Development of a machine learning-based framework for predicting failures in heat supply networks

Dauren Darkenbayev[1,4], Gulnar Balakayeva[2], Uzak Zhapbasbayev[3], Mukhit Zhanuzakov[2]
[1]Department of Computational Sciences and Statistics, Faculty of Mechanics and Mathematics, Al-Farabi Kazakh National University, Almaty, Kazakhstan
[2]Department of Computer Science, Faculty of Information Technologies,Al-Farabi Kazakh National University, Almaty, Kazakhstan
[3]Laboratory "Modeling in Energy Sector", Satbayev University, Almaty, Kazakhstan
[4]Department of Computer Science, Institute of Physics, Mathematics and Digital Technologies Kazakh National Women's Teacher Training University, Almaty, Kazakhstan

Article Info

Article history:

Received Mar 21, 2025
Revised Aug 20, 2025
Accepted Sep 27, 2025

Keywords:

Anomaly detection
Ensemble learning
Heat supply
Predictive maintenance
Real-time monitoring

ABSTRACT

The increasing complexity and scale of heat supply systems leads to a higher risk of failures, which may cause significant economic and environmental consequences. This study develops a predictive mathematical framework for the early detection of emergency conditions in heat supply networks (HSNs) using machine learning (ML). The proposed approach is based on the LightGBM gradient boosting (GB) algorithm, chosen for its high accuracy and efficiency in handling large datasets. Real operational data (temperature, pressure, flow, and vibration) were considered. Data preprocessing, feature engineering (including SHAP analysis), and hyperparameter tuning with grid search and 5-fold cross-validation improved prediction quality. The model achieved accuracy of 85%, F1-score of 0.82, and receiver operating characteristic (ROC)-area under the curve (AUC) of 0.96, outperforming logistic regression (LR) and decision trees. The framework may be integrated into monitoring systems for predictive maintenance, reducing downtime and optimizing costs.

Corresponding Author:

Dauren Darkenbayev
Department of Computational Sciences and Statistics, Faculty of Mechanics and Mathematics
Al-Farabi Kazakh National University
Almaty, Kazakhstan
Email: dauren.kadyrovich@gmail.com

1. INTRODUCTION

Modern technological systems such as industrial production, transportation, and energy infrastructure are characterized by increasing complexity and automation [1]–[3]. While this improves efficiency, it also increases vulnerability to failures, leading to economic losses, environmental harm, and safety hazards [4]–[6]. Heat supply networks (HSNs), as critical urban infrastructure, are especially prone to risks due to their wide distribution and dynamic operation. Early prediction of pre-emergency states in HSNs is therefore a research priority.

Traditional diagnostic and maintenance methods (rule-based systems and statistical models) depend on thresholds and expert knowledge [7]–[10], but they are insufficient for nonlinear and high-dimensional data. For example, Rahal *et al.* [11] analyzed heat losses but noted scalability limits. Ukoba [12] applied time-series anomaly detection, but robustness was low under changing loads.

Machine learning (ML) offers more flexibility. Support vector machines (SVM) and random forests (RF) have been used for anomaly detection with moderate success [13], [14]. Artificial neural networks

(ANNs) improve flexibility but require high resources and lack interpretability [15]. Gradient boosting (GB), particularly LightGBM, shows strong predictive ability in power grids and industrial systems. LightGBM efficiently handles large datasets and complex nonlinear dependencies, using gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB). Despite these strengths, little research applies GB to HSN fault prediction [16].

This study addresses this gap by proposing a LightGBM-based predictive framework for HSNs, integrating SHAP for interpretability. Main contributions:
- Data-driven methodology for early detection of pre-emergency conditions using multi-dimensional data (temperature, pressure, flow, and vibration).
- Comparative analysis of logistic regression (LR), RF, SVM, ANN, and LightGBM.
- SHAP integration for interpretability.
- Validation using a large synthetic dataset with cross-validation and performance metrics (accuracy, F1, and receiver operating characteristic (ROC)-area under the curve (AUC)).

