International Information Technology University

UDC: 004.8:697.9 On manuscript right

DAURENBAYEVA NURKAMILYA ALDANGAROVNA

Smart Fault Detection System for building microclimate control

8D06102 – Computer and Software Engineering

Thesis for the degree of doctor of
Philosophy (PhD)

Scientific consultant
Doctor of Phys.- Math.sc.
Associate Professor, SDU University
Atymtayeva L.B.
Foreign consultant
Coordinator Professor, Polytechnic University of Coimbra
Mateus Mendes

Republic of Kazakhstan
Almaty, 2025

CONTENT

NORMATIVE REFERENCES

This thesis uses references to the following standards:

Instructions for the preparation of a thesis and an abstract, Higher Attestation Commission of the Ministry of Education and Science of the Republic of Kazakhstan dated September 28, 2004 No. 377-3 y.

GOST 7.32-2001. Report on research work. Structure and design rules.

GOST 7.1-2003. Bibliographic record. Bibliographic description. General requirements and rules of compilation.

ST RK 34.005-2002. Information Technology. Basic terms and definitions (first edition).

ST RK. 34.015-2002. Information Technology. Set of standards for automated systems. Terms of reference for creating an IS (first edition).

ST RK 34.027-2006. Information Technologies. Classification of software tools (first edition).

ST RK 34.014-2002. Information Technology. A set of standards for automated systems. Automated systems. Terms and definitions.

DESIGNATIONS AND ABBREVIATIONS

AI	Artificial Intelligence
ARIMA	Autoregressive Integrated Moving Average
ASHRAE	American Society of Heating, Refrigerating and Air-Conditioning Engineers
BMS	Building Management System
C	Current
CO_2	Carbon dioxide
CO_{2out}	Carbon dioxide outside
CRISP-DM	Cross-Industry Standard Process for Data Mining
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DP	Dew-point
DP_{out}	Dew-point outside
FDD	Fault Detection and Diagnosis
HVAC	Heating, Ventilation, and Air Conditioning
IEA	International Energy Agency
KDD	Knowledge Discovery in Databases
L	Light
LSTM	Long Short-Term Memory
ML	Machine Learning
MLP	Multilayer Perceptron
OSEMN	Obtain, Scrub, Explore, Model, Interpret
P	Pressure
P_{out}	Pressure outside
PCA	Principal Component Analysis
T	Temperature
TVOC	Total Volatile Organic Compounds
V	Voltage
H	Humidity
H_{out}	Humidity outside

DEFINITIONS

Aftershock (A) - weak earthquakes following a main earthquake, potentially affecting the stability of structures.

BME280 (Atmospheric Pressure, Humidity, and Temperature Sensor) - a sensor that measures atmospheric pressure, humidity, and temperature with high accuracy, commonly used for weather stations and environmental monitoring.

BH1750 (Digital Light Sensor) - a sensor that measures ambient light intensity in lux, used to assess lighting conditions in environments such as smart lighting systems.

CCS811 (Digital Air Quality Sensor) - a sensor that detects the concentration of CO_2 and total volatile organic compounds (TVOCs) in the air, useful for air quality monitoring.

CO_2 - Carbon dioxide, important for the climate and respiration, harmful at high concentrations.

CO_2 outside (CO_{2out}) - the level of carbon dioxide outside, indicating air pollution.

Current (C) - the flow of electric charge through a conductor, necessary for devices to function.

Dew-point (DP) - the temperature at which water vapor begins to condense into water.

Dew-point outside (DP_{out}) - the temperature outside at which condensation of moisture begins.

ESP8266 Module - a low-cost Wi-Fi microchip with full TCP/IP stack and microcontroller capabilities, used to connect devices to the internet for IoT applications.

Fault - means a «malfunction», «error», or «failure» that disrupts system operation. In building microclimate systems, faults include sensor failures, deviations in temperature or humidity, and control errors affecting comfort and energy efficiency.

Humidity (H) - the amount of water vapor in the air, affecting the perception of temperature.

Humidity outside (H_{out}) - the water vapor content in the air outside, influencing the surrounding environment.

HVAC (Heating, Ventilation, and Air Conditioning) is an acronym that refers to systems responsible for heating, ventilation, and air conditioning. These systems regulate temperature, humidity, and air quality within buildings to create comfortable and safe indoor environments for living and working.

Light (L) - electromagnetic radiation visible to the human eye, important for visibility and plant growth.

ML8511 (UV Sensor) - a sensor designed to measure ultraviolet (UV) radiation levels, typically used for monitoring UV exposure and environmental conditions related to UV light.

Microclimate parameters - essential environmental factors like temperature, humidity, air quality, light intensity, airflow, CO2 levels, and nutrient levels,

monitored and controlled within a microclimate system to create an optimal environment for various applications.

Microclimate system - a specialized setup controlling and regulating environmental conditions within a confined space, using components like sensors, heaters, and ventilation, to maintain specific parameters such as temperature and humidity for various applications.

MPU6050 (3-Axis Gyroscope and Accelerometer) - a sensor that combines a 3-axis gyroscope (measuring rotational movement) and a 3-axis accelerometer (measuring linear acceleration), often used in motion detection and orientation tracking in various applications like drones or fitness devices.

Power (Pw) - the rate at which energy is used, important for the operation of microclimate systems.

Pressure (P) - the force exerted by the air on a surface, dependent on altitude and weather conditions.

Pressure outside (P_{out}) - atmospheric pressure outside, which changes based on weather.

Temperature (T) - a measure of thermal energy in the environment, affecting comfort and health. Temperature outside - the air temperature outside at a given moment.

TVOC - volatile organic compounds that pollute the air and affect health.

UV-radiation (UVr) - A part of the light spectrum, can be harmful to health with prolonged exposure.

Voltage (V) - the difference in electrical potential, which causes current to flow.

Note on Terminology:

In this work, the author translates the English term **«fault»** into Russian as **«ошибка»**. This translation is used to describe any deviation from normal operation in the building microclimate control system, including faults, failures, and anomalies affecting system performance.

INTRODUCTION

General characteristics of the research

This thesis 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. (2023) [1] 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) [2] considered fault diagnosis in BAS-based HVAC systems under conditions of incomplete sensor data and proposed an Improved Fireworks Algorithm–Long Short-Term Memory (IFWA-LSTM) approach to recover missing measurements before applying ICA-KNN for fault detection. Li et al. (2024) [3] developed Fault-Tolerant Control (FTC) using Bayesian inference, confirming its effectiveness in energy saving and comfort maintenance. Zhang et al. (2024) [4] 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.

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.

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 [5].

Research aim and objectives

The aim is to develop and validate machine learning and statistical methods for reliable fault detection in microclimate systems of residential and non-residential buildings to improve energy efficiency.

The objectives are:

– Review existing monitoring and fault diagnosis techniques.

– Develop an experimental data 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.

Scientific novelty

– Developed a unified diagnostic framework combining DBSCAN, PCA, and Z-score to process high-frequency, unlabeled microclimate data.

– Adapted CRISP-DM methodology to Kazakhstan's building specifics and high-dimensional sensor datasets.

– Proposed a method to detect operational deviations through outlier identification and clustering.

– Introduced quantitative reliability assessment using MTBF and reliability function R(t).

– Improved analysis accuracy and visualization via data cleaning and dimensionality reduction.

Scientific results

– Comprehensive review of modern microclimate monitoring and fault detection, emphasizing machine learning.

– Developed and deployed an experimental multiparametric sensor monitoring system in residential and non-residential buildings.

– Monitored 11 key parameters following ASHRAE and IEA standards.

– Applied Z-score based outlier removal to enhance data quality.

– Used PCA for dimensionality reduction, improving data visualization.

– Demonstrated DBSCAN's superiority over K-Means in anomaly detection with noisy, multidimensional unlabeled data.

– Adapted CRISP-DM methodology for real-time fault detection with unlabeled and noisy data.

– Reliability assessment showed residential systems have MTBF ≈ 103 hours and 24-hour reliability ≈ 79.2%, while non-residential systems have MTBF ≈ 45 hours and reliability ≈ 58.6%.

Key contributions for defense

– Developed an experimental microclimate data acquisition model with multiparametric sensors measuring a wide range of parameters per ASHRAE standards.

– Adapted CRISP-DM for fault detection in real-world unlabeled microclimate data with minimal preprocessing.

– Validated DBSCAN's effectiveness and optimized parameters (eps=1.1, min_samples=5) for similar feature spaces.

– Demonstrated higher operational stability of residential microclimate systems via MTBF and reliability metrics.

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 at the PAMDAS 2025 – International Conference on Physical Asset Management and Data Science, Coimbra Institute of Engineering (ISEC), Polytechnic University of Coimbra, Portugal.

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 provides a detailed examination of the theoretical and practical aspects of indoor microclimate management in buildings. It explores the key concepts and the importance of maintaining optimal microclimate conditions for comfort, energy efficiency, and the sustainable operation of engineering systems. The chapter also analyzes the main challenges in controlling indoor environmental parameters and presents modern approaches and technologies for microclimate management, including Building Management Systems (BMS) that enable intelligent monitoring and optimization of indoor conditions.

Chapter 2 discusses the CRISP-DM methodology and its application in the study. The main stages - from understanding the problem to model deployment - are described in detail. The choice of CRISP-DM is justified as a flexible and universal framework, which is further adapted to the specific tasks of microclimate analysis and control in buildings.

Chapter 3 reviews machine learning methods and fault detection algorithms applicable to building microclimate management. It demonstrates that supervised learning methods are effective when labeled data are available, while unsupervised methods are suitable for detecting new or hidden faults. Particular attention is given to the DBSCAN algorithm, which is robust to noise and does not require a predefined number of clusters, as well as to the PCA method, which enables dimensionality reduction and identification of key dependencies among microclimate parameters.

Chapter 4 presents the experimental setup and architecture of the microclimate monitoring system. Experiments were conducted in both residential and non-residential buildings using the NodeMCU (ESP8266) platform and more than 16 sensors measuring temperature, humidity, CO_2, illuminance, and other parameters. The system ensured stable data collection and transmission, enabling the identification of anomalies and analysis of microclimate characteristics in different building types.

Chapter 5 presents the results of data analysis and visualization for microclimate parameters, including temperature, humidity, dew point, pressure, CO_2, TVOC, UV radiation, and power. Data cleaning was performed using the Z-score method, and key correlations were identified: in residential buildings, temperature correlated with dew point and energy consumption, while CO_2 correlated with TVOC; in non-residential buildings, humidity correlated with dew point, and indoor and outdoor pressures were strongly related. The processed data were prepared for further modeling and analysis.

Chapter 6 describes the data preprocessing procedures conducted prior to modeling and analysis. The raw microclimate data, including temperature, humidity, dew point, pressure, CO_2, TVOC, power, current, voltage, illuminance, and UV radiation, were cleaned of missing values, outliers, and duplicates using the Z-score technique. After preprocessing, the standard deviation of most parameters decreased, indicating improved stability and uniformity of the dataset. A comparison of

indicators before and after cleaning confirmed the improved data quality and suitability for machine learning model development.

Chapter 7 presents the results of fault detection in microclimate control systems using machine learning methods. Data from residential and non-residential buildings were analyzed with PCA and clustering algorithms such as K-means and DBSCAN. PCA effectively reduced data dimensionality while preserving essential information and revealed significant differences between building types. Clustering results showed that K-means was ineffective for anomaly detection, whereas DBSCAN successfully identified outliers, including erroneous readings of temperature, pressure, voltage, and illuminance. The findings confirmed the applicability of statistical and intelligent methods for fault diagnosis and reliability enhancement of building microclimate systems.

The **Discussion** section analyzes the obtained results, compares them with existing studies, and outlines the prospects for further research and practical application of the proposed approach.

The **Conclusion** summarizes the main findings of the dissertation and provides practical recommendations for implementing intelligent fault diagnosis systems in building microclimate management.

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 92 pages, excluding the appendices.

In all publications related to the dissertation, the author played the leading role, including the development of the research concept, data analysis, interpretation of the results, and preparation of the articles.

1 CHALLENGES IN BUILDING MICROCLIMATE CONTROL

Microclimate Control as a Global Issue

Over the last decade, rising energy consumption and CO_2 emissions have increased the global focus on energy efficiency. Buildings consume up to 40% of global energy, with residential and commercial sectors using over 60% of electricity. In Kazakhstan, 90% of energy in housing services goes to building operations. Microclimate control -managing temperature, humidity, airflow, and air quality-is crucial for comfort and energy savings. However, most modern systems regulate only temperature. Comprehensive microclimate systems are essential for sustainable environments in buildings, agriculture, and industry, aiming to reduce energy use and system failures [6].

Microclimate Control in Greenhouses

A greenhouse serves as a protective environment for vegetation, shielding it from adverse weather conditions while allowing sunlight penetration. Despite their apparent simplicity, greenhouses constitute intricate systems. Their primary function lies in maintaining internal climates within specific parameters conducive to plant growth. However, achieving this goal is challenged by external factors such as wind speed, solar radiation, temperature, and humidity fluctuations. Microclimate control systems play a pivotal role in mitigating these influences, safeguarding vegetation, optimizing growth, and minimizing risks of disease and pests. Yet, challenges persist, particularly in the diffuse nature of air temperature control methods and the energy-intensive operation of multiple actuators for ventilation, heating, and humidity regulation. Hence, the adoption of efficient energy systems becomes imperative to alleviate operational costs and enhance sustainability [7-9].

Microclimate Control in Hospitals

In recent decades, substantial investments have been directed towards advancing air purification technologies in hospitals, encompassing air disinfection, modern technological solutions, and state-of-the-art equipment. Designing engineering systems for medical institutions poses unique challenges owing to their demanding environments and specific operational requirements. Medical facilities, ranging from general hospitals to specialized clinics and diagnostic centers, necessitate meticulous attention to heating, ventilation, and air conditioning (HVAC) systems due to stringent hygiene standards and the risk of nosocomial infections. The evolving nature of medical technologies further complicates system design, requiring innovative engineering solutions to meet evolving healthcare demands. Moreover, architectural shifts towards more compact and integrated hospital designs present additional challenges in maintaining distinct cleanliness zones and preventing cross-contamination. Consequently, ensuring optimal microclimate conditions, including temperature, humidity, air quality, and airflow, emerges as a critical concern in hospital design, renovation, and ongoing maintenance efforts. Despite these challenges, ongoing research endeavors and adherence to established protocols and

standards aim to enhance microclimate control in hospitals, promoting safety, comfort, and well-being for both patients and healthcare providers [9-13].

Microclimate optimization in occupied spaces focuses on managing key environmental factors - such as temperature, humidity, CO_2 levels, and pollutants- through advanced HVAC systems. Effective control of these parameters enhances human health, comfort, and productivity while improving energy efficiency. Challenges include uneven indoor conditions, external environmental influences, and technical limitations. Scientific research worldwide supports the development of adaptive, sustainable microclimate control solutions. Collaboration and technological integration are essential for creating healthier, more comfortable, and energy-efficient indoor environments.

1.1 Building Microclimate Management

Building microclimate management refers to the process of controlling and regulating the indoor environment within buildings to ensure optimal conditions for occupants in terms of temperature, humidity, air quality, and other factors. This management is crucial for maintaining comfort, health, and productivity levels within buildings while also striving for energy efficiency and environmental sustainability.

Key aspects of building microclimate management include:

Temperature Regulation: Maintaining a comfortable temperature range within buildings, typically between 20 and 25 °C (68-77°F), depending on factors such as the season, building occupancy, and activities conducted within the space.

Humidity Control: Managing relative humidity levels to prevent discomfort, mold growth, and damage to building materials. The recommended indoor humidity range is typically between 30 and 60%.

Air Quality Monitoring: Ensuring good indoor air quality by monitoring and controlling pollutants such as carbon dioxide (CO_2), volatile organic compounds (VOCs), particulate matter, and other contaminants.

Ventilation: Providing adequate ventilation to supply fresh outdoor air and remove stale indoor air, preventing the buildup of pollutants and maintaining oxygen levels.

Energy Efficiency: Implementing energy-efficient HVAC (Heating, Ventilation, and Air Conditioning) systems, insulation, and building design strategies to minimize energy consumption while still meeting comfort requirements.

Automation and Control Systems: Utilizing building management systems (BMS) or smart technologies to automate and optimize microclimate control processes, including scheduling, setpoint adjustments, and fault detection.

Effective building microclimate management requires a holistic approach that considers the interactions between building systems, occupant behavior, and external environmental conditions. By implementing advanced control strategies, leveraging data-driven insights, and integrating sustainable design principles, building operators can create healthier, more comfortable, and energy-efficient indoor environments.

Building Management Systems (BMS)

BMS are centralized systems designed to monitor and control various building services and systems. These systems are integral to managing the microclimate and overall performance of buildings.

BMS integrates and controls multiple building systems, including HVAC (Heating, Ventilation, and Air Conditioning), lighting, security, fire safety, energy management, and others. By centralizing control and communication, BMS allows for coordinated operation and optimization of these systems.

Key functions of BMS include continuous monitoring of parameters such as temperature, humidity, occupancy, air quality, energy consumption, and equipment status. Based on predefined setpoints and control algorithms, the system adjusts HVAC operation, lighting levels, and other parameters to maintain desired conditions efficiently.

Modern BMS platforms often offer remote access and control capabilities, enabling building operators to monitor and manage building systems from anywhere via web-based interfaces or mobile applications. This facilitates proactive troubleshooting, quick response to alarms, and optimization of system performance remotely.

Energy management is a critical aspect of BMS, with the system optimizing HVAC operation, scheduling equipment usage, implementing demand response strategies, and identifying energy-saving opportunities. By analyzing energy data and implementing conservation measures, BMS helps reduce energy consumption and operating costs.

Advanced BMS platforms incorporate Fault Detection and Diagnostics (FDD) algorithms to detect equipment faults, inefficiencies, and abnormal operating conditions in real-time. This helps prevent system failures, reduce downtime, and optimize maintenance efforts.

Data analytics tools within BMS analyze data collected from building systems, sensors, and meters to identify trends, patterns, and anomalies. Customizable reports and dashboards facilitate informed decision-making and performance tracking.

Scalability and flexibility are essential features of BMS, allowing for easy expansion, upgrades, and integration with emerging technologies. The system should adapt to changes in building infrastructure, occupancy patterns, and operational requirements over time.

Ultimately, BMS aims to enhance occupant comfort, productivity, and well-being by maintaining optimal indoor environmental conditions and ensuring a safe and healthy indoor environment. User-friendly interfaces and personalized control options contribute to a positive occupant experience.

In summary, Building Management Systems are essential for optimizing building performance, enhancing energy efficiency, and providing comfortable, safe, and sustainable indoor environments. Leveraging advanced technology, automation, and data-driven insights, BMS helps buildings operate more efficiently and cost-effectively. The authors in [14] analyzed different types of smart temperature regulators used in Building Energy Management Systems (BEMS) and concluded

that factors such as device placement, sensor sensitivity, and power source are critical for effective building energy management.

1.2 Importance of Microclimate

Microclimate control is a very important feature for different types of buildings. Fault prediction and detection is an important challenge for the optimal performance of microclimate control systems.

Microclimate Control as a Global Problem

Over the past decade, the rapid growth in energy consumption and the associated Carbon Dioxide (CO_2) emissions, with a reduction in the amount of the planet's fuel and energy resources, have led research scientists to pay special attention to finding energy efficiency solutions.

According to reports by the United States Department of Energy (DOE), among all sectors of the economy that consume significant amounts of energy, buildings and residential complexes use more than one-third (up to 40%) of the total world energy consumption. In addition, more than 60% of the electricity consumed goes to the sector of residential and commercial buildings.

On a global scale, commercial buildings use approximately 41% percent of primary energy consumed worldwide, including the United States, Europe, and Asia. Those numbers are still expected to rise over the next 20 years.

About 90% of the total energy consumption by the housing and communal services sector of the Republic of Kazakhstan is spent on building exploitation. Residential buildings are characterized by the highest energy consumption: 50-55%.

Industrial buildings use somewhat less, 35-45% and civilian buildings account for about 10%.

In housing and civil engineering, energy efficiency reserves reach up to 40%. In this regard, measures to reduce heat and energy waste are of great importance for the Republic.

In general, the state of a microclimate system is described mainly by the environment air parameters: air temperature, relative humidity, amount of carbon dioxide, as well as air mobility, NH_4 (ammonia) content, H_2S (hydrogen sulfide) and bacterial contamination.

To cope with increasing demand, different control strategies are being incorporated into the existing infrastructures, to sustain the demand for electricity in residential and commercial buildings.

Currently, the most widely used automated microclimate control systems will provide just temperature regulation. That increases people's comfort and modern intelligent control allows energy consumption optimization. Nonetheless, there are still other variables for optimal comfort and well-being. Hence, a good microclimate control system goes beyond temperature, namely controlling also humidity and speed of moving air.

Importance of Microclimate Control for People.

Indoor microclimate conditions depend on a number of factors, such as climate zone; season of the year; type of equipment used; type of facilities; air exchange conditions; size of the room to be climatized; and number of people inside the room.

In [9] studied the microclimate of dwellings in a harsh climate and proved that both regional and seasonal differentiation of the thermal state of a person is necessary to create thermal comfort in a residential room. Naturally, in winter time, a higher amount of energy is necessary to maintain air temperature in the room. Such a temperature indicator relieves the physiological fatigue of people coming from low ambient temperatures. Figure 1. shows a room and the variables that are important to monitor. The variables considered are:

Qceil – Heat loss through the ceiling
Qd – Heat loss through door;
QEo – Heat loss through exhaust openings;
Qh – Human heat dissipation;
Qwin – Heat loss through windows;
Qw – Heat loss through walls;
tR – Radiation temperature;
+t – Room temperature;
−t – Outside temperature;
+fi – Room humidity;
−fi – Outside humidity;
−v – Outside air velocity;
+v – Room air velocity;
Qfl – Heat loss through the floor;
1 – Heating system;
G – Building structure thickness;

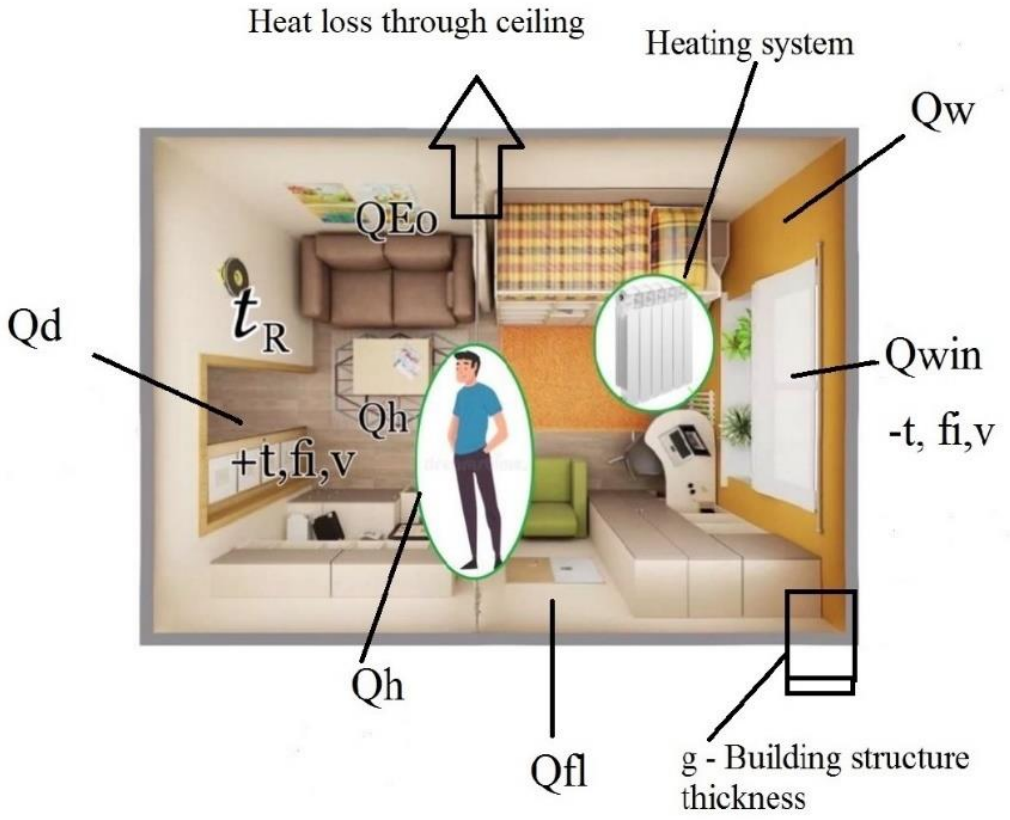

Figure 1 - Room microclimate variables. Diagram based on [10].

For this reason, in the winter season in the first construction and climatic zone, hygienists recommend maintaining indoor air temperatures in the range 23-24^{0}C.

Human thermal comfort is also highly correlated to the local temperature in which individual parts of the body are located, and in particular, the head and legs.

The floor temperature affects the thermal state of a person most strongly. Direct contact with the cold floor leads to colds.

In this regard, in residential premises, the temperature on the floor surface can be lower than the average air temperature, but should not be lower by more than two degrees.

With an increase in the difference of temperature between the air and the interior surfaces, the radiant cooling of a person increases. That can cause a violation of the thermoregulation of the human body.

According to physiological observations thermal comfort in a residential room is considered to be achieved only when the air temperature is not higher than the interior surfaces by more than 2-3^{0}C.

The normalized difference for residential premises determines the dew loss on the wall surface to a greater extent than the thermal comfort of a person.

The comfortable values of indoor air mobility depend on the combination of air temperature, humidity, radiation situation in the room, type of work and season of the year.

The totality of the considered characteristics of the microclimate and their permissible ranges, established by hygienists, describe the conditions that need to be created in the room so that a person experiencing a thermal state of neutrality (that is, could not determine whether he is warm or cold), which is usually assessed as comfortable.

The microclimate in the premises is formed due to the disturbing effects of the external environment and the technological process inside the building, which is countering outer effects by heating or cooling, as well as controlling other microclimate variables.

The peculiarity of microclimate systems is that they consume a large amount of energy resources, including thermal and electrical energy and sometimes tap water.

1.3 Challenges of Microclimate Management

Microclimate management in buildings faces several challenges that can impact the ability to maintain optimal indoor conditions and achieve energy efficiency. Some of the key challenges are described below.

Complexity of Indoor Environments

Indoor spaces are dynamic and diverse, with variations in occupancy, activities, and environmental conditions. Managing microclimates requires understanding and addressing these complexities to ensure consistent comfort and air quality.

Variability in Occupant Preferences

Occupants have different comfort preferences and thermal sensitivities, making it challenging to establish universal temperature and humidity settings that satisfy everyone. Balancing individual preferences with energy efficiency goals can be difficult.

Energy Consumption

Heating, cooling, and ventilating buildings consume a significant amount of energy, contributing to both operational costs and environmental impacts. Optimizing energy use without compromising comfort requires sophisticated control strategies and efficient building systems.

Air Quality Concerns

Indoor air quality can be compromised by pollutants such as VOCs, particulate matter, allergens and microbial contaminants. Effective management of ventilation, filtration, and pollutant sources is essential for maintaining healthy indoor environments.

Climate Change and Extreme Weather Events

Climate change can lead to more frequent and intense heatwaves, cold spells, and extreme weather events, challenging traditional HVAC systems' ability to cope with fluctuating outdoor conditions and maintain comfort indoors.

Aging Infrastructure and Equipment

Many buildings have outdated HVAC systems and infrastructure, leading to inefficiencies, reliability issues, and increased maintenance requirements. Retrofitting and upgrading these systems can be costly and disruptive.

Integration of Smart Technologies

Incorporating smart technologies and IoT (Internet of Things) devices into building management systems offers opportunities for improved control and automation. However, integrating these technologies seamlessly and ensuring data security and privacy present challenges for building owners and operators.

Data Management and Analysis

Collecting, analyzing, and interpreting data from various sensors and building systems is crucial for optimizing microclimate management. However, managing large volumes of data, ensuring data accuracy, and extracting actionable insights pose challenges for building managers.

Addressing these challenges requires a multidisciplinary approach that combines expertise in building science, HVAC engineering, data analytics, and occupant behavior. By leveraging advanced technologies, adopting sustainable design practices, and prioritizing occupant comfort and well-being, building owners and operators can overcome these challenges and create healthier, more energy-efficient indoor environments [6, 9].

1.4 Summary

The chapter delves into the complexities surrounding the management of microclimates within built environments. It begins by emphasizing the significance of microclimate control, highlighting its crucial role in ensuring optimal environmental conditions for various applications such as agriculture, building automation, and healthcare. The chapter then proceeds to discuss the challenges inherent in microclimate management, including the intricacies of BMS and the importance of fault detection systems in building automation. Furthermore, it explores modern technologies that offer intelligent solutions for microclimate management, such as IoT sensors and machine learning algorithms.

Despite the advancements in technology, the chapter also identifies significant gaps in existing research, indicating areas where further investigation and innovation are needed. These gaps encompass various aspects, including the integration of renewable energy sources, adaptive control strategies, and user-centered design principles. By addressing these gaps, researchers, engineers, and policymakers can develop more robust, efficient, and sustainable microclimate control systems that meet the evolving needs of society and the environment. Overall, the chapter provides a comprehensive overview of the challenges and opportunities in building microclimate control, laying the groundwork for future research and development in this critical field.

2 CRISP-DM METHODOLOGY

In the realm of data analysis and knowledge discovery, various methodologies have been developed to guide practitioners through the intricacies of extracting insights from data. This section provides a review of three prominent methodologies: Knowledge Discovery in Databases (KDD), Cross-Industry Standard Process for Data Mining (CRISP-DM), and OSEMN. Each methodology offers a structured approach to data analysis, aiming to facilitate the systematic exploration, modeling, and interpretation of data to derive actionable insights. By examining the key principles and phases of these methodologies, this review aims to provide readers with a comprehensive understanding of the foundational frameworks underpinning data analysis processes.

Among the methodologies reviewed, the CRISP-DM emerges as the preferred choice for data analysis projects. CRISP-DM offers a structured and comprehensive framework comprising six distinct phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. This methodical approach ensures that data analysis initiatives are conducted systematically, aligning with business objectives and yielding actionable insights. CRISP-DM's emphasis on iterative refinement and collaboration fosters transparency and ensures that the resulting models are robust and effective. Its versatility and widespread adoption make CRISP-DM the gold standard in the field of data mining and analysis, earning it the reputation as the go-to methodology for practitioners across diverse industries.

2.1 Review of Existing Methodologies

The CRISP-DM methodology serves as a comprehensive guide for developing intelligent fault diagnosis systems tailored to building microclimate control. This structured approach encompasses six key phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

A diagram of the CRISP-DM process that shows the six key phases and indicates the important relationships between them is shown in Figure 2.

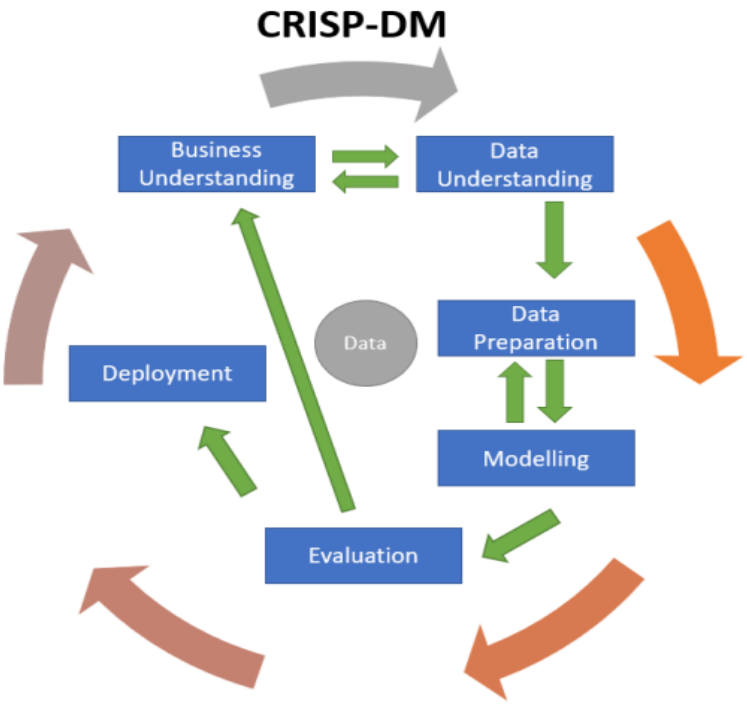

Figure 2 - Phases of the CRISP-DM reference model

Note - Compiled based on source data Cornellius Yudha Wijaya. CRISP-DM Methodology For Your First Data Science Project. TDS Archive, Medium. Retrieved, from https://medium.com/data-science (data access 05.10.2024)

The process begins with Business Understanding, where stakeholders' objectives, constraints, and challenges related to building microclimate control are identified. This phase lays the foundation for aligning the project with organizational goals and ensuring its relevance to the target environment.

Next, in the Data Understanding phase, relevant data sources such as sensor data, historical performance records, and maintenance logs are collected and analyzed. This stage aims to gain insights into the available data and assess its quality, completeness, and suitability for modeling.

In Data Preparation, the collected data undergoes preprocessing to enhance its quality and prepare it for analysis. Techniques like data cleaning, feature engineering, and normalization are applied to optimize the dataset for modeling.

The Modeling phase involves selecting appropriate machine learning algorithms and building predictive models to diagnose faults in building microclimate control systems. This step requires iterative experimentation with various algorithms and parameter settings to develop accurate and robust models.

Once models are developed, they undergo Evaluation to assess their performance using metrics like accuracy, precision, and recall. This phase ensures that the models meet the required standards for fault diagnosis in real-world scenarios.

Finally, in the Deployment phase, the developed fault diagnosis system is integrated into operational environments. Continuous monitoring and maintenance

are essential to ensure the systems ongoing effectiveness and reliability in optimizing energy efficiency, enhancing occupant comfort, and ensuring building safety [15-19].

By following the CRISP-DM methodology, organizations can systematically develop and deploy intelligent fault diagnosis systems tailored to building microclimate control, thereby addressing key challenges and optimizing building performance.

KDD is a methodology akin to CRISP-DM, designed by Gregory Piatetsky-Shapiro, founder of the popular blog site KDnuggets. It offers a systematic approach to extracting insights and knowledge from data, with clear outcomes delineated at each stage (Figure 3).

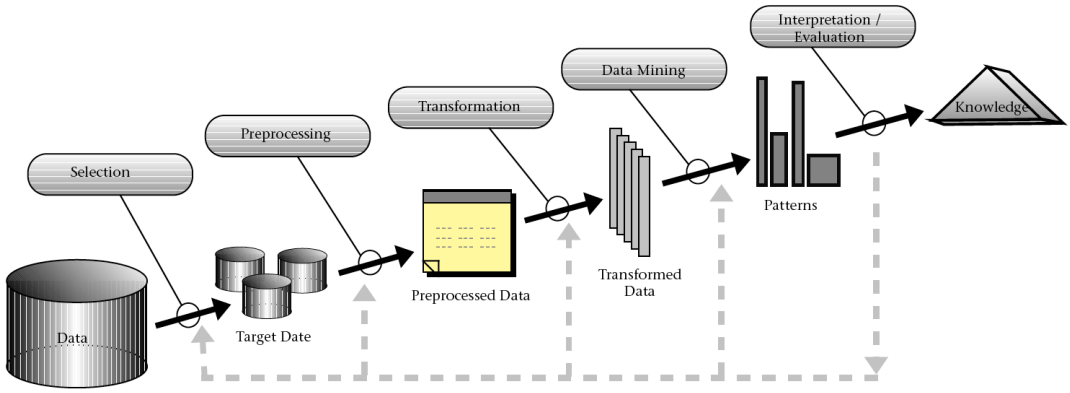

Figure 3 - Knowledge Discovery in Databases (KDD)

Note - Compiled based on source data GeeksforGeeks. URL: https://www.geeksforgeeks.org/dbms/kdd-process-in-data-mining/ (data access: 05.10.2024)

In the Selection stage, akin to CRISP-DM's Business Understanding, the focus is on gaining a comprehensive understanding of the project's objectives and data intricacies. This involves reviewing past projects and literature to glean insights from similar endeavors, thereby informing the project's direction.

