Vol. 01 (2025) 35-53, DOI: 10.61552/EC.2025.003

Energy Catalyst

www.jme.aspur.rs

Machine Learning-Based Predictive Models for Indoor Air Quality and Thermal Comfort: Bridging Sensor Data and Human Perception in Healthcare Facilities

Tajul Rosli Razak[a,d,*], Mohammad Hafiz Ismail[b], Hasila Jarimi[c,d], Mohd Shahrul Mohd Nadzir[e], Tianhong Zheng[d], Zhang Yanan[d], Emy Zairah Ahmad[f], Ubaidah Syafiq[c], Norasikin Ahmad Ludin[c], Noor Muhamad Abd. Rahman[g], Mohd Haikal Jamaludin[g], Saffa Riffat[d]

[a]*Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Malaysia,*
[b]*Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 02600 Arau, Malaysia,*
[c]*Solar Energy Research Institute. The National University of Malaysia, 43600 Bangi Selangor, Malaysia,*
[d]*Department of Architecture and Built Environment, The University of Nottingham, University Park, Nottingham, NG7 2RD, United Kingdom,*
[e]*Department of Earth Sciences and Environment, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia,*
[f]*Faculty of Electrical Technology and Engineering, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal Melaka, Malaysia,*
[g]*Engineering Services Division, Ministry of Health Malaysia, 62590, Putrajaya, Malaysia*

Keywords:

Indoor air quality, Thermal comfort
Machine learning, Predictive model
Artificial intelligence

* Corresponding author:

Tajul Rosli Razak
E-mail: tajulrosli@uitm.edu.my

Received: 8 February2025
Revised: 6 April 2025
Accepted: 16 May 2025

A B S T R A C T

Poor indoor air quality (IAQ) and inadequate thermal comfort continue to challenge healthcare facilities in Malaysia, especially in rural areas where poor ventilation and high humidity are prevalent. These environmental stressors negatively impact patient outcomes, staff well-being, and operational efficiency. Existing solutions often overlook the integration of subjective human perception with objective sensor data, resulting in limited adaptability and responsiveness. This study introduces a data-driven, human-centric framework that combines environmental sensor measurements with user-reported comfort feedback to develop predictive models for IAQ and thermal comfort. A comprehensive machine learning pipeline was implemented using four algorithms—Random Forest (RF), XGBoost, Artificial Neural Network (ANN), and Support Vector Machine (SVM)—to model both continuous IAQ indicators and categorical thermal preferences. Experimental results show that Random Forest achieved the best overall performance, with the lowest root mean squared error (RMSE = 14.35) in regression and the highest classification accuracy (87.5%) in predicting thermal preference.

Statistical validation confirmed that Random Forest and XGBoost performed similarly in regression, while Random Forest and SVM showed no significant differences in classification accuracy. These findings validate Random Forest as a robust and consistent model across both tasks. This study contributes a validated, AI-enhanced framework for intelligent environmental monitoring in healthcare settings, emphasizing the integration of subjective and objective data streams. The approach supports personalized, data-driven interventions and offers practical insights for facility managers, clinicians, and policymakers aiming to optimize indoor conditions. Future work will focus on scaling the dataset across diverse facilities and climates, enabling real-time deployment, and incorporating explainable AI techniques to enhance model transparency and stakeholder trust.

1. INTRODUCTION

Indoor air quality (IAQ) and thermal comfort are fundamental determinants of health and well-being, particularly crucial in healthcare facilities where environmental conditions directly influence patient recovery rates and staff performance. Research has shown that healthcare facilities require specialized attention to thermal comfort due to their unique operational requirements and vulnerable occupants [1]. In Malaysia and other tropical regions, healthcare facilities face significant challenges, including inadequate ventilation, excessive humidity, and air pollution, which deteriorate IAQ and thermal comfort [2]. These environmental challenges compromise operational efficiency and significantly impact patients' and healthcare workers' physical and mental well-being.

Recent advancements in environmental monitoring technologies have led to sophisticated sensor-based systems. However, current approaches predominantly focus on collecting isolated sensor data, creating a significant gap between objective measurements and occupants' subjective experiences. Research indicates that understanding individual differences in thermal comfort perception is crucial for developing practical solutions that enhance both occupant satisfaction and workplace productivity [3]. This is particularly relevant in healthcare settings, where thermal conditions can significantly affect operating room functionality and patient recovery environments [4].

The emergence of machine learning (ML) as a powerful analytical tool has revolutionized the processing of complex, multi-dimensional data, enabling sophisticated predictive modeling for real-time monitoring and decision-making. Recent studies have demonstrated ML's capability to improve building energy prediction and environmental monitoring [5]. The application of various ML techniques in indoor air quality detection has shown promising results in creating more responsive and efficient monitoring systems [6]. Despite these advances, systematic reviews indicate that integrating ML approaches in indoor air quality management, especially in healthcare settings, remains an emerging field with significant potential for development [7].

This research introduces an innovative methodology that leverages machine learning (ML) algorithms to integrate real-time environmental sensor data with human perception surveys, addressing critical gaps in indoor air quality (IAQ) and thermal comfort assessment within Malaysian healthcare facilities. Traditional approaches often fail to incorporate subjective human responses, limiting their ability to provide holistic and adaptive environmental solutions. By employing advanced ML techniques—including Artificial Neural Networks (ANN), Random Forest, XGBoost, and Support Vector Machines (SVM)—this study enhances predictive modeling for IAQ and thermal comfort monitoring. These algorithms were selected based on their proven capabilities in handling complex environmental datasets, with recent studies highlighting the complementary strengths of tree-based methods and neural networks in modeling building environmental conditions [8].

The study focuses on key environmental parameters, including temperature, relative humidity, CO_2 levels, particulate matter (PM2.5 and PM10), and volatile organic compounds (VOCs), which are analyzed in conjunction with established thermal comfort models [9]. The integration of objective environmental measurements with subjective comfort assessments provides a comprehensive dataset that strengthens predictive accuracy and supports proactive decision-making in healthcare settings.

The key contributions of this study are as follows:

(a) Development of a novel human-centric AI framework that fuses subjective comfort feedback with real-time environmental sensor readings to optimize indoor environment monitoring.

(b) Implementation and validation of ML-based predictive models, identifying Random Forest as the most robust and consistent model across regression and classification tasks for IAQ and thermal comfort prediction.

(c) Empirical evidence demonstrating the effectiveness of combining subjective and objective data to enhance environmental management strategies and inform human-centric, data-driven interventions.

This study offers valuable insights into improving IAQ and thermal comfort in healthcare environments by bridging the gap between human perception and sensor-based monitoring. The findings support intelligent, real-time environmental management strategies, empowering facility managers and policymakers to implement more effective interventions that enhance patient recovery, staff productivity, and overall well-being.

2. BACKGROUND

This section presents a background review of indoor air quality (IAQ), thermal comfort, and the application of machine learning-based predictive models. It provides the foundational understanding necessary for contextualizing the study's objectives and highlights the importance of integrating these domains for intelligent environmental monitoring.

2.1 Indoor Air Quality

Poor indoor air quality (IAQ) remains a significant challenge in Malaysian healthcare facilities, particularly in rural areas where inadequate ventilation and high humidity levels exacerbate these issues. Research indicates that healthcare facilities often experience elevated levels of indoor pollutants, including volatile organic compounds (VOCs) and particulate matter. For instance, Baudet et al. highlight that cleaning and disinfecting products, extensively used in healthcare settings to mitigate infection risks, are significant sources of VOCs, with prevalent ethanol and isopropanol [10]. Furthermore, Yusup et al. note that in many Asian countries, including Malaysia, buildings frequently lack adequate ventilation systems, leading to higher indoor pollution levels than outdoor environments [11]. This is particularly concerning in rural healthcare facilities, where natural ventilation may be insufficient due to structural limitations and environmental factors.

Humidity is critical in IAQ, especially in tropical climates like Malaysia's. High humidity can promote the growth of mold and fungi, which are detrimental to respiratory health. Rahman et al. discuss the lack of standardized regulations for acceptable levels of indoor fungi, which can vary significantly between countries [12]. This inconsistency poses a challenge for rural healthcare facilities in Malaysia, where resources for monitoring and remediation may be limited. Additionally, high humidity can exacerbate the effects of other pollutants, leading to an increased incidence of Sick Building Syndrome (SBS) among healthcare workers and patients [13].

The implications of poor IAQ in healthcare settings are profound, affecting not only patient safety but also the health of healthcare workers. Poor IAQ has been linked to various health issues, including respiratory problems and increased susceptibility to infections [14, 15]. The need for adequate ventilation systems is underscored by the findings of Hellgren et al., who emphasize that good IAQ is crucial for protecting patients and healthcare personnel from airborne contaminants [16]. Maintaining adequate IAQ becomes even more pronounced in rural areas, where access to advanced HVAC systems may be limited.

2.2 Thermal Comfort

Thermal comfort in Malaysian healthcare facilities, mainly in rural areas, remains a critical challenge due to inadequate ventilation and high humidity. The tropical climate of Malaysia, characterized by high temperatures and humidity levels, significantly impacts the thermal comfort of occupants in healthcare settings. This issue is exacerbated in rural areas where infrastructure may not support effective climate control measures.

Research indicates that the design and operation of healthcare facilities play a crucial role in achieving thermal comfort. For instance, Koutroumpi emphasizes the importance of understanding occupant adaptive behaviors in naturally ventilated wards, noting that humidity and airflow are critical factors often overlooked in existing thermal comfort assessments [17]. This is particularly relevant in rural healthcare facilities where natural ventilation may be the primary means of cooling, leading to discomfort during peak humidity periods.

