Applied Thermal Engineering 273 (2025) 126556

Contents lists available at ScienceDirect

Applied Thermal Engineering

journal homepage: www.elsevier.com/locate/apthermeng

Research Paper

A comprehensive method for exploratory data analysis and preprocessing the ASHRAE database for machine learning

Amir Rahmanparast [a], Muhammed Milani [b], Muhammet Camci [c,d,*], Yakup Karakoyun [e], Ozgen Acikgoz [a], Ahmet Selim Dalkilic [a]

[a] Department of Mechanical Engineering, Mechanical Engineering Faculty, Yildiz Technical University, Istanbul 34349, Turkiye
[b] Department of Computer Sciences, Engineering and Natural Sciences Faculty, Bandirma Onyedi Eylul University, Bandirma 10200, Turkiye
[c] Department of Mechanical Engineering, Engineering Faculty, Siirt University, Siirt 56100, Turkiye
[d] Department of Mechanical Engineering, Engineering Faculty, King Mongkut's University of Technology Thonburi (KMUTT), Bangkok 10140, Thailand
[e] Department of Mechanical Engineering, Engineering Faculty, Van Yuzuncu Yil University, Van 65080, Turkiye

ARTICLE INFO

Keywords:
Random forest
SHAP
ASHRAE Global thermal comfort database
PMV
Thermal comfort
Machine learning

ABSTRACT

Thermal comfort prediction is crucial for building energy efficiency and occupant comfort. ML methods are commonly used to predict thermal comfort. This research presents a comprehensive process for exploring and preprocessing the ASHRAE Database, providing a substantial dataset comprising 107,583 records of thermal comfort observations to create ML algorithms that can estimate Fanger's PMV. With the most detailed cleaning and preprocessing stages in the literature, which included the imputation of missing values and the management of outliers, the final dataset is reduced to 55,443 records for the analyses. For practical applications and indoor comfort assessments, its estimation offers significant advantages due to its speed, ease of use, and cost-effectiveness. This study aimed to investigate which parameters are important in Fanger's PMV model and which subset of variables is best for variable selection using different feature selection and analysis methods. The T_a and T_r had a high correlation value of 0.92, indicating a robust link between these two variables. The study employed Feature importance, the SelectKBest, SHAP, P-box, and PDP analyses, which showed consistency and suggested condensing the first six elements into three, and also was validated with the Chinese Database with 41,977 entries. The study targeted three parameters: T_a, clo, and M, using less expensive and simple measurement devices. To evaluate the accuracy of the research performance, RF and SVM models were created based on these three parameters. The results indicated that they have the accuracies of 85% and 70%, respectively, which are far better than the conventional models.

1. Introduction

Forecasting thermal comfort is essential for enhancing energy efficiency in structures and safeguarding occupant well-being. Although machine learning (ML) approaches are widely employed to forecast thermal comfort, these models frequently encounter difficulties in real-time adaptation, particularly in the regulation of heating, ventilating, and air conditioning systems (HVAC). Elements include age, environmental factors, and individual comfort judgments that exacerbate the difficulty.

As interest in data-driven models grows, more complex algorithms are being used to predict thermal comfort, which has a wide range of input parameters. However, these developed models may not always be suitable for real-time applications, such as the instant control of HVAC. The difficulty of accurately estimating thermal comfort in buildings is compounded by the complex interactions between various influencing factors. Traditional HVAC systems often fail to account for real-time changes in occupancy and external conditions. As smart buildings become more common over time, there is a growing demand for energy-efficient HVAC solutions aiming both to maintain comfort and to reduce energy consumption. The ASHRAE RP-884 [1] database plays a key role in informing standards like ISO 7730 [2] and ASHRAE 55 [3], but current models still face limitations in their ability to adapt. Therefore, it can be said that there is a need for more advanced, responsive thermal comfort prediction systems.

