Machine learning based predictive fault detection in microclimate systems

Nurkamilya Daurenbayeva[1,†], Symbat Mamanova[1,†], Artem Bykov[1,†], Akbota Akim[1,†], Bagdat Yagaliyeva[2,*,†]

[1] International Information Technology University, 34/1 Manas St., Almaty, 050000, Kazakhstan
[2] Satbayev University, 22 Kanysh Satpayev St., Almaty, 050013, Kazakhstan

Abstract

This study examines the application of machine learning methods for monitoring and fault detection in climate control systems. Using a hardware complex to collect data on microclimate parameters such as temperature, humidity, dew point, illumination, voltage and current, an analysis is performed using the CRISP-DM methodology. Standardization techniques Z-score and the principal component method (PCA), as well as clustering and classification methods are used for data processing. The goal is to timely identify and predict deviations in the operation of microclimate systems, which contributes to increased energy efficiency and indoor comfort. Studies have shown that the developed Multilayer perceptron model detects up to 80% of anomalies in microclimate data (recall), while gradient boosting provides lower rates (Precision=0.67 with Recall=0.40) for detecting rare failures. The proposed integrated approach makes it possible to timely diagnose deviations in the operation of microclimate systems and thereby increase the energy efficiency of control systems. Thus, the use of machine learning methods in climate control provides more reliable fault diagnosis.

Keywords

machine learning, microclimate, fault detection, principal component method, clustering, classification, energy efficiency

1. Introduction

Management of indoor microclimate is essential for maintaining energy efficiency and ensuring occupant comfort. In Kazakhstan, buildings account for approximately 43% of total energy consumption, with the residential sector representing the largest share [1]. To improve energy efficiency, it is necessary to develop management strategies that consider multiple environmental parameters, not only temperature, but also humidity, dew point, power, illumination, vibration, voltage, and current[2,3]. An important aspect of this process is the detection and diagnosis of faults in microclimate systems using machine learning methods. These techniques enable real-time prediction and identification of deviations, thereby ensuring more efficient energy use and maintaining optimal indoor comfort.

The problem addressed in this study is the lack of comprehensive solutions capable of processing multiple microclimate parameters in real-time while effectively detecting hidden anomalies. Existing methods [4, 5], either fail to account for noise in the data or focus only on specific aspects of fault detection, limiting their applicability in dynamic and complex environments.

The aim of this study is to apply machine learning methods and statistical approaches for monitoring and detecting faults in indoor microclimates using modern data collection and analysis techniques. The research objectives include reviewing existing approaches, developing a hardware

[1] STIoT 2025: Workshop on Smart Technologies and IoT, November 19-20, 2025, Almaty, Kazakhstan
* Corresponding author.
† These authors contributed equally.
n.daurenbayeva@iitu.edu.kz (N. Daurenbayeva); s.mamanova@iitu.edu.kz (S. Mamanova); a.bykov@iitu.edu.kz (A. Bykov); aakim@iitu.edu.kz (A. Akim); bagdat.yagaliyeva@gmail.com (B. Yagaliyeva)
0000-0003-0341-4017 (N. Daurenbayeva); 0009-0001-7277-5492 (S. Mamanova); 0000-0002-9563-5185 (A. Bykov); 0009-0009-2144-7091 (A. Akim); 0000-0003-4644-2261 (B. Yagaliyeva)

system for data collection, conducting detailed data analysis, developing machine learning models for anomaly detection, and testing the proposed solutions in indoor environments. The scientific novelty lies in proposing a new methodology for fault detection in microclimate control systems using advanced machine learning methods for accurate and timely diagnosis. The practical significance of the work consists in improving energy efficiency and enhancing indoor comfort through the implementation of an intelligent fault detection system.

The novelty of this study lies in the integration of the CRISP-DM methodology with a combined application of Principal Component Analysis (PCA), clustering, and classification techniques for microclimate-related tasks. This approach enhances the efficiency of fault detection and diagnosis by leveraging the strengths of dimensionality reduction, unsupervised learning, and supervised classification within a structured data analysis framework.

The integration of the CRISP-DM methodology with PCA and machine learning algorithms for real-time fault detection.

In study [6], an automated fault detection method for building ventilation systems was developed to identify abnormal energy consumption. Utilizing a dynamic building model and a statistical approach based on the Chernoff bound approximation, the method detects deviations from normal operation and classifies them by urgency levels. A two-month analysis revealed high-urgency periods, leading to the identification of a faulty occupancy counter. This method shows promise in automatically detecting anomalies in ventilation systems.