Moreover, studies have shown that the thermal comfort levels in Malaysian healthcare facilities often do not meet the recommended standards. Chaloeytoy reports that the neutral temperature for hospitals in tropical climates like Malaysia is around 26.4°C, with a comfortable range between 25.3°C and 28.2°C [18]. However, many facilities struggle to maintain these temperatures, especially during the hotter months, leading to discomfort among patients and staff. Faraj's findings further support this, indicating that even with design improvements, achieving year-round thermal comfort remains challenging in Malaysian environments [19].

The impact of inadequate ventilation on thermal comfort cannot be overstated. While energy-efficient, Huang and Hwang highlight that natural ventilation strategies often fail to provide adequate thermal comfort in hot and humid conditions [20]. This is particularly concerning in rural healthcare facilities where mechanical ventilation systems may be lacking or poorly maintained. The interplay between ventilation rates and thermal comfort is critical; insufficient airflow can lead to overheating, adversely affecting the health and well-being of patients and healthcare workers [21].

2.3 Machine Learning – Predictive Model

Machine learning (ML) has emerged as a powerful tool for predicting indoor air quality (IAQ) and thermal comfort, leveraging its ability to analyze complex datasets and identify patterns that traditional methods may overlook. Various predictive models have been employed in this domain, including artificial neural networks (ANN), support vector machines (SVM), and ensemble learning techniques.

One of the prominent applications of ML in thermal comfort monitoring is ANN. For instance, Palladino et al. developed a simplified algorithm that combines outputs from different ANN models to predict indoor thermal conditions and the Predicted Mean Vote (PMV) index without requiring indoor monitored data [22]. Similarly, Tardioli et al. highlighted the effectiveness of advanced ensemble machine-learning methods for predicting thermal perception, suggesting that deep learning techniques can enhance prediction accuracy in naturally ventilated buildings [23]. Moreover, De et al. utilized ensemble-based methods to predict various thermal comfort metrics, demonstrating the versatility of ML in adapting to different environmental conditions [24].

In addition to ANN, SVM has been effectively applied to predict thermal comfort. Ju et al. employed SVM to predict thermal comfort in vehicles, considering factors such as temperature and humidity, which significantly influence passenger satisfaction [25]. This approach aligns with findings from Liu et al., who proposed personal thermal comfort models using wearable sensors, emphasizing the importance of individual feedback in enhancing prediction accuracy [26]. Furthermore, the integration of IoT with ML has been shown to facilitate real-time monitoring and control of thermal comfort, as demonstrated by Nascimento and Lopes, who utilized ML algorithms to analyze historical data for predictive modeling [27].

The rationale for employing machine learning approaches in IAQ and thermal comfort monitoring stems from their ability to handle large volumes of data and their adaptability to various environmental conditions. For instance, Wong et al. noted that ML models could forecast IAQ by analyzing historical profiles of IAQ parameters, thereby enabling proactive measures to improve public health [28]. Additionally, Huang et al. discussed the application of random forests for predicting thermal comfort in underground train carriages, illustrating the broad applicability of ML across different environments [29]. The ability of ML to continuously learn and improve from new data further enhances its utility in dynamic settings, such as smart buildings, where environmental conditions can change rapidly.

3. METHODOLOGY

This section outlines the methodology employed to develop a machine learning-based predictive model for indoor air quality (IAQ) and thermal comfort monitoring in a Malaysian healthcare setting. The approach is anchored in a real-world case study conducted at Pusat Hemodialysis Sungai Lembing, Pahang, providing both environmental sensor data and subjective feedback from occupants. Section 3.1 describes the case study site and data collection strategies, including the integration of environmental measurements with human perception data. Section 3.2 presents the machine learning framework, detailing each phase of the model development process—from data preparation and exploratory data analysis (EDA) to model training, evaluation, and statistical validation. This methodological structure ensures that the developed models are both data-driven and contextually relevant, enabling accurate, human-centric environmental monitoring in healthcare environments.

3.1 A Case Study: Pusat Hemodialysis Sungai Lembing, Pahang

Pusat Hemodialysis Sungai Lembing, as shown in Figure 1, is a public healthcare facility in the historic town of Sungai Lembing in Pahang, Malaysia. This clinic specializes in providing dialysis treatment to patients with chronic kidney disease, serving the local population, particularly those in rural and underserved areas.

Sungai Lembing, known for its lush natural surroundings and high humidity, presents unique environmental challenges that influence indoor air quality (IAQ) and thermal comfort. The clinic operates under these constraints, where maintaining optimal indoor conditions is critical for patient care and staff productivity. Poor IAQ and thermal discomfort can lead to adverse health outcomes, lower treatment efficacy, and diminished well-being for both patients undergoing hemodialysis and healthcare professionals.

(a) Pusat Hemodialysis Sungai Lembing, Pahang.

(b) Pusat Hemodialysis Sungai Lembing, Pahang.

Fig. 1. Pusat Hemodialysis Sungai Lembing, Pahang, Malaysia.

The Pusat Hemodialysis Sungai Lembing is a single-storey healthcare facility with an estimated total floor area of approximately 250–300 square meters, housing various functional zones including a dialysis treatment bay, multiple consultation and procedure rooms, a store, pantry, utility rooms, and staff rest areas. The main activity zone is the central Dialysis Bay, accommodating multiple patient stations where routine treatment occurs daily between 8:00 AM to 6:00 PM. The building is oriented with its main entrance facing southwest, as depicted in the architectural plan (see Figure 1- (a) and (b)). This case study offers a unique opportunity to explore integrating digital monitoring tools with machine learning models to address environmental challenges in rural healthcare facilities. It aims to evaluate how dynamic predictive modeling can improve IAQ and thermal comfort, ultimately enhancing the clinic's operational efficiency and patient care outcomes.

3.2 Machine learning-based predictive model for IAQ and thermal comfort monitoring

This section introduces a comprehensive machine learning (ML) framework for monitoring indoor air quality (IAQ) and thermal comfort. We outline multiple strategies in detail in the subsequent subsections to achieve this objective.

A. Data Collection

The initial step focuses on collecting datasets specific to this case study, targeting indoor air quality (IAQ) parameters and thermal comfort conditions within Pusat Hemodialysis Sungai Lembing, located in Pahang. This data collection effort is designed to comprehensively understand the clinic's environmental conditions, forming a solid foundation for analysis and developing predictive models. To achieve this objective, two distinct datasets were collected: Dataset 1 – Environmental Sensor Data, which includes objective measurements from installed sensors, and Dataset 2 – Human Perception Data, which captures subjective feedback from patients and staff regarding their comfort and IAQ perceptions.

a) Dataset 1 - Environmental Sensor Data

This dataset comprises objective indoor air quality (IAQ) measurements and thermal comfort parameters collected from environmental sensors installed within Pusat Hemodialysis Sungai Lembing. Key features include temperature, relative humidity, carbon dioxide (CO2) levels, particulate matter (PM2.5 and PM10), and volatile organic compound (VOC) concentrations. These continuous time-series measurements provide valuable insights into the clinic's environmental dynamics and serve as the backbone for real-time analysis and predictive modeling.

We used a custom-configured sensor device, AiRSYNC (as shown in Fig. 2), strategically deployed within the clinic to collect environmental data. AiRSYNC offers a cost-effective and reliable solution for monitoring indoor air quality (IAQ) and thermal comfort in healthcare settings. It measures essential IAQ parameters, including particulate matter (PM2.5 and PM10), volatile organic compounds (VOCs), carbon dioxide (CO2), and nitrogen dioxide (NO2), along with thermal comfort indicators such as temperature, relative humidity, and Mean Radiant Temperature (MRT). Furthermore, the system supports real-time monitoring and visualization of environmental conditions, providing actionable insights to optimize IAQ and thermal comfort. This proactive approach enhances patient care and staff well-being by ensuring a more comfortable and healthier indoor environment.

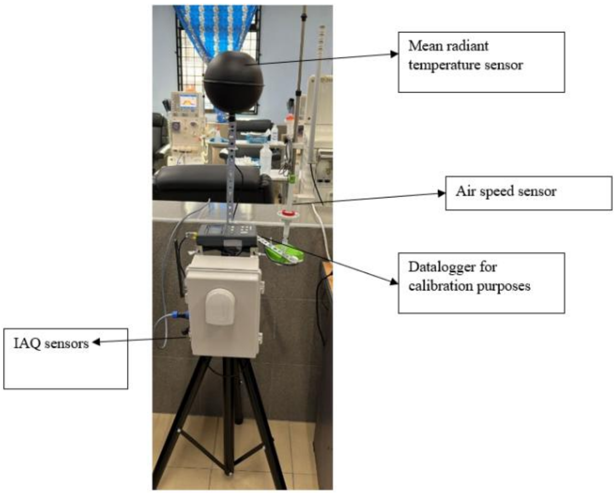

Fig. 2. The IAQ and Thermal Comfort Sensor (AiRSYNC) used in this study.

Fig. 3. The IAQ and Thermal Comfort Sensor (AiRSYNC) used in this study.

Figure 3 illustrates the installation process of AiRSYNC at Pusat Hemodialysis Sungai Lembing, located in Pahang. The image highlights key components of the system being set up in the healthcare facility, showcasing its integration into the existing infrastructure to enhance air quality and operational efficiency. To ensure accurate environmental monitoring, the AiRSYNC sensor unit was strategically installed within the Dialysis Bay, approximately 1.5 meters above the ground, at a central location away from direct windows, exterior doors, and heat-generating equipment (e.g., medical machinery, sterilizers). This placement was chosen to capture representative indoor conditions where patients and staff spend extended periods. The data collection occurred during the inter-monsoon period (April to June 2024), characterized by hot and humid tropical conditions with average outdoor temperatures ranging from 28°C to 34°C.