Moving to Preprocessing, the emphasis is on data cleaning to rectify issues like missing data, outliers, and irrelevant entries. This stage lays the groundwork for subsequent transformations and ensures that the data is primed for further analysis.

The Transformation stage refines the data, focusing on feature engineering, addressing multicollinearity, and ensuring data normalization. This step prepares the data for modeling by shaping it into the necessary format.

Data Mining is where the processed data is utilized for modeling, employing various techniques to derive actionable insights. This phase leverages algorithms to extract valuable information from the databases, aligning with the project's objectives.

In Interpretation and Evaluation, insights gained from the modeling stage are used to make predictions and answer project objectives. Effective communication of methods and results to stakeholders, including non-technical audiences, is vital in this phase.

Similar to CRISP-DM, KDD is an iterative process, allowing for adjustments based on emerging issues, new insights, or evolving project objectives. This iterative nature ensures that the final result meets the project's requirements and delivers successful outcomes.

The OSEMN (Obtain, Scrub, Explore, Model, Interpret) process stands as a cornerstone in the field of Data Science, offering a structured approach to tackle complex analytical challenges. It serves as a robust framework to address the intricate nature of bee monitoring data, providing clarity and organization throughout the analytical journey (Figure 4).

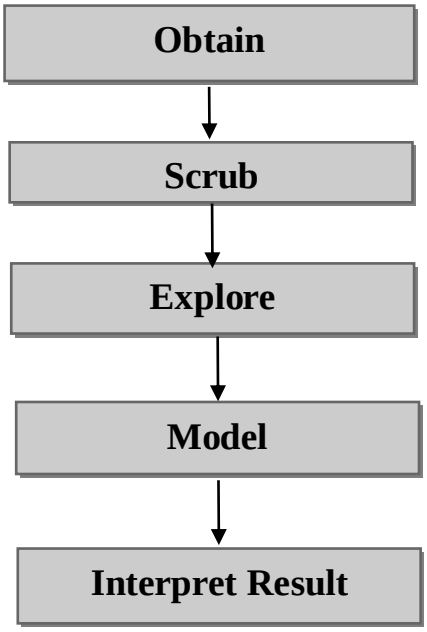

Figure 4 - OSEMN (Obtain, Scrub, Explore, Model, Interpret)

Note - Compiled by the author based on the source data [23]

Obtain Data:

Beehive data is sourced from sensors strategically placed both inside and outside hives, capturing crucial metrics like temperature, humidity, weight, noise levels, and environmental conditions such as temperature, humidity, and CO2 levels.

Sensor data is aggregated into nodes, each comprising a group of sensors connected to a common microcontroller, allowing for the collection of specific data types.

Scrub Data:

The collected data undergoes preprocessing steps to ensure its integrity and usability.

Actions include merging disparate data columns into a unified table, cleansing data of invalid values, normalizing data to account for variations in types and ranges, and processing extreme values using techniques like the RANSAC algorithm to detect outliers.

Explore Data (EDA):

EDA involves uncovering patterns, relationships, and anomalies within the dataset.

By dividing data into «smooth» and «rough» components, EDA techniques aim to extract meaningful insights while identifying potential outliers or irregularities that warrant further investigation.

Model Data:

Modelling entails developing mathematical expressions of model parameters to predict trends or classify data.

Machine learning algorithms, ranging from logistic regression to random forest, are employed to build predictive models using preprocessed and structured data.

Interpret Data:

The final stage involves interpreting results to derive actionable insights and draw meaningful conclusions.

Through rigorous evaluation and analysis, researchers assess the effectiveness of the study methodology, ensuring reproducibility and validity of findings.

The utilization of the OSEMN workflow in bee monitoring not only streamlines the analytical process but also facilitates comprehensive data exploration and interpretation. By adhering to this structured methodology, researchers can unlock valuable insights into bee behavior, health, and environmental interactions, ultimately contributing to the advancement of beekeeping practices and environmental conservation efforts.

The OSEMN model, also known as «MSEMiN» offers a straightforward yet effective approach to the data science process, emphasizing ease of navigation between its sections.

Mine: Extracting data from different sources.

Scrub: Cleaning and preparing data for analysis.

Explore: Exploring data to discover patterns and trends.

Model: Creating mathematical models for prediction or classification of data.

iNterpret: Interpreting data analysis results to derive meaningful insights.

This simplicity facilitates adaptability, allowing practitioners to revisit earlier stages if encountered with barriers during modeling work.

In the Obtain stage, akin to KDD’s Selection and CRISP-DM’s Business/Data Understanding, the focus is on aligning with stakeholders to acquire pertinent data essential for answering project queries effectively [23].

Scrubbing, the subsequent stage, entails cleaning the data by addressing outliers, normalization, and feature engineering. This ensures that the data is primed for exploration and analysis in the subsequent stages.

The Explore stage mirrors CRISP-DM’s Data Understanding, employing visualizations like histograms and heatmaps to gain insights into the data's distribution and check for normality and collinearity. This step aids in meeting assumptions and balancing data to reduce modeling errors.

Transitioning to the Model stage, practitioners utilize the meticulously prepared data to employ machine learning algorithms, defining success markers and focusing on those yielding favorable results. Flexibility is paramount here, as insights from

this stage may necessitate revisiting earlier stages for data restructuring or refinement.

The Interpret stage, akin to CRISP-DM's Evaluation, involves communicating results to stakeholders in an accessible manner, ensuring comprehension across technical and non-technical audiences. This facilitates discussions, identifies areas for improvement, and advocates for additional resources if needed.

Overall, the OSEMN model provides a structured framework for the data science process, offering flexibility for iterative refinement and adaptation as required. Understanding and exploring various methodologies enable practitioners to select the most suitable approach for their projects, ensuring successful outcomes through systematic checkpoints and continuous improvement.

Table 1 - Comparison of Data Mining Methodologies: CRISP-DM, KDD, and OSEMN

Aspect	CRISP-DM	KDD	OSEMN
Stages	6 stages	6 stages	5 stages
Iterative	Yes	Yes	Yes
Business Understanding	Gather project objectives and requirements	Gain understanding of project objectives	Obtain project goals and stakeholder alignment
Data Understanding	Assess data sources and quality	Assess data sources and relevance	Obtain and understand data
Data Preparation	Clean, transform, and prepare data	Preprocess and clean data	Scrub data and prepare for analysis
Modeling	Build and evaluate models	Use algorithms to mine data	Apply machine learning models
Evaluation	Assess model performance and insights	Interpret and evaluate results	Evaluate model performance and results
Deployment	Deploy model into production	Present findings to stakeholders	Share results and discuss implications

Table 1 - Comparison of Data Mining Methodologies: CRISP-DM, KDD, and OSEMNTable 1. provides a concise overview of the similarities and differences between these methodologies in terms of their stages and objectives within the data science process.

2.2 CRISP-DM Methodology Selection and Application Process

The selection of an appropriate methodology is paramount for the successful execution of any research endeavor, particularly in data-intensive studies. This section elucidates the rationale behind the adoption of the CRISP-DM methodology for the present dissertation work, as well as its application process within the context of the study.

CRISP-DM was chosen as the guiding framework for this research project due to several key considerations. Firstly, CRISP-DM enjoys widespread recognition and acceptance as an industry-standard methodology in the fields of data mining and

machine learning. Its established reputation underscores its suitability for guiding rigorous and systematic data analysis processes.

Flexibility is another compelling factor that influenced the selection of CRISP-DM. The methodology's inherent flexibility allows for adaptation to the unique characteristics of the dataset and research objectives, ensuring that the analytical approach remains agile and responsive to evolving research needs throughout the study.

A defining feature of CRISP-DM is its iterative nature, which comprises distinct phases including business understanding, data understanding, data preparation, modeling, evaluation, and deployment. This iterative approach facilitates continuous refinement and improvement throughout the research process, ensuring that insights gleaned from earlier phases inform subsequent analyses and decision-making.

CRISP-DM places a strong emphasis on establishing a comprehensive understanding of the business context and objectives before delving into technical analyses. This emphasis on business understanding ensures alignment between data mining efforts and the overarching goals of the research project, ultimately enhancing the relevance and impact of the findings.

The delineation of clear phases and deliverables within CRISP-DM further enhances its utility as a guiding framework for this dissertation work. By providing structured guidelines and milestones for each phase of the data mining process, CRISP-DM facilitates effective project management and communication within the research team, thereby ensuring that the study progresses in a systematic and organized manner.

In summary, the selection of the CRISP-DM methodology for this dissertation work is driven by its established reputation, flexibility, iterative nature, emphasis on business understanding, and provision of clear phases and deliverables. By adhering to the structured approach delineated by CRISP-DM, this research project aims to conduct rigorous data analyses that contribute to advancing knowledge in the field of data mining and machine learning.

2.3 Summary

This chapter provides an overview of the CRISP-DM methodology, highlighting its relevance and application in data analysis. The chapter outlines the six key phases of the methodology- Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment - demonstrating how each phase contributes to a structured and iterative approach for solving complex data analysis problems. Emphasizing the flexibility and adaptability of CRISP-DM, this chapter explains why it is widely adopted in various industries and how its systematic framework ensures alignment with business objectives. By following the CRISP-DM methodology, this research aims to achieve comprehensive and actionable insights, thereby advancing knowledge in the field of data mining and machine learning.

3 MICROCLIMATE SYSTEM OPTIMIZATION WITH MACHINE LEARNING AND FAULT DETECTION

In recent years, the intersection of machine learning techniques and environmental science has led to significant advancements in optimizing microclimate systems. The delicate balance of environmental conditions within confined spaces, such as buildings or greenhouses, plays a crucial role in various aspects, including human comfort, energy efficiency, and agricultural productivity. To achieve optimal performance and maintain desired conditions, continuous monitoring, fault detection, and adaptive control mechanisms are essential.

This chapter explores the integration of machine learning algorithms with fault detection techniques to enhance the optimization of microclimate systems. By leveraging supervised and unsupervised learning approaches, along with time series analysis, we can extract meaningful insights from sensor data streams. These insights not only aid in identifying anomalies and faults but also provide valuable information for proactive maintenance and system improvement.

The utilization of supervised learning algorithms enables the detection of known patterns and deviations from expected behavior. By training models on labeled datasets, we can develop robust fault detection systems capable of accurately identifying abnormal conditions and potential issues within microclimate systems. Furthermore, unsupervised learning techniques offer the ability to uncover hidden patterns and anomalies in the data without prior labeling, providing a more flexible approach for fault detection in dynamic environments.

Time series analysis plays a vital role in understanding the temporal dynamics of microclimate data. By analyzing historical trends and patterns, we can forecast future conditions, detect recurring patterns, and identify deviations from expected temporal behavior. This temporal awareness enhances the effectiveness of fault detection algorithms by considering the context of past observations and predicting future trends.

Moreover, the integration of machine learning-based fault detection with BMS offers seamless control and automation capabilities. By incorporating adaptive learning mechanisms, microclimate systems can dynamically adjust their operation based on real-time feedback and changing environmental conditions. This adaptive control not only optimizes energy consumption and resource utilization but also enhances overall system resilience and reliability [24-26].

3.1 Fault Detection and Diagnosis

FDD play a key role in high-cost and safety-critical processes, such as in microclimate systems. Early detection, or even prediction, of process faults is very important to plan and execute maintenance interventions to help avoid abnormal event progression [27]. Early interventions can help maximize equipment availability and minimize maintenance costs [28].

Fault detection can be accomplished through various means.

With the growing demand for smart building infrastructure and organization maintenance, automatic fault detection has attracted attention from both academia and industry.

Figure 5 shows the hierarchy of FDD methods according to [29]. As shown in the Figure 5, there are methods based on quantitative models, methods based on the history of processes, and methods based on rules. Quantitative model-based methods are classified into detailed and simplified physical models. Process history-based approaches are divided into knowledge-based and data-driven methods, which include expert systems, pattern classification, causal analysis, statistical approaches, feature extraction, and machine learning. Rule-based methods represent another category that applies predefined thresholds and if–then rules. Such fault detection techniques have been effectively used to improve reliability in power system components (Abd-Alkader et al., 2021) [30].

There are various fault detection methods. For example, Xiangjun et al., [31] used several types of fault detection methods based on information fusion, artificial intelligence (AI), neural networks, fuzzy algorithms, and genetic algorithms. The authors note that different fault information can serve as a good source for processing different fault detection methods. Thus, it is possible to reduce the influence of the interfering signal, eliminate the limitations of a single protection circuit, and improve the accuracy and reliability of fault detection by integrating and combining all kinds of fault information.

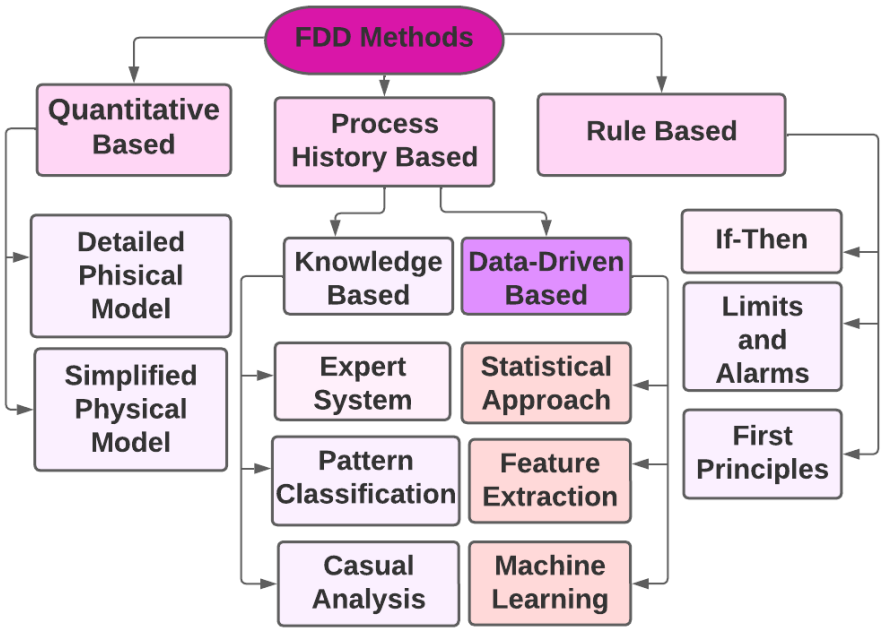

Figure 5 - Fault detection and diagnosis methods

Note – Compiled by the author based on the source [29]

3.2 Fault Diagnosis Algorithms

Diagnosis algorithms are effective methods for identifying and classifying faults in microclimate control systems.

In the context of error diagnosis, there are a number of techniques used to detect and identify problems in various systems. Some commonly used error diagnosis methods are listed below.

1) Principal Component Analysis (PCA);

The PCA algorithm is used to highlight the most significant components in the data and reduce dimensionality. Used to identify underlying patterns and anomalies.

2) K-Nearest Neighbors (KNN);

The KNN algorithm classifies objects based on the majority of classes among their k-nearest neighbors. Used to identify anomalies and classify by similarity.

3) Decision tree method;

4) Random Forest;

Random Forest is an ensemble method that builds many decision trees and combines them for higher accuracy and robustness. Used to diagnose errors in systems with large amounts of data.

5) Support Vector Machine (SVM);

SVM is used to separate classes by finding the optimal hyperplane. Used to classify and diagnose errors in data with nonlinear dependencies.

6) Logistic Regression;

Logistic regression is used to model the likelihood of an object being assigned to a particular class. Widely used in error diagnosis and classification.

The decision tree algorithm constructs a tree structure representing a sequence of questions for classification. Used to identify the causes of faults.

These algorithms can be used either independently or in combination, depending on the specific problem of diagnosis and data characteristics (Table 2).

Table 2 - Machine Learning Algorithms for Microclimate Analysis

Algorithm	Description	Typical Applications
Linear Regression	A simple regression algorithm used to model the relationship between one or more independent variables and a dependent variable. It assumes a linear relationship between the predictors and the target variable.	Predicting microclimate variables such as temperature, humidity, or air pressure based on historical data.
Decision Trees	A tree-based algorithm that recursively splits the data into subsets based on the value of input features. It's intuitive, easy to interpret, and can handle both numerical and categorical data. Decision trees can be used for both regression and classification tasks.	Classifying microclimate patterns based on sensor data, identifying factors influencing microclimate variations.
Random Forest	An ensemble learning method that builds multiple decision trees and combines their predictions to improve accuracy and reduce overfitting. It's robust, scalable, and suitable for high-dimensional datasets.	Predicting microclimate variables, feature selection, identifying important predictors.

Continuation of the table 2

Algorithm	Description	Typical Applications
Support Vector Machines (SVM)	A supervised learning algorithm that finds the optimal hyperplane that best separates classes in high-dimensional space. It's effective in handling complex datasets with clear margin of separation between classes. SVM can be used for both classification and regression tasks.	Classifying microclimate patterns, predicting microclimate variables.
k-Nearest Neighbors (k-NN)	A simple algorithm that classifies new data points based on the majority class of their nearest neighbors in feature space. It's non-parametric and requires no training phase. k-NN can be used for both classification and regression tasks.	Identifying similar microclimate patterns, anomaly detection, clustering.
Neural Networks	A set of algorithms inspired by the structure and function of the human brain. Neural networks consist of interconnected nodes (neurons) organized in layers. They're capable of learning complex patterns and relationships from data and can be applied to various tasks, including regression, classification, and clustering.	Predicting microclimate variables, pattern recognition, anomaly detection, feature extraction.
Gaussian Mixture Models (GMM)	A probabilistic model that represents the distribution of data as a mixture of Gaussian distributions. GMMs can capture complex data distributions and are often used for clustering and density estimation.	Clustering similar microclimate conditions, identifying underlying distributions in data.

Fault Diagnosis performance metrics

Major performance Metrics for Diagnostic systems include, among others,

– False positives;

– False negatives;

– Time delay – time span between the initiation and the detection of a fault (failure) event;

– Percent isolation to one line-replaceable unit (LRU).

Table 3 - Decision Matrix for Fault Detection Evaluation

Outcome	Fault (F_1)	No Fault(F_0)	Total
Positive (D_1) (detected)	a Number of detected faults	B Number of false alarms	$a+b$ Total number of alarms
Negative (D_0) (not detected)	c Number of missed faults $a+c$ Total number of faults	d Number of correct rejections $b+d$ Total number of fault-free cases	$c+d$ Total number of non-alarms $a+b+c+d$ Total number of cases

Performance requirements for diagnostic algorithms typically specify the maximum allowable number of false positives and false negatives as a percentage of the total faults present in a particular subsystem over its expected life. For example, a typical requirements may be stated as «no more than 5 percent of false positives and no more than 3 percent of false negatives». A case should be made for the relative significance of these two metrics: False negatives may present major risks to the

health of the equipment under test. Missed fault conditions may lead to a catastrophic failure resulting in loss of life or loss of the system. False positive, on the other hand, could be accommodated with an unavoidable loss of confidence on the part of the system operator as to the effectiveness of the diagnostic tools. For this reason, tradeoffs that naturally arise between the two metrics favor a stricter false-negative requirements. The time delay matrix is most significant not only as an early warning to the operator of an impending failure but also in terms of providing a sufficient time window that allows a prognostic algorithm to perform its intended task. Fault detection events may be evaluated through the decision matrix. It is based on a hypothesis-testing methodology and represents the possible fault detection combinations that my occur. From this matrix, the detection metrics can be computed readily. The probability of detection (POD) given a fault assesses the detected faults over all potential fault cases.

3.3 Machine Learning Algorithms

Machine learning is a unified algorithmic framework designed to identify computational models that accurately describe empirical data and the phenomena underlying it, with little or no human involvement. While still a young discipline with much more awaiting discovery than is currently known, today machine learning can be used to teach computers to perform a wide array of useful tasks including automatic detection of objects in images (a crucial component of driver-assisted and self-driving cars), speech recognition (which powers voice command technology), knowledge discovery in the medical sciences (used to improve our understanding of complex diseases), and predictive analytics(leveraged for sales and economic forecasting), to just name a few.

Machine learning algorithms work by searching through a set of possible prediction models for the model that best captures the relationship between the descriptive features and target feature in a dataset. An obvious criteria for driving this search is to look for models that are consistent with the data. There are, however, at least two reasons why simply searching for consistent models is not sufficient for learning useful prediction models. First, when we are dealing with large datasets, it is likely that there is noise in the data, and prediction models that are consistent with noisy data make incorrect predictions. Second, in the vast majority of machine learning projects, the training set represents only a small sample of the possible set of instances in the domain. As a result, machine learning is an ill-posed problem, that is, a problem for which a unique solution cannot be determined using only the information that is available. Comparison of Supervised and Unsupervised Learning Algorithms shows in Table 4.

Supervised Learning

Machine learning is defined as an automated process that extracts patterns from data. To build the models used in predictive data analytics applications, we use supervised machine learning. Supervised machine learning techniques automatically learn a model of the relationship between a set of descriptive features and a target feature based on a set of historical examples, or instances. We can then use this

model to make predictions for new instances. These two separate steps are shown in 6.

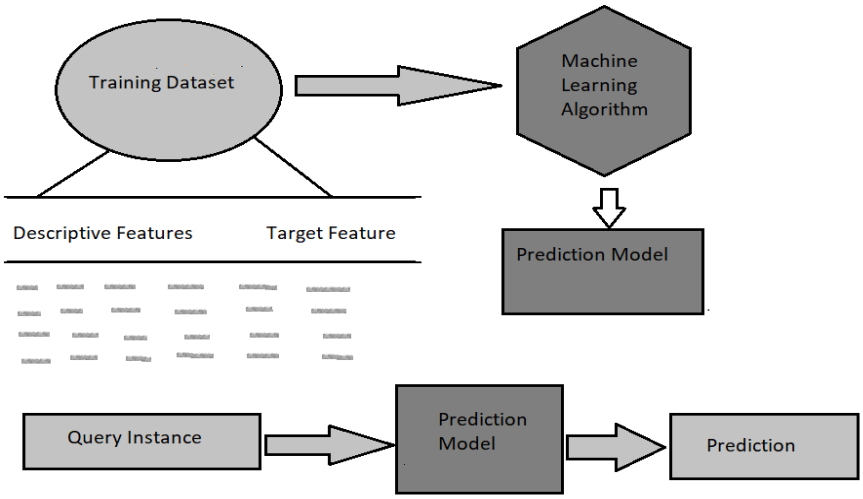

Figure 6 - The two steps in supervised machine learning: learning a model from a set of historical instances and using a model to make predicting

Note - Compiled by the author based on the source data

Supervised learning is a valuable technique for developing smart fault detection systems in buildings' microclimate control. In this approach, data collected from various sensors within the building's HVAC system, such as temperature, humidity, and pressure sensors, is labeled based on known fault conditions or anomalies. This labeled data is then used to train supervised learning algorithms, such as decision trees, random forests, SVM, or neural networks. The trained model learns the patterns associated with normal operation as well as various fault conditions.

Once trained, the model is deployed in the smart fault detection system to monitor real-time sensor data continuously. When deviations from normal operating conditions or the presence of faults are detected, the system generates alerts or notifications to building operators or maintenance personnel. These alerts can be sent via email, SMS, or integrated into building management systems (BMS) for prompt action.

The performance of the supervised learning model is evaluated using metrics such as accuracy, precision, recall, F1-score, or area under the receiver operating characteristic curve (AUC-ROC). Cross-validation techniques are employed to ensure the robustness and generalization of the model. Additionally, the model is continuously updated and refined based on feedback from real-world operations and additional labeled data. This iterative process helps improve fault detection accuracy and reduce false positives, ultimately enhancing building efficiency, reducing energy consumption.

In the context of predicting faults in building microclimate systems, three widely used machine learning methods were analyzed: the Multilayer Perceptron (MLP), Support Vector Machine (SVM), and Gradient Boosting. Each of these algorithms has distinct characteristics, architectural features, and practical applications, which

makes them suitable for different scenarios in detecting rare faults in engineering systems.

Unsupervised Learning

Unsupervised learning offers a powerful approach to developing smart fault detection systems for microclimate control in buildings. In this context, unsupervised learning algorithms can analyze sensor data without the need for predefined labels or categories. Instead, these algorithms autonomously identify patterns, anomalies, and deviations in the data, which may signal faults or malfunctions in the building's HVAC system.

For example, clustering algorithms like K-means or hierarchical clustering can group similar sensor data points together, revealing typical patterns of operation within the building's microclimate system. Anomaly detection techniques, such as isolation forests or Gaussian mixture models, can flag data points that deviate significantly from the norm, indicating potential faults or abnormalities in the HVAC system.

Dimensionality reduction methods like PCA or t-distributed Stochastic Neighbor Embedding (t-SNE) can help visualize complex sensor data and uncover underlying structures or correlations. Additionally, unsupervised learning can uncover hidden patterns or trends in sensor data that may not be immediately apparent, providing valuable insights for fault diagnosis and system optimization.

By integrating unsupervised learning techniques into smart fault detection systems, building operators can continuously monitor sensor data in real-time, adapting to changing conditions and detecting new types of faults or anomalies as they arise. This adaptive approach enhances the system's ability to identify and diagnose faults accurately, ultimately improving building efficiency, reducing energy consumption, and ensuring occupant comfort. Summarizing the application of unsupervised learning in smart fault detection systems for microclimate control in buildings shows in Table 5.

Table 4 - Comparison of Supervised and Unsupervised Learning Algorithms

Aspect	Supervised Learning	Unsupervised Learning
Objective	Learns patterns and relationships with labeled data	Identifies hidden structures and patterns in unlabeled data
Data Requirement	Requires labeled data	Works with unlabeled data
Use Cases	Prediction, classification, regression	Clustering, dimensionality reduction, anomaly detection
Algorithm Selection	Linear regression, decision trees, SVM, neural networks	K-means, hierarchical clustering, PCA, anomaly detection
Interpretability	Model predictions are interpretable	Results may be less interpretable, rely on visualizations
Data Exploration vs. Prediction	Focuses on prediction or classification	Emphasizes data exploration and discovery
Model Evaluation	Accuracy, precision, recall, F1-score, MSE	Silhouette score, reconstruction fault, anomaly score

A table summarizing the application of unsupervised learning in smart fault detection systems for microclimate control in buildings:

Table 5 – Applications of Unsupervised Learning in Microclimate Analysis

Application	Description
Clustering Analysis	Algorithms like K-means or hierarchical clustering group similar sensor data points, revealing typical patterns of operation within the building's microclimate system.
Anomaly Detection	Techniques such as isolation forests or Gaussian mixture models flag data points deviating significantly from the norm, indicating potential faults or abnormalities in the HVAC system.
Dimensionality Reduction	Methods like Principal Component Analysis (PCA) or t-distributed Stochastic Neighbor Embedding (t-SNE) help visualize complex sensor data and uncover underlying structures or correlations.
Pattern Discovery	Unsupervised learning uncovers hidden patterns or trends in sensor data, providing valuable insights for fault diagnosis and system optimization.
Continuous Monitoring	Building operators can continuously monitor sensor data in real-time, adapting to changing conditions and detecting new types of faults or anomalies as they arise, enhancing fault detection accuracy and system efficiency.

This Table 5 provides a concise overview of how unsupervised learning techniques are applied in smart fault detection systems for microclimate control in buildings, highlighting their various applications and benefits.

In this study, two clustering methods were used: K-means and DBSCAN, each with its own characteristics and advantages. K-means is a classical clustering method based on minimizing the sum of squared distances between data points and cluster centroids. This method divides the data into a predefined number of clusters, and its results depend on the initial centroids, which can affect the stability of the solution. It works well with well-separated and compact clusters but may struggle with clustering data containing noise or clusters of arbitrary shapes.

On the other hand, DBSCAN is a density-based clustering algorithm that does not require a predefined number of clusters and can detect clusters of arbitrary shape. It relies on two key parameters: ε (epsilon) – the radius of the neighborhood of a point, and min_samples – the minimum number of points required within the neighborhood to form a cluster. DBSCAN is effective for working with data containing noise and outliers, as it automatically separates noisy points from clusters. This makes it especially useful for tasks where the data might include anomalies or have a complex structure, such as microclimate data.

Unlike K-means, DBSCAN does not require specifying the number of clusters in advance and automatically labels outliers as noise. This allows it to adapt more flexibly to different data structures, identifying both primary and rare patterns. In the analysis of microclimate system data, DBSCAN demonstrated its effectiveness in classifying system states and identifying anomalous situations, such as sensor failures or system malfunctions. This is particularly useful for diagnostics and early detection of potential issues in the systems.

Therefore, for analyzing microclimate data, especially for fault detection, DBSCAN proved to be the more suitable method. It adapts to various data types and can handle data containing noise or outliers effectively, something that is not always achievable with K-means. The results are given in Chapter 7.

In recent years, there has been rapid progress in the development of intelligent building climate control methods aimed at improving energy efficiency and ensuring comfortable indoor conditions for occupants. A systematic review of artificial intelligence technologies applied to enhance the efficiency and reliability of HVAC systems is presented in [32–34]. The authors note that modern approaches utilize machine learning algorithms to analyze large volumes of operational data and enable early fault detection. In particular, Random Forest-based techniques discussed in [33] demonstrate high accuracy in diagnosing faults in hotel buildings, while hybrid methods using wireless sensor networks [34–36] ensure robustness and autonomy of monitoring systems.

Considerable attention has been devoted to the study of classification and clustering methods for fault diagnosis. Publications [37–45] propose models based on multilayer neural networks, PCA, sensitivity analysis, and linear discriminant analysis (LDA) for diagnosing faults in refrigeration and air handling units. These studies demonstrate that the use of statistical and data-driven models enables effective anomaly detection in sensor data, reducing energy consumption and improving the reliability of microclimate control systems. In particular, [44–47] show that clustering and association rule mining methods can be applied to construct adaptive and self-learning diagnostic systems.

Along with traditional machine learning approaches, deep and hybrid neural network models are being actively implemented. Studies [52–55] describe approaches based on convolutional neural networks (CNNs), auto-associative neural networks (AANNs), and generative adversarial networks (GANs) for automatic fault diagnosis in ventilation and cooling systems. These models provide a high degree of generalization and the ability to capture complex nonlinear relationships between microclimate parameters. Research works [56–60] explore feature selection methods, handling of imbalanced datasets, and the use of virtual sensors, which is particularly important when data are limited or partially missing.

A distinct line of research is represented by studies [61–67], which propose semi-supervised and modular approaches for fault diagnosis in variable refrigerant flow (VRF) and air conditioning systems. Such methods combine the strengths of expert knowledge and machine learning algorithms, enabling the construction of interpretable and efficient diagnostic models. Meanwhile, [68–70] highlight the role of cloud computing, big data analytics, and parallel algorithms in ensuring scalability and seamless integration of intelligent diagnostic systems into smart building infrastructures.

Summarizing the findings from studies [32–70], it can be concluded that modern building microclimate control systems are shifting from traditional reactive strategies toward intelligent, predictive, data-driven models. Machine learning methods-from classical algorithms such as PCA, SVM, and KNN to advanced deep learning and ensemble models-have proven to be effective tools for fault detection,

fault prediction, and energy optimization. The integration of such intelligent diagnostic solutions significantly enhances building energy efficiency, reliability, and occupant comfort, establishing intelligent fault detection as a key direction in the advancement of smart building microclimate technologies.

Based on the analysis of the literature [32-70], it can be concluded that data-driven intelligence methods, including both classical algorithms (such as PCA, SVM, and KNN) and modern approaches (such as deep learning, ensemble methods, and generative networks), are actively applied to ensure reliable and accurate fault detection in building microclimate systems, thereby improving energy efficiency and occupant comfort.

3.4 PCA approach with mathematical description

Principal Component Analysis, is a popular technique for analyzing large datasets with a high number of dimensions/features per observation. It reduces the dimensionality of datasets, allowing for easier visualization and analysis. One advantage of PCA is its ability to mitigate the «curse of dimensionality», which can slow down data processing, particularly in tasks like training machine learning models. By reducing the total number of features while preserving the most important information, PCA can improve computational efficiency without significantly impacting model performance. Additionally, PCA is an unsupervised method, meaning it does not require labeled data for training, making it versatile and widely applicable in various domains. Its working principle involves selecting features that contribute the most variance to the data distribution, known as principal components, enabling efficient dimensionality reduction while retaining critical information.

Recent studies, such as Attouri et al. (2024), [70] have proposed improved Kernel PCA-based approaches to enhance fault detection and monitoring in industrial applications. Their research demonstrates that kernel PCA significantly improves sensitivity to nonlinear faults compared to conventional PCA methods, enabling more accurate and robust fault diagnosis in real-time monitoring systems.

As well as visually inspecting scatter plots, we can calculate formal measures of the relationship between two continuous features using covariance and correlation. For two features, a and b, in a dataset of n instances, the sample covariance between a and b is:

$$cov(a,b) = \frac{1}{n-1} \sum_{i=1}^{n} ((a_i - a^-) \times (b_i - b^-)) \quad (1)$$

where a_i and b_i are values of features *a* and *b* for the i^{th} instance in a dataset, and a^- b^- are sample means of features *a* and *b*. Covariance values fall into the range $[-\infty, \infty]$ where negative values indicate a negative relationship, positive values indicate a positive relationship, and values near zero indicate that there is little or no relationship between the features.

Covariance is measured in the same units as the features that it measures. As are result, comparing the covariance between pairs of features only makes sense if each

pair of features is composed of the same mixture of units. Correlation is a normalized form of covariance that ranges between -1 and +1. We calculate the correlation between two features by dividing the covariance between the two features by the product of their standard deviations. The correlation between two features, *a* and *b*, can be calculate as:

$$corr(a,b) = \frac{cov(a,b)}{sd(a) \times sd(b)} \tag{2}$$

Where $cov(a,b)$ is the covariance between features *a* and *b* *sd(a)* and *sd(b)* are the standard deviations of *a* and *b* respectively. Because correlation is normalized, it is dimensionless and, consequently, does not suffer from the interpretability difficulties associated with covariance. Correlation values fall into the range $[-1,1]$, where values close to -1 indicate a very strong negative correlation (or covariance), values close to 1 indicate a very strong correlation, and values around 0 indicate no correlation. Feature that have no correlation are said to be independent.

