ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

ARCHIVIO ISTITUZIONALE
DELLA RICERCA

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

IoT-Based Indoor Thermal Comfort Prediction Using Multivariate Statistical Analysis

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:
Afif, O., Ingenito, G., Bellomo, M., Romagnoli, A., Tartagni, M., Romani, A. (2025). IoT-Based Indoor Thermal Comfort Prediction Using Multivariate Statistical Analysis. IEEE SENSORS JOURNAL, 25(8), 14358-14369 [10.1109/JSEN.2025.3545360].

Availability:
This version is available at: https://hdl.handle.net/11585/1008498 since: 2025-05-14

Published:
DOI: http://doi.org/10.1109/JSEN.2025.3545360

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/).
When citing, please refer to the published version.

(Article begins on next page)

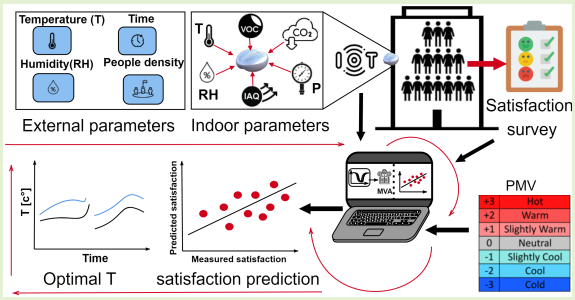
10 November 2026

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2025.3545360

IoT-Based Indoor Thermal Comfort Prediction Using Multivariate Statistical Analysis

Oumaima Afif, *Graduate Student Member, IEEE*, Gaetano Ingenito, Marco Bellomo, Andrea Romagnoli, Marco Tartagni, *Member, IEEE*, and Aldo Romani, *Member, IEEE*

***Abstract*— Emerging internet of things (IoT) technologies, combined with sensors and data analytics, enable intelligent management of buildings, significantly improving energy efficiency and optimizing thermal comfort for occupants. In this paper, we present a novel modeling framework that integrates IoT-based monitoring with multivariate statistical analysis (MVA) to assess indoor thermal comfort and predict the optimal temperature (T) that maximizes occupant satisfaction within a given space. By continuously monitoring real-time environmental conditions, 11 indoor and outdoor variables are collected every 20 minutes from three different locations via an IoT network. Additionally, surveys on occupant satisfaction and thermal comfort are conducted four times daily over a 21-day period to evaluate overall indoor comfort. Both linear models, using standard regression, and non-linear models, adopt locally weighted regression (LWR) methods, which was exploited for the first time in this context to predict indoor comfort using current and historical environmental data. The model achieves a prediction error of 9.6% in predicting user satisfaction at the selected site. Another predictive model is developed to predict the optimal T that maximizes occupant satisfaction. Notably, the model demonstrates significant improvements in occupant satisfaction across three different rooms, increasing from 5.18% to 100%, from 12.69% to 91.53%, and from 58.48% to 100%, respectively. This study highlights the significant role of IoT and MVA in uncovering intricate correlations within complex data sets. To our knowledge, this is the first time MVA has been applied to indoor wellness models, where neural networks (NNs) are currently the most used approach with higher computational complexity. The proposed framework sets the stage for the optimum control of heating, ventilation and air conditioning (HVAC) systems in buildings. By leveraging IoT and MVA, this framework provides actionable insights for HVAC system optimization, reducing unnecessary energy consumption while maintaining occupant comfort. This research demonstrates a practical, interpretable, and scalable alternative to NNs-based models, making it highly suitable for real-time building management and smart climate control systems.**



***Index Terms*— Environmental monitoring, Internet of Things, machine learning, multivariate analysis, indoor wellness, digital twins.**

I. Introduction

THE rise in stress levels exerts a profound influence on individuals' physical and mental well-being. Several studies have demonstrated that this phenomenon is especially noticeable in work environments [1]–[3], given that approximately 90% of individuals spend the majority of their lives indoors [4]. Many aspects of an individual's life can impact their well-being, including their social situation, physical and mental health, financial and occupational status, and their surrounding environment [5]. The latter is also referred to as environmental wellness and is influenced by the physical and chemical components of the environment [4] [6]. The international labour organisation (ILO) has acknowledged that mental health can be significantly impacted by indoor environmental wellness [7]. Within this framework, several initiatives undertaken by researchers have placed a growing emphasis on human well-being. For instance, in [8]–[12], children's physical health and psychological well-being in a school environment have been investigated. These studies illustrate the relationship between changes in cognitive function and various environmental factors, with a particular focus on concentrations of carbon dioxide (CO_2) and volatile organic compounds (VOCs). On the one hand, these works confirm that an appropriate environment can increase the comfort, productivity, and well-being of the occupants. On the other hand, there are some shortcomings, such as the use of data from different sources, which can introduce errors due to different calibration, accuracy, and filtering techniques.

O. Afif, G. Ingenito, M. Tartagni, and A. Romani are with the Advanced Research Center on Electronic Systems (ARCES) of the University of Bologna, Cesena Campus, 47522, Cesena Italy (E-mail:{oumaima.afif2, marco.tartagni, aldo.romani}@unibo.it, gaetano.ingenito@studio.unibo.it). A. Romagnoli and M. Bellomo are with Taua Società Benefit (E-mail: {a.romagnoli, m.bellomo}@taua.tech).

This study was carried out within the Agritech National Research Center and received funding from the European Union Next-GenerationEU (PNRR) – Mission 4 Component 2, Investment 1.4 – D.D. 1032 17/06/2022, CN00000022. The authors would like to acknowledge the financial support by the Chips Joint Undertaking (Chips JU) under grant agreement No. 101112286 (LoLiPoP IoT)

© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2025.3545360

Other research that has been conducted within the same demographic (with a focus on medical conditions) indicated that exposure to contaminated air heightens the susceptibility to acquiring illnesses [13]. From another perspective, the study conducted in [14] focused on the integration of biosensor data and environmental parameters, emphasizing the promotion of psychological well-being. The main purpose of the latter work was to meticulously examine the correlation that exists between the merged data and the productivity metrics within the workforce. However, the group of participants had to follow strict guidelines to ensure the accuracy of the study, such as sitting in a specific place, which could affect the results.

Indeed, the notable increase in the widespread availability of sensors designed to monitor gaseous, biological, chemical, and particle pollutants is creating new opportunities for real-time monitoring within buildings [15]. The great evolution of wireless communications, and the rise of internet of things (IoT) technologies, offer unprecedented opportunities for ubiquitous data acquisition, transmission, and processing, either remotely or at the edge. Yet, the task of choosing dependable sensors poses a substantial challenge for researchers in the field [16]. Along with the benefits of low-cost sensing technologies, multiple challenges have been a point of focus for authorities and agencies working in the field of importance performance analysis (IPA) management. There are still numerous questions (including energy and efficiency management) on the use of the available low-cost sensors [3]. An additional issue may be found in the application of standard indoor conditions that may not ensure users' satisfaction [17]. In fact, the ability to tailor comfort conditions to specific needs generally increases the satisfaction and comfort of the occupants in indoor environments, with positive individual and social outcomes [18].

The predicted mean vote (PMV) model, developed by [19], evaluates thermal comfort by using six factors: four environmental variables (temperature (T), relative humidity (RH), mean radiant temperature (MRT), and air velocity) and two personal factors (metabolic rate and clothing insulation). It generates a score ranging from -3 to 3, representing sensations from cold to hot. The PMV model is not easily applied in smart buildings due to the complex interactions between indoor T, RH, and heating, ventilation, and air conditioning (HVAC) systems. Factors like the difference between HVAC setpoints and indoor T can vary over time. Therefore, a building management system (BMS) may simulate comfort and energy consumption based on different HVAC configurations, to identify optimal settings that ensure comfort while minimizing energy usage. Consequently, there is an increasing demand for the creation of new thermal comfort models utilizing low-computation machine learning (ML) that connect comfort levels with controllable HVAC operations for the effective management of smart buildings through IoT nodes [20], [21].

The proposed framework advances IoT-based indoor monitoring by combining multivariate analysis (MVA) with IoT technologies for real-time comfort prediction. It provides a computationally efficient alternative to traditional ML models [22] and enables edge-based data processing. Additionally, it integrates HVAC control for energy-efficient management, enhancing both occupant comfort and sustainability.