1	Time	CO2(ppm)	HCHO(ppb)	TVOC(ppm)	AirFlow(m/s)	PM2.5(ug/m3)	PM10(ug/m3)	AT('C)	MRT('C)	RH(%)
2	2024-11-06	740.25	2.2	2.7377008	0	1	2	28	28.25	75
3	2024-11-06	743.8125	2.4	2.62851406	0	1	2	28	27.5	75
4	2024-11-06	738.5625	2.4	2.58333333	0	1	2	28	27.75	75
5	2024-11-06	742.6875	2.4	2.66616466	0	1	1	28	27.5	75
6	2024-11-06	748.3125	2.2	3.04643574	0	1	2	28	28	75

Fig. 4. Time-stamp measurement dataset.

Dataset 1 (as shown in Figure 4) presents time-stamped measurements of environmental parameters collected through the AiRSYNC device to evaluate indoor air quality (IAQ) and thermal comfort conditions at Pusat Hemodialysis Sungai Lembing. A total of 14,216 time-stamped entries were recorded, providing a comprehensive and high-resolution profile of the indoor environment. Table 1 summarizes the key features captured in this dataset.

Table 1. Description of Environmental Parameters.

Parameter	Description
Time	When the measurement was taken.
CO_2 (ppm)	Amount of carbon dioxide in the air.
HCHO (ppb)	The level of formaldehyde, a harmful gas.
TVOC (ppm)	Total amount of indoor air pollutants.
AirFlow (m/s)	Speed of air movement.
PM2.5 (μg/m^3)	Small dust particles are in the air.
PM10 (μg/m^3)	Larger dust particles are in the air.
AT (°C)	Temperature of the air.
MRT (°C)	Heat from surrounding surfaces.
RH (%)	Humidity level in the air.

b) Dataset 1 - Environmental Sensor Data

This dataset encompasses subjective evaluations of IAQ, and thermal comfort gathered through surveys and feedback from patients and clinic staff. It includes qualitative and quantitative responses reflecting personal comfort levels, perceived air freshness, and overall satisfaction with the indoor environment. These data points offer a human-centric perspective, complementing the objective sensor data and facilitating the integration of human factors into the predictive model.

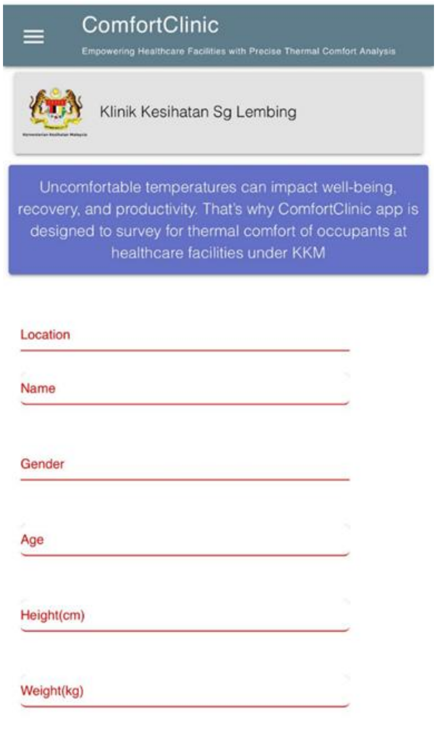

Fig. 5. The main page of ComfortClinic.

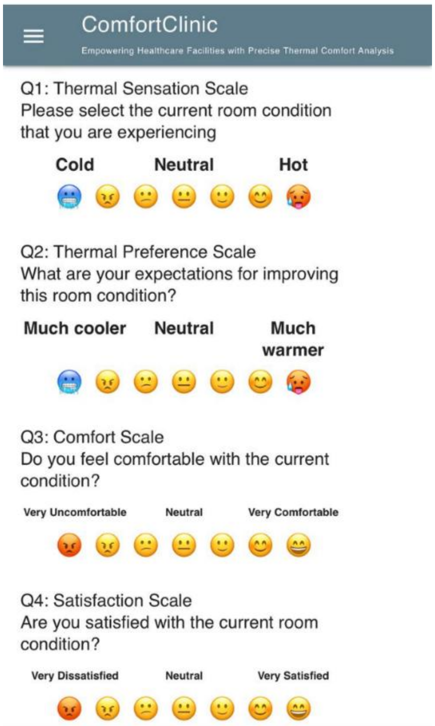

Fig. 6. Example Likert scale question using emoji.

To effectively capture the perceptions of both patients and staff at the clinic, we employed an interactive approach using a mobile application. Participants can easily access the survey by scanning a QR code, ensuring convenience and accessibility. To enhance user engagement, the survey incorporates emojis as response options for each question, providing a visually appealing and intuitive way for users to express their opinions and feedback, as shown in Figures 5 and 6.

ID	Time	Gender	Age	Weight	Height	Occupatio	Clothing Ite	Clothing La	Thermal Se	Thermal Pr	Comfort Sc	Satisfaction Scale	
288	11-06-24 8:02	Mariati	female	45	146	54	ppk	Long Sleev	1	-3	0	0	3
289	11-06-24 8:15	phoon yok	female	83	154	45	tidak beke	Light Shirt,	1	1	1	1	1
290	11-06-24 8:17	cheong kw	female	49	157	48	suri rumah	Light Shirt,	1	1	1	1	1
291	11-06-24 8:18	lai sin yin	female	80	160	51	tidak beke	Light Shirt,	1	1	1	1	1
292	11-06-24 8:19	siti. habiba	female	41	145	38	tidak beke	Light Shirt,	1	1	1	1	1

Fig. 7. Subjective data captured from patients and staff at Pusat Hemodialysis Sungai Lembing.

This dataset (as shown in Figure 7) comprises subjective data collected from patients and clinic staff at Pusat Hemodialysis Sungai Lembing. A total of 91 survey responses were gathered throughout the data collection period, offering valuable insights into individual experiences of thermal comfort and indoor air quality perception. Each response was timestamped and later synchronized with the nearest environmental sensor readings to enable integrated analysis. This synchronized dataset allows for meaningful correlation between subjective human feedback and objective environmental conditions, supporting the development of predictive comfort and air quality models. The key attributes of this dataset are detailed in Table 2.

Table 2. Description of Subjective Dataset.

Attribute	Description
ID	Unique code for each participant.
Time	Time when the response was recorded.
Name	Participant label (anonymized).
Gender	Participant's gender (e.g., male, female).
Age	Age in years.
Height	Height in centimeters (cm).
Weight	Weight in kilograms (kg).
Occupation	Participant's job or activity type.
Clothing Insulation (Icl)	Type of clothing worn, measured in Clo.
Thermal Sensation	How the participant feels (cold to hot scale).
Thermal Preference	Whether the participant wants it warmer, cooler, or no change.
Comfort Satisfaction Scale	Satisfaction with comfort (1 = yes, 0 = no).

B. Data Intergration

The integration of Dataset 1 (environmental sensor data) and Dataset 2 (human perception data) is a pivotal aspect of this study, enabling a holistic analysis of indoor air quality (IAQ) and thermal comfort at Pusat Hemodialysis Sungai Lembing. However, integrating these datasets poses challenges, mainly due to differences in their data collection methods and frequencies. To address this, the *Temporal Synchronization* method was employed, providing an effective solution for aligning the two datasets and ensuring compatibility for comprehensive analysis.

Temporal synchronization was a key methodology to seamlessly integrate Dataset 1 (sensor data) and Dataset 2 (human perception data). Both datasets were timestamped during the data collection phase to maintain precise chronological order and enable effective alignment.

Dataset 1, which captures environmental sensor readings (e.g., CO_2 levels, temperature, relative humidity, etc.), records data continuously at fixed intervals. Conversely, Dataset 2 comprises subjective survey responses from patients and staff gathered whenever participants complete the survey by scanning a QR code. This discrepancy in the timing and frequency of data collection necessitates a robust synchronization method.

A time-matching algorithm was used to align these datasets. The algorithm pairs each entry in Dataset 2 (survey responses) with the nearest corresponding timestamp in Dataset 1 (sensor readings). For example:

- If a patient submits a response at 10:02:15, the algorithm identifies and matches the closest available sensor data recorded around 10:02 in Dataset 1.
- This ensures that human perceptions are accurately associated with the environmental conditions experienced during their response.

This step is critical as it enables meaningful comparisons between objective environmental parameters and subjective comfort evaluations. Through this synchronization, the integrated dataset provides a reliable foundation for subsequent correlation analysis, predictive modeling, and deriving actionable insights into IAQ and thermal comfort conditions within the clinic.

