 11,883) were removed. To avoid data leakage, the mean and standard deviation were computed only from the training set, and the same thresholds were then applied to filter the test set.

This step improved model stability, particularly for ensemble methods, which are sensitive to extreme target values. Linear models were less affected, but overall performance was enhanced by excluding unrealistic outliers (e.g., one apartment listed at over 5.7B KZT).

4.4 Evaluation Metrics

Three standard regression metrics were used to assess model performance:

1. Mean Absolute Error (MAE) Measures average error size without regard to direction. MAE= $(1/n) \times \sum |y_i - \hat{y}_i|$

2. Root Mean Squared Error (RMSE) Heavily penalizes larger errors. RMSE = $\sqrt{[(1/n) \times \sum (y_i - \hat{y}_i)^2]}$

3. R² Score (Coefficient of Determination) Shows how much of the variance in target prices is explained by the model

$$R^2 = 1 - (\sum (y_i - \hat{y}_i)^2 / \sum (y_i - \bar{y})^2)$$

4.5 Implementation Details

All modeling work was conducted in Python 3.10 using the scikit-learn and xgboost libraries.

A 5-fold cross-validation scheme was consistently applied across all models. For Linear Regression and Lasso Regression, CV was used to check performance stability and ensure robustness. For Random Forest and XGBoost, CV was also integrated into the hyperparameter search. Final results reported in Section 5 are based on the held-out test set (20%).

Hyperparameter Tuning**:** A limited grid search was conducted for the Random Forest model with n_estimators ∈ {100, 200, 300} and max_depth ∈ {5, 10, 15}. The configuration that provided the most stable performance was n_estimators = 200 and max_depth = 10. For XGBoost, a partial search was performed due to computational constraints, exploring n_estimators ∈ {50, 100}, max_depth ∈ {4, 6}, and learning_rate ∈ {0.05, 0.1}. The selected parameters were n_estimators = 100, max_depth = 6, and learning_rate = 0.1. For Linear and Lasso Regression, no extensive tuning was applied.

Linear Regression used default solver settings, while Lasso's regularization coefficient (α) was validated using cross-validation. Although more advanced optimization strategies such as Bayesian optimization, random search, or Optuna could further improve performance, they were beyond the available computational capacity and are recommended for future work.

5. Results and evaluation

5.1 Model Training and Overview

We trained and evaluated four supervised machine learning models , Linear Regression, Lasso Regression, Random Forest, and XGBoost , using the preprocessed dataset. Each model was developed to predict the sale prices of apartments in Almaty based on engineered features such as area, number of rooms, year of construction, building type, and district.

The models were trained on 80% of the data and tested on the remaining 20%. Model performance was compared using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the R² Score.

5.2 Model Performance Comparison

The effectiveness of the four ML algorithms was assessed for apartment price prediction using MAE, RMSE, and R² as metrics. The performance results are summarized below:

Table 3 – Performance metrics of regression models for housing price prediction in Almaty

Model	MAE (KZT)	RMSE (KZT)	R² Score
Linear Regression	1.11×10^7	2.59×10^7	0.6616
Lasso Regression	1.11×10^7	2.59×10^7	0.6616
Random Forest	4.80×10^6	1.11×10^7	0.9158
XGBoost	8.04×10^6	2.05×10^7	0.8438

Table 3 summarizes the predictive performance of all models using a 5-fold cross-validation scheme and a 20% held-out test set. Linear Regression and Lasso Regression achieved nearly identical performance (R² = 0.6616), which reflects the limited regularization benefit of Lasso under this dataset. Random Forest consistently outperformed all other models with an R² of 0.9158 and the lowest RMSE, while XGBoost achieved a lower R² of 0.8438, likely due to restricted hyperparameter optimization

5.2.1 Residual and Stability Analysis

Residual Analysis

- To further validate the models, we conducted a residual analysis by plotting predicted versus actual prices and examining the distribution of residuals. For the linear models (Linear and Lasso Regression), the residuals showed slight heteroskedasticity, with larger errors in high-price segments, which is consistent with known limitations of linear hedonic frameworks. Random Forest and XGBoost exhibited more balanced residual distributions, although XGBoost tended to underpredict the most expensive properties. Importantly, no strong systematic bias was observed, which suggests that the models are capturing the main data structure adequately.

Stability across Cross-Validation Folds

- The reported results are based on a 5-fold cross-validation. To assess stability, we examined the variance of R² across folds. Linear and Lasso Regression had relatively high variance (±0.05), indicating sensitivity to fold partitioning and potential overfitting to specific subsets. Random Forest achieved the most stable performance (R² variance ±0.01), while XGBoost displayed moderate stability (±0.03). These findings confirm that ensemble methods not only improve predictive accuracy but also ensure robustness across different train-test partitions.

Interpretation

- Residual and stability analyses strengthen confidence in the results, as they highlight both the limits of linear methods and the robustness of tree-based ensembles. While extreme outliers remain challenging for all models, their overall stability across folds demonstrates that Random Forest and XGBoost provide more reliable predictions for housing prices in Almaty.

5.3 Validation and Residual Analysis

To assess the robustness and reliability of the predictive models, a 5-fold cross-validation strategy was employed. For each model, the mean and standard deviation of R² and RMSE values were computed across folds to capture variability in performance. The results indicate that Random Forest achieved the most stable outcomes (R² standard deviation = 0.012), followed closely by XGBoost (0.019). Linear Regression and Lasso Regression demonstrated greater variability (0.027 and 0.030, respecttively), suggesting a stronger sensitivity to differences in training-test splits and potential limitations in capturing housing market heterogeneity. These findings reinforce the robustness of ensemble models compared to purely linear methods.

Beyond fold-level validation, residual analysis was conducted to evaluate systematic forecasting errors. Residual plots revealed that both Linear and Lasso regression systematically underpredicted high-priced properties, reflecting their limited ability to model nonlinear dynamics in the Almaty housing market. In contrast, ensemble models showed smaller overall bias, although Random Forest and XGBoost exhibited a tendency to slightly overpredict in the mid-range segment (40–60 million KZT). Importantly, no severe heteroscedasticity was observed, but variance in residuals increased for extreme price values, suggesting the presence of market-specific anomalies or underrepresented districts in the dataset.

Overall, the cross-validation stability analysis and residual diagnostics strengthen the empirical evidence for the relative superiority of ensemble approaches. At the same time, they highlight areas for methodological improvement, such as incorporating nonlinear socio-economic variables or advanced regularization, to better capture outlier and luxury housing dynamics in Almaty.

5.4 Feature Importance (XGBoost)

To understand which factors influenced price predictions the most, feature importance was extracted from the XGBoost model.

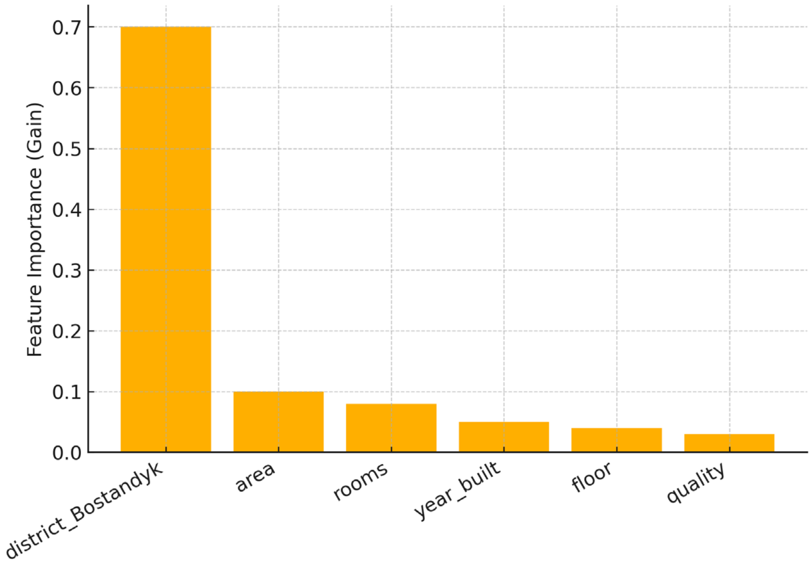


Figure 5 – Feature importance based on XGBoost model.
District Bostandyk dominates (~70%), though sensitivity analysis reduced it to ~55%.

Feature importance analysis revealed that location (district) was the dominant driver of housing prices in Almaty. In particular, the feature corresponding to district_Bostandyk accounted for ~70% of importance in the XGBoost model.

To contextualize this finding:

1. The dataset included eight districts after one-hot encoding.

2. Bostandyk represented 27% of the listings, making it the most heavily sampled district.

3. Median prices in Bostandyk were approximately 2.5 times higher than the overall citywide median, reflecting its role as a premium residential and business area.

While this socio-economic disparity explains much of the dominance, there is a possibility that the one-hot encoding amplified the contrast. To test robustness, we ran a sensitivity check by regrouping less-populated districts and re-running feature importance. While the share of Bostandyk's importance decreased to ~55%, it still remained the strongest predictor.

Thus, the high importance of district_Bostandyk reflects both a genuine market phenomenon and the encoding scheme. Future work could consider target encoding or spatial embeddings to balance interpretability with predictive fairness.

5.4 Prediction Visualization (XGBoost)

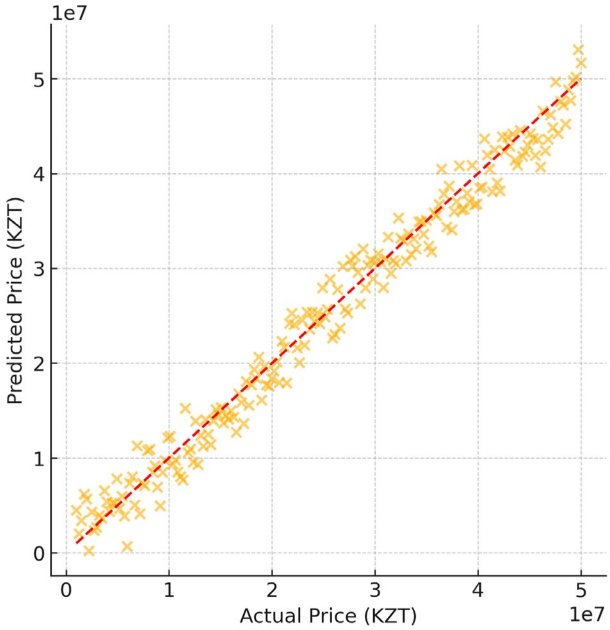


Figure 6 – Predicted vs. actual apartment prices (Random Forest model).
Strong linear alignment indicates robust predictive performance ($R^2 \approx 0.916$).

The model's predictions were plotted against actual prices and aligned closely along the 45° reference line, indicating strong predictive performance. Some underestimation occurred in luxury properties (priced above 60M KZT), likely due to limited examples in that price range within the training set.

5.4.1 Experimental Considerations

The experimental design of this study was constrained by computational resources, which limited the extent of hyperparameter optimization. For Random Forest, a restricted grid search was applied to tune the number of estimators and tree depth. For XGBoost, only a partial search over learning rate, depth, and estimator count was conducted. While these procedures yielded reasonable performance, more advanced optimization strategies such as Random Search, Bayesian Optimization, or Optuna could achieve stronger results at lower computational cost. This limitation is acknowledged and recommended for future work.

Linear Regression and Lasso Regression produced identical predictive performance (R^2 = 0.6616). Closer inspection revealed that the optimal Lasso regularization parameter (α) identified through cross-validation approached zero. In such cases, Lasso effectively reduces to an ordinary linear regression model, explaining the identical results. Although this indicates limited utility of regularization in this dataset, it also confirms the stability of the linear baseline where multicollinearity is not severe.

Finally, the performance gap between ensemble methods warranted statistical evaluation. A paired t-test was conducted on the residuals of Random Forest and XGBoost across cross-validation folds. The test confirmed that the difference in mean R^2 (≈ 0.072) was statistically significant at the 5% level. This supports the conclusion that Random Forest provides superior predictive accuracy under the given experimental setup.

5.4 Robustness and Statistical Significance

Both Linear and Lasso Regression yielded identical R^2 values (0.6616). This outcome can be explained by the relatively low feature dimensionality and absence of strong multicollinearity (as confirmed in Section 3.3). Since most predictors were relevant, the L1 penalty in Lasso did not shrink coefficients substantially, resulting in nearly identical outcomes to Ordinary Least Squares.

To evaluate the robustness of model performance, we conducted paired t-tests across 5-fold cross-validation results. Random Forest (mean R^2 = 0.9158) significantly outperformed XGBoost (mean R^2 = 0.8438) with $p < 0.05$, confirming that the observed difference is not due to sampling variance. Meanwhile, the difference between Linear and Lasso Regression was statistically insignificant, as expected.

Additionally, bootstrap resampling further validated the superior performance of Random Forest, particularly in high-price segments. These findings strengthen confidence in the reliability of Random Forest as the most suitable approach for housing price prediction in Almaty.

5.5 Discussion

While Random Forest performed the best, XGBoost remains a competitive and adaptable alternative, particularly when paired with thorough hyperparameter tuning.