II. Related work

Monitoring indoor thermal comfort is a key area of research for evaluating human well-being. Early studies in this field relied on historical data from specific locations, combined with environmental measurements and occupant feedback on comfort levels. These factors not only affect people's comfort but also impact the performance of equipment and can help optimize energy usage in HVAC systems. In [23], the effects of thermal comfort on employee productivity within a specific company are evaluated by collecting data on T, RH, air velocity, biometric information, and individual satisfaction levels from the workspace. Various classification and regression algorithms are applied to identify the most effective method for analyzing each parameter, such as random forest (RF), regression tree (RT), support vector machine (SVM), and k-nearest neighbors (KNN).

In [24], the authors employ a non-invasive sensor network that allows for unobtrusive monitoring of elderly individuals, thereby respecting their comfort while capturing valuable data. The approach is personalized, as it incorporates individual behavioral data, and uses well-being indices derived from daily surveys, enhancing the relevance and accuracy of its assessments. Moreover, by integrating ML algorithms—specifically RF and RT—the research provides a data-driven method for estimating user well-being. However, a limitation is the relatively small sample size, with only eight individuals across five apartments, which may affect the generalizability of the results. The focus on elderly participants limits the findings' applicability to other demographics.

In [22], the authors propose a learning-based solution leveraging ML techniques and IoT-enabled pervasive sensing. Their approach involves the development of an intelligent thermal comfort management (iTCM) system, which collects environmental data via a wireless sensor network and monitors vital signs using wearable devices. A key strength is the use of ubiquitous sensing technologies for comprehensive data collection. This enables the development of more accurate predictive models like the intelligent thermal comfort neural network (iTCNN), which significantly outperforms traditional comfort models such as the PMV by up to 13.1%. However, a primary challenge is the potential complexity and cost associated with deploying pervasive IoT infrastructure across diverse building settings, which may limit scalability.

In [26], an IoT-based approach to smart building management, focusing on thermal comfort, is presented using deep learning methodologies. The study introduces a fine-grained architecture where DNN are employed to model complex interactions between various comfort factors and building operations, showcasing enhanced predictive capabilities compared to traditional models. The fine-grained modeling approach allows for the evaluation of each contributing factor to thermal comfort independently, resulting in notably improved prediction accuracy. The incorporation of IoT-generated data ensures that the model can provide adaptive and real-time predictions, highlighting a significant advancement in smart

© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted,
but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2025.3545360

TABLE I

LITERATURE REVIEW OF THERMAL COMFORT PREDICTION METHODS

Reference	Machine Learning Method	Input Parameters (sensed quantities)	Predicted Parameters	Complexity
[24]	RF, RT	T, RH, lighting	User well-being indices	Moderate
[22]	Neural network (NN)	T, RH, heart rate	Thermal comfort	High
[25]	Back-propagation neural network (BPNN)	T, RH, lighting	Indoor quality, thermal comfort	High
[26]	Deep neural networks (DNN)	CO_2, HVAC Status, T, RH	Thermal comfort factors	High
[27]	KNN, SVM, classification and regression tree (CART)	T, RH, CO_2, people density	Thermal comfort index	High
[28]	Deep meta-learning (DML)	T, RH, physiological data	Personal thermal comfort	High
[29]	Convolutional neural network (CNN)	T, CO_2	Indoor occupancy patterns	Moderate
[30]	KNN, decision trees (DT), hybrid artificial neural network (ANN)	T, CO_2, RH, light levels	Room occupancy accuracy	Moderate
[31]	iTCNN	T, RH	Thermal comfort satisfaction	High
[32]	Conventional ML, deep learning (DL)	T, datasets (ASHRAE RP-884)	Thermal sensation votes	High
[33]	RF, RT	T, RH, lighting	User well-being indices	Moderate
This paper	MVA	T, RH, VOCs, indoor air quality (IAQ), static IAQ, CO_2, pressure (P), people density, PMV, time, External T and RH	Occupancy satisfaction, optimal T	Moderate

building technology. However, a key limitation is the potential complexity and computational requirements of the DNN models, which may pose challenges in resource-constrained environments or in buildings lacking robust IoT infrastructure.

Recent studies reflect a shift towards comprehensive, ML-driven frameworks that analyze the combined effects of T, RH, lighting, and air quality. An optimal indoor environmental quality prediction model is proposed in [25], employing ML classification algorithms to understand occupant satisfaction across various environmental parameters. The study centers on the use of a back-propagation neural network for predictive modeling. While this method has proven effective, alternative approaches with lower computational demands could further optimize the system's performance and efficiency. Exploring lighter machine learning models or simpler algorithms might reduce computational costs while maintaining accuracy, especially in environments where resource constraints are a concern. Table I presents a detailed comparison of our proposed method with other related studies.

To address the identified research gap, this paper presents a novel integration of IoT and low-computation ML techniques to enable predictive environmental analysis for indoor spaces. The primary focus is on enhancing comfort, occupant-perceived indoor environmental quality, and satisfaction, which is the thermal condition assessed using PMV 7-point thermal scale, within corporate environments, although the system is also adaptable for residential use. Importantly, the proposed method relies on sole environmental parameters, with no need to measure sensitive personal information.

To the best of the author's knowledge and based on the latest literature review in [34], this study is the first to apply MVA in this particular context. Integrating MVA into such applications involves dimensionality reduction and feature extraction, which helps reduce computational burdens while retaining crucial information [35]. Unlike ML models, which are resource-intensive and often lack transparency, the MVA methods used in this study, such as principal component analysis (PCA), partial least square (PLS), and locally weighted regression (LWR), are computationally efficient and interpretable, making them particularly well-suited for IoT environments. While ML, especially DL, excels in capturing complex patterns, its sensitivity to noise and significant computational demands pose challenges for real-time IoT applications [36], [37]. MVA relies on linear algebra and statistical methods, which results in lower computational complexity compared to traditional ML and DL models [38], [39]. ML methods generally require more computational power than MVA due to the need for training algorithms and hyperparameter optimization. DL models are the most resource-intensive due to their deep architectures, requiring specialized hardware like graphics processing units (GPUs) or tensor processing unitss (TPUs). These models also tend to have opaque decision-making processes, which can be problematic in regulated environments.

In addition, MVA excels in real-time or near-real-time applications, which are common in many IoT fields such as environmental monitoring, Because it uses closed-form solutions and matrix computations, MVA delivers faster results compared to more complex models. In IoT applications where both speed and interpretability are essential, MVA provides clear mathematical relationships, offering transparency in decision-making. This transparency is especially beneficial in sectors where regulatory compliance and clear decision-making processes are crucial. The developed model of this study not only predicts the optimal T but also compares it with real-time data, enabling a future implementation on a microcontroller connected to a thermostat to maintain ideal T conditions that maximize occupant satisfaction.

The main contributions are as follows:

© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2025.3545360

1) This study stands out by employing an IoT device equipped with multiple sensors to collect seven key environmental parameters—indoor T, indoor RH, IAQ, P, CO_2 and VOCs concentration, and static IAQ—alongside external factors such as external T, RH, time, room occupancy density, and assessments of 89 occupants' well-being.
2) A hybrid MVA methodology is proposed to leverage PLS regression alongside standard regression techniques to effectively capture the relationships between the variables presented in Table II and occupancy satisfactions of the thermal comfort. PLS is particularly efficient in modeling linear relationships, making it an ideal first-step analysis tool. However, to address potential nonlinearities and enhance model precision, we propose the application of LWR, which serves as a localized extension of PLS. This dual-approach methodology provides a more robust framework for exploring complex relationships, particularly in cases where global linear models may fall short.
3) PMV values are used to assess occupant comfort and to estimate predicted percentage of dissatisfied (PPD) based on Fanger's model. This approach offers a comprehensive framework for optimizing indoor environmental quality by aligning T settings with occupant preferences, enhancing both comfort and energy efficiency.

The modeling framework presented in this paper aims at addressing the following research questions:

1) How can IoT-based environmental data and MVA be combined to measure and predict indoor comfort?
2) Can the optimal T that maximizes occupant satisfaction be accurately predicted?

The rest of the paper is organized as follows: Section III delves into the IoT framework used in the study. Section IV provides a detailed explanation of the methodologies, including devices, data acquisition processes, and the multivariate analysis techniques applied. Section V discusses the data preparation and preprocessing approaches. Section VI presents an analysis of the results. Lastly, Section VII summarizes the study, highlighting key findings, implications, and potential applications.