ID	Time	Name	Gender	Age	Height	Weight	Occupatic	Clothing It	Clothing	Thermal	Therm.	Comfort	Satisfa	CO2(ppm)	HCHO(ppt	TVOC(ppn	AirFlow(m	PM2.5(ug/	PM10(ug/	AT(°C)	MRT(°C)	RH(%)
288	11/06/2024 08:02	Mariati	female	45	146	54	ppk	Long Sleev	1	-3	0	0	3	850.6875	8.6	1.378514	0	2	3	25	25	65
289	11/06/2024 08:15	phoon yol	female	83	154	45	tidak beke	Light Shirt	1	1	1	1	1	840.9375	4.4	1.126255	0	2	2	24	24.5	66
290	11/06/2024 08:17	cheong kv	female	49	157	48	suri ruma	Light Shirt	1	1	1	1	1	842.0625	6.4	1.118725	0.862621	0	1	24	24.75	66
291	11/06/2024 08:18	lai sin yin	female	80	160	51	tidak beke	Light Shirt	1	1	1	1	1	870.9375	7.4	1.126255	0.721439	1	2	24	24	66
292	11/06/2024 08:19	siti. habib	female	41	145	38	tidak beke	Light Shirt	1	1	1	1	1	868.3125	8.2	1.11496	0.591374	1	2	24	24	65
293	11/06/2024 08:20	wong cho	male	71	155	56	tidak beke	Light Shirt	1	1	1	1	1	848.625	6.8	1.141315	0.788139	2	3	24	24.25	64
294	11/06/2024 08:21	rosli oma	male	50	175	75	bekerja se	Light Shirt	1	1	1	1	1	815.4375	5.4	1.14508	0	1	1	24	24.5	64
295	11/06/2024 08:22	selamah	female	58	160	63	tidak beke	Long Sleev	1	1	1	1	1	863.25	4.2	1.141315	0.683642	2	3	24	24.75	64
296	11/06/2024 08:23	ashraq ya	male	38	173	98	bekerja se	Light Shirt	1	1	1	1	1	874.3125	3.6	1.12249	0.853728	1	2	24	24.25	63
297	11/06/2024 08:24	saemah y	female	63	150	48	tidak beke	Long Sleev	1	1	1	1	1	847.875	3.2	1.13002	0.662521	1	2	24	24.5	63
298	11/06/2024 08:25	goh check	male	50	160	58	tidak beke	Light Shirt	1	1	1	1	1	861.9375	2.8	1.156376	0.684754	1	2	24	23.75	62
299	11/06/2024 08:25	ang chin	male	64	160	64	tidak beke	Light Shirt	1	1	1	1	1	861.9375	2.8	1.156376	0.684754	1	2	24	23.75	62
300	11/06/2024 08:26	syed syan	male	44	180	87	ppp	Light Shirt	1	1	1	1	1	849.375	2.8	1.10743	0	0	1	24	24.5	62
301	11/06/2024 08:31	ASLINA BI	female	38	155	75	JURURAW	Light Shirt	2	0	0	0	0	888.75	2.8	1.13002	0.784804	1	2	24	24.5	62
302	11/06/2024 10:23	Rohaya bi	female	41	150	67	Jururawat	Light Shirt	1	-1	-1	0	0	857.0625	4.8	0.708333	0.576922	3	5	24	24.5	60
303	11/07/2024 08:13	Mariati	female	45	146	54	ppk	Long Sleev	1	0	0	3	3	863.25	9.8	1.156376	0.849281	1	1	24	24.75	68
304	11/07/2024 08:57	wan tima	female	72	150	49	suri ruma	Long Sleev	1	1	1	1	1	854.4375	10.8	0.915412	0.574699	0	0	24	24.75	62

Fig. 8. Representative sample of the merged dataset.

Figure 8 illustrates a representative sample of the merged dataset, which has been processed to incorporate temporal synchronization. This process focuses primarily on aligning the timestamps from the two separate datasets to ensure consistency and accurate integration. The resulting dataset ensures that subsequent analyses can be conducted without temporal mismatches or inconsistencies by applying temporal synchronization. For example, this enables accurate correlations, causality studies, or predictive modeling across the variables sourced from both datasets.

This merged and synchronized dataset is the primary foundation for further analysis, ensuring meaningful and reliable results. The alignment of temporal data also improves the interpretability of findings, as all comparisons and conclusions are drawn from data aligned to the same timeline.

C. Exploratory Data Analysis

The initial data quality assessment included checks for missing values, outliers, and inconsistencies to ensure reliability for subsequent analysis. The results confirmed that the dataset was complete, with no missing values or irregularities, and the summary statistics indicated it was clean and ready for further processing.

```
print(missing_values)
              ID             Time           Location               Name             Gender
               0                0                  0                  0                  0
             Age           Height             Weight         Occupation     Clothing_Items
               0                0                  0                  0                  0
 Clothing_Layers Thermal_Sensation Thermal_Preference     Comfort_Scale Satisfaction_Scale
               0                0                  0                  0                  0
         CO2_ppm         HCHO_ppb           TVOC_ppm         AirFlow_ms          PM2.5_ugm3
               0                0                  0                  0                  0
       PM10_ugm3             AT_C              MRT_C         RH_percent
               0                0                  0                 91
```

Fig. 9. Check missing values using R programming.

The dataset was checked for missing values to ensure its readiness for analysis. Missing values were identified and removed to ensure data completeness. Character variables were systematically converted into a numerical format to facilitate compatibility with machine learning models requiring numeric input. The results showed no missing values were complete for most variables, such as *ID, Time, Age,* and environmental data like *CO2_ppm*. However, the *RH_percent* (relative humidity) column had 91 missing values, indicating the need for further preprocessing. This initial assessment confirms the dataset's overall cleanliness while identifying a specific area requiring attention for data imputation or removal before proceeding to analysis.

```
> print(summary_data)
       ID            Time             Location             Name               Gender            Age
 Min.   :288.0   Length:91          Length:91          Length:91          Min.   :27.00   Min.   :145.0
 1st Qu.:310.5   Class :character   Class :character   Class :character   1st Qu.:41.00   1st Qu.:150.0
 Median :333.0   Mode  :character   Mode  :character   Mode  :character   Median :58.00   Median :155.0
 Mean   :333.0                                                            Mean   :54.88   Mean   :157.2
 3rd Qu.:355.5                                                            3rd Qu.:69.00   3rd Qu.:160.0
 Max.   :378.0                                                            Max.   :83.00   Max.   :180.0
     Height         Weight            Occupation        Clothing_Items  Clothing_Layers   Thermal_Sensation
 Min.   :36.0   Length:91          Length:91          Min.   :1.000   Min.   :-3.0000   Min.   :-1.0000
 1st Qu.:49.0   Class :character   Class :character   1st Qu.:1.000   1st Qu.: 1.0000   1st Qu.: 1.0000
 Median :64.0   Mode  :character   Mode  :character   Median :1.000   Median : 1.0000   Median : 1.0000
 Mean   :64.1                                         Mean   :1.088   Mean   : 0.7912   Mean   : 0.8462
 3rd Qu.:75.0                                         3rd Qu.:1.000   3rd Qu.: 1.0000   3rd Qu.: 1.0000
 Max.   :98.0                                         Max.   :2.000   Max.   : 1.0000   Max.   : 2.0000
 Thermal_Preference Comfort_Scale    Satisfaction_Scale    CO2_ppm          HCHO_ppb          TVOC_ppm
 Min.   :-1.0000    Min.   :-1.000   Min.   :695.1      Min.   : 1.00   Min.   :0.2415   Min.   :0.0000
 1st Qu.: 1.0000    1st Qu.: 1.000   1st Qu.:798.8      1st Qu.: 3.80   1st Qu.:0.6989   1st Qu.:0.0000
 Median : 1.0000    Median : 1.000   Median :821.6      Median : 5.80   Median :0.8740   Median :0.0000
 Mean   : 0.9011    Mean   : 0.956   Mean   :820.4      Mean   : 6.11   Mean   :0.8394   Mean   :0.2508
 3rd Qu.: 1.0000    3rd Qu.: 1.000   3rd Qu.:850.1      3rd Qu.: 8.80   3rd Qu.:0.9324   3rd Qu.:0.5586
 Max.   : 3.0000    Max.   : 3.000   Max.   :888.8      Max.   :12.40   Max.   :1.3785   Max.   :0.8626
   AirFlow_ms       PM2.5_ugm3       PM10_ugm3          AT_C            MRT_C         RH_percent
 Min.   :0.000   Min.   :0.000   Min.   :23.00   Min.   :23.75   Min.   :60.00   Mode:logical
 1st Qu.:0.000   1st Qu.:1.000   1st Qu.:24.00   1st Qu.:24.25   1st Qu.:62.50   NA's:91
 Median :1.000   Median :1.000   Median :25.00   Median :24.75   Median :68.00
 Mean   :1.022   Mean   :1.604   Mean   :24.81   Mean   :24.99   Mean   :66.34
 3rd Qu.:1.000   3rd Qu.:2.000   3rd Qu.:25.50   3rd Qu.:25.75   3rd Qu.:69.00
 Max.   :7.000   Max.   :9.000   Max.   :27.00   Max.   :28.50   Max.   :72.00
```

Fig. 10. Check on distribution and outlier.

The dataset's summary statistics (as in Figure 10) were generated to provide an overview of the variables' distributions and key metrics. The output includes minimum, maximum, mean, median, and quartiles for numeric variables and the length, class, and mode for categorical variables. Variables like *Age, Height,* and *Weight* show reasonable ranges, with *Age* varying from 27 to 83 years and *Height* ranging from 145 to 180 cm. Environmental measurements, including *CO_2_ppm* and *PM2.5_ugm3*, display appropriate value distributions, with no extreme outliers detected. Additionally, thermal comfort-related metrics like *Thermal_Sensation, Comfort_Scale,* and *Satisfaction_Scale* range from -1 to 1, indicating uniform coding of these subjective perceptions. The summary also confirms that some columns, such as *RH_percent*, contain missing values (91 NAs), flagged earlier for preprocessing. These insights suggest that the dataset is mostly clean and ready for analysis, aside from addressing the missing values.