Two tools that can be useful are the correlation matrix and covariance matrix. A covariance matrix contains a row and column for each feature, and each element of the matrix lists the covariance between the corresponding pairs of features. As a result, the elements along the main diagonal list the covariance between a feature and itself, in other words, the variance of the feature. The covariance matrix, usually denoted as $\sum$, between a set of continuous features, $\{a, b, \ldots., z\}$, is given as

$$\sum_{\{a,b,\ldots,z\}} = \begin{bmatrix} var(a) & cov(a,b) & \ldots & cov(a,z) \\ cov(b,a) & var(b) & \ldots & cov(b,z) \\ \ldots & \ldots & \ldots & \ldots. \\ cov(z,a) & cov(z,b) & & var(z) \end{bmatrix} \tag{3}$$

Similarly, the correlation matrix is just a normalized version of the covariance matrix and shows the correlation between each pair of features:

$$correlation\ matrix = \begin{bmatrix} corr(a,a) & corr(a,b) & \ldots & corr(a,z) \\ corr(b,a) & corr(b,b) & \ldots & corr(b,z) \\ \ldots & \ldots & \ldots & \ldots. \\ corr(z,a) & corr(z,b) & & corr(z,z) \end{bmatrix} \tag{4}$$

Step1. Data

We consider a dataset having n features or variables denoted by $X_1\ ;\ X_2; \ldots;\ X_n$.
Let there be N examples
Let the values of the i^{th} feature X_i be $X_{i1}\ ;\ X_{i2}; \ldots;\ X_{iN}$, as represented in Table 7.

Table 6 - Examples of Feature Configurations

Features	Example 1	Example 2 …	Example N
$\boldsymbol{X_1}$	X_{11}	X_{12}	X_{1n}
$\boldsymbol{X_2}$	X_{21}	X_{22} …	X_{2n}
⋮		…	
$\boldsymbol{X_i}$	X_{i1}	X_{i2} …	X_{in}
⋮			
$\boldsymbol{X_n}$	X_{n1}	X_{n2} …	X_{nN}

$$\overline{X_\iota} = \frac{1}{N}(X_{i1} + X_{i2} + \cdots + X_{iN}) \quad (5)$$

Step 2. Calculate the covariance matrix

$$Cov(X_i, X_j) = \frac{1}{N-1}\sum_{k=1}^{N}(X_{ik} - \overline{X_\iota})(X_{jk} - \overline{X_J}) \quad (6)$$

$$S = \begin{bmatrix} Cov(X_1,X_1) & Cov(X_1,X_2) & \dots & Cov(X_1,X_n) \\ Cov(X_2,X_1) & Cov(X_2,X_2) & \dots & Cov(X_2,X_n) \\ \vdots & \vdots & & \vdots \\ Cov(X_n,X_1) & Cov(X_n,X_2) & \dots & Cov(X_n,X_n) \end{bmatrix} \quad (7)$$

Step 3. Calculate the eigenvalues and eigenvectors of the covariance matrix.

1) Set up the equation: This is a polynomial equation of degree n. It has n real roots and these roots are the eigenvalues of S.

$$\det(S - \lambda I) = 0 \quad (8)$$

where:

det - (determinant): The determinant of a square matrix, which is a scalar value that provides important properties about the matrix, such as whether it is invertible.

C - the covariance matrix, which is a symmetric square matrix that represents the relationships between different variables in a dataset.

λ (lambda) - the eigenvalues of matrix which represent the amount of variance captured along the corresponding eigenvectors.

I - The identity matrix of the same size as, which has ones on the diagonal and zeros elsewhere.

2) If $\lambda = \lambda'$ is an eigenvalue. Then the corresponding eigenvector is a vector

$$U = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

(9)

Such that

$$(S - \lambda' I)\, U = 0 \tag{10}$$

Step 4. Calculate the eigenvalues and eigenvectors of the covariance matrix

We now normalize the eigenvectors. Given any vector X we normalize it by dividing X by its length. The length (or, the norm) of the vector

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \tag{11}$$

Is defined as:

$$\|X\| = \sqrt{{x_1}^2 + {x_2}^2 + \cdots + {x_n}^2} \tag{12}$$

We compute the n normalized eigenvectors $e_1, e_2, \ldots, e_n$ by

$$e_i = \frac{1}{\|U_i\|} U_i, \qquad i = 1,2, \ldots, n. \tag{13}$$

Step 5. Derive new data set

Order the eigenvalues from highest to lowest.

The unit eigenvector corresponding to the largest eigenvalues is the first principal component.

1) Let the eigenvalues in descending order be $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ and let the corresponding unit eigenvector be $e_1, e_2, \ldots, e_n$.

2) Choose a positive integer ρ such that $1 \leq \rho \leq n$.

3) Choose the eigenvectors corresponding to eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_p$ and form the following $p \times n$ matrix (we write the eigenvectors as row vectors):

$$F = \begin{bmatrix} {e_1}^T \\ {e_2}^T \\ \vdots \\ {e_p}^T \end{bmatrix} \tag{14}$$

Step 6. Derive new data set

4) We form the following $n \times N$ matrix:

$$X = \begin{bmatrix} X_{11} - \overline{X_1} & X_{12} - \overline{X_1} & \dots & X_{1N} - \overline{X_1} \\ X_{21} - \overline{X_2} & X_{22} - \overline{X_2} & \dots & X_{2N} - \overline{X_2} \\ \vdots & \vdots & & \vdots \\ X_{n1} - \overline{X_n} & X_{n2} - \overline{X_n} & \dots & X_{nN} - \overline{X_n} \end{bmatrix} \quad (15)$$

5) Next compute the matrix:

$$X_{new} = FX. \quad (16)$$

Note that this is a $p \times N$ matrix. This gives us a dataset of N samples having ρ features.

3.5 Mathematical justification explaining variance PCA

Let X be an $n \times p$ matrix whose column correspond to p variables and whose rows correspond to n samples or measurements. What does it mean to introduce new variables? In a very general sense, it means coming up with mapping (mathematical function)
$f: R^p \rightarrow R^q$ from the old variables to the new ones, such that it has an inverse $f: R^q \rightarrow R^p$, which restores the original data. The function f is then applied to each row of X to get the new data matrix, Z. This formulation is too general, however. For example, because R^p is isomoprhic (as a set) to R, we can encode everything in a single variable. But the relationship between the old variables and the new one very non-trivial. The new variable, even though it encodes precisely the same information, would tell us nothing meaningful about the data. Therefore, we need to impose some restrictions on the nature of f. Here we require that f be a linear function; so that $f(x) = xA$, where x is a row of X and A is a $p \times q$ matrix.

The new data matrix can be then computed as $Z = X\,A$. Now it's no longer possible to squeeze everything into a single variable; for the inverse mapping to exist we need $q \geq p$.

Because we are trying to reduce the dimensionality of the data, not expand it, we'll stick with $q = p$, so A is an invertible $p \times p$ square matrix.

But even an invertible linear transformation is too general for PCA because it does not preserve the total variance. From now on, we shall assume that the sample mean of each column of X is 0. This can be achieved by subtracting from each column of X have zero mean, so do the column of $Z = X\,A$:

$$\left(\frac{1}{n} \quad \frac{1}{n} \quad \dots \quad \frac{1}{n}\right) Z = \left(\frac{1}{n} \quad \frac{1}{n} \quad \dots \quad \frac{1}{n}\right) XA = (0\ 0\ \dots 0)A = (0\ 0\ \dots 0) \quad (17)$$

Then the covariance matrix of Z is

$$Cov(Z) = \frac{1}{n} Z'Z = \frac{1}{n}(AX)'(AX) = A'\left(\frac{1}{n} X'X\right) A = A'Cov(X)A \quad (18)$$

The total variance of X is the sum of the diagonal elements in the covariance matrix $Cov(X)$, i.e. its trace $Tr(Cov(X))$. In general, a transformation $Cov(X) \to A'Cov(X)A$ does not preserve the trace of the matrix $Cov(X)$. For instance, multiplying all variables by a constant number c is a linear transformation with matrix $c \cdot I$. This transformation will inflate the total variance by a factor of c^2.

To ensure that the total variance does not change, we require that $Cov(X) \to A'Cov(X)A$ is a similarity transformation, which always preserves the trace of the matrix it transforms. This is equivalent to saying that $A^{-1} = A'$, or that A is orthogonal. Preserving the total variance is not only (or even an important) reason to require that the change of variables is an orthogonal transformation. As we saw earlier, the transformation $c \cdot I$ inflates the total variance by the factor of c^2, but it does so by uniformly inflating the variance of each variable. So when we compute the fraction of the total variance explained by the variables, that common factor cancels out. The real problem is that we could rescale individual variables. Consider a 2×2 matrix

$$A = \begin{pmatrix} a_{11} & 0 \\ 0 & a_{22} \end{pmatrix} \tag{19}$$

Unless a_{11} and a_{22} are ± 1, A is not orthogonal. So, in general, A will not preserve the total variance of every matrix. However, for any given matrix X, there are many possible values of a_{11} and a_{22} that will preserve X's total variance: they form an ellipse

$$a_{11}^2 Var(X_{\cdot 1}) + a_{22}^2 Var(X_{\cdot 2}) = Var(X_{\cdot 1}) + Var(X_{\cdot 2}) \tag{20}$$

Recall that the objective of PCA is make the first variable explain the maximum fraction of the total variance. By choosing a_{22} close to zero (and inferring a_{11} from the above equation), we can make the fraction of variance "explained" by the first principal component arbitrarily close to 1 without transforming the data in any meaningful way. This kind of cheating is made impossible to maximize the explained variance. The typical use of PCA is to keep only the first $k < p$ principal components. Because PCA is an orthogonal transformation, this corresponds to projecting the data from its original p - dimensional space to a k -dimensional subspace. The remaining $p - k$ components are lost in this projection; so it makes sense to minimize the variability of the data in those directions. Because the total variance is constant, minimizing the variance of the last $p - k$ variables is the same as maximizing the variance of the first k variables. The choices we make in PCA are motivated precisely by this objective:

1. PCA itself is designed to maximize the variance of the first k components, and minimize the variance of the last $p - k$ components, compared to all other orthogonal transformations.

2. We choose the first k components, and not just some k components, because they have the highest variance out of all principal components.

3. We try to choose k big enough to make the lost information – the variance of the last $p - k$ components - sufficiently small [77-78].

PCA proves to be a robust and versatile technique for dimensionality reduction, especially in high-dimensional data scenarios where interpretability and computational efficiency are crucial. Its mathematical foundation – rooted in the computation of covariance matrices, eigenvalues, and eigenvectors – allows PCA to transform original features into a reduced set of orthogonal components that capture the most significant variance within the data. As demonstrated in the studies by Shlens [71] and Wang [72], PCA not only enhances visualization and modeling performance but also serves as an essential tool in fault detection systems without the necessity of explicit process modeling. Further investigations by Jung et al. [73] and Wang & Xiao [74] illustrate how PCA, sometimes in combination with other techniques such as LDA, can be successfully applied to detect sensor and mechanical faults in industrial processes. Moreover, advanced variants like Kernel PCA and optimized component selection, as discussed by Zhao & Wang [75] and Wang et al. [76], highlight the growing sophistication in tailoring PCA-based methods to specific diagnostic and control applications. Overall, PCA's theoretical and practical strengths make it a key method in modern machine learning and intelligent system design.

3.6 Summary

The chapter focuses on optimizing microclimate systems through the utilization of machine learning and fault detection techniques. It covers a range of topics including fault detection algorithms, machine learning algorithms such as supervised and unsupervised learning, as well as time series analysis. The integration of these techniques with building management systems is highlighted, emphasizing the importance of adaptive learning mechanisms in achieving optimal performance.

Additionally, the chapter introduces data preprocessing techniques such as Z-score normalization and dimensionality reduction using PCA, which enables effective data visualization and simplification. Clustering methods, including K-means and DBSCAN, are applied to detect outliers and segment operational states in both cold and hot seasons. The explained variance derived from PCA confirms the suitability of reducing features to two principal components while preserving most of the informational content. These combined approaches support real-time fault diagnostics and contribute to the development of intelligent, self-adaptive microclimate control systems.

The presented approaches are based on the authors' previous studies, which explored clustering techniques, machine learning integration, and fault detection for microclimate systems in residential and non-residential buildings [77-79]. These works demonstrated the applicability of PCA for dimensionality reduction, DBSCAN for anomaly detection, and data-driven models for optimizing microclimate management, forming the methodological foundation of this chapter.

4 EXPERIMENTAL SETUP AND SYSTEM ARCHITECTURE

4.1 Microclimate environments

The primary objective of the experiment was to detect and analyze potential faults or anomalies in microclimate parameters both inside the premises and in the surrounding outdoor area. To achieve this, a hardware setup was employed, consisting of over 16 sensors capable of measuring various microclimate characteristics, including temperature, humidity, carbon dioxide levels, and others.

The experimental methodology involved continuous data collection from the sensors at a high update frequency, enabling the prompt detection and recording of any deviations from the norm or abnormal situations in the microclimate. The real-time data collected was then transmitted to Google Sheets for further analysis.

The analysis of the obtained data included identifying anomalous values, discrepancies, or outliers in the microclimate parameters, which could indicate possible errors in the operation of the microclimate control system or abnormal situations requiring intervention. A scientific approach to data analysis enabled not only the identification of specific faults but also the recognition of trends and patterns in their occurrence, contributing to the development of effective strategies for managing and monitoring the microclimate in buildings [77-79].

In the context of collecting data from sensors in microclimate parameters, the use of the NodeMCU microcontroller represents an attractive alternative to the traditional Arduino platform. NodeMCU has several advantages, including a built-in Wi-Fi module and operation at 3.3 Volts. These features enable the creation of wireless data collection systems and ensure compatibility with a wide range of microclimate sensors, making it the preferred choice for various monitoring and management tasks in buildings, premises, or outdoor areas. Thus, the use of NodeMCU in microclimate monitoring systems presents a promising research direction, opening up new opportunities for the development of efficient and flexible solutions in the field of climate technologies.

The experiment was carried out in two different locations: a residential building and a non-residential building. Each of these places has its unique characteristics and factors influencing the indoor microclimate.

In the context of a residential building, the microclimate can be more variable, as each house may have its own heating, air conditioning, and ventilation systems. Factors such as geographical location, building type, and room size can also influence microclimate parameters.

In non-residential buildings, particularly those frequented by children, maintaining an optimal indoor microclimate is crucial due to their heightened sensitivity to environmental factors. Children are more susceptible to variations in temperature and humidity, which can affect their comfort and health. Additionally, factors such as the number of children present, their activity levels, and the presence of specialized areas like play zones and sleeping quarters can influence the indoor climate. Therefore, it’s essential to consider these elements when designing and managing the microclimate of such spaces to ensure a healthy and comfortable environment for young occupants.

Thus, in each of these places, differences in microclimate parameters may be observed due to various factors, including structural features, space usage, and user characteristics. Conducting the experiment in multiple locations makes it more representative and allows for consideration of the diversity of conditions in which the microclimate control system is used. The 7 shows Hardware Systems Indoors and Outdoors of a Residential building and Non-Residential building.

Figure 7 - Installed Hardware Complex Inside and Outside Rooms in a Residential building, in a Non-Residential building [80]

Table 6 provides detailed information about the experimental setup conducted in a single-storey Residential building equipped with central heating and air conditioning systems. It includes specifications of the house layout, heating and air conditioning units, sensor placements, data collection methods, and any other relevant details pertaining to the experimental environment.

Table 7 Overview of Environmental Monitoring and Experimentation in a Non-Residential building Setting: This table presents an overview of the environmental monitoring and experimentation conducted within a Non-Residential building setting. It outlines the setup utilized for monitoring various environmental parameters within the Non-Residential building, such as temperature, humidity, air quality, and noise levels. Additionally, it may include details about the experimental procedures, equipment used, sensor placements, and any other pertinent information related to the conducted experiments.

Table 6 - Description of Experimental Setup in a Single-Storey Residential building with Central Heating and Air Conditioning Systems

Parameter	Description
House Type	Residential building
Floors	Single-storey
Area	Approximately 150 square meters
Rooms	Living room, kitchen, three bedrooms, bathroom, wardrobe

Continuation of the table 6

Parameter	Description
Systems	Central heating and air conditioning system, ventilation system
Sensors	Temperature, humidity, CO2 level, and illumination sensors installed in various rooms of the house
Experiments	Variation of air temperature and humidity, activation and deactivation of the air conditioning system, emulation of faults in the heating and ventilation systems
Data Analysis	Statistical analysis, machine learning, time series forecasting
Objective	Study of the impact of factors on the indoor microclimate, detection and diagnosis of faults in the microclimate control systems
Results	Valuable data on the functioning of the microclimate and the effectiveness of control and management methods of the microclimate in private houses

Table 7- Overview of Environmental Monitoring and Experimentation in a Non-Residential building Setting

Parameter	Description
House Type	Non-Residential building
Location	Inside and outside the Non-Residential building building
Sensors	Installed inside and outside to measure temperature, humidity, CO2 level, and illumination
Experiments	Various scenarios of indoor air quality and temperature changes, activation of ventilation systems
Data Analysis	Statistical analysis, machine learning, time series analysis
Objective	Assessment of indoor environment quality and comfort in the Non-Residential building
Results	Obtaining valuable information about microclimatic conditions and effectiveness of interventions

A comparative Analysis of Parameters between Residential buildings and Non-Residential buildings: This table provides a comparative analysis of various parameters between Residential buildings and Non-Residential buildings. It may include comparisons of environmental parameters such as temperature, humidity, air quality, and noise levels, as well as structural differences, occupancy patterns, and energy consumption profiles between the two types of settings. The table aims to highlight similarities and differences in environmental conditions and characteristics between Residential buildings and Non-Residential buildings for the purpose of the conducted research.

Table 8 – A Comparative Analysis of Parameters between Residential buildings and Non-Residential buildings

Parameter	Residential building	Non-Residential building
Location	Located in rural or suburban areas, away from densely populated areas	Typically located in urban or suburban areas, closer to residential neighborhoods
Purpose	Residential dwelling for a family or individuals	Educational facility for young children
Size	Typically larger in size, with multiple rooms and outdoor space	Generally smaller in size, with classrooms, play areas, and administrative offices

Continuation of the table 8

Parameter	Residential building	Non-Residential building
Systems	Equipped with central heating, air conditioning, and ventilation systems for comfort	Often equipped with heating and air conditioning systems, as well as ventilation for comfort and air quality control
Environment	Focus on creating a comfortable living space with emphasis on privacy and relaxation	Emphasis on creating a stimulating learning environment with facilities for play and education
Sensors	Sensors may be installed to monitor indoor environmental parameters such as temperature, humidity, and air quality	Sensors may be installed to monitor indoor environmental parameters to ensure a safe and comfortable learning environment
Activities	Activities may include relaxation, recreation, and household tasks	Activities are primarily focused on early childhood education, play, and socialization
Experimentation	Experiments may involve studying the impact of various factors on indoor comfort and efficiency of home systems	Experiments may involve studying the impact of environmental factors on child development and learning outcomes
Data Analysis	Data analysis may focus on optimizing home systems for energy efficiency and comfort	Data analysis may focus on improving the learning environment and ensuring the safety and well-being of children
Objective	To create a comfortable and efficient living environment for occupants	To provide a safe, stimulating, and nurturing environment for children to learn and grow

4.2 Hardware Utilized for Experimental Setup and Data Collection

Hardware Complex Based on NodeMCU for Microclimate Parameters Measurement

The hardware complex assembled on the NodeMCU microcontroller platform represents a powerful tool for measuring and monitoring microclimate parameters. NodeMCU is based on the ESP8266 microcontroller and provides extensive capabilities in the field of Internet of Things (IoT).

NodeMCU Features:

ESP8266 Microcontroller: NodeMCU is equipped with the ESP8266 microcontroller with Xtensa architecture, providing high performance and efficient device management.

Wi-Fi Support: The microcontroller supports Wi-Fi IEEE 802.11 b/g/n standards, allowing the device to exchange data wirelessly with external sources.

Programming Flexibility: NodeMCU supports programming both using the Lua language and the Arduino IDE, providing versatility and accessibility for developers.

Integrated Peripheral Devices: Built-in analog and digital inputs/outputs, as well as USB and GPIO connectors, provide the ability to connect and interact with various sensors and devices.

Library and Module Support: NodeMCU has wide support for libraries and modules such as HTTP and MQTT clients, facilitating interaction with external servers.

Wi-Fi Firmware Update: The ability to update firmware over Wi-Fi (OTA) allows for convenient deployment of firmware changes remotely.

Embedded System Description:

NodeMCU is an embedded system specifically designed for measuring microclimate parameters. This embedded system has limited computational resources, ensuring stable execution of microclimate measurement functions, programmatically controlled sensor interaction, and reliable operation over an extended period.

Thus, the hardware complex based on NodeMCU provides a convenient and efficient solution for measuring microclimate parameters, ensuring accuracy and reliability in various conditions.

4.3 Sensor Deployment Strategy

The sensor deployment strategy plays a key role in developing a smart microclimate data acquisition system in buildings. It ensures comprehensive coverage of critical environmental parameters and affects the overall reliability of the system. Important factors include the building's characteristics, external influences, sensor redundancy in case of failure, and data synchronization. The main sensors used in constructing the hardware system are shown in Figures 8–14.

Figure 8- Full-fledged platform based on the ESP8266 module

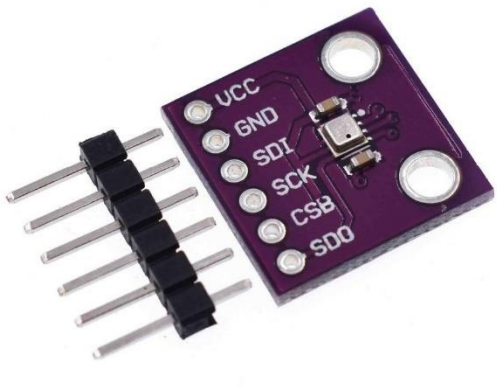

Figure 9 - Atmospheric pressure, humidity and temperature sensor BME280

The BME280 sensor is a compact and efficient device that measures temperature, humidity, and atmospheric pressure, making it ideal for microclimate research. Its high accuracy, low power consumption, and ease of integration with microcontrollers and IoT platforms enable real-time environmental monitoring. The

sensor’s versatility allows it to be used in various research fields, providing reliable data crucial for understanding microclimate dynamics.

Figure 60- Digital sensor for air quality monitoring CCS811

The CCS811 digital sensor is essential for air quality research, offering real-time measurement of volatile organic compounds (VOCs) and equivalent CO_2 (eCO_2) levels. Its high accuracy and reliability make it ideal for assessing indoor and outdoor air pollution, ventilation efficiency, and health impacts. Easily integrated with IoT platforms, it supports remote monitoring and is widely applicable in environmental science, public health, and building studies.

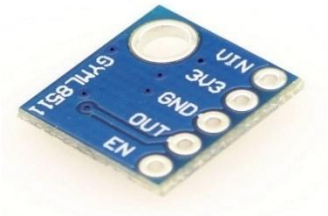

Figure 11 - UV sensor ML8511

The ML8511 UV sensor is crucial for monitoring ultraviolet (UV) radiation levels in various environments. It measures UVA and UVB radiation, helping assess health risks like skin damage and cancer, as well as environmental impacts like ozone depletion and climate change. It supports research in occupational safety, solar energy, and public health interventions. With easy integration into IoT platforms, the ML8511 provides accurate, real-time data for a range of applications, making it an essential tool for UV radiation studies and protective strategies.

Figure 12 - Digital light sensor BH1750

The BH1750 digital light sensor is crucial for measuring ambient light levels in various environments. It provides accurate, reliable data for environmental monitoring, building performance analysis, and occupant comfort studies. Researchers use it to assess lighting conditions' impact on health, productivity, and well-being, as well as to optimize daylighting strategies and energy efficiency. The sensor also supports urban planning by evaluating outdoor light levels in public spaces. Its integration with IoT platforms allows for real-time data collection, making it an essential tool for research in lighting and its effects.(https://learn.adafruit.com/adafruit-bh1750-ambient-light-sensor/overview)

Figure 13 - Multifunctional ammeter

Figure 14- 3 Axis gyroscope and accelerometer MPU6050

During the implementation of the building's microclimate monitoring system, sensors were strategically installed both indoors and outdoors to ensure accurate, representative data collection. Key parameters included temperature, humidity, pressure (BME280), CO_2 and TVOC (CCS811), UV radiation (ML8511), light intensity (BH1750), motion (MPU6050), and energy consumption (ammeter). Indoor sensors were placed to avoid interference from heat sources and ensure coverage of areas with high human activity or environmental impact. Outdoor sensors were shielded from direct weather influences to maintain measurement accuracy. This approach enabled reliable real-time monitoring and effective analysis of the building's environmental and energy conditions. Sensor placement in

Table .

To ensure accurate measurements, sensors should be placed strategically based on their function and environmental conditions. Temperature, humidity, and pressure

sensors (BME280) should be installed indoors, away from heat sources and drafts, and outdoors on walls shielded from sunlight. CO_2 and TVOC sensors (CCS811) should be placed in areas prone to air pollution, like kitchens and offices. Ultraviolet radiation sensors (ML8511) should be installed outdoors, away from shading objects. Light sensors (BH1750) are best positioned in representative areas such as near windows. Motion sensors (MPU6050) should be mounted on moving or vibrating objects. Energy sensors should be placed near electrical systems. All sensors should be installed at a height of 1.5 meters, away from heat, sunlight, and high-traffic areas.

Table 9 - Sensor placement

Parameter	Location for internal sensors	Location for outdoor sensors
Temperature	In the center of the room, away from windows and heating devices	On external walls protected from rain and sun
Humidity	In corners, near windows and ventilation	On external walls protected from weather conditions
Pressure	On the walls, away from heat sources and convection	On the outside wall of the building, in a protected location
CO2, TVOC	In places with possible sources of air pollution	Not required
Ultraviolet	Not required for internal monitoring	On the roof or walls not obscured by other objects
Illumination	In the center of the room, near windows or artificial light sources	On the facades of buildings, away from buildings and trees
Energy parameters	On distribution boards and near electrical appliances	Not required
Movement (MPU6050)	On objects with possible movements (e.g. doors, walls)	Not required

The conducted experiment demonstrated the effectiveness of the microclimate monitoring system based on the NodeMCU platform and a set of sensors including BME280, CCS811, ML8511, BH1750, MPU6050, and a multifunctional ammeter. The system ensured accurate data collection in both residential and non-residential environments. NodeMCU's built-in Wi-Fi, programming flexibility, and compact design make it an ideal choice for Internet of Things (IoT) applications [81-82].

The selected sensors are well-supported in technical documentation and literature [83–87], confirming the validity of their application. The strategic deployment of sensors provided reliable measurements and enabled the detection of anomalies in microclimate conditions, contributing to improved comfort, safety, and energy efficiency in buildings.

4.4 Summary

This chapter describes the experimental setup used for microclimate monitoring in various environments, such as residential and non-residential buildings. The main goal of the experiment was to identify faults in microclimate systems using over 16 sensors measuring parameters like temperature, humidity, carbon dioxide levels, and others. The system is based on the NodeMCU microcontroller, which allows for data

collection from sensors and transmission for analysis. The experiment was conducted in both residential and non-residential buildings, where differences in microclimate were influenced by structural features and the use of spaces. Various heating and air conditioning systems were used, and data collection and analysis methods included statistical techniques and machine learning.

The chapter also provides details on building types, equipment, and sensor placement strategies, as well as a comparative analysis of microclimate parameters between different building types. The NodeMCU-based system proved effective for detecting anomalies and diagnosing faults in microclimate systems, demonstrating the importance of IoT devices in such research.

5 DATA UNDERSTANDING AND VISUALIZATION

5.1 Variables monitored

In this chapter, we delve into the exploration and visualization of the dataset outlined in Appendix A. This dataset comprises microclimate parameters such as Indoor and Outdoor Temperature, Indoor and Outdoor Humidity, Dew-point, Pressure, TVOC, Power, Current, Voltage, Aftershock, CO2, UV-radiation. Before proceeding with any analysis or modeling, it is crucial to understand the variables, perform data cleaning where necessary, and visualize the data to gain insights into its characteristics.

The variables measured are the following. Temperature: Measurement of ambient temperature in degrees Celsius.

Indoor and Outdoor Humidity: Levels of humidity indoors and outdoors, expressed as a percentage.

Dew-point: Dew-point temperature, indicating the temperature at which air becomes saturated and dew forms.

Pressure: Atmospheric pressure recorded in millibars or other appropriate units.

TVOC (Total Volatile Organic Compounds): Concentration of volatile organic compounds in the air, often measured in parts per million (ppm).

Power, Current, Voltage: Electrical parameters representing power consumption, current flow, and voltage levels.

Aftershock: Data related to seismic activity or aftershocks, if applicable.

CO2: Carbon dioxide levels in the air, usually measured in parts per million (ppm).

UV radiation refers to electromagnetic radiation with a wavelength shorter than that of visible light, but longer than X-rays. It is typically measured in microwatts per square centimeter (μW/cm²) using specialized UV radiation meters or sensors.

Describing Variables:The dataset consists of several variables capturing various aspects of the microclimate. Graphs illustrating the changes in all microclimate parameters over time have been created. To gain a comprehensive understanding of the data, we meticulously examined each variable, noting their data types, range, distribution, and any potential outliers or missing values.

Standard Microclimate Parameters for Indoor Environments

To ensure thermal comfort, air quality, and energy efficiency in building microclimate systems, it is essential to refer to internationally recognized standards and regulations. The table below summarizes the normative values for the key parameters considered in this study, based on sources such as ASHRAE, ISO, WHO, and relevant national standards (e.g., GOST, SNIP).

Table 10- Recommended indoor environmental parameters and their standard sources

Parameter	Description	Recommended Values	Standard Source
T	Indoor air temperature	Winter: 20–24 °C; Summer: 23–26 °C	ASHRAE 55, GOST 30494-2011
Tout	Outdoor air temperature	Depends on the climate zone	-

Continuation of the table 10

Parameter	Description	Recommended Values	Standard Source
H	Indoor relative humidity	30–60 %, up to 70 % in summer	ASHRAE 55, GOST 30494-2011
Hout	Outdoor relative humidity	-	-
DP	Indoor dew point temperature	≤ 16 °C	ASHRAE 55
DPout	Outdoor dew point temperature	-	-
CO_2	Indoor carbon dioxide concentration	≤ 1000 ppm (up to 1500 ppm temporarily)	ASHRAE 62.1, ISO 16000-26
CO_2out	Outdoor CO_2 concentration	~400 ppm	-
TVOC	Total Volatile Organic Compounds	≤ 0.3–0.5 mg/m³ (up to 1 mg/m³ temporarily)	ISO 16000-6, WHO
P	Indoor air pressure	~1013 hPa (normal atmospheric pressure)	ISO 2533
Pout	Outdoor air pressure	-	-
V	Air velocity	≤ 0.2–0.3 m/s	ASHRAE 55, GOST 30494-2011
C / L	Illuminance (lighting level)	300–500 lx (offices), ≥ 100 lx (corridors)	SNIP 23-05-95, EN 12464-1
A	Noise level	≤ 40 dB (residential), ≤ 50 dB (offices)	SNIP 23-03-2003, GOST 12.1.003-83
Pwr	Power consumption	Monitored per energy management standards	ISO 50001
UVr	Ultraviolet radiation	≤ 0.1 mW/cm²	ISO 15858, ICNIRP
Note - For outdoor parameters (Tout, Hout, DPout, CO_2out, Pout), no standard limits are imposed, as these are used for comparative or contextual analysis			

Table 11 – Overview of Indoor Climate Standards and Controlled Parameters

Abbreviation	Full Name	Relevant Microclimate Parameters
ASHRAE 55	Thermal Environmental Conditions for Human Occupancy	T, H, DP, V, A
ASHRAE 62.1	Ventilation for Acceptable Indoor Air Quality	CO_2, CO_2out, TVOC, V
ISO 16000-6	Indoor Air – Part 6: Determination of VOCs in Indoor and Test Chamber Air	TVOC
ISO 16000-26	Indoor Air – Part 26: Assessment of the Indoor Air Quality Using CO_2 Concentration	CO_2, CO_2out
ISO 50001	Energy Management Systems – Requirements with Guidance for Use	Pwr, C
ISO 2533	Standard Atmosphere	T, P, H, DP, Tout, Pout, Hout, DPout
ISO 15858	UV-C Devices – Safety Information	UVr
ICNIRP	Guidelines on Limits of Exposure to Ultraviolet Radiation (180–400 nm)	UVr
GOST 30494-2011	Residential and Public Buildings – Indoor Climate Parameters	T, H, DP, CO_2, V, L

Continuation of the table 11

Abbreviation	Full Name	Relevant Microclimate Parameters
GOST 12.1.003-83	Occupational Safety Standards System – Noise – General Safety Requirements	A
SNIP 23-05-95	Natural and Artificial Lighting	L
SNIP 23-03-2003	Noise Protection	A
EN 12464-1	Light and Lighting – Lighting of Workplaces – Part 1: Indoor Workplaces	L
Note: The recommended values are based on ASHRAE 55, GOST 30494-2011, and other standards (see References [88-94]).		

Official reports from the Statistics Committee of the Republic of Kazakhstan [95] were referenced to illustrate national trends in energy consumption and housing services, providing contextual support for the study.

The recommended indoor environmental parameters and their corresponding standard sources are listed in Table 10.

An overview of indoor climate standards and the microclimate parameters they address is provided in Table 11.

5.2 Data visualization

Visualization:

Visualizing the data is paramount to comprehend its underlying patterns and relationships. We employed a variety of visualization techniques, including histograms, scatter plots, and box plots, to explore the distribution and relationships between variables. These visualizations provided valuable insights into. All figures about microclimate parametres given in Appendix A.

Data Cleaning:

Upon initial inspection, we identified several data quality issues that required cleaning. These included [list common data cleaning tasks performed, such as handling missing values, correcting data formats, and removing duplicates]. Additionally, we addressed outliers that could significantly affect the analysis results.

During the data cleaning process, several data quality issues were identified upon initial inspection, including handling missing values, correcting data formats, removing duplicates, and addressing outliers that could significantly affect the analysis results. Z-score was used in this stage, calculated using the formula:

$$Z= X−μ/ σ$$

where:

Z - is the Z-score of the data point,

X - is the individual data point,

μ - is the mean (average) of the dataset,

σ (sigma) - is the standard deviation of the dataset.

The indoor temperature shows a slight increase due to the placement of the hardware complex in the kitchen of a private house. Cooking processes and the

presence of a large number of people indoors naturally lead to a rise in temperature. Conversely, the outdoor temperature exhibits somewhat higher readings, attributed to the placement of the hardware complex on the sunny side. Data collection occurred 4 seasons. Graphs illustrating the changes in all microclimate parameters over time have been created. For the remaining microclimate parameters, graphs are presented in Appendix A (Figure A1- Figure A31) Original and Cleaned Data for Residential and Non-residential Buildings).

The analysis of raw data for the dew point in a residential space during the autumn and winter periods yielded the following results. In the autumn period, dew point values ranged from 0 to 20°C, with most values falling between 5 and 10°C, though occasional peaks reached 20°C. During the winter period, a sharp decrease in dew point values was observed, indicating a significant drop in temperature and possible changes in humidity levels inside the building. These fluctuations may be linked to variations in air temperature and indoor humidity. However, further data processing and analysis will be necessary for more precise conclusions.