TABLE II

VARIABLES CONSIDERED FOR THE STUDY

Variable	unit	Variable	unit
T	[°C]	P	[bar]
RH	[%]	CO_2 concentration	[ppm]
IAQ, Static IAQ[1]	[] []	VOC	[ppm]
External T	[°C]	External RH	[%]
People density	$[\frac{\#\text{ people}}{\text{m}^2}]$	Time	[Min]

III. SYSTEM ARCHITECTURE

This study is based on extensive acquisitions performed with the Tauanito INAIR indoor air quality sensor, an optimized wireless sensor node for IoT-based environmental monitoring applications developed by Taua [40]. It supports scalable integration of multiple sensors, includes an embedded global positioning system (GPS) module for localization, and offers comprehensive data measurement capabilities with seamless transmission via cellular networks through the Internet. The internal device structure is shown in Fig. 1 along with the sensed quantities, and the data management architecture.

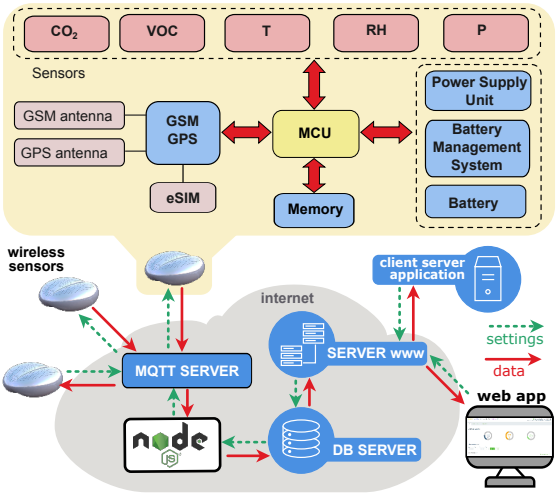


Fig. 1. Schematic representation of the device architecture and data transmission.

A Bosch BME680 digital sensor [41] measures gas, RH, P, and T. The gas sensor reacts to a broad range of VOCs, by changing its internal resistance, due to the adsorption and subsequent oxidation/reduction of these compounds on its sensitive layers. VOCs are chemicals that easily evaporate into the air, released by sources like industries and household products, and their long-term exposure can cause health issues and contribute to air pollution [42]. Then, the device outputs the resistance value, which is also influenced by environmental parameters (e.g. RH). The raw values are then processed with the support of dedicated software libraries to calculate an IAQ index, which reflects the air quality based on VOC concentrations and environmental conditions, and consists of an index ranging from 0 to 500, with a resolution of 1. RH, P, and T are measured respectively with absolute accuracies of $\pm 3\%$, ± 0.6 hPa, and 0.5 °C. Moreover, an ExplorIR-M miniature CO_2 from Global Sensing Solutions is used to measure CO_2 concentrations up to 5% and with a typical accuracy of ± 70 ppm, by using a non-dispersive infrared lamp to direct waves of light through a tube filled with an air sample.

The core of the system is an STMicroelectronics STM32L0 microcontroller unit (MCU), powered by an internal lithium-ion rechargeable battery (3.7 V and 2000 mAH capacity) with a battery management system. Battery usage is optimized at MCU level with dedicated algorithms, which adjust device settings and disable the supply rails of sensors to limit power consumption according to the required needs. The MCU operates in different ultra-low power modes when not all device functionalities are needed. In these modes, the high-frequency 8 MHz clock oscillator is turned off, leaving only a 32.768 kHz

[1]Is a filtered (trimmed) version of the IAQ, the best representative of the measurement in case the device is fixed in place.

© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted,
but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2025.3545360

quartz oscillator for the internal real-time clock (RTC) active. Furthermore, the MCU wakes up only when a task needs to be executed. A custom scheduler has been developed to minimize power consumption and enable the core only when a task must run. To maximize battery duration, cellular data transmission is disabled when the battery voltage drops below 3.5 V, while sensor management and data acquisitions are kept active by saving data to the on-board ultra-low power static memory until the device is recharged. When the battery voltage drops below 3.3 V, the device shuts down all services, including measurements, to preserve the internal battery. The sensor node manages power settings, maximizing battery life up to 2 years with one acquisition per day, and up to 2 months with acquisitions every 20 minutes. Device dimensions are $137.5 \times 110 \times 60$ mm.

The system is equipped with a GPS module for location tracking and implements cellular data transmission with a global system for mobile communication (GSM) module, supporting general packet radio service (GPRS) operation. Additional non-volatile memories are also available.

Sensor data are collected by the MCU through serial interfaces and stored internally. Data are periodically transferred on the Internet through the cellular connection by exploiting the Message Queuing Telemetry Transport (MQTT) application-layer protocol, designed for efficient messaging in IoT systems on top of transmission control protocol (TCP)/internet protocol (IP), and leveraging TCP's reliable data transmission. Data is managed on the cloud by dedicated servers, including a remote database server, and a web server, which allows client-server applications to retrieve data and set sensor parameters. To facilitate user accessibility, the ecosystem features a user-friendly web-based application, used in this work, which provides authenticated access to the acquired data via the Internet, ensuring general data protection regulation (GDPR) compliance and incorporating an encryption layer.

IV. Methodology

A. Data collection

The datasets used in this work include different types of data: sensor data, external environment data, and user information. For the acquisition of internal variables, the sensor nodes are placed in different areas within the Cesena Campus of the University of Bologna, as shown in Fig. 2, which is characterized by an A3 building energy rating (BER), corresponding to a yearly primary energy consumption of 50 up to 75 kWh per square meter. Specifically, the first sensor (S_1) is relocated weekly between two rooms, room A (R_A) and room B (R_B), while the second sensor (S_2) remains in a third room C (R_C), for the entire acquisition period. Student Room R_A and Cafeteria Room R_B with 425 m^2 and 246.2 m^2 respectively have a larger peak concentration of individuals of more than 100 people, while Staff Room R_C, area around 487.5 m^2, hosts the static sensor with a sparse population of typically 26 people. Additional data regarding external T and RH are collected through the free VisualCrossing database [43]. A measurement campaign is set up: sensor data are continuously collected every 20 minutes and uploaded to the cloud every 2 hours for a period of 21 days during working days and hours, from 09:00 to 18:00. External parameters are collected hourly during the same period. The measured quantities used in this study are summarized in Table. II. Furthermore, the occupancy of the rooms is manually evaluated daily on working days from 09:00 to 18:00, at four different times: 09:00, 12:00, 15:00, and 18:00.

B. Assessment of well-being

Typically, two metrics are used to assess environmental well-being: the first is a self-rated overall satisfaction on a scale ranging from 1 (not satisfied) to 10 (perfectly satisfied); the second is the PMV, introduced by Fanger [19], which represents the average thermal sensation rating on a standard scale for a large group of individuals under various combinations of thermal environmental variables [44]. More specifically, PMV represents the average response of a significant number of individuals on a 7-point thermal sensation scale ranging from +3 (hot) to -3 (cold), with 0 indicating a neutral sensation. To evaluate the thermal comfort index PMV, the calculation methods from ANSI/ASHRAE Standard 55 [45] and ISO 7730 [46] can be used. The model and equations for calculation of PMV can be found in [47], [48].

During the measurement campaign, both metrics are determined through a questionnaire automatically sent at the times disclosed in Section IV-A, to a group of users through an automated process based on messaging applications, to ensure minimal intrusion and to increase the likelihood of a higher response rate. A group of 46 individuals is employed throughout the test for room R_A, 27 for room R_B, and 16 for room R_C. The data are treated in an aggregated manner.

V. Data Preprocessing and Predictive Models

In general, we aim to establish a relationship between a j-th measured wellness y_j (a broader concept encompassing both thermal and air quality aspects that influence overall well-being assessment), with $j = 1, \ldots, N$, and the environmental variables x_{ij}, where $i = 1, \ldots, K$, using a linear or nonlinear model $\mathcal{F}$ such that $y_j = \mathcal{F}(x_{1j}, \ldots, x_{Kj})$. In our case, $K = 11$, representing variables such as T, RH, CO_2, VOCs, acquisition time, and others. Moreover, N refers to the number of samples used to build the data set. Once the model is assessed, we can predict wellness $\hat{y}_p$ for a new set of P samples using the same model, such that $\hat{y}_p = \mathcal{F}(x_{1p}, \ldots, x_{Kp})$, where $p = 1, \ldots, P$. More details on the algorithms selected are provided in Appendix I.