Modern HVAC diagnostic methods can be categorized into traditional machine learning algorithms, deep neural networks, and hybrid models, which outperform classical approaches. However, their effectiveness heavily depends on data quality, making them highly sensitive to noise and missing values, which limits their applicability in real-world scenarios.

The study hypothesizes that the integration of PCA with supervised learning models can improve the accuracy and robustness of fault detection in dynamic building environments.

In another study [2], researchers assessed indoor comfort conditions in a hospital using quality indicators such as temperature, humidity, air speed, illuminance, and air quality. The model proposed by the ISO 7730 standard was used to assess hygrothermal comfort.

The study A Machine Learning Approach to Microclimate Monitoring and Fault Detection [7] investigated the application of machine learning techniques for microclimate monitoring. It demonstrated the limitations of Principal Component Analysis (PCA) in handling noisy data and proposed a general architecture for a fault detection system. The present work builds upon this research by analyzing classification algorithms for fault diagnosis, offering a comprehensive evaluation of their performance and effectiveness in microclimate control systems.

Furthermore, the integration of machine learning techniques has been explored for predictive maintenance in various industries, including building systems. Machine learning algorithms have been successfully applied to analyze large datasets from sensors to detect early signs of equipment failures, enabling timely maintenance and reducing unexpected downtimes.

These studies highlight the potential of combining advanced data collection systems with machine learning methodologies to enhance the efficiency and reliability of microclimate management in buildings.

2. Materials and research methods

The study utilized microclimate data collected from sensors in residential and non-residential buildings during both the hot and cold seasons. Data was recorded every minute, resulting in over 500,000 entries per season. Prior to analysis, preprocessing steps were applied, including missing data imputation, outlier removal (Z-score method), and normalization. This approach ensured data reliability for further application of machine learning methods in fault detection. Multilayer perceptron (MLP), gradient boosting, and support vector machines (SVM) are widely used in building microclimate control systems to improve energy efficiency and provide comfortable conditions. These machine learning methods are used to predict energy consumption, detect

anomalies in the operation of heating, ventilation, and air conditioning (HVAC) systems, and optimize climate parameters. The methodology for monitoring and fault detection in microclimate parameters using statistical methods and machine learning is illustrated in Figure 1. For example, gradient boosting is used to model energy consumption in buildings, which allows for more accurate load forecasting and efficient resource management. SVM and MLP are used to classify climate system states and detect faults, which facilitates timely maintenance and reduces energy costs. Thus, the integration of these algorithms into climate control systems contributes to the creation of smart buildings with optimal energy consumption and improved living and working conditions.

Data overview

The dataset consisted of approximately 1 million samples collected from 16 sensors (temperature, humidity, CO_2, light, power, voltage, current) across two building types during summer and winter seasons.

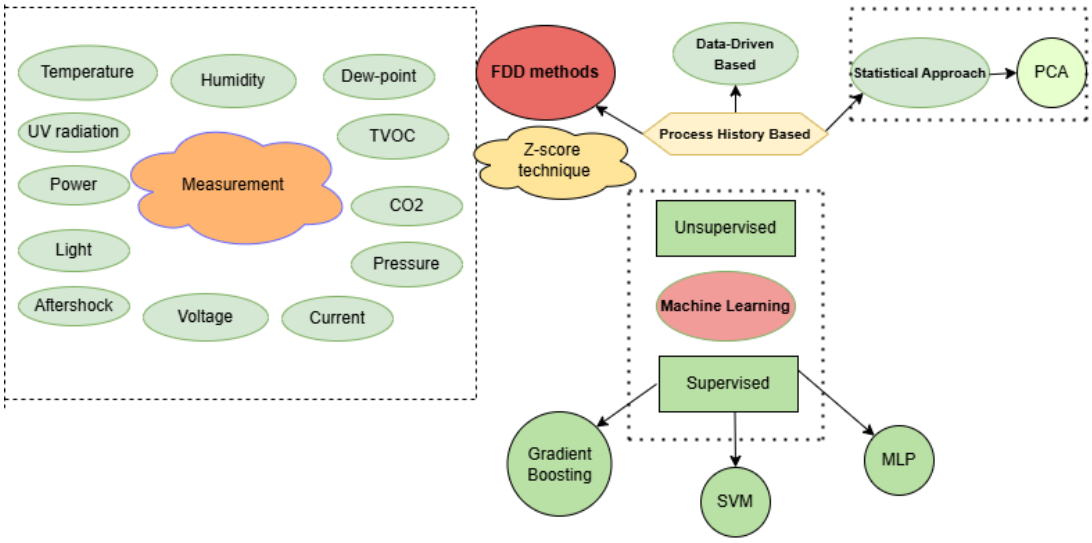

Figure 1: Methodology for Monitoring and Fault Detection in Microclimate Parameters Using Statistical Methods and Machine Learning.