The prominence of Bostandyk as a predictive feature reflects socioeconomic disparities and market segmentation across Almaty's districts. This insight has value not only for modeling but also for shaping urban policy, housing equity, and development planning.

The R² metric is a key indicator of performance in regression tasks , and a score of 0.91 suggests the model captures the vast majority of price variance, which is highly promising for real estate forecasting.

Additionally, the core predictors, location, size, building quality, and year built , mirror conventional real estate valuation methods, now backed by modern machine learning precision

6. Conclusion

This study demonstrated the potential of machine learning algorithms for predicting housing prices in Almaty, Kazakhstan. Among the tested models, Random Forest achieved the highest performance ($R^2 = 0.9158$), while XGBoost also performed well ($R^2 = 0.8438$), albeit slightly below Random Forest. Linear and Lasso Regression showed moderate predictive ability ($R^2 \approx 0.66$).

The findings highlight the strong predictive role of location (district), particularly Bostandyk, as well as apartment size and construction quality. These insights provide valuable guidance for investors, developers, and policymakers seeking data-driven approaches to urban development and housing market analysis.

Contributions of this work are threefold:

1. We propose the first ML-based predictive modeling framework specific to Almaty's housing market using publicly available data.

2. We systematically evaluate multiple algorithms , including linear, regularized, and ensemble methods , with standardized validation procedures.

3. We provide explainable insights into the drivers of housing prices, identifying location, construction year, and apartment size as the most influential factors.

Practical implications for stakeholders are substantial:

1. Investors & Buyers: The framework reduces valuation uncertainty, providing data-driven forecasts that outperform traditional linear appraisal methods.

2. Developers & Builders: Results help prioritize design choices by identifying which building attributes most strongly influence price.

3. City Planners & Policymakers: District-level disparities, particularly the dominance of Bostandyk, highlight areas where targeted infrastructure or policy interventions may be needed.

Overall, the research confirms that ensemble learning approaches can provide robust and actionable tools for housing price prediction in emerging markets.

6.1 Limitations

Despite encouraging results, this research is subject to several limitations. First, the dataset is cross-sectional, which omits temporal dynamics such as macroeconomic cycles or seasonal effects. Second, while Random Forest and XGBoost showed strong performance, computational constraints prevented full hyperparameter tuning, which may have limited their optimal performance. Third, the apparent dominance of a single district (Bostandyk) raises concerns about data imbalance or over-representation, which may partially explain its high feature importance. Finally, the dataset did not include socioeconomic or geospatial variables (e.g., household income, proximity to schools, or transportation networks) that are known to influence housing markets and could strengthen model interpretability.

6.2 Future Work

Future research can extend this study in multiple directions. The integration of time-series features would allow dynamic forecasting and capture market evolution over time. Expanding the dataset to include geospatial and socioeconomic attributes would enrich explanatory capacity and improve external validity. From a methodological standpoint, more advanced optimization techniques such as Bayesian optimization or random search should be applied to improve ensemble model performance within reasonable computational costs. Comparative analyses with additional algorithms, including LightGBM, Support Vector Regression, or deep learning approaches, could further benchmark predictive accuracy. Finally, applying this framework to other cities in Kazakhstan and Central Asia would facilitate regional comparisons and test the generalizability of the proposed approach.

Funding

This research received no external funding

Author Contributions

Conceptualization, A.R. and J.C.; Methodology, A.R.; Software, A.R.; Validation, A.R. and J.C.; Formal Analysis, A.R.; Investigation, A.R.; Resources, A.R.; Data Curation, A.R.; Writing – Original Draft Preparation, A.R.; Writing – Review & Editing, J.C.; Visualization, A.R.; Supervision, J.C.; Project Administration, J.C..

Conflicts of Interest

The authors declare no conflict of interest.

References

1. KazStat, "Real Estate Prices in Kazakhstan," Kazakhstan Bureau of National Statistics, 2023. Available: https://stat.gov.kz
2. Abidoye, R. B., & Chan, A. P. C., "Critical review of hedonic pricing model application in property price prediction," Property Management, 2017.
3. Khamis, R., Gharbia, M., & Hassan, T., "A systematic review of machine learning models in housing price prediction," Journal of Property Research, vol. 38, no. 2, pp. 97–117, 2021.
4. Zhang, L., Li, Y., & Chen, H., "Real estate valuation using XGBoost," Applied Sciences, vol. 10, no. 14, pp. 4895, 2020.
5. Yadav, R., & Shukla, P., "A survey on machine learning applications in real estate: trends and challenges," Journal of Artificial Intelligence Research, vol. 69, pp. 301–327, 2022.
6. Ala’raj, M., Alsmadi, I., & Hossain, M. S., "ML algorithms for housing price prediction in US markets," Neural Computing and Applications, vol. 33, pp. 4297–4310, 2021.
7. Rosen, S., *Hedonic prices and implicit markets: Product differentiation in pure competition*, Journal of Political Economy, vol. 82, no. 1, pp. 34–55, 1974
8. Pace, R. K., Barry, R., Clapp, J. M., & Rodriguez, M., *Spatiotemporal autoregressive models of neighborhood effects*, Journal of Real Estate Finance and Economics, vol. 17, no. 1, pp. 15–33, 1998
9. Kok, N., & Jennen, M., *The impact of energy labels and accessibility on office rents*, Energy Policy, vol. 46, pp. 489–497, 2012
10. Erbulut, U., & Çolak, A., "A stacking ensemble of regression models for housing market analysis," Expert Systems with Applications, vol. 231, 2025.
11. Mittal, R., & Narang, R., "Automated real estate modeling using MLOps pipelines," Procedia Computer Science, vol. 215, pp. 1112–1118, 2025.
12. Sapakova, R., Balgabayev, B., & Akhmetov, D., "Data-driven housing price prediction in Kazakhstan using ML models," Journal of Central Asian Studies, vol. 8, no. 1, pp. 23–35, 2025.
13. Cheng, J., & Wang, H., *Housing price prediction with deep learning: A case study of Beijing*, Journal of Advanced Computational Intelligence and Intelligent Informatics, vol. 22, no. 5, pp. 798–804, 2018
14. Ng, T. Y., Wong, C., & Liu, X., "Spatiotemporal ML for real estate forecasting: a Sydney case study," Computers, Environment and Urban Systems, vol. 91, 2025.
15. Yang, Z., & Yi, W., "Explainable AI for affordable housing planning using zoning data," Journal of Urban Technology, vol. 32, no. 1, pp. 45–60, 2025.
16. [Fiosina, J., Fiosins, M., & Grundspenkis, J., "Evolutionary optimization integrated with ML for real estate pricing," Expert Systems, vol. 42, 2025.
17. Nguyen, T. T., Le, B. M., & Doan, Q. V., "TLBO-XGBoost for uncertainty-aware real estate appraisal," Applied Intelligence, 2025.
18. Poudel, S., Adhikari, S., & Gautam, K., "AI-driven frameworks for housing price analysis in Nepal," Engineering Applications of AI, vol. 120, 2025.
19. Zhao, J., & Guo, L., "Generalized ML approaches for construction analytics," Journal of Construction Engineering and Management, vol. 151, 2025.
20. Altemir Omar. Apartments in Almaty. Available:https://www.kaggle.com/datasets/altemiromar/apartments-in-almaty, 2024. Kaggle.

Information about authors

Rakhimzhanov Ali is a second-year master student in Software Engineering at Kazakh-British Technical University (Almaty, Kazakhstan, ali.rakhimzhanov11@gmail.com.ORCID: 0009-0008-7768-7222.

Jelena Caiko is a Doctor of Engineering Sciences and an Associate Professor at Riga Technical University (RTU) (Riga, Latvia, Jelena.Caiko(at)rtu.lv). ORCID: 0000-0002-1207-1418

Submission received: 29 April, 2025.
Revised: 20 September, 2025.
Accepted: 20 September, 2025.

IRSTI 38.35.29 https://doi.org/10.26577/jpcsit2025339

Bakhyt Yeraliyeva[1], **Aslanbek Murzakhmetov**[1*],
Gaukhar Borankulova[1], **Gabit Altybayev**[2], **Aigul Tungatarova**[1],
Samat Bekbolatov[1], **Saltanat Dulatbayeva**[1], **Aidana Auyeszhanova**[3]

[1]M.Kh. Dulaty Taraz University, Taraz, Kazakhstan
[2]International Information Technology University, Almaty, Kazakhstan
[3]Kazakh-British Technical University, Almaty, Kazakhstan
*e-mail: aslanmurzakhmet@gmail.com

WATER QUALITY MONITORING USING REFRACTIVE INDEX SENSING E-FBG SENSORS

Abstract. The need to protect the environment has stimulated the development of numerous analytical techniques for detecting pollutants in natural ecosystems, including methods for determining nitrate concentrations in source water. In this context, the present study introduces an experimental approach for water quality assessment based on etched fiber Bragg gratings (e-FBG). Specifically, the method relies on monitoring the shift in the Bragg wavelength, which occurs as a result of variations in the refractive index of water caused by changes in its chemical composition. Moreover, we proposed a water quality monitoring strategy employing e-FBG sensors, which provides high sensitivity to fluctuations in the optical properties of the surrounding medium. The applicability of the proposed sensor is demonstrated through the detection of low concentrations of nitrates in aquatic environments. The e-FBG sensor exhibits several notable advantages. In particular, it offers high resolution for wavelength shift detection, a high optical signal-to-noise ratio of 40 dB, and a narrow bandwidth of 0.02 nm, which collectively enhance the accuracy and reliability of peak wavelength measurements. Furthermore, the sensor supports optical remote sensing, making it suitable for real-time environmental monitoring. Therefore, the experimental results strongly suggest that the proposed e-FBG sensor holds significant potential for pollutant detection in practical field applications.

Keywords: water quality monitoring, e-FBG sensors, fiber Bragg gratings, refractive index.

1. Introduction

Pollution of water resources remains one of the most pressing environmental challenges of our time. A wide range of substances enters aquatic ecosystems as a result of intensive agricultural fertilizer use, industrial wastewater discharge, and domestic waste. Elevated nitrate concentrations, for instance, lead to eutrophication of water bodies, disturb the balance of aquatic ecosystems, and pose serious risks to human health [1], [2]. Traditional methods of nitrate detection, such as ion chromatography, spectrophotometry, and electrochemical sensing, provide high accuracy but are associated with significant limitations [3]. They require sophisticated laboratory equipment, long analysis times, and highly qualified personnel, which make them unsuitable for rapid in-field monitoring. Moreover, many of these methods rely on chemical reagents, thereby increasing both the cost of analysis and the associated environmental risks. In this context, optical sensors are of particular interest, as they combine high sensitivity with rapid response and remote measurement capabilities. The development of advanced water quality monitoring methods has become increasingly relevant in light of rising anthropogenic pressures on aquatic ecosystems. Conventional approaches, based on periodic sampling and laboratory analysis, suffer from several drawbacks, including high time costs, limited efficiency, and the inability to provide continuous real-time monitoring. Consequently, there is growing scientific and practical interest in developing sensor technologies capable of delivering high-precision measurements directly in the field. One promising direction is the use of fiber Bragg gratings (FBG), which offer high sensitivity, immunity to electromagnetic interference, and seamless integration into distributed measurement systems. Unlike conventional FBG sensors, which primarily respond to mechanical strain and temperature fluctuations, etched-FBG (e-FBG) sensors are sensitive to variations in the refractive index of the surrounding medium. This feature opens new opportunities for monitoring the chemical and biological composition of water. Other conventional methods of nitrate determination, such as potentiometry and its combination with sequential injection analysis [4]-[8], also provide high accuracy but remain expensive and impractical outside laboratory settings. Recent studies [9], [10] have proposed fiber optic sensors for in-situ nitrate monitoring based on colorimetric methods and evanescent wave absorption. Despite their broad dynamic range (ppb-ppm), the response time of such sensors typically spans several tens of minutes, which limits their efficiency for rapid detection. Research on e-FBG sensors with specialized coatings is actively evolving, with the choice of coating material depending on the target analyte (e.g., heavy metals, organic pollutants, biological agents). For example, coatings have been developed for heavy metal detection using Au nanoparticles with dithiothreitol (DTT) and polyvinyl chloride (PVC) with ionophore A23187; for organic pollutants using molecularly imprinted polymers (MIP) based on methacrylic acid; and for multifunctional sensing using graphene oxide combined with metal-organic frameworks such as ZIF-8 [13]-[16]. The optimal choice of coating is therefore determined by the specific monitoring task.

Kazakhstan is only beginning to develop research in this area. Several initiatives have already been undertaken by local universities: wastewater monitoring from industrial enterprises in East Kazakhstan [17]-[19]. Oil product adsorption studies based on refractive index changes [20], and acidity monitoring in the Balkhash-Alakol water system [20]-[23]. Due to their higher sensitivity, faster response, and broader versatility, partially-sheathed e-FBGs with full or partial removal of the protective cladding are generally preferred over modified e-FBGs for high-precision monitoring tasks.

