Advanced IoT-Enabled Indoor Thermal Comfort Prediction Using SVM and Random Forest Models

Nurtileu Assymkhan, Amandyk Kartbayev
School of Information Technology and Engineering, Kazakh-British Technical University, Almaty, Kazakhstan

***Abstract*—Predicting thermal comfort within indoor environments is essential for enhancing human health, productivity, and well-being. This study uses interdisciplinary approaches, integrating insights from engineering, psychology, and data science to develop sophisticated machine learning models that predict thermal comfort. Traditional methods often depend on subjective human input and can be inefficient. In contrast, this research applies Support Vector Machines (SVM) and Random Forest algorithms, celebrated for their precision and speed in handling complex datasets. The advent of the Internet of Things (IoT) further revolutionizes building management systems by introducing adaptive control algorithms and enabling smarter, IoT-driven architectures. We focus on the comparative analysis of SVM and Random Forest in predicting indoor thermal comfort, discussing their respective advantages and limitations under various environmental conditions and building designs. The dataset we used included comprehensive thermal comfort data, which underwent rigorous preprocessing to enhance model training and testing—80% of the data was used for training and the remaining 20% for testing. The models were evaluated based on their ability to accurately mirror complex interactions between environmental factors and occupant comfort levels. The results indicated that while both models performed robustly, Random Forest demonstrated greater stability and slightly higher accuracy in most scenarios. The paper proposes potential strategies for incorporating additional predictive features to further refine the accuracy of these models, emphasizing the promise of machine learning in advancing indoor comfort optimization.**

Keywords*—Heating; building energy management; thermal comfort; IoT; Support Vector Machine; Random Forest

I. Introduction

Optimizing built environments for human habitation crucially involves predicting thermal comfort, a significant challenge intensified by climate change. As climate change escalates, extreme weather events become more frequent and severe, heightening the need for effective management of indoor thermal conditions. The importance of accurately predicting thermal comfort is underscored by its substantial impact on human health, productivity, and overall well-being. Inadequate thermal environments, characterized by excessive heat or cold, can result in discomfort, fatigue, and health complications, adversely affecting an individual's quality of life and reducing productivity in various environments such as workplaces, educational institutions, and homes.

Moreover, the economic consequences of neglecting thermal comfort are significant. Suboptimal indoor climates lead to heightened energy consumption as occupants frequently use heating or cooling systems to alleviate discomfort. This increased reliance on HVAC (Heating, Ventilation, & Air Conditioning) systems not only results in higher utility bills but also contributes to environmental strain. Consequently, there is an urgent need to develop reliable predictive models that can accurately forecast occupants' thermal comfort preferences under varying environmental conditions and architectural designs. Such models must incorporate a range of factors, including ambient temperature, humidity levels, clothing insulation, metabolic rates, and individual preferences, to deliver precise assessments of thermal comfort levels.

Addressing this imperative necessitates interdisciplinary collaboration among architects, engineers, psychologists, and data scientists. Integrating insights from environmental science, human physiology, and behavioral psychology is essential for developing effective predictive models. By harnessing advancements in sensor technology, data analytics, and machine learning algorithms, these models can be refined to provide real-time insights into the dynamics of thermal comfort. This enables building managers and occupants to optimize indoor environments, thereby enhancing well-being and promoting sustainable resource utilization.

To further understand the impact of thermal comfort, let's explore a detailed example. Temperature plays a critical role in human well-being, akin to how a rise in body temperature can signal illness, indicating that something is amiss. Similarly, room temperature significantly affects comfort and, consequently, our ability to function optimally.

Consider a scenario on a hot summer day: you begin to prepare for lessons or study in your room. To create a quiet environment, you close the door to block out noise and shut the window to keep out the heat. However, this action inadvertently leads to a reduction in airflow and available space, causing an increase in carbon dioxide levels as it accumulates in the room. This buildup of carbon dioxide decreases the oxygen levels, leading to a rise in room temperature. Consequently, you may start to feel distracted and lethargic, a direct result of the diminished air quality and increased warmth. This situation can be remedied by simply opening the door to improve ventilation. This action helps to balance the air quality and regulate the room's temperature, restoring a more comfortable and conducive environment for studying. This example underscores the importance of managing thermal comfort to maintain productivity and well-being in indoor spaces. The process of predicting thermal comfort involves the analysis of various factors, including temperature, humidity, air velocity, and clothing insulation.

Traditional methods, based on human comfort models, tend to be subjective and time-consuming.

In this paper, we explore the application of Support Vector Machines (SVM) and Random Forest algorithms for predicting thermal comfort in buildings, aiming to assess their effectiveness and compare their performance across different scenarios. The goal is to provide a thorough understanding of how these machine learning algorithms can aid building designers and facility managers in optimizing indoor environments and enhancing occupant comfort. Our research is structured around a series of hypotheses that guide the experimental design:

- Data Preparation: We hypothesize that removing NaN values and establishing a threshold for the minimum number of observations per feature will improve model accuracy by ensuring the data quality and relevance of the features used.
- Feature Encoding: We will evaluate the suitability of different encoding strategies as OneHotEncoder, LabelEncoder, and Word2Vec, to determine how best to handle categorical variables. The choice of encoder may significantly impact the performance of our models, depending on the nature of the data.
- Feature Selection: The SelectKBest model will be utilized to identify the most relevant features for predicting thermal comfort. This method is expected to highlight the variables most closely linked to the outcomes, thereby streamlining the modeling process.
- Feature Variants: Post feature selection, we will focus on variants of the filtered features that are closely associated with temperature prediction. This step is crucial for refining the model's focus and enhancing its predictive accuracy regarding thermal comfort.

Through this structured approach, we aim to validate our hypotheses and draw meaningful conclusions about the utility of the algorithms in the context of thermal comfort prediction, potentially offering actionable insights for the design and management of building environments. Both SVM and RF are supervised learning algorithms capable of being trained on datasets consisting of thermal comfort parameters alongside corresponding human feedback.

This paper is structured as follows: Section II provides a literature review, contextualizing our study within existing research. Section III describes the methodology employed, detailing the techniques used to analyze data. Section IV presents the findings of the study, supported by relevant tables and illustrations. Section V discusses the implications of these results. Finally, Section VI offers a conclusion, summarizing the key outcomes and proposing directions for future research.

II. Related Works

The Internet of Things (IoT) is revolutionizing the building management systems (BMS) industry, with forecasts predicting up to 125 billion connected devices by 2030. Despite these advancements, current BMS solutions often lack flexibility, especially in terms of feedback control options. To fully leverage the potential of IoT, adaptive control algorithms and modular architectures are being explored. The authors have introduced the "Semantically-Enhanced IoT-enabled Intelligent Control System" (SEMIoTICS) architecture, which enhances control system capabilities through redundancy and automatically adjusts configurations based on quality-of-service criteria [1]. Additionally, Model Predictive Control (MPC) is becoming increasingly popular for optimizing energy efficiency and comfort in HVAC systems. Nonetheless, the use of nonlinear models introduces significant computational challenges. In response, research has shifted towards linear controllers that utilize Jacobian linearization. A notable innovation in this field is a bilinear model for nonlinear MPC, designed to minimize energy costs while maintaining comfort levels. However, the computational intensity of this model poses challenges for its application in real-time control settings [2].