2. MATERIALS AND METHODS

2.1. Predicting failures in heat networks

Failures (leaks, pipe breaks, and equipment damage) disrupt heating and cause severe consequences [17]. Predicting them is complex due to nonlinear interactions of pressure, temperature, and flow [18]-[23] ML is applied for:
- Condition diagnostics (classification of malfunctions using sensor data).
- Maintenance optimization (time-series based predictive schedules).
- Heat flow modeling (simulation of network performance).
- Clustering (identifying high-risk sections).

2.2. Gradient boosting method

GB combines weak learners (decision trees) into a strong predictor by minimizing loss iteratively.

```
Algorithm 1. GB procedure
Input: training set $\{(x_i, y_i)\}$, learning rate $\eta$, iterations $M$
    1. Initialize model with constant prediction.
    2. For t = 1 … M:
       a) Compute negative gradient (pseudo-residuals)
       b) Fit weak learner $h_t(x)$
       c) Update model: $F_t(x) = F_{t-1}(x) + \eta \cdot h_t(x).\ t = 1 \ldots M$:
    3. Output $FM(x)$.
```

The optimization objective:

$$L(y, f(x)) = E\left[(y - f(x))^2\right] \tag{1}$$

where:
- y - true output (accident/non-accident),
- $f(x)$ - model prediction,
- x - feature vector (pressure, temperature, flow, and vibration).

2.3. Dataset description

The dataset included operational data:
- Source parameters (plant output),
- Pipeline parameters (pressure, temperature, and diameter),
- Maintenance (cleaning schedules),
- Environment (ambient temperature and humidity).

Normalization was applied, with training/test split and 5-fold cross-validation.

3. RESULTS

3.1. Conceptual framework

Figure 1 illustrates the developed conceptual framework for predictive maintenance of HSNs using ML algorithms. The framework includes four key stages:

- Initialization and normalization of input parameters,
- Forecast generation for specified time intervals,
- Visualization of predicted outcomes for operator interpretation,
- Decision support for proactive maintenance planning.

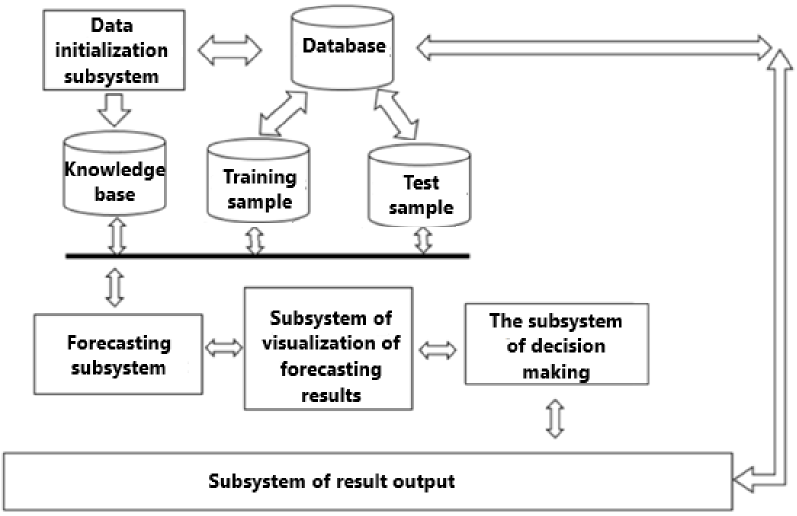

Data initialization subsystem
Database
Knowledge base
Training sample
Test sample
Forecasting subsystem
Subsystem of visualization of forecasting results
The subsystem of decision making
Subsystem of result output

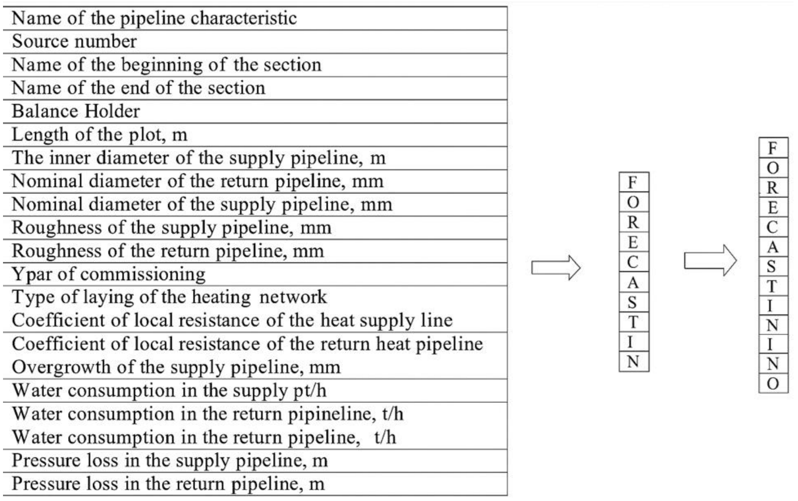
Figure 1. Conceptual scheme of predicting and decision-making