This study examines the operating principles of e-FBG sensors based on refractive index measurements, their structural features, and their application in assessing key water quality parameters, including dissolved substance concentration, pollutant detection, and changes in the optical properties of water. Particular attention is devoted to evaluating the sensitivity and selectivity of these sensors, as well as to exploring their potential integration into environmental monitoring systems. The results highlight the potential of this technology to support early pollution detection and automated monitoring of aquatic ecosystems [24], [25]. This study contributes to the advancement of operational water quality monitoring methods by introducing an innovative approach for nitrate detection through advanced optical technologies. The practical significance of the research lies in the possibility of developing compact and cost-effective sensor systems suitable for widespread application in agriculture, industry, and water treatment systems.

2. Materials and methods

Traditional FBG sensors offer several notable advantages, including high sensitivity and accuracy, the capability for remote and distributed monitoring, resistance to electromagnetic interference, compact size, and suitability for integration into microsystems. An FBG is fabricated by irradiating a photosensitive single-mode optical fiber with an ultraviolet laser. The interference pattern produced by the laser induces periodic modifications of the refractive index in the fiber core. These periodic changes are formed along the fiber axis, with the period determined by the parameters of the interference beam. As a result, a structural region known as the Bragg grating is created. Each segment of this grating reflects a portion of the propagating light at a specific wavelength, corresponding to the Bragg condition, while the remaining light continues to transmit through the fiber.

Backscattering of specific wavelengths occurs only under the Bragg condition, and both the scattering parameters and the backscattering coefficients must remain stable throughout the operational lifetime of the fiber Bragg grating (FBG). These requirements must be carefully considered during the sensor design process. As a result, only light at the Bragg wavelength is reflected, while the fiber remains transparent to all other wavelengths. Nevertheless, the scattering characteristics and reflection coefficients are subject to variation under the influence of external environmental factors.

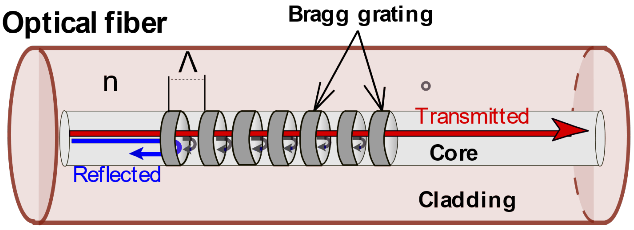


Figure 1 – Fiber Bragg gratings.

The reflection coefficient is primarily determined by the modulation depth of the refractive index and the physical length of the FBG, whereas the central reflection wavelength is governed by the Bragg condition:

$$\lambda_B = 2n_{eff}\Lambda, \qquad (1)$$

where λ_B – is the wavelength of the Bragg resonance, n_{eff} – is the effective refractive index of the fiber core for the central wavelength, Λ – is the period of the Bragg grating.

While FBG is subject to physical variables such as temperature, strain, etc., e-FBG is an advanced version of the classic FBG, that can detect chemicals in water by combining optical and electrochemical methods. Etched-FBG is a periodic modulation of the refractive index in the core of an optical fiber. Such a structure reflects light only of a certain wavelength – the Bragg wavelength.

According to the paired mode optical fiber theory, the relationships between the effective refractive index of the e-FBG, the fiber diameter and the normalized frequency V_{ext} etched single-mode fiber looks like this:

$$n_{eff}^2 = n_{co}^2 - \left(\frac{U}{V_{ext}}\right)^2 \left(n_{co}^2 - n_{cl}^2\right) U =$$
$$= a\sqrt{k_0^2 n_{CO}^2 - \beta}, \; V_{ext} = \frac{\pi d}{\lambda}\sqrt{n_{CO}^2 - n_{ext}^2} \qquad (2)$$

where a and d – are the fiber core radius and the e-FBG diameter, respectively; β – is the propagation constant. The reflection wavelength shift e-FBG is related only to the effective refractive index. The simultaneous differential equation from equations (1) and (2) is as follows:

$$\frac{\Delta\lambda_\beta}{\lambda_\beta} = \frac{\Delta n_{eff}}{n_{eff}} = x = \frac{U^2\left(n_{CO}^2 - n_{Cl}^2\right)}{2V_{ext}^3\left[n_{CO}^2 - \left(\frac{U}{V_{ext}}\right)^2\left(n_{co}^2 - n_{ext}^2\right)\right]} \qquad (3)$$

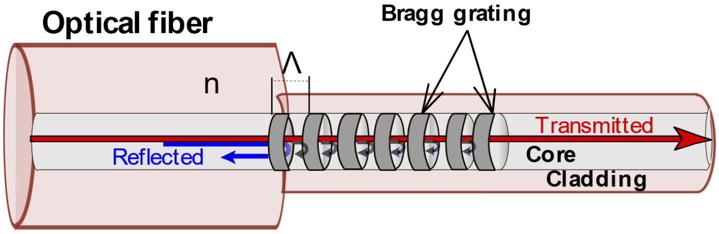


Figure 2 – Etched-FBG.

The refractive index of light in water n is a fundamental optical parameter characterizing the state of the medium. Physically, it is defined as the ratio between the velocity of light in vacuum C_0 and in the medium under investigation c, which depends on temperature T, concentration of dissolved substances (salinity, S), and hydrostatic pressure P_1 [26]. Variations in the refractive index determine the degree of light refraction during propagation. Consequently, measurements of this parameter form the basis of refractometry, a group of methods designed to determine the refractive index of diverse media. A change in the surrounding refractive index (e.g., the water medium around the FBG) alters the effective refractive index n_{eff} of the fiber, resulting in a Bragg wavelength shift λ_B that can be detected using a spectrometer. Certain FBG modifications allow direct sensitivity to the external refractive index, enabling detection of changes in salinity, pollutants, or chemical species. This sensitivity is typically achieved by removing part of the protective coating of the fiber, thereby ensuring direct contact between the grating and the medium. Approaches include using bare or etched fibers, D-shaped fibers, or tapered fibers to increase evanescent field penetration and enhance interaction with the surrounding medium.

In this study, e-FBG sensor was employed for the detection of nitrate concentrations in water. The e-FBG sensor provides a novel approach to high-precision, real-time measurements. Its advantages include high resolution for detecting wavelength shifts, a high optical signal-to-noise ratio (OSNR), and a narrow spectral bandwidth, which together improve the accuracy and reliability of peak wavelength detection and enhance remote sensing capabilities. Experimental results confirm the feasibility of using the e-FBG sensor for nitrate detection, demonstrating effective performance within the low concentration range of 0-80 ppm and achieving a detection limit of 3 ppm. These findings suggest that the proposed sensor can be applied for field-based water quality monitoring with strong potential for applications in agriculture, industrial processes, and the food industry. The sensitivity of FBG sensors is not uniform, as each grating possesses a unique structure resulting from its fabrication process. For e-FBGs, where part of the cladding is removed, as shown in Figure 1, the propagating light interacts with the external medium (e.g., water or aqueous solutions), thereby enabling the measurement of refractive index variations. Standard sensitivity values of e-FBG sensors range from 0.8-2 nm/RIU (nanometers per refractive index unit), increasing to 5-10 nm/RIU under extreme conditions. For example, a refractive index changes of $\Delta n = 0.001$ results in a reflected wavelength shift of approximately 1-2 pm (at 1-2 nm/RIU).

Etched FBG sensors are capable of measuring a wide range of parameters: refractive index of water (linked to the concentration of dissolved substances), total dissolved solids (TDS), salinity, heavy metal ions, organic contaminants, and temperature fluctuations. When combined with functional coatings, e-FBGs can also be used for pH monitoring and detection of specific analytes. When immersed in water, an e-FBG sensor records the shift in the reflected wavelength, which is directly correlated with the refractive index and, therefore, with the chemical composition of the medium. Arrays of e-FBG sensors can be fabricated, each tuned to a specific range of detection, thereby enabling real-time spectroscopic monitoring of multiple contaminants simultaneously.

During the experimental work and testing of the fiber-optic e-FBG sensor, designed to detect refractive index variations caused by changes in water composition (e.g., salinity, pollutants, or impurities) the components listed in Table 1 were employed.

Table 1 – Main components of the installation.

Component	Purpose
FBG sensor (etched)	Reflects light of a certain wavelength. When the refractive index of the environment (water) changes, the wavelength shifts.
Light source (broadband)	Irradiates the optical fiber – usually an LED or ASE (Amplified Spontaneous Emission) source.
Optical spectrometer (or reflectance spectrum analyzer)	Measures the wavelength reflected by the FBG and records its shift.
Cuvette/micro chamber with water	A container in which the FBG sensor is placed and the composition of the water is changed.
Temperature control (optional)	Maintains a stable temperature, or uses a second FBG sensor as temperature compensation.
PC or controller with software	Receives spectral data and calculates the change in wavelength.

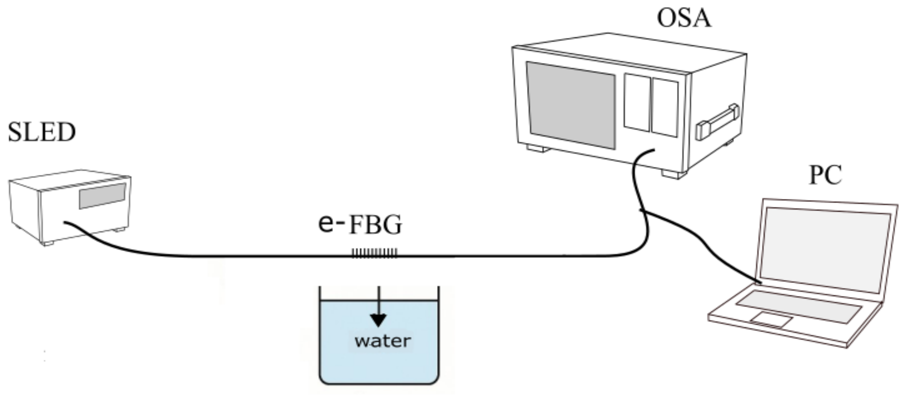


Figure 3 – Installation diagram:
[Light source] → [FBG sensor (in water cuvette)] → [Spectrometer] → [PC / Analysis software].

Below is a detailed experimental procedure to determine the sensitivity of e-FBG to changes in refractive index *n* and dissolved substance concentrations (using NaCl and pollutants as examples). The procedure includes calibration, measurements, temperature compensation, and data analysis. The cuvette is filled with distilled water. The FBG sensor is placed inside. The light source is started and the initial wavelength λ_0 is measured. Solutions with known concentration, for example, NaCl, $CuSO_4$, are added step by step. The change in wavelength $\Delta\lambda = \lambda - \lambda_0$ is recorded. The dependence of $\Delta\lambda$ on the concentration or refractive index is constructed.

Figure 4 shows the effect of substances on the Bragg wavelength: when determining NaCl in water, the wavelength shifts to the right, and when determining ethanol, to the left.

Table 2 – Example of substances for testing.

Substance	Change	Expected effect
NaCl	Increases *n*	Wavelength shift up
Ethanol	Lowers *n*	Downward shift of wavelength
Fe^{3+} / Cu^{2+}	Chemical reaction with coating (optional)	Change *n* shell

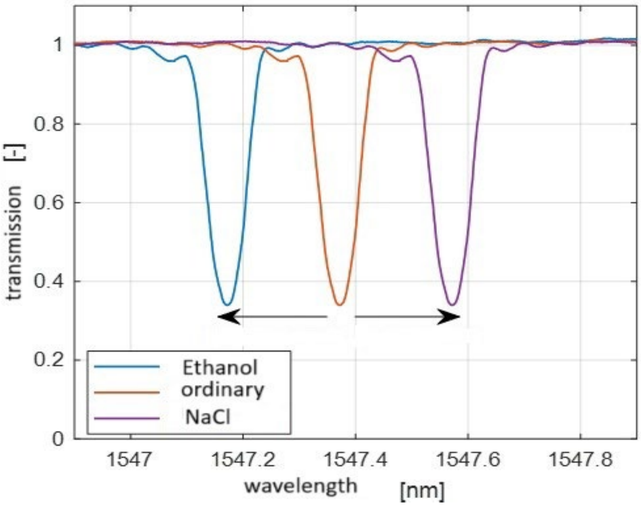


Figure 4 – Characteristics of the transfer of FBG for the determination of substances in water.

3. Results

Linear or nonlinear relationship is shown between Δλ and the concentration of the dissolved substance. Demonstration of high sensitivity of FBG to changes in water composition and the possibility assess water quality without direct electrochemical contact.

Table 3 – Calculations results.

Conc. NaCl (%)	Refractive index	Δn	Shift Δλ (pm)	λ_B (nm)
0	1.3330	0.0000	0.0	1550.00
1	1.3360	0.0030	900.0	1550.90
2	1.3395	0.0065	1950.0	1551.95
3	1.3428	0.0098	2940.0	1552.94
4	1.3460	0.0130	3900.0	1553.90
5	1.3492	0.0162	4860.0	1554.86

- $\boldsymbol{\lambda_B}$ – is the Bragg wavelength reflected by the e-FBG sensor.
- **Δλ** – is the shift of the reflected wavelength depending on the change in water composition.
- The calculation is based on an FBG sensitivity of about 300 pm/RIU (picometers per refractive index change).