ML techniques have become a fundamental element in the development of thermal comfort prediction systems. Despite this, thermal

* Corresponding author at: Department of Mechanical Engineering, Engineering Faculty, Siirt University, Siirt 56100, Turkiye.
E-mail address: mcamci@siirt.edu.tr (M. Camci).

https://doi.org/10.1016/j.applthermaleng.2025.126556
Received 15 November 2024; Received in revised form 22 March 2025; Accepted 21 April 2025
Available online 21 April 2025

Nomenclature

ASHRAE 55	American society of heating, refrigerating and air-conditioning engineers
BNN	Bayesian neural network
clo	Clothing insulation
EDA	Exploratory data analysis
HVAC	Heating, ventilating and air conditioning systems
ISO EN 7730	International standard
M	Metabolic rate, (W/m^2)
ML	Machine learning
PDP	Partial dependence plots
PMV	Predicted mean vote
PPD	Percentage people dissatisfied
RF	Random Forest
RH	Relative humidity, (%)
SVM	Support vector machines
T_a	Air temperature, (℃)
T_g	Globe temperature, (℃)
T_r	Mean radiant temperature, (℃)
V_a	Air velocity, (m/s)

comfort remains a subjective experience influenced by numerous factors, including environmental conditions and physiological data, which can be different from person to person in terms of age, gender, race, habits, and activity. Numerous ML-based models concentrate on a single aspect of thermal comfort, which can lead to conflicting predictions when multiple factors are considered. This matter is especially critical for at-risk groups, such as the elderly, whose heat perception may alter with age, possibly resulting in severe health hazards such as hypo- and hyperthermia.

Dear's ASHRAE Research Project 884 [1] focused on adaptive thermal comfort, involving nearly 21,000 sets of raw thermal comfort data from research groups worldwide. The data underwent cleaning and standardization processes before being incorporated into a database. Thermal comfort indicators were re-calculated implementing ASHRAE-sponsored software, and clothing insulation (clo) values were converted to ASHRAE 55. The database has potential uses beyond adaptive thermal comfort modeling, such as developing empirical thermal indices and validating laboratory-based comfort models and standards. Accessible through a World Wide Web website, the database is now in the public domain in various formats. The ASHRAE Global Thermal Comfort Database II [4] project, initiated in 2014, aims to advance HVAC science and art by utilizing open-source research resources. The database, which contains 81,846 objective indoor climate observations and subjective assessments of building occupants, is designed to help with various questions about thermal comfort in field settings. Its web-based interface allows users to filter on criteria such as building typology, occupancy type, demographic variables, indoor thermal environmental criteria, calculated comfort indices, environmental control criteria, and outdoor meteorological data.