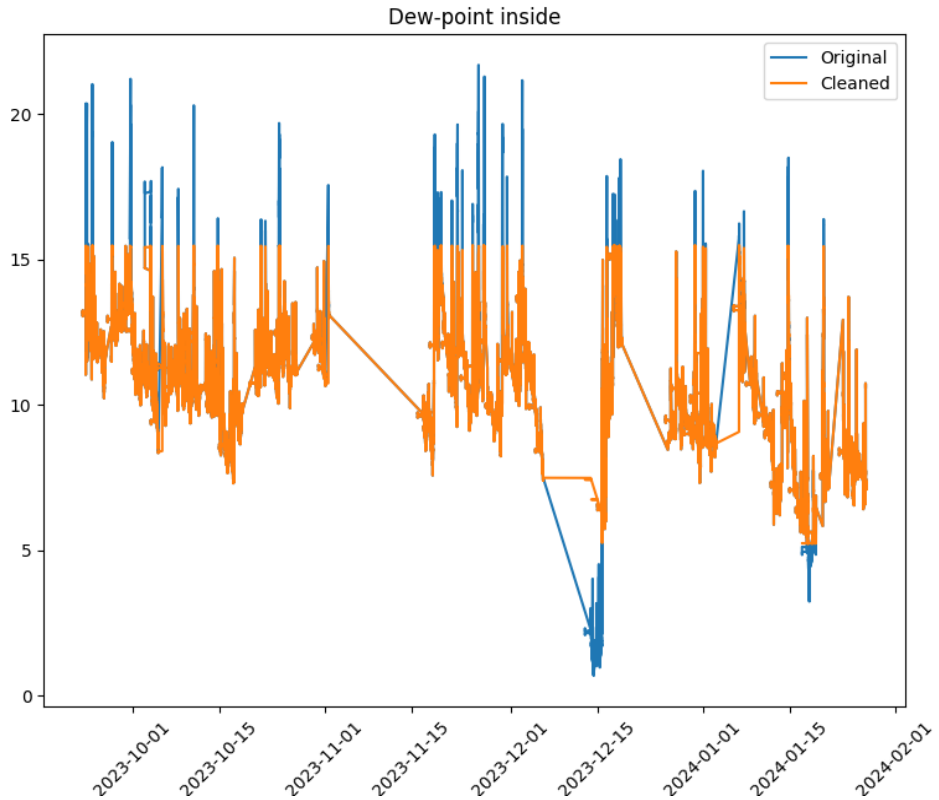

Figure 15 – Dew-point inside data after cleaning data

After applying the Z-score technique, noticeable changes occurred in the graph. In the autumn period, the values now range from 8°C to 16°C, with the data being cleaned of outliers and anomalous values. This suggests that the extreme values, previously observed outside this range, were likely outliers that were removed. Similarly, in the winter period, the data was cleaned, and the values now range from 5°C to 16°C. The application of the Z-score method effectively reduced the impact of extreme deviations, providing a more accurate representation of the typical fluctuations of the dew point during these seasons. This improvement indicates that the data has become more reliable for further analysis and model building.

In Appendix A, we provide a detailed overview of the dataset, including its structure, variable definitions, and initial data exploration findings. Through rigorous data understanding and visualization, we established a solid foundation for subsequent analysis and modeling tasks. In some graphs, it is apparent that two parameters are combined, such as CO2 levels, indoor and outdoor temperatures, and so forth. During monitoring, errors occurred intermittently, possibly due to data transmission issues via Wi-Fi, power outages in private residences, or sensor malfunctions. Elevated readings can be observed in certain parameters. Like as example in Figure presented graph after cleaning data, all graphs after using Z-score shows in Appendix B.

5.3 Variable's Correlations

In the correlation analysis, residential buildings show strong positive links between indoor temperature and dew-point (0.74) and strong negative correlations with outdoor temperature (-0.68) and energy consumption (-0.68). CO_2 and TVOC are highly correlated (0.85), while current, power, and voltage show near-perfect correlation (0.99).

In non-residential buildings, indoor humidity strongly correlates with dew-point (0.95) and indoor–outdoor pressure (0.93). CO_2 and TVOC moderately correlate with their outdoor counterparts, and voltage, current, and power remain nearly perfectly correlated (0.98).

Overall, residential buildings show stronger interactions among indoor comfort parameters, while non-residential buildings are more influenced by outdoor conditions and energy use.

As shown in recent microclimate studies [96], data visualization techniques such as histograms, scatter plots, and box plots are essential for identifying spatial and temporal patterns in building energy consumption.

During the cleaning phase, issues such as missing values, outliers, and inconsistent formats were identified and addressed. In particular, the Z-score technique was employed to detect and remove statistical outliers, which significantly improved the data quality and allowed for more accurate representation of seasonal changes in indoor parameters such as the dew point [97-98].

Table 12 - Comparison of Correlation Patterns Between Residential and Non-Residential Buildings

Parameter	Residential Building Correlation	Non-Residential Building Correlation
T	Strong positive correlation with DP (0.74), negative with Tout (-0.68), negative with Pwr (-0.68)	Strong positive correlation with DP (0.40), moderate positive with Tout (0.40)
Tout	Moderate negative correlation with T (-0.68), moderate positive with Energy (0.52)	Moderate positive correlation with DPout (0.65), moderate negative with Pout (-0.64)
H	Strong positive correlation with DP (0.69)	Strong positive correlation with DP (0.95), moderate positive with Hout (0.66)

Continuation of the table 12

Parameter	Residential Building Correlation	Non-Residential Building Correlation
Hout	No significant correlation noted	Moderate positive correlation with H (0.66)
DP	Strong positive correlation with T (0.74), H (0.69)	Strong positive correlation with H (0.95), moderate positive with DPout (0.65)
DPout	Moderate positive correlation with DP (0.51)	Moderate positive correlation with DP (0.65)
TVOC	Strong positive correlation with CO2 (0.85)	Moderate positive correlation with CO2 (0.71), strong with CO2out (0.79)
CO2	Strong positive correlation with TVOC (0.85)	Moderate positive correlation with CO2out (0.71), TVOC (0.79)
CO2out	No significant correlation noted	Moderate positive correlation with CO2 (0.71), TVOC (0.79)
P	Strong positive correlation with Pout (0.95)	Strong positive correlation with Pout (0.93)
Pout	Strong positive correlation with P (0.95)	Strong positive correlation with P (0.93)
C	Strong positive correlation with Pwr (0.99), V (0.99)	Strong positive correlation with Pwr (0.98), V (0.98)
V	Strong positive correlation with C (0.99), Pwr (0.99)	Strong positive correlation with C (0.98), Pwr (0.98)
A	Strong positive correlation with Pwr (0.99)	Strong positive correlation with Pwr (0.98)
Pwr	Strong positive correlation with C (0.99), V (0.99)	Strong positive correlation with C (0.98), V (0.98)
UVr	No significant correlation noted	No significant correlation noted
L	No significant correlation noted	No significant correlation noted

The entire preprocessing pipeline aligns with best practices in environmental data science and building microclimate research, particularly when dealing with multi-seasonal indoor/outdoor monitoring [99-100].

5.4 Summary

This section provided an in-depth exploration and visualization of the dataset, which included the identification and handling of data quality issues such as missing values, incorrect formats, duplicates, and outliers. Various visualization techniques, including histograms, scatter plots, and box plots, were employed to gain insights into the distribution and relationships between the variables. The Z-score method was used to clean datasets. Key patterns and trends, such as seasonal variations and anomalies likely caused by sensor malfunctions or data transmission issues, were identified. Correlation analysis highlighted significant relationships between different microclimate parameters, such as the strong positive correlation between indoor humidity and dew-point, the strong negative correlation between indoor temperature and energy consumption, and the near-perfect correlation between voltage, current, and power consumption. Additionally, the analysis revealed the strong positive correlation between indoor and outdoor pressure, as well as the close relationship

between CO2 and TVOC concentrations. Overall, the data understanding and visualization process established a solid foundation for further modeling and analysis.

6 DATA PREPARATION

6.1 The data preparation phase

In the phase of Data Preparation for an intelligent fault detection system in building microclimate control, the second stage entails the preprocessing of collected data to render it amenable for subsequent analysis and modeling. This phase encompasses a series of operations aimed at refining the raw data, including outlier and anomaly detection and removal, imputation of missing values, conversion of categorical features into numerical representations, standardization or normalization of feature scales, and partitioning the dataset into training and testing subsets.

The primary objective of this phase is to ensure that the data is suitably prepared for utilization in machine learning algorithms or other analytical methodologies. Effective data preparation is paramount for achieving accurate and reliable outcomes in subsequent analysis and model development.

Data preparation encompasses the systematic cleaning, transformation, and structuring of raw data to conform to analytical or modeling requirements. This process encompasses diverse tasks such as managing missing data, eliminating duplicate entries, encoding categorical variables, standardizing or normalizing numerical attributes, and partitioning the dataset for model training and evaluation. The overarching aim of data preparation is to guarantee the integrity, completeness, and appropriateness of the data for robust analysis and modeling endeavors.

6.2 Data cleaning

After data cleaning, various microclimate parameters showed improvements in stability. For temperature, both indoor and outdoor, the range of values narrowed, indicating more stable conditions. Humidity and dew point also became more consistent, with decreased standard deviations, especially for outdoor conditions. TVOC and CO2 concentrations showed reduced variability, indicating more homogeneous indoor air quality. Pressure, both indoor and outdoor, as well as electrical parameters like current, voltage, and power, showed slight reductions in standard deviation, signaling more stable conditions. The aftershock data and UV radiation also became more stable after cleaning, with significant reductions in standard deviation. Finally, light conditions demonstrated a noticeable reduction in variability, indicating more consistent lighting conditions. Tables 13 and 14 provide a comparison of the microclimate parameters before and after cleaning. These tables present the following metrics for each parameter: minimum, maximum, standard deviation (STD), and average (AVG).

Table 13 - Comparison of Microclimate Parameters Before and After Cleaning (Min, Max, STD, AVG) using Residential building datasets

Microclimate parameters	Before Cleaning Data				After Cleaning Data			
	Min	Max	STD	AVG	min	max	STD	AVG
1	2	3	4	5	6	7	8	9
T	18.05	37.43	31.51	1.99	28.53	34.49	31.56	1.66
T_{out}	-20.76	65.94	13.61	12.68	-7.73	33.08	9.64	10.90
H	19.35	84.68	3.08	27.38	22.76	32.00	2.33	27.22

Continuation of the table 13

1	2	3	4	5	6	7	8	9
H_{out}	0.00	87.75	19.89	50.77	20.94	80.55	16.87	52.29
DP	0.69	21.68	2.56	10.35	6.51	14.19	1.95	10.43
DP_{out}	-59.8	15.39	6.45	-0.03	-9.71	9.64	4.79	0.86
TVOC	0.00	3582	91.70	22.35	0.0	159.00	29.95	15.51
CO_2	0.00	3341.0	245.64	516.33	400.0	882.00	121.16	480.79
CO_{2out}	0.00	17590.0	376.82	699.92	400.0	1262.0	245.17	643.27
P	494.60	723.04	3.90	705.60	699.76	711.44	3.36	705.59
P_{out}	695.01	724.39	3.97	706.29	0.0	3.69	0.32	0.58
C	0.00	3.69	0.32	0.58	0.19	1.06	0.31	0.58
V	0.00	251.30	12.43	224.34	205.7	242.90	10.40	224.79
A	18.00	65423.0	1898.48	65214.38	63428.0	65423.00	322.82	65275.84
Pwr	0.00	782.30	56.22	85.89	22.2	170.20	54.00	85.08
UVr	-7.47	0.00	0.27	-7.22	-7.47	-6.83	0.12	-7.23
L	0.00	54612.0	13724.76	5880.24	0.0	26467.0	7071.95	3412.38

Table 14 - Comparison of Microclimate Parameters Before and After Cleaning (Min, Max, STD, AVG) using Non-Residential building datasets

Microclimate parameters	Before Cleaning Data				After Cleaning Data			
	min	max	STD	AVG	min	max	STD	AVG
T	21.85	45653.00	165.59	27.25	21.85	33.47	1.32	26.65
T_{out}	-17.93	71.96	15.41	14.11	-8.99	37.22	11.01	12.12
H	8.34	45641.00	110.90	25.97	8.34	77.35	7.17	25.70
H_{out}	0.00	80.17	20.46	41.50	10.83	72.18	18.22	43.17
DP	-11.26	18.74	4.68	4.81	-2.20	11.83	4.15	5.06
DP_{out}	-63.46	13.47	7.82	-2.25	-13.97	9.48	5.71	-0.88
TVOC	0	0	0	0	0	0	0	0
CO_2	0	0	0	0	0	0	0	0
CO_{2out}	0.00	18012.0	511.88	814.98	400.00	1582.00	377.00	769.93
P	464.93	706.68	4.11	690.82	684.66	696.98	3.14	690.75
P_{out}	0.00	708.52	86.98	680.43	679.84	708.52	4.10	691.69
C	0.00	1.65	0.16	0.08	0.0	0.32	0.06	0.02
V	0.00	242.40	16.13	231.26	210.10	242.40	3.89	232.31
A	1.00	65402.00	4394.73	64732.11	58791	65402.0	638.29	65018.71
Pwr	0.00	329.60	30.61	14.87	0.00	60.60	6.20	2.79
L	0.00	54612.00	16044.90	7502.95	0.00	31570.00	8266.53	4128.19
UVr	0.00	9.62	1.19	9.29	9.22	9.62	0.07	9.44

After cleaning the data for all parameters considered (temperature, humidity, dew point, TVOC, CO_2, pressure, current, voltage, seismic activity, power, UV radiation and irradiance), a decrease in standard deviation is observed, indicating more stable conditions both inside and outside the building, while the average values for most parameters remain virtually unchanged, indicating improved homogeneity and reduced variability in the data.

As demonstrated by Capozzoli et al. (2015), thorough data cleaning and outlier detection significantly enhance the accuracy of automated anomaly detection in building energy management systems.

6.3 Summary

The Data Preparation chapter focuses on the processes required to prepare raw microclimate data for analysis and modeling. Key steps include cleaning the data by addressing issues such as missing values, outliers, and inconsistencies. After cleaning, various microclimate parameters, such as temperature, humidity, CO_2 levels, and pressure, exhibited more stable conditions, with reduced variability and more consistent values. For example, the standard deviation of temperature and humidity decreased, suggesting a more uniform and reliable dataset. The chapter also compares pre- and post-cleaning data, highlighting the improvements made in terms of stability and data quality. Tables with comparisons of min, max, standard deviation, and average values before and after cleaning for both the Residential and Non-Residential buildings datasets were presented. These results underscore the importance of data cleaning in ensuring accurate and reliable outcomes for subsequent analysis and machine learning model development.

7 FAULT DETECTION RESULTS

In modern integrated microclimate systems, intelligent methods such as multi-agent systems are widely applied to coordinate between various rooms in residential and non-residential buildings, optimizing the balance between comfort and energy consumption (Altayeva & Omarov, 2018). Thanks to distributed control and the ability to adapt to changing conditions, these systems enhance flexibility, adaptability, and scalability of microclimate management. To improve the reliability and robustness of such systems, the implementation of effective fault diagnosis methods is required, enabling timely detection of hidden equipment failures and reduction of excessive energy consumption – an essential factor for maintaining high comfort levels and minimizing operational costs.

This study proposes a methodology for automatic fault detection and diagnosis in HVAC systems using modern machine learning techniques and statistical approaches, including clustering algorithms (K-means, DBSCAN) and PCA. This comprehensive approach aims to increase the resilience, reliability, and energy efficiency of microclimate systems in both residential and commercial premises, while facilitating real-time fault prevention and mitigation, thereby significantly reducing the risk of system failures and extending equipment lifespan.

Unlike the work of Uskenbayeva et al. (2022), which primarily focuses on air quality control using neural networks, the present research broadens the functionality of microclimate systems by integrating fault diagnosis methods. This integration not only ensures the maintenance of a comfortable and safe indoor environment but also significantly enhances system operational efficiency by reducing energy consumption and maintenance costs, aligning with contemporary demands for energy efficiency and sustainable development.

In this chapter, we present the results from applying PCA and clustering techniques to both the Residential building and Non-Residential building datasets. These results offer insights into how the data can be effectively reduced in dimensionality, as well as how different patterns and structures emerge when clustering the data.

PCA is a powerful technique used for dimensionality reduction and data exploration. By transforming the data into a set of linearly uncorrelated variables known as principal components (PCs), we can capture the most significant patterns of variance within the dataset. The Figures 16-19 shows Using PCA model for Residential and Non-Residential buildings for Hot and Cold seasons.

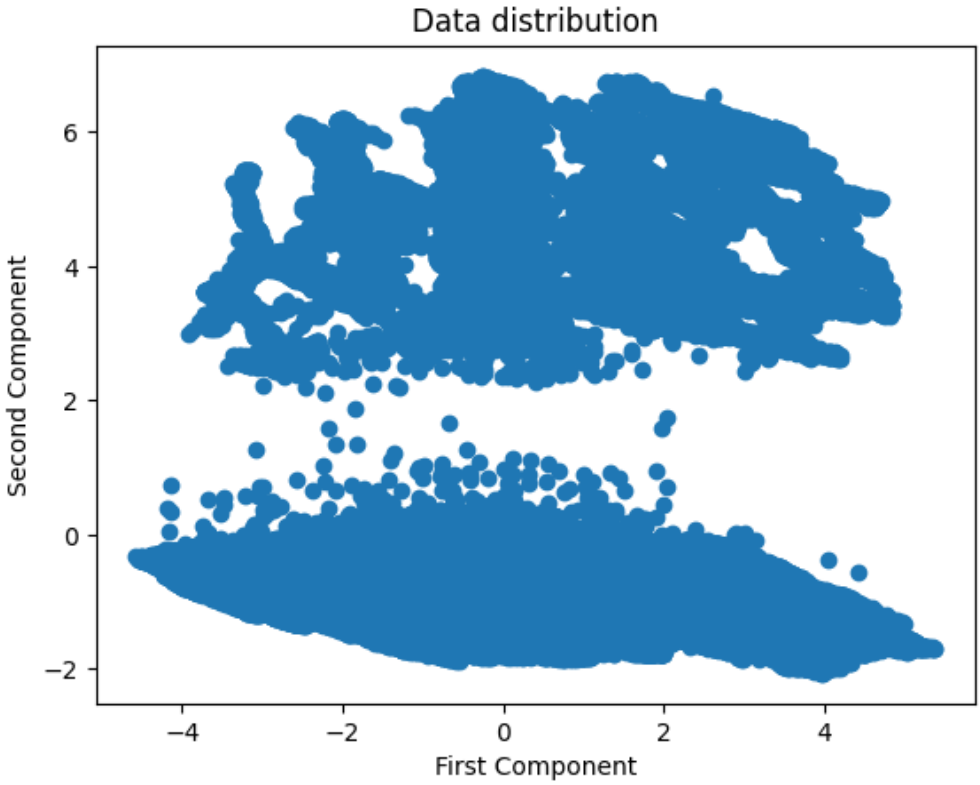

Figure 16 - Using PCA model for Residential building (Cold season)

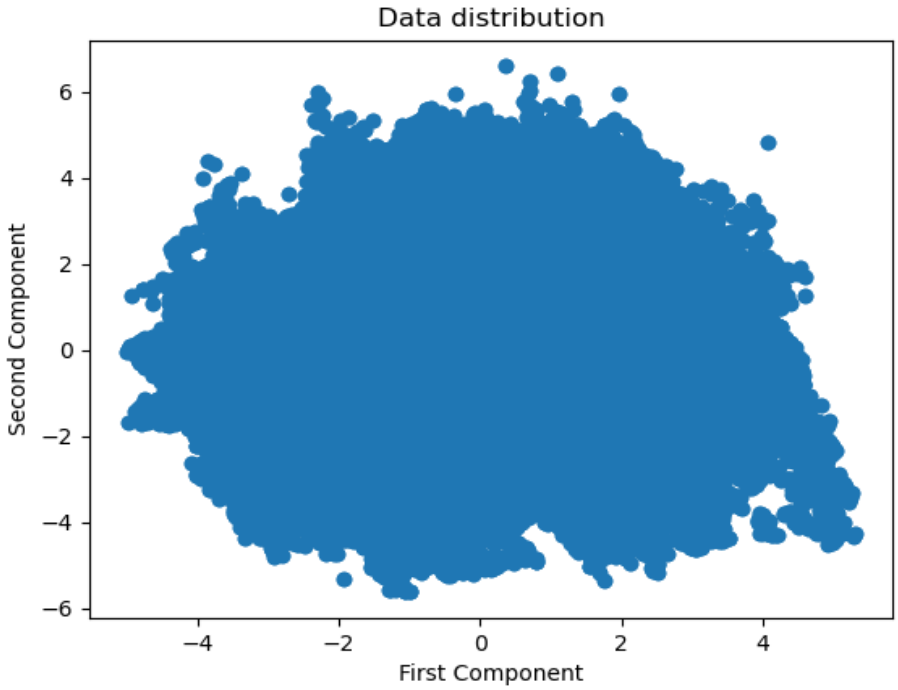

Figure 17 - Using PCA for Residential building (Hot season)

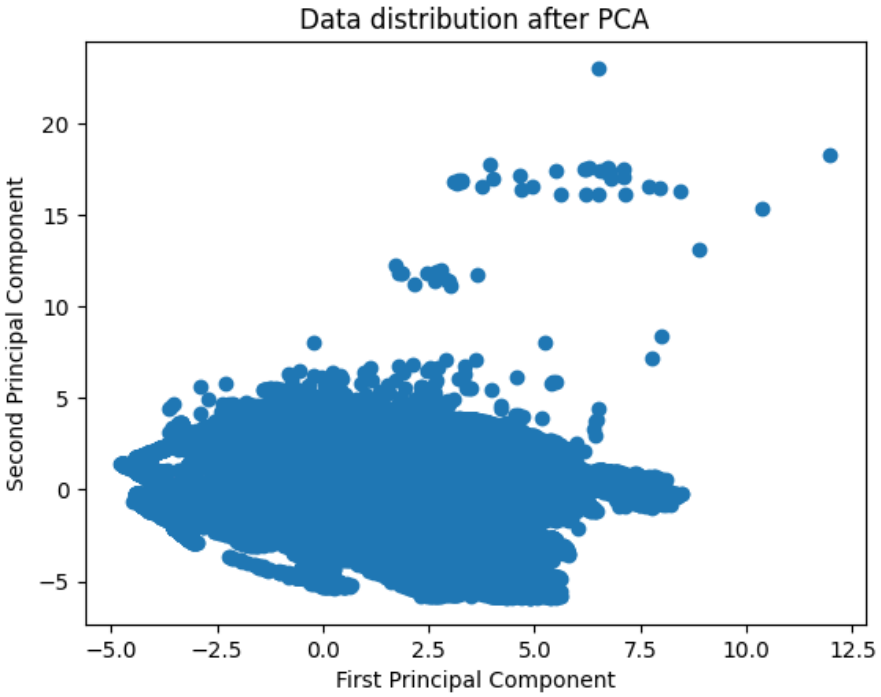

Figure 18 - Using PCA model for Non-residential building (Cold season)

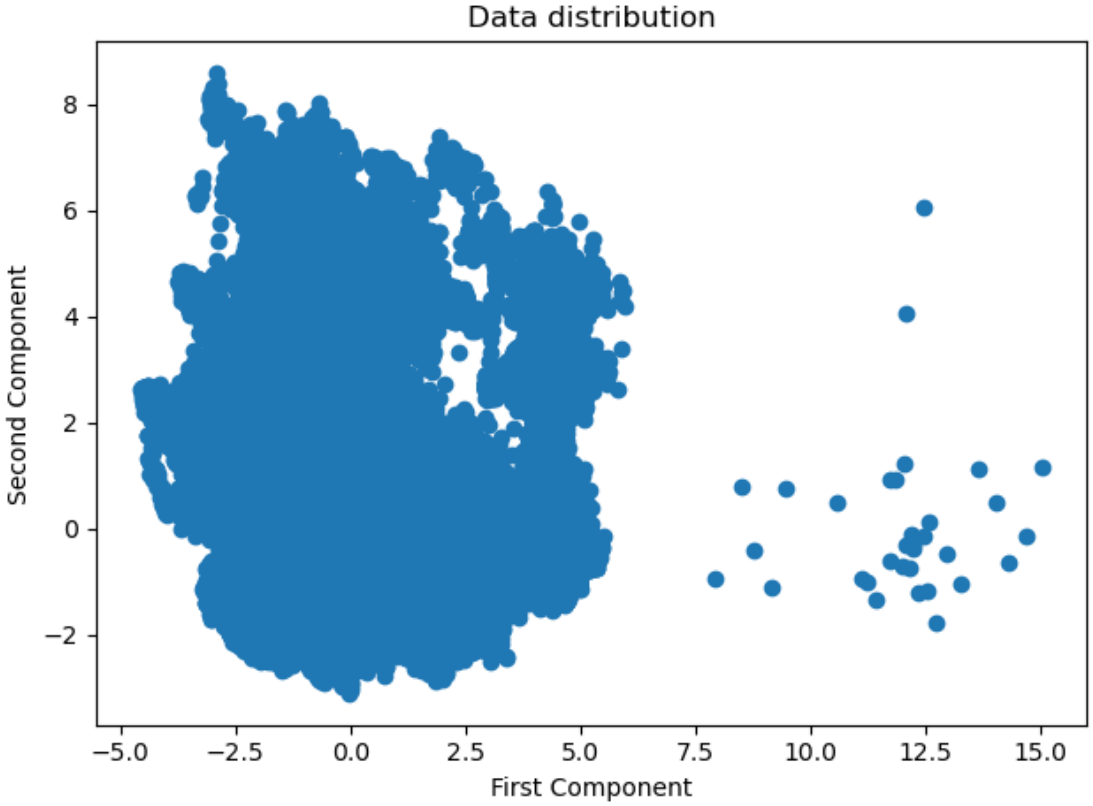

Figure 19 – Using PCA for Non-Residential building (Hot season)

7.1 PCA explained variance

PCA explained variance indicates the proportion of the dataset's total variance that is captured by each principal component. It helps determine how many components are needed to retain most of the information while reducing dimensionality with minimal loss.

Analysis and Interpretation

PCA for residential and non-residential buildings show how the main components explain the variability of the data in each dataset. Let's take a closer look at these results and their interpretation.

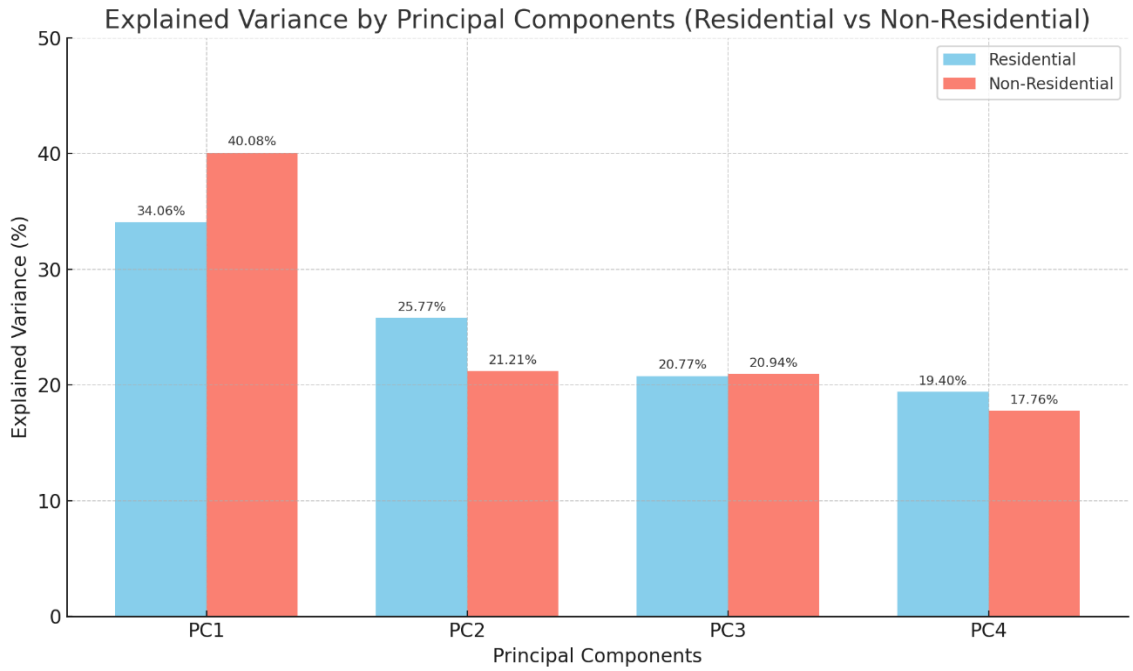

Figure 20 – Explained Variance Ratio of Principal Components for Residential and Non-Residential Buildings

The PCA revealed that the first four principal components account for 100% of the total variance in both residential and non-residential building datasets, indicating a highly effective dimensionality reduction with no information loss.

In residential buildings, the first two components explain more than 50% of the total variability, while in non-residential buildings, they account more than 60%. This suggests that the data structure can be reliably represented in a two-dimensional space without significant loss of information. Notably, the first principal component in non-residential buildings explains a larger portion of the variance (40.08%), indicating a dominant influence – likely due to external environmental conditions. In contrast, the residential data shows a more balanced variance distribution across components, reflecting more complex and interconnected relationships among indoor parameters (Figure 20).

Overall, PCA successfully highlighted the key structural differences between the datasets and confirmed the feasibility of feature dimensionality reduction for subsequent modeling and visualization tasks.

7.2 Clustering Results by using K-means and DBSCAN methods

Clustering techniques were applied to the datasets to identify inherent groupings within the data. The results provide insights into how different points in the dataset are grouped together and how these clusters correspond to the characteristics of the underlying data.

The study of clustering methods for identifying errors in microclimate parameters using machine learning algorithms. Specifically, the clustering algorithm was applied to cluster data on the microclimate in Residential building during the hot and cold seasons, which is one of the most commonly used clustering methods (Figures 21, 22) and cold season (Figures 23, 24), in Non-residential building, the clusters were unified during the cold (Figures 25, 26) and hot seasons (Figures 27, 28).

The K-means algorithm operates on the principle of minimizing within-cluster variance and maximizing between-cluster variance. The clustering process involves several steps:

1) Initialization – the number of clusters *k* is specified, and initial cluster centroids are randomly chosen.
2) Assignment – each data point is assigned to the cluster whose centroid is closest.
3) Update – the center of each cluster is recalculated as the mean of all points assigned to that cluster.
4) Iteration – steps 2 and 3 are repeated until the centroids stabilize or the changes become negligible.

However, when applying the K-means method to the microclimate data, clusters were obtained for both individual and combined groups of spaces during the warm and cold seasons. The algorithm did not effectively identify outliers, which limits its applicability in this task, where anomaly detection is crucial. This highlights the need for more suitable clustering methods, such as DBSCAN, which can more effectively identify outliers and provide more accurate data segmentation.

Clustering Results for Residential building

For the Residential building dataset, a clustering algorithm formed two clusters, with the following distribution:

Cluster 0: 185,172 points

Cluster 1: 28,790 points

Out of Clusters: 4 points

The clustering results indicate a clear distinction between two major groups within the Residential building dataset. The first cluster (Cluster 0) contains the majority of the data points, while the second cluster (Cluster 1) represents a smaller, but still significant, portion. This clear division may reflect differences in certain characteristics between the two groups (Figures 23, 25 Residential building for cold and hot seasons).

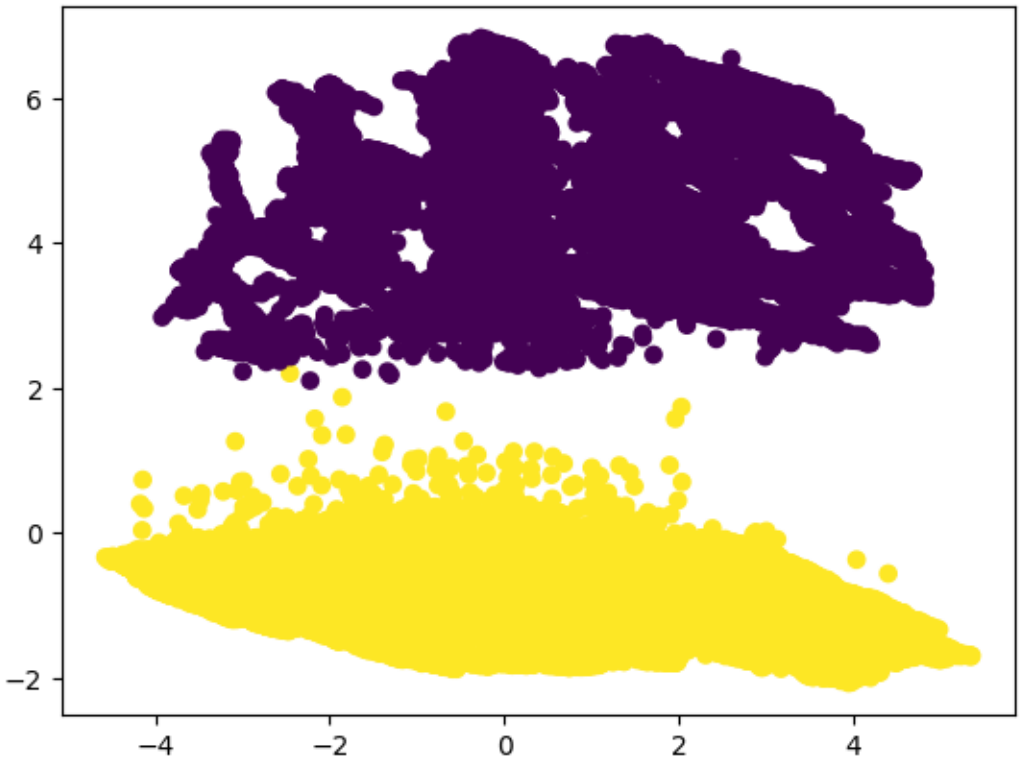

Figure 21 - Using K-means for Residential building (Cold season)

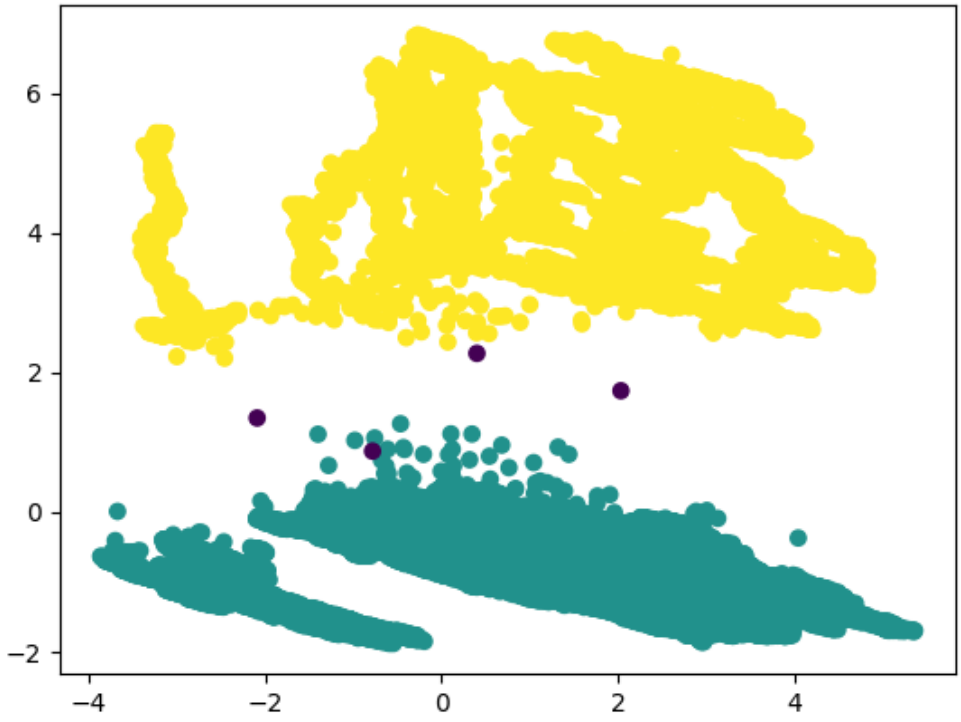

Figure 22 - Using DBSCAN for Residential building (Cold season)

Table 17 – Residential and Non-Residential buildings result after using DBSCAN

Parameter	Residential (Mean / SD / Min / Max)	Non-residential (Mean / SD / Min / Max)	DBSCAN Result
T (°C)	27.64/164.81/25.49/ 45652.00	27.63/164.51/25.43/45652.00	Outliers detected (extreme Max)
Tout (°C)	24.51 / 13.82 / 0.0 / 64.66	24.48 / 13.81 / 0.0 / 64.66	No anomaly
H (%)	32.77 / 2.60 / 26.04 / 40.49	32.70 / 2.61 / 26.04 / 40.49	Stable clusters
Hout (%)	35.62 / 20.39 / 0.0 / 82.69	35.60 / 20.37 / 0.0 / 82.69	Stable clusters
DP (°C)	9.22 / 1.18 / 5.70 / 12.86	9.20 / 1.20 / 5.56 / 12.86	No anomaly
P (hPa)	687.37 / 1.52 / 684.18 / 692.03	687.37 / 1.51 / 684.18 / 692.03	Normal density
Pout (hPa)	666.42 / 119.52 / 0.0 / 693.12	666.48 / 119.36 / 0.0 / 693.12	Some outliers (Min = 0.0)
CO2	0.0 / 0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0 / 0.0	No data / sensor off
TVOC	0.0 / 0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0 / 0.0	No data / sensor off
V (V)	227.89 / 31.58 / 0.0 / 0.54	227.92 / 31.55 / 0.0 / 0.54	Outliers (Min = 0.0)
C (A)	0.066 / 0.144 / 0.0 / 96.80	0.0657 / 0.1434 / 0.0 / 96.80	Wide range, but dense
Pwr (W)	12.50 / 26.89 / 0.0 / 96.80	12.46 / 26.85 / 0.0 / 96.80	Reflects usage pattern
UVr	9.19 / 1.65 / 0.0 / 9.60	9.19 / 1.65 / 0.0 / 9.60	No anomaly
L	10548.45/17980.22 / 0.0 / 54612.0	10541.23 / 17948.36 / 0.0 / 54612.0	Outliers (High variance)

The anomalies detected by the DBSCAN method include an extreme temperature value (Max > 45,000°C), indicating a measurement error, the presence of a zero in the minimum value of outdoor pressure (Pout), which is an anomaly, a zero in the minimum value of voltage (V), which is atypical under normal operation, and high variability and extremes in illuminance (L). Note: CO2 and TVOC are missing from all records, so DBSCAN did not analyze them. Anomalies were identified as outliers not belonging to high-density clusters.