This structure emphasizes the integration of data-driven models into operational workflows to enable real-time monitoring and risk assessment. The reliability of such a system fundamentally depends on the quality of the dataset, which was carefully curated and split into training and test samples to preserve class distributions.

3.2. Dataset formation

As shown in Figure 2, the training dataset was formed by grouping features into categories:

- Source characteristics (e.g., thermal power plant output),
- Pipeline characteristics (e.g., pressure, temperature profiles),
- Maintenance parameters (e.g., chemical cleaning schedules),
- Environmental factors (e.g., ambient temperature, humidity).



Name of the pipeline characteristic
Source number
Name of the beginning of the section
Name of the end of the section
Balance Holder
Length of the plot, m
The inner diameter of the supply pipeline, m
Nominal diameter of the return pipeline, mm
Nominal diameter of the supply pipeline, mm
Roughness of the supply pipeline, mm
Roughness of the return pipeline, mm
Ypar of commissioning
Type of laying of the heating network Coefficient of local resistance of the heat supply line
Coefficient of local resistance of the return heat pipeline Overgrowth of the supply pipeline, mm
Water consumption in the supply pt/h
Water consumption in the return pipineline, t/h Water consumption in the return pipeline, t/h
Pressure loss in the supply pipeline, m
Pressure loss in the return pipeline, m

Figure 2. Dataset structure categories

This categorization facilitates the identification of high-risk areas prone to failures, such as:

- Pipeline sections exceeding critical pressure levels,
- Zones with cavitation risks due to gravity-induced flow,
- Regions where vaporization may occur from insufficient pressure,
- Subscriber buildings facing insufficient heat delivery during peak loads.

By incorporating such domain-specific knowledge, the dataset ensures realistic representation of operational conditions, a critical factor in achieving robust model performance.

3.3. Comparative analysis of machine learning models

Table 1 presents the comparative performance of four ML models evaluated using accuracy and F1-score metrics. GB (LightGBM) demonstrated superior predictive capability with an accuracy of 85% and an F1-score of 0.82, outperforming LR (82% accuracy, F1=0.80), decision trees (79% accuracy, F1=0.76), and linear regression (76% accuracy, F1=0.74).

Table 1. Comparison of models

Algorithms	Accuracy (%)	AUC	F1-score
GB	85	0.96	0.82
LR	82	0.98	0.80
Decision tree	79	0.92	0.76
Linear regression	76	0.90	0.74

Figure 3 shows the ROC curve for four machine learning models applied to the problem of anomaly detection in heating networks. The X-axis shows the false positive rate (FPR), and the Y-axis shows the true positive rate (TPR). The diagonal line corresponds to a random classifier and serves as a benchmark for assessing the quality of the models.

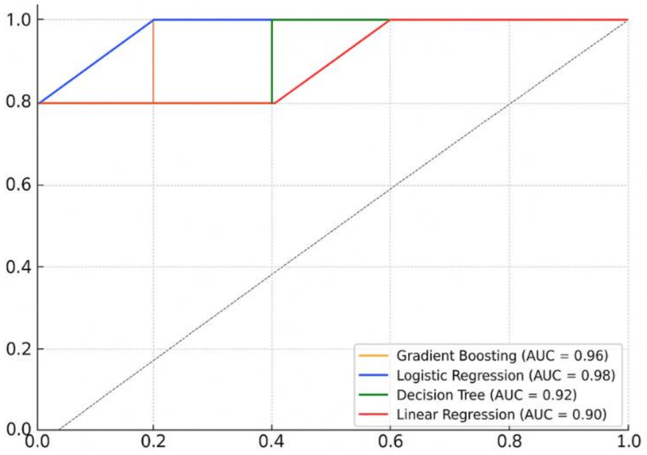


Figure 3. ROC curves of ML models

The curves show that all algorithms demonstrate predictive ability above chance. The logistic regression model demonstrates the best classification quality, with the highest area under the curve (AUC=0.98), followed by GB (AUC=0.96), decision tree (AUC=0.92), and linear regression (AUC=0.90). The closer a model's ROC curve is to the upper left corner of the graph, the higher its accuracy; thus, logistic regression and boosting demonstrate the most consistent ability to detect anomalies. Figure 4 provides a visualization of model performance across metrics, highlighting the stability and higher precision-recall balance achieved by GB.