The literature review below summarizes previous research and points out research gaps. It does this by comparing existing works in a structured way and making a clear statement about unresolved problems to strengthen the study's justification. Farhan et al. [5] proposed a new ML method for determining a person's thermal comfort profile. They use a feature vector to identify the optimal feature set and train a classifier to provide a thermal sensation class. The method's accuracy is 76.7 %, two times greater than the widely used Fanger's model [6], which only achieves 35.4 %. The research also suggests that an individual's age and outside temperature, which Fanger's model [6] does not consider, significantly impact thermal comfort. Lu et al. [7] have extensively studied thermal comfort theory, focusing on both static and adaptive models. Adaptive thermal comfort emphasizes interactions between occupants and the indoor environment using data-driven methods. The ASHRAE RP884 dataset is a key achievement in this area. This study purposes to construct a thermal comfort model implementing RP 884 of the three main climate zones and simulate a tabular Q-learning temperature set-point control system. The numerical thermal comfort model has the best recall at 49.3 %, outperforming predicted mean vote (PMV)'s 43 %. Q-learning-based temperature control can bring occupants' comfort levels within reasonable bounds. According to Parkinson et al. [8], the ASHRAE Global Thermal Comfort collection II has released a quality assurance exercise on the first generation of adaptive comfort standards. The study validated the adaptive comfort model for naturally ventilated buildings using 60,321 comfort questionnaire records and measurement data. The results suggest future improvements to the standards, adaptive comfort theory, and building operating techniques. The study found that adaptive comfort mechanisms apply to all building occupants, as thermal environmental exposures drive indoor adaptation. This presents an opportunity for integrating air conditioning practices into an adaptive framework through set-point nudging in building automation systems. Wang et al.'s [9] study used the ASHRAE Global Thermal Comfort Database II to assess thermal comfort in buildings. The most common measures include thermal sensation, thermal preference, comfort, and acceptability. The study suggests that personal experience of thermal conditions in the built environment should be included in surveys. Logistic regression and Support vector machines (SVM) were used to predict thermal acceptability and preference with thermal feeling and comfort. The results provide insights into selecting subjective thermal metrics for efficient data collection on occupants' thermal experience, which could aid in designing field research, chamber experiments, and human-building interaction interfaces. Luo et al. [10] used three data sampling techniques and nine ML algorithms to predict thermal sensation votes in the ASHRAE Global Thermal Comfort Database II. Random Forest (RF) performed the best, with an accuracy of 66.3 % and 61.1 % for 3-point and 7-point votes, correspondingly. ML algorithms are typically more accurate than the PMV model. The top six crucial features for thermal sensation votes prediction are age, metabolic rate (M), clo, air velocity (V_a), relative humidity (RH), and air temperature (T_a). RF can predict thermal sensation votes with an overall accuracy of 63.6 %, just 2.6 % less accurate than using 12 input features. The study also discussed issues like splitting training and testing data, encoding techniques, and hyperparameter tweaking in ML comfort models. Zhou et al. [11] developed a self-correcting model using the RP-884 thermal comfort database and support vector machine technique. The model, which uses variables like V_a, garment insulation, M, and T_a, significantly reduced inaccurate results and improved the fitting degree by 83.7 % and the sum of squares for residuals by 96.4 % compared to the PMV model. The model can distinguish between thermal comfort comebacks in natural ventilation and air-conditioning systems, eliminating the need for separate models. An open-access platform was created to help use ML algorithms for thermal comfort data study. Gao et al. [12] proposed DeepComfort, a deep reinforcement learning-based system for controlling thermal comfort in buildings. They propose a cost-minimization problem considering both HVAC system energy consumption and occupant comfort. They present a deep deterministic policy gradient-based approach for optimal thermal comfort management and a deep feedforward neural network-based approach for occupant thermal comfort forecasting. They tested their method in a simulation environment and found that it can increase thermal comfort prediction performance by 14.5 %, lower HVAC energy consumption by 4.31 %, and increase occupant thermal comfort by 13.6 %. According to the study of Ma et al. [13], thermal comfort is a critical aspect of indoor environments, influenced by both known and unquantifiable factors. A predictive model for occupant thermal preference was created using a Bayesian neural network (BNN) algorithm, combining measurements and past knowledge. The BNN model outperformed traditional models like the adaptive comfort model and PMV but predicted less accurately.