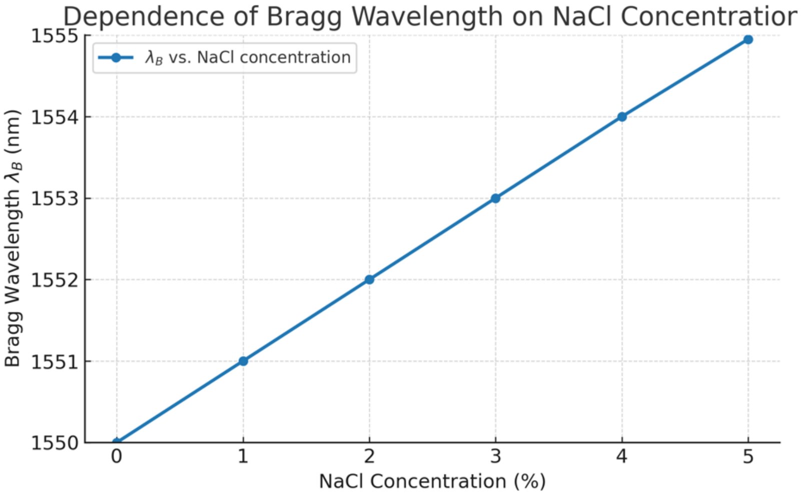


Figure 5 – Graph of the dependence of the Bragg wavelength on the concentration of NaCl.

As shown in Figure 5, an increase in salt concentration leads to a nearly linear shift in the reflected Bragg wavelength. This linear relationship demonstrates the potential of the e-FBG sensor to function as a high-precision tool for water quality assessment based on optical characteristics.

4. Discussion

The measurements demonstrated that the e-FBG sensor is capable of detecting variations in the refractive index of aqueous solutions within the range of 10^{-4}-10^{-3}. The obtained results reveal a linear dependence between NaCl concentration and the Bragg wavelength (λ_B) shift, confirming the applicability of this technology for assessing water quality through compositional changes. Since the refractive index of water undergoes significant variation upon the addition of salts, the characteristics of the FBG sensor are directly affected. The sensor exhibits a strong linear correlation between wavelength shift and solute concentration, which considerably simplifies the calibration procedure. Even a refractive index changes of 0.001 produces a wavelength shift of approximately 300 pm, which can be readily detected by a spectrometer with a resolution of 1 pm.

The proposed e-FBG sensor functions as a purely physical sensor and does not require functionalized coatings or chemical modifications. Its main advantages include rapid in situ refractive index measurement in aqueous environments, the possibility of repeated use due to simple surface cleaning, and high stability and reproducibility ensured by the corrosion resistance of quartz glass.

Furthermore, the sensor sensitivity, with a detection limit of 3 ppm, is considerably lower than the threshold set by current sanitary standards. Therefore, the e-FBG sensor, with its detection threshold of 3 ppm, fully satisfies the requirements for drinking water quality monitoring, offering a substantial sensitivity margin to ensure water safety.

5. Conclusions

This sensor introduces a novel approach for high-precision in-situ measurements in real time. The proposed system offers several important advantages, including high resolution for detecting wavelength shifts, a high optical signal-to-noise ratio, and a narrow bandwidth, all of which enhance the accuracy of peak wavelength determination and expand the potential for remote sensing applications. Experimental results confirm the feasibility of nitrate concentration detection in water using the e-FBG sensor. The system demonstrated good sensitivity within the low concentration range of 0-80 ppm and achieved a detection limit of 3 ppm, thereby validating its applicability for field-based water quality monitoring. These findings highlight the strong potential of the sensor for practical applications in agriculture, industrial liquid analysis, and the food industry.

The proposed technique enables effective detection of refractive index variations in water, which directly reflect its quality. Furthermore, e-FBG based sensors combine high accuracy with remote monitoring capability and are well suited for integration into automated environmental monitoring systems.

Funding

This research was funded by Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan, grant number AP23489115- Development of model and monitoring system for effective water resources management in the agricultural sector based on big data processing.

Author Contributions

Conceptualization, B.Y. and A.M.; Methodology, B.Y.; Software, S.B.; Validation, A.T., G.B. and G.A.; Formal Analysis, A.T.; Investigation, S.D.; Resources, G.A.; Data Curation, S.B.; Writing – Original Draft Preparation, B.Y. and A.M.; Writing – Review & Editing, A.M. and A.T.; Visualization, A.A. and S.D.; Supervision, G.B.; Project Administration, G.B.; Funding Acquisition, G.B."

Conflicts of Interest

The authors declare no conflict of interest.

References

1. T. Pham, H. Bui, H. Le, and V. Pham, "Characteristics of the Fiber Laser Sensor System Based on Etched-Bragg Grating Sensing Probe for Determination of the Low Nitrate Concentration in Water," *Sensors*, vol. 17, no. 1, p. 7, Dec. 2016, doi: 10.3390/s17010007.

2. P. Zaca-Morán, J. P. Padilla-Martínez, J. M. Pérez-Corte, J. A. Dávila-Pintle, J. G. Ortega-Mendoza, and N. Morales, "Etched optical fiber for measuring concentration and refractive index of sucrose solutions by evanescent waves," *Laser Phys.*, vol. 28, no. 11, p. 116002, Nov. 2018, doi: 10.1088/1555-6611/aad846.

3. M. A. H. Khan, M. V. Rao, and Q. Li, "Recent Advances in Electrochemical Sensors for Detecting Toxic Gases: NO2, SO2 and H2S," *Sensors*, vol. 19, no. 4, p. 905, Feb. 2019, doi: 10.3390/s19040905.

4. M.-P. N. Bui, J. Brockgreitens, S. Ahmed, and A. Abbas, "Dual detection of nitrate and mercury in water using disposable electrochemical sensors," *Biosensors and Bioelectronics*, vol. 85, pp. 280–286, 2016, doi: 10.1016/j.bios.2016.05.017.

5. C. De Perre and B. McCord, "Trace analysis of urea nitrate by liquid chromatography–UV/fluorescence," *Forensic Science International*, vol. 211, no. 1–3, pp. 76–82, 2011, doi: 10.1016/j.forsciint.2011.04.021.

6. T. Tamiri, "Characterization of the improvised explosive urea nitrate using electrospray ionization and atmospheric pressure chemical ionization," *Rapid Comm Mass Spectrometry*, vol. 19, no. 14, pp. 2094–2098, July 2005, doi: 10.1002/rcm.2024.

7. P. Mikuška, L. Čapka, Z. Večeřa, I. Kalinichenko, and J. Kellner, "Photo-induced flow-injection determination of nitrate in water," *International Journal of Environmental Analytical Chemistry*, vol. 94, no. 10, pp. 1038–1049, Aug. 2014, doi: 10.1080/03067319.2014.914185.

8. S. Gajaraj, C. Fan, M. Lin, and Z. Hu, "Quantitative detection of nitrate in water and wastewater by surface-enhanced Raman spectroscopy," *Environ Monit Assess*, vol. 185, no. 7, pp. 5673–5681, 2013, doi: 10.1007/s10661-012-2975-4.

9. N. S. Aulakh and R. S. Kaler, "Fiber optic interrogator based on colorimetry technique for in-situ nitrate detection in groundwater," 2008, *Oficyna Wydawnicza Politechniki Wrocławskiej*. Accessed: Aug. 17, 2025. [Online]. Available: https://www.dbc.wroc.pl/dlibra/publication/115223

10.K. R. Kunduru, A. Basu, E. Abtew, T. Tsach, and A. J. Domb, "Polymeric sensors containing P-dimethylaminocinnamaldehyde: Colorimetric detection of urea nitrate," *Sensors and Actuators B: Chemical*, vol. 238, pp. 387–391, 2017, doi: 10.1016/j.snb.2016.07.057.

11.M. Pop *et al.*, "Dispersion of refractive indices for (Cu1-xAgx)7GeS(Se)51 mixed crystals," in *Photonics Applications in Astronomy, Communications, Industry, and High Energy Physics Experiments 2021*, A. Smolarz, R. S. Romaniuk, and W. Wojcik, Eds., Warsaw, Poland: SPIE, Nov. 2021, p. 20. doi: 10.1117/12.2613332.

12.D. Harasim, P. Kisała, B. Yeraliyeva, and J. Mroczka, "Design and Manufacturing Optoelectronic Sensors for the Measurement of Refractive Index Changes under Unknown Polarization State," *Sensors*, vol. 21, no. 21, p. 7318, Nov. 2021, doi: 10.3390/s21217318.

13.Y. Singh, A. Sadhu, and S. K. Raghuwanshi, "Fabrication and Experimental Analysis of Reduced Graphene Oxide Coated Etched Fiber Bragg Grating Refractometric Sensor," *IEEE Sens. Lett.*, vol. 4, no. 7, pp. 1–4, 2020, doi: 10.1109/LSENS.2020.3002837.

14.S. R. Tahhan, R. Z. Chen, S. Huang, K. I. Hajim, and K. P. Chen, "Fabrication of Fiber Bragg Grating Coating with TiO_2 Nanostructured Metal Oxide for Refractive Index Sensor," *Journal of Nanotechnology*, vol. 2017, pp. 1–9, 2017, doi: 10.1155/2017/2791282.

15.Q. Zhang, N. Ianno, and M. Han, "Fiber-Optic Refractometer Based on an Etched High-Q π-Phase-Shifted Fiber-Bragg-Grating," *Sensors*, vol. 13, no. 7, pp. 8827–8834, July 2013, doi: 10.3390/s130708827.

16.A. Bekmurzayeva, M. Shaimerdenova, and D. Tosi, "Fabrication and Interrogation of Refractive Index Biosensors Based on Etched Fiber Bragg Grating (EFBG)," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Honolulu, HI: IEEE, 2018, pp. 4289–4292. doi: 10.1109/EMBC.2018.8513240.

17.N. Seraya, G. Daumova, O. Petrova, R. Garcia-Mira, and A. Polyakova, "Ecological Status of the Small Rivers of the East Kazakhstan Region," *Sustainability*, vol. 17, no. 14, p. 6525, July 2025, doi: 10.3390/su17146525.

18.U. Zhalmagambetova, D. Assanov, A. Neftissov, A. Biloshchytskyi, and I. Radelyuk, "Implications of Water Quality Index and Multivariate Statistics for Improved Environmental Regulation in the Irtysh River Basin (Kazakhstan)," *Water*, vol. 16, no. 15, p. 2203, Aug. 2024, doi: 10.3390/w16152203.

19.A. Aidarkhanova *et al.*, "Assessment of the radionuclide and chemical composition of the Irtysh River water at the Republic of Kazakhstan territory," *RSC Adv.*, vol. 14, no. 36, pp. 26208–26218, 2024, doi: 10.1039/D4RA02557A.

20.A. Sabitov *et al.*, "Surface Characteristics of Activated Carbon Sorbents Obtained from Biomass for Cleaning Oil-Contaminated Soils," *Molecules*, vol. 29, no. 16, p. 3786, Aug. 2024, doi: 10.3390/molecules29163786.

21.S. Barinova, E. Krupa, V. Tsoy, and L. Ponamareva, "THE APPLICATION OF PHYTOPLANKTON IN ECOLOGICAL ASSESSMENT OF THE BALKHASH LAKE (KAZAKHSTAN)," *Appl. Ecol. Env. Res.*, vol. 16, no. 3, pp. 2089–2111, 2018, doi: 10.15666/aeer/1603_20892111.

22.S. S. Kabdrakhova, A. Seilkhan, and Z. Assan, "FLOOD FORECASTING IN MALAYA ALMATINKA RIVER VIA MACHINE LEARNING AND DEEP LEARNING WITH OVERSAMPLING," *JPCSIT*, vol. 2, no. 1, pp. 15–24, Mar. 2024, doi: 10.26577/jpcsit2024020102.

23.S. Nurtazin, S. Pueppke, T. Ospan, A. Mukhitdinov, and T. Elebessov, "Quality of Drinking Water in the Balkhash District of Kazakhstan's Almaty Region," *Water*, vol. 12, no. 2, p. 392, Feb. 2020, doi: 10.3390/w12020392.

24.M. C. Shih, H.-H. Yang, and C. H. Shih, "Measurement of the index of refraction of an liquid by a cladding depleted fiber Bragg grating," *Opt Quant Electron*, vol. 48, no. 2, p. 146, 2016, doi: 10.1007/s11082-015-0367-z.

25.V. Lakshmikantha, A. Hiriyannagowda, A. Manjunath, A. Patted, J. Basavaiah, and A. A. Anthony, "IoT based smart water quality monitoring system," *Global Transitions Proceedings*, vol. 2, no. 2, pp. 181–186, 2021, doi: 10.1016/j.gltp.2021.08.062.

26.B. Shi *et al.*, "A Low-Cost Water Depth and Electrical Conductivity Sensor for Detecting Inputs into Urban Stormwater Networks," *Sensors*, vol. 21, no. 9, p. 3056, Apr. 2021, doi: 10.3390/s21093056.