Another articles introduce a cutting-edge reinforcement learning (RL)-based approach for HVAC systems integrated into the Transactive Energy Simulation Platform (TESP). Utilizing the Deep Deterministic Policy Gradients (DDPG) algorithm, this method focuses on intelligent and granular control of HVAC operations by optimizing a cost function that seeks a balance between electricity costs and end-user dissatisfaction. The approach includes a market price prediction model developed using Artificial Neural Networks (ANN), a DDPG-based RL control algorithm, and both implementation and testing phases within the TESP framework [3]. Further the authors present a simulation model that incorporates both high-level and low-level controllers for a passenger car's air conditioning system. This model prioritizes occupant thermal comfort and the precise regulation of the physical system. They also introduce an Eco-Cooling Strategy employing MPC to optimize control inputs. The strategy is designed to achieve efficient cooling, reduce energy consumption, and maintain comfort. The simulation results underscore the critical role of control settings in effective thermal management [4].

Fuzzy logic-based models are increasingly utilized to control air conditioning systems at variable speeds, optimizing energy consumption and enhancing thermal comfort. Implemented in hardware such as microcontrollers, VLSI chips, and EDA tools, these controllers precisely manage temperature and humidity levels, effectively regulating fan and compressor speeds. Integrated with other techniques, they significantly improve energy efficiency and system performance [5]. Ref. [6] explores a range of HVAC control strategies, from classical PID controllers to advanced MPC. They address challenges in system simulation, control implementation, artificial intelligence integration, and energy savings, introducing the LAMDA controller to enhance real-time responsiveness and self-adjustment based on contextual information, further refining control accuracy and efficiency in HVAC systems.

The escalating energy consumption in commercial buildings, particularly through HVAC systems, has spurred increased research into optimizing energy efficiency. Despite advancements in HVAC technologies that have enhanced Demand Response (DR) programs, challenges remain in the application of model predictive control techniques. Recent studies have utilized machine learning methods, including Reinforcement Learning and Supervised Learning, to improve these systems [7]. Research in BEM has particularly focused on optimizing HVAC operations through various innovative approaches. Key developments include dynamic demand response controllers, mixed-integer nonlinear optimization models, stochastic programs, multi-objective optimization models, occupancy-based controllers, and incentive-based DR controllers. Additional methodologies explored include event-based control, mutual information frameworks, and MPC [8]. Furthermore, this paper [9] introduces a three-layered model designed for optimizing energy consumption in smart homes, incorporating data collection, prediction, and optimization phases. The model employs an Alpha Beta filter for reducing noise, Dynamic Evolving Neural Network (DELM) for dynamic parameter prediction, and fuzzy controllers for making refined control decisions. This integrated approach not only addresses static user parameters but also enhances both comfort and energy efficiency.

One study introduces an innovative model that omits gender and age factors in assessing thermal comfort, focusing instead on six key thermal factors: air temperature, mean radiant temperature, relative humidity, air speed, clothing insulation, and metabolic rate. This model, developed using Supervised Machine Learning, is tailored for application in a commercial building environment [10]. Another study conducted in Bilbao, Spain, at the KUBIK energy efficiency research facility, examines human thermal perception in response to external temperatures to enhance indoor comfort and reduce energy consumption [11]. Further research evaluates indoor thermal comfort using the Fanger method and adhering to ASHRAE Standard 55, emphasizing real-world conditions to promote well-being, productivity, and energy conservation in buildings [12].

Additionally, a study introduces a model based on multiple preferences for predicting group thermal comfort in shared spaces. This model integrates individual preferences and environmental parameters, segments occupants by Body Mass Index (BMI), predicts individual comfort zones, and adjusts settings to achieve group satisfaction [13]. Overall, optimizing thermal comfort in buildings is crucial for enhancing occupant well-being, productivity, and energy efficiency. Effective assessment models take into account variables such as air temperature, humidity, radiant temperature, and air speed, with the ASHRAE 55 standards providing guidelines for acceptable conditions.

Alternative models such as ANN, hybrid ANN-fuzzy systems, SVM, decision trees, and Bayes networks offer enhanced flexibility and accuracy in predicting thermal comfort [14]. Thermal comfort is a key component of indoor environmental quality, which can be categorized into static, adaptive, and data-driven models. Static models, like the Predicted Mean Vote (PMV), incorporate environmental and personal factors but have recognized limitations due to their lack of adaptability to individual responses. Adaptive models account for psychological and behavioral adaptations, enhancing their responsiveness to occupant preferences. Data-driven models leverage real-time data from sensor technologies for dynamic and responsive assessments of thermal comfort [15].

Further advancements are seen in the development of a building thermal model that utilizes low-resolution data from smart thermostats, improving accuracy and applicability across different seasons. This approach transforms traditional empirical models into a data-driven framework by using surrogate features to approximate internal heat gains. The model's design allows for implementation on either edge devices or cloud infrastructure, facilitating efficient data collection, model learning, and deployment [16].

Research continues to evolve with studies focusing on innovative cooling technologies such as Thermoelectric Air Ducts, with neural network models demonstrating high accuracy in predicting comfort parameters in dynamic settings. Understanding the interplay between climatic variables, occupant comfort, and system performance is fundamental [17]. Overall, the prediction of thermal comfort and optimization of energy use in buildings are critical for ensuring occupant satisfaction and achieving energy efficiency. Key factors influencing comfort include metabolic rate, clothing insulation, and air temperature.

Deep feedforward neural networks and reinforcement learning models are increasingly utilized to predict comfort levels, which is essential for monitoring and optimizing HVAC energy consumption in building operations [18]. A novel methodology employing machine learning, data mining, and statistical techniques has been developed to create predictive models for Combined Heat, Cooling, and Power (CHCP) systems. This methodology encompasses four stages: data preparation, data engineering, model building, and model evaluation. Data preparation includes retrieving failure events, labeling instances, and compiling a comprehensive dataset. Data engineering focuses on improving data representation through feature extraction and selection. The model building phase employs machine learning algorithms for various classification and regression tasks, while model evaluation assesses time to failure and other performance metrics to ensure the model's suitability [19].

Another innovative study explores thermal comfort in indoor environments through a novel approach called Relative Thermal Sensation (RTS). This method views thermal sensation as a continuous function of time, offering a more detailed understanding of human thermal perception. The study introduces a 3-point Relative Thermal Sensation Scale (RTSS) to collect real-time data on thermal sensations, capturing subtle changes that traditional discrete scales might overlook. Additionally, the research integrates RTS data with Absolute Thermal Sensation data from modified versions of the ASHRAE 7-point thermal sensation scale, enhancing the comprehensive understanding of thermal comfort [20].

Interpretable thermal comfort systems are being developed to enhance energy efficiency and occupant satisfaction in smart building environments. Traditional models such as the PMV often lack interpretability, posing challenges for building operators who need to understand the mechanisms influencing thermal comfort. To address this, researchers are integrating machine learning techniques that promote model transparency, such as Partial Dependence Plots (PDP) and SHAP values.

These tools allow operators to comprehend how environmental conditions affect human comfort and to evaluate the significance of various features under different scenarios. Furthermore, these interpretable machine learning algorithms are also being used to create surrogate models that replicate and potentially improve upon existing thermal comfort models, making them more accessible and actionable for building management [21].