The study suggests that integrating window opening/closing behavior and subjective evaluation measures with thermal comfort models can enhance predictive performance. Overall, thermal comfort is a critical aspect of indoor comfort. Park and Park [14] developed a prediction model for personal thermal comfort using ensemble transfer learning to overcome challenges in collecting individual features. Wearable wristbands and sensors were used to gather physiological and environmental data, and a pre-trained model was created using deep learning and ML methods. The ensemble transfer learning approach improved the accuracy of predicting thermal comfort for two target participants, with 95 % accuracy and an F1 score for the first subject and 85 % and 83 % for the second subject, respectively. Lala et al.'s [15] study used multitask learning to predict thermal comfort in elementary school classrooms. 512 students participated in tests, and DeepComfort, a deep learning model, was suggested. Tested on the ASHRAE Global Thermal Comfort database II and dataset, DeepComfort showed good F1 scores, accuracy (around 90 %), and generalization ability. It performed better than six popular metric-specific single-task ML methods, marking the first time multitask learning was used in a classroom. Bai et al. [16] conducted a study on the performance of 10 ML models, including traditional and ensemble models, in predicting building occupants' thermal comfort. The models were trained using ASHRAE Global Thermal Comfort Database II features and analyzed using RF and gradient boosting decision trees. Ensemble models showed the most significant improvement, reaching 40 % to 60 % during training. RF and deep cascade forest models also showed significant benefits in predicting thermal preferences, offering guidelines for automated thermal environment control in buildings. Lan et al.'s study [17] utilized the ASHRAE Global Thermal Comfort Database II to analyze thermal comfort in classrooms. The study identified eleven key factors influencing comfort, using a hybrid model combining Bayesian optimization and extreme gradient boosting. The findings revealed 80 % of impacts and 20 % of interacting effects, with some interactive effects being stronger than the main ones. The study aimed to develop a more robust personal comfort model to guide HVAC design and regulation development for thermal environment and energy-saving needs. Feng et al. [18] developed an ML method to accurately project building occupants' thermal comfort levels using the ASHRAE RP-884 database. They pre-processed the dataset using an imputation method and selected relevant input characteristics using the Relif algorithm. Four ensemble models were developed to assess the connection between thermal comfort levels and input features. A decision-making method was used to assess the credibility of each estimator, and the final PMV was obtained using Genetic Algorithm optimized coefficients. The results showed promising results with a low Root Mean Squared Error value of 0.157. Lamberti et al. [19] developed a data-driven model using 61,710 subjective reactions from field study environmental characteristics stored in two ASHRAE databases. The model produced two simpler and moreaccurate models, which improved prediction compared to PMV and other regression models. The models determined comfort areas at 90 %, 80 %, and 70 % of thermal acceptability, indicating that the prediction error is below the 90 % allowable range. These models enable occupant-centered HVAC control and thermal comfort estimations, allowing for more adaptable building management standards that lower energy consumption without compromising interior comfort. Yang et al. [20] analyzed data from the ASHRAE RP-884 and the Chinese Thermal Comfort Database to understand indoor thermal environment characteristics in China. They found significant differences between China and the global database. The Chinese population showed greater adaptability, with 80 % acceptable temperature ranges in cooled buildings, 15.99 °C to 27.26 °C in heated buildings, and 15.69 °C to 30.78 °C in naturally ventilated buildings. A thermal adaptation model was developed based on the Chinese Thermal Comfort Database for China's naturally ventilated structures. Park and Woo [21] presented a method to minimize the dimensionality of factors for PMV prediction using ML, aiming to facilitate rapid PMV computations without sacrificing predictive accuracy. This study identified the most significant characteristics for PMV using PCA, Best Subset, and Gini Importance, comparing each model against the others. The findings indicated that PCA and ANN attained the best accuracy of 89.70 %, while the combination of Best Subset and RF exhibited the most rapid prediction performance overall.

Some recent research has focused on data cleaning and processing for ML applications using the ASHRAE Global Thermal Comfort Database II [4]. Based on the provided literature above, which used the ASHRAE databases and implemented ML techniques, this study is necessary because it can reduce the number of input variables required for calculation, which reduces the amount of measurement equipment that needs to be set up. This paper discusses the best way to select features for Fanger's PMV [6] estimation, which can help in building a derived PMV estimation model. In addition to this, our literature review, as presented in Table 1, indicates that aside from Park and Woo's study [21], which achieved 89.7 % accuracy with 25,261 reduced data entries and three thermal comfort parameters requiring costly measurement equipment, no other study approaches our 85 % accuracy derived from 55,443 data entries utilizing three cost-effective thermal comfort parameters from the ASHRAE Database. In this context, the proposed ML-based feature selection and estimation techniques may be used to reduce the number of PMV parameters.

With comprehensive data collection, some preprocessing, including careful data analysis and preparation, is crucial to ensure the accuracy and reliability of ML models. To the best of the authors' knowledge, the literature review above indicates that no ML study has a more detailed data analysis and preprocessing than the one presented. The dataset is optimized for future modeling by thoroughly analyzing its structure, identifying patterns, and correcting issues such as missing data and anomalies. This strategy involves performing Exploratory Data Analysis (EDA) [22], understanding distributions, multicollinearity reduction, and using data cleaning methods such as standardization and coding. Furthermore, this study highlights the valid environmental parameters by filtering and carefully transforming the data, thus improving the quality and efficiency of thermal comfort prediction models. These strategies provide a robust framework for the model-building process that delivers reliable and significant results.

On the other hand, thermal comfort prediction is essential for improving energy efficiency in buildings and protecting occupant well-being. Although ML approaches are often used to predict thermal comfort, these models often face difficulties in real-time adaptation, especially in the regulation of HVAC systems. Factors such as age, environmental conditions, and individual comfort assessment exacerbate this problem. In reality, many metrics are difficult to measure, but in environments where thermal comfort is compromised, such as kindergartens or areas with multiple age groups/conditions, it becomes impossible. Therefore, it is believed that the method presented in the current study, which shows that PMV can be predicted with fewer factors, can offer advantages in terms of usability and cost. In this study, several methods showed consistency and suggested decreasing the six conventional thermal comfort elements to a smaller number of critical ones. In simple terms, in complex environments where many parameters are involved in determining thermal comfort, eliminating some of the less influential variables can significantly increase the efficiency of the system.