After data cleaning, the mean values of most parameters remained largely unchanged, while standard deviation generally decreased, indicating more stable conditions. Temperature, humidity, and pressure exhibited reduced variability both indoors and outdoors. TVOC and CO2 levels showed decreased standard deviation, with TVOC also exhibiting a reduction in mean concentration, suggesting improved air quality. Electrical parameters (current, voltage, and power) demonstrated enhanced stability. Light, UV radiation, and aftershock data showed significant reductions in standard deviation, reflecting more consistent environmental conditions. Overall, data cleaning resulted in reduced variability, enhancing data reliability and interpretability.

The PCA analysis for residential and non-residential buildings demonstrates that the first two principal components account for the majority of data variance: 59.83% for residential and 61.29% for non-residential buildings. In both cases, the first four components capture nearly all variability, confirming their significance in describing the data structure. However, PCA results can be sensitive to noise, necessitating cautious interpretation of explained variance. Overall, the distribution of explained variance appears reasonable and supports the applicability of PCA for dimensionality reduction in these datasets (Figure 2).

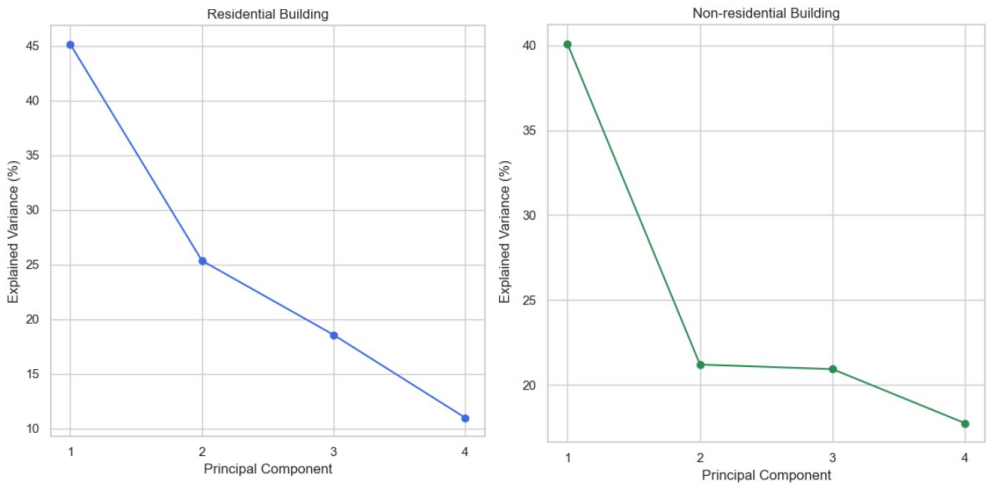

Figure 2: Analysis of Explained Variance Across Principal Components for Buildings.

During the training of the Multi-Layer Perceptron (MLP) model, the following training loss values were recorded at each iteration. These values demonstrate how the model's performance improved as it learned and adjusted its weights through gradient-based optimization techniques, such as stochastic gradient descent.

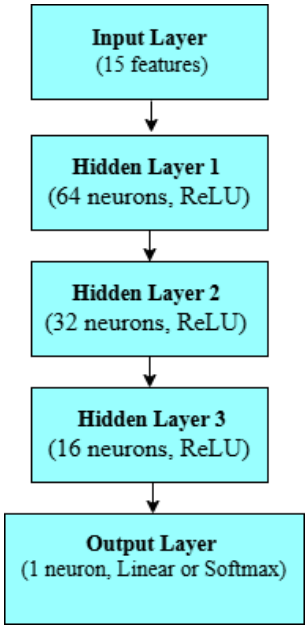

Figure 3: Schematic architecture of a Multilayer Perceptron (MLP).

The image presents a schematic architecture of a Multilayer Perceptron (MLP) designed for analyzing microclimate parameters (Figure 3).