A. Model selections

The analysis procedures for selecting the preprocessing method are applied to the calibration data of the predictive models, detailing the metrics obtained under three basic approaches: no preprocessing, mean-centering, and auto-scaling. After extensive trials, we selected auto-scaling due to its consistently low root mean square error of cross-validation (RMSECV) and root mean square error in prediction (RMSEP) across all the techniques used in the subsequent analysis.

© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2025.3545360

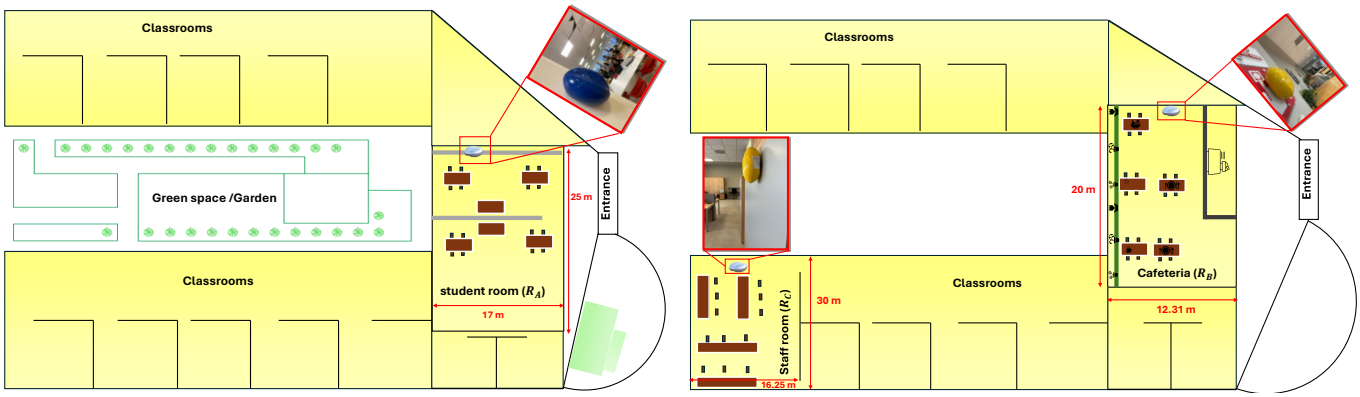


Fig. 2. On-site pictures of the INAIR devices on a simplified map of the Cesena campus: (a) Student Room (R_A) on the ground floor; (b) Cafeteria Room (R_B) and Staff Room (R_C) on the first floor.

- PCA setup: PCA is utilized to identify correlations between physical quantities and their levels, as well as to assess statistical differences among various locations.
- PLS setup: PLS is used to predict satisfaction. The model is calibrated using four principal components (PCs). The non-linearity is particularly intrinsic between the satisfaction scores and the environmental variables. In addition, to predict the optimal T that maximizes thermal comfort, a new data set is constructed that includes all environmental variables except T, which is replaced by the PMV scores. The decision to use PMV scores instead of satisfaction scores is driven by the inherent non-linear relationship between satisfaction and T, as can be seen in Fig. 4. For example, low scores can indicate either a feeling of excessive coldness or a feeling of excessive heat. After building the model, the regression vector ($\boldsymbol{\beta}$) is extracted to represent the linear relationship between the new data and the optimal T.
- LWR setup: LWR is utilized in the same way as PLS, but differently it best handles the non-linearity of the data. The model is calibrated using 10 local points and 4 PCs.

For better data visualization, a few initial PCs often capture most of the variance, making a specific number unnecessary [49]. However, when predicting variables from new data, especially with models like PLS, or LWR, a stricter selection of variables is crucial to distinguish useful information from noise. Cross-validation (CV) is used to achieve this since it allows testing the prediction ability of the model without needing other acquisitions [50]. The central concept of the CV is to partition the input datasets into subsets, set them aside, and construct a model with the remaining data. In each round of CV, a portion of the dataset (i.e., the test set) is set aside, and the model is trained on the remaining data (i.e., the training set). The trained model is then validated on the excluded data points, ensuring that each test evaluates the model's performance on unseen data. Specifically, CV divides the dataset into J subsets. For each subset ($j = 1, \ldots, J$), the model is trained on the data from the remaining $J-1$ subsets and tested on the j-th subset. Finally, the omitted subset is used as input to the model and the predicted results are compared with the actual values, allowing an initial estimate of the error to be made using the predicted residual sums of square (PRESS) parameter calculated as:

$$PRESS = \sum_{i=1}^{I} \sum_{j=1}^{J} (e_{ij}^{r})^2 \tag{1}$$

I is the total number of observations. where e_{ij}^{r} is the prediction error for the j-th response variable of the i-th observation using a number r of components. From that, the RMSECV is determined:

$$RMSECV = \sqrt{\frac{PRESS}{IJ}}. \tag{2}$$

Finally, it is important to emphasize that for practical applications it is highly recommended to evaluate the predictive performance of the model using an external test set that has not been used in either the calibration or the CV phase. The parameter derived from this process is called RMSEP. When the test set is carefully selected, RMSEP serves as the parameter that provides the most accurate insight into the predictive ability of the model. The analysis procedures include selecting the preprocessing methods applied to the calibration data of the predictive model detailing the metrics obtained under three basic preprocessing methods: no preprocessing, mean centering, and autoscaling. As previously mentioned, after extensive testing, we decided to use autoscaling for all techniques, as it gave the best results in terms of RMSECV and RMSEP for both the CV and test sets, respectively.

B. Model, data structure and mathematical tools

In the end, we have the matrix $\mathbf{Y}$ containing the satisfaction score, obtained through a questionnaire. In the same periods as before, the matrix of independent variables $\mathbf{X}$ contains $N = 3024$ observations, and $K = 11$ variables are constructed, each with its respective unit of measurement, as shown in Table II. Questionnaire scores recorded four times per day are linearly interpolated to estimate values at 20-minute intervals, assuming slow changes over time. The data are collected as previously mentioned in Section IV-A and then preprocessed to remove any possible artifacts or outliers

© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted,
but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2025.3545360

caused by involuntary touches to the device. Therefore, all observations with at least one variable outside the 99% range are eliminated. To account for the different sampling times, each observation is linearly resampled under the assumption of local slow dynamics, allowing for comparison between observations. Two class groups are created, with the first class associating each observation with a specific location. The second class is defined on the basis of different times of day, as mentioned above.

Fig. 3 shows the variation of the variable across rooms and time. The full dataset was used to create both the training and test sets, with 80% allocated for training and 20% for testing, randomly selected. CV was performed using a Venetian blind method with $s = 10$ splits and uniform thickness for all analyses performed on the data. More generally, for a dataset of size N, applying s splits with thickness t, each data entry at index $((k-1)\cdot s + t + (m-1)\cdot t)$ is assigned to the same CV group. Here, k ranges from 1 to $\lfloor N/s \rfloor$, and m ranges from 1 to $\lfloor s/t \rfloor$ [51].

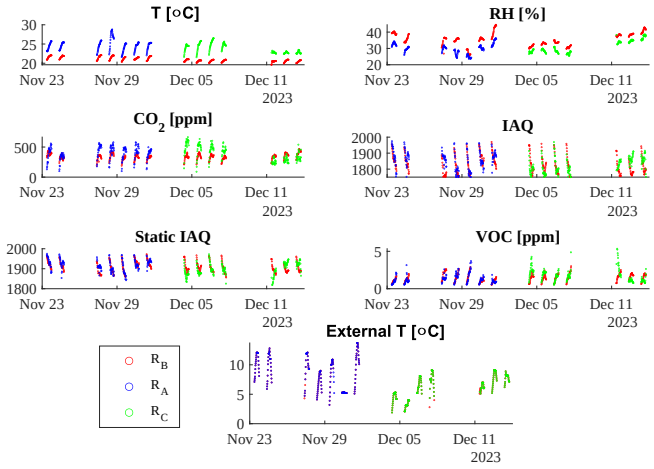


Fig. 3. Variables versus time of acquisition for each room.