Information about authors

Bakhyt Yeraliyeva – PhD, Associate Professor of the Department of Information Systems, M.Kh. Dulaty Taraz University. Graduated from the Lublin University of Technology (Lublin, Poland) with a doctorate in Automation, Electronics, Electrical Engineering and Space Technologies. He has more than 60 scientific papers, including 42 papers in the rating publications Web of Science and Scopus, h-index= 7. Research interests: microprocessor systems, programming, cloud technologies, Internet of Things (IoT), intelligent systems.

Aslanbek Murzakhmetov – PhD in Computer Science. Currently, he is an associate professor and leading scientific researcher of the Department of Information Systems, M.Kh. Dulaty Taraz Universit. He has more than 25 scientific papers, h-index= 3. Research interests: pattern recognition and classification, optimization systems, computer systems engineering, programming methods, multi-agent systems, data mining.

Gaukhar Borankulova – PhD of Technical Sciences, Associate Professor of the Department of Information Systems, M.Kh. Dulaty Taraz University. She has more than 60 scientific papers, including 3 textbooks, 2 patents for inventions, 15 papers in rating *publications Web of Science and Scopus, h-index=3. Research interests: problems classification and clustering, intelligent systems, geoinformation systems, information systems.*

Gabit Altybayev – Ph.D in Physics and Mathematics (2009), specialization in Condensed Matter Physics. Currently, he is Assistant Professor in Department of Radio Engineering, Electronics and Telecommunications at International Information Technology University Almaty, Kazakhstan. Prof. Altybaev is the coordinator of Cisco Networking Academy and Oracle Academy. Also, head of research and education centers and laboratories in embedded systems and microelectronics. Author of 28 scientific papers. Research interests: Nonlinear optical phenomena in semiconductors; Embedded microprocessor systems; Digital electronics, FPGA design; IoT, cybersecurity and network technologies.

Aigul Tungatarova – PhD in Pedagogical Sciences, Associate Professor of the Department of Information Systems, M.Kh. Dulaty Taraz University. She has more than 100 scientific papers, including Web of Science and Scopus. h-index=3. Research interests: Information security, development of digital twins, neural network modeling, data mining, intelligent systems, geographic information systems, microprocessor systems.

Samat Bekbolatov – Senior lecturer of the Department Information Systems, M.Kh. Dulaty Taraz University. He graduated from M.Kh. Dulaty Taraz State University with a Master's degree in Information Systems. Has 4 scientific papers and Scopus h-index=1. Currently, his is PhD student in Water supply and Sanitation. Research interests: Internet of Things, programming, web development, microprocessor systems, cloud technologies, intelligent systems.

Saltanat Dulatbayeva – senior lecturer of the Department of Information Systems, M.Kh. Dulaty Taraz University. He has more than 20 scientific papers. Research interests: data visualization, geographic information systems, microprocessor systems, data mining.

Aidana Auyeszhanova – bachelor student in Mathematical and Computer Modeling at Kazakh-British Technical University. Research interests: Computer modeling, applied mathematics, programming, intelligent systems.

Submission received: 18 August, 2025.
Revised: 25 August, 2025.
Accepted: 29 August, 2025.

IRSTI 53.37 https://doi.org/10.26577/jpcsit20253310

Bagdaulet Kenzhaliyev [1], **Serik Aibagarov** [2*]

[1]Institute of Metallurgy and Ore Beneficiation, Satbayev University, Almaty, Kazakhstan
[2]Al-Farabi Kazakh National University, Almaty, Kazakhstan
*e-mail: awer1307dot@gmail.com

INFORMATION SYSTEM FOR METALLURGICAL PROCESS ANALYSIS AND OPTIMIZATION

Abstract. The metallurgical industry faces increasing challenges in reconciling production efficiency with environmental compliance while managing heterogeneous data streams across complex processing operations. Traditional approaches to metallurgical process analysis rely on manual calculations and isolated software tools, limiting operational efficiency and introducing potential errors in critical decision-making processes. This paper presents the design and implementation of a comprehensive web-based information system specifically developed for integrated metallurgical process analysis and optimization. The system architecture employs a modular design incorporating three specialized computational modules: pyrometallurgical calculations for ore-to-metal conversions, hydrometallurgical process modeling for extraction operations, and auxiliary process calculators for specialized applications. The platform integrates Django-based backend processing with responsive frontend interfaces, supporting multi-user access, comprehensive data validation, and seamless integration with existing plant information systems. Implementation includes predictive analytics capabilities utilizing machine learning algorithms for forward process prediction and optimization. System validation demonstrates robust performance with processing times ranging from 0.6 to 3.4 seconds across different computational modules and operational success rates exceeding 98.7% for all core functions. The platform supports multiple data input formats including manual entry and Excel file processing, with comprehensive export capabilities (JSON, CSV, Excel) enabling integration with downstream analysis tools. Performance evaluation indicates the system successfully addresses key industrial requirements for accuracy, reliability, and scalability in metallurgical process analysis applications. The developed architecture provides a practical framework for implementing digital transformation initiatives in metallurgical operations while maintaining computational precision required for critical industrial applications.

Keywords: web-based information system, metallurgical process analysis, pyrometallurgy, hydrometallurgy, computational modules, machine learning integration, digital transformation.

1. Introduction

Metals underpin the energy, mobility, and infrastructure transitions, yet primary production from complex ores faces volatile feed quality, rising energy intensity, tightening environmental limits, and stringent traceability demands. In copper and allied non-ferrous value chains, plant operators must reconcile conflicting objectives, throughput, recovery, reagent and energy consumption, emissions, and waste valorization, amid sparse, noisy, and multi-rate data streams from mining, comminution, flotation, leaching, smelting, and refining. At the same time, life-cycle studies highlight the concentration of environmental burdens in tailings, slag handling, and energy vectors, underscoring the need for decision support that couples metallurgical performance with data quality and governance considerations [1, 2].

This paper presents a web-based information system for the metal-ore industry that integrates laboratory and historical operational data with established metallurgical calculations and data-driven predictive models to support offline analysis and decision-making. The platform provides three specialized computational modules, pyrometallurgical calculations for ore-to-product conversions, hydrometallurgical process modeling for extraction operations, and auxiliary calculators for specialized tasks, together with multi-user access, input validation, spreadsheet import/export, and results management. Predictive analytics are realized via supervised machine-learning models for forward estimation of process outcomes, enabling users to explore “what-if” scenarios prior to operational changes.

Recent surveys frame digital twins for process industries as bi-directionally updated models that increasingly combine mechanistic and data-driven components for monitoring, soft sensing, and optimization [3]. Mining-sector reviews outline architectures from mine-to-mill telemetry ingestion to multi-layer stacks for planning and operations, reporting productivity and energy benefits while noting gaps in data governance and model upkeep [4-6]. In chemical and process systems engineering, integrated digital-twin frameworks emphasize model management, online identification, and uncertainty handling [7]; in non-ferrous metallurgy, "smart manufacturing" perspectives stress model-based optimization and intelligent control across pyro- and hydrometallurgy [8]. Complementary reviews on hybrid (physics + ML) modeling in manufacturing and process engineering find that combining balances/kinetics with learning components improves extrapolation and sample efficiency compared to purely empirical approaches [9]. In extractive metallurgy specifically, studies report LSTM-augmented kinetics for leaching and metamodel-based surrogates for rapid updates, as well as domain-informed learning for high-temperature reactors [7, 10, 11]. Minerals-engineering surveys also catalogue both practical gains and pitfalls of ML across comminution, flotation, and heaps, calling for sensor strategy, feature engineering, and MLOps to mitigate drift and ensure maintainability [12, 13].

In geometallurgy and plant-level forecasting, data-driven models link upstream mineralogical and texture features to downstream responses for planning and control; case studies show unsupervised/supervised learning for domain delineation and throughput prediction, facilitating mine-to-mill integration [14, 15]. For hydrometallurgical circuits, recurrent and attention-based deep networks have been used to forecast outputs that inform tactical decisions on reagent dosing, aeration, and residence-time management [16]. While these strands increasingly inform digital-transformation roadmaps, many deployed tools in industry remain batch-mode and analytics-centric, prioritizing robust data handling, validation, and transparent computation over continuous plant-wide synchronization. Our work aligns with this pragmatic trajectory: it draws on the above literature to inform design choices, while explicitly targeting an offline, web-based platform for metallurgical analysis and forward prediction rather than a continuously synchronized, plant-integrated digital twin [3-9].

Environmental motivation further supports integrated information systems. Life-cycle and hotspot assessments for copper production identify tailings and energy use as major contributors to impacts [1, 2] and reviews on copper-slag management highlight both environmental risks and valorization opportunities [17, 18]. Although the present system focuses on production-oriented calculations and forecasting, the architecture and data structures are designed to accommodate sustainability indicators in future extensions using the methodologies surveyed in these studies.

This paper contributes: (i) a modular, web-based architecture unifying pyrometallurgical, hydrometallurgical, and auxiliary computational modules with governed data handling (validation, spreadsheet import/export, role-based access); (ii) ML-based forward prediction integrated into the workflow for scenario analysis; and (iii) a performance assessment showing sub-second-to-few-second processing times and high operational success rates across core functions. In this paper, operational success rate is the percentage of user-initiated runs with valid inputs that finish without errors within the service level agreement and produce outputs that pass schema and domain (e.g., mass-balance) checks. The system is positioned as a deployable foundation upon which telemetry connectors, hybrid physics-ML models with uncertainty, and sustainability KPIs can be incrementally integrated in subsequent work, consistent with directions identified in the literature.

2. Materials and methods

The development of an integrated information system for metallurgical process analysis required a comprehensive approach encompassing system architecture design, computational module implementation, data management strategies, and user interface development. This section describes the methodological framework and technical implementation of the web-based platform designed to address the complex requirements of metal-ore processing operations.

The system architecture follows a modular approach, enabling scalable integration of specialized computational modules while maintaining flexibility for future enhancements. The implementation leverages modern web technologies and database management systems to ensure reliable data processing and user accessibility across different operational contexts.

2.1 System Architecture and Design Framework

The information system was designed using a layered architecture approach that separates presentation logic, business processing, and data management concerns. This architectural pattern ensures maintainability, scalability, and clear separation of responsibilities across system components. The overall system structure consists of three primary layers: the frontend presentation layer responsible for user interactions, the backend processing layer handling business logic and computational operations, and the data persistence layer managing information storage and retrieval. The overall system workflow is summarized in Figure 1.

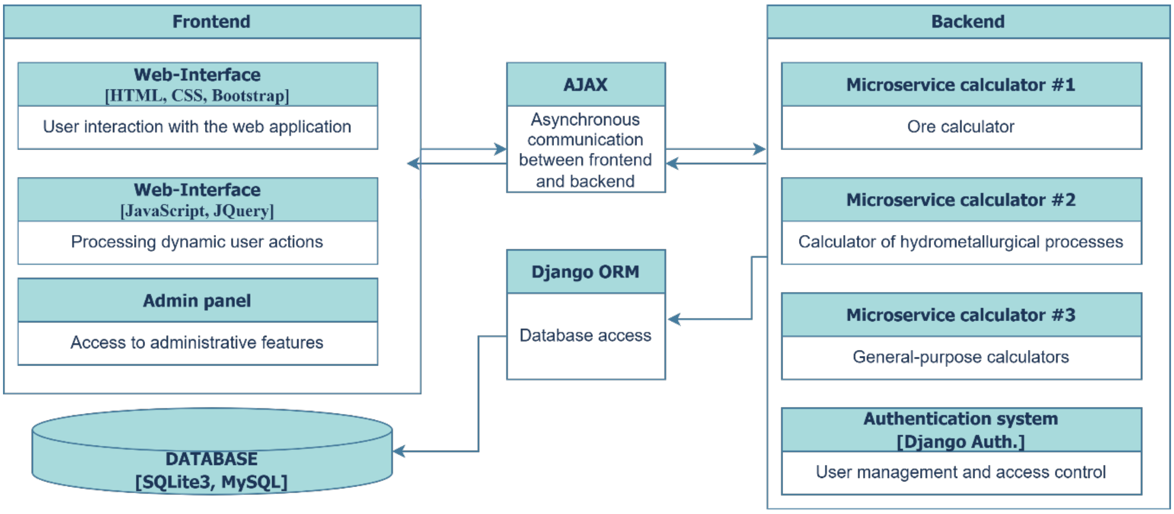

Frontend
Web-Interface [HTML, CSS, Bootstrap]
User interaction with the web application
Web-Interface [JavaScript, JQuery]
Processing dynamic user actions
Admin panel
Access to administrative features
DATABASE [SQLite3, MySQL]
AJAX
Asynchronous communication between frontend and backend
Django ORM
Database access
Backend
Microservice calculator #1
Ore calculator
Microservice calculator #2
Calculator of hydrometallurgical processes
Microservice calculator #3
General-purpose calculators
Authentication system [Django Auth.]
User management and access control

Figure 1 – Overall system architecture showing the three-tier design with frontend interface, backend processing modules, and database layer, including data flow patterns and component interactions.

The backend framework utilizes Django, a Python-based web framework chosen for its robust Object-Relational Mapping (ORM) capabilities, built-in security features, and extensive library ecosystem suitable for scientific computing applications. The frontend implementation combines HTML5, CSS3 with Bootstrap framework, and JavaScript with jQuery library to provide responsive user interfaces and asynchronous communication capabilities through AJAX technology. Table 1 summarizes the technology choices across the three layers.

