ABSTRACT

of the PhD thesis by Daurenbayeva Nurkamilya Aldangarovna on «Smart Fault Detection System for building microclimate control», submitted for the degree of Doctor of Philosophy (PhD) in the EP 8D06102 – Computer and Software Engineering

General characteristics of the research. This dissertation focuses on studying and practically implementing automated fault detection and diagnosis methods for building microclimate systems using machine learning algorithms. The primary emphasis is on clustering and statistical techniques that enable anomaly detection in unlabeled data. To this end, an experimental data collection and analysis model was developed, employing multiparametric sensor modules in both residential and non-residential buildings, following IEA and ASHRAE standards.

Relevance of the research. In recent years, methods of automatic diagnosis for Heating, Ventilation and Air Conditioning (HVAC) systems have been developing particularly intensively. Wang et al. (2024) showed that sensor bias reduces cooling efficiency in data centers and proposed a hybrid approach based on Random Forest and Bayesian inference. Zhao et al. (2023) considered diagnosis under incomplete data and applied IFWA-LSTM (Improved Fuzzy Weighted Average – Long Short-Term Memory) for data recovery. Li et al. (2024) developed Fault-Tolerant Control (FTC) using Bayesian inference, confirming its effectiveness in energy saving and comfort maintenance. Zhang et al. (2024) proposed a transfer learning method based on energy and mass balance, which improved diagnostic quality under different operating conditions.

These studies reflect a global trend: the development of robust diagnostic methods capable of operating with noisy, incomplete, and non-stationary data. However, most such works were carried out either on simulated data or in specialized systems (data centers, large public buildings) with high levels of automation.

In contrast, this research is based on real operational data from residential and non-residential buildings in Kazakhstan, where automation is limited and measurements are characterized by high noise levels and seasonal fluctuations. This predetermined the choice of methods that do not require prior labeling: statistical cleaning (Z-score), dimensionality reduction (PCA), and unsupervised clustering (DBSCAN). Such an approach makes it possible to detect faults even under high uncertainty and ensures practical applicability in Kazakhstan’s residential sector.

Additionally, the use of MTBF (Mean Time Between Failures) and the reliability function R(t) provides the opportunity to quantitatively assess the condition and durability of microclimate systems, which is particularly important for autonomous monitoring and fault prevention.

The relevance of this study was clarified in the Address to the People of Kazakhstan by the President of the Republic of Kazakhstan K.K. Tokayev dated

September 8, 2025. The document highlighted the difficult state of housing and communal infrastructure and the need for its digital transformation. Special attention was paid to the introduction of monitoring and intelligent management systems, as well as the use of artificial intelligence and «Smart cities» technologies to increase energy efficiency and infrastructure sustainability.

State of the problem. Previous studies have demonstrated scientific interest in topics such as intelligent control systems and fault detection. The works of both domestic and foreign authors contributed to the formation of a theoretical foundation in this field. However, many issues remain open, particularly those related to the practical implementation of unsupervised learning methods, processing of multidimensional sensor data, and adaptation of algorithms to local climatic and infrastructural conditions. To date, the level of research in this area, considering the specifics of Kazakhstan, remains extremely limited. Therefore, the topic requires further study, with a focus on identifying new approaches, methods, and mechanisms for the development and functioning of such systems. These factors determined the choice of the dissertation topic, as well as its aim and research objectives.

Although intelligent control and HVAC fault detection research provides a strong theoretical base, much relies on synthetic data, limiting real-world applicability. Studies considering Kazakhstan's unique climatic and infrastructural conditions are limited, underscoring the necessity for localized approaches and methodologies. This gap defines the focus and objectives of the present research.

Research aim and objectives

The purpose of the dissertation is to develop a data-driven approach based on statistical and machine learning methods to assess reliability and detect hidden faults in building microclimate systems of residential and non-residential buildings.

The objectives are:

– Review existing monitoring and fault diagnosis techniques.

– Develop an experimental setup acquisition system using multiparametric sensors.

– Collect, preprocess, and clean sensor data (including outlier removal with Z-score).

– Apply Principal Component Analysis (PCA) for dimensionality reduction and visualization.

– Compare DBSCAN and K-Means clustering for anomaly detection effectiveness.

– Assess system reliability via Mean Time Between Failures (MTBF) and reliability function R(t).

Object and subject of research

Object: Microclimate (HVAC) systems in residential and non-residential buildings.