The study utilized Feature Importance, the SelectKBest, SHapley Additive ExPlanations (SHAP), Probability boxes (P-box), and Partial Dependence Plots (PDP) analyses, which demonstrated consistency and indicated the reduction of the first six elements into three. The research focused on three elements: T_a, clo, and M, employing cost-effective and straightforward measuring instruments. The aim of this study is to predict indoor thermal comfort as explained above using ML models developed from a reduced set of parameters in the related database. To evaluate the precision of the research performance, RF and SVM models were developed based on these three. Finally, ML-based models have high accuracy, but their limited interpretability usually causes problems

Table 1
Accuracies of some ML works using the ASHRAE Dataset for thermal comfort predictions.

Researcher	Inputs	Outputs	Methods	Dataset	Data number	Accuracy (%), R^2 (—)
The current study	clo, M, T_a	PMV	RF, SVM	ASHRAE II	55,443	85 % for RF 70 % for SVM
Park & Woo [21]	clo, M, T_g	PMV	ANN, LSTM, RF	ASHRAE II	25,261	ANN: PCA: 89.70 %; Best Subset: 88.96 %; Gini Importance: 87.83 %
Lu et al. [7]	Environmental parameters like T_a, T_{out}, V_a, RH, T_r, clo, M, etc.	TSV	KNN, RF, SVM	ASHRAE RP884	5576	KNN 49.3 % SVM 48.7 % RF 48.7 %
Farhan et al. [5]	T_a, V_a, RH, T_r, clo, M	TCV	Deep NN	ASHRAE RP884	11,164	MSE is 1.16
Wang et al. [9]	Subjective thermal comfort vote (TPV, TSV, PMV, TAV, AMV, AMP)	TCV, TPV	LR, SVM	ASHRAE II	16,795	TCV 64 % TPV 87 %
Luo et al. [10]	T_a, V_a, RH, SET, clo, M, Age, Sex, T_{out}, Season, Operation mode, building type, etc	TSV	LR, NB, ANN, KNN, AB, DT, GBM, RF, SVM	ASHRAE II	10,618	3 Point 60–66 % 7 Point 52–57 %
Ma et al. [11]	T_a, T_r, T_{out}, T_{op}, RH, RH_{out}, V_a, clo, M, Weight, Age, Operation mode, etc.	TPV	BNN	ASHRAE II	78,113	0.693

in practice due to the use of insufficient amounts of data. This study provides a comprehensive method for EDA and preprocessing in the ASHRAE Global Thermal Comfort Database, which contains 107,583 data entries, to enhance the accuracy of ML applications in predicting thermal comfort.

2. Data analysis and preprocessing

Data analysis and preprocessing are the most important parts of an ML pipeline because they highly affect the performance and reliability of the predictive model. At this stage, the dataset structure, the interrelations among variables, and any intrinsic patterns that may influence the model's accuracy are examined thoroughly. The data's essential qualities, such as primary trends, variances, and interrelationships among aspects, may be discerned from these. Furthermore, this procedure may reveal possible flaws, such as multicollinearity, class imbalance, and anomalies that might possibly impair the model's predictions. As the raw data flows through the system, it preprocesses it. It guarantees that the dataset is fit for the modeling phase and of very great quality. Normalizing, encoding categorical variables, and imputation of missing values are among preprocessing methods.

This study proceeded through the following steps: Data Preparation and Analysis (In the first phase of the research, 107,583 data points were examined from the ASHRAE Global Thermal Comfort Database. Missing data were calculated and processed for outliers accordingly. Feature Selection (Using methods such as feature importance, SelectKBest, SHAP, P-box, and PDP analyses, the most relevant features were selected to predict the PMV value. The obtained parameters were tried on two models and good results were obtained. After checking with SHAP and PDP analyses, the model with the highest accuracy score was chosen as a good candidate to guess the PMV. The accuracy of the prediction was measured by the R^2 value. HVAC systems can utilize these models to enhance thermal comfort and boost energy efficiency.