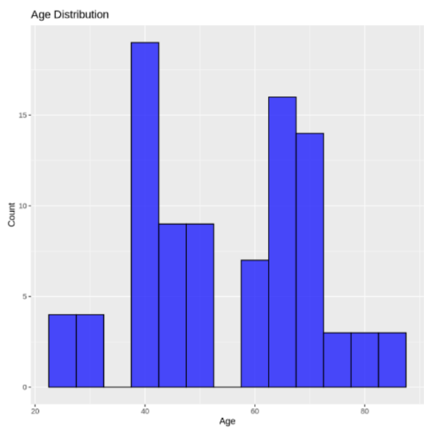

Fig. 11. Histogram of Age Distribution.

The histogram in Figure 11 shows the age distribution of individuals in the dataset. The x-axis represents age ranges, while the y-axis shows the count of individuals in each age group. The data suggests a bimodal distribution, with two prominent age groups centered around the 40–45 range, which has the highest frequency, and another between 55–60. A noticeable gap between these groups indicates fewer individuals in the 45–55 range. Additionally, minor counts are observed in the youngest age group (20–30) and older age brackets (70–80). This distribution suggests that the population sample may predominantly consist of middle-aged and early seniors.

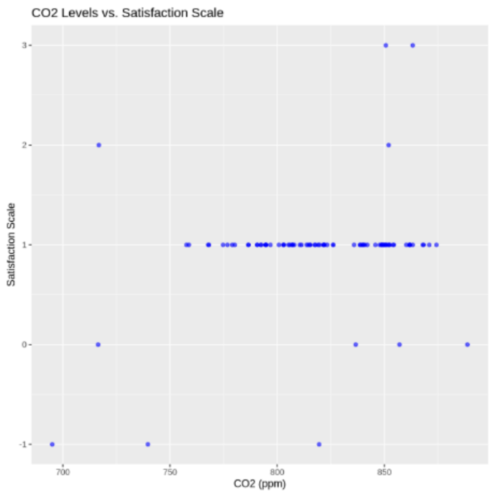

Fig. 12. Scatterplot relationship between CO2 and satisfaction scale.

The scatterplot in Figure 12 illustrates the relationship between *CO_2 levels (ppm)* and *satisfaction scale*, with *CO_2 levels* ranging from 700 to 850 ppm and satisfaction scores spanning from -1 to 3. Most data points cluster tightly at a satisfaction scale of 1 across the *CO_2* range, indicating this is the most common satisfaction level regardless of *CO_2* concentration. Outliers exist at other satisfaction levels, including a few points at scales -1, 0, 2, and 3, but these are sparse and do not show a consistent pattern with *CO_2 levels*. There is no strong relationship between *CO_2 levels* and the satisfaction scale. However, the dominance of a satisfaction scale of 1 may reflect a bias or trend in the data.

The correlation matrix of environmental parameters (see Figure 13) reveals several strong associations among key indoor environmental variables. Notably, *Mean Radiant Temperature (MRT)* and *Air Temperature (AT)* exhibit a strong positive correlation, confirming their mutual dependence in defining thermal conditions. Likewise, *PM2.5* and *PM10* concentrations are strongly correlated, which is expected given their shared particulate origin. Airflow also shows moderate positive correlations with both *PM2.5* and *PM10*, suggesting that increased air movement may be associated with the distribution or resuspension of particulate matter, rather than a reduction. Meanwhile, *CO_2*, *TVOC*, and *formaldehyde (HCHO)* display relatively weak correlations with other parameters, indicating that these pollutants may arise from different sources or follow independent dispersion patterns. These insights can support more focused interventions targeting specific environmental factors influencing indoor air quality and thermal comfort.

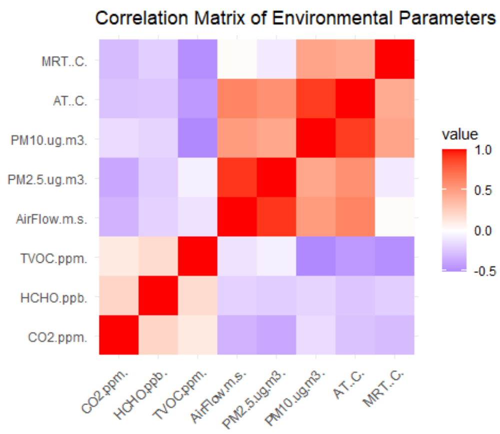

Fig. 13. Correlation matrix of environmental parameters measured at Pusat Hemodialysis Sungai Lembing.

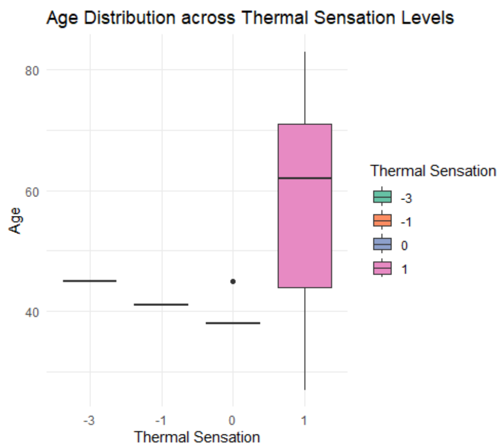

Fig. 14. Boxplot Age Distribution across Thermal Sensation Levels.

From the boxplot in Figure 14 depicting age distribution across thermal sensation levels, it is clear that most age groups lean toward a thermal sensation of +1 (slightly warm), with a broader age range and higher median in this category than others. The thermal sensation levels -3 (cold) and -1 (slightly cool) exhibit narrower age distributions, primarily involving younger individuals. Level 0 (neutral) has a lower median age, indicating that middle-aged participants might feel neutral thermal sensations more often. This suggests that perceived thermal comfort may vary with age, with older individuals favoring warmer conditions.

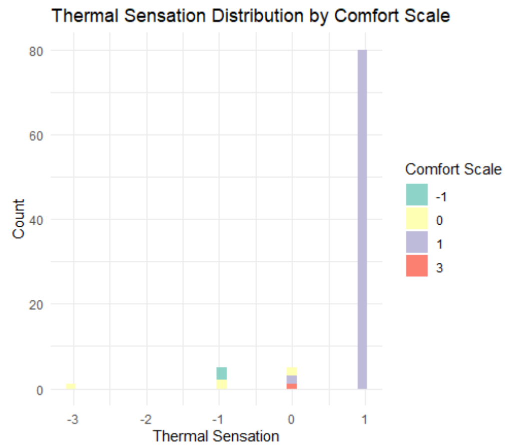

Fig. 15. Bar plot Thermal Sensation Distribution by Comfort Scale.

The bar plot in Figure 15 shows the distribution of thermal sensations categorized by the comfort scale. Most participants report a thermal sensation of +1 (slightly warm), with the highest count across all comfort scale levels. Most responses are associated with a comfort scale of 1 (comfortable) within this sensation. Thermal sensations of -3, -1, and 0 are less frequently reported, with fewer counts across all comfort scales, indicating that extreme cold or neutral thermal sensations are less common. This suggests that participants predominantly experience slightly warm conditions, often perceived as comfortable. This pattern could guide strategies to maintain thermal environments within this preferred range.

The exploratory data analysis (EDA) revealed that most occupants felt neutral to slightly warm, with comfort closely linked to air temperature and mean radiant temperature. While thermal comfort was influenced mainly by temperature, air quality factors like CO_2 and TVOC showed more complex patterns. Age also affected thermal sensation, suggesting that comfort varies by demographic. These insights highlight the importance of including both environmental and personal factors in the machine learning (ML) model to improve its accuracy and adaptability. The findings will guide feature selection to ensure the ML model captures real-world conditions and supports smarter HVAC control and building design for better indoor comfort.

D. *Developing a Predictive Model via R Programming*

To develop a predictive model, the process follows a structured approach:

Step 1 – Data Preparation

During data preparation, several steps were taken to make the dataset ready for machine learning. Categorical and character variables were converted into factors and then into numeric values. Numeric variables (except the target) were standardized using z-score normalization. Missing values were removed by omitting rows with incomplete data. The dataset was split into predictor variables (X) and target variables (Y), with predictors transformed into a matrix format suitable for models like XGBoost. A check was performed to ensure the target variable had more than one unique value—if not, all data was kept in the training set to avoid errors. Finally, the data was split into 80% training and 20% testing sets using a stratified method to keep the distribution balanced. These steps ensured the data was clean, consistent, and ready for modeling.

Step 2 – Feature Selection

Random Forest was chosen for feature selection due to its ability to handle nonlinear relationships, its robustness to multicollinearity, and its effectiveness in ranking feature importance based on impurity reduction. It is well-suited for high-dimensional data, efficiently identifying the most relevant predictors while discarding less significant ones. Additionally, its ensemble nature reduces overfitting, ensuring stable and generalizable feature selection. Using Random Forest (please see Figure 16), we improve model accuracy, enhance interpretability, and optimize computational efficiency, making it a reliable choice for selecting the most influential features in the dataset.

```
# Random Forest Feature Importance (Using CO2.ppm. as reference target)
rf_feat <- randomForest(X_train, Y_train$CO2.ppm., importance = TRUE)
feat_importance <- importance(rf_feat)
top_features <- rownames(feat_importance)[order(feat_importance[, 1], decreasing = TRUE)][1:10]

# Use only selected features for modeling
X_train <- X_train[, top_features]
X_test <- X_test[, top_features]
```

Fig. 16. Snippet on Feature Selection in R.