VI. Results and Discussion

A. Satisfaction prediction

The first statistical analysis performed on the preprocessed data is the PCA (with $A = 4$ PCs) applied to the entire data set to visualize the correlation between the variables. The correlation analysis and the loadings plot show an expected correlation between T and people's satisfaction. In particular, as shown in Fig. 4, there is a clear negative correlation between the two parameters, a positive one between T and PMV, and a complete uncorrelation with air quality and VOC concentration. The latter is evidence that although VOC are potentially harmful to organisms, they may not be directly perceived by humans. The use of the PLS model is first aimed at predicting satisfaction scores. However, it also serves as a way to identify the relationship between satisfaction and the variables, mainly T. It is constructed using 80% of the original data, randomly selected with 4 PCs, and then applied to the remaining data set for validation. A CV is performed. The error for CV is $RMSECV = 0.6204$. Applying the model

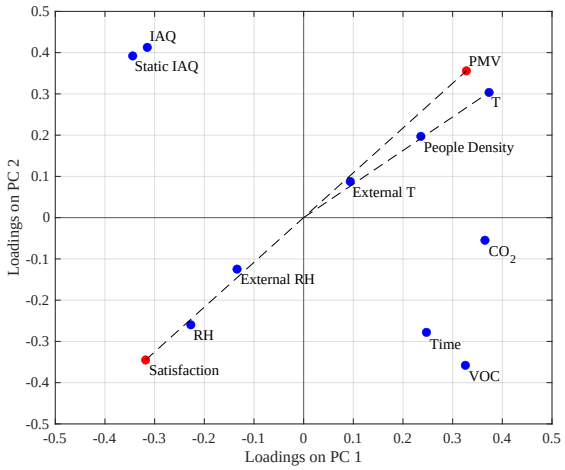


Fig. 4. PCA loadings plot of the first two PCs of all the variables of Table II.

to the validation set yields an error of $RMSEP = 0.6261$. In Fig. 5, a comparison is made between the actual measured satisfaction scores and the predicted scores. As expected, the model shows suboptimal performance on the scores due to the inherent nonlinearity between the model inputs and outputs. The error on the validation set is $RMSEP = 0.4688$.

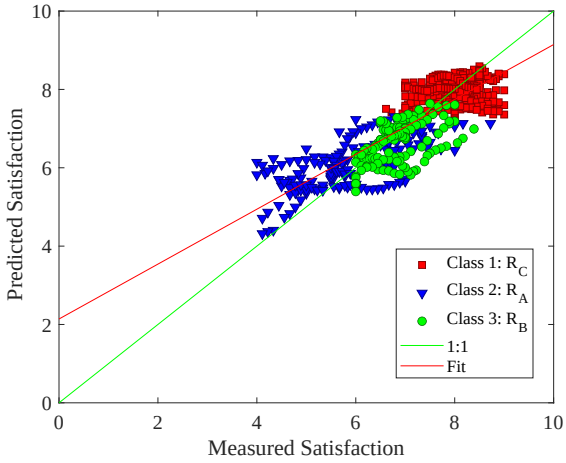


Fig. 5. Predicted vs. observed satisfaction of PLS model.The green and the red lines are the ideal and the model's fit respectively.

Among the new contributions, the study uses LWR as a technique for model improvement. Recall that the LWR model is a non-parametric model that is best suited for nonlinear problems. The LWR model is built with 4 PCs and 10 local points. The obtained CV error is $RMSECV = 0.4843$, while the prediction error is $RMSEP = 0.0960$. Fig. 6 shows the regression curve generated by the model. This observation supports the hypothesis of non-linearity between the input and output of the model, due to the intrinsic nature of satisfaction scores. For example, consider the reference observation indexed as '40', its nearest neighbors, listed as '$39-38-37-11-36-12-17-13$', and highlight the specific observations most closely associated with it in score space. Each neighbor contributes to the prediction based on

© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2025.3545360

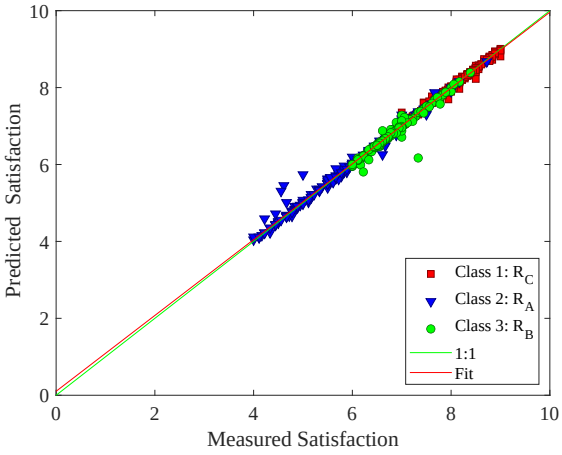


Fig. 6. Predicted vs. observed satisfaction in the LWR model.

its proximity to '40', with closer observations receiving higher weights through a kernel function. This insight allows for a nuanced understanding of the local context surrounding the observations and informs the weighting scheme used in the LWR analysis.

B. Optimal T prediction

This section aims to predict the optimal T at which the percentage of dissatisfied individuals in the R_C and target rooms is minimized, or equivalently, at which the average satisfaction is maximized. This is done by exploiting the linear relationship between the PMV scores and the other quantities, especially the T. This relation is the first assumed prior, since the definition of PMV already distinguishes between cold and hot feelings. To achieve this, another PLS model is developed using the variables RH, external RH, external T, CO_2, IAQ, static IAQ, VOC, people density, and PMV to predict the optimal T by considering 6 PCs. This model learns to predict the T values through the contribution of the environment and the PMV. A second data set is constructed as a copy of the first, except that all PMV values are set to 0. The reasoning behind this is that when the PMV is at the ground level, it corresponds to the ideal satisfaction. The newly created data set is applied to the previously calibrated model to predict the optimal T. The entire framework is shown in Fig. 7. The decision to utilize PMV scores instead of satisfaction scores is predicated on two factors. Primarily, there is an inherent non-linear relationship between satisfaction and T as shown in Fig. 4. Consequently, a given level of low satisfaction may correspond to either a high or low T level. Secondly, the PMV model is a standardized measure, with a neutral value representing the maximum satisfaction in a room and accounting for the thermal sensation.

To calculate the satisfaction rate a PLS model is built using the variables (Optimal T, RH, external T, external RH, VOCs, CO_2, IAQ and static IAQ to predict the PMV. The predicted PMV values are then used to evaluate the PPD according to Fanger's model. Thus, if the PMV values are in the range $[-0.5, +0.5]$ (or equivalent for PPD ($\leq$ 10%)) [19], [45], [46], it indicates that the individuals find the environment satisfactory. Fig. 8 shows a comparison between the predicted and measured T. The analysis reveals significant p-values: $p_C = 0.0214$, $p_A = 2.17e-10$, and $p_B = 6.01e-12$, thus indicating that the model's predictions closely align with the measured values. The distribution of PMV scores, shown in Fig. 9, has a mean of 0.10 and a standard deviation of 0.17, indicating overall satisfaction. The predicted satisfaction percentage, based on the area under the curve within the optimal PMV, is calculated to be 98.82%. The model led to significant improvements in occupant satisfaction after the prediction of the optimal T for each study space. Specifically, for R_C satisfaction increased from 58.48% to 100%, for R_A from 5.18% to 100%, and for R_B from 12.69% to 91.53% demonstrating the model's effectiveness in enhancing comfort across varied environments. While ML models, such as those used in [32], offer valuable insights into thermal comfort prediction, their statistical results (e.g., p-value = 0.106) and higher variability (standard deviation = 4.058) highlight limitations in precision and consistency. In contrast, our MVA-based framework demonstrates significantly lower variability (standard deviation = 0.17) and stronger statistical significance, delivering high-precision predictions that are well-suited for IoT environments. Although the sensors in this study were carefully deployed to account for different environmental conditions, people densities expanding the dataset to include a wider variety of building types and regions in future work would further strengthen the model's generalizability and adaptability to diverse real-world conditions.

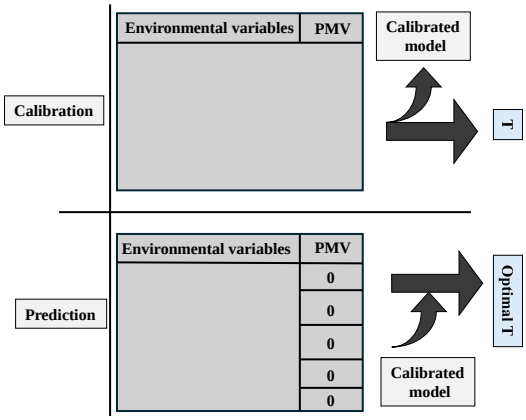


Fig. 7. Framework of the optimal temperature prediction model.