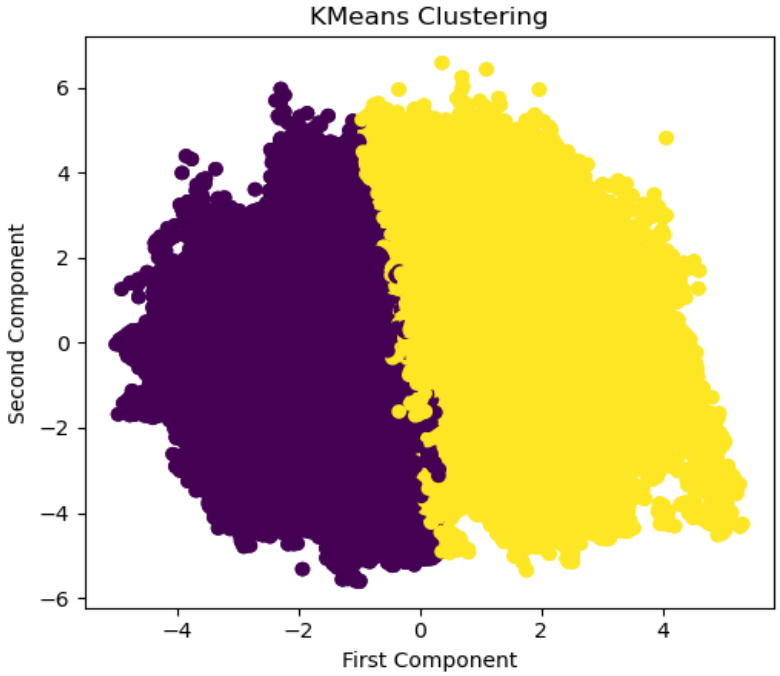

Figure 23 - Using K-means for Residential building (Hot season)

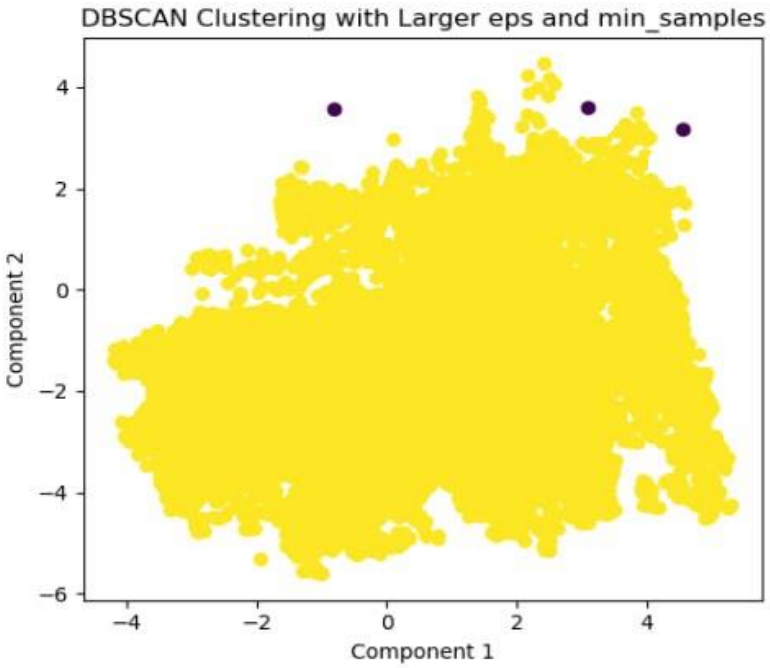

Figure 24 - Using DBSCAN for Residential building (Hot season)

Number of clusters formed: 1
Number of points in cluster 0: 76,643
Number of points in cluster -1: 3
Number of points outside clusters: 3

Clustering Results for Non-Residential building
For the Non-Residential building dataset, the clustering algorithm identified three distinct clusters, with the following distribution:
Cluster 0: 952 points
Cluster 1: 194,842 points
Cluster 2: 29,363 points
Out of Clusters: 9 points

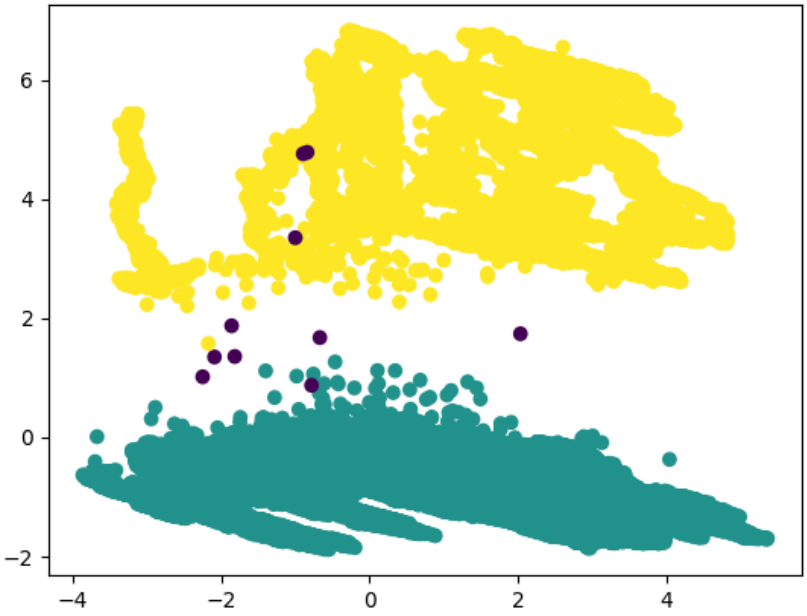

Figure 25 - Using DBSCAN for Non-Residential building (Cold season)

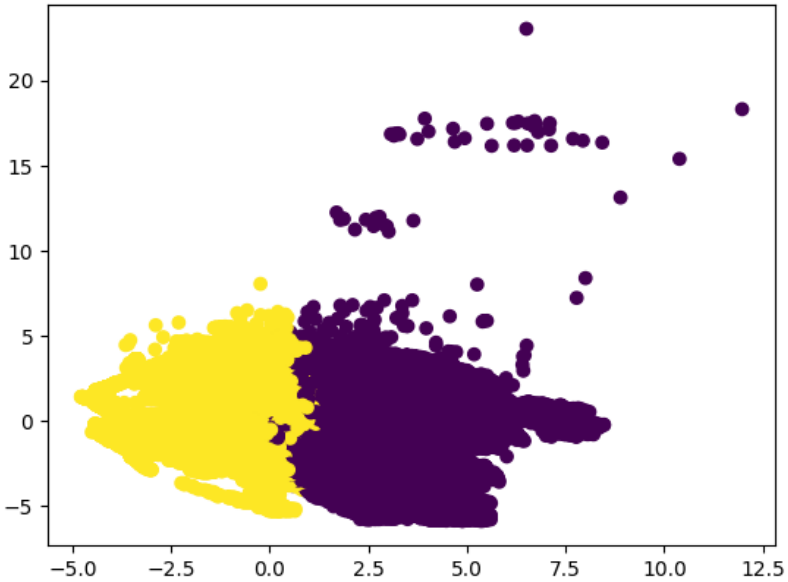

Figure 26 - Using K-means for Non-Residential building (Cold season)

These results show that the Non-Residential building data has a more complex clustering structure than the Residential building dataset, with three main groups forming distinct clusters. Cluster 1 contains the overwhelming majority of data points, while Clusters 0 and 2 represent much smaller groups. This further highlights the variation and complexity inherent in the Non-Residential building dataset (Figures 26, 28 Non-residential for hot and cold seasons).

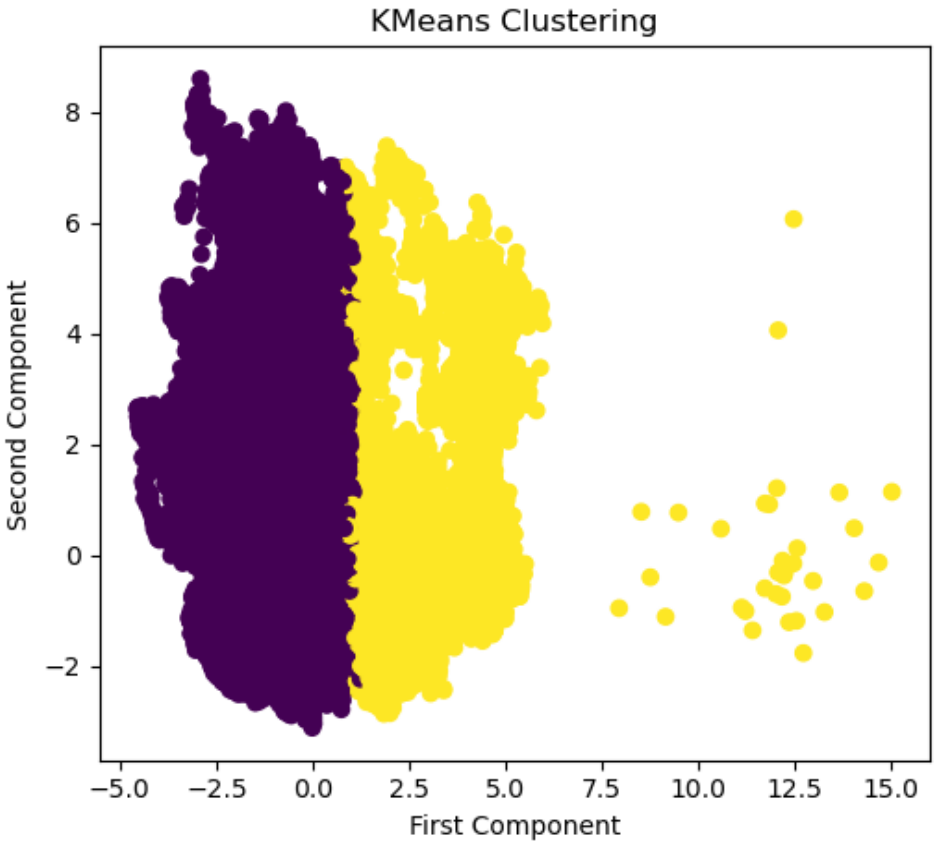

Figure 27 – Using K-means for Non-Residential building (Hot season)

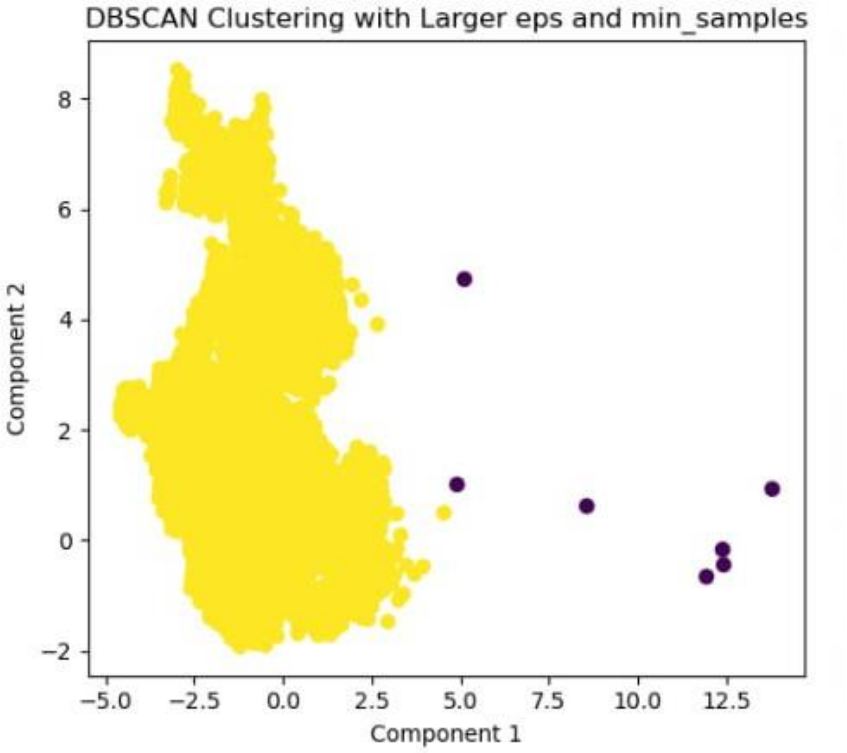

Figure 28 – Using DBSCAN for Non-Residential building (Hot season)

Number of clusters formed: 1
Number of points in cluster 0: 76915

Number of points in cluster -1: 7
Number of points outside clusters: 7

In the context of applying the DBSCAN algorithm, anomalies were identified as outliers that do not conform to the main data structure. These outliers include extreme values, such as a temperature exceeding 45,000°C, indicating a possible measurement error; zero values for pressure and voltage, which are inconsistent with the operation of microclimate systems; and high variability and extreme values in illumination, suggesting instability in the measurements. These outliers do not belong to clusters with high density and were therefore classified as anomalies, requiring further analysis and data correction [106].

7.3 Summary

This chapter presents the results of applying PCA and clustering methods (K-means and DBSCAN) to microclimate data from residential and non-residential buildings. PCA demonstrated high effectiveness in reducing data dimensionality while preserving 100% of the variance, with the first principal component explaining the largest portion of the data's variability. Clustering using K-means identified data groups, but the algorithm did not detect outliers, limiting its application for anomaly detection tasks. In contrast, DBSCAN successfully identified anomalies, such as extreme temperature and pressure values, and high variability in certain parameters. These results confirm the need to use DBSCAN for accurate data segmentation and anomaly detection in microclimate systems.

DISCUSSION

This thesis focuses on the application of machine learning methods for fault diagnosis in building microclimate systems, with an emphasis on improving energy efficiency and reliability. The DBSCAN algorithm was chosen for data clustering and anomaly detection, enabling effective identification of faults in heating, ventilation, HVAC systems based on unlabeled sensor data. The application of this algorithm in real-world building operations demonstrated its high effectiveness in identifying anomalies and normal states of the system, despite the lack of formal accuracy metrics.

In this study, the author focuses on enhancing the reliability and energy efficiency of microclimate systems using DBSCAN, while in the work of Wang et al., particular attention is given to adjusting the false alarm rate and improving the fault diagnosis of chillers for more effective management of heating, ventilation, and air conditioning systems.

A significant aspect of the research was testing various values of the eps parameter for the DBSCAN algorithm. The experiments showed that the correct choice of this parameter is crucial for obtaining accurate clustering results, as it determines the model's sensitivity to anomalies and its ability to exclude noise data. The experiments revealed that the selection of eps must be tailored to the specific dataset, as different data types require different approaches for detecting patterns and deviations.

To enhance the quality of clustering and improve the interpretability of results, data preprocessing was performed using the Z-score method to remove outliers. This step proved to be vital, as removing outliers significantly improved the clustering results and enhanced the accuracy of the analysis. The Z-score method provided cleaner and more understandable data for subsequent analysis and visualization.

Additionally, to reduce data dimensionality and improve cluster visualization, PCA was applied. This approach greatly simplified the interpretation of results and made them more visual, which is crucial when dealing with unlabeled data and large volumes of information. The use of PCA provided improved cluster visualization, facilitating more accurate fault diagnosis and accelerating the analysis process. PCA proved effective for both residential and non-residential buildings, with the first four components capturing 100% of the variance. In residential buildings, the first two components explain over half of the variability, while in non-residential ones, they cover more than 60%. This means the data can be reliably visualized in two dimensions. The higher influence of the first component in non-residential buildings suggests stronger effects from external conditions, whereas the more even distribution in residential data points to more complex indoor relationships.

One of the key findings of this work is the high practical significance of the proposed methodology. The application of DBSCAN and preprocessing methods (Z-score, PCA) not only achieved meaningful results in diagnosing faults in microclimate systems but also demonstrated the potential of machine learning methods to improve building energy efficiency. This approach allows for the rapid detection of system anomalies, which can significantly reduce energy consumption

and enhance the reliability of operational processes. Methodology for detecting faults in microclimate systems: statistical approaches and machine learning methods shows in Figure 29.

The CRISP-DM methodology used in this study proved to be an effective tool for structuring the process of data collection, preparation, analysis, and machine learning model implementation. A notable aspect is that the proposed methods can be utilized in the early stages of analysis, even when precise labeled data are not yet available, making the DBSCAN algorithm a useful tool for preliminary diagnosis and monitoring.

However, it is important to note that the chosen approach requires further research to improve clustering methods and adapt them to various types of microclimate systems. Specifically, future studies should focus on enhancing the robustness of the DBSCAN algorithm to noise and optimizing its parameters for use with different types of sensor data.

Thus, the results of this research confirm the effectiveness of using machine learning methods, particularly the DBSCAN algorithm, for monitoring and diagnosing faults in building microclimate systems. The findings can serve as a foundation for further research and the development of intelligent microclimate control systems aimed at improving energy efficiency and the reliability of buildings.

Experiments were conducted on the data to adjust the eps parameter in the DBSCAN algorithm, including testing various values of this parameter and evaluating clustering metrics. This approach allowed for the selection of the optimal eps value, which provided the best clustering results, effectively identifying anomalous points and correctly separating the data into clusters.

Testing different eps values demonstrated that the proper selection of this parameter is critical for achieving a balanced model sensitivity, ensuring high accuracy in anomaly detection while minimizing misclassification of noise. At the same time, this process required additional computational resources and, in some cases, proved to be sensitive to the characteristics of the data.

Thus, experiments with eps parameter tuning and clustering metric evaluation demonstrated the effectiveness of the DBSCAN algorithm for fault diagnosis in microclimate systems. However, for different data sets, additional parameter tuning and validation are required to achieve optimal model performance.

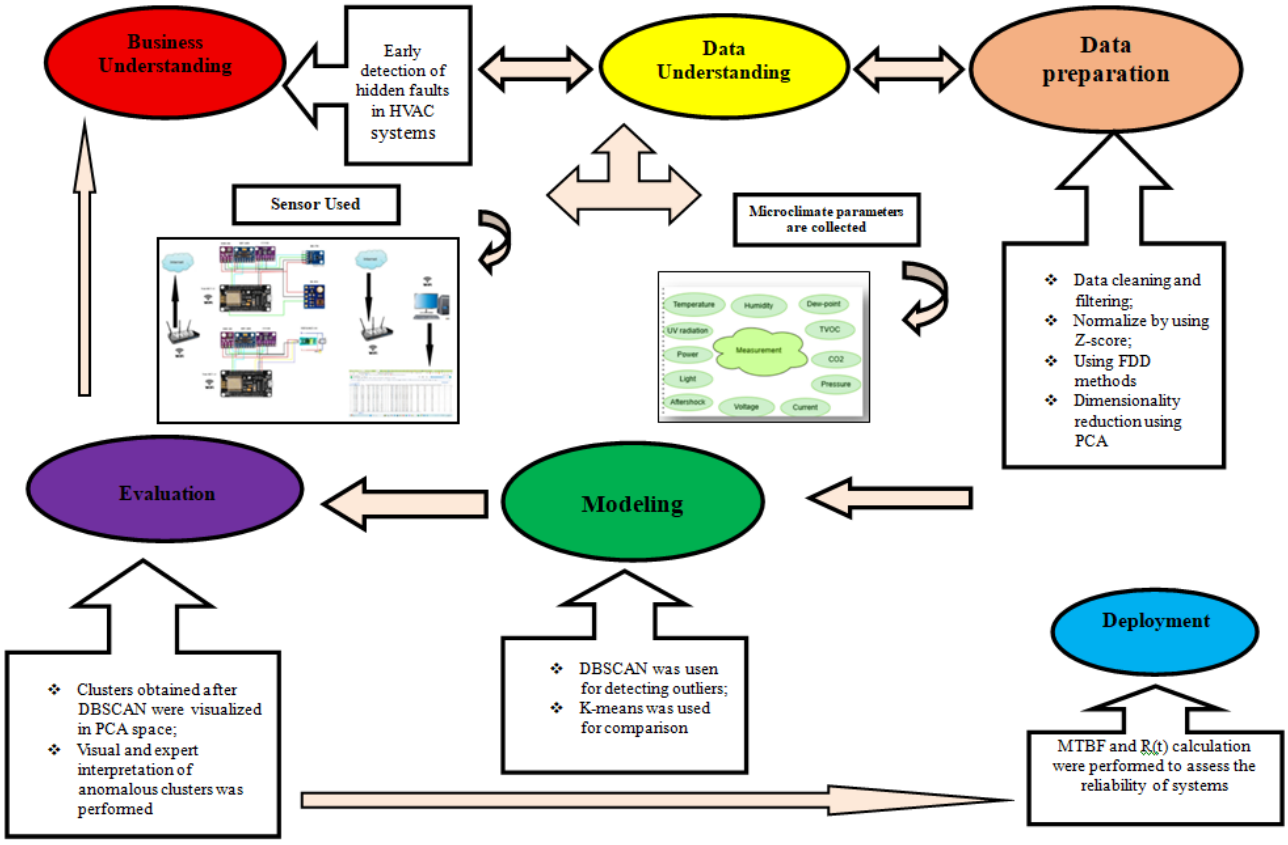

Figure 29 - Unsupervised Fault Detection Process (CRISP-DM Adaptation)

Main Scientific Contributions

This dissertation produced several important contributions that demonstrate the applicability of machine learning methods for fault detection in building microclimate systems. One of the key achievements was the development of an experimental data acquisition system based on custom-assembled sensors, designed in accordance with international standards such as those of the IEA and ASHRAE. This system was deployed in both residential and non-residential buildings, ensuring a diverse and realistic dataset for analysis.

The research followed the CRISP-DM methodology, which provided a structured framework from problem definition and data collection to model development and evaluation. A central focus was placed on the DBSCAN clustering algorithm, which proved to be highly effective in detecting system faults under conditions where labeled training data were unavailable. A detailed analysis of DBSCAN's core parameters, particularly eps and min_samples, confirmed their critical role in the algorithm's ability to accurately identify anomalous behavior and system irregularities.

In addition, data preprocessing played a vital role in improving the quality of the results. Z-score normalization was used to remove outliers, while PCA was applied to reduce data dimensionality and enhance the interpretability of clustering outcomes. These steps significantly improved diagnostic accuracy and facilitated clearer visualization of system states.

PCA showed that the first four components explain 100% of the variance in both residential and non-residential building data, enabling effective dimensionality reduction without information loss. In non-residential buildings, the first two components account for over 60% of the variance, with the first component explaining 40.08%, indicating a dominant factor likely related to external conditions. In residential buildings, the variance is more evenly distributed, reflecting more complex and interconnected indoor parameters.

The research also demonstrated the feasibility of implementing a low-cost, practical monitoring solution based on affordable sensors and open-source software. This makes the approach accessible for real-world deployment in a wide range of building environments.

To evaluate the operational stability of HVAC systems in residential and non-residential buildings, we calculated two key reliability metrics:

Mean Time Between Failures (MTBF)

Reliability function R(t) - the probability that the system operates without failure for a given time t.

Step 1: Fault Detection

Outlier events were detected using DBSCAN clustering. Each detected group of anomalies is considered a «failure»event. The number of outliers per building type and season:

Residential (Cold Season): 4 faults

Residential (Hot Season): 3 faults

Non-Residential (Cold Season): 9 faults

Non-Residential (Hot Season): 7 faults

Total faults:

N_{res} = 7 (residential)

$N_{non\text{-}res}$ = 16 (non-residential)

Step 2: Mean Time Between Failures (MTBF)

The MTBF is defined as:

$MTBF = T_{total} / N$

Where:

MTBF is the Mean Time Between Failures

Ttotal is the total observation time (in hours): 30 days × 24 = 720 hours

N is the total number of detected failure events

Calculations:

$MTBF_{res} = 720 \text{ h} / 7 \approx 102.9$ hours

$MTBF_{non\text{-}res} = 720 \text{ h} / 16 = 45$ hours

Step 3: Reliability Estimation

The reliability function R(t) assumes an exponential failure distribution:

$R(t) = e^{-t/MTBF}$

Where:

R(t) is the probability the system operates without failure for time t

t is the duration of interest (e.g., 24 hours)

Results for t = 24 hours:

$R_{res}(24) = e^{-24/102.86} \approx e{-}0.2333 \approx 0.792$ (или 79.2%)

$R_{non\text{-}res}(24) = e^{-24/45} \approx e{-}0.5333 \approx 0.586$ (или 58.6%)

To sum up residential systems demonstrate higher reliability (MTBF ≈ 103 h, R24 ≈ 79.2%) compared to non-residential systems (MTBF ≈ 45 h, R24 ≈ 58.6%). This reflects the more consistent and stable operation of residential HVAC systems, as opposed to the intermittent use in non-residential contexts. These findings further support the relevance of outlier detection and data-driven reliability analysis in microclimate system monitoring.

Based on the calculated MTBF and reliability function R(t), the following conclusion can be drawn:

Microclimate systems in residential buildings demonstrate higher reliability compared to non-residential ones. The Mean Time Between Failures (MTBF) for residential buildings is approximately 103 hours, while for non-residential buildings it is only 45 hours. This indicates that faults and anomalies occur significantly less frequently in residential settings. The probability of failure-free operation over a 24-hour period is also higher in residential buildings (~79.2% vs. ~58.6%).

Possible reasons include:

- More stable and continuous operation (24/7) in residential environments;
- Less exposure to abrupt external changes and operating cycles;
- A higher number of irregular peaks and outliers in non-residential buildings during working hours, increasing the likelihood of detected anomalies.

To enhance the reliability of HVAC systems in non-residential buildings, early fault detection and adaptive diagnostic mechanisms should be implemented, particularly during periods of fluctuating loads.

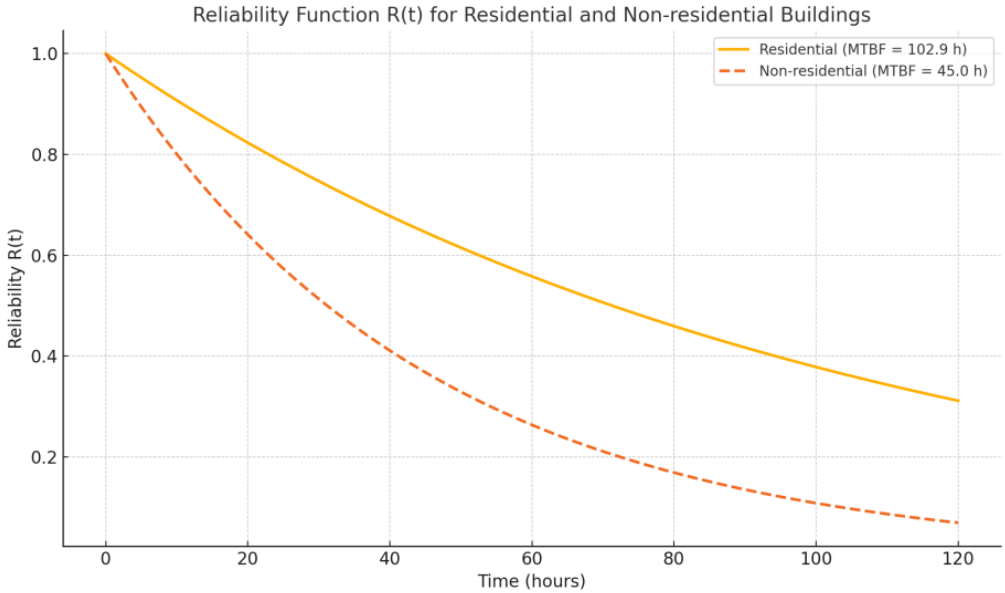

Figure 30 – Reliability Function R(t) for Residential and Non-residential Buildings

The results of this research confirm that unsupervised machine learning techniques - particularly clustering and dimensionality reduction - can be effectively applied to detect faults in building microclimate systems. The proposed methodology

is especially well-suited for real-world scenarios where labeled data are limited or unavailable, enabling early detection of abnormal system behavior without the need for complex or expensive infrastructure.

From a practical standpoint, the approach supports improvements in energy efficiency, operational reliability, and system safety. The use of open-source tools and affordable hardware makes the solution scalable and adaptable to different types of buildings and climate control systems.

Overall, this dissertation contributes to the advancement of intelligent building management and offers a foundation for future research in data-driven diagnostics.

CONCLUSION

This thesis examined the applicability of unsupervised machine learning techniques - particularly clustering algorithms - for detecting faults in building microclimate systems. The motivation behind the study was to enhance the reliability, energy efficiency, and safety of HVAC systems, especially in real-world scenarios where labeled data are rarely available.

To support this, a custom microclimate data collection system was developed using affordable sensors, designed in accordance with international standards such as those from the IEA and ASHRAE. The system was deployed in both residential and non-residential buildings, providing a robust and diverse dataset for analysis. The research followed the CRISP-DM methodology, ensuring a structured workflow from data acquisition through preprocessing, modeling, and evaluation.

A key focus of the study was the DBSCAN clustering algorithm. Through systematic tuning of its parameters - particularly *eps* and *min_samples*- the algorithm proved capable of detecting anomalies and hidden faults in unlabeled datasets. Visual inspection and expert validation confirmed that the model could reliably distinguish between normal and abnormal system behavior.

Data preprocessing, including outlier removal via Z-score normalization and dimensionality reduction through PCA, significantly improved the clarity and interpretability of clustering results. These steps enhanced the model's effectiveness in identifying performance deviations in HVAC systems.

All research objectives were achieved, and the findings confirmed that unsupervised learning methods can provide a strong foundation for intelligent, data-driven fault diagnostics. Beyond academic contributions, the proposed approach offers practical value - it enables the development of low-cost, scalable monitoring solutions using open-source tools and widely available sensor technologies.

In summary, this work contributes to the advancement of intelligent building automation by demonstrating a viable and adaptable methodology for early fault detection, which can lead to improved energy performance and safer building operation.

REFERENCES

1 Wang J., Tian Y., Qi Z., Zeng L., Wang P., Yoon S. Sensor fault diagnosis and correction for data center cooling system using hybrid multi-label Random Forest and Bayesian inference // Building and Environment. – 2023. – Vol. 249. – Article No. 111124. – DOI: 10.1016/j.buildenv.2023.111124.

2 Zhao T., Zhang B., Li M., Liu G., Wang P. Handling fault detection and diagnosis in incomplete sensor measurements for BAS-based HVAC system // Journal of Building Engineering. – 2023. – Vol. 80. – Article No. 108098. – DOI: 10.1016/j.jobe.2023.108098.

3 Li G., Wang C., Liu L., Fang X., Kuang W., Xiong C. Study on sensor fault-tolerant control for central air-conditioning systems using Bayesian inference with data increments // Sensors. – 2024. – Vol. 24, No. 4. – Article No. 1150. – DOI: 10.3390/s24041150.

4 Zhang B., Rezgui Y., Luo Z., Zhao T. Fault detection research on novel transfer learning-based method for cross-condition, cross-system and cross-operation in public building HVAC sensors // Energy. – 2024. – Vol. 313. – Article No. 133704. – DOI: 10.1016/j.energy.2024.133704.

5 Tokayev, K.K. Kazakhstan in the Era of Artificial Intelligence: Current Challenges and Solutions through Digital Transformation // Address by the President of the Republic of Kazakhstan to the People of Kazakhstan. 08.09.2025. – Astana: Official website of the President of Kazakhstan. https://www.akorda.kz/en/. 01.10.2025.

6 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, Vol.16, No.8. -P. 21. – DOI: https://doi.org/10.3390/en16083508.

7 Ganzhur M., Ganzhur A., Kobylko A., Fathi D. Automation of microclimate in greenhouses // E3S Web of Conferences. – 2020. – Vol. 210. – P. 1–6. – Article No. 05004. – DOI: https://doi.org/10.1051/e3sconf/202021005004.

8 Mukazhanov Y., Kamshat Z., Orazbayeva A., Shayhmetov N., Alimbaev C. Microclimate Control in Greenhouses // Proceedings of the 17th International Multidisciplinary Scientific GeoConference SGEM. – Vienna, Austria, 2017. – Vol. 17, No. 62. – P. 699–704. – DOI: https://doi.org/10.5593/sgem2017/62/S27.089.

9 Житов В. Г. Исследование и обеспечение параметров микроклимата жилых и общественных зданий методами оптимального планирования эксперимента: диссертация кандидата технических наук. – Иркутский государственный технический университет. – Иркутск, Россия, 2007. – 180 с. – DOI: 61:07-5/3220.

10 Cannistraro G., Bernardo E. Monitoring of the indoor microclimate in hospital environments: A case study of the Papardo hospital in Messina // International Journal of Heat and Technology. – 2017. – Vol. 35, Suppl. 1. – P. S456–S465. – DOI: https://doi.org/10.18280/ijht.35Sp0162.

11 Hoxha A., Dervishi M. G., Bici M. E. Evaluation of microclimate in regional hospital in Berat // IOSR Journal of Dental and Medical Sciences. – 2014. – Vol. 13, No. 1. – P. 96–101.