3.4. LightGBM predictions with technical parameters

The strong performance of LightGBM can be attributed to its ability to capture complex nonlinear relationships and effectively handle class imbalance via its GOSS and EFB techniques. The relatively high F1-score (0.82) indicates a good balance between precision and recall, which is critical in pre-emergency detection where both false positives (unnecessary maintenance) and false negatives (missed failures) have significant operational consequences.

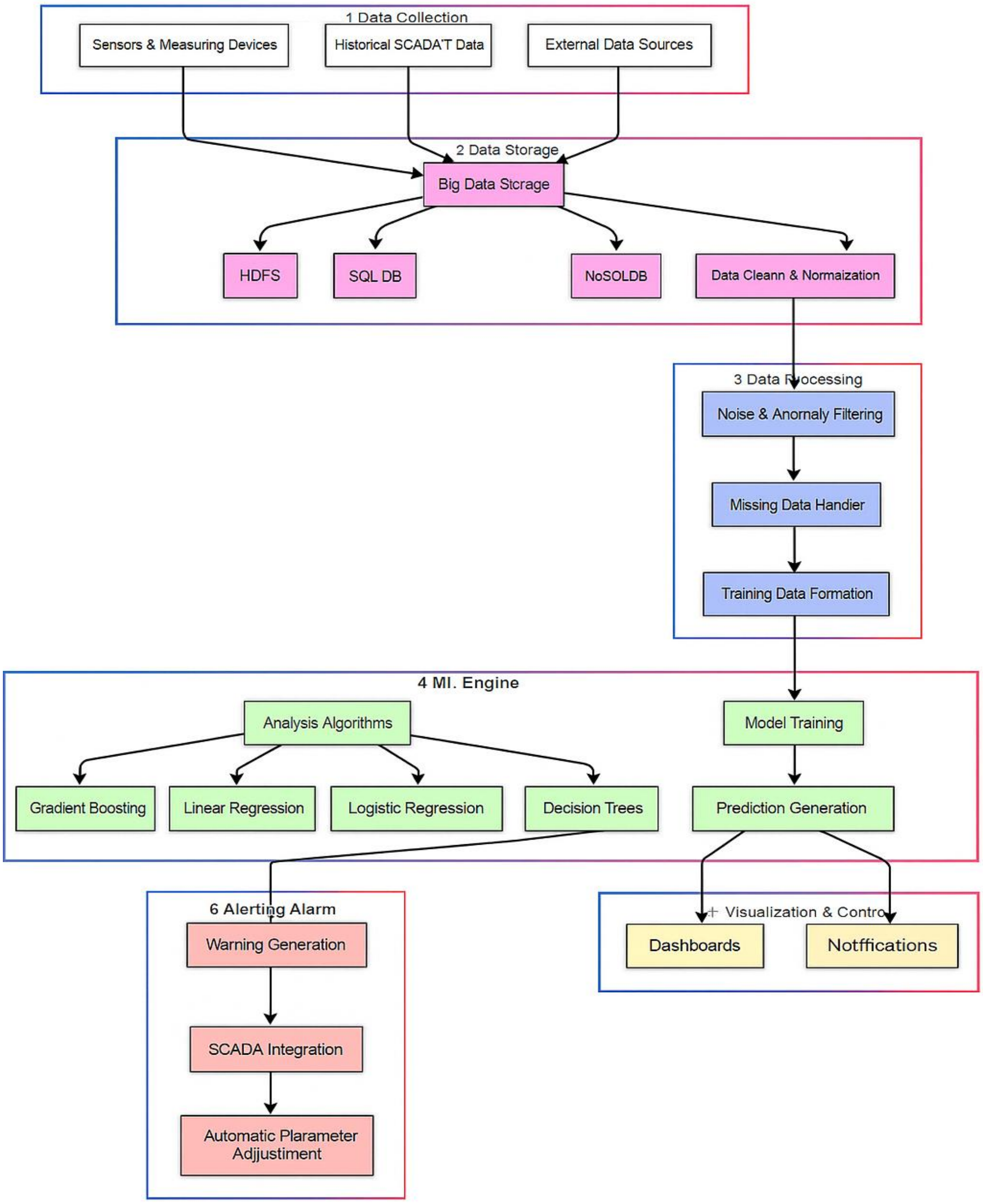


Figure 4. Architecture of predictive maintenance system

Compared to other studies, Xue *et al*. [24] reported a 78% accuracy using SVM on district heating fault datasets, and Xue *et al*. [25] achieved 80% accuracy using RF. Thus, the LightGBM approach in this study surpasses these benchmarks, demonstrating improved generalization.

However, the study's reliance on a synthetic dataset introduces potential limitations. While the data reflects realistic operational patterns, the absence of real-world noise and unmodeled system behaviors could lead to overestimated performance. Future work should validate the model on actual operational datasets from district heating systems to assess robustness.

3.5. Results of LightGBM classification

Table 2 provides example predictions from the LightGBM model, demonstrating its classification of operational states (accident/non-accident) based on input parameters. This table illustrates how LightGBM correctly identifies pre-emergency states, enabling timely intervention to mitigate potential failures. The results highlight the potential of LightGBM-based predictive maintenance frameworks to enhance the reliability of HSNs. With its superior accuracy and interpretability (via SHAP analysis), the proposed system can support decision-making processes, reduce unplanned downtimes, and optimize resource allocation.

Table 2. LightGBM classification with technical parameters

Pressure (MPa)	Temperature (°C)	Flow rate (m^3/h)	Result	Confidence
1.2	85	Accident	Accident	0.91
2.5	70	No accident	No accident	0.87
3.8	90	Accident	Accident	0.93