VII. Conclusion

This study introduced an innovative modeling framework that integrates IoT technologies with MVA techniques to assess indoor comfort and predict the optimal T for maximizing occupant satisfaction. The study employed both linear PLS and non-linear LWR regression models. The LWR model demonstrated superior predictive accuracy, achieving a RMSEP of 9.6%. After predicting the optimal T, the study shows that applying it would significantly improve occupant satisfaction levels across different rooms, increasing from 5.18% to 100% in R_A, 12.69% to 91.53% in R_B, and 58.48% to 100% in R_C .

© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted,
but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2025.3545360

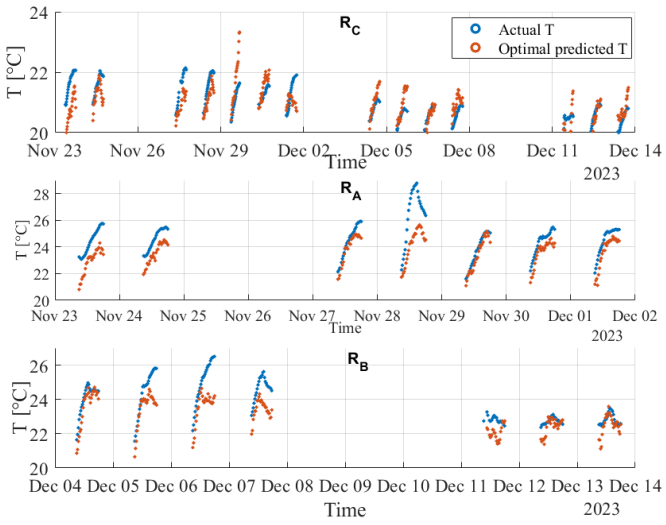


Fig. 8. The predicted versus. the actual T.

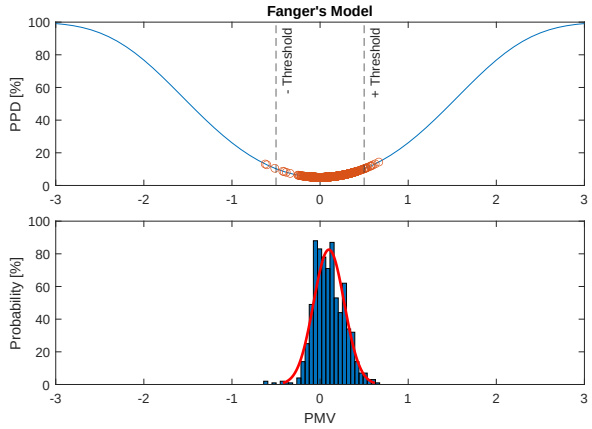


Fig. 9. PPD scores using the Fanger's model onto the predicted PMV.

The framework leverages real-time data from IoT sensors and advanced statistical methods to provide actionable insights for optimizing energy efficiency and occupant well-being. The results of MVA applied to experimental data measured by an IoT device sensing multiple environmental quantities, were presented. First, data analysis allows us to clearly understand the correlation between variables. Second, the model can classify different environments based on IoT environmental data. Third, a nonlinear predictive model of wellness is derived, which could be a basis for digital twins of indoor environments. Unlike ML models requiring convolutions, the presented models could be implemented in a sum-of-weighted product that may be easily embedded in platforms with low computational power, e.g. at the edge. This research highlights the valuable contribution of combining IoT and MVA methodologies in understanding the intricacies of data, leading to a better understanding of the psycho-physical facets of individuals' daily experiences. The analysis conducted at the building level shows that environmental data and the model can reliably predict users' perception of comfort. As the next step, the model can help understand how to operate the building to maximize comfort or identify trade-offs with consumed energy to allow savings. Then, the proposed framework enables real-time control and integration into digital twin environments by applying predictive modeling to dynamic HVAC adjustments. The approach is scalable to various building types by incorporating more training data.

However, a key challenges of this study are the data scope: the study was conducted in three distinct locations within the same campus, each differing significantly in terms of occupancy patterns and population diversity. A total of 3024 valid data instances were collected for 11 variables between November 22 and December 11. However, the dataset is limited to a single building, which may restrict the model's generalizability. To enhance applicability, future research should expand data collection to include a broader range of buildings, such as industrial facilities and commercial spaces, while also extending the study duration to capture long-term trends. Additional aspects to consider are seasonality and HVAC scheduling: the HVAC system on the considered building follows a predefined operational schedule, consisting of a hot season, a cold season, and two intermediate periods. Due to time constraints, this study was limited to the cold season, preventing an assessment of the model's adaptability to different climatic conditions. To improve its robustness, future studies should incorporate data across various seasons. Additionally, it is unable to fully capture the time-dependent nature of environmental variables. Future research will aim to further optimize the methods discussed by incorporating the time-dependent nature of environmental variables using dynamic approaches. The feasibility of implementation at the edge will be explored, enabling local real-time data collection, analysis, and control with IoT technologies, e.g. by pursuing real-time adjustments of T and RH, through smart connected thermostats. An additional development includes evolving the system into a cloud-based digital twin for continuous indoor environment optimization. This approach will ensure dynamic optimization of occupant comfort and well-being based on real-time user preferences and environmental conditions.

APPENDIX I
MULTIVARIATE ANALYSIS TECHNIQUES

In this appendix, a mathematical summary of multivariate analysis techniques used in this paper is provided.

1) PCA: is a MVA method used to analyze a data where observations are characterized by multiple inter-correlated quantitative dependent variables [50], [52]. In this work, it is applied to uncover connections between the physical variables shown in Table II and occupant satisfaction indicators, as defined previously. Each of the N observations can be represented as a single point in a K-dimensional space, where K is a number of variables. When K becomes too big, it is often difficult to get insight into the data. PCA solves this problem by finding a new A-dimensional reference system (A is the number of PCs, where generally $A < K$) that still contains most of the total variance. From a geometrical point of view, these new directions called PCs, define an A-dimensional subspace of the original K-dimensional data space, where the amount of information that each new PC adds, in the form of

© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2025.3545360

explained variance, is included in descending order in the PCs. Generally, a data matrix $\mathbf{X}(N \times K)$ can be decomposed into two arbitrary matrices, $\mathbf{T}(N \times A)$ (scores) and $\mathbf{W}(A \times K)$ (loadings), so that $\mathbf{T} = XW'$, where $\mathbf{W}$ is an orthonormal matrix and $(\cdot)'$ is the transpose operator. Following classical regression theory, it could be shown that an optimal matrix $\mathbf{W_{opt}}$ minimizes the error residual matrix $\mathbf{E}(N \times K)$ as presented in the following [50]:

$$\mathbf{X} = \mathbf{T}\mathbf{W}'_{\mathbf{opt}} + \mathbf{E}. \tag{3}$$

The objective is to discover the optimal projection of $\mathbf{X}$ (for prediction) along the directions of loadings, accomplished through a dot product operation as $\hat{\mathbf{X}} = \mathbf{T}\mathbf{W}'_{\mathbf{opt}}$, where $\hat{\mathbf{X}}$ corresponds to the approximation of $\mathbf{X}$ using A PCs. It is demonstrated that the loadings correspond to the eigenvectors of the correlation matrix $\mathbf{X}\mathbf{X}'$.

2) PLS: In contrast to PCA, PLS is tailor-made for regression analysis. Alongside the $\mathbf{X}$ datasets, it incorporates an output matrix $\mathbf{Y}(N \times M)$ where M represents the number of outputs. This matrix contains values of the variable of interest associated with each of the N observations from the $\mathbf{X}$ dataset. PLS is used to predict the wellness metrics, namely satisfaction and PMV by uncovering the linear relationship between the $\mathbf{X}$ and $\mathbf{Y}$ matrix. PLS discerns slightly distinct directions from PCs, referred to as latent variables (LVs), that attempt to simultaneously maximize the variance of $\mathbf{X}$, the variance of $\mathbf{Y}$, and the covariance between the two. Mathematically, this can be explained by the fact that the score matrix in PLS predicts not only $\mathbf{X}$ but also $\mathbf{Y}$ [53]. PLS has been developed from chemometrics applications but has been successfully used for extracting information from sensors for a wide range of applications, from mass spectroscopy to precision agriculture and infrastructure engineering [54]–[56].