12 Fabbri K., Gaspari J., Vandi L. Indoor Thermal Comfort of Pregnant Women in Hospital: A Case Study Evidence // Sustainability. – 2019. – Vol. 11, No. 23. – P. 1–14. – Article No. 6664. – DOI: https://doi.org/10.3390/su11236664.

13 Czarniecki W., Kopacz M., Okołowicz W., Gajewski J., Grzędziński E. Investigations of the microclimate in hospital wards // Energy and Buildings. – 1991. – Vol. 16, No. 5. – P. 727–733. – DOI: https://doi.org/10.1016/0378-7788(91)90044-4.

14 Daurenbayeva N., Atymtayeva L., Nurlanuly A. Choosing the intelligent thermostats for the effective decision making in BEMS // Proceedings of the 17th International Conference on Electronics, Computer and Computation (ICECCO-2023). – 2023. – P. 1–4. – DOI: https://doi.org/10.1109/ICECCO58239.2023.10147131.

15 Costa C. J., Aparicio J. T. The evolution of CRISP-DM for Data Science: Methods, Processes and Frameworks // SBC Reviews on Computer Science. – 2020.

16 Studer S., Bui T. B., Drescher C., Hanuschkin A., Winkler L., Peters S., Müller K.-R. Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology. – 2020. – DOI: https://doi.org/10.48550/arXiv.2003.05155.

17 Owuor D. O., Runkler T., Laurent A., Orero J., Menya E. Ant Colony Optimization for Mining Gradual Patterns. – 2022. – DOI: https://doi.org/10.48550/arXiv.2208.14795.

18 Ayele W. Y. Adapting CRISP-DM for Idea Mining: A Data Mining Process for Generating Ideas Using a Textual Dataset. – 2021. – DOI: https://doi.org/10.48550/arXiv.2105.00574.

19 Plotnikova V., Dumas M., Milani F. P. Adaptations of data mining methodologies: a systematic literature review // PeerJ Computer Science. – 2021. – Vol. 6. – Article ID: e267. – DOI: 10.7717/peerj-cs.267.

20 Molina-Coronado B., Mori U., Mendiburu A., Miguel-Alonso J. Survey of Network Intrusion Detection Methods from the Perspective of the Knowledge Discovery in Databases Process. – 2020. – DOI: https://doi.org/10.48550/arXiv.2001.09697.

21 Cornellius Yudha Wijaya. CRISP-DM Methodology For Your First Data Science Project. TDS Archive, Medium. https://medium.com/data-science.17.10.2023

22 GeeksforGeeks. -URL: https://www.geeksforgeeks.org/dbms/kdd-process-in-data-mining/.05.10.2024.

23 Dineva, Kristina & Atanasova, Tatiana. (2018). OSEMN process for working over data acquired by iot devices mounted in beehives.

24 Qiang G., Tang S., Hao J., Di Sarno L., Wu G., Ren S. Building automation systems for energy and comfort management in green buildings: A critical review and future directions // Renewable and Sustainable Energy Reviews. – 2023. – Vol. 179. – Article No. 113301.

25 Ahsan M., Shahzad W., Arif K. M. AI-Based Controls for Thermal Comfort in Adaptable Buildings: A Review // Buildings. – 2024. – Vol. 14, No. 11. – P. 3519.

26 Bian Y., Fu X., Gupta R. K., Shi Y. Ventilation and Temperature Control for Energy-efficient and Healthy Buildings: A Differentiable PDE Approach. – 2024. – DOI: https://doi.org/10.48550/arXiv.2403.08996

27 Park Y.J., Fan S.K., Hsu C.Y. A review on fault detection and process diagnostics in industrial processes // Processes. – 2020. – Vol. 8, No. 9. – Article 1123. – DOI: https://doi.org/10.3390/pr8091123.

28 Lau M., Liu Y., Yu Y. On detection conditions of double faults related to terms in Boolean expressions // Annual International Computer Software and Applications Conference. – 2006. – Vol. 1. – P. 403-410. – DOI: https://doi.org/10.1109/COMPSAC.2006.82.

29 Hyvärinen J., Kärki S. IEA Annex 25. Real time simulation of HVAC systems for building optimization, fault detection and diagnosis: Building optimization and fault diagnosis source book. Technical report. – Espoo, Finland: VTT Building Technology, 1996. – 96 p.

30 Abd-Alkader M. Y., Ebid A. M., Mahdi I., Abdelrashed Nosseir I. Application of using fault detection techniques in different components in power systems // Future Engineering Journal. – 2021. – Vol. 2, No. 2. – P. 24–40. – DOI: 10.54623/fue.fej.2.2.4.

31 Xiangjun Z., Yuanyuan W., Yao X. Faults Detection for Power Systems // In: Zhang W. (Ed.) Fault Detection. – Rijeka, Croatia: IntechOpen, 2010. – Chapter 4. – DOI: https://doi.org/10.5772/10133.

32 Merabet G. H., Essaaidi M., Ben Haddou M., Qolomany B., Qadir J., Anan M., Al-Fuqaha A., Abid M. R., Benhaddou D. Intelligent building control systems for thermal comfort and energy-efficiency: A systematic review of artificial intelligence-assisted techniques // Renewable and Sustainable Energy Reviews. – 2021. – Vol. 135. – Article No. 110185.

33 Matetic, I., Štajduhar, I., Wolf, I., Palaic, D., Ljubic, S. Random Forests Model for HVAC System Fault Detection in Hotel Buildings // Advances in Computational Intelligence. Lecture Notes in Computer Science. – 2023. – Vol. 14308. – P. 651- 662. – DOI: 10.1007/978-3-031-43085-5_52.

34 Panda, R. R., Gouda, B. S., Panigrahi, T. Efficient fault node detection algorithm for wireless sensor networks // In Proceedings of the 2014 International Conference on High Performance Computing and Applications (ICHPCA). – India: Bhubaneswar, 2014, (22–24 December) – P. 1–5. – DOI: https://doi.org/10.1109/ICHPCA.2014.7045308.

35 Panda, M., Khilar, P. M. Distributed Byzantine fault detection technique in wireless sensor networks based on hypothesis testing. Comput. Electr. Eng. – 2015 – 48:270-285. – DOI: https://doi.org/10.1016/j.compeleceng.2015.06.024.

36 Yu, T., Akhtar, A. M., Wang, X., Shami, A. Temporal and spatial correlation based distributed fault detection in wireless sensor networks // In Proceedings of the 2015 IEEE 28th Canadian Conference on Electrical and Computer Engineering (CCECE). – Halifax, NS, Canada, 2015, (3–6 May). – P. 1351–1355. – DOI: https://doi.org/10.1109/CCECE.2015.7129475.

37 Dey M., Rana S. P., Dudley S. A case study-based approach for remote fault detection using multi-level machine learning in a smart building // Smart Cities. – 2020. – Vol. 3, No. 2. – P. 401–419. – DOI: 10.3390/smartcities3020022.

38 Miljković D. Fault detection methods: A literature survey // Proceedings of the 34th International Convention MIPRO. – Opatija, Croatia, 2011 (23–27 May). – P. 750–755. – IEEE. – DOI: 10.23919/MIPRO.2011.5967078.

39 Aleem S. A., Shahid N., Naqvi I. H. Methodologies in power systems fault detection and diagnosis // Energy Systems. – 2015. – Vol. 6, No. 1. – P. 85–108. – DOI: 10.1007/s12667-014-0129-1.

40 Durán C., Sanjuan M. On-Line early fault detection of a centrifugal chiller based on data-driven approach // Proceedings of the ASME 2016 Energy Sustainability Conference. – 2016. – Vol. 1. – Article No. V001T11A009. – DOI: 10.1115/ES2016-59291.

41 Gao Y., Liu S., Li F., Liu Z. Fault detection and diagnosis method for cooling dehumidifier based on LS-SVM NARX model // International Journal of Refrigeration. – 2015. – Vol. 61. – P. 10–20. – DOI: 10.1016/j.ijrefrig.2015.08.020.

42 Hu Y., Li G., Chen H., Li H., Liu J. Sensitivity analysis for PCA-based chiller sensor fault detection // International Journal of Refrigeration. – 2015. – Vol. 63. – P. 133–145. – DOI: 10.1016/j.ijrefrig.2015.11.006.

43 Li D., Hu G., Spanos C. A data-driven strategy for detection and diagnosis of building chiller faults using linear discriminant analysis // Energy and Buildings. – 2016. – Vol. 128. – P. 519–529. – DOI: 10.1016/j.enbuild.2016.07.010.

44 Yan R., Ma Z., Kokogiannakis G., Zhao Y. A sensor fault detection strategy for air-handling units using cluster analysis // Automation in Construction. – 2016. – Vol. 70. – P. 77–88. – DOI: 10.1016/j.autcon.2016.06.005.

45 Montazeri A., Kargar M. Fault detection and diagnosis in air handling using data-driven methods // Journal of Building Engineering. – 2020. – Vol. 31. – Article No. 101388. – DOI: 10.1016/j.jobe.2020.101388.

46 Yan K., Huang J., Shen W., Ji Z. Unsupervised learning for fault detection and diagnosis of air-handling units // Energy and Buildings. – 2019. – Article No. 109689. – DOI: 10.1016/j.enbuild.2019.109689.

47 Liu J., Shi D., Li G., Xie Y., Li K., Liu B., Ru Z. Data-driven and association rule mining-based fault diagnosis and action mechanism analysis for building chillers // Energy and Buildings. – 2020. – Article No. 109957. – DOI: 10.1016/j.enbuild.2020.109957.

48 Motomura A., Miyata S., Adachi S., Akashi Y., Lim J., Tanaka K., Kuwahara Y. Fault evaluation process in HVAC system for decision making of how to respond to system faults // IOP Conference Series: Earth and Environmental Science. – 2019. – Vol. 294. – Article No. 012054. – DOI: 10.1088/1755-1315/294/1/012054.

49 Bigaud D., Charki A., Caucheteux A., Titikpina F., Tiplica T. Detection of faults and drifts in the energy performance measurement of a building using Bayesian networks // Journal of Dynamic Systems, Measurement, and Control. – 2019. – Vol. 141, No. 11. – Article No. 111009. – DOI: 10.1115/1.4043922.

50 Li D., Zhou Y., Hu G., Spanos C. J. Handling incomplete sensor measurements in fault detection and diagnosis for building HVAC systems // IEEE Transactions on Automation Science and Engineering. – 2019. – Vol. 16, No. 4. – P. 1–14. – DOI: 10.1109/TASE.2019.2948101.

51 Beghi A., Brignoli R., Cecchinato L., Menegazzo G., Rampazzo M., Simmini F. Data-driven fault detection and diagnosis for HVAC water chillers // Control Engineering Practice. – 2016. – Vol. 53. – P. 79–91. – DOI: 10.1016/j.conengprac.2016.04.018.

52 Liu J., Zhang M., Wang H., Zhao W., Liu Y. Sensor fault detection and diagnosis method for AHU using 1-D CNN and clustering analysis // Computational Intelligence and Neuroscience. – 2019. – Vol. 2019. – Article No. 5367217. – DOI: 10.1155/2019/5367217.

53 Zhong C., Dai Y., Jin N., Lou B. Energy efficiency solutions for buildings: Automated fault diagnosis of air-handling units using generative adversarial networks // Energies. – 2019. – Vol. 12, No. 3. – Article No. 527. – DOI: 10.3390/en12030527.

54 Behravan A., Abboush M., Obermaisser R. Deep learning application in mechatronics systems' fault diagnosis: A case study of the demand-controlled ventilation and heating system // 2019 Advances in Science and Engineering Technology International Conferences (ASET). – 2019. – P. 1–5. – DOI: 10.1109/ICASET.2019.8714453.

55 Elnour M., Meskin N., Al-Naemi M. Sensor fault diagnosis of multi-zone HVAC systems using auto-associative neural network // 2019 IEEE Conference on Control Technology and Applications (CCTA). – 2019. – P. 1–6. – DOI: 10.1109/CCTA.2019.8920554.

56 Liu J., Li G., Liu B., Li K., Chen H. Knowledge discovery of data-driven-based fault diagnostics for building energy systems: A case study of the building variable refrigerant flow system // Energy. – 2019. – Vol. 174. – P. 503–514. – DOI: 10.1016/j.energy.2019.02.161.

57 Fan Y., Cui X., Han H., Lu H. Chiller fault diagnosis with field sensors using the technology of imbalanced data // Applied Thermal Engineering. – 2019. – Vol. 159. – Article No. 113933. – DOI: 10.1016/j.applthermaleng.2019.113933.

58 Eom Y. H., Yoo J. W., Hong S. B., Kim M. S. Refrigerant charge fault detection method of air-source heat pump system using convolutional neural network for energy saving // Energy. – 2019. – Article No. 115877. – DOI: 10.1016/j.energy.2019.115877.

59 Mattera C., Quevedo J., Escobet T., Shaker H., Jradi M. A method for fault detection and diagnostics in ventilation units using virtual sensors // Sensors. – 2018. – Vol. 18, No. 11. – Article No. 3931. – DOI: 10.3390/s18113931.

60 Yan K., Ma L., Dai Y., Shen W., Ji Z., Xie D. Cost-sensitive and sequential feature selection for chiller fault detection and diagnosis // International Journal of Refrigeration. – 2018. – Vol. 86. – P. 401–409. – DOI: 10.1016/j.ijrefrig.2017.11.003.

61 Yan, K., Zhong, C., Ji, Z., & Huang, J. (2018). Semi-supervised Learning for Early Detection and Diagnosis of Various Air Handling Unit Faults. Energy and Buildings. – DOI:10.1016/j.enbuild.2018.10.016

62 Chen, Yimin & Wen, Jin & Chen, Taiyu & Pradhan, Ojas. (2018). Bayesian Networks for Whole Building Level Fault Diagnosis and Isolation.

63 Li, Guannan & Chen, Huanxin & Hu, Yunpeng & Wang, Jiangyu & Guo, Yabin & Liu, Jiangyan & Li, Haorong & Huang, Ronggeng & Lv, Hang & Li, Jiong. (2017). An improved decision tree-based fault diagnosis method for practical variable refrigerant flow system using virtual sensor-based fault indicators. Applied Thermal Engineering. 129. 10.1016/j.applthermaleng.2017.10.013.

64 Li, W.-T., UL Hassan, N., Khan, F., Yuen, C., & Keow, Y. (2019). Data driven model for performance evaluation and anomaly detection in integrated air source heat pump operation. 2019 IEEE International Conference on Industrial Technology (ICIT), 1280–1285. – DOI:https://doi.org/10.1109/ICIT.2019.8755022.

65 Turner, W. J. N., Staino, A., & Basu, B. (2017). Residential HVAC fault detection using a system identification approach. Energy and Buildings, 151, 1–17. https://doi.org/10.1016/j.enbuild.2017.06.008

66 Sun, S., Li, G., Chen, H., Huang, Q., Shubiao, S., & Hu, W. (2017). A hybrid ICA-BPNN-based FDD strategy for refrigerant charge faults in variable refrigerant flow system. Applied Thermal Engineering, 127, 91–99. – DOI:https://doi.org/10.1016/j.applthermaleng.2017.08.047

67 Guo, Yabin & Li, Guannan & Chen, Huanxin & Hu, Yunpeng & Li, Haorong & Liu, Jiangyan & Hu, Min & Hu, Wenju. (2017). Modularized PCA method combined with expert-based multivariate decoupling for FDD in VRF systems including indoor unit faults. Applied Thermal Engineering. 115. 10.1016/j.applthermaleng.2017.01.008.

68 Chang, Long & Wang, Hong & Wang, Lingfeng. (2017). Cloud-Based parallel implementation of an intelligent classification algorithm for fault detection and diagnosis of HVAC systems. 1-6. – DOI:10.1109/ISC2.2017.8090835.

69 Chen, Yimin & Wen, Jin. (2017). A whole building fault detection using weather-based pattern matching and feature-based PCA method. 4050-4057. – DOI:10.1109/BigData.2017.8258421.

70 Attouri K., Mansouri M., Hajji M., Kouadri A., Bensmail A., Bouzrara K., Nounou H. Improved fault detection based on kernel PCA for monitoring industrial applications // Journal of Process Control. – 2024. – Vol. 133. – Article No. 103143. – DOI: 10.1016/j.jprocont.2023.103143.

71 Shlens J. A tutorial on principal component analysis // arXiv.org. – 2014. – Mathematics, Computer Science. – DOI: 10.48550/arXiv.1404.1100.

72 Wang T. Fault detection and diagnosis based on principal component analysis // Signal Processing for Fault Detection and Diagnosis in Electric Machines and Systems. – 2020. – Chapter 5. https://digital-library.theiet.org. 05.04.2025.

73 Jung D.Y., Lee S.M., Wang H.M. Fault detection method with PCA and LDA and its application to induction motor // Journal of Central South University. – 2010. – Vol. 17, No. 6. – P. 1238–1242. – DOI: 10.1007/s11771-010-0625-y.

74 Wang F., Xiao F. Detection and diagnosis of AHU sensor faults using principal component analysis method [Electronic resource] // PolyU Scholars Hub. - 2003. -URL: https://research.polyu.edu.hk/en/publications/detection-and-diagnosis-of-ahu-sensor-faults-using-principal-comp.26.04.2025.

75 Zhao X., & Wang X. A fault detection algorithm based on wavelet denoising and KPCA // Advances in Future Computer and Control Systems. – 2012. – Vol. 159. – P. 311–317. – DOI: 10.1007/978-3-642-29387-0_46.

76 Wang H., Zhou H., & Hang B. Number selection of principal components with optimized process monitoring performance // Proceedings of the 43rd IEEE Conference on Decision and Control (CDC). – 2004. – Vol. 4. – P. 4726–4731. – DOI: 10.1109/CDC.2004.1429537

77 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. -Vol. 19. -P. 327–334. – DOI: https://doi.org/10.18576/amis/190209.

78 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. -2024. -№ 4 (149). -С. 202–214. – DOI: https://doi.org/10.32523/2616-7263-2024-149-4-202-214.

79 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, No 2. -DOI: https://doi.org/10.54309/IJICT.2024.18.2.008.

80 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.

81 Espressif Systems. ESP8266 RTOS SDK Documentation. https://docs.espressif.com. 10.11.2022.

82 NodeMCU Firmware. https://github.com.10.11.2022.

83 Bosch Sensortec. BME280: Integrated Environmental Sensor. https://www.bosch-sensortec.com. 11.11.2022.

84 ams-OSRAM. CCS811: Digital Gas Sensor Solution for Monitoring Indoor Air Quality. https://www.ams-osram.com .12.11.2022.

85 SparkFun Electronics. ML8511 UV Sensor Hookup Guide. https://learn.sparkfun.com.13.11.2022.

86 ROHM Semiconductor. BH1750FVI: Digital Ambient Light Sensor IC for I2C Bus Interface. https://www.mouser.com. 14.11.2022.

87 TDK InvenSense. MPU-6050: Six-Axis (Gyro + Accelerometer) MEMS MotionTracking™ Devices. https://invensense.tdk.com. 15.11.2022.

88 International Energy Agency (IEA). The Future of Heat Pumps. 2022. – 82 p. URL: https://www.iea.org/reports/the-future-of-heat-pumps.10.04.2024.

89 ASHRAE. ASHRAE Handbook – HVAC Applications. Atlanta, GA: American Society of Heating, Refrigerating and Air-Conditioning Engineers, 2021. – 1000 p.

90 U.S. Department of Energy. Buildings Energy Data Book. https://www.energy.gov.28.05.2024.

91 International Energy Agency. World Energy Outlook. https://www.iea.org.28.05.2024.

92 ASHRAE. ASHRAE Handbook – Fundamentals. Atlanta, GA: American Society of Heating, Refrigerating and Air-Conditioning Engineers, 2021.

93 ASHRAE. Standard 55-2020: Thermal Environmental Conditions for Human Occupancy. Atlanta, GA: ASHRAE, 2020.

94 ASHRAE. Standard 62.1-2019: Ventilation for Acceptable Indoor Air Quality. Atlanta, GA: ASHRAE, 2019.

95 Statistics Committee of the Republic of Kazakhstan. Reports and statistics on energy consumption and housing services. https://stat.gov.kz.10.05.2024.

96 Dougherty T.R., Jain R.K. Invisible walls: Exploration of microclimate effects on building energy consumption in New York City // Sustainable Cities and Society. – 2023. – Vol. 90. – P. 104364. – DOI: 10.48550/arXiv.2208.03017.

97 Indoor Environment Data Time-Series Reconstruction Using Autoencoder Neural Networks. https://doi.org/10.48550/arXiv.2009.08155. 15.02.2023.

98 Essentials of Data Visualization. https://imarticus.org. 15.02.2023.

99 Z-Score for Outlier Detection in Python. https://www.geeksforgeeks.org. 15.02.2023.

100 Outliers Detection Using IQR, Z-Score, LOF, and DBSCAN. https://www.analyticsvidhya.com.15.02.2023.

101 Capozzoli A., Lauro F., & Khan I. Fault detection analysis using data mining techniques for a cluster of smart office buildings // Expert Systems with Applications. – 2015. –Vol.42, No.9 –P. 4324–4338. – DOI:10.1016/j.eswa.2015.01.010.

102 Altayeva A., Omarov B. Design of a multiagent-based smart microgrid system for building energy and comfort management // Turkish Journal of Electrical Engineering & Computer Sciences. – 2018. – Vol. 26, No. 5. – P. 2714–2725. – DOI: 10.3906/elk-1711-163.

103 Altayeva A., Uskenbayeva R. Intelligent microclimate control in smart building // Bulletin of Satbayev University. Series «Technical Sciences». – 2019. – No. 1 (131). – P. 105–110.

104 Altayeva A. Multi-agent based microclimate control in residential buildings // In: Internet Conference, Satbayev University. – 2018. – DOI: 10.31643/2018.051.

105 Altayeva A., Uskenbayeva R. Agent-based intelligent decision-making system for energy consumption // In: IV International Scientific and Practical Conference «Global Science and Innovations 2019: Central Asia». – Astana, 2019. – P. 198–200.

106 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.

APPENDIX A
Original data (Residential building)

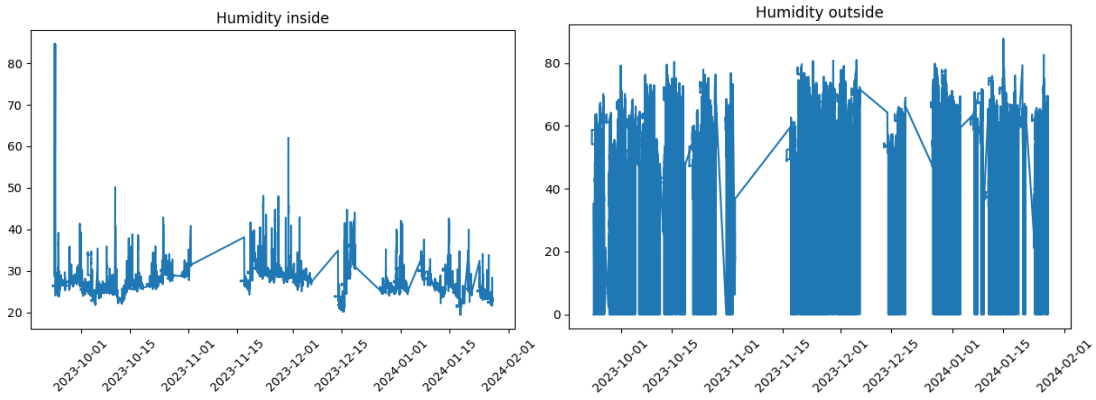

Figure A1 - Room Humidity Outdoor and Humidity Data Over Time Graph

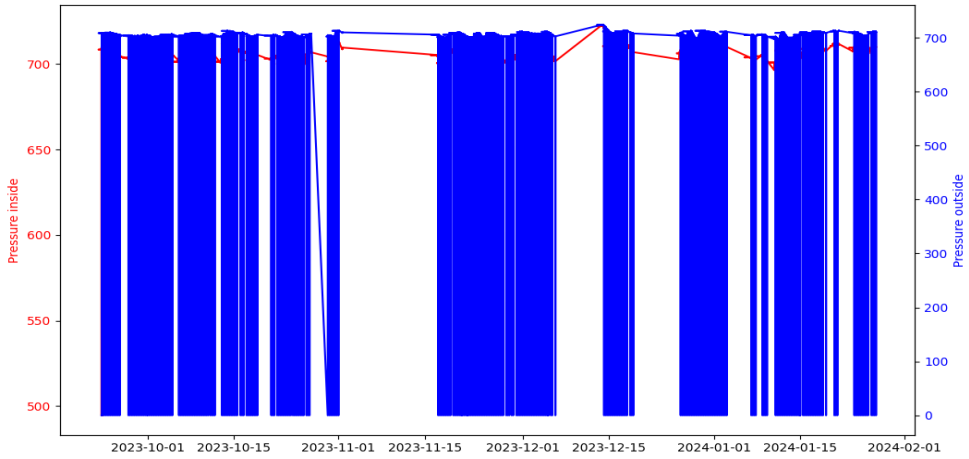

Figure A2- Combined Indoor and Outdoor Pressure Trends

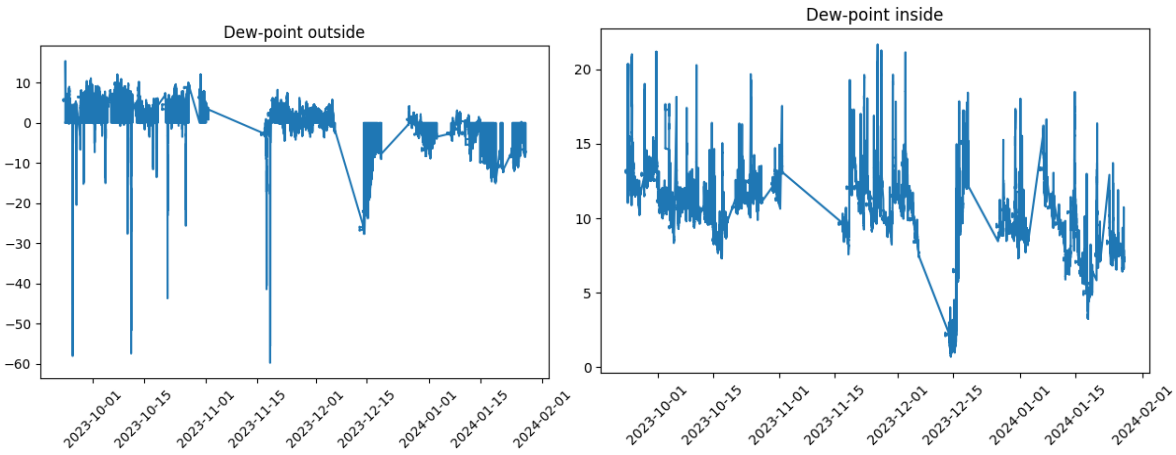

Figure A3 - Room Dew-point

Outdoor Dew-point Data

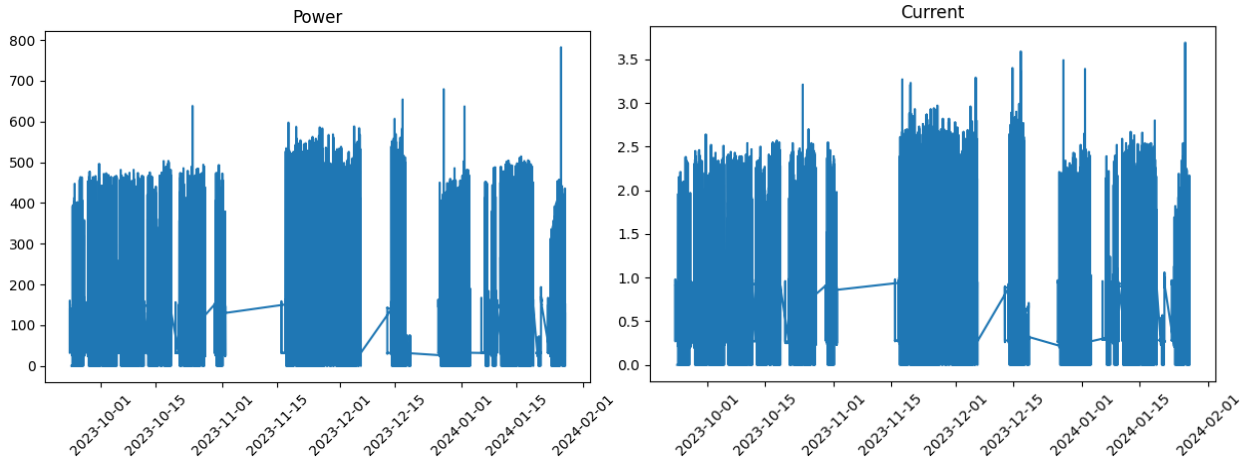

Figure A4 - Power Consumption and Current Over Time

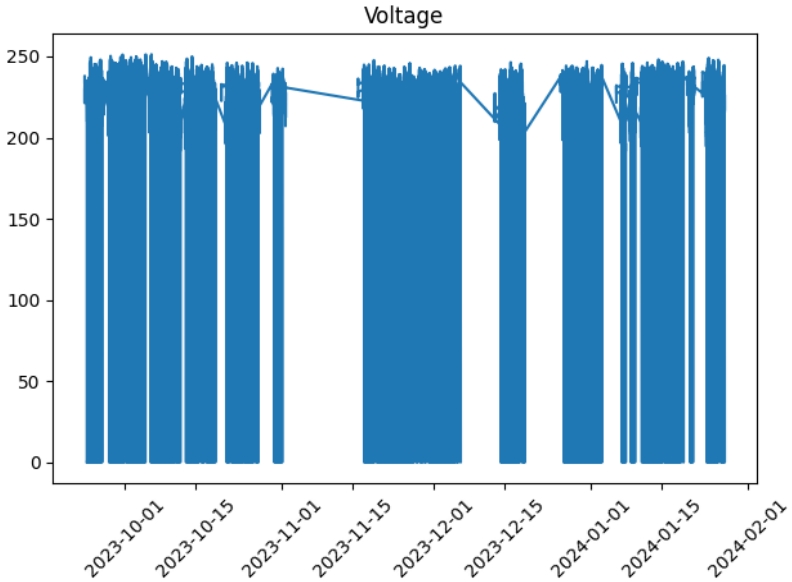

Figure A5 - Temporal Trends in Voltage Variation

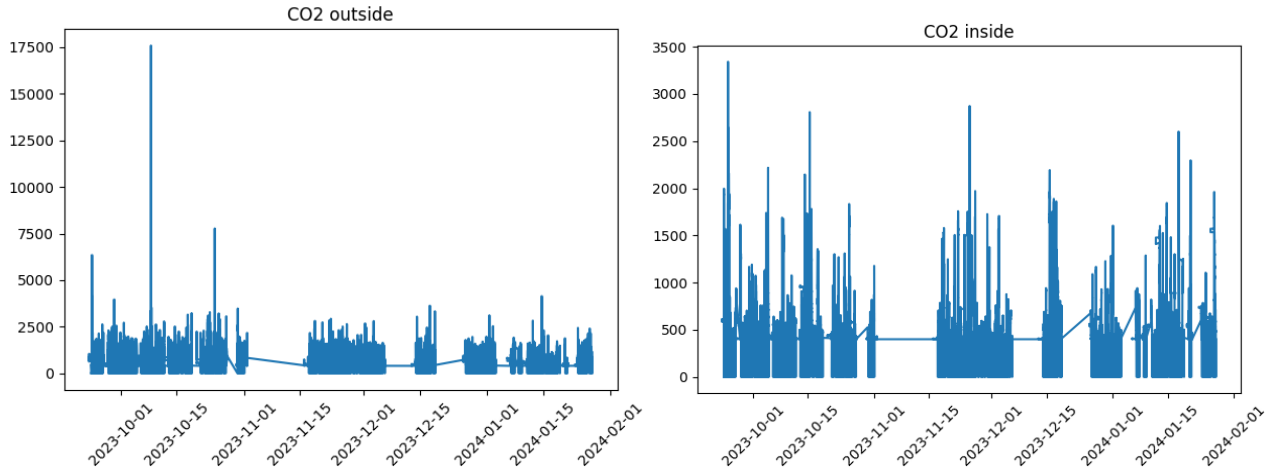

Figure A6 - Outdoor and indoor CO_2 Data

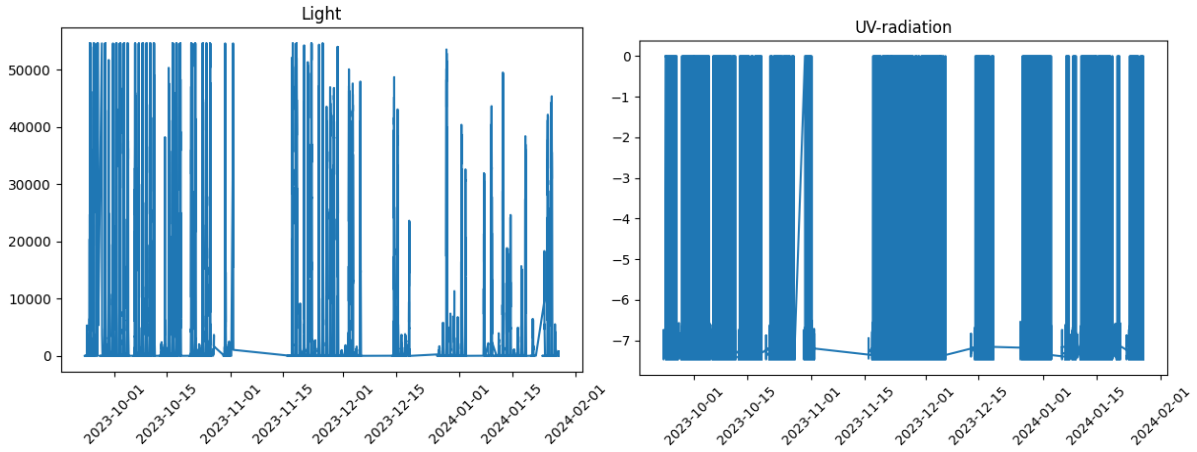

Figure A7 - Temporal Trends in Light Intensity and UV Radiation Exposure Over Time

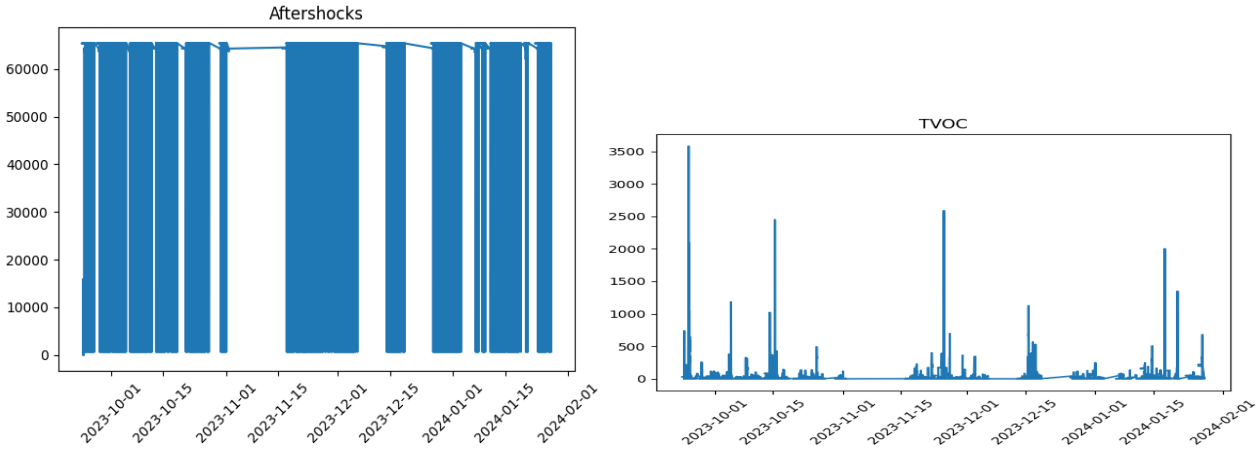

Figure A8 - Aftershock Intensity and TVOC Over Time

For Non-Residential building (original data)