2.1. Data definition

The dataset [1,4] being analyzed consists of 107,583 data points, specifically aimed at predicting the PMV, a critical indicator of thermal comfort. The variables that affect this forecast include clo, M, T_a, Mean radiant temperature (T_r), RH, and V_a. This phase conducts a thorough analysis of the dataset, identifying significant patterns and potential issues like missing data and outliers that require resolution before building a reliable model. Table 2 presents a comprehensive summary of the dataset, encompassing the total number of observations, the proportion of missing data, and summary statistics for each variable.

This dataset outlines numerous environmental and personal factors that contribute to the prediction of PMV. However, there are several challenges that the model needs to deal with prior to its development. The presence of large proportions of missing data within the ASHRAE dataset is of great importance, particularly for factors such as T_r, which have a 70 % missing rate. This could potentially affect the reliability of the model. Besides, the existence of outliers in some variables such as T_a and V_a imply potential errors in data entry or a unique condition during data collection, necessitating careful consideration during data preprocessing. Other than those two, the other data appears to be generally consistent, with most variables showing moderate variability around their means. Proper handling of missing data and outliers is critical for the path to an accurate and robust PMV prediction model.

2.2. Exploratory data analysis

EDA is an initial, essential step in the data science process that leads to a profound understanding of structure, patterns, and possible anomalies residing in data. In this step, one tries to summarize the main features and results of the data using various statistical and visualization techniques, primarily graphical means [22]. EDA not only aids in the identification of trends and relationships between variables, but it is also critical for uncovering latent patterns, outliers, and potential biases that can jeopardize an analysis's conclusions. Through iterative exploration

Table 2
Preliminary analyses of the raw ASHRAE Database having 107,583 data entries for the cleaning process of the modified ASHRAE Database.

Variable	Data Number	Missing (%)	Mean	Std. Dev.	Min	25th Percentile	Median	75th Percentile	Max
PMV	66,734	38.0 %	0.14	0.94	−3.0	−0.4	0.1	0.66	3.0
T_a	99,911	7.1 %	24.50	3.78	0.6	22.3	23.9	26.4	63.2
RH	97,762	9.1 %	47.55	15.76	0.0	35.3	47.2	59.4	100.0
T_r	32,473	70.0 %	24.60	4.26	1.2	22.4	24.0	27.0	148.1
V_a	89,892	16.4 %	0.18	0.41	0.0	0.04	0.09	0.19	56.17
clo	99,663	7.4 %	0.67	0.29	0.0	0.49	0.63	0.78	2.89
M	90,419	16.0 %	1.21	0.25	0.65	1.06	1.20	1.20	6.83

with EDA, important insights will emerge, providing guidance on the selection of the most appropriate ML algorithms, thereby building predictive models in this data that are both accurate and robust.

2.2.1. Distribution analysis

Descriptive statistics includes distribution analysis, which aids in comprehending the distribution of each variable within the dataset. Each feature's distribution will allow us to derive or identify key measures of central tendency, spread, and shape of our data, important for selecting suitable statistical methods and ML models. It is often drawn-out histograms, along with box plots and density plots, to visualize the frequency and variability of data points. Analyzing the type of distribution reveals that skewness, kurtosis, or the presence of outliers significantly impacts the outcome of prediction models. Therefore, it provides a fundamental understanding that aids in future data preprocessing and modeling strategies.