Mathematically, in PLS the score matrix $\mathbf{T}$ of the $\mathbf{X}$ matrix is not only a good predictor of $\mathbf{X}$ but also of $\mathbf{Y}$ $(N \times M)$:

$$\mathbf{Y} = \mathbf{T}\mathbf{C}' + \mathbf{F} \tag{4}$$

where $\mathbf{F}\,(N \times M)$ is the error matrix that indicates the difference between the observed values in $\mathbf{Y}$ and the values predicted by the model. In (4), from a geometrical point of view, $\mathbf{C}\,(M \times A)$ is a matrix that identifies the directions in the A-dimensional subspace $\mathbf{T}$ that maximize the covariance with the matrix $\mathbf{Y}$. Therefore, (3) and (4) can be merged so that the model estimates an array of coefficients $\boldsymbol{\beta}(K \times M)$ which allows an easy prediction of the variable of interest, $\hat{\mathbf{Y}}$ starting from a newly acquired observation, $\mathbf{X}$:

$$\hat{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta}. \tag{5}$$

3) LWR: A non-parametric regression technique based on the premise that smooth, continuous non-linear functions can be locally approximated by piecewise linear functions. It uses local linear approximations to predict a dependent variable $\mathbf{Y}$ from independent variables $\mathbf{X}$, based on a specified number of PCs [57]–[60]. For each sample in the dataset, the regression selects the q-closest points, and these points are assigned weight values based on their distance to $\mathbf{X}_i$. The regression is performed concerning these weighted values, the value of the fitted regression surface at $\mathbf{X}$ is then called the fitted value, i.e. $g(x_i)$ so that:

$$y_i = g(x_i) + \epsilon_i \tag{6}$$

where ϵ is an independent normal variable with a null mean. The method estimates the function $g(x)$ using only a limited number of samples, $g(x)$ is locally linear within each subset of points but can represent a non-linear function overall. from the calibration set within a certain distance threshold from the reference point $\mathbf{X}_i$. The chosen metric for selecting the points is the Mahalanobis distance since it considers the correlation between scores with the following equation: [61]

$$D_M = \sqrt{(x-\mu)^T \mathbf{S}^{-1}(x-\mu)} \tag{7}$$

where μ is the mean value and $\mathbf{S}$ is the covariance matrix, more details of the LWR kernel function can be found in [62].

REFERENCES

[1] Y. Chida and A. Steptoe, "Positive psychological well-being and mortality: a quantitative review of prospective observational studies," *Psychosomatic medicine*, vol. 70, no. 7, pp. 741–756, 2008.
[2] D. C. Ganster and C. C. Rosen, "Work stress and employee health: A multidisciplinary review," *Journal of management*, vol. 39, no. 5, pp. 1085–1122, Feb. 2013.
[3] J. Saini, M. Dutta, and G. Marques, "Sensors for indoor air quality monitoring and assessment through internet of things: a systematic review," *Environmental Monitoring and Assessment*, vol. 193, no. 2, p. 66, Jan. 2021.
[4] M. Arif, M. Katafygiotou, A. Mazroei, A. Kaushik, E. Elsarrag *et al.*, "Impact of indoor environmental quality on occupant well-being and comfort: A review of the literature," *International Journal of Sustainable Built Environment*, vol. 5, no. 1, pp. 1–11, June. 2016.
[5] R. A. Kiefer, "An integrative review of the concept of well-being," *Holistic nursing practice*, vol. 22, no. 5, pp. 244–252, Sept. 2008.
[6] M. Belloni, L. Carrino, and E. Meschi, "The impact of working conditions on mental health: Novel evidence from the uk," *Labour Economics*, vol. 76, p. 102176, Jun. 2022.
[7] International Labour Organization, "Workplace health promotion and well-being (occupational safety and health)," 2019, Accessed: Apr. 23, 2024. [Online]. Available: https://www.ilo.org/safework/areasofwork/workplace-health-promotion-and-well-being/lang--en/index.htm
[8] E. Oldham and H. Kim, "Ieq field investigation in high-performance, urban elementary schools," *Atmosphere*, vol. 11, no. 1, p. 81, Jan 2020.
[9] N. R. Kapoor, A. Kumar, T. Alam, A. Kumar, K. S. Kulkarni, and P. Blecich, "A review on indoor environment quality of indian school classrooms," *Sustainability*, vol. 13, no. 21, p. 11855, Oct. 2021.
[10] S. S. Korsavi, A. Montazami, and D. Mumovic, "The impact of indoor environment quality (ieq) on school children's overall comfort in the uk; a regression approach," *Building and Environment*, vol. 185, p. 107309, Nov. 2020.
[11] A. Ortiz Perez, B. Bierer, L. Scholz, J. Wöllenstein, and S. Palzer, "A wireless gas sensor network to monitor indoor environmental quality in schools," *Sensors*, vol. 18, no. 12, p. 4345, Dec. 2018.
[12] M. Muladi, S. Sendari, and T. Widiyaningtyas, "Real time indoor air quality monitoring using internet of things at university," in *2nd Borneo International Conference on Applied Mathematics and Engineering (BICAME)*, Balikpapan, Indonesia, Dec. 2018, pp. 169–173.
[13] J. Weng, Y. Zhang, Z. Chen, X. Ying, W. Zhu, and Y. Sun, "Field measurements and analysis of indoor environment, occupant satisfaction, and sick building syndrome in university buildings in hot summer and cold winter regions in china," *International Journal of Environmental Research and Public Health*, vol. 20, no. 1, p. 554, Dec. 2022.
[14] S. Van Der Valk, T. Myers, I. Atkinson, and K. Mohring, "Sensor networks in workplaces: Correlating comfort and productivity," in *IEEE Tenth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)*, Singapore, Apr. 2015, pp. 1–6.
[15] S. J. Emmerich, K. Y. Teichman, and A. K. Persily, "Literature review on field study of ventilation and indoor air quality performance verification in high-performance commercial buildings in north america," *Science and Technology for the Built Environment*, vol. 23, no. 7, pp. 1159–1166, Mar. 2017.

© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted,
but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2025.3545360

[16] W. T. Honeycutt, M. T. Ley, and N. F. Materer, "Precision and limits of detection for selected commercially available, low-cost carbon dioxide and methane gas sensors," *Sensors*, vol. 19, no. 14, p. 3157, July. 2019.

[17] L. C. Tagliabue, F. R. Cecconi, S. Rinaldi, and A. L. C. Ciribini, "Data driven indoor air quality prediction in educational facilities based on iot network," *Energy and Buildings*, vol. 236, p. 110782, Apr. 2021.

[18] Z. Gou, D. Prasad, and S. S.-Y. Lau, "Are green buildings more satisfactory and comfortable?" *Habitat International*, vol. 39, pp. 156–161, July. 2013.

[19] P. Fanger, "Thermal comfort. analysis and applications in environmental engineering." *Thermal comfort. Analysis and applications in environmental engineering.*, 1970.

[20] J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang, and W. Zhao, "A survey on internet of things: Architecture, enabling technologies, security and privacy, and applications," *IEEE internet of things journal*, vol. 4, no. 5, pp. 1125–1142, Mar. 2017.

[21] L. Liu, Z. Chang, X. Guo, S. Mao, and T. Ristaniemi, "Multiobjective optimization for computation offloading in fog computing," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 283–294, Dec. 2017.

[22] W. Hu, Y. Wen, K. Guan, G. Jin, and K. J. Tseng, "itcm: Toward learning-based thermal comfort modeling via pervasive sensing for smart buildings," *IEEE Internet of Things Journal*, vol. 5, no. 5, pp. 4164–4177, July. 2018.

[23] E. Laftchiev and D. Nikovski, "An iot system to estimate personal thermal comfort," in *IEEE 3rd World Forum on Internet of Things (WF-IoT)*, Reston, VA, USA, Dec. 2016, pp. 672–677.

[24] S. Casaccia *et al.*, "Measurement of users' well-being through domotic sensors and machine learning algorithms," *IEEE Sensors Journal*, vol. 20, no. 14, pp. 8029–8038, Mar. 2020.

[25] Q. Jiang, J. Liu, and X. Yang, "Optimization of indoor quality and thermal comfort for university classrooms using data-based machine learning," in *E3S Web of Conferences*, vol. 546. EDP Sciences, July. 2024, p. 02011.