4.1	75	No accident	No accident	0.89
5.4	95	Accident	Accident	0.95

4. CONCLUSION

This study developed a LightGBM-based predictive framework for HSNs, achieving accuracy=85%, F1=0.82, ROC-AUC=0.96. The model outperforms traditional approaches, confirming its suitability for predictive maintenance. Future work: validation on real operational datasets from different HSNs, field trials across diverse heating networks, IoT integration with real-time sensors, and development of lightweight real-time implementations for deployment.

ACKNOWLEDGMENTS

This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No BR24992907 for 2024-2026).

FUNDING INFORMATION

The authors acknowledge financial support from the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. BR24992907, 2024–2026), which funded this research.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Dauren Darkenbayev	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Gulnar Balakayeva	✓	✓			✓	✓		✓	✓	✓		✓	✓	✓
Uzak Zhapbasbayev	✓			✓			✓		✓	✓			✓	✓
Mukhit Zhanuzakov	✓		✓	✓			✓		✓	✓	✓			

C	: **C**onceptualization	I	: **I**nvestigation	Vi	: **Vi**sualization
M	: **M**ethodology	R	: **R**esources	Su	: **Su**pervision
So	: **So**ftware	D	: **D**ata Curation	P	: **P**roject administration
Va	: **Va**lidation	O	: Writing -**O**riginal Draft	Fu	: **Fu**nding acquisition
Fo	: **Fo**rmal analysis	E	: Writing - Review &**E**diting		

CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict of interest regarding the publication of this paper.

DATA AVAILABILITY

All data supporting the findings of this study on heating network emergency prediction are included within the article and its supplementary materials.

REFERENCES

[1] D. Stecher *et al.*, "Creating a labeled district heating data set: From anomaly detection towards fault detection," *Energy*, vol. 313, pp. 1-13, 2024, doi: 10.1016/j.energy.2024.134016.

[2] S. Aslam, P. P. Aung, A. S. Rafsanjani, and A. P. P. A. Majeed, "Machine learning applications in energy systems: current trends, challenges, and research directions," *Energy Informatics*, vol. 8, no. 52, 2025, doi: 10.1186/s42162-025-00524-6.