Subject: Data-driven methods for fault detection and reliability analysis.

Theoretical and methodological framework. The research employs modern machine learning techniques such as DBSCAN clustering, Z-score statistical analysis, PCA for dimensionality reduction, and the CRISP-DM methodology to structure the data analysis pipeline. The implementation uses Python with libraries including scikit-learn, pandas, and matplotlib.

Information base. The study is based on experimentally collected sensor data from residential and non-residential buildings. Parameters measured every 10 seconds include temperature, humidity, CO_2 concentration, volatile organic compounds (TVOC), pressure, electrical current and voltage, power consumption, UV radiation, illumination, among others, using IoT sensor devices.

The scientific novelty lies in the development of a unified diagnostic approach that combines DBSCAN, PCA, and Z-score algorithms for processing unlabeled microclimate data. The CRISP-DM methodology was adapted for multidimensional sensor data, which enhanced the accuracy and robustness of the analysis. The proposed approach enables the detection of deviations from normal operating conditions through outlier identification and cluster formation, as well as the assessment of microclimate system reliability using MTBF and R(t) metrics. The effectiveness of the approach was confirmed through an analysis of data informativeness based on explained variance.

The following scientific results were obtained during the dissertation research, defining its scientific novelty:

– A comprehensive review of modern approaches to microclimate monitoring and fault diagnosis of HVAC systems was conducted, with a focus on data processing algorithms and machine learning methods.

– An experimental microclimate monitoring model was developed and implemented for residential and non-residential buildings using multiparametric sensor modules.

– Monitoring was performed across 11 parameters (temperature, humidity, CO_2, TVOC, illuminance, power consumption, etc.) in accordance with ASHRAE and IEA standards.

– Data collection, preliminary processing, and outlier removal using the Z-score method were carried out, which improved the accuracy of subsequent analysis.

– Principal Component Analysis (PCA) was applied to reduce data dimensionality while preserving feature informativeness and facilitating visualization.

– A comparison of DBSCAN and K-Means clustering algorithms was conducted for anomaly detection in noisy unlabeled data; DBSCAN was shown to have advantages when working with complex multidimensional datasets containing noise points.

– The CRISP-DM methodology was adapted for fault detection tasks in noisy, unlabeled data with minimal preprocessing, including applicability for real-time operation.

– For the first time, the reliability of microclimate systems was assessed based on real data: Residential buildings showed MTBF ≈ 103 hours and R(24) ≈ 79.2%; non-residential buildings showed MTBF ≈ 45 hours and R(24) ≈ 58.6%. This indicates higher system reliability in the residential sector.

Key contributions for defense:

– An experimental setup for the collection and analysis of microclimate sensor data in residential and non-residential buildings was developed and implemented. The setup includes measurements of temperature, humidity, CO_2 levels, illuminance, atmospheric pressure, dew point, air velocity, seismic vibrations (aftershocks), TVOC concentration, ultraviolet radiation (UVr), and electricity consumption in accordance with ASHRAE standards.

– The CRISP-DM methodology was adapted for fault detection in unlabeled real microclimate data with minimal preprocessing. A comparison of DBSCAN and K-Means algorithms was conducted, confirming the advantage of DBSCAN for anomaly detection in multiparametric noisy datasets.

– The reliability of microclimate systems was assessed using MTBF and R(t) metrics, revealing higher operational stability of systems in residential buildings compared to non-residential ones.

Theoretical and practical significance. This dissertation advances theoretical understanding of dimensionality reduction and density-based clustering applied to high-dimensional, unlabeled microclimate sensor data. Practically, it proposes a cost-effective, real-time monitoring model enabling early fault detection to reduce energy waste, equipment wear, and improve occupant comfort. The methodology is suitable for integration into existing Building Management Systems, supporting sustainable and energy-efficient smart building development.

Dissemination and publications. The main findings and scientific contributions of this research were presented and discussed at seminars held by the Department of Computer Engineering at the International University of Information Technologies (2021–2025), the Department of Information Systems at Suleyman Demirel University (SDU, 2024–2025), the Polytechnic Institute of Coimbra, Portugal (2023–2024), and the Department of Artificial Intelligence at the Financial University under the Government of the Russian Federation (2024–2025). The author also participated in the 17th International Conference on Electronics, Computer and Computation (ICECCO), 2023 and International conference on Physical Asset Management and Data Science, (PAMDAS) 2025.