Key elements of the diagram:

- Input layer with 15 neurons, corresponding to microclimate parameters such as temperature, humidity, pressure, CO2, and others;
- Three hidden layers with varying numbers of neurons (e.g., 64, 32, and 16), utilizing the ReLU activation function;
- Output layer with a single neuron, which can perform either regression (linear activation) or classification (softmax);
- Arrows indicating the flow of data between layers.

This MLP architecture enables the model to capture complex dependencies in the data and make accurate predictions.

To evaluate the effectiveness of models in detecting anomalies in microclimate systems, it is crucial to analyze key metrics such as precision, recall, and F1-score for the rare class. Since rare failures can critically impact system performance, their accurate classification is of great importance.

The experimental results indicate that the Multilayer Perceptron (MLP) demonstrated a high recall for rare faults (recall = 80%), highlighting its ability to detect a significant number of failures. However, the model exhibited low precision (precision = 9%), leading to a high number of false positives. In contrast, Gradient Boosting achieved higher precision (67%) but lower recall (40%), suggesting that while it reduces false alarms, it may miss some failures.

This contrast in model performance reflects fundamental differences in their behavior:

1. MLP is highly sensitive to anomalies, making it preferable for tasks where missing a failure is unacceptable, even at the cost of increased false positives.
2. Gradient Boosting provides a more balanced performance, reducing false alarms but potentially overlooking some anomalies. Comparative Analysis of SVM, MLP and Gradient Boosting Classifiers for Fault Detection in Microclimate Systems have shown below (Table 1).

Statistical significance of model differences was assessed using paired t-tests between the F1-scores of compared classifiers.

Therefore, model selection should be based on the specific requirements of the task. If minimizing missed faults is a priority (e.g., in safety-critical monitoring systems), MLP is the better choice. Conversely, if avoiding false alarms is more critical (e.g., in automated control systems), Gradient Boosting is more suitable.

Incorporating this analysis into the study allows for a more objective assessment of the strengths and weaknesses of the models, enabling informed decision-making when selecting an algorithm for monitoring microclimate parameters.

Table 1

Comparative Analysis of SVM, MLP and Gradient Boosting Classifiers for Fault Detection in Microclimate Systems

Metric	SVM	Gradient Boosting	MLP	Comments
Model Accuracy	1.00	1.00	1.00	All models classify dominant classes perfectly
Precision for rare faults (-1)	0.09	0.67	0.09	Gradient Boosting significantly reduces false positives
Recall for rare faults (-1)	0.80	0.40	0.80	SVM and MLP detect more rare faults but have lower precision
F1-score for rare faults (-1)	0.16	0.50	0.16	Gradient Boosting provides a better balance of precision and recall
Macro Average (P/R/F1)	0.70/0.93/0.72	0.89/0.80/0.83	-	SVM emphasizes recall, Gradient Boosting achieves balanced metrics
Weighted Average (P/R/F1)	1.00/1.00/1.00	1.00/1.00/1.00	-	All models classify dominant classes perfectly

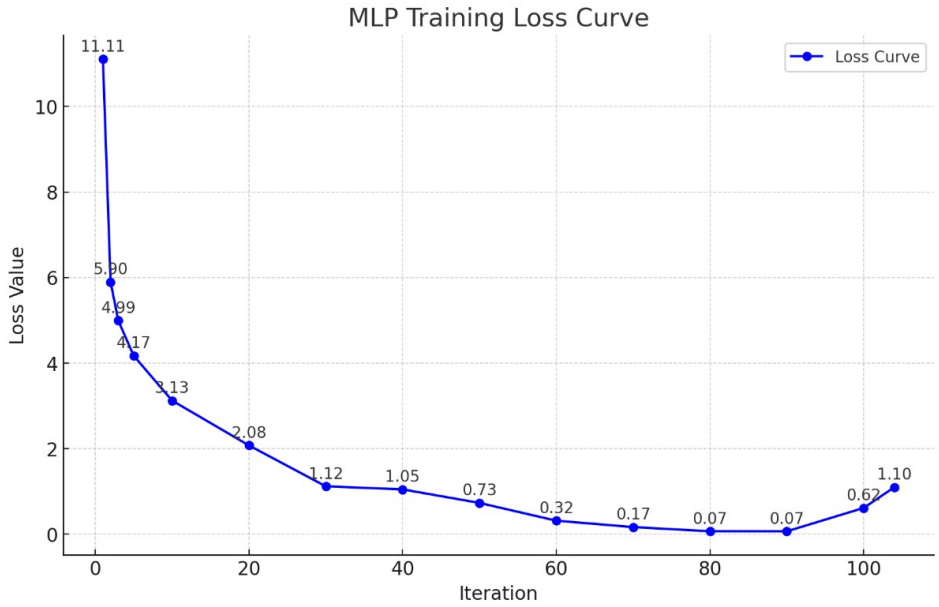

Figure 4: Evolution of loss during one successful training for Building (MLP)

The graph illustrates the change in the loss function during the training of the Multi-Layer Perceptron (MLP) model. The X-axis represents training iterations, while the Y-axis shows the loss function value. At the beginning of training, a high initial loss is observed (Iteration 1) as the model starts adjusting its parameters. Evolution of loss during one successful training for Building (MLP) presents in Figure 4.