III. Methodology

A. Dataset

The dataset, sourced from the ASHRAE and available on Kaggle [22], comprises 70 columns and 107,583 rows, containing data collected globally from 1995 to 2015. Initially, an examination of the dataset description led to a filtering process. This revealed that some columns contained sparse data. Consequently, a threshold was set at 60,000 rows; data points below this limit were discarded. Additionally, it was necessary to address missing values. Despite starting with 107,583 rows, the removal of rows with NaN values was essential to ensure data integrity.

Another analytical approach considered was the use of the Interquartile Range (IQR) method to identify and eliminate outliers, further refining the dataset's quality (see Fig.1).

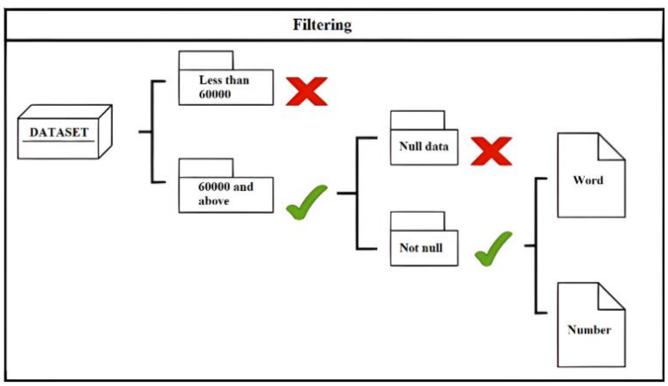

Fig. 1. Data filtering scheme.

Regarding the conversion of text data to numeric form, as shown in Fig. 2, two encoding options were evaluated: LabelEncoder and OneHotEncoder. The decision to proceed with OneHotEncoder was based on its superior performance in preliminary results [23], effectively transforming categorical text data into a usable format for machine learning models.

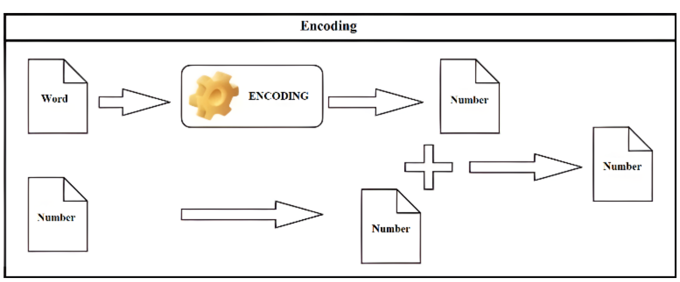

Fig. 2. Encoding scheme for the conversion of text data to numeric form.

In the feature selection process, as shown in Fig. 3, two methods were considered: using the SelectBest library or selecting based on correlation with a predefined threshold. The chosen method was to use correlations, specifically setting a boundary above 50% to determine relevant features. The final set of features selected includes Age, Clothing insulation (Clo), Sex, Metabolic rate (Met), Thermal preference, Year, Season, Köppen climate classification, Cooling strategy at the building level, City, Predicted Percentage of Dissatisfied (PPD), Air temperature (C), Outdoor monthly air temperature (C), Relative humidity (%), and Air velocity (m/s). This selection represents the culmination of extensive testing with various combinations of features, all of which will be detailed in the Experiments section of our study.

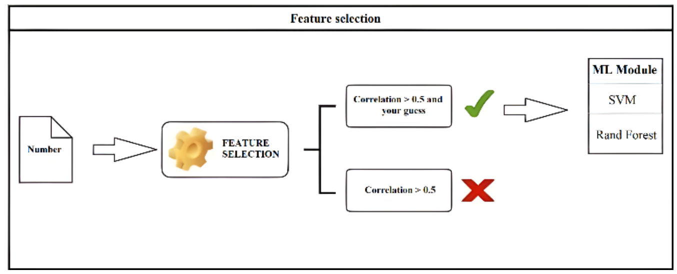

Fig. 3. Feature selection.

These features were instrumental in enhancing the predictive accuracy of our models. For the experimental setup, the dataset was divided into 80% for training and 20% for testing. Typically, thermal comfort ratings in the dataset ranged from 1 to 6. Another hypothesis tested was the conversion of these label values into integers. By reducing the range of thermal comfort ratings from six to three distinct categories, we observed a significant improvement in model accuracy. This transformation simplifies the model's classification task, enabling more precise predictions.

B. Inter Quartile Range (IQR)

The Interquartile Range (IQR) is a measure of statistical dispersion that is calculated as the difference between the third quartile ($Q3$) and the first quartile ($Q1$) of a dataset. Mathematically, it is defined as:

$$IQR = Q3 - Q1 \tag{1}$$

where $Q1$ is the median of the lower half of the dataset and $Q3$ is the median of the upper half of the dataset. It is particularly useful in identifying and dealing with outliers, which are data points that significantly differ from the rest of the dataset. Here's how the IQR is calculated and how it can be used to remove outliers:

1) Calculation of IQR:

- Firstly, you need to arrange your dataset in ascending order.
- Then, find the median of the dataset, which is the middle value when the data is sorted. If the dataset has an odd number of observations, the median is the middle value. If it has an even number of observations, the median is the average of the two middle values.

- Divide the dataset into two halves at the median. The lower half contains all the values less than or equal to the median, and the upper half contains all the values greater than or equal to the median.
- Find the median of each half. This gives you the first quartile (*Q1*) and the third quartile (*Q3*) of the dataset, respectively.
- The *IQR* is then calculated as the difference between *Q3* and *Q1: IQR = Q3 - Q1.*

2) Identifying outliers using IQR:

- Outliers can be detected using the IQR method by considering values that lie below $Q1 - 1.5 \times IQR$ or above $Q3 + 1.5 \times IQR$. These values are considered to be significantly different from the rest of the dataset.
- Values below $Q1 - 1.5 \times IQR$ or above $Q3 + 1.5 \times IQR$ are commonly referred to as lower and upper bounds, respectively.
- Any data points falling outside these bounds can be considered outliers.

3) Removing outliers using IQR:

- Once outliers are identified using the IQR method, you can choose to remove them from the dataset to improve the robustness of your analysis or model.
- Outliers can be removed by filtering the dataset to exclude any observations that fall outside the lower and upper bounds defined by $Q1 - 1.5 \times IQR$ and $Q3 + 1.5 \times IQR$, respectively.
- After removing outliers, the dataset may be more representative of the underlying distribution and less influenced by extreme values.

4) Considerations:

- While the IQR method is effective in identifying and removing outliers, it's important to exercise caution and consider the context of the data.
- Outliers may sometimes carry valuable information or be indicative of rare but important events. Therefore, the decision to remove outliers should be made judiciously based on the specific goals of the analysis or model.
- Additionally, the choice of the multiplier (1.5 in the conventional method) used to define the bounds can be adjusted depending on the desired level of sensitivity to outliers.

In summary, the *IQR* is a useful statistical measure for assessing the spread of a dataset and identifying outliers. By calculating the *IQR* and defining bounds based on it, outliers can be effectively detected and removed, leading to a more robust analysis or model.