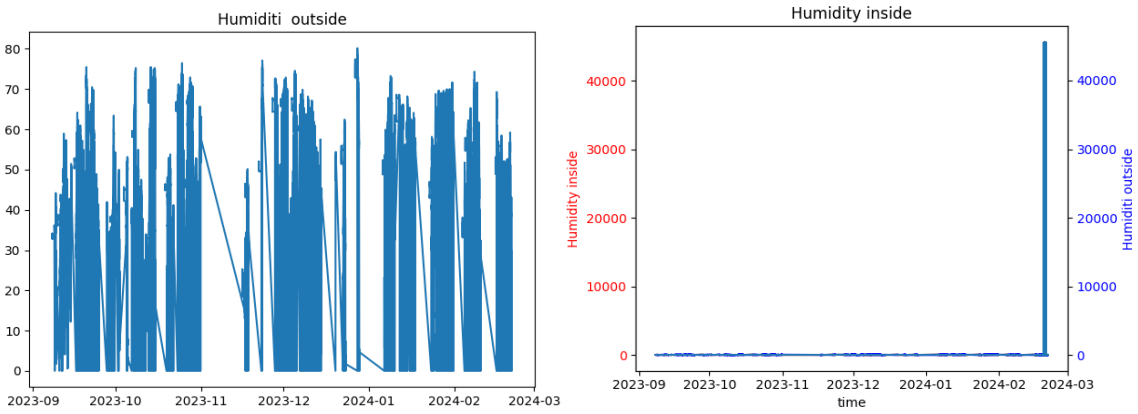

Figure A9 - Outside Humidity and inside Humidity Over Time Graph

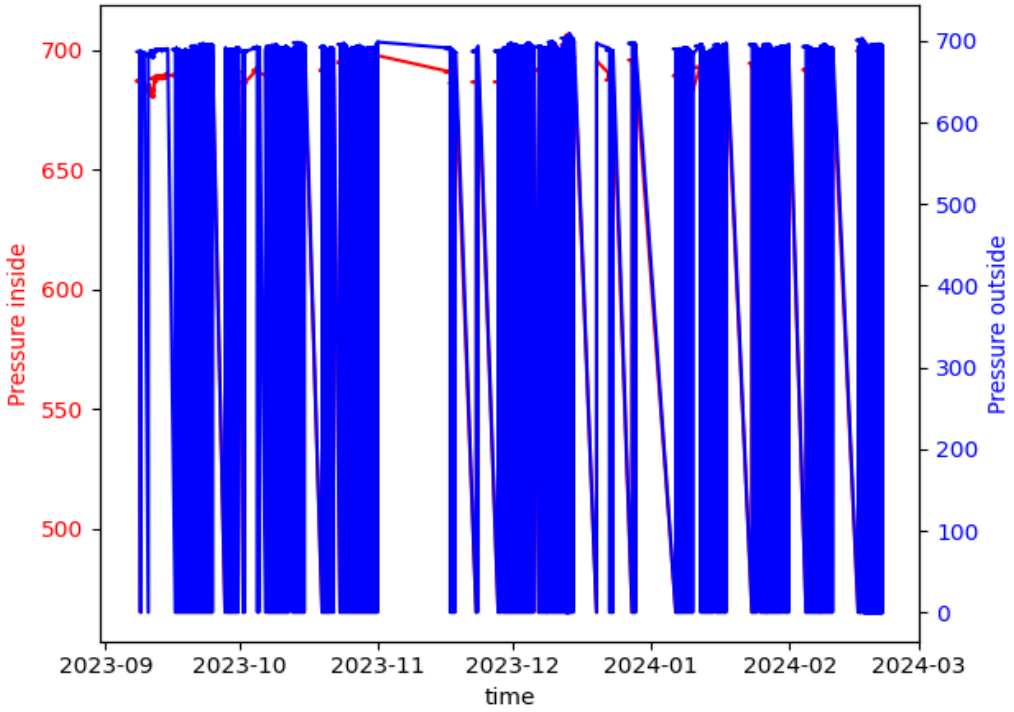

Figure A10 - Combined Pressure Over Time Graph

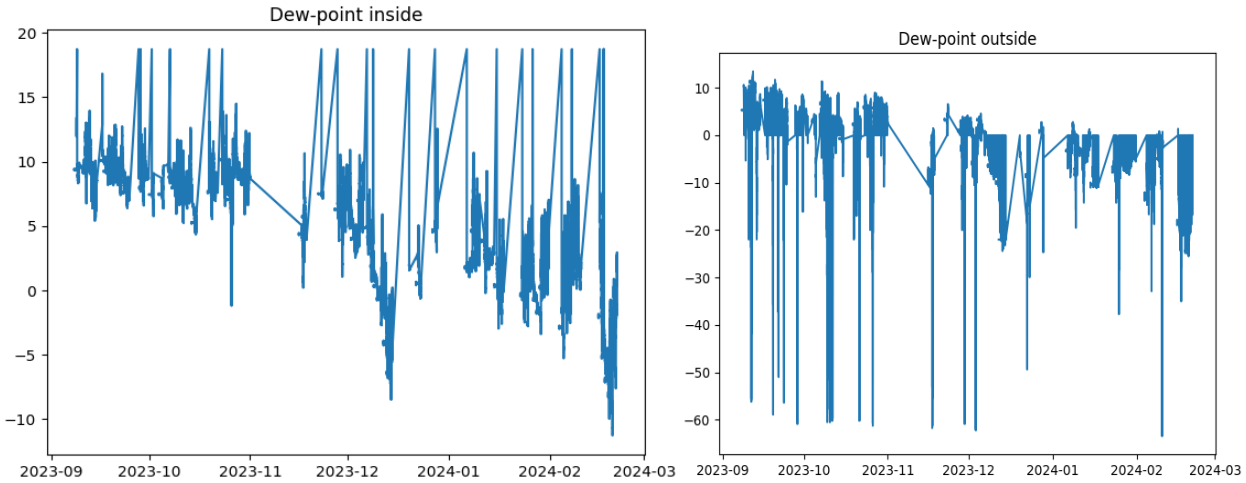

Figure A11 - Inside Dew-point and Outside Humidity Over Time Graph

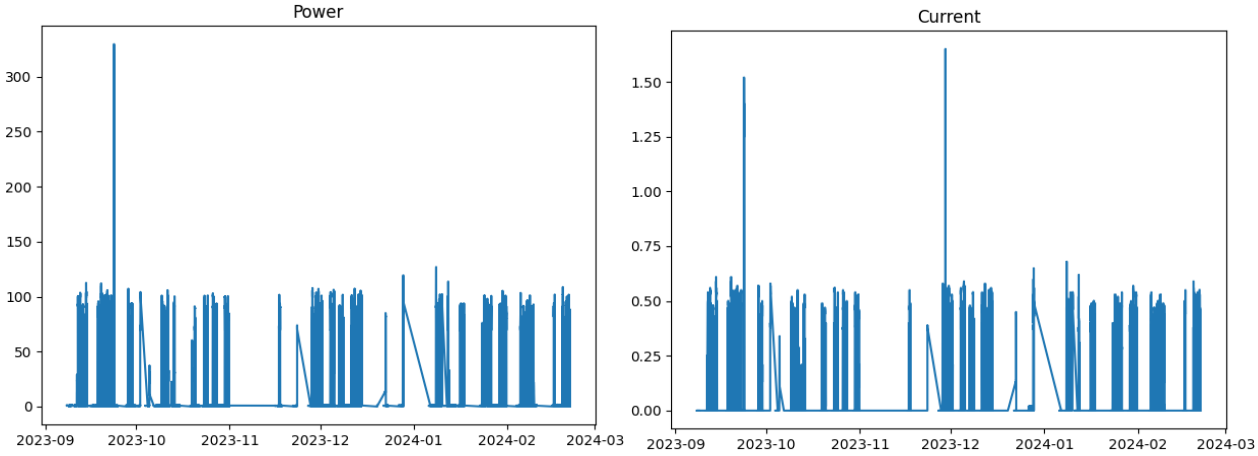

Figure A12 – Power and Current Over Time Graph

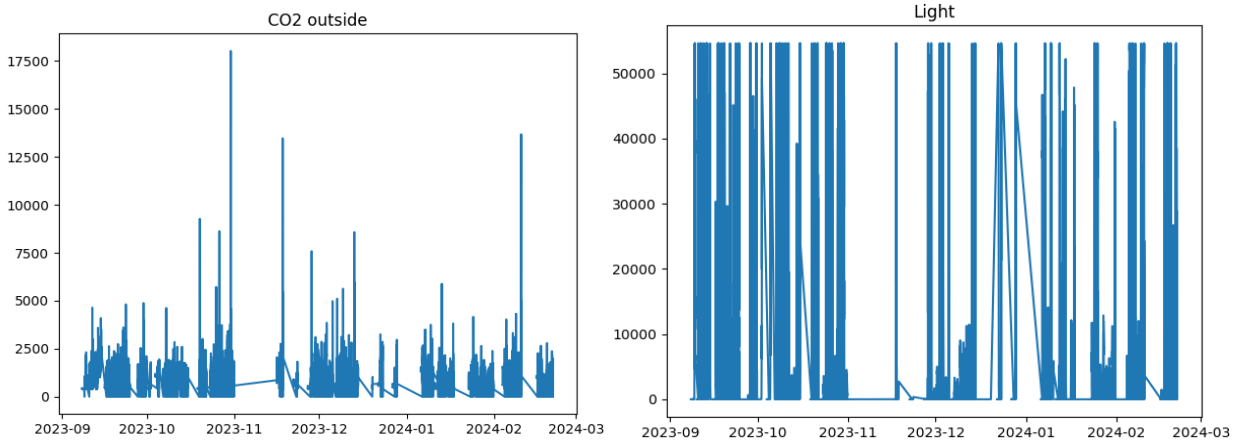

Figure A13 - Outside CO_2 and Light Over Time Graph

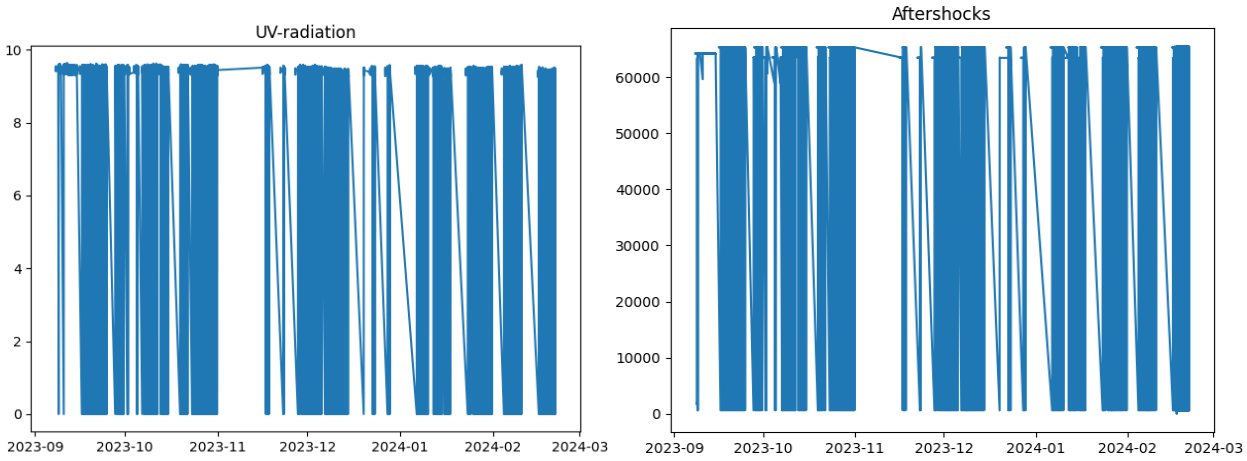

Figure A14- UV-radiation and Aftershocks Over Time Graph

(Cleaning data, Residential building)

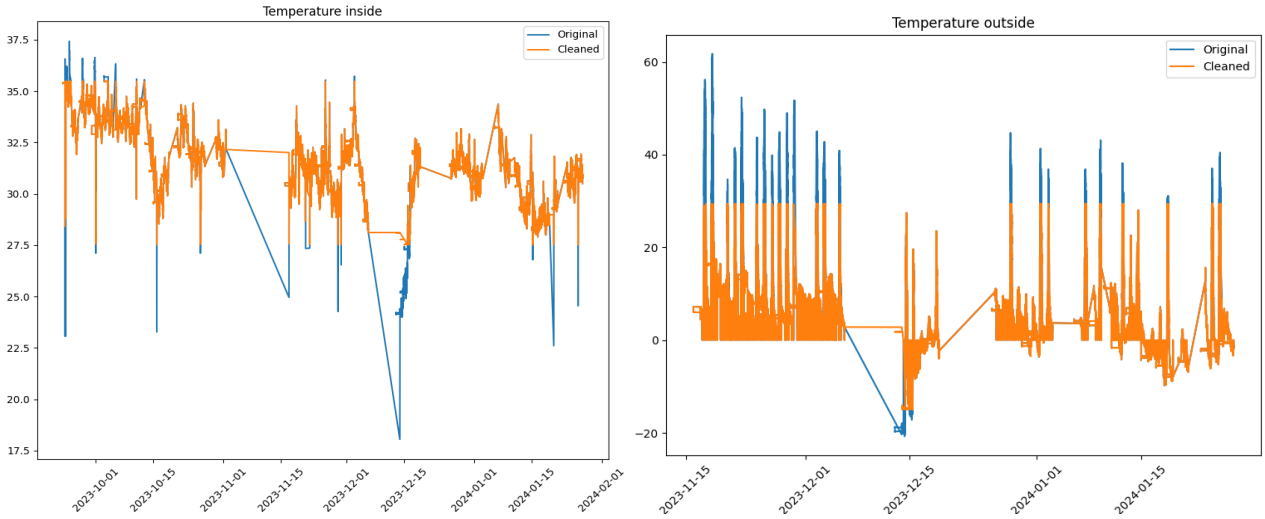

Figure A15 – Room indoor and Outdoor Temperature

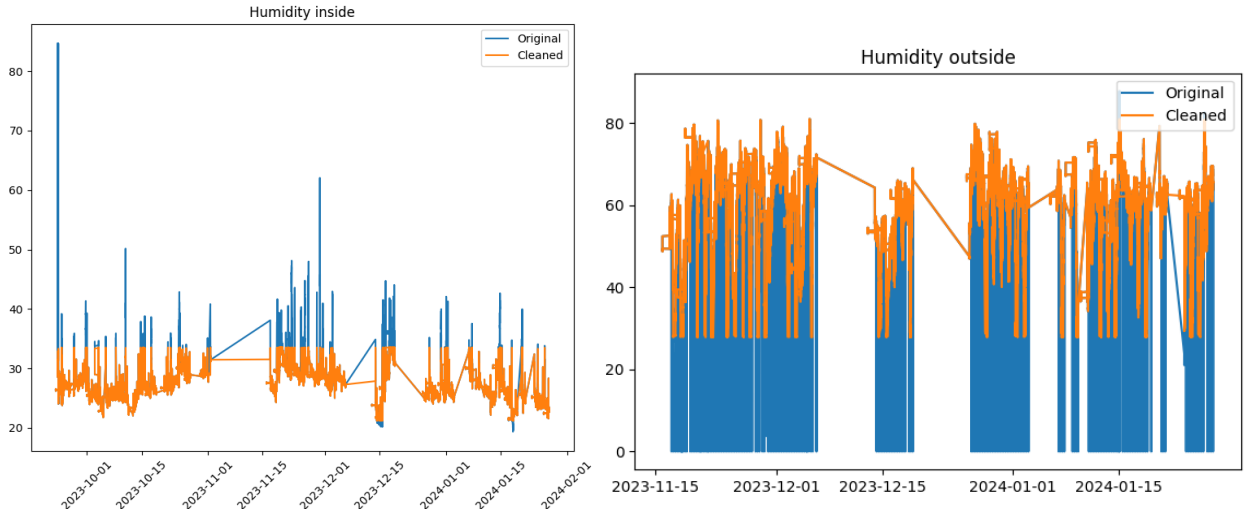

Figure A16 – Room indoor and Outdoor Humidity Over Time Graph after cleaning

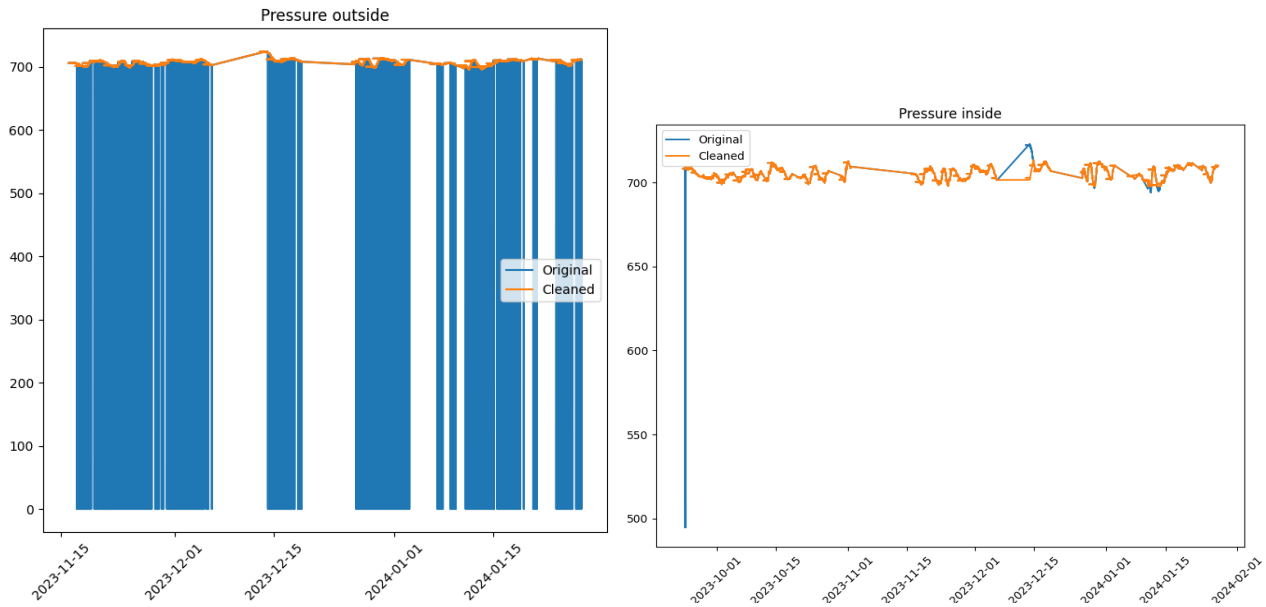

Figure A17 – Outdoor and Indoor Pressure Data after cleaning

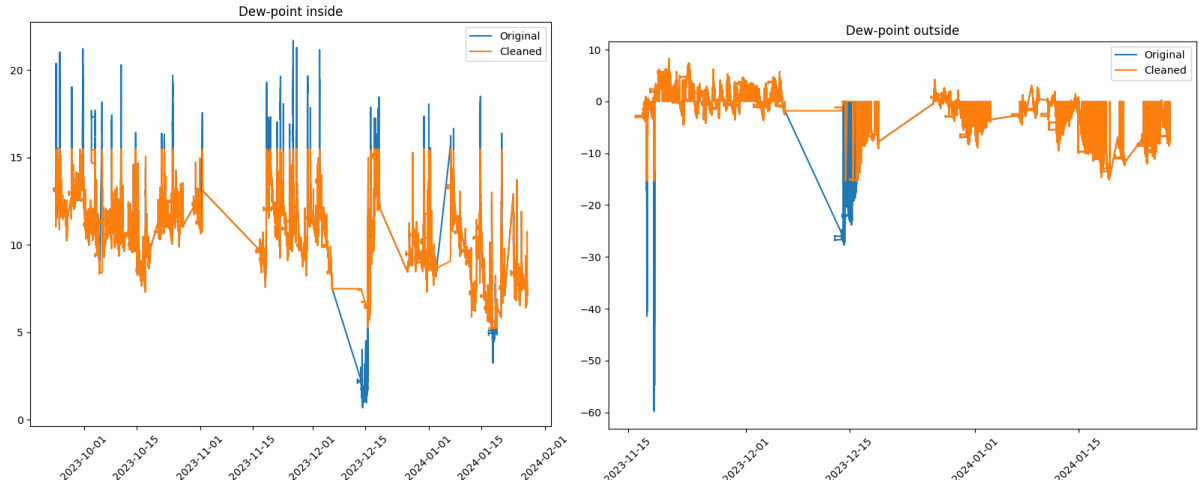

Figure A18 – Room and Outdoor Dew-point Over Time Graph after cleaning

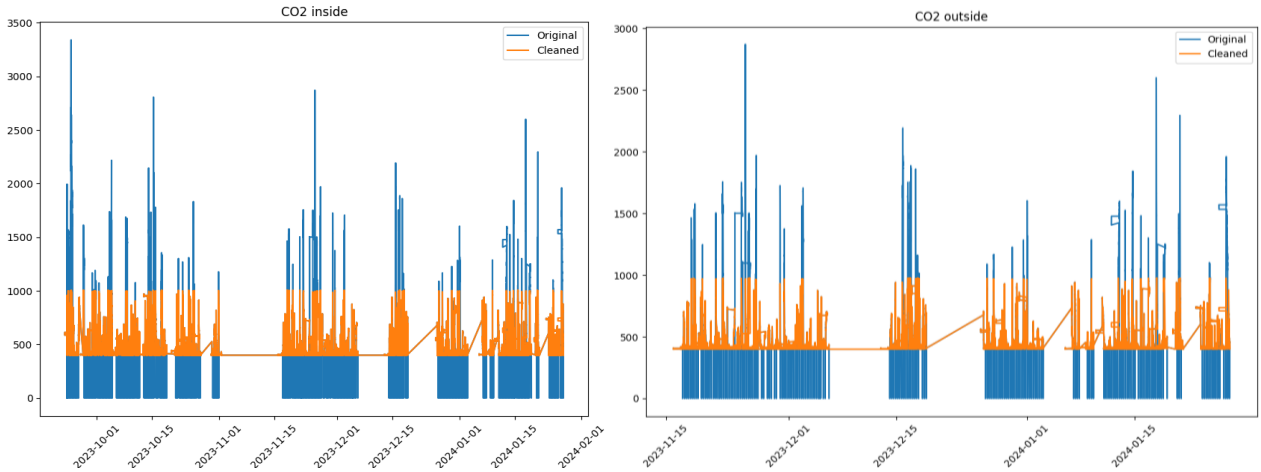

Figure A19 - Room inside and outside CO2 data after cleaning

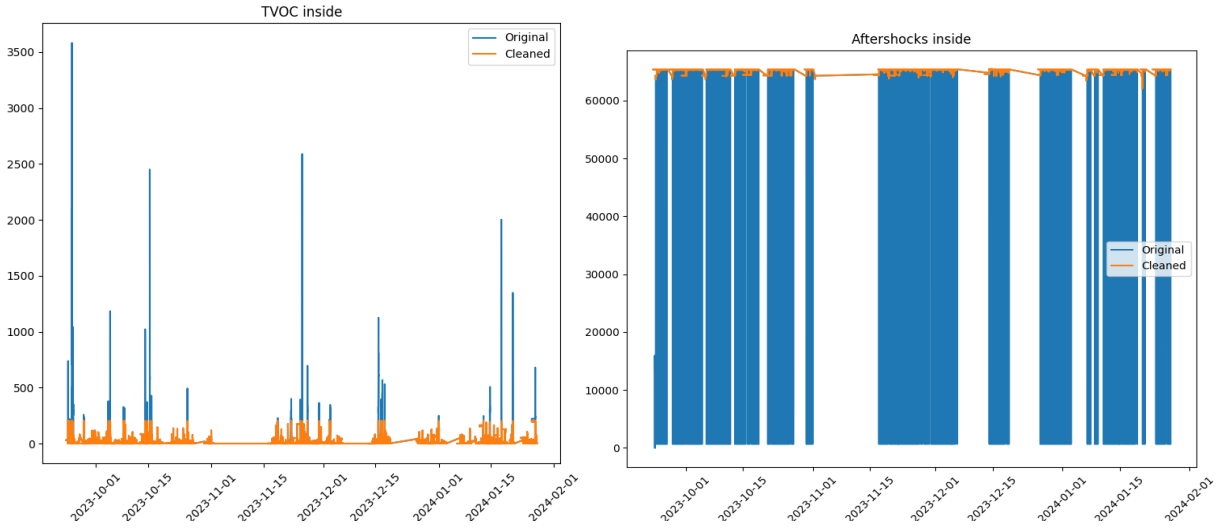

Figure A20 - TVOC and aftershock data after cleaning

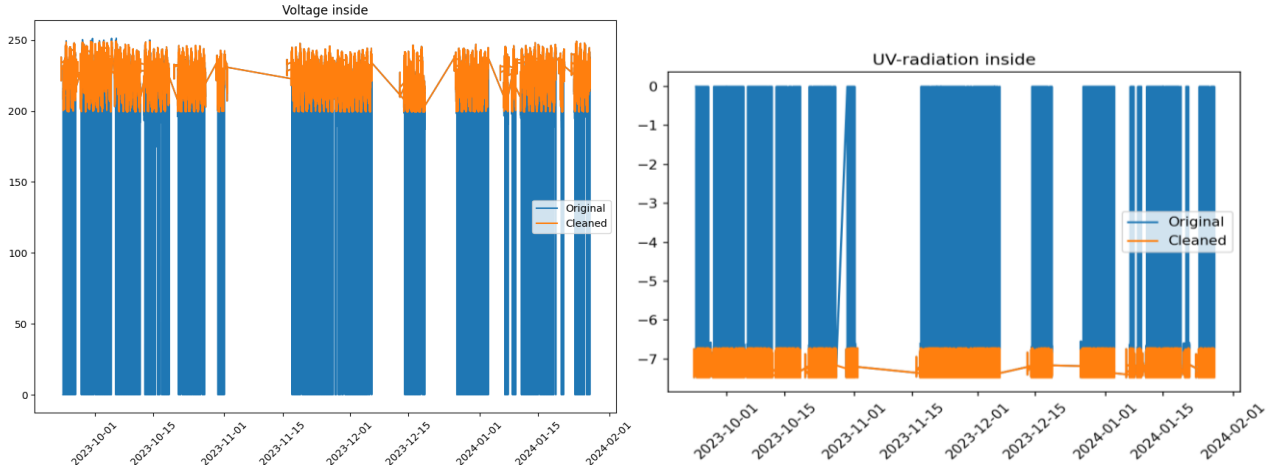

Figure A21 - Temporal Trends in Voltage Variation and UV Radiation Exposure

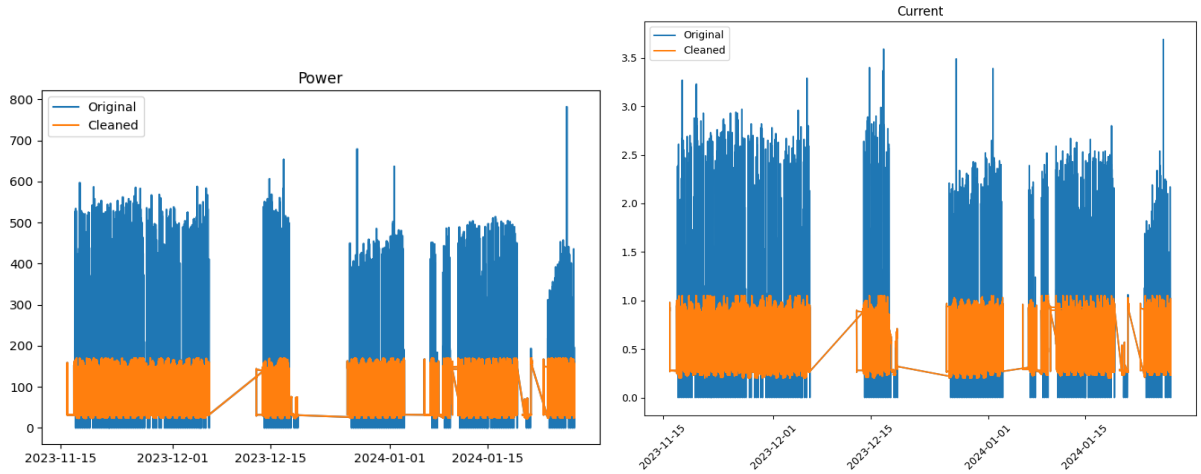

Figure A22- Power Consumption and Electrical Usage Over Time after cleaning

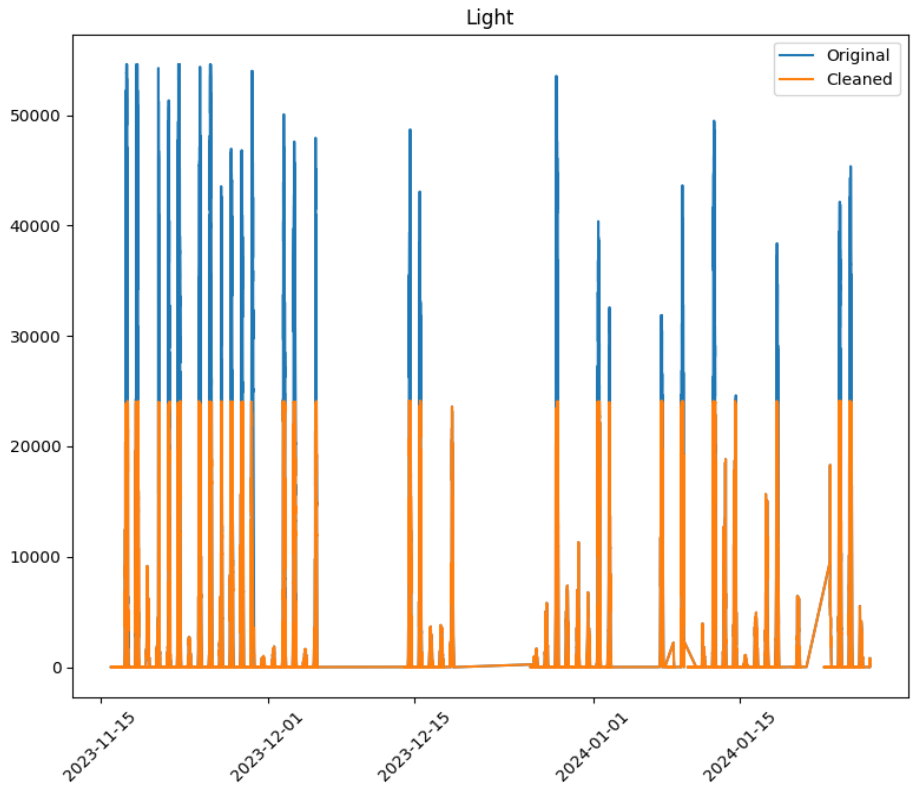

Figure A23 - Temporal Trends in Light Intensity after cleaning

For Non-Residential building(Cleaning data)