[3] O. R. Ajao, "Optimizing Energy Infrastructure with AI Technology," *Open Journal of Applied Sciences*, vol. 14, no. 12, pp. 3516-3544, 2024, doi: 10.4236/ojapps.2024.1412230.

[4] G. A. Susto, A. Schirru, S. Pampuri, S. McLoone, and A. Beghi, "Machine learning for predictive maintenance: A multiple classifier approach," in *2016 IEEE Asian Hardware-Oriented Security and Trust (AsianHOST)*, Yilan, Taiwan, 2016, pp. 1-6, doi: 10.1109/AsianHOST.2016.7835555.

[5] K. Vahldiek, B. Rüger, and F. Klawonn, "Leakages in District Heating Networks—Model-Based Data Set Quality Assessment and Localization," *Sensors*, vol. 22, no. 14, 2022, doi: 10.3390/s22145300.

[6] W. Sun, "Anomaly Detection Analysis for District Heating Apartments," *Journal of Applied Science and Engineering*, vol. 21, no. 1, 2018, doi: 10.6180/jase.201803_21(1).0005.

[7] M. Tang, Y. Li, H. Zhang, S. Wang, and X. Chen, "Cost Sensitive LightGBM Based Online Fault Detection Framework for Wind Turbine Gearboxes," *Frontiers in Energy Research*, vol. 9, 2021, doi: 10.3389/fenrg.2021.701574.

[8] X. Lv, F. Liu, M. Jiang, and L. Jia, "Fault diagnosis of power transformers based on dissolved gas analysis and improved LightGBM hybrid integrated model with dual branch structure," *IET Electric Power Applications*, vol. 18, no. 12, pp. 2008-2020, 2024, doi: 10.1049/elp2.12528.

[9] R. Song, Y. Liu, and H. Zhou, "An Improved LightGBM Based Method for Series Arc Fault Detection in Photovoltaic Grid Connected Systems," *Electronics*, vol. 14, no. 18, pp. 1-25, 2025, doi: 10.3390/electronics14183593.

[10] J. Wang, J. Chi, Y. Ding, H. Yao, and Q. Guo, "Based on PCA and SSA LightGBM oil immersed transformer fault diagnosis method," *PLOS ONE*, vol. 20, no. 2, pp. 1-14, 2025, doi: 10.1371/journal.pone.0314481.

[11] M. Rahal, B. S. Ahmed, R. Renström, R. Stener, and A. Würtz, "Data Driven Heat Pump Management: Combining Machine Learning with Anomaly Detection for Residential Hot Water Systems," *Neural Computing Applied*, pp. 1-27, 2025, doi: 10.1007/s00521-025-11318-y.

[12] K. Ukoba, "Optimizing renewable energy systems through artificial intelligence," *Energy & Environment*, vol. 35, no. 7, pp. 3833-3879, 2024, doi: 10.1177/0958305X241256293.

[13] H. Bahlawan *et al*., "Detection and identification of faults in a District Heating Network," *Energy Conversion and Management*, vol. 266, p. 115837, 2022, doi: 10.1016/j.enconman.2022.115837.

[14] P. Koul and Y. Siddaramu, "Machine Learning in Production Engineering: A Comprehensive Review," *International Journal of Multidisciplinary Research in Arts, Science and Technology*, vol. 3, no. 5, May 2025, doi:10.61778/ijmrast.v3i5.131

[15] A. Altybay, A. Nakiskhozhayeva, and D. Darkenbayev, "Numerical Simulation and Parallel Computing of the Acoustic Wave Equation," *AIP Conferences Proceedings*, vol. 3085, no. 1, pp. 1-7, 2024, doi: 10.1063/5.0194676.

[16] D. Darkenbayev, A. Altybay, Z. Darkenbayeva, and N. Mekebayev, "Intelligent Data Analysis on an Analytical Platform," *Informatyka, Automatyka, Pomiary w Gospodarcei Ochronie Srodowiska*, vol. 14, no. 1, pp. 119–122, 2024, doi: 10.35784/iapgos.5423.

[17] J. van Dreven, V. Boeva, S. Abghari, H. Grahn, J. Al Koussa, and E. Motoasca, "Intelligent Approaches to Fault Detection and Diagnosis in District Heating: Current Trends, Challenges, and Opportunities," *Electronics*, vol. 12, no. 6, 2023, doi: 10.3390/electronics12061448.