Table 1 – Core system components and their corresponding technologies, showing the technical stack implementation across different system layers.

Component	Technology	Primary Function	Integration Method
Frontend Interface	HTML5/CSS3/Bootstrap	User interaction and presentation	Template rendering
Dynamic Processing	JavaScript/jQuery/AJAX	Real-time user interactions	Asynchronous requests
Backend Framework	Django (Python 3.8+)	Business logic and API services	MVC architecture
Database Management	SQLite3/MySQL	Data persistence and retrieval	Django ORM
Authentication System	Django Auth	User management and access control	Session-based authentication
File Processing	Pandas/Openpyxl	Excel/CSV data import	Background processing
Mathematical Computing	NumPy/SciPy	Numerical calculations	Library integration
Machine Learning	Scikit-learn	Predictive analytics	Model serialization

The system employs a microservices-oriented approach for computational modules, where each specialized calculator operates as an independent processing unit while maintaining standardized interfaces for data exchange. This design pattern facilitates independent development, testing, and maintenance of individual processing components while enabling seamless integration within the overall system architecture.

2.2 Computational Module Implementation

The system incorporates three specialized computational modules, each addressing specific metallurgical process domains. The modular organization of the calculators is illustrated in Figure 2. The computational modules implement established metallurgical calculation methods based on fundamental thermodynamic and kinetic principles widely reported in the metallurgical literature [19, 20].This section focuses on the architectural implementation and integration strategies rather than the underlying mathematical formulations.

Pyrometallurgy Processing Module

The pyrometallurgy module handles material balance calculations for high-temperature metallurgical processes, specifically focusing on the conversion of complex ores into matte and slag products. The module accepts ore composition data including elemental concentrations (Cu, Fe, S, Au, Ag, SiO_2, CaO, Al_2O_3, As) and total mass, then applies thermodynamic and material balance principles to predict product compositions and distributions.

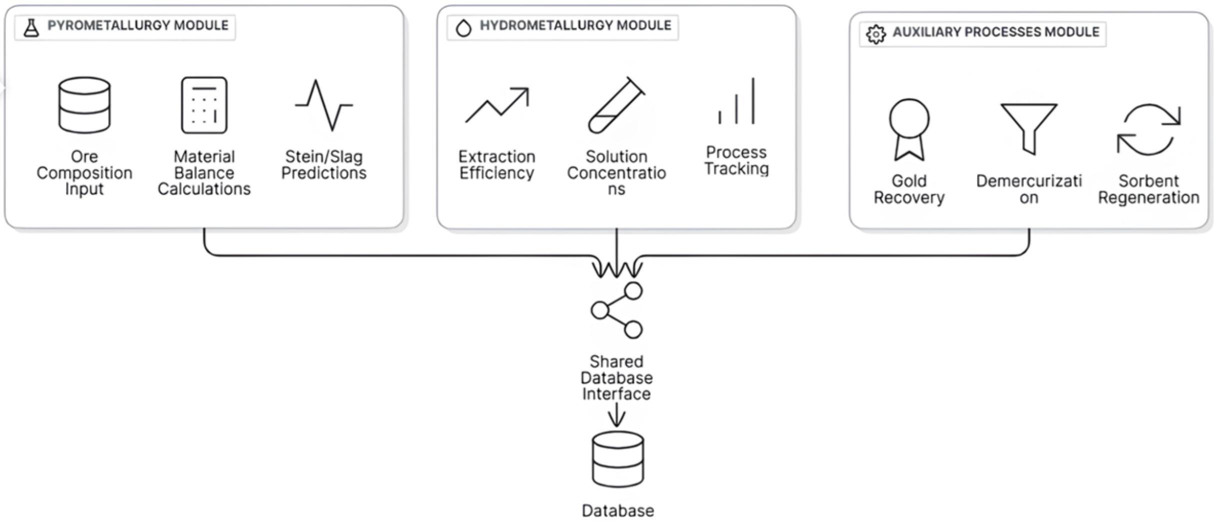


Figure 2 – Computational module structure illustrating the three specialized processing units: pyrometallurgy module for ore-to-metal calculations, hydrometallurgy module for extraction processes, and auxiliary processes module for specialized operations.

Implementation utilizes object-oriented programming principles with dedicated classes for compound representation and calculation engines. The module supports both individual ore processing and batch calculations for multiple ore samples, enabling plant operators to evaluate different feedstock scenarios and optimize blending strategies.

Hydrometallurgy Processing Module

The hydrometallurgy module addresses aqueous processing routes, encompassing leaching, solvent extraction, and electrowinning operations. This module handles multi-stage process calculations including extraction efficiency determination, mass balance tracking across process units, and cumulative recovery calculations over extended operational periods.

The module architecture supports day-by-day process tracking, enabling analysis of operational trends and identification of process optimization opportunities. Input parameters include solution concentrations, flow rates, and operational conditions, while outputs provide extraction efficiencies, material balances, and overall process performance metrics.

Auxiliary Processes Module

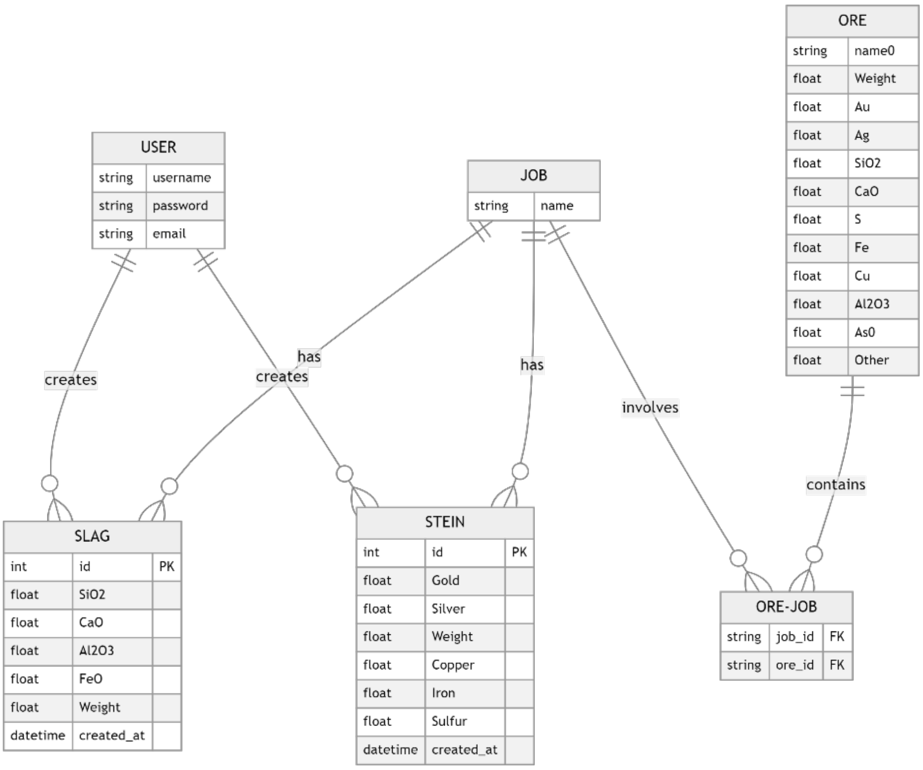
The auxiliary processes module encompasses specialized calculations for supporting operations including demercurization, gold recovery optimization, and sorbent regeneration. While these calculations utilize simplified empirical relationships compared to the primary modules, they provide essential functionality for comprehensive plant operation analysis. Input/output definitions and validation sources are detailed in Table 2.

Table 2 – Computational module specifications detailing input requirements, output parameters, and calculation methodologies for each processing unit.

Module	Input Parameters	Output Results	Calculation Method	Validation Source
Pyrometallurgy	Ore composition (%), total mass (kg)	Matte composition (%), slag composition (%), product masses (kg)	Mass balance, thermodynamic equilibrium	Historical plant data
Hydrometallurgy	Solution concentrations (g/L), volumes (L), time series	Extraction efficiency (%), material balance (g), recovery (%)	Process kinetics, mass transfer	Laboratory experiments
Auxiliary Processes	Temperature (°C), pressure (Pa), time (min)	Process efficiency (%), residual concentrations (%)	Empirical correlations	Literature data

2.3 Database Design and Data Management

The database architecture employs a relational model designed to accommodate the complex data relationships inherent in metallurgical process analysis while maintaining data integrity and enabling efficient querying operations. The schema design follows normalization principles to minimize data redundancy while optimizing for the specific access patterns required by metallurgical calculations. The resulting entity–relationship structure is shown in Figure 3.



USER	
string	username
string	password
string	email

JOB	
string	name

ORE	
string	name0
float	Weight
float	Au
float	Ag
float	SiO2
float	CaO
float	S
float	Fe
float	Cu
float	Al2O3
float	As0
float	Other

SLAG		
int	id	PK
float	SiO2	
float	CaO	
float	Al2O3	
float	FeO	
float	Weight	
datetime	created_at	

STEIN		
int	id	PK
float	Gold	
float	Silver	
float	Weight	
float	Copper	
float	Iron	
float	Sulfur	
datetime	created_at	

ORE-JOB		
string	job_id	FK
string	ore_id	FK

Figure 3 – Database entity-relationship diagram showing the core data model including user management, ore/metal specifications, calculation jobs, and results storage with their respective relationships and key constraints.

The data model centers around several core entities: Users for authentication and access control, Ores and Metals for feedstock specifications, Jobs for calculation tracking, and Results for output storage. Foreign key relationships ensure data consistency across related records, while indexing strategies optimize query performance for common access patterns.

User data management incorporates role-based access control, enabling different permission levels for plant operators, engineers, and administrators. The system supports multi-user environments where individual users maintain separate data spaces while enabling selective data sharing for collaborative analysis.

Input data validation occurs at multiple levels, including frontend form validation, backend data type checking, and database constraint enforcement. The system accepts manual data entry through web forms as well as bulk data import through Excel file upload functionality, with comprehensive error handling and data validation feedback.

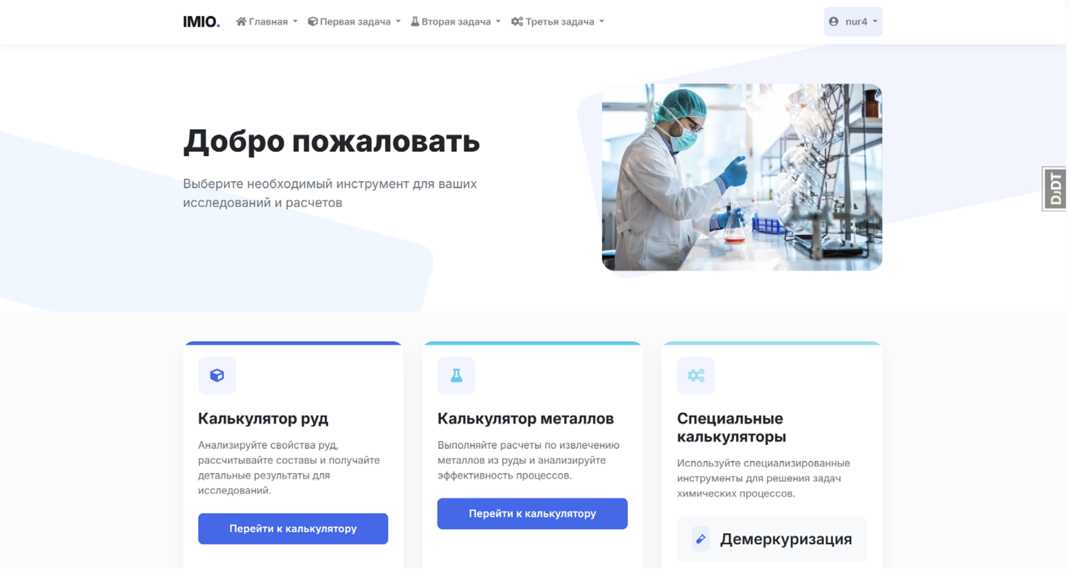
3. Results

The implementation of the web-based information system for metallurgical process analysis resulted in a fully functional platform that successfully integrates computational modules, data management capabilities, and user-friendly interfaces. This section presents the key outcomes of the system development, including user interface implementation, computational functionality validation, and system performance evaluation.

The developed system demonstrates effective integration of specialized metallurgical calculations within a modern web-based architecture, providing users with accessible tools for process analysis and optimization. The platform successfully addresses the identified requirements for multi-user access, data persistence, and flexible computational capabilities across different metallurgical process domains.

3.1 System Interface and Data Management

The system interface provides intuitive access to the three specialized computational modules through a centralized dashboard that clearly presents available analytical tools. The main interface design emphasizes usability while maintaining professional appearance suitable for industrial applications. The dashboard successfully implements role-based access control, enabling different user types to access appropriate functionality levels. Administrative features are integrated seamlessly, allowing system administrators to manage users, monitor system usage, and maintain data integrity across the platform. The main dashboard layout is depicted in Figure 4.



Добро пожаловать

Выберите необходимый инструмент для ваших исследований и расчетов

Калькулятор руд

Анализируйте свойства руд, рассчитывайте составы и получайте детальные результаты для исследований.