The main results of the research were presented in the following works:

1. **Daurenbayeva, N**., Nurlanuly, A., Atymtayeva, L., Mendes, M. Survey of Applications of Machine Learning for Fault Detection, Diagnosis and Prediction in Microclimate Control Systems. Energies 2023, *16*, 3508. https://doi.org/10.3390/en16083508.

2. **Daurenbayeva, N.,** Atymtayeva, L., Nurlanuly, A., Bykov, A., Akhmetov, B., Shuitenov, G., Turusbekova, U. A Machine Learning Approach to Microclimate Monitoring and Fault Detection. AMIS 2025, 19, 327-334. https://doi.org/10.18576/amis/190209.

3. **Дауренбаева, Н.А.,** Нұрланұлы, А., Атымтаева, Л.Б., Быков, А.А., Ергалиев, Д.С., Әбдірашев, Ө.К. Микроклимат параметрлерін кластеризациялау: әдістер мен математикалық сипаттамалар // ENU Bulletin (Л.Н. Гумилев ЕНУ Хабаршысы), Technical Sciences And Technology Series, №4 (149)/ 2024 pp. 202-214. https://doi.org/10.32523/2616-7263-2024-149-4-202-214.

4. **Daurenbayeva, N.,** Atymtayeva, L., Nurlanuly, A. 17th International Conference on Electronics Computer and Computation (ICECCO-2023). Choosing the intelligent thermostats for the effective decision making in BEMS. 1-4. 10.1109/ICECCO58239.2023.10147131.

5. **Daurenbayeva, N.,** Atymtayeva, L., Mendes, M., Nurlanuly, A. & Yagalieva B., (2025, July 17–18). Machine learning approach to fault detection in microclimate system at residential and non-residential buildings. Paper presented at the PAMDAS 2025 – International Conference on Physical Asset Management and Data Science, Coimbra Institute of Engineering (ISEC), Polytechnic University of Coimbra, Portugal.

6. **Daurenbayeva, N.A.,** Atymtayeva, L.B., Lutsenko, N.S., Nurlanuly A. Integration of machine learning for microclimate management optimization in buildings: perspectives and opportunities. International Journal of Information and Communication Technologies, 2024. Vol. 5. Is. 2. https://doi.org/10.54309/IJICT.2024.18.2.008.

7. **Дауренбаева, Н.А.,** Атымтаева, Л.Б., Ыбытаева, Г.С., Нұрланұлы, А. Свидетельство на право охраны программы для ЭВМ № 41781 Республики Казахстан. Аппаратный комплекс для реального мониторинга параметров микроклимата с интегрированным датчиком сейсмического воздействия / заявка 04.01.2024; публикация 05.01.2024.

Chapter overview

Chapter 1 discusses the theoretical and practical aspects of building microclimate control. It covers key concepts, the importance of microclimate for comfort and energy efficiency, and major challenges in its regulation. An overview of modern Building Management Systems (BMS) is provided.

Chapter 2 justifies the choice of the CRISP-DM methodology for structuring the data analysis process. Existing approaches are reviewed, and the implementation of this methodology within the research is described.

Chapter 3 explores algorithms for fault detection and diagnosis in microclimate systems, including machine learning methods. It explains the working principles of the DBSCAN algorithm and the use of Principal Component Analysis (PCA) for dimensionality reduction.

Chapter 4 presents the architecture of the experimental setup, including the description of residential and non-residential premises, equipment used, and sensor placement strategy for data collection.

Chapter 5 focuses on the analysis of the collected data: monitored microclimate parameters are described, visualization methods are presented, and correlations between variables are analyzed.

Chapter 6 outlines the data preprocessing steps, including cleaning, normalization, and handling of missing values, which ensured the reliability of subsequent modeling.

Chapter 7 contains the main results of clustering and fault diagnosis using PCA, K-means, and DBSCAN algorithms. Interpretations and comparisons of the results are provided.

The Discussion section analyzes the results, compares them with other studies, and outlines future application prospects.

The Conclusion summarizes the research, highlights key findings, and offers practical recommendations for the implementation of intelligent microclimate diagnostic systems in buildings.

Structure and Volume of the thesis

The dissertation includes an introduction, seven main chapters, a literature review, a description of the methodological framework, presentation of results with subsequent discussion, and a conclusion. Additionally, the work contains an appendix presenting supplementary materials that complement the main content of the study. The dissertation contains 30 illustrations and 17 tables. The total length of the main text is 90 pages, excluding the appendices.