Key Training Stages:

- Iteration 1 – High initial loss, the model is just beginning to learn.
- Iteration 2-10 – Significant error reduction, the model learns rapidly.
- Iteration 20-50 – Stabilization, model quality improves.
- Iteration 60-90 – Minimal loss values, the model reaches high accuracy.
- Iteration 100-104 – Loss increases, signs of overfitting appear.

At the end of training, early stopping was applied: «Training loss did not improve more than tol=0.000100 for 10 consecutive epochs. Stopping». This means the training process was halted because the loss value did not improve by more than 0.0001 over 10 consecutive epochs. This mechanism prevents overfitting and allows the model to reach optimal performance without excessive iterations.

Thus, the graph provides a clear visualization of the MLP training process, highlighting the phases of rapid error reduction, stabilization, achievement of high accuracy, and the emergence of overfitting signs in the final iterations.

3. Results

The application of machine learning (ML) techniques in fault detection for indoor microclimate systems has yielded promising results. The study demonstrates that both Support Vector Machines (SVM) and Gradient Boosting classifiers achieve perfect overall accuracy in identifying faults. However, their performance varies when detecting rare faults (class -1). SVM exhibits high recall (0.80) but low precision (0.09), indicating a higher rate of false positives. In contrast, the Gradient Boosting classifier achieves a more balanced performance with a precision of 0.67 and recall of 0.40, resulting in a higher F1-score (0.50) for rare faults.

Gradient Boosting achieved an F1-score of 0.50 for rare fault detection, outperforming MLP (0.16) and SVM (0.16). This indicates its higher balance between precision and recall.

The Principal Component Analysis (PCA) conducted on both residential and non-residential buildings reveals that the first two principal components account for approximately 60% of the data variance. This significant reduction in dimensionality facilitates the identification of patterns and anomalies within the microclimate parameters. However, it is essential to consider that PCA results can be sensitive to noise, necessitating cautious interpretation of the explained variance.

The training process of the Multi-Layer Perceptron (MLP) model indicates a substantial decrease in loss values over iterations, reflecting the model's capacity to learn complex patterns in the data. Nonetheless, the observed increase in loss after a certain point suggests potential overfitting, highlighting the importance of implementing early stopping criteria to maintain model generalization.

These findings underscore the potential of ML methods in enhancing fault detection and energy efficiency in indoor microclimate systems. Future research should focus on refining these models, addressing challenges such as noise sensitivity and overfitting, and exploring their applicability in diverse building environments.

4. Conclusion

In this study, the authors developed a hardware system for collecting data on microclimate parameters, which is integrated with the building management system [8]. Data was collected in real time using various sensors measuring microclimate parameters such as temperature, humidity, pressure, CO_2 concentration, and others. The collected data was processed using the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, which provides a structured approach to data analysis [9].

At the data preparation stage, the Z-score technique was used for normalization, which made it possible to identify and exclude outliers, improving the quality of the analysis. To reduce the dimensionality of the data and identify the main factors affecting the microclimate, the principal component analysis (PCA) was used. The analysis showed that the first two principal components explain 59.83% of the data variance for residential buildings and 61.29% for non-residential buildings, which confirms their significance in describing the data structure. However, given the sensitivity of PCA to noise, the results should be interpreted with caution.

Classification methods including multilayer perceptron (MLP), support vector machine (SVM) and gradient boosting were used to detect faults and predict faults in the microclimate system [10]. MLP is a neural network with 15 neurons in the input layer corresponding to the microclimate parameters, three hidden layers with different numbers of neurons (e.g. 64, 32 and 16) and an output layer with one neuron performing either regression or classification. During the training of MLP, a decrease in the loss function value was observed, indicating an improvement in the model performance.