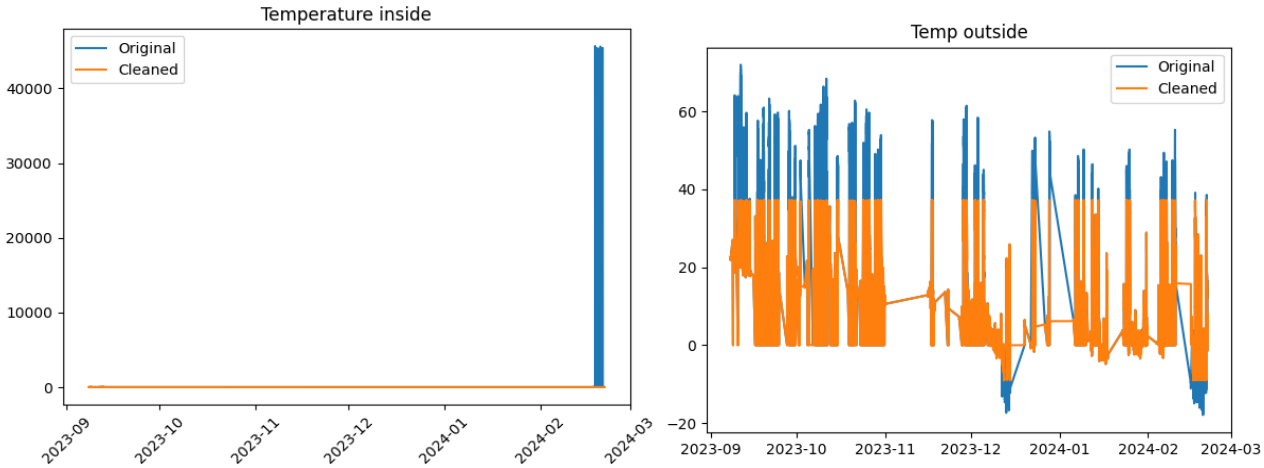

Figure A24 – Inside and Outside temperature Over Time Graph

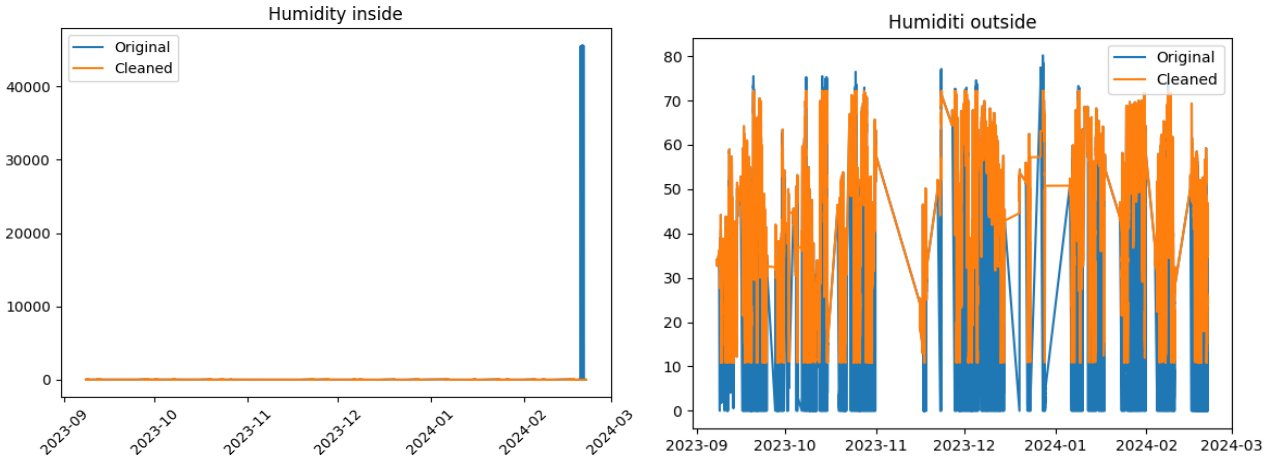

Figure A25 – Inside and Outside Humidity Over Time Graph

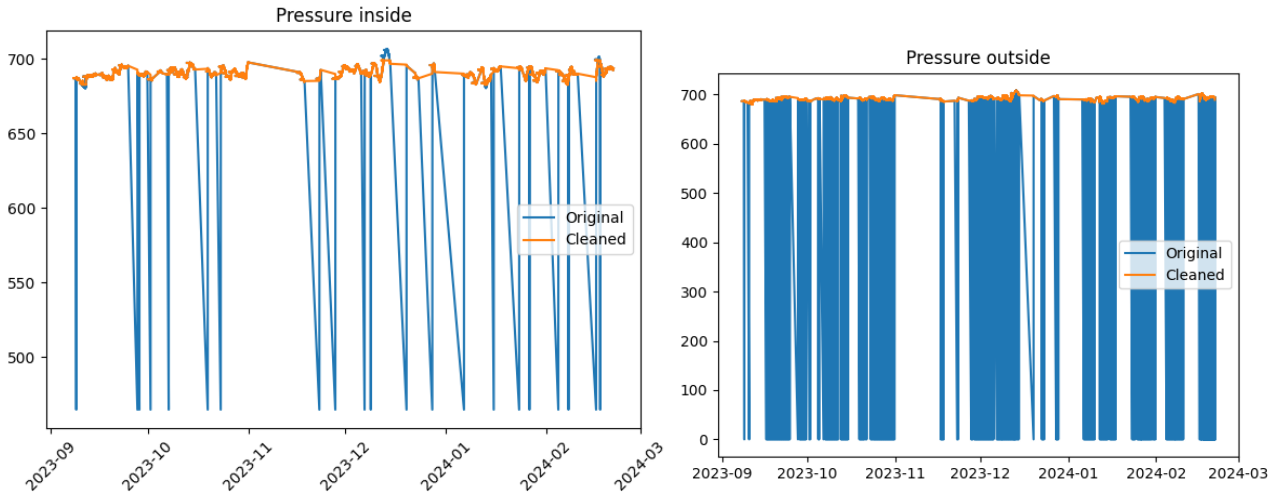

Figure A26 – Inside and Outside Pressure Over Time Graph

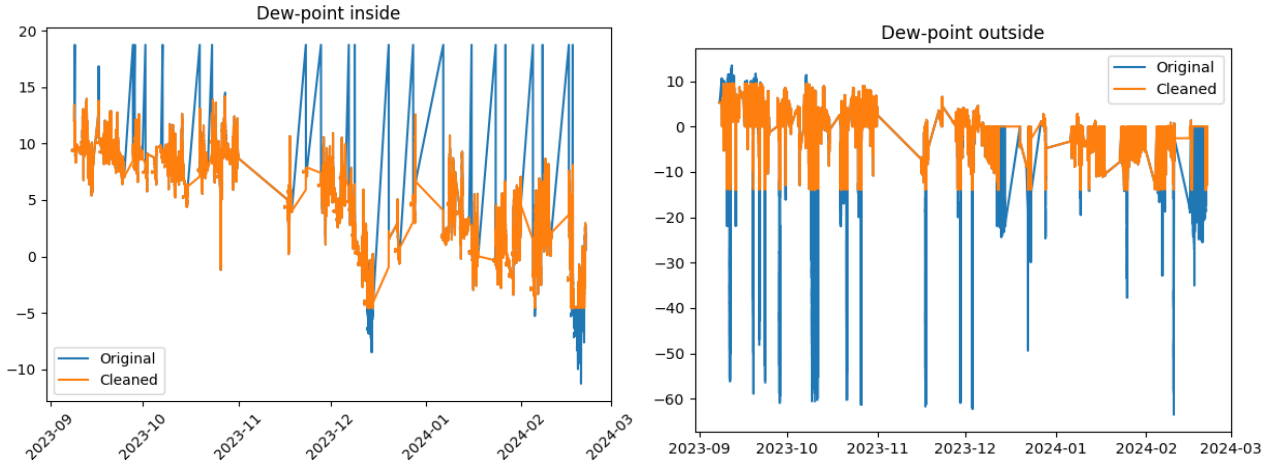

Figure A27 - Inside and outside Dew-point Over Time Graph

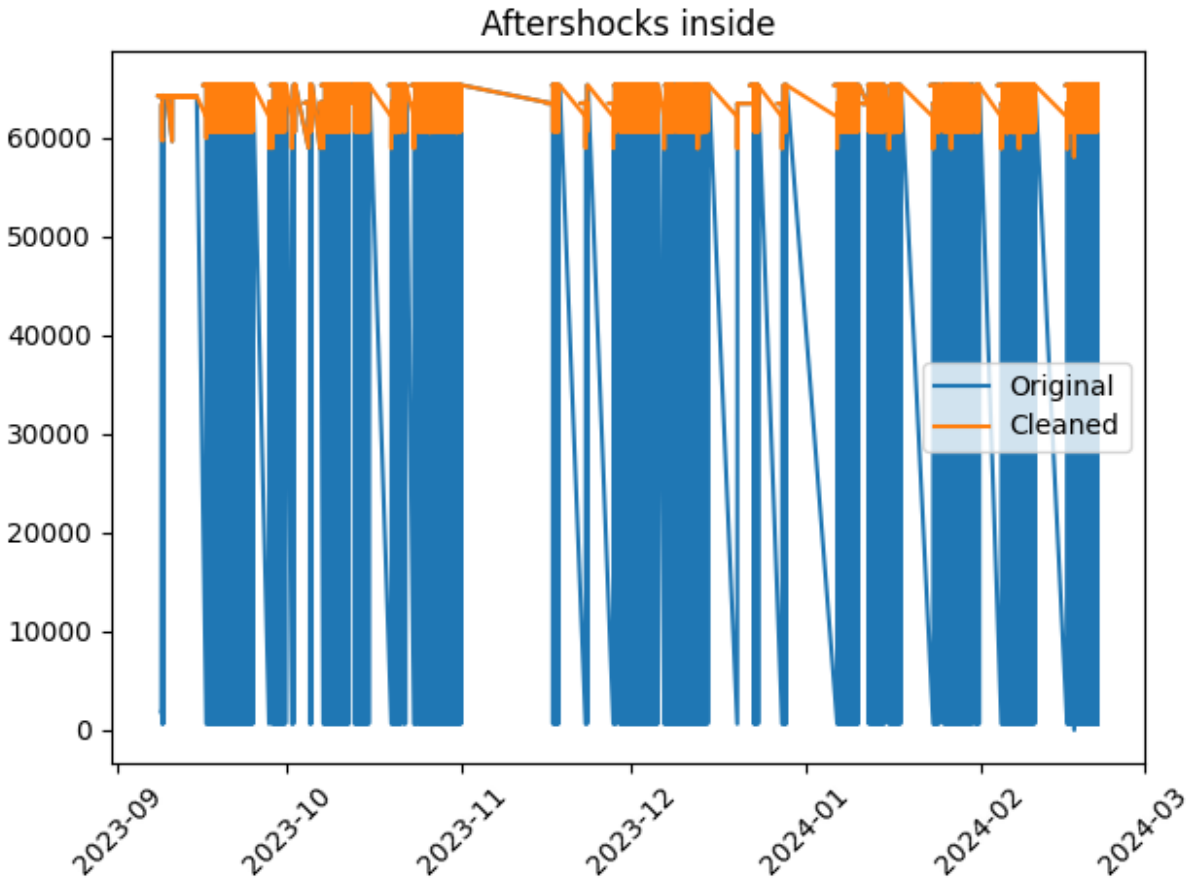

Figure A28 - Combined aftershocks Over Time Graph

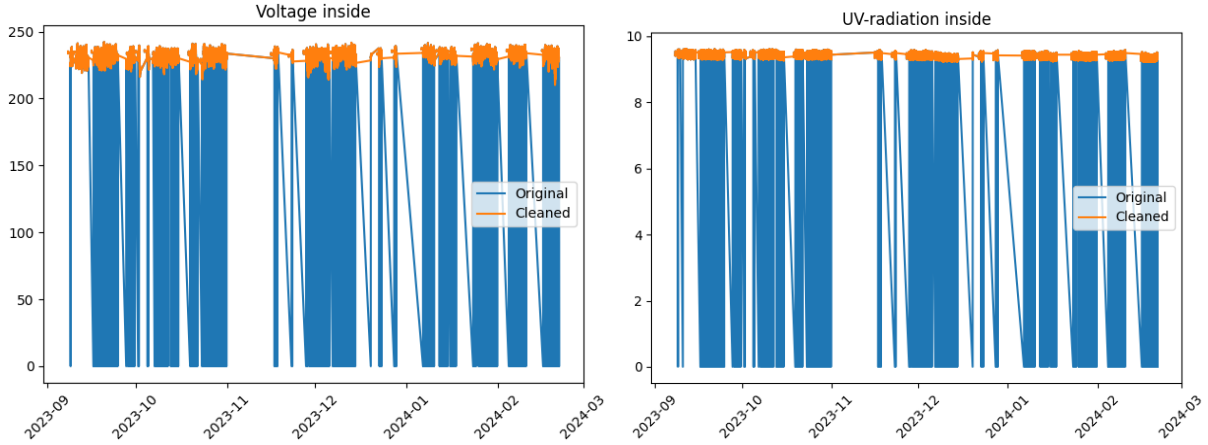

Figure A29 - Voltage and UV-radiation Over Time Graph

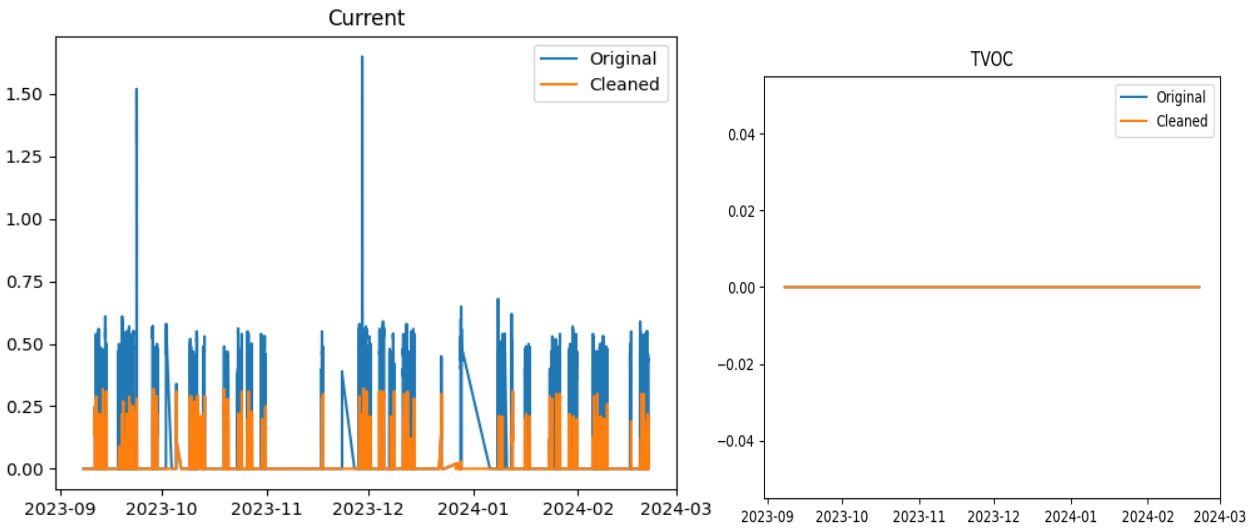

Figure A30 - Current and TVOC Over Time Graph

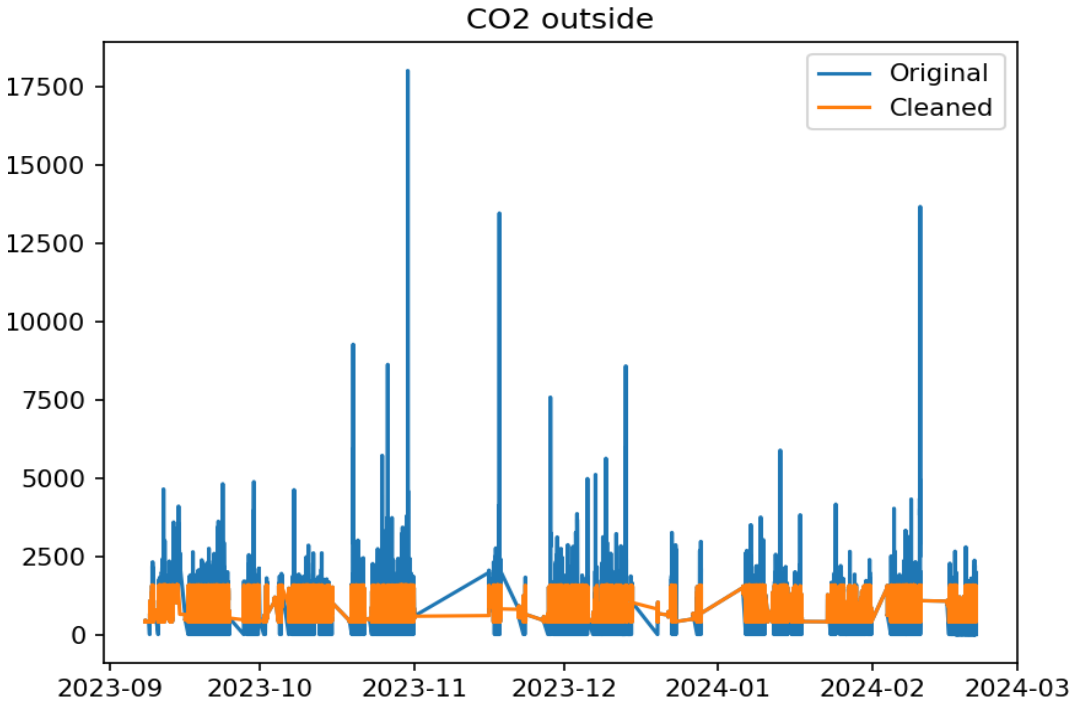

Figure A31 - Outside CO_2 Over Time Graph

APPENDIX B

Program code

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from scipy import stats
#csv = pd.read_csv('march_may.csv', sep=',', header=0, encoding='utf-8')
dfp1 = pd.read_csv('C:/Users/user/Desktop/диссертация/test/march_may1.csv', sep=',', header=0, encoding='utf-8')
#dfp1 = csv.iloc[:260497].copy() <--- This split was necessary because you had dates in the other dataset that used dot and slash
#dfp2 = csv.iloc[260498:].copy()
print('')
print('Data loaded 1st part - tail:', dfp1.tail())
#print('Data loaded 2nd part - head:', dfp2.head())
# Join date and time in the same string
dfp1['Date Time'] = dfp1['Date']+' '+dfp1['Time']
#dfp2['Date Time'] = dfp2['Date']+' '+dfp2['Time']
# Convert date and time from sring to datetime   <--- In this dataset only dot is used
#dfp1['Datetime'] = pd.to_datetime(dfp1['Date Time'], format='%d/%m/%Y %H:%M:%S')
dfp1['Datetime'] = pd.to_datetime(dfp1['Date Time'], format='%d.%m.%Y %H:%M:%S')
#dfp1 = dfp1.drop('Date Time', axis=1)
dfp2 = dfp1.drop('Date Time', axis=1)
# In[]
# Merge into one dataframe


df = pd.concat([dfp1,dfp2])
#df.set_index('Datetime')
#df['Datetime'] = df['Datetime'] #.dt.tz_localize('UTC').dt.tz_convert('Asia/Almaty')
# In[] Find non-numerics and remplace them
mask = pd.to_numeric(df['Voltage'], errors='coerce').isna()
print('Non-numeric values in Voltage (just to see):',mask.sum())
non_numerics = df[mask]
print(non_numerics)
# Replace non-numerics by nan
df2 = df.copy()
df2['Voltage'].iloc[mask] = 0
mask = pd.to_numeric(df2['Voltage'], errors='coerce').isna()
print('Non-numeric values in the cleaned Voltage:',mask.sum())
# In[] Plot just one variable
plt.plot(df2['Datetime'], df2['Voltage'].astype(float))
plt.title('Power')
#plt.xticks(rotation=45)  # Rotate the text if you want
plt.show()
# In[]

plt.figure()
plt.plot(df2['Datetime'], df2['Light'].astype(float))
plt.title('Light')
plt.show()

# In[] Treat non-numeric
```

```
        plt.rcParams["figure.autolayout"] = True

        fig, ax1 = plt.subplots()

        color = 'red'
        ax1.set_xlabel('time')
        ax1.set_ylabel('Temperature',color = color)
        ax1.plot(df2['Datetime'], df2['Temperature'].astype(float), color = color)
        ax1.tick_params(axis='y',labelcolor = color)

        ax2 = ax1.twinx()  # instantiate a second axes that shares the same x-axis

        color = 'blue'
        ax2.set_ylabel('Temperature outside',color = color)  # we already handled the x-label with ax1
        ax2.plot(df2['Datetime'], df2['Pressure outside'].astype(float),color=color)
        ax2.tick_params(axis='y',labelcolor=color)

        #fig.tight_layout()  # otherwise the right y-label is slightly clipped
        plt.show()

        Data Cleaning
        ''' plot clean and original data '''
        def original_clean_plot(df,dfc,var):
            plt.plot(df['Datetime'], df[var],label = 'Original')
            plt.plot(dfc['Datetime'], dfc[var],label = 'Cleaned')
            plt.legend()
            plt.title(var)
            plt.show()

        ''' show statistics for one variable '''
        def show_stats(df, var):
            print('\nStats for ',var)
            print('Minimum Maximum Mean STD')

print(np.min(df[var].astype(float)), np.max(df[var].astype(float)) )

        ''' Clean all variables in the dataset and store result in new dataframe'''


        def clean_data(df):
            print('\n\nCleaning data:')
            dfnumeric = df.copy()  # Dataframe cleaned from non-numeric
            dfc = df.copy()  # Dataframe cleaned from non-numeric and discrepants
            for c in df.columns:
                if c in ['Date', 'Time', 'Datetime']:
                    continue
                print('Cleaning ', c)

                mask = pd.to_numeric(df[c], errors='coerce').isna()
                # Replace non-numerics by nan
                dfc.loc[mask, c] = np.nan
                dfnumeric.loc[mask, c] = np.nan
                dfnumeric[c] = dfnumeric[c].astype(float)
                dfc[c] = dfc[c].astype(float)
                z = stats.zscore(dfc[c], nan_policy='omit')
                # Replace all samples where z > 1.5 by nan
                dfc.loc[np.abs(z) > 1.5, c] = np.nan
```

```
            dfc[c] =
dfc[c].interpolate(method='linear')
        return dfnumeric, dfc


    dfnumeric, dfcleaned =
clean_data(df) # Dataframes cleaned
from non-numeric and from
discrepants too

    # In[] Show plots

    original_clean_plot(dfnumeric,df
cleaned,'Current')
    plt.figure()
    original_clean_plot(dfnumeric,df
cleaned,'Humidity outside')
    plt.figure()
    original_clean_plot(dfnumeric,df
cleaned,'Dew-point outside')


    # In[] Show statistical values

    show_stats(dfnumeric,'Temperat
ure')
    show_stats(dfcleaned,'Temperatu
re')
    show_stats(dfcleaned,'Dew-point
outside')

    # In[] Show correlation matrix

    corr = dfcleaned.corr()

    # In[] PCA

    # In[] PCA

    from sklearn.decomposition
import PCA
    from sklearn.preprocessing
import StandardScaler

    # Identify non-numeric columns
    non_numeric_cols =
dfcleaned.select_dtypes(include=['obje
ct']).columns
    print('Non-numeric columns:',
non_numeric_cols)

    # Convert columns to numeric
and drop non-numeric columns
    dfcleaned =
dfcleaned.apply(pd.to_numeric,
errors='coerce')
    dfcleaned =
dfcleaned.dropna(axis=1, how='all')  #
Drop columns with all NaN values

    # Prepare data for PCA
    dfpca = dfcleaned.drop(['Date',
'Time', 'Datetime'], axis=1,
errors='ignore')
    dfpca = dfpca.dropna()  # Ensure
no NaN values remain

    # Standardize features
    scaler = StandardScaler()
    dfpcasc =
scaler.fit_transform(dfpca)

    # Perform PCA
    pca = PCA(n_components=4)
    comp =
pca.fit_transform(dfpcasc)

    # Plot PCA components
    plt.scatter(comp[:, 0], comp[:,
1])
    plt.title('Data distribution')
    plt.xlabel('First Component')
    plt.ylabel('Second Component')
    plt.show()
    # In[] k-means
    from sklearn.cluster import
KMeans
    import matplotlib.pyplot as plt

    # Apply KMeans clustering
```

```
        kmeans_model =
KMeans(n_clusters=2,
random_state=42)  # Добавлен
random_state для воспроизводимости
        kmeans_model.fit(comp)

        print('Centroids:\n',
kmeans_model.cluster_centers_)

        # Plot KMeans clusters
        plt.figure()
        plt.scatter(comp[:, 0], comp[:, 1],
c=kmeans_model.labels_,
cmap='viridis')  # Отображаем по
первым двум компонентам
        plt.title('KMeans Clustering')
        plt.xlabel('First Component')
        plt.ylabel('Second Component')
        plt.colorbar(label='Cluster
Label')
        plt.show()
        # In[] k-means

        from sklearn.cluster import
KMeans

        # Apply KMeans clustering
        kmeans_model =
KMeans(n_clusters=2)
        kmeans_model.fit(comp)

        print('Centroids:\n',
kmeans_model.cluster_centers_)

        # Plot KMeans clusters
        plt.figure()
        plt.scatter(comp[:, 0], comp[:, 1],
c=kmeans_model.labels_,
cmap='viridis')
        plt.title('KMeans Clustering')
        plt.xlabel('First Component')
        plt.ylabel('Second Component')
        plt.colorbar(label='Cluster
Label')
        plt.show()
        # In[]
        import pickle
        import matplotlib.pyplot as plt

        # Загрузка модели из файла
        fname =
'C:/Users/user/Desktop/диссертация/te
st/country-house-DBSCAN-0.81.bin'
        with open(fname, 'rb') as file:
            model = pickle.load(file)

        # Проверка наличия атрибута
labels_
        if hasattr(model, 'labels_'):
            print('Предсказания:\n',
model.labels_)

            # Визуализация кластеров
            plt.figure()
            plt.scatter(chunk[:, 0],
chunk[:, 1], c=model.labels_)
            plt.title('DBSCAN Clustering')
            plt.xlabel('Feature 1')
            plt.ylabel('Feature 2')
            plt.colorbar(label='Cluster
Label')
            plt.show()

            # Анализ кластеров
            nclusters = set(model.labels_)
            print('Количество
образованных кластеров: ',
len(nclusters) - 1)

            for j in nclusters:
                print('  Точки в кластере ',
j, ': ', len(model.labels_[model.labels_
== j]))

            print('Вне кластеров: ',
len(model.labels_[model.labels_ == -
1]))

            # Статистические
параметры для каждого кластера
```

```
        chunk = dfpca[l: -l]  # Предполагается, что dfpca уже определен

        for j in nclusters:
            print('  Кластер ', j)
            cluster = chunk[model.labels_ == j]
            print(cluster.describe())

        # Статистика для кластеров 1, 2 и 3, а также для точек вне кластеров
        if 0 in nclusters:
            cluster1 = chunk[model.labels_ == 0]
            cluster1data = cluster1.describe()
            print('Статистика кластера 1:\n', cluster1data)

        if 1 in nclusters:
            cluster2 = chunk[model.labels_ == 1]
            cluster2data = cluster2.describe()
            print('Статистика кластера 2:\n', cluster2data)

        if 2 in nclusters:
            cluster3 = chunk[model.labels_ == 2]
            cluster3data = cluster3.describe()
            print('Статистика кластера 3:\n', cluster3data)

        # Статистика для точек вне кластеров
        nocluster = chunk[model.labels_ == -1]
        noclusterdata = nocluster.describe()
        print('Статистика вне кластеров:\n', noclusterdata)
    else:
        print("Модель не содержит атрибут labels_. Проверьте, была ли модель обучена перед сохранением.")

    # In[]
    # Импорт необходимых библиотек
    import pickle
    import matplotlib.pyplot as plt
    from sklearn.cluster import DBSCAN

    # Загрузка модели из файла
    fname = 'C:/Users/user/Desktop/диссертация/test/country-house-DBSCAN-0.81.bin'
    with open(fname, 'rb') as file:
        model = pickle.load(file)

    # Вывод предсказанных меток кластеров
    print('Предсказания:\n', model.labels_)

    # Визуализация кластеров
    plt.figure()
    plt.scatter(chunk[:, 0], chunk[:, 1], c=model.labels_)
    plt.title('DBSCAN Clustering')
    plt.xlabel('Feature 1')
    plt.ylabel('Feature 2')
    plt.colorbar(label='Cluster Label')
    plt.show()

    # Анализ кластеров
    nclusters = set(model.labels_)
    print('Количество образованных кластеров: ', len(nclusters) - 1)

    for j in nclusters:
```

```
            print('  Точки в кластере ', j, ':
', len(model.labels_[model.labels_ ==
j]))

        print('Вне кластеров: ',
len(model.labels_[model.labels_ == -
1]))

        # Статистические параметры
для каждого кластера
        chunk = dfpca[l: -l]  #
Предполагается, что dfpca уже
определен

        for j in nclusters:
            print('  Кластер ', j)
            cluster = chunk[model.labels_
== j]
            print(cluster.describe())

        # Вывод статистики для
кластера 1, 2 и 3, а также для точек
вне кластеров
        if 0 in nclusters:
            cluster1 =
chunk[model.labels_ == 0]
            cluster1data =
cluster1.describe()
            print('Статистика кластера
1:\n', cluster1data)

        if 1 in nclusters:
            cluster2 =
chunk[model.labels_ == 1]
            cluster2data =
cluster2.describe()
            print('Статистика кластера
2:\n', cluster2data)

        if 2 in nclusters:
            cluster3 =
chunk[model.labels_ == 2]
            cluster3data =
cluster3.describe()
            print('Статистика кластера
3:\n', cluster3data)

        # Статистика для точек вне
кластеров
        nocluster = chunk[model.labels_
== -1]
        noclusterdata =
nocluster.describe()
        print('Статистика вне
кластеров:\n', noclusterdata)

        # In[] DBSCAN
        import pickle
        import matplotlib.pyplot as plt
        from sklearn.cluster import
DBSCAN

        # Загрузка модели из файла
        fname =
'C:/Users/user/Desktop/диссертация/te
st/country-house-DBSCAN-0.81.bin'
        with open(fname, 'rb') as file:
            model = pickle.load(file)

        # Проверка наличия атрибута
labels_
        if hasattr(model, 'labels_'):
            print('Предсказания:\n',
model.labels_)

            # Визуализация кластеров
            plt.figure()
            plt.scatter(chunk[:, 0],
chunk[:, 1], c=model.labels_)
            plt.title('DBSCAN Clustering')
            plt.xlabel('Feature 1')
            plt.ylabel('Feature 2')
            plt.colorbar(label='Cluster
Label')
            plt.show()

            # Анализ кластеров
            nclusters = set(model.labels)
```

```
        print('Количество
образованных кластеров: ',
len(nclusters) - 1)

        for j in nclusters:
            print('  Точки в кластере ',
j, ': ', len(model.labels_[model.labels_
== j]))

        print('Вне кластеров: ',
len(model.labels_[model.labels_ == -
1]))

        # Статистические
параметры для каждого кластера
        chunk = dfpca[l: -l]  #
Предполагается, что dfpca уже
определен

        for j in nclusters:
            print('  Кластер ', j)
            cluster =
chunk[model.labels_ == j]
            print(cluster.describe())

        # Статистика для кластеров
1, 2 и 3, а также для точек вне
кластеров
        if 0 in nclusters:
            cluster1 =
chunk[model.labels_ == 0]
            cluster1data =
cluster1.describe()
            print('Статистика кластера
1:\n', cluster1data)

        if 1 in nclusters:
            cluster2 =
chunk[model.labels_ == 1]
            cluster2data =
cluster2.describe()
            print('Статистика кластера
2:\n', cluster2data)

        if 2 in nclusters:
            cluster3 =
chunk[model.labels_ == 2]
            cluster3data =
cluster3.describe()
            print('Статистика кластера
3:\n', cluster3data)

        # Статистика для точек вне
кластеров
        nocluster =
chunk[model.labels_ == -1]
        noclusterdata =
nocluster.describe()
        print('Статистика вне
кластеров:\n', noclusterdata)
    else:
        print("Модель не содержит
атрибут labels_. Проверьте, была ли
модель обучена перед
сохранением.")
```

```
    # In[] Cluster analysis

    import pickle
    file = open(fname, 'rb')
    # dump information to that file
    model = pickle.load(file)
    # close the file
    file.close()
    print('Predições:\n',model.labels)

    plt.figure()
    plt.scatter(chunk[:,0], chunk[:,1],
c=model.labels_)


    # In[] Cluster analysis

    nclusters = set(model.labels)
```

```
        print('Number of clusters
formed: ',len(nclusters)-1)
        for j in nclusters:
            print('  Points in cluster ',j,
                  ':
',len(model.labels_[model.labels_==j
]))

        print('Out of clusters: ',
len(model.labels_[  model.labels_==-1
]) )

        import pickle
        import matplotlib.pyplot as plt
        from sklearn.cluster import
DBSCAN

        # Загрузка модели из файла
        fname =
'C:/Users/user/Desktop/диссертация/te
st/country-house-DBSCAN-0.81.bin'
        with open(fname, 'rb') as file:
            model = pickle.load(file)

        # Вывод предсказанных меток
кластеров
        print('Предсказания:\n',
model.labels_)

        # Визуализация кластеров
        plt.figure()
        plt.scatter(chunk[:, 0], chunk[:,
1], c=model.labels_)
        plt.title('DBSCAN Clustering')
        plt.xlabel('Feature 1')
        plt.ylabel('Feature 2')
        plt.colorbar(label='Cluster
Label')
        plt.show()

        # In[] statistical parameters per
cluster

        chunk = dfpca[l : -l]

        for j in nclusters:
            print('  Cluster ',j)
            cluster = chunk [
model.labels_==j  ]
            print(cluster.describe())

        cluster1 = chunk [
model.labels_==0  ]
        cluster1data = cluster1.describe()

        cluster2 = chunk [
model.labels_==1  ]
        cluster2data = cluster2.describe()

        cluster3 = chunk [
model.labels_==2  ]
        cluster3data = cluster3.describe()

        nocluster = chunk [
model.labels_==-1  ]
        noclusterdata =
nocluster.describe()
        # In[] DBSCAN

        from sklearn.cluster import
DBSCAN
        import pickle

        #x: Data,
        #y: Labels

        #model = DBSCAN(eps=1.1,
min_samples=5)

        # I'll get just a slice of the dataset
to make DBSCAN faster
        l = int(len(comp) / 3 * 0.9)
        chunk = comp[l : -l]
        #model.fit(chunk)
        # In[] Save into pickle
        import pickle
        fname =
'C:/Users/user/Desktop/диссертация/te
st/country-house-DBSCAN-0.81.bin'
        file = open(fname, 'wb')
```

```
        # dump information to that file
        pickle.dump(model, file)
        # close the file
        file.close()

        # In[] Cluster analysis
        import pickle
        file = open(fname, 'rb')
        # dump information to that file
        model = pickle.load(file)
        # close the file
        file.close()
        print('Predições:\n',model.labels_
)
        plt.figure()
        plt.scatter(chunk[:,0], chunk[:,1],
c=model.labels_)
        plt.show()
        # In[] Cluster analysis
        nclusters = set(model.labels_)
        print('Number of clusters
formed: ',len(nclusters)-1)
        for j in nclusters:
            print('  Points in cluster ',j,
                  ':
',len(model.labels_[model.labels_==j
]))
        print('Out of clusters: ',
len(model.labels_[  model.labels_==-1
]) )
        # In[] statistical parameters per
cluster
        chunk = dfpca[l : -l]
        for j in nclusters:
            print('  Cluster ',j)
            cluster = chunk [
model.labels_==j  ]
            print(cluster.describe())
        cluster1 = chunk [
model.labels_==0  ]
        cluster1data = cluster1.describe()
        cluster2 = chunk [
model.labels_==1  ]
        cluster2data = cluster2.describe()
        cluster3 = chunk [
model.labels_==2  ]
        cluster3data = cluster3.describe()
        nocluster = chunk [
model.labels_==-1  ]
        noclusterdata =
nocluster.describe()
```

APPENDIX C

Copyright certificate

ҚАЗАҚСТАН РЕСПУБЛИКАСЫ

РЕСПУБЛИКА КАЗАХСТАН

СВИДЕТЕЛЬСТВО

О ВНЕСЕНИИ СВЕДЕНИЙ В ГОСУДАРСТВЕННЫЙ РЕЕСТР ПРАВ НА ОБЪЕКТЫ, ОХРАНЯЕМЫЕ АВТОРСКИМ ПРАВОМ

№ 41781 от «5» января 2024 года

Фамилия, имя, отчество, (если оно указано в документе, удостоверяющем личность) автора (ов):
ДАУРЕНБАЕВА НУРКАМИЛЯ АЛДАНГАРОВНА, АТЫМТАЕВА ЛЯЗЗАТ БАХИТОВНА, ЫБЫТАЕВА ГАЛИЯ СЕЙТКАЛИЕВНА, НҰРЛАНҰЛЫ АЛМАС

Вид объекта авторского права: **программа для ЭВМ**

Название объекта: **Аппаратный комплекс, предназначенный для реального мониторинга параметров микроклимата с интегрированным датчиком сейсмических воздействий**

Дата создания объекта: **04.01.2024**

Қазақстан Республикасының https://www.kazpatent.kz/kz/ сайтының "Авторлық құқық" бөлімінде тексеруге болады https://copyright.kazpatent.kz

Подлинность документа возможно проверить на сайте kazpatent.kz в разделе «Авторское право» https://copyright.kazpatent.kz

Подписано ЭЦП

Е. Оспанов

Participation Certificate – International Conference

№ 2023-05/055

SULEYMAN DEMIREL UNIVERSITY

ICECCO 2023

CERTIFICATE

of sucessful presentation at ICECCO 2023, this certifies that

Nurkamilya Daurenbayeva, Lyazzat Atymtayeva and Almas Nurlanuly

has successfully presented a research paper titled «CHOOSING THE INTELLIGENT THERMOSTATS FOR THE EFFECTIVE DECISION MAKING IN BEMS» at the 17th International Conference on Electronics, Computer and Communication (ICECCO 2023) held from 1-2 June 2023 in Kaskelen.

Rector of SDU

Igenbayev Alimzhan

Certificate of Research Internship

Instituto Superior de Engenharia
Politécnico de Coimbra

Certificate of Stay

ISEC – Coimbra Institute of Engineering

CERTIFICATE OF STAY

Coimbra Institute of Engineering (ISEC) - Polytechnic Institute of Coimbra (P COIMBRA 02) hereby states that the student **__Nurkamilya A. Daurenbayeva _____**, from__International Information Technology University, Almaty_[*University*],__[*Erasmus+code*]-Kazakhstan[*country*], has attended our institution and successfully completed her research as PHD student:

☐ **Erasmus+ Studies Mobility**

X **Erasmus+ Training Mobility**

Period of stay:

Arrival date: _01__ /_04__ /_2024___ Departure date: _30__ /04_ /2024__

Coimbra, 30 de Abril de 2024

International Coordinator's signature

_____

Prof. Luís Castro, *PhD*

1 de 1
Mod6_304.03
SISTEMA INTERNO DE GARANTIA DA QUALIDADE