[18] C. Hermans, J. Al Koussa, T. Van Oevelen, and D. Vanhoudt, "Fault detection for district heating substations: Beyond three sigma approaches," *Smart Energy*, vol. 1, pp. 1-9, 2024, doi: 10.1016/j.segy.2024.100159.

[19] L. Manservigi, H. Bahlawan, E. Losi, M. Morini, P. R. Spina, and M. Venturini, "A diagnostic approach for fault detection and identification in district heating networks," *Energy*, vol. 251, 2022, doi: 10.1016/j.energy.2022.123988.

[20] A. Rafati and H. R. Shaker, "Predictive maintenance of district heating networks: A comprehensive review of methods and challenges," *Thermal Science and Engineering Progress*, vol. 53, pp. 1-20, 2024, doi: 10.1016/j.tsep.2024.102722.

[21] N. Dimitropoulos *et al*., "Forecasting of short-term PV production in energy communities through Machine Learning and Deep Learning algorithms," in *2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA)* Chania, Crete, Greece, 2021, pp. 1-6, doi: 10.1109/IISA52424.2021.9555544.

[22] T. Capotosto, A. R. di Fazio, S. Perna, F. Conte, G. Iannello, and P. de Falco, "Day-ahead forecast of PV systems and end-users in the contest of renewable energy communities," in *2022 AEIT International Annual Conference (AEIT)*, Rome, Italy, Oct. 2022, pp. 1-6, doi: 10.23919/AEIT56783.2022.9951849.

[23] D. Mazzeo *et al*., "Artificial intelligence application for the performance prediction of a clean energy community," *Energy*, vol. 232, p. 120999, 2021, doi: 10.1016/j.energy.2021.120999.

[24] P. Xue *et al*., "Fault detection and operation optimization in district heating substations based on data mining techniques," *Applied Energy*, vol. 205, pp. 926–940, 2017, doi: 10.1016/j.apenergy.2017.08.035.

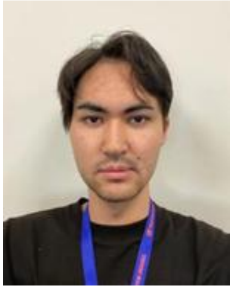
[25] P. Xue, Y. Jiang, Z. Zhou, X. Chen, X. Fang, and J. Liu, "Machine learning-based leakage fault detection for district heating networks," *Energy and Buildings*, vol. 223, p. 110161, Sep. 2020, doi: 10.1016/j.enbuild.2020.110161.

BIOGRAPHIES OF AUTHORS

Dauren Darkenbayev Ph.D. associate professor of the Department of Computational Sciences and Statistics, Faculty of Mechanics and Mathematics, Al-Farabi Kazakh National University and Department of Computer Science Kazakh National Women's Teacher Training University, Almaty, Kazakhstan. Research interests: big data processing, mathematical and computer modeling, development of computer systems for the educational process, machine learning, and deep learning in inverse problems. He can be contacted at email: dauren.kadyrovich@gmail.com and dauren.darkenbayev1@gmail.com.

Gulnar Balakayeva Doctor of Physical and Mathematical Sciences, Professor of the Department of Computer Science, Faculty of Information Technologies and Department of Computational Sciences and Statistics, Faculty of Mechanics and Mathematics, Al-Farabi Kazakh National University, Almaty, Kazakhstan. Research interests: big data processing, mathematical and computer modeling, and development of computer systems for the educational process. She can be contacted at email: gulnardtsa@gmail.com.

Uzak Zhapbasbayev Doctor of Technical Sciences, Professor of Satpayev University. Head of Laboratory. Almaty, Kazakhstan. Research interests: fundamental and applied problems of fluid and gas mechanics, thermal physics, thermodynamics, mathematical and computer modeling of energy, ecology, and oil industry problems. He can be contacted at email: u.zhapbasbayev@satbayev.university.



Mukhit Zhanuzakov Ph.D. student at the Department of Computer Science, Faculty of Information Technologiesal-Farabi Kazakh National University, Almaty, Kazakhstan. Research interests: big data, algorithms and data structures, software, machine learning, fault-tolerance, and reliability. He can be contacted at email: zhanmuha01@gmail.com.