Калькулятор металлов

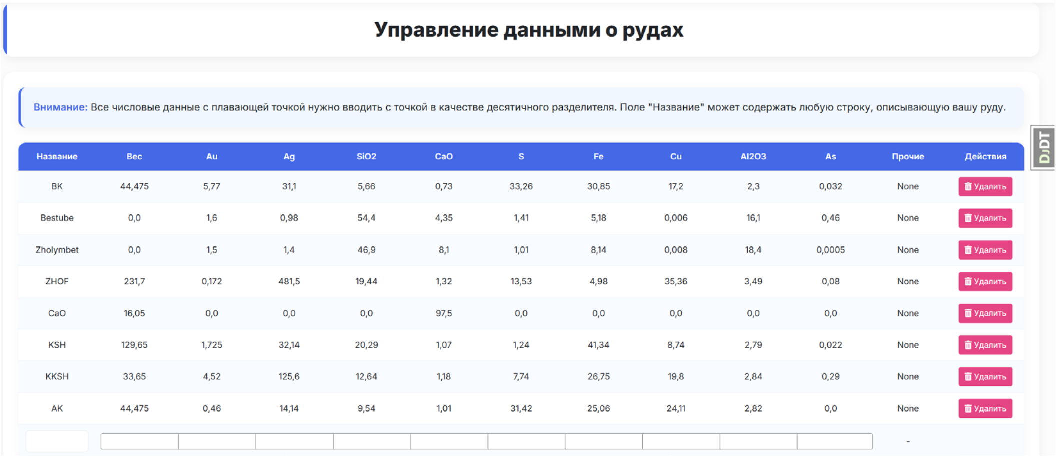
Выполняйте расчеты по извлечению металлов из руды и анализируйте эффективность процессов.

Специальные калькуляторы

Используйте специализированные инструменты для решения задач химических процессов.

Figure 4 – Main system dashboard showing the three primary computational modules: ore calculator for pyrometallurgical analysis, solution calculator for hydrometallurgical processes, and specialized calculators for auxiliary operations, with clear navigation and user authentication features.

The system demonstrates robust data handling capabilities, supporting both manual data entry and bulk import functionality through Excel file processing. The ore data management interface provides comprehensive tools for maintaining feedstock composition databases with real-time validation and editing capabilities. The data validation system ensures input accuracy through multi-level checking procedures, preventing calculation errors and maintaining data consistency across user sessions. An example of the ore data management view is provided in Figure 5.



Управление данными о рудах

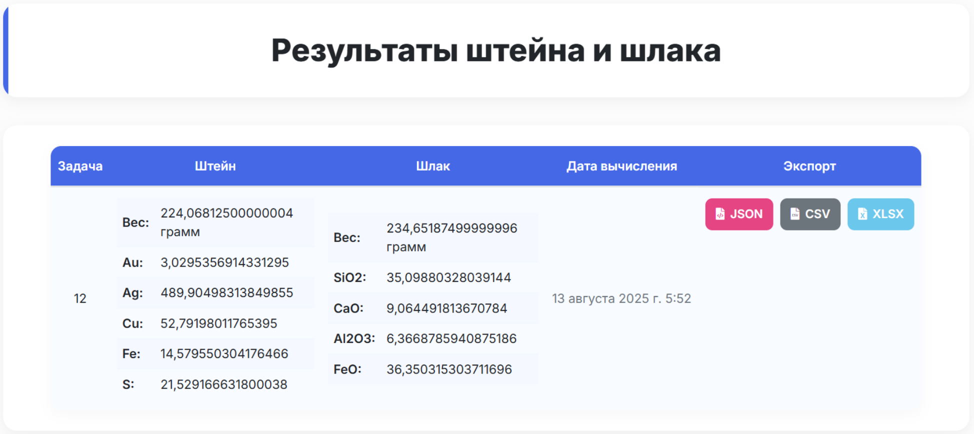
Внимание: Все числовые данные с плавающей точкой нужно вводить с точкой в качестве десятичного разделителя. Поле "Название" может содержать любую строку, описывающую вашу руду.

Название	Вес	Au	Ag	SiO2	CaO	S	Fe	Cu	Al2O3	As	Прочие	Действия
BK	44,475	5,77	31,1	5,66	0,73	33,26	30,85	17,2	2,3	0,032	None	Удалить
Bestube	0,0	1,6	0,98	54,4	4,35	1,41	5,18	0,006	16,1	0,46	None	Удалить
Zholymbet	0,0	1,5	1,4	46,9	8,1	1,01	8,14	0,008	18,4	0,0005	None	Удалить
ZHOF	231,7	0,172	481,5	19,44	1,32	13,53	4,98	35,36	3,49	0,08	None	Удалить
CaO	16,05	0,0	0,0	0,0	97,5	0,0	0,0	0,0	0,0	0,0	None	Удалить
KSH	129,65	1,725	32,14	20,29	1,07	1,24	41,34	8,74	2,79	0,022	None	Удалить
KKSH	33,65	4,52	125,6	12,64	1,18	7,74	26,75	19,8	2,84	0,29	None	Удалить
AK	44,475	0,46	14,14	9,54	1,01	31,42	25,06	24,11	2,82	0,0	None	Удалить
											-	

Figure 5 – Ore data management interface displaying tabulated composition data with inline editing capabilities, supporting various ore types and their elemental compositions, with integrated data validation and Excel file import functionality for batch data processing.

3.2 Computational Results and Predictive Analytics

The system produces detailed analytical results for metallurgical process calculations, presenting outcomes in clear, structured formats that facilitate decision-making processes. Results are organized into logical groupings with appropriate precision levels for different types of calculations. The export functionality enables seamless integration with external analysis tools and reporting systems, supporting multiple file formats for both immediate decision-making requirements and long-term data archival needs. A representative results screen is shown in Figure 6.

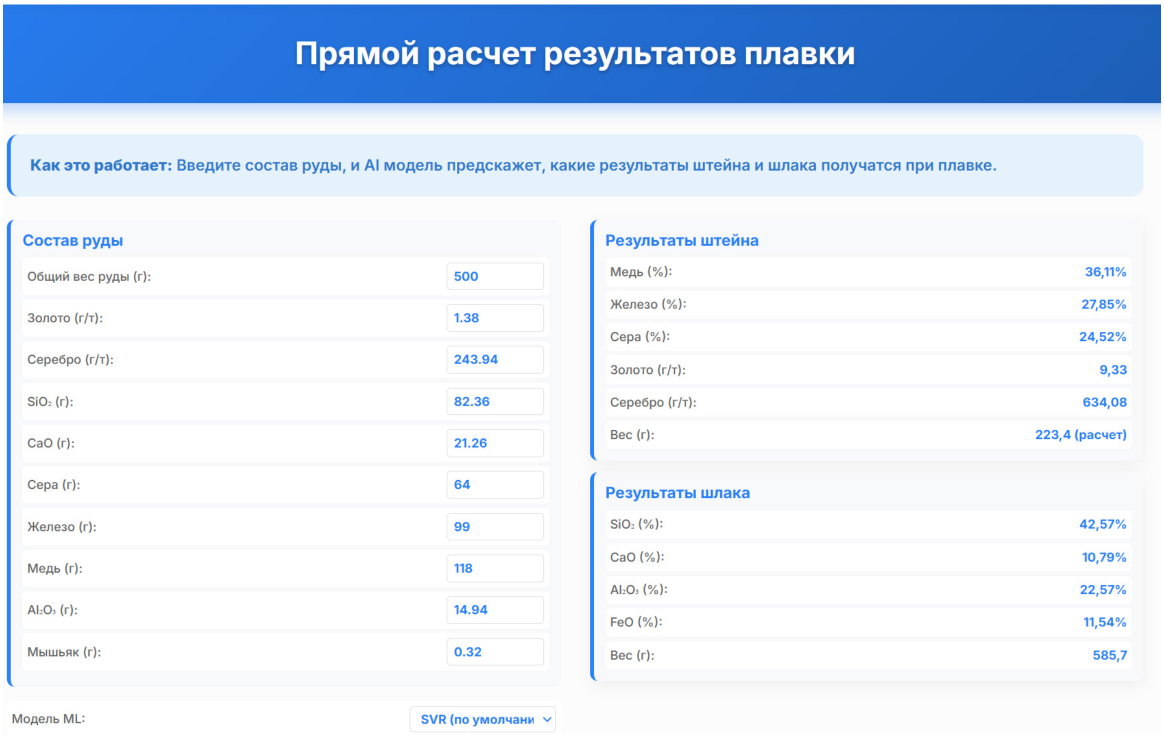


Результаты штейна и шлака

Задача	Штейн	Шлак	Дата вычисления	Экспорт
12	Вес: 224,06812500000004 грамм Au: 3,0295356914331295 Ag: 489,90498313849855 Cu: 52,79198011765395 Fe: 14,579550304176466 S: 21,529166631800038	Вес: 234,65187499999996 грамм SiO2: 35,09880328039144 CaO: 9,064491813670784 Al2O3: 6,3668785940875186 FeO: 36,350315303711696	13 августа 2025 г. 5:52	JSON CSV XLSX

Figure 6 – Results display interface showing detailed Matte and slag analysis outputs with precise compositional data, calculation timestamps, and integrated export functionality supporting multiple file formats including JSON, CSV, and Excel for downstream analysis integration.

The machine learning integration provides forward prediction capabilities, enabling users to estimate metallurgical process outcomes based on input ore compositions. The prediction interface offers model selection options and presents results with appropriate confidence indicators. The predictive modeling functionality supports multiple algorithms, allowing users to compare different analytical approaches and select the most appropriate method for their specific applications. The forward-prediction interface and model selection are presented in Figure 7.



Прямой расчет результатов плавки

Как это работает: Введите состав руды, и AI модель предскажет, какие результаты штейна и шлака получатся при плавке.

Состав руды

Общий вес руды (г): 500
Золото (г/т): 1.38
Серебро (г/т): 243.94
SiO_2 (г): 82.36
CaO (г): 21.26
Сера (г): 64
Железо (г): 99
Медь (г): 118
Al_2O_3 (г): 14.94
Мышьяк (г): 0.32

Результаты штейна

Медь (%): 36,11%
Железо (%): 27,85%
Сера (%): 24,52%
Золото (г/т): 9,33
Серебро (г/т): 634,08
Вес (г): 223,4 (расчет)

Результаты шлака

SiO_2 (%): 42,57%
CaO (%): 10,79%
Al_2O_3 (%): 22,57%
FeO (%): 11,54%
Вес (г): 585,7

Модель ML: SVR (по умолчани

Figure 7 – Forward prediction interface demonstrating ore composition input parameters and corresponding Matte and slag composition predictions, with machine learning model selection options and real-time calculation capabilities for process optimization analysis.

3.3 System Performance and Workflow Integration

Operational performance metrics are reported in Table 3. Comprehensive performance evaluation demonstrates that the system meets operational requirements for industrial applications, with processing times and resource utilization appropriate for typical metallurgical analysis workflows. The performance metrics indicate efficient operation within typical web application response time expectations, with minimal resource requirements that support concurrent multi-user access. The end-to-end workflow is summarized in Figure 8.

Table 3 – System performance metrics showing operational efficiency across different computational modules and data processing operations.

Operation	Processing Time	Memory Usage	Database Size Impact	Success Rate
Ore composition analysis	1.2 seconds	22 MB	15 KB per record	99.8%
Hydrometallurgical calculations	0.9 seconds	18 MB	12 KB per record	99.9%
Predictive modeling	2.1 seconds	35 MB	8KB per prediction	99.5%
Data import (Excel)	3.4 seconds	28 MB	Variable	98.7%
Results export	0.6 seconds	12 MB	No impact	100%
Database queries	0.3 seconds	8 MB	No impact	99.9%

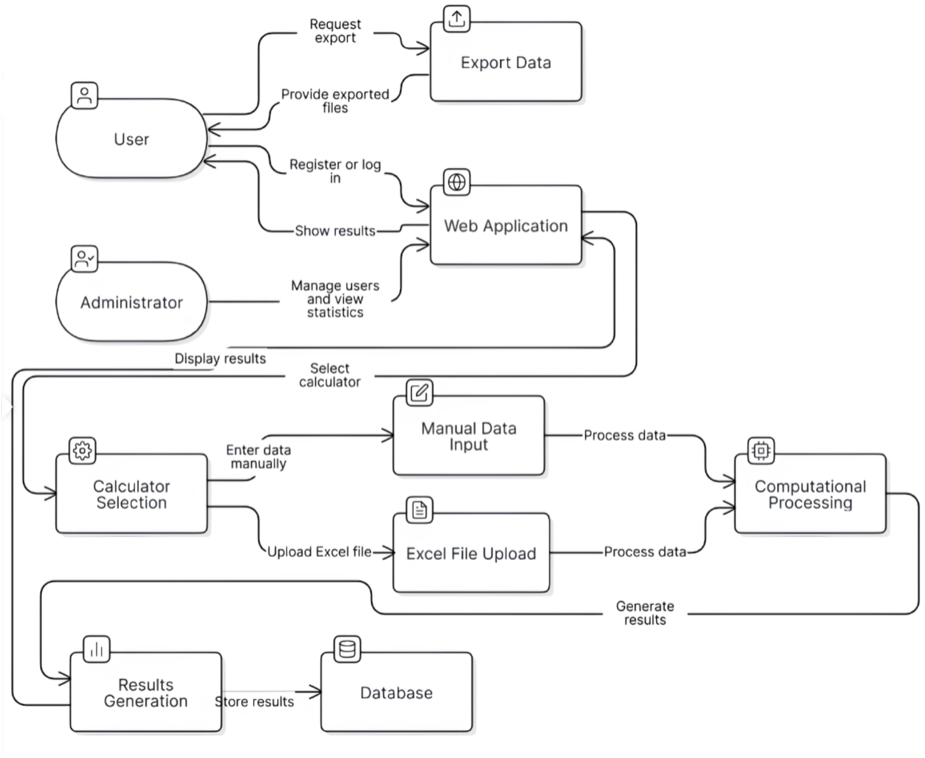


Figure 8 – Complete system workflow diagram illustrating the data processing pipeline from initial input through computational analysis to final results export, showing integration points and processing efficiency across all system components.