A comparative analysis of SVM and gradient boosting models for fault detection in microclimate systems showed that both models demonstrate high overall accuracy. However, SVM has high sensitivity (recall) to rare faults, but low accuracy, which leads to a large number of false positives. While gradient boosting provides a more balanced performance for rare faults (faults), with higher accuracy but lower sensitivity.

The clustering methods used to analyze the microclimate data have been described in detail in other works by the authors [11],[12]. These methods allowed to reveal hidden patterns and structures in the data, which contributes to more effective microclimate management in buildings.

Thus, an integrated approach, including data collection and pre-processing, application of dimensionality reduction, classification and clustering methods, allows to effectively detect and predict faults in microclimate systems, ensuring the reliability and sustainability of microclimate management in buildings.

5. Limitations and future work

The proposed fault detection system can be integrated into existing BMS architectures to enable real-time anomaly monitoring. Future work will focus on expanding the dataset, improving noise resistance, and developing explainable AI mechanisms for fault interpretation.

Acknowledgements

This research has been funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP23489999).

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] Y. Mukazhanov, Z. Kamshat, A. Orazbayeva, N. Shayhmetov, and C. Alimbaev, «Microclimate Control in Greenhouses», in *Proc. 17th Int. Multidisciplinary Scientific GeoConference SGEM*, Vienna, Austria, 27–29 Nov. 2017, pp. 699–704. doi: 10.5593/sgem2017/62/S27.089.

[2] D. Camuffo, A. Bernardi, G. Sturaro, and A. Valentino, «The microclimate inside the Pollaiolo and Botticelli rooms in the Uffizi Gallery, Florence», *Journal of Cultural Heritage*, vol. 3, pp. 155–161, 2002. doi: 10.1016/S1296-2074(02)01171-8.

[3] A. Denizopoulou and Z. Andreopoulou, «Monitoring pollution level and microclimate conditions in a naturally ventilated livestock building using open-source device», *Journal of Environmental Protection and Ecology*, vol. 20, pp. 562–570, 2020.

[4] Cannistraro G., Cannistraro M., Galvagno A., Trovato G. (2017). Analysis and measures for energy savings in operating theaters, IJHT, Vol. 35, No. Sp. 1.

[5] A. Nacer, B. Marhic, L. Delahoche, and J. B. Masson, «ALOS: Automatic learning of an occupancy schedule based on a new prediction model for a smart heating management system», *Building and Environment*, vol. 142, pp. 484–501, 2018.

[6] M. Bang, S. Engelsgaard, E. Alexandersen, M. Skydt, H. R. Shaker, and M. Jradi, "Novel Real-Time Model-Based Fault Detection Method for Automatic Identification of Abnormal Energy Performance in Building Ventilation Units," *Energy and Buildings*, vol. 183, pp. 238–251, 2018. doi: 10.1016/j.enbuild.2018.11.006.

[7] N. Daurenbayeva, L. Atymtayeva, A. Nurlanuly, A. Bykov, B. Akhmetov, G. Shuitenov, and U. Turusbekova, «A Machine Learning Approach to Microclimate Monitoring and Fault Detection», *Applied Mathematics and Information Sciences*, vol. 19, no. 2, pp. 327–334, 2025.

[8] F. Farmani, M. Parvizimosaed, H. Monsef, and A. Rahimi-Kian, «A conceptual model of a smart energy management system for a residential building equipped with CCHP system», *Electric Power and Energy Systems*, pp. 523–536, 2021.

[9] K. Dineva, «OSEMN process for working over data acquired by IoT devices mounted in beehives», *Current Trends in Natural Sciences*, 2018.

[10] B. Mateus, M. Mendes, J. T. Farinha, A. B. Martins, and A. M. Cardoso, «Data Analysis for Predictive Maintenance Using Time Series and Deep Learning Models — A Case Study in a Pulp Paper Industry», in *Proc. IncoME-VI and TEPEN 2021*, Springer, Berlin/Heidelberg, Germany, pp. 11–25, 2023.

[11] N. Daurenbayeva, L. Atymtayeva, and A. Nurlanuly, «Choosing the intelligent thermostats for the effective decision making in BEMS», in *Proc. IEEE ICECCO 2023*, pp. 1–4, 2023. doi: 10.1109/ICECCO58239.2023.10147131.

[12] N. Daurenbayeva, A. Nurlanuly, L. Atymtayeva, and M. Mendes, «Survey of Applications of Machine Learning for Fault Detection, Diagnosis and Prediction in Microclimate Control Systems», *Energies*, pp. 1–21, 2023.