C. Applied Algorithms

SVM is a robust supervised machine learning algorithm well-suited for both classification and regression tasks. In thermal comfort prediction, SVM is employed to delineate the complex interrelationships between various environmental factors—like temperature, humidity, and air velocity—and human thermal comfort responses. The algorithm focuses on maximizing the margin between classes in classification tasks or minimizing the error in regression, all while effectively controlling for overfitting. By training on labeled datasets that encapsulate environmental conditions and corresponding thermal comfort ratings, SVM learns to accurately predict thermal comfort levels based on specific environmental inputs.

Random Forest is another versatile machine-learning algorithm capable of handling both classification and regression challenges. It operates on an ensemble learning principle, utilizing multiple decision trees to construct a more accurate and robust model, as shown in Fig. 4. The process involves extensive data preparation, including cleaning, handling missing values, and appropriate transformations to fit the model. Random Forest uses random sampling to select data subsets for training each tree, employs recursive partitioning for tree creation, and integrates a voting mechanism to aggregate the predictions from various trees. This method is particularly effective in modeling nonlinear relationships and interactions among different environmental variables. Random Forest's ability to generate reliable predictions is enhanced by its ensemble approach, which provides a comprehensive view of thermal comfort across varying conditions.

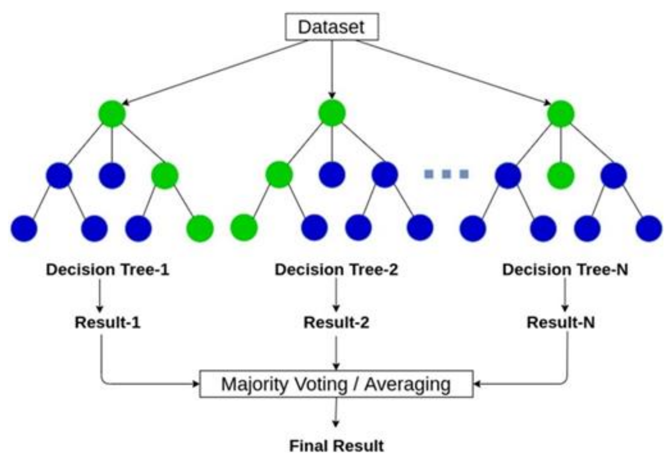

Fig. 4. Multiple decision trees of the Random Forest algorithm.

Both SVM and Random Forest are adept at capturing the nuanced dynamics between environmental parameters and thermal responses, making them invaluable for predicting thermal comfort in diverse settings. These models stand out for their robustness against overfitting, ensuring consistent reliability across different datasets and environmental scenarios. While SVM offers clear decision boundaries facilitating easier interpretation of the factors influencing thermal comfort, Random Forest provides insights into feature importance through its aggregated decision trees, although individual tree interpretations are less straightforward.

The flexibility of SVM and Random Forest models allows for the accommodation of various data types, making them ideal for integration with different environmental sensors and monitoring systems in thermal comfort assessment. An innovative approach within this domain is utilizing the 'Thermal preference' column as an alternative predictive variable instead of the conventional 'Thermal comfort' scale, moving away from traditional models that categorize comfort into six distinct levels to a more simplified three-level scale, which could potentially streamline the prediction process and enhance model accuracy.

D. Integration with IoT

The IoT component of the system is integral to enhancing building management by deploying a comprehensive network of sensors throughout the facility. These sensors are designed to monitor a variety of environmental conditions in real-time, including temperature, humidity, CO2 levels, and occupancy rates. The data collected by these IoT sensors is then transmitted to a central server, where it is stored and analyzed. For efficient and reliable data transfer, wireless communication protocols such as Wi-Fi, Bluetooth, or LoRaWAN are utilized.

As part of the system design of the controller, a thorough selection of hardware components and parameters was conducted. The designed printed circuit board (PCB) features include:

- A PCB thickness of 1.5 mm;
- A copper foil thickness of 35 μm;
- Glass epoxy laminated with foil;
- Epoxy-urethane varnish;
- Minimum conductor width and spacing of 0.3 mm, with a power bus width of 0.4 mm;
- Minimum hole diameter of 0.352 mm, with mounting hole diameter of 2 mm;
- Geometric dimensions of the board: 50.7 mm × 39.1 mm × 6.81 mm.

Following the selection of the electronic base and PCB parameters, a Raspberry Pi compatible topology was developed using DipTrace, as shown in Fig. 5. All components were positioned as closely as possible to minimize the board's size. Metallized holes were created in the corners of the board for mounting in the enclosure. The connections for power sources were placed on the left side of the board, and the connection for the battery was located on the right side.

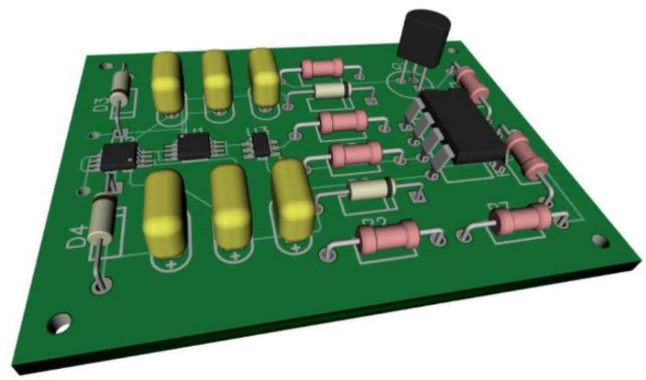

Fig. 5. Schematic of the IoT controller developed using DipTrace software.

The AI models within the system leverage this real-time data to continuously refine their predictions and immediately adjust the building's HVAC system to achieve optimal thermal comfort. A key feature of this setup is its feedback loop mechanism, which plays a critical role in maintaining desired thermal conditions. The AI algorithms actively process the incoming data from the IoT sensors and either make recommendations or directly control the HVAC system's operations, as shown in Fig. 6.

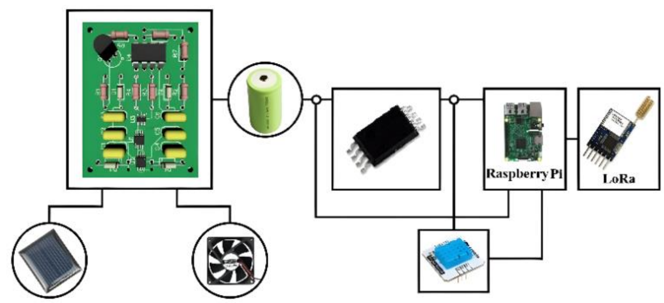

Fig. 6. General design of the system.

The device, powered by a rechargeable battery (referred to as the "slave device" in our model), collects data from sensors and sends this data to the master device. In our case, we use temperature and humidity sensors, which allow for monitoring in the environment to solve specific tasks. The only requirement for using this topology is that all slave devices must be no more than 100 meters away from the master device.

The topology of this model implies that each slave device is only aware of the master device's existence and there is no data transmission between two slave devices. This arrangement eliminates one of the main drawbacks: unreliable communication between two devices. The master device, in turn, is connected to the global network, and therefore, it only structures and redirects the data to the final destination—a database. There is a possibility that the master device may fail. However, nothing prevents connecting two master devices, through which sensor data is transmitted, and writing it into backup databases.