Importance of thermal comfort elements such as age, individual heat sensitivity, and subjective variables are known in the literature. The ASHRAE Global Thermal Comfort Database includes these parameters. However, when the issue is about the derivation of PMV, the well-known formula by Fanger [6] does not include them. In addition, reduction of the six parameters to three, excluding occupant feedback loops (self-reported comfort levels) originally integrated into real-time predictive adjustments has been performed in the current work. Finally, it is proved that the use of three parameters was enough for the accuracy of ML models' predictions of PMV. These other personal or subjective traits above were left out of the model, as well as most of the occupant feedback loops, in other words, self-reported comfort levels. This is because the goal of this study is to make the complex inputs used by Fanger in the PMV formula [6] predictable using only a few basic variables, such as T_a, clo, and M. The integration of feedback loops into real-time predictive adjustments, or more specifically, subjective personal factors, is the primary cause of inconsistencies in the ASHRAE Database. This is due to the subjective nature of individual decisions, which are influenced by their character and daily mood. Although metabolic activity can give some information about a person, inclusion of things that aren't easily observable, like age or sensitivity, has not been taken in to consideration, because they would have made the developed current method too complicated and had to be left out of this study. Conversely, addressing user feedback mechanisms, such as self-reported comfort levels, will undoubtedly be a vital future step in the context of real-time adaptation and tailored comfort requirements in building automation. Incorporating user input to improve instantaneous PMV estimations might boost predictions and facilitate the development of tailored comfort solutions. Therefore, it was preferred to be better to leave out the subjective or demographic variables from the integration. This is because the goal of the study is to simplify Fanger's formula's [6] large number of input measurements and reduce the number of variables while keeping the