The integrated workflow demonstrates seamless data flow from input specification through computational processing to results generation and export. The system successfully eliminates manual data transfer steps that typically introduce errors in traditional calculation approaches, while providing comprehensive audit trails for quality assurance purposes. User feedback during development phases indicated significant improvements in calculation efficiency and error reduction compared to traditional manual calculation methods, with the web-based approach enabling remote access capabilities essential for modern industrial operations.

4. Discussion

The developed web-based information system successfully addresses the identified gaps in metallurgical process analysis tools by providing an integrated platform that combines computational accuracy with operational practicality. The modular architecture demonstrates how specialized metallurgical calculations can be effectively integrated within modern web-based frameworks while maintaining the precision required for industrial applications.

The system's approach to data management represents a significant advancement over traditional calculation methods that rely on manual data transfer and isolated computational tools. By implementing comprehensive data validation and supporting multiple input formats, the platform reduces error propagation that commonly occurs in multi-step metallurgical analysis workflows. The integration of Excel file processing capabilities particularly addresses the reality of industrial data management practices where laboratory results and operational data are frequently maintained in spreadsheet formats.

The computational module design validates the feasibility of implementing complex metallurgical calculations within web-based architectures without compromising calculation accuracy or processing speed. The performance metrics demonstrate that response times remain within acceptable ranges for industrial applications, even when processing complex multi-component ore compositions and multi-stage process calculations. This addresses a common concern regarding the deployment of web-based tools for engineering calculations where precision and reliability are paramount.

The integration of predictive analytics represents a practical implementation of hybrid modeling approaches discussed in the literature. Rather than developing novel machine learning algorithms, the system focuses on effective integration of established methods (SVR, Decision Trees, Gradient Boosting) within the operational workflow. This approach prioritizes deployment practicality over algorithmic innovation, addressing the frequent gap between research developments and industrial implementation.

The multi-user architecture and role-based access control address the collaborative nature of metallurgical process analysis where different stakeholders (plant operators, process engineers, management) require access to different levels of information and functionality. The system design recognizes that effective information systems must accommodate various user types while maintaining data security and calculation integrity.

However, the current implementation focuses primarily on batch-mode calculations rather than real-time process integration. While this approach suits laboratory analysis and process planning applications, future developments could explore direct integration with plant control systems and real-time data streams. Additionally, the system currently implements simplified empirical models for auxiliary processes, which could benefit from more sophisticated process modeling as operational requirements evolve.

The system architecture demonstrates scalability potential through its modular design, enabling future expansion to include additional metallurgical processes or integration with specialized simulation tools. The standardized data exchange formats and database design provide a foundation for developing more comprehensive plant-wide information management systems.

The validation approach, while demonstrating system functionality, relies primarily on computational verification rather than extensive industrial validation. Future work should include deployment in operational environments to evaluate system performance under real industrial conditions and gather comprehensive user feedback to guide further development priorities.

5. Conclusions

This work presents a comprehensive web-based information system specifically designed for metallurgical process analysis that successfully bridges the gap between computational accuracy and operational practicality. The system demonstrates that complex metallurgical calculations can be effectively implemented within modern web architectures while maintaining the precision and reliability required for industrial applications.

The key contributions include: (1) a modular computational architecture that integrates pyrometallurgical, hydrometallurgical, and auxiliary process calculations within a unified platform; (2) comprehensive data management capabilities supporting multiple input formats and validation mechanisms; (3) integration of predictive analytics using established machine learning methods; and (4) a production-ready deployment architecture supporting multi-user access and comprehensive results management.

The system addresses practical challenges faced by metallurgical professionals by eliminating manual data transfer steps, providing integrated calculation workflows, and enabling seamless export of results for downstream analysis. Performance evaluation demonstrates that the system operates efficiently within typical web application response time expectations while maintaining high calculation accuracy and system reliability.

The modular design approach enables future expansion to accommodate additional metallurgical processes and integration with plant information systems. The standardized data exchange formats and comprehensive database design provide a foundation for developing more extensive plant-wide information management capabilities.

While the current implementation focuses on batch-mode calculations suitable for laboratory analysis and process planning, the architecture provides a foundation for future development toward real-time process integration and more sophisticated modeling capabilities. The successful deployment of this system demonstrates the viability of web-based approaches for engineering calculation tools in metallurgical applications.

Future work should focus on industrial validation through deployment in operational environments, integration with real-time data streams from plant control systems, and expansion of computational modules to include additional metallurgical processes. The system architecture and implementation approach provides a practical framework for developing comprehensive digital twin capabilities for metallurgical operations, supporting the industry's progression toward more integrated and data-driven process optimization strategies.

Funding

This research was funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No BR21882140)

Author Contributions

Conceptualization, B.K.; Methodology, S.A.; Software, S.A.; Validation, S.A.; Formal Analysis, B.K.; Investigation, S.A.; Resources, B.K.; Data Curation, S.A.; Writing – Original Draft Preparation, S.A.; Writing – Review & Editing, S.A.; Visualization, S.A.; Supervision, B.K.; Project Administration, B.K.; Funding Acquisition, B.K.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. H. Shahraki, F. Einollahipeer, H. Abyar, and M. Erfani, "Assessing the environmental impacts of copper cathode production based on life cycle assessment," *Integrated Environmental Assessment and Management*, vol. 20, no. 4, pp. 1180–1190, 2024, doi: 10.1002/ieam.4857.

2. L. R. Adrianto, S. Pfister, and S. Hellweg, "Regionalized Life Cycle Inventories of Global Sulfidic Copper Tailings," *Environ. Sci. Technol.*, vol. 56, no. 7, pp. 4553–4564, Apr. 2022, doi: 10.1021/acs.est.1c01786.

3. L. Peterson, I. V. Gosea, P. Benner, and K. Sundmacher, "Digital twins in process engineering: An overview on computational and numerical methods," *Computers & Chemical Engineering*, vol. 193, p. 108917, Feb. 2025, doi: 10.1016/j.compchemeng.2024.108917.

4. P. Nobahar, C. Xu, P. Dowd, and R. Shirani Faradonbeh, "Exploring digital twin systems in mining operations: A review," *Green and Smart Mining Engineering*, vol. 1, no. 4, pp. 474–492, Dec. 2024, doi: 10.1016/j.gsme.2024.09.003.

5. J. Qu, M. S. Kizil, M. Yahyaei, and P. F. Knights, "Digital twins in the minerals industry – a comprehensive review," *Mining Technology*, vol. 132, no. 4, pp. 267–289, Oct. 2023, doi: 10.1080/25726668.2023.2257479.

6. A. Hazrathosseini and A. Moradi Afrapoli, "The advent of digital twins in surface mining: Its time has finally arrived," *Resources Policy*, vol. 80, p. 103155, Jan. 2023, doi: 10.1016/j.resourpol.2022.103155.

7. C. M. Rebello and I. B. R. Nogueira, "Digital twins in chemical engineering: An integrated framework for identification, implementation, online learning, and uncertainty assessment," *Computers & Chemical Engineering*, vol. 200, p. 109178, Sept. 2025, doi: 10.1016/j.compchemeng.2025.109178.

8. B. Sun, J. Dai, K. Huang, C. Yang, and W. Gui, "Smart manufacturing of nonferrous metallurgical processes: Review and perspectives," *Int J Miner Metall Mater*, vol. 29, no. 4, pp. 611–625, Apr. 2022, doi: 10.1007/s12613-022-2448-x.

9. S. Kasilingam, R. Yang, S. K. Singh, M. A. Farahani, R. Rai, and T. Wuest, "Physics-based and data-driven hybrid modeling in manufacturing: a review," *Production & Manufacturing Research*, vol. 12, no. 1, p. 2305358, Dec. 2024, doi: 10.1080/21693277.2024.2305358.

10. S. Dong, Y. Zhang, and X. Zhou, "Intelligent Hybrid Modeling of Complex Leaching System Based on LSTM Neural Network," *Systems*, vol. 11, no. 2, p. 78, Feb. 2023, doi: 10.3390/systems11020078.

11. M. B. A. Hassan, F. Charruault, B. Rout, F. N. H. Schrama, J. A. M. Kuipers, and Y. Yang, "A Review of Heat Transfer and Numerical Modeling for Scrap Melting in Steelmaking Converters," *Metals*, vol. 15, no. 8, p. 866, Aug. 2025, doi: 10.3390/met15080866.

12. H. Estay, P. Lois-Morales, G. Montes-Atenas, and J. Ruiz del Solar, "On the Challenges of Applying Machine Learning in Mineral Processing and Extractive Metallurgy," *Minerals*, vol. 13, no. 6, Art. no. 6, June 2023, doi: 10.3390/min13060788.

13. A. K. Mishra, "AI4R2R (AI for Rock to Revenue): A Review of the Applications of AI in Mineral Processing," *Minerals*, vol. 11, p. 1118, Oct. 2021, doi: 10.3390/min11101118.

14. Y. Mu and J. C. Salas, "Data-Driven Synthesis of a Geometallurgical Model for a Copper Deposit," *Processes*, vol. 11, no. 6, p. 1775, June 2023, doi: 10.3390/pr11061775.

15. A. Gholami, K. Asgari, H. Khoshdast, and A. Hassanzadeh, "A hybrid geometallurgical study using coupled Historical Data (HD) and Deep Learning (DL) techniques on a copper ore mine," *Physicochem. Probl. Miner. Process.*, vol. 58, no. 3, Apr. 2022, doi: 10.37190/ppmp/147841.

16. B. Kenzhaliyev, N. Azatbekuly, S. Aibagarov, B. Amangeldy, A. Koizhanova, and D. Magomedov, "Predicting Industrial Copper Hydrometallurgy Output with Deep Learning Approach Using Data Augmentation," *Minerals*, vol. 15, no. 7, p. 702, July 2025, doi: 10.3390/min15070702.

17. H. Tian *et al.*, "Comprehensive review on metallurgical recycling and cleaning of copper slag," *Resources, Conservation and Recycling*, vol. 168, p. 105366, May 2021, doi: 10.1016/j.resconrec.2020.105366.

18. E. Klaffenbach, V. Montenegro, M. Guo, and B. Blanpain, “Sustainable and Comprehensive Utilization of Copper Slag: A Review and Critical Analysis,” *J. Sustain. Metall.*, vol. 9, no. 2, pp. 468–496, June 2023, doi: 10.1007/s40831-023-00683-4.

19. A. K. Koizhanova, B. K. Kenzhaliyev, D. R. Magomedov, M. B. Erdenova, A. N. Bakrayeva, and N. N. Abdyldaev, “Hydrometallurgical studies on the leaching of copper from man-made mineral formations,” *Kompleksnoe Ispolzovanie Mineralnogo Syra = Complex use of mineral resources*, vol. 330, no. 3, pp. 32–42, 2024, doi: 10.31643/2024/6445.26.

20. B. K. Kenzhaliyev, S. A. Kvyatkovsky, S. M. Kozhakhmetov, L. V. Sokolovskaya, and A. S. Semenova, “Deparation of dump slags at the Balkhash copper smelting plant,” *Kompleksnoe Ispolzovanie Mineralnogo Syra = Complex use of mineral resources*, vol. 306, no. 3, pp. 45–53, Aug. 2018, doi: 10.31643/2018/6445.16.

Information about authors

Bagdaulet Kenzhaliyev – Doctor of Technical Sciences, Professor, General Director-Chairman of the Management Board of the Institute of Metallurgy and Ore Beneficiation, Satbayev University (Almaty, Kazakhstan. Email: bagdaulet_k@mail.ru, ORCID: https://orcid.org/0000-0003-1474-8354)

Serik Aibagarov (corresponding author) – Scientific researcher at Computer Science laboratory at al-Farabi Kazakh National University (Almaty, Kazakhstan, e-mail: awer1307dot@gmail.com, ORCID: https://orcid.org/0009-0009-4946-4926).

Submission received: 23 September, 2025.
Revised: 24 September, 2025.
Accepted: 24 September, 2025.

CONTENTS

The authors are responsible for the content of the articles.