For instance, if the system detects any deviations from set comfort levels, it is programmed to make necessary adjustments to temperature, humidity, or airflow. This dynamic adjustment ensures that thermal comfort is not only achieved but sustained, adapting to both environmental changes and occupancy patterns within the building. A Raspberry Pi connected to a LoRa module serves as the master device. The role of the slave device is performed by a system composed of a microcontroller, a LoRa module, and sensors, all powered by a rechargeable battery. The collection of analog values produced by temperature and humidity sensors is facilitated by an integrated analog-to-digital converter. A fully charged battery can support the operation of the devices for up to 30 days.

The algorithm is implemented on the Raspberry Pi, which emulates the operation of a microcontroller. This device continuously listens on the 866 MHz frequency, which is used for transmitting data from the sensors. Selecting a suitable and efficient microcontroller will be part of future research. Managing comfortable environmental levels can also be controlled through an application as shown in Fig. 7. The interface allows us to easily manage the environment of a room. On the display, it can be seen current room conditions including temperature, humidity, and comfort level. To customize these settings, we can use the "Adjust Settings" section on the right, and the "Apply Settings" button to apply the new conditions.

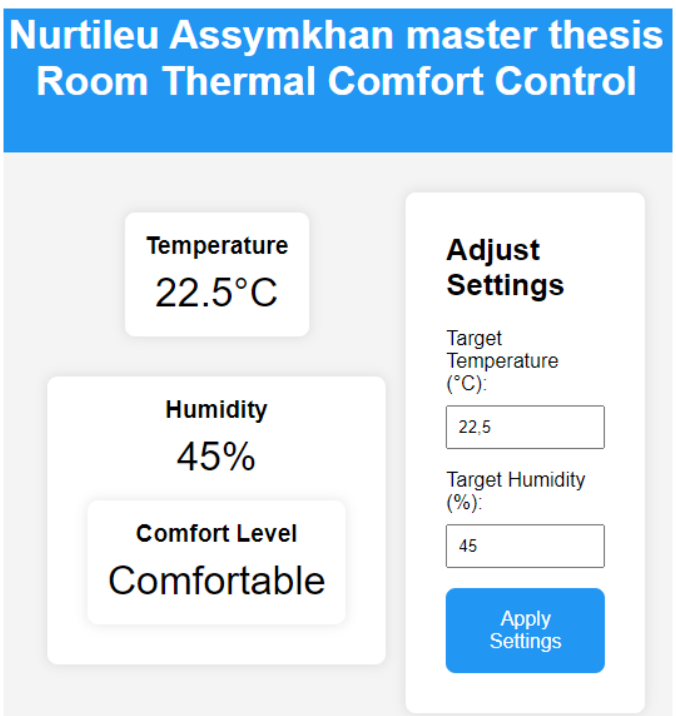

Fig. 7. GUI of the managing application.

IV. RESULTS

After an initial filtering process, our dataset was reduced from 70 to 21 columns. We continued to refine our feature selection by using correlations and deliberately avoided incorporating Fanger's features. Further filtration using both correlation analysis and the SelectKbest model, which assists in identifying the most impactful features, led us to define three distinct sets of features:

- First Set (17 features): Age, Sex, Metabolic rate (Met), Thermal preference, Thermal sensation, Clothing insulation (Clo), Subject's height (cm), Subject's weight (kg), Year, Season, Köppen climate classification, Building type, Cooling strategy at building level, Air temperature (C), Outdoor monthly air temperature (C), Relative humidity (%), and Air velocity (m/s).
- Second Set (9 features): Age, Sex, Met, Clo, Year, Season, Air temperature (C), Relative humidity (%), Air velocity (m/s).
- Third Set (15 features): Age, Clo, Sex, Met, Thermal preference, Year, Season, Köppen climate classification, Cooling strategy at building level, City, Predicted Percentage of Dissatisfied (PPD), Air temperature (C), Outdoor monthly air temperature (C), Relative humidity (%), Air velocity (m/s).

Following the feature selection, our dataset consisted of 17 columns and 6,765 rows. In the initial modeling phase, we utilized all 17 features, which yielded unsatisfactory results. Subsequent iterations with 9 and then 15 of the 17 features also failed to significantly improve outcomes. These iterations allowed us to test our hypotheses; notably, the IQR method enhanced model accuracy by approximately 3-4%, and the method of reducing label values improved accuracy by 20-23%.

Adjusting the parameters of our models led to more promising configurations. For the SVM model, optimal settings were identified as a radial basis function (RBF) kernel with gamma set to 0.001 and C set to 3. For the Random Forest model, the best parameters were found to be 300 estimators with a maximum depth of 15. These parameters maximized accuracy.

Additionally, comparing the impact of using LabelEncoder versus OneHotEncoder on the dataset revealed a difference in performance of 2-4%. This discrepancy influenced our decision to favor OneHotEncoder. Our tests on data standardization, using both the StandardScaler and MinMaxScaler, indicated that standardization did not significantly alter the accuracy, which remained relatively stable. Tables I, II, and III below present the initial results of our prediction efforts, illustrating the performance of each feature set and modeling approach:

TABLE I. ITERATION OF 17 FEATURES

Model	Accuracy	Precision	Recall	F1 score
SVM	0.509	0.451	0.509	0.436
RF	0.543	0.505	0.543	0.5

TABLE II. ITERATION OF 9 FEATURES

Model	Accuracy	Precision	Recall	F1 score
SVM	0.507	0.461	0.507	0.438
RF	0.526	0.513	0.526	0.49

TABLE III. ITERATION OF 15 FEATURES

Model	Accuracy	Precision	Recall	F1 score
SVM	0.533	0.448	0.533	0.433
RF	0.54	0.475	0.539	0.482

Based on the initial results, we further pursued enhancing model accuracy by employing the hypotheses formulated earlier in our study. The implementation of the IQR method was a particular focus, aimed at refining the data by removing outliers, which are often a source of prediction error. Tables IV, V, and VI below display the outcomes of applying the IQR method, which has further streamlined the model training process. These tables illustrate the effect of this technique on the overall performance of the models:

TABLE IV. ITERATION OF 17 FEATURES WITH *IQR*

Model	Accuracy	Precision	Recall	F1 score
SVM	0.522	0.44	0.522	0.441
RF	0.548	0.517	0.548	0.504

TABLE V. ITERATION OF 9 FEATURES WITH *IQR*

Model	Accuracy	Precision	Recall	F1 score
SVM	0.507	0.44	0.383	0.424
RF	0.52	0.501	0.52	0.479

TABLE VI. ITERATION OF 15 FEATURES WITH *IQR*

Model	Accuracy	Precision	Recall	F1 score
SVM	0.563	0.539	0.563	0.425
RF	0.57	0.494	0.57	0.5

Building on the improvements, which enhanced model accuracy by approximately 2-5%, our next step involves reducing label values to further increase the accuracy. This simplifies the output space of the model, potentially making it easier for the algorithms to distinguish between different states of thermal comfort. The Tables VII, VIII, IX shows the result of this approach:

TABLE VII. ITERATION OF 17 FEATURES WITH REDUCING LABELS

Model	Accuracy	Precision	Recall	F1 score
SVM	0.715	0.644	0.715	0.614
RF	0.744	0.708	0.744	0.704

TABLE VIII. ITERATION OF 9 FEATURES WITH REDUCING LABELS

Model	Accuracy	Precision	Recall	F1 score
SVM	0.688	0.598	0.688	0.569
RF	0.699	0.657	0.699	0.645

TABLE IX. ITERATION OF 15 FEATURES WITH REDUCING LABELS

Model	Accuracy	Precision	Recall	F1 score
SVM	0.78	0.608	0.78	0.683
RF	0.78	0.719	0.78	0.727

We utilized Random sampling to select subsets of the dataset for training individual decision trees within our Random Forest model. By integrating strategies such as feature reduction, IQR, and Random sampling, we have enhanced the construction and performance of our decision trees. These trees are built using recursive partitioning that methodically splits the data into increasingly specific subsets. This splitting is based on the feature values that most effectively differentiate the categories of the target variable.