[26] W. Zhang, W. Hu, and Y. Wen, "Thermal comfort modeling for smart buildings: A fine-grained deep learning approach," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2540–2549, Sept. 2018.

[27] Y. T. Acquaah, B. Gokaraju, R. C. Tesioro, G. Monty, and K. Roy, "Occupancy and thermal preference-based hvac control strategy using multisensor network," *IEEE Sensors Journal*, vol. 23, no. 11, pp. 11 785–11 795, Apr. 2023.

[28] L. Yu, S. Chen, S. Qin, and Y. Liu, "Deep meta-learning for personal thermal comfort modeling in office buildings," in *China Automation Congress (CAC)*. Xiamen, China: IEEE, Nov. 2022, pp. 3988–3993.

[29] W. Zhang, P. W. Tien, J. Calautit, and Y. Wu, "Building occupancy prediction through machine learning for enhancing energy efficiency, air quality and thermal comfort: Review and case study," in *Proc. Applied Energy Symposium 2021: Low carbon cities and urban energy systems*, Matsue, Japan, Sep. 2021.

[30] S. A. Abdel-Razek, H. S. Marie, A. Alshehri, and O. M. Elzeki, "Energy efficiency through the implementation of an ai model to predict room occupancy based on thermal comfort parameters," *Sustainability*, vol. 14, no. 13, p. 7734, June. 2022.

[31] W. Hu, Y. Luo, Z. Lu, and Y. Wen, "Heterogeneous transfer learning for thermal comfort modeling," in *Proceedings of the 6th ACM international conference on systems for energy-efficient buildings, cities, and transportation*, Nov. 2019, pp. 61–70.

[32] S. U. Rehman *et al.*, "Personalisedcomfort: a personalised thermal comfort model to predict thermal sensation votes for smart building residents," *Enterprise Information Systems*, vol. 16, no. 7, p. 1852316, Nov. 2022.

[33] H. Hijry *et al.*, "Real time worker stress prediction in a smart factory assembly line," *IEEE Access*, Aug. 2024.

[34] A. Soleimanijavid, I. Konstantzos, and X. Liu, "Challenges and opportunities of occupant-centric building controls in real-world implementation: A critical review," *Energy and Buildings*, p. 113958, Apr. 2024.

[35] T. Rato *et al.*, "A systematic comparison of pca-based statistical process monitoring methods for high-dimensional, time-dependent processes," *AIChE Journal*, vol. 62, no. 5, pp. 1478–1493, Sept. 2016.

[36] X. Chang, F. Nie, Y. Yang, C. Zhang, and H. Huang, "Convex sparse pca for unsupervised feature learning," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 11, no. 1, pp. 1–16, July. 2016.

[37] M. S. Mahdavinejad, M. Rezvan, M. Barekatain, P. Adibi, P. Barnaghi, and A. P. Sheth, "Machine learning for internet of things data analysis: A survey," *Digital Communications and Networks*, vol. 4, no. 3, pp. 161–175, Aug. 2018.

[38] H. Ij, "Statistics versus machine learning," *Nat Methods*, vol. 15, no. 4, p. 233, Apr. 2018.

[39] S. Makridakis, E. Spiliotis, V. Assimakopoulos, A.-A. Semenoglou, G. Mulder, and K. Nikolopoulos, "Statistical, machine learning and deep learning forecasting methods: Comparisons and ways forward," *Journal of the Operational Research Society*, vol. 74, no. 3, pp. 840–859, 2023.

[40] Tāua s.r.l. società benefit, "Tāua - internet of things and web of things solutions," 2024. [Online]. Available: https://www.taua.it/en/

[41] Bosch Sensortec, "Bme680 integrated environmental unit," 2024. [Online]. Available: https://www.bosch-sensortec.com/products/environmental-sensors/gas-sensors/bme680/

[42] G. C. Pratt, K. Palmer, C. Y. Wu, F. Oliaei, C. Hollerbach, and M. J. Fenske, "An assessment of air toxics in Minnesota." *Environmental Health Perspectives*, vol. 108, no. 9, pp. 815–825, Sept. 2000.

[43] Visual Crossing Corporation, "Weather data & weather api," 2024. [Online]. Available: https://www.visualcrossing.com/

[44] J. Van Hoof, M. Mazej, and J. L. Hensen, "Thermal comfort: research and practice," *Frontiers in Bioscience-Landmark*, vol. 15, no. 2, pp. 765–788, 2010.

[45] ASHRAE, "Thermal environmental conditions for human occupancy," 2023. [Online]. Available: https://www.ashrae.org/technical-resources/bookstore/standard-55-thermal-environmental-conditions-for-human-occupancy.

[46] International Organization for Standardization, "Ergonomics of the thermal environment - analytical determination and interpretation of thermal comfort using calculation of the pmv and ppd indices and local thermal comfort criteria," https://webstore.ansi.org/standards/iso/isodis77302023, 2023, iSO/DIS 7730:2023.

[47] S. Zare, N. Hasheminezhad, K. Sarebanzadeh, F. Zolala, R. Hemmatjo, and D. Hassanvand, "Assessing thermal comfort in tourist attractions through objective and subjective procedures based on iso 7730 standard: A field study," *Urban climate*, vol. 26, pp. 1–9, Dec. 2018.

[48] B. E. Ainsworth, W. L. Haskell, A. S. Leon, D. R. Jacobs Jr, H. J. Montoye, J. F. Sallis, and R. S. Paffenbarger Jr, "Compendium of physical activities: classification of energy costs of human physical activities." *Medicine and science in sports and exercise*, vol. 25, no. 1, pp. 71–80, Jan. 1993.

[49] R. Bro and A. K. Smilde, "Principal component analysis," *Analytical methods*, vol. 6, no. 9, pp. 2812–2831, Mar. 2014.

[50] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, Aug. 1987.

[51] S. Wold, "Cross-validatory estimation of the number of components in factor and principal components models," *Technometrics*, vol. 20, no. 4, pp. 397–405, Apr. 1978.

[52] L. Pucci, E. Testi, E. Favarelli, and A. Giorgetti, "Human activities classification using biaxial seismic sensors," *IEEE Sens. Lett.*, vol. 4, no. 10, pp. 1–4, Aug. 2020.

[53] S. Wold, M. Sjöström, and L. Eriksson, "Pls-regression: a basic tool of chemomedtrics," *Chemometrics and intelligent laboratory systems*, vol. 58, no. 2, pp. 109–130, Oct. 2001.

[54] L. Franceschelli *et al.*, "Real-time gas mass spectroscopy by multivariate analysis," *Scientific Reports*, vol. 13, no. 1, p. 6059, Apr. 2023.

[55] ——, "Non-intrusive microwave technique for direct detection of concrete compressive strength monitoring by multivariate modeling," *Measurement*, vol. 206, p. 112332, Jan. 2023.

[56] O. Afif *et al.*, "A versatile, machine-learning-enhanced rf spectral sensor for developing a trunk hydration monitoring system in smart agriculture," *Sensors*, vol. 24, no. 19, p. 6199, Sept. 2024.

[57] Z. Wang, T. Isaksson, and B. R. Kowalski, "New approach for distance measurement in locally weighted regression," *Analytical Chemistry*, vol. 66, no. 2, pp. 249–260, 1994.

[58] T. Naes, T. Isaksson, and B. Kowalski, "Locally weighted regression and scatter correction for near-infrared reflectance data," *Analytical Chemistry*, vol. 62, no. 7, pp. 664–673, 1990.

[59] W. S. Cleveland and S. J. Devlin, "Locally weighted regression: an approach to regression analysis by local fitting," *Journal of the American statistical association*, vol. 83, no. 403, pp. 596–610, 1988.

[60] H.-G. Müller, "Weighted local regression and kernel methods for nonparametric curve fitting," *Journal of the American Statistical Association*, vol. 82, no. 397, pp. 231–238, 1987.

[61] M. Malkauthekar, "Analysis of euclidean distance and manhattan distance measure in face recognition," in *Third International Conference on Computational Intelligence and Information Technology (CIIT)*, Mumbai, India, Oct. 2013, pp. 503–507.

[62] V. Centner and D. L. Massart, "Optimization in locally weighted regression," *Analytical Chemistry*, vol. 70, no. 19, pp. 4206–4211, 1998.

© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.