Step 3 – Modeling Approach

Two types of models are used: (1) Multi-output regression predicts multiple continuous targets simultaneously, such as *CO_2, PM2.5,* and *Satisfaction_Scale*, leveraging algorithms like Random Forest, XGBoost, Neural Networks, and SVM. (2) Multi-class classification is applied to predict *Thermal_Preference*. In this categorical variable, multiple labels can be associated with a single instance using Random Forest and SVM models.

1. *Multi-output regression for CO2, PM2.5, and Satisfaction_Scale*

In phase, four different regression models were implemented to predict CO2 concentration: Random Forest (RF), Extreme Gradient Boosting (XGBoost), Artificial Neural Networks (ANN), and Support Vector Machines (SVM), as can be seen in Figure 17. The Random Forest model was trained using an ensemble learning approach with 500 decision trees to improve robustness and reduce overfitting. The XGBoost model, known for its efficiency and performance in handling structured data, was trained using a gradient boosting framework with 500 iterations, optimizing for squared error loss. To implement ANN, a feedforward neural network with five hidden nodes was trained using backpropagation, with a linear output activation function and a maximum of 500 iterations to ensure convergence. Finally, the SVM model was trained using a radial basis function (RBF) kernel, which helps capture complex, non-linear relationships between features and CO2 levels. By employing these diverse models, the study aimed to compare their predictive performance and determine the most effective approach for estimating indoor CO2 concentration in the given environmental setting.

```
# Regression Models
xgb_train_reg <- xgb.DMatrix(data = as.matrix(X_train), label = Y_train$CO2.ppm.)
rf_model <- randomForest(X_train, Y_train$CO2.ppm., ntree = 500)
xgb_model <- xgboost(data = xgb_train_reg, nrounds = 500, objective = "reg:squarederror")
ann_model <- nnet(X_train, Y_train$CO2.ppm., size = 5, linout = TRUE, maxit = 500)
svm_model <- svm(X_train, Y_train$CO2.ppm., kernel = "radial")
```

Fig. 17. Snippet on multi-output regression for all RF, XGboost, ANN, and SVM.

2. *Multi-class classification for Thermal_Preference*

For the classification task, four different machine learning models were employed to predict thermal preference: Random Forest (RF), Extreme Gradient Boosting (XGBoost), Artificial Neural Networks (ANN), and Support Vector Machines (SVM) – please see Figure 18. The RF model was trained using 500 decision trees, leveraging ensemble learning to improve classification accuracy. The XGBoost classifier was implemented with 500 boosting iterations and utilized the softmax function to handle multi-class classification, ensuring efficient learning of complex patterns. The ANN classifier was structured as a feedforward neural network with five hidden neurons and a softmax activation function, optimizing for multi-class classification through iterative training with a maximum of 500 iterations. The SVM model was trained using a radial basis function (RBF) kernel, effectively handling non-linear decision boundaries in high-dimensional spaces. By comparing these models, the study aimed to identify the most suitable classification approach for accurately predicting occupants' thermal preferences based on environmental and demographic factors.

```
# Classification Models
rf_class <- randomForest(X_train, Y_train$Thermal.Preference, ntree = 500)
xgb_class <- xgboost(data = xgb_train_class, nrounds = 500, objective = "multi:softmax", num_class = length(unique(Y_train_xgb)))
ann_class <- nnet(X_train, class.ind(Y_train$Thermal.Preference), size = 5, softmax = TRUE, maxit = 500)
svm_class <- svm(X_train, Y_train$Thermal.Preference, kernel = "radial")
```

Fig. 18. Snippet on multi-class classification for all RF, XGboost, ANN, and SVM.

Step 4 – Performance Evaluation

Two key metrics were utilized for performance evaluation: Root Mean Squared Error (RMSE) for regression models and accuracy for classification models.

1. Multi-Output Regression Models

The provided R code snippet in Figure 19 computes the predicted values for different regression models to evaluate their performance using Root Mean Square Error (RMSE). Predictions are generated for Random Forest (pred_rf_reg), Extreme Gradient Boosting (pred_xgb_reg), Artificial Neural Network (pred_ann_reg), and Support Vector Machine (pred_svm_reg) models using the test dataset (X_test)—the xgb.DMatrix conversion ensures compatibility with the XGBoost model. These predictions can then be compared against target values to assess each model's accuracy and generalization performance.

```
# Compute RMSE for Regression Models
pred_rf_reg <- predict(rf_model, X_test)
pred_xgb_reg <- predict(xgb_model,
xgb.DMatrix(data = as.matrix(X_test)))
pred_ann_reg <- predict(ann_model, X_test)
pred_svm_reg <- predict(svm_model, X_test)
```

Fig. 19. Snippet on computed RMSE for regression models.

The computed RMSE values (see Table 3) reflect the performance of four regression models in predicting the target variables, including CO_2, PM2.5, and Satisfaction Scale in this multi-output regression case study. Lower RMSE values indicate better predictive accuracy. Random Forest achieved the best performance with an RMSE of 14.35, outperforming XGBoost, which recorded an RMSE of 17.52. The Support Vector Machine (SVM) model showed moderate performance with an RMSE of 21.41. At the same time, the Artificial Neural Network (ANN) exhibited the highest RMSE at 28.77, suggesting greater difficulty in capturing the underlying data patterns. These results highlight Random Forest as the most effective regression model for this case study, demonstrating superior predictive capability compared to the other models evaluated.

Table 3. Computed RMSE for all models.

Model	RMSE
Random Forest	14.35412
XGBoost	17.51975
ANN	28.77451
SVM	21.41230

2. Multi-class Classification models

The provided R code snippet (see Figure 20) computes the predicted class labels for different classification models to evaluate their accuracy. Predictions are obtained for Random Forest (pred_rf_class), Extreme Gradient Boosting (pred_xgb_class), Artificial Neural Network (pred_ann_class), and Support Vector Machine (pred_svm_class) models using the test dataset (X_test)—the xgb.DMatrix conversion ensures compatibility with the XGBoost model, while the type = "class" argument in the neural network model ensures class label predictions. These predicted labels can be compared against actual target labels to assess the classification performance of each model.

```
# Compute Accuracy for Classification Models
pred_rf_class <- predict(rf_class, X_test)
pred_xgb_class <- predict(xgb_class,
xgb.DMatrix(data = as.matrix(X_test)))
pred_ann_class <- predict(ann_class, X_test,
type = "class")
pred_svm_class <- predict(svm_class, X_test)
```

Fig. 20. Snippet on computed Accuracy for classification models.

The computed accuracy values (see Table 4) indicate the classification performance of four models for the multi-class classification task predicting Thermal Preference. Random Forest and Support Vector Machine (SVM) achieved the highest accuracy of 87.5%, demonstrating strong and comparable predictive capabilities. The Artificial Neural Network (ANN) followed with an accuracy of 75.0%, indicating reasonable performance but slightly lower than the top two models. In contrast, XGBoost performed poorly in this classification task, with an accuracy of only 6.25%, suggesting challenges in capturing the underlying patterns within the data. These findings highlight Random Forest and SVM as the most effective classification models for this case study, with ANN also providing acceptable performance.

Table 4. Computed accuracy for classification models.

Model	Accuracy
Random Forest	0.8750
XGBoost	0.0625
ANN	0.7500
SVM	0.8750

Step 5 – Visualization

Step 5 presents visualizations of the results from the previous step, highlighting the RMSE values for regression models and accuracy scores for classification models. These visual representations provide a more precise comparison of model performance for regression and classification tasks.

1. Regression Models

The bar chart in Figure 21 illustrates the RMSE values for four regression models that predict indoor air quality indicators. Among the models, Random Forest achieved the lowest RMSE, approximately 14.35, indicating the most accurate predictive performance in this regression task. XGBoost followed closely with an RMSE of about 17.52, still within an acceptable predictive range. The SVM model demonstrated moderate performance with an RMSE of around 21.41, suggesting some limitations in capturing complex relationships within the dataset. In contrast, the ANN model exhibited the highest RMSE, exceeding 28, indicating substantial difficulty in modeling the underlying data patterns. These results confirm that Random Forest and XGBoost are more reliable for regression in this study, while ANN may require further tuning or more data to improve performance.

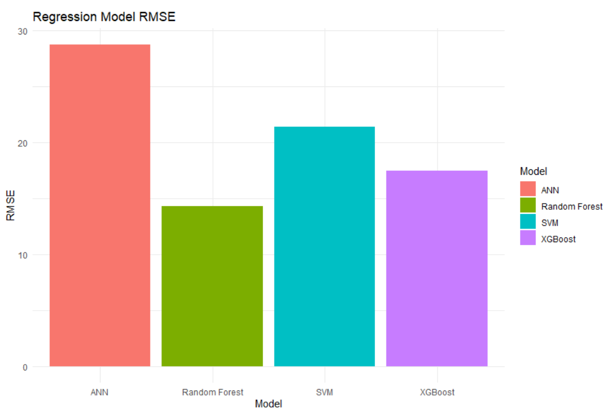

Fig. 21. Bar chart of Regression Model – RMSE .

2. Classification models

The bar chart in Figure 22 illustrates the classification accuracy of four models for predicting thermal preference. Random Forest and SVM achieved the highest accuracy of 87.5%, indicating strong and reliable performance in classifying thermal comfort levels. The ANN model followed closely with an accuracy of 75%, showing reasonably good performance, though slightly less effective than the top models. In contrast, XGBoost exhibited poor performance, with an accuracy of only 6.25%, highlighting its limitations in handling this specific multi-class classification problem within the dataset. These results suggest that Random Forest and SVM are the most suitable models for classification in this context, offering both stability and predictive accuracy.