The process is further refined through selective feature selection, which concentrates on the most impactful variables. This allows the model to focus on the data elements that are most predictive of the outcomes, significantly enhancing the overall performance of the model. These integrations contribute to a more efficient predictive tool, suitable for complex scenarios in smart building environments. After incorporating the feature-reduced model, further simplifying the feature space, we observed the following results, as in Tables X, XI, XII:

TABLE X. ITERATION OF 17 FEATURE-REDUCED LABELS AND *IQR*

Model	Accuracy	Precision	Recall	F1 score
SVM	0.726	0.598	0.726	0.621
RF	0.733	0.678	0.733	0.688

TABLE XI. ITERATION OF 9 FEATURE-REDUCED LABELS AND *IQR*

Model	Accuracy	Precision	Recall	F1 score
SVM	0.706	0.498	0.706	0.584
RF	0.717	0.668	0.717	0.653

TABLE XII. ITERATION OF 15 FEATURE-REDUCED LABELS AND *IQR*

Model	Accuracy	Precision	Recall	F1 score
SVM	0.835	0.697	0.835	0.76
RF	0.821	0.738	0.821	0.766

The implications of these findings are significant, especially in the context of predictive accuracy in environmental modeling for predicting thermal comfort levels in smart building systems. The Receiver Operating Characteristic (ROC) curves graph, presented in Fig. 8, provide a visual comparison of the performance of two machine learning models: SVM and Random Forest (RF). These curves are essential tools in evaluating the models by plotting the True Positive Rate (sensitivity) against the False Positive Rate (1-specificity) at various threshold settings. The area under the curve (AUC) serves as a summary measure of the model's ability to discriminate between positive and negative classes.

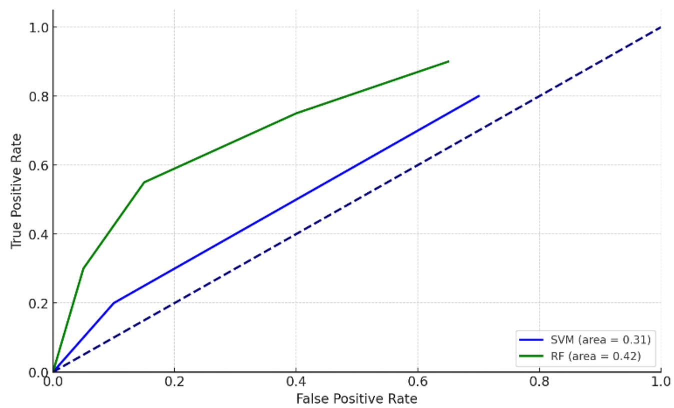

Fig. 8. ROC comparison for both the SVM and RF.

In this analysis, the SVM model demonstrates an AUC of 0.72, while the RF model exhibits a slightly superior AUC of 0.84. This suggests that the RF model has a better overall performance in distinguishing between the classes under study, likely due to its ensemble nature, which typically provides a more robust prediction by averaging multiple decision processes.

The following boxplots, depicted in Fig. 9, below show the distribution of cross-validation accuracy scores for SVM and RF models across various feature sets and conditions. The SVM exhibits a broader range of accuracy variations, especially with the 15-feature set. The increase in accuracy when reducing label values suggests that SVM benefits significantly from a simplified output space, potentially due to reduced complexity in the decision boundary formation. The RF shows tighter accuracy distributions and higher median accuracies across all feature sets, indicating better stability and robustness. The performance improvements with label reduction demonstrate RF's effectiveness in handling more straightforward, cleaner data. Further exploring more sophisticated data preprocessing techniques such as feature scaling, transformation (like log transformation for skewed data), or anomaly detection methods could help to handle outliers more dynamically than just applying IQR.

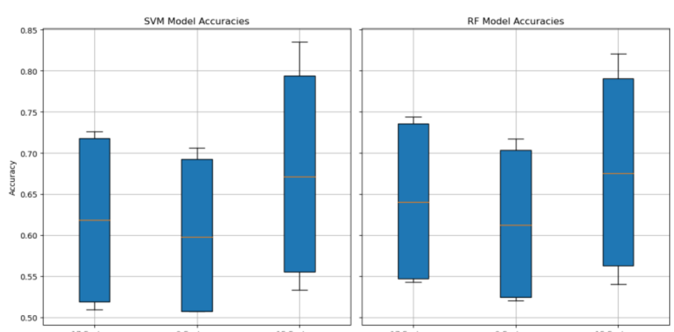

Fig. 9. Boxplots of cross-validation accuracy scores.

The implications of these findings are significant, especially in the context of predictive accuracy in environmental modeling, such as predicting thermal comfort levels in smart buildings. The higher AUC for the RF model indicates a higher likelihood of correctly classifying the thermal comfort levels as satisfactory or unsatisfactory, which is crucial for developing systems that can dynamically adjust to maintain or achieve desired comfort states. Moreover, the relatively lower performance of the SVM could be attributed to its sensitivity to the choice of kernel and the tuning of its parameters, which might not have been optimal in this scenario. These insights not only aid in selecting the appropriate model for deployment but also highlight the importance of model tuning and feature selection, reinforcing the need for ongoing model adjustment in practical applications to achieve the best outcomes.

V. Discussion

This research assesses the effectiveness of Random Forest and SVM algorithms across different feature sets in predicting thermal comfort and thermal preference. We introduced eight new features in our analysis, while seven features were consistent with those used in prior studies. When comparing the outcomes for guessing Thermal comfort versus Thermal preference, the performance gap between them was relatively narrow, ranging from 1-3%, with Random Forest generally exhibiting greater stability.

Specifically, in the scenarios where we tested sets with 9 and 15 features, alternative versions of our models initially led in performance. However, a significant shift occurred when we simplified the prediction scale from six to three Thermal comfort values, which resulted in our primary model configuration achieving superior results. This simplification appeared to enhance the model's ability to discriminate between different levels of comfort effectively.

Moving forward, we plan to incorporate additional features to refine the models' accuracy further. One promising candidate is Heart Rate Variability (HRV), which has potential implications for assessing physiological responses to thermal environments. Moreover, I am keen to explore the capabilities of neural networks and deep learning techniques, inspired by their success in related fields. My intention is to experiment with various advanced algorithms such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, Autoencoders, and Deep Belief Networks (DBNs). These methodologies, referenced in papers [18] and [10], will form the basis of future research efforts aimed at enhancing the predictive accuracy of thermal comfort models.

While the reduction of the thermal comfort scale from six levels to three markedly improved the models' discriminative capabilities, some researchers argue that this simplification might obscure subtle nuances in human comfort perception [24]. Critics suggest that while a simpler output space indeed facilitates more accurate classifications by reducing the complexity the model must manage, it potentially oversimplifies human experiences, which could be better captured with a more granular scale [25].

The introduction of the IQR method led to an approximate 3-4% improvement in model accuracy. However, it is essential to note that while IQR can effectively reduce outlier influence, it may bring limitations to valid extreme cases that are crucial for understanding the full spectrum of environmental impacts on thermal comfort [26]. The more substantial impact came from reducing label values, which boosted accuracy by 20-23%. This dramatic increase underscores the pivotal role of thoughtful statistical techniques in predictive model development. Yet, there remains a debate over whether such methods compromise the depth of data insights for the sake of model performance [27].

The system's architecture, employing a master-slave device setup, has shown efficient data management and transmission capabilities. Slave devices, powered by rechargeable batteries and strategically placed not more than 100 meters from a Raspberry Pi master device, efficiently transmit sensor data. However, some experts raise concerns about the scalability and maintenance of such setups in larger or more complex building environments [28]. Critics also question the reliance on Raspberry Pi for critical real-time data processing, citing potential limitations in processing power and storage compared to more robust computing solutions [29].

The capability for direct user interaction with the building management system through an application is hailed for its user-centered design, merging comfort with energy efficiency. Nevertheless, this approach raises questions about the trade-offs between user control and automated system efficiency, with some arguing that excessive user interaction might lead to less optimal energy use [30].

This study's significant contributions to environmental control and smart building management highlight the intersection of advanced computational techniques with practical IoT implementations. Yet, the ongoing exploration of new features and modeling techniques also points to a field that is constantly evolving, with ongoing debates about the best balance between accuracy, user experience, and system reliability [31]. We hope, these discussions are crucial as they push the boundaries of what smart building systems can achieve, ensuring they meet both current and future demands effectively.

VI. Conclusion

This study has explored the application of Random Forest and SVM algorithms to predict thermal comfort and preference, utilizing a refined feature set that integrates both newly introduced variables and established ones from prior research. The performance differential between predicting thermal comfort and thermal preference was relatively minimal, typically within 1-3%, with Random Forest demonstrating superior stability and robustness across varied feature sets. A pivotal enhancement in model performance was observed when the complexity of the thermal comfort scale was reduced from six to three levels, which notably improved the model's ability to discriminate between different comfort states more effectively.

To address the shortcomings in existing research, our paper showcases the advantages of the proposed techniques over traditional methods. For instance, the paper introduces clever adjustments to the thermal comfort prediction models, such as the reduction of the thermal comfort scale from six to three levels, which has shown to improve the accuracy of SVM and Random Forest models. This simplification not only enhances model precision but also makes these models more adaptable to various data types, which is essential for integrating environmental sensors in building management systems. Additionally, the novel use of the 'Thermal preference' column as a predictive variable instead of the standard 'Thermal comfort' scale offers a more streamlined and effective approach to predicting thermal comfort. By providing a thorough comparative analysis of these modifications against conventional methods, the paper highlights the practical implications in improving thermal comfort assessments.

While the simplification of the thermal comfort scale has yielded significant improvements, it also raises questions about the potential limitations of oversimplifying the nuances of human thermal perception. The use of the IQR method has also shown to improve model accuracy modestly; however, its tendency to remove valid extreme data points could limit understanding the broader impacts of environmental variables on thermal comfort. Moreover, the substantial increase in accuracy from reducing label values underscores the critical role of sophisticated statistical techniques in developing effective predictive models, though this approach has sparked debate regarding the depth and granularity of data interpretation.

The architectural design of our IoT-based system, featuring a master-slave configuration, has proven effective in data management and transmission, albeit with some concerns about scalability and dependency on limited-capability devices like the Raspberry Pi for critical processing tasks. Additionally, the system's design allowing direct user interaction via an application exemplifies a user-centered approach that harmoniously blends comfort with energy efficiency, though it also invites scrutiny over the potential for suboptimal energy usage due to excessive manual interventions.

Future enhancements are planned through the integration of additional predictive variables such as HRV, which holds promise for assessing physiological responses to varying thermal conditions. The potential of neural networks and deep learning will also be explored to leverage their proven capabilities in similar domains. Techniques such as CNNs, LSTM networks, and DBNs will be investigated to further refine the accuracy and efficiency of our models.

The contributions of this research to the fields of environmental control and smart building management are significant, illustrating the powerful synergy between advanced computational methods and practical IoT implementations. The ongoing development and refinement of these models push the boundaries of what smart building systems can achieve, ensuring they not only meet current demands but are also well-prepared for future challenges.

References

[1] G. M. Milis, C. G. Panayiotou and M. M. Polycarpou, "IoT-Enabled Automatic Synthesis of Distributed Feedback Control Schemes in Smart Buildings," in *IEEE Internet of Things Journal*, vol. 8, no. 4, pp. 2615-2626, 15 Feb.15, 2021, doi: 10.1109/JIOT.2020.3019662.

[2] S. I. Khather, M. A. Ibrahim, and A. I. Abdullah, "Review and performance analysis of nonlinear model predictive control—current prospects, challenges and future directions," *Journal Européen des Systèmes Automatisés*, vol. 56, no. 4, pp. 593-603, 2023. doi: 10.18280/jesa.560409.

[3] B. Liu, M. Akcakaya and T. E. Mcdermott, "Automated Control of Transactive HVACs in Energy Distribution Systems," in *IEEE Transactions on Smart Grid*, vol. 12, no. 3, pp. 2462-2471, May 2021, doi: 10.1109/TSG.2020.3042498.

[4] H. Wang, M. R. Amini, Q. Hu, I. Kolmanovsky and J. Sun, "Eco-Cooling Control Strategy for Automotive Air-Conditioning System: Design and Experimental Validation," in *IEEE Transactions on Control Systems Technology*, vol. 29, no. 6, pp. 2339-2350, Nov. 2021, doi: 10.1109/TCST.2020.3038746.

[5] Z. A. Shah, H. F. Sindi, A. Ul-Haq and M. A. Ali, "Fuzzy Logic-Based Direct Load Control Scheme for Air Conditioning Load to Reduce Energy Consumption," in *IEEE Access*, vol. 8, pp. 117413-117427, 2020, doi: 10.1109/ACCESS.2020.3005054.

[6] L. Morales Escobar, J. Aguilar, A. Garcés-Jiménez, J. A. Gutierrez De Mesa and J. M. Gomez-Pulido, "Advanced Fuzzy-Logic-Based Context-Driven Control for HVAC Management Systems in Buildings," in *IEEE Access*, vol. 8, pp. 16111-16126, 2020, doi: 10.1109/ACCESS.2020.2966545.

[7] D. Azuatalam, W.-L. Lee, F. de Nijs, and A. Liebman, "Reinforcement learning for whole-building HVAC control and demand response," *Energy and AI*, vol. 2, 2020, doi: 10.1016/j.egyai.2020.100020

[8] H. Mansy and S. Kwon, "Optimal HVAC Control for Demand Response via Chance-Constrained Two-Stage Stochastic Program," in *IEEE Transactions on Smart Grid*, vol. 12, no. 3, pp. 2188-2200, May 2021, doi: 10.1109/TSG.2020.3037668.