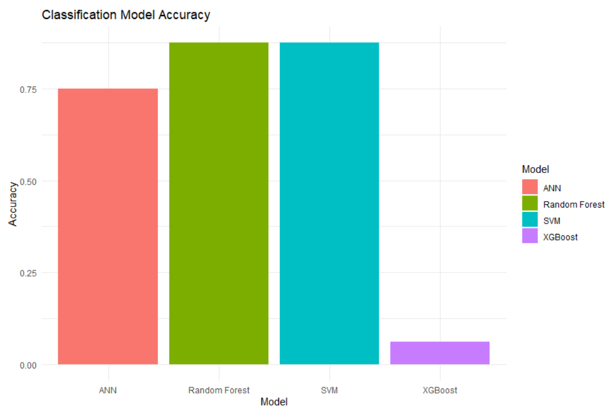

Fig. 22. Bar chart of Classification Model – Accuracy.

Step 6 – Statistical Validation of Model Performance

To assess the statistical significance of the observed differences in model performance, we conducted a 5-fold cross-validation for both regression and classification tasks. We applied paired t-tests for regression models to compare RMSE values across folds. For classification models, McNemar's test was used to determine whether the differences in prediction accuracy were statistically significant. This statistical validation step ensures that performance differences are not due to random variation.

1. Regression model

We focused statistical validation specifically on the Random Forest and XGBoost models. This decision was based on their superior performance in regression tasks, as evidenced by their relatively lower RMSE values than other models, as presented in Table 3. A standard practice in machine learning evaluation is to compare the top-performing models statistically to assess whether observed differences are statistically significant. This approach avoids unnecessary inflation of Type I error that may arise from multiple pairwise tests involving underperforming models such as ANN and SVM.

Table 5. Statistical Comparison of RMSE between Random Forest and XGBoost (5-Fold Cross-Validation).

Test	Test Statistic	p-value	95% Confidence Interval	Mean Difference	Interpretation
Paired t-test	t = 0.844	0.446	[-8.50, 15.93]	3.71	No statistically significant difference
Wilcoxon signed-rank test	V = 8	1.000	Not applicable	Not reported	No statistically significant difference

A paired t-test and Wilcoxon signed-rank test (as in Table 5) were conducted to evaluate whether the observed difference in RMSE between Random Forest and XGBoost was statistically significant. As shown in Table 5, neither test revealed a significant difference ($p > 0.05$), suggesting that the performance difference between the two models is not statistically meaningful.

2. *Classification model*

We performed McNemar’s test specifically between the Random Forest and Support Vector Machine (SVM) models because both achieved the same highest classification accuracy (87.5%) in predicting thermal preference, as shown in Table 4. This test is appropriate when two classifiers yield similar overall accuracy, and we want to determine whether their differences in prediction patterns are statistically significant. Conducting the test between these two top-performing models avoids unnecessary comparisons and reduces the risk of inflated Type I error. Furthermore, since the XGBoost model performed poorly (6.25% accuracy), it was excluded from this statistical comparison due to its limited relevance and unbalanced prediction behavior, which would bias the analysis.

Table 6. McNemar’s Test Contingency Table for Random Forest and SVM Classification Results.

	SVM Incorrect	SVM Correct
RF Incorrect	2	0
RF Correct	0	14

The McNemar’s test (see Table 6) was used to assess whether the differences in prediction behavior between the Random Forest and SVM classifiers were statistically significant. However, the result returned NaN for the chi-squared statistic and NA for the p-value. This typically occurs when all disagreements lie in a single cell (i.e., no discordant pairs in opposite directions), violating McNemar's test's assumptions.

In this case, Random Forest and SVM predictions differed in only 2 instances, misclassified by both models. Since there were no cases where RF was correct and SVM incorrect (or vice versa), the test could not compute a valid statistic. This outcome suggests high consistency between the two models, reinforcing that their classification behavior is nearly identical despite having separate algorithmic foundations.

In summary, statistical tests showed no significant difference between the models' performance. In regression, paired t-tests and Wilcoxon tests confirmed that Random Forest and XGBoost had similar RMSE values. For classification, McNemar’s test showed that Random Forest and SVM had identical accuracy with perfect agreement. This suggests that Random Forest is a reliable and consistent model for both regression and classification tasks, while XGBoost and SVM also remain strong alternatives depending on the prediction goal

4. DISCUSSION

This study presents a comprehensive machine learning (ML) framework to predict indoor air quality (IAQ) and thermal comfort in a healthcare facility setting, specifically within Pusat Hemodialysis Sungai Lembing, Pahang. The discussion draws directly from the systematic approach outlined in Section 3.2, highlighting the logical progression from raw data collection to model development and statistical validation.

The first critical step was the integration of two complementary datasets—objective environmental sensor data and subjective human perception data. By applying temporal synchronization, we ensured that each subjective response from patients or staff was aligned with the nearest corresponding environmental measurement. This fusion of data provided a unified view of both the physical environment and human experience, enabling a human-centric modeling approach that is rarely explored in the context of tropical healthcare facilities.

Following data integration, exploratory data analysis (EDA) played a vital role in understanding the distribution, quality, and underlying patterns in the dataset. EDA revealed no major data quality issues, aside from some missing values in relative humidity (RH), which were appropriately handled. The distribution of demographic data, particularly age, was found to influence thermal sensation, with older participants tending to prefer warmer environments. Furthermore, visualizations such as scatterplots and correlation matrices highlighted key environmental variables like air temperature, MRT, and TVOCs, which strongly relate to comfort levels. These insights directly informed the feature selection and model development steps.

In Step D, the predictive modeling process began with R programming and data preparation. This included normalization of numerical predictors, conversion of categorical variables into numerical format, and stratified data partitioning to maintain representativeness in training and testing sets. Random Forest was employed for feature importance analysis, helping to reduce dimensionality and focus the models on the most influential predictors. This approach not only enhances model performance but also ensures computational efficiency.

The study deployed four machine learning algorithms—Random Forest, XGBoost, Artificial Neural Networks (ANN), and Support Vector Machines (SVM)—to perform both regression (predicting CO_2, PM2.5, and satisfaction scale) and classification (predicting thermal preference). For regression, Random Forest achieved the lowest RMSE of 14.35, followed by XGBoost (17.52), while ANN recorded the highest error, indicating its limitation with the given data. In classification, Random Forest and SVM both achieved an accuracy of 87.5%, with ANN performing reasonably well (75%) and XGBoost lagging significantly (6.25%).

To ensure the validity of these results, Step 5 introduced statistical validation. Paired t-tests and Wilcoxon signed-rank tests were conducted for the top-performing regression models, Random Forest and XGBoost, revealing no statistically significant difference, indicating that either model can be a strong candidate depending on system requirements. Similarly, McNemar's test between Random Forest and SVM for classification also showed no statistically significant difference in prediction behavior, affirming both models' reliability.

These findings lead to several key contributions of the paper:

1. *Methodological Integration:* The study proposes a structured and reproducible end-to-end pipeline, combining objective and subjective data sources—a notable advancement for environmental monitoring in healthcare.
2. *Data Fusion with Temporal Synchronization:* Using timestamp-based merging of disparate data sources enables meaningful cross-analysis, allowing ML models to reflect real-world occupant-environment interactions.
3. *Model Benchmarking in Tropical Healthcare Context:* This study benchmarks widely used ML algorithms in a Malaysian clinic setting, a context that is underrepresented in the current literature.
4. *Statistically Validated Insights*: Unlike prior studies that stop at RMSE or accuracy metrics, this study includes formal statistical tests to support the robustness and generalizability of its results.
5. *Support for Real-World Implementation*: Using interpretable models like Random Forest and statistically sound validation makes the approach highly suitable for operational deployment in smart healthcare facilities.

In summary, this study highlights the effectiveness of ML-based predictive modeling for IAQ and thermal comfort monitoring in a Malaysian healthcare setting. The structured methodology, validated results, and human-centric integration significantly contribute to developing intelligent, responsive indoor environments in clinical settings.

5. CONCLUSION

This study successfully developed and evaluated a machine learning-based predictive framework for monitoring indoor air quality (IAQ) and thermal comfort in Malaysian healthcare. The study enabled a holistic assessment of occupant comfort and air quality by integrating temporally synchronized environmental sensor data and human perception responses. The structured development process—from data preprocessing and exploratory analysis to model implementation and statistical validation—resulted in reliable and interpretable insights into environmental comfort dynamics.

Key findings indicate that Random Forest consistently outperformed other models in both regression and classification tasks. It achieved the lowest root mean squared error (RMSE = 14.35) in predicting continuous IAQ variables and the highest classification accuracy (87.5%) in predicting thermal preferences, showing strong generalizability and model stability. XGBoost and SVM also showed promise in specific tasks but lacked consistent superiority. Statistical tests confirmed no significant performance difference between Random Forest and its closest competitors, validating its robustness as a default choice for real-time environmental modeling in similar contexts.

Despite these encouraging results, the study has a few limitations. The dataset was limited to a single rural healthcare facility and a specific time window, limiting its generalizability across different climates, building types, and seasons. Additionally, the classification models may be influenced by class imbalance, and complex models like ANN proved computationally intensive, which could hinder real-time deployment in resource-constrained settings.

Future work should focus on expanding data collection across diverse healthcare facilities and climatic zones in Malaysia to improve model adaptability. Implementing real-time integration with HVAC systems and validating model performance under live operating conditions will be crucial for operationalizing these predictive insights. Exploring hybrid approaches, such as combining Random Forest with fuzzy logic or reinforcement learning, may enhance adaptability and decision-making in dynamic environments. Continued emphasis on explainable AI methods will support broader acceptance in regulatory healthcare contexts.