[9] A. Rabinowitz, F. M. Araghi, T. Gaikwad, Z. D. Asher, and T. H. Bradley, "Development and evaluation of velocity predictive optimal energy management strategies in intelligent and connected hybrid electric vehicles," *Energies*, vol. 14, 5713, 2021. doi: 10.3390/en14185713.

[10] F. H. Mohamed Salleh, M. b. Saripuddin and R. bin Omar, "Predicting Thermal Comfort of HVAC Building Using 6 Thermal Factors," *2020 8th International Conference on Information Technology and Multimedia (ICIMU)*, Selangor, Malaysia, 2020, pp. 170-176, doi: 10.1109/ICIMU49871.2020.9243466.

[11] N. Morresi *et al*., "Sensing Physiological and Environmental Quantities to Measure Human Thermal Comfort Through Machine Learning Techniques," in *IEEE Sensors Journal*, vol. 21, no. 10, pp. 12322-12337, 15 May15, 2021, doi: 10.1109/JSEN.2021.3064707.

[12] R. Widiastuti, J. Zaini, W. Caesarendra, D. Shona Laila and J. Candra Kurnia, "Prediction on the Indoor Thermal Comfort of Occupied Room Based on IoT Climate Measurement Open Datasets," *2020 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, Jakarta, Indonesia, 2020, pp. 40-45, doi: 10.1109/ICIMCIS51567.2020.9354277.

[13] Z. Zhang, B. Lin, Y. Geng, H. Zhou, X. Wu, and C. Zhang, "The effect of group perception feedbacks on thermal comfort," *Energy and Buildings*, vol. 254, 2022, doi: 10.1016/j.enbuild.2021.111603.

[14] F. Hani Mohamed Salleh and M. binti Saripuddin, "Monitoring Thermal Comfort Level of Commercial Buildings' Occupants in a Hot-Humid Climate Country Using K-nearest Neighbors Model," *2020 5th International Conference on Power and Renewable Energy*, Shanghai, China, 2020, pp. 209-215, doi: 10.1109/ICPRE51194.2020.9233145.

[15] M. Khalil, M. Esseghir and L. Merghem-Boulahia, "An IoT Environment for Estimating Occupants' Thermal Comfort," *2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications*, London, UK, 2020, pp. 1-6, doi: 10.1109/PIMRC48278.2020.9217157.

[16] X. Zhang, M. Pipattanasomporn, T. Chen and S. Rahman, "An IoT-Based Thermal Model Learning Framework for Smart Buildings," in *IEEE Internet of Things Journal*, vol. 7, no. 1, pp. 518-527, Jan. 2020, doi: 10.1109/JIOT.2019.2951106.

[17] K. Irshad, A. I. Khan, S. A. Irfan, M. M. Alam, A. Almalawi and M. H. Zahir, "Utilizing Artificial Neural Network for Prediction of Occupants Thermal Comfort: A Case Study of a Test Room Fitted With a Thermoelectric Air-Conditioning System," in *IEEE Access*, vol. 8, pp. 99709-99728, 2020, doi: 10.1109/ACCESS.2020.2985036.

[18] G. Gao, J. Li and Y. Wen, "DeepComfort: Energy-Efficient Thermal Comfort Control in Buildings Via Reinforcement Learning," in *IEEE Internet of Things Journal*, vol. 7, no. 9, pp. 8472-8484, Sept. 2020, doi: 10.1109/JIOT.2020.2992117.

[19] N. Smatov, R. Kalashnikov, and A. Kartbayev, "Development of context-based sentiment classification for intelligent stock market prediction," *Big Data Cogn. Comput.*, vol. 8, 51, 2024. doi: 10.3390/bdcc8060051.

[20] Z. Wang, H. Onodera and R. Matsuhashi, "Proposal of Relative Thermal Sensation: Another Dimension of Thermal Comfort and Its Investigation," in *IEEE Access*, vol. 9, pp. 36266-36281, 2021, doi: 10.1109/ACCESS.2021.3062393.

[21] N. Cibin, A. Tibo, H. Golmohamadi, A. Skou, and M. Albano, "Machine learning-based algorithms to estimate thermal dynamics of residential buildings with energy flexibility," *Journal of Building Engineering*, vol. 65, 2023, doi: 10.1016/j.jobe.2022.105683.

[22] C. Miller, B. Picchetti, C. Fu, and J. Pantelic, "Limitations of machine learning for building energy prediction: ASHRAE Great Energy Predictor III Kaggle competition error analysis," *Science and Technology for the Built Environment*, vol. 28, no. 5, pp. 610-627, 2022. doi: 10.1080/23744731.2022.2067466.

[23] S. Biswas and H. Rajan, "Fair preprocessing: towards understanding compositional fairness of data transformers in machine learning pipeline," in *Proc. 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2021)*, New York, USA, 2021, pp. 981–993. doi: 10.1145/3468264.3468536.

[24] Q. Zhao, Z. Lian, and D. Lai, "Thermal comfort models and their developments: A review," *Energy and Built Environment*, vol. 2, no. 1, pp. 21-33, 2021. doi: 10.1016/j.enbenv.2020.05.007.

[25] C. Aliferis and G. Simon, "Overfitting, underfitting and general model overconfidence and under-performance pitfalls and best practices in machine learning and AI," in *Artificial Intelligence and Machine Learning in Health Care and Medical Sciences*, Springer, 2024, ch. 10. doi: 10.1007/978-3-031-39355-6_10.

[26] N. Eslamirad, A. Sepúlveda, F. De Luca, and K. Sakari Lylykangas, "Evaluating outdoor thermal comfort using a mixed-method to improve the environmental quality of a university campus," *Energies*, vol. 15, no. 4, 1577, 2022. doi: 10.3390/en15041577.

[27] A. Hassan, "An effective ensemble-based framework for outlier detection in evolving data streams," *International Journal of Advanced Computer Science and Applications*, vol. 13, pp. 315-329, 2022. doi: 10.14569/IJACSA.2022.0131135.

[28] S.-Y. Chuang, N. Sahoo, H.-W. Lin, and Y.-H. Chang, "Predictive maintenance with sensor data analytics on a Raspberry Pi-based experimental platform," *Sensors*, vol. 19, 3884, 2019. doi: 10.3390/s19183884.

[29] A. Carvalho, C. Machado and F. Moraes, "Raspberry Pi Performance Analysis in Real-Time Applications with the RT-Preempt Patch," *2019 Latin American Robotics Symposium (LARS), 2019 Brazilian Symposium on Robotics (SBR) and 2019 Workshop on Robotics in Education (WRE)*, Rio Grande, Brazil, 2019, pp. 162-167, doi: 10.1109/LARS-SBR-WRE48964.2019.00036.

[30] A. M. Coutinho Demetrios, D. De Sensi, A. F. Lorenzon, K. Georgiou, J. Nunez-Yanez, K. Eder, and S. Xavier-de-Souza, "Performance and energy trade-offs for parallel applications on heterogeneous multi-processing systems," *Energies*, vol. 13, 2409, 2020. doi: 10.3390/en13092409.

[31] R. Almutairi, G. Bergami, and G. Morgan, "Advancements and challenges in IoT simulators: A comprehensive review," *Sensors*, vol. 24, 1511, 2024. doi: 10.3390/s24051511.