In conclusion, this study demonstrates that data-driven, statistically validated, and human-centric AI models hold strong potential for improving environmental monitoring and comfort optimization in Malaysian healthcare environments.

Acknowledgement

We would like to acknowledge the Royal Society (ICA\R1\201236) and UEM Edgenta SDN BHD for the financial support for this research. Besides, we would like to acknowledge Universiti Kebangsaan Malaysia (RS 2020-006) and (RS 2024_006) for the corporation to this research. The co-author, Dr. Hasila Jarimi, would also like to thank UKRI for the Marie Skłodowska-Curie Actions (MSCA) Postdoctoral Fellowship Guarantee Funding (Ref: 101151868), which enabled her to complete this paper during her fellowship.

REFERENCES

[1] J. Khodakarami and N. Nasrollahi, "Thermal comfort in hospitals – A literature review," *Renewable and Sustainable Energy Reviews*, vol. 16, no. 6, pp. 4071–4077, Aug. 2012, doi: 10.1016/J.RSER.2012.03.054.

[2] A. A. Aziz, D. Sumiyoshi, and Y. Akashi, "Low cost humidity controlled air-conditioning system for building energy savings in tropical climate," *Journal of Building Engineering*, vol. 11, pp. 9–16, May 2017, doi: 10.1016/J.JOBE.2017.03.005.

[3] Z. Wang *et al.*, "Individual difference in thermal comfort: A literature review," *Build Environ*, vol. 138, pp. 181–193, Jun. 2018, doi: 10.1016/J.BUILDENV.2018.04.040.

[4] C. A. Balaras, E. Dascalaki, and A. Gaglia, "HVAC and indoor thermal conditions in hospital operating rooms," *Energy Build*, vol. 39, no. 4, pp. 454–470, Apr. 2007, doi: 10.1016/J.ENBUILD.2006.09.004.

[5] C. Fan, Y. Sun, Y. Zhao, M. Song, and J. Wang, "Deep learning-based feature engineering methods for improved building energy prediction," *Appl Energy*, vol. 240, pp. 35–45, Apr. 2019, doi: 10.1016/J.APENERGY.2019.02.052.

[6] H. Q. Flayyih, J. Waleed, and A. M. Ibrahim, "IoT-based AI Methods for Indoor Air Quality Monitoring Systems: A Systematic Review," *International Journal of Computing and Digital Systems*, vol. 16, no. 1, pp. 813–826, Jul. 2024, doi: 10.12785/IJCDS/160159.

[7] A. Singh, M. Islam, and N. Dinh, "Forecasting Indoor Air Quality Using Machine Learning Models," *Digest of Technical Papers - IEEE International Conference on Consumer Electronics*, 2024, doi: 10.1109/ICCE59016.2024.10444387.

[8] M. W. Ahmad, M. Mourshed, and Y. Rezgui, "Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption," *Energy Build*, vol. 147, pp. 77–89, Jul. 2017, doi: 10.1016/J.ENBUILD.2017.04.038.

[9] Y. H. Yau and B. T. Chew, "A review on predicted mean vote and adaptive thermal comfort models," vol. 35, no. 1, pp. 23–35, Oct. 2012, doi: 10.1177/0143624412465200.

[10] A. Baudet, E. Baurès, O. Blanchard, P. Le Cann, J. P. Gangneux, and A. Florentin, "Indoor Carbon Dioxide, Fine Particulate Matter and Total Volatile Organic Compounds in Private Healthcare and Elderly Care Facilities," *Toxics*, vol. 10, no. 3, Mar. 2022, doi: 10.3390/TOXICS10030136.

[11] Y. Yusup, M. I. Ahmad, and N. Ismail, "Indoor Air Quality of Typical Malaysian Open-air Restaurants," *Environment and Pollution*, vol. 3, no. 4, Aug. 2014, doi: 10.5539/EP.V3N4P10.

[12] W. A. Rahman, H. Rosli, S. N. Baharuddin, and B. Salleh, "Incidence and remediation of fungi in a sick building in Malaysia: a case study," *Aerobiologia (Bologna)*, vol. 28, no. 2, pp. 275–283, Jun. 2012, doi: 10.1007/S10453-011-9226-Y.

[13] I. A. Rahman, J. C. P. Putra, and S. Nagapan, "Correlation of Indoor Air Quality with Working Performance in Office Building," *Mod Appl Sci*, vol. 8, no. 6, pp. 153–160, 2014, doi: 10.5539/MAS.V8N6P153.

[14] M. El-Sharkawy and M. Noweir, "Indoor air quality levels in a University Hospital in the Eastern Province of Saudi Arabia," *J Family Community Med*, vol. 21, no. 1, pp. 39–47, Apr. 2014, doi: 10.4103/2230-8229.128778.

[15] M. N. N. Rodrigo, Y. Kosala, B. A. K. S. Perera, and C. Dalugoda, "Mitigation of Hospital Acquired Infections in Developing Countries through the Provision of a Better IAQ," *Engineer: Journal of the Institution of Engineers, Sri Lanka*, vol. 51, no. 1, p. 39, Jan. 2018, doi: 10.4038/ENGINEER.V51I1.7286.

[16] U. M. Hellgren, M. Hyvärinen, R. Holopainen, and K. Reijula, "Perceived indoor air quality, air-related symptoms and ventilation in Finnish hospitals," *Int J Occup Med Environ Health*, vol. 24, no. 1, pp. 48–56, 2011, doi: 10.2478/S13382-011-0011-5.

[17] S. K. Koutroumpi, "Identifying Gap of Knowledge about Thermal Comfort in Naturally Ventilated Wards in Hot-humid Settings," *Global Journal of Engineering Sciences*, vol. 6, no. 2, Jul. 2020, doi: 10.33552/GJES.2020.06.000632.

[18] K. Chaloeytoy, A. Dubsok, and S. Kittipongvises, "Evaluating Thermal Comfort in Ward Areas of the Public Hospital in the Tropical Climates," *E3S Web of Conferences*, vol. 530, May 2024, doi: 10.1051/E3SCONF/202453005009.

[19] L. Faraj, M. F. Z. Bin Jaafar, and Z. Shari, "Evaluating Ventilation for Thermal Comfort in Two Classrooms Types in Malaysian Secondary Classrooms," *International Journal of Sustainable Development and Planning*, vol. 18, no. 7, pp. 2069–2078, Jul. 2023, doi: 10.18280/IJSDP.180709.

[20] K. T. Huang and R. L. Hwang, "Parametric study on energy and thermal performance of school buildings with natural ventilation, hybrid ventilation and air conditioning," *Indoor and Built Environment*, vol. 25, no. 7, pp. 1148–1162, Nov. 2016, doi: 10.1177/1420326X15609773.

[21] F. Azizpour, S. Moghimi, E. Salleh, S. Mat, C. H. Lim, and K. Sopian, "Thermal comfort assessment of large-scale hospitals in tropical climates: A case study of University Kebangsaan Malaysia Medical Centre (UKMMC)," *Energy Build*, vol. 64, pp. 317–322, 2013, doi: 10.1016/J.ENBUILD.2013.05.033.

[22] D. Palladino, I. Nardi, and C. Buratti, "Artificial Neural Network for the Thermal Comfort Index Prediction: Development of a New Simplified Algorithm," *Energies (Basel)*, vol. 13, no. 17, Sep. 2020, doi: 10.3390/EN13174500.

[23] G. Tardioli, R. Filho, P. Bernaud, and D. Ntimos, "An Innovative Modelling Approach Based on Building Physics and Machine Learning for the Prediction of Indoor Thermal Comfort in an Office Building," *Buildings*, vol. 12, no. 4, Apr. 2022, doi: 10.3390/BUILDINGS12040475.

[24] A. De *et al.*, "Implementing At-Scale Adaptive Thermal Comfort Controls for Mixed Mode Building Using Machine Learning," *FARU Journal*, vol. 9, no. 2, pp. 12–19, Dec. 2022, doi: 10.4038/FARUJ.V9I2.172.

[25] Y. J. Ju, J. R. Lim, and E. S. Jeon, "Prediction of AI-Based Personal Thermal Comfort in a Car Using Machine-Learning Algorithm," *Electronics (Basel)*, vol. 11, no. 3, Feb. 2022, doi: 10.3390/ELECTRONICS11030340.

[26] S. Liu, S. Schiavon, H. P. Das, M. Jin, and C. J. Spanos, "Personal thermal comfort models with wearable sensors," *Build Environ*, vol. 162, Sep. 2019, doi: 10.1016/J.BUILDENV.2019.106281.

[27] M. Nascimento and P. Lopes, "Internet of things and machine learning applied to the thermal comfort of internal environments," *Indoor and Built Environment*, vol. 31, no. 9, pp. 2274–2290, Nov. 2022, doi: 10.1177/1420326X221097338.

[28] L. T. Wong, K. W. Mui, and T. W. Tsang, "Updating Indoor Air Quality (IAQ) Assessment Screening Levels with Machine Learning Models," *Int J Environ Res Public Health*, vol. 19, no. 9, May 2022, doi: 10.3390/IJERPH19095724.

[29] K. Huang, S. Lu, X. Li, and W. Chen, "Using random forests to predict passengers' thermal comfort in underground train carriages," *Indoor and Built Environment*, vol. 32, no. 2, 2022, doi: 10.1177/1420326X221110046.