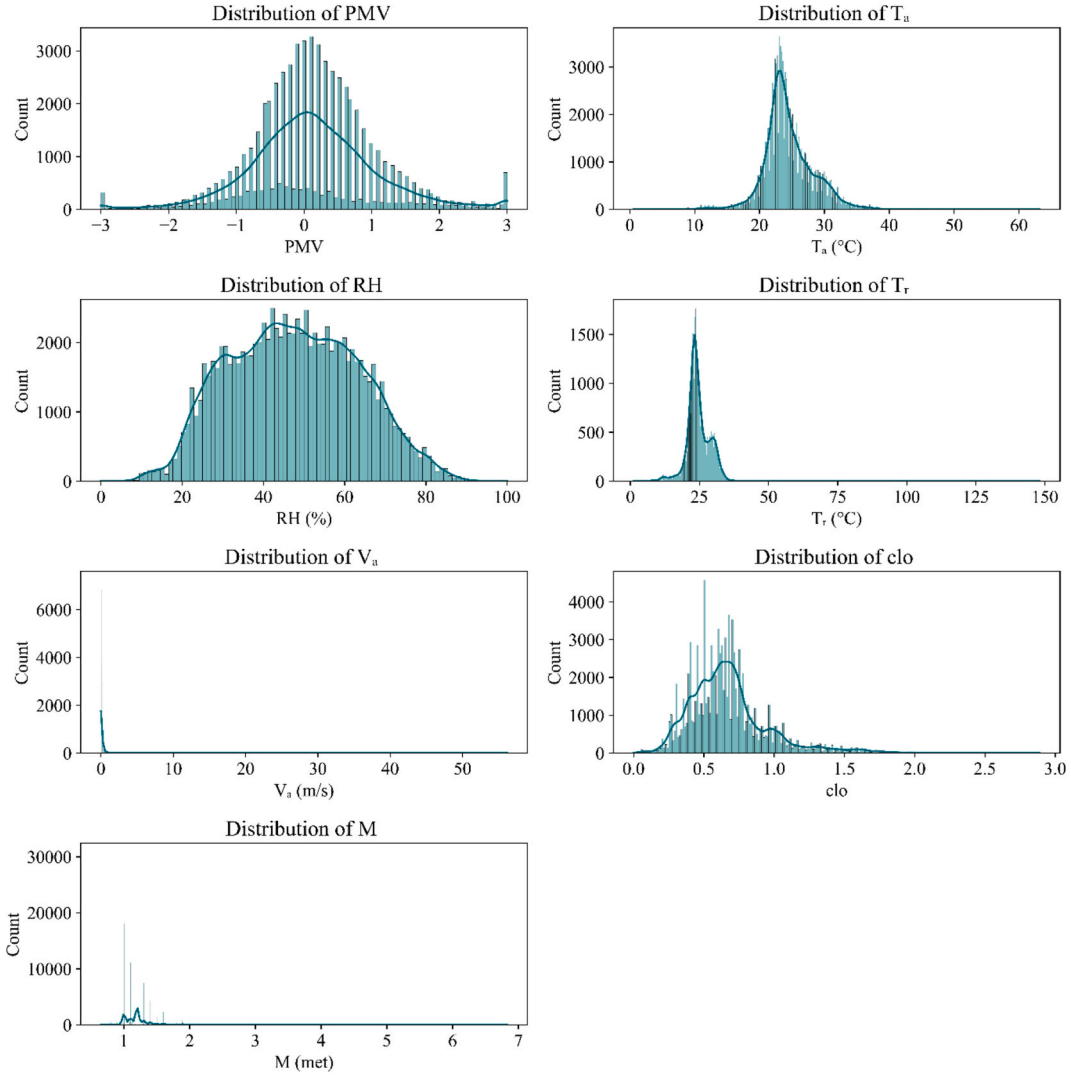


Fig. 1. Histograms of the raw ASHRAE Database regarding distributions for the key variables in thermal comfort and PMV values before data cleaning as a first step.

accuracy of the predictions. As a future work, evaluation of the distinct variations maybe subjected in other works.

Fig. 1 displays the distribution of the primary predictor variables (clo, M, T_a, T_r, RH, and V_a) in conjunction with the target variable PMV. The majority of variables exhibit distributions that are either normal or closely approximate normal, with the exception of V_a and T_r, which display lengthy tails and possible outliers. These observations indicate that it may be necessary to either transform these variables or deal with the outliers.

An analysis of key variable distributions reveals important characteristics for the modeling process: the target variable PMV displays an almost Gaussian distribution centered around zero. This is important since it enables better performance for most ML algorithms. On the other hand, the input variables present a varied distribution; T_a and T_r are pretty close to normal, maybe slightly skewed, with T_r showing a very marked peak at lower values. Also notable for skewness are the distributions for clo and RH, which are bimodal and probably indicative of lurking potential subgroups in the data. Distributions for M and V_a are also highly skewed, displaying the long tail characteristics that indicate outliers. These give a strong signal that some careful preprocessing, especially for dealing with skewness and outliers, is very appropriate to ensure robust model performance.

2.2.2. Correlation matrix

A correlation matrix is key to giving linear dependencies between pairs of variables in the data set for ML and statistical analysis. This, therefore, points to the quantities by which different variables correlate, guiding the process of feature selection and dimensionality reduction. It identifies features that might be redundant or highly predictive of one another. For example, a high correlation between two variables could indicate multicollinearity, which usually adversely affects the performance of some ML algorithms [23]. Such features and attributes were important for improving model interpretability and ensuring that input features all make independent contributions to the model's prediction task.

The correlation matrix displayed in Fig. 2 provides insights into the relation between the predictors and the target variable PMV. The heatmap demonstrates a substantial correlation between T_a and T_r with PMV, indicating that they have a major effect on thermal comfort. In contrast, factors such as clo and V_a exhibit weaker associations, suggesting that their influence may be more nuanced or contingent upon specific circumstances.

A close examination of the correlation matrix in Fig. 2 reveals several intriguing relationships between the variables. The most obvious one is a strong positive correlation between T_a and T_r, which equals 0.92—this relationship may point to the possibility of multicollinearity. Moreover, PMV has quite a high correlation with both T_a (0.78) and T_r (0.76), indicating that thermal comfort largely depends on the temperature-related variables. On the other hand, clo is negatively correlated with both T_a (−0.47) and T_r (−0.50), which represents an adaptive likely behavior in which clo decreases as temperature increases. Other variables, such as RH and V_a, show weaker correlations, signifying lesser direct interaction. To sum up, Fig. 2's correlation matrix shows some critical linear relationships between variables that are related to temperature. These relationships may affect how well a model works when there is multicollinearity.

2.3. Data cleaning and preprocessing

Data cleaning and preprocessing are essential phases in every data-driven projects since they establish the basis for precise and significant analysis. It involves a methodical examination of missing values, anomalies, and discrepancies in data. If these processes are not unaddressed, skewed outcomes and diminished model efficacy can be observed. Furthermore, preparation procedures like normalization, scaling, and feature transformation are essential to convert the data into forms that satisfy the specific requirements of different algorithms, hence facilitating the effective processing of inputs by the generated models. These steps will enhance the quality, dependability, and interpretability of the data collection, yielding more robust and reliable results. Consequently, data cleaning and preparation are imperative for any endeavor aimed at yielding successful outcomes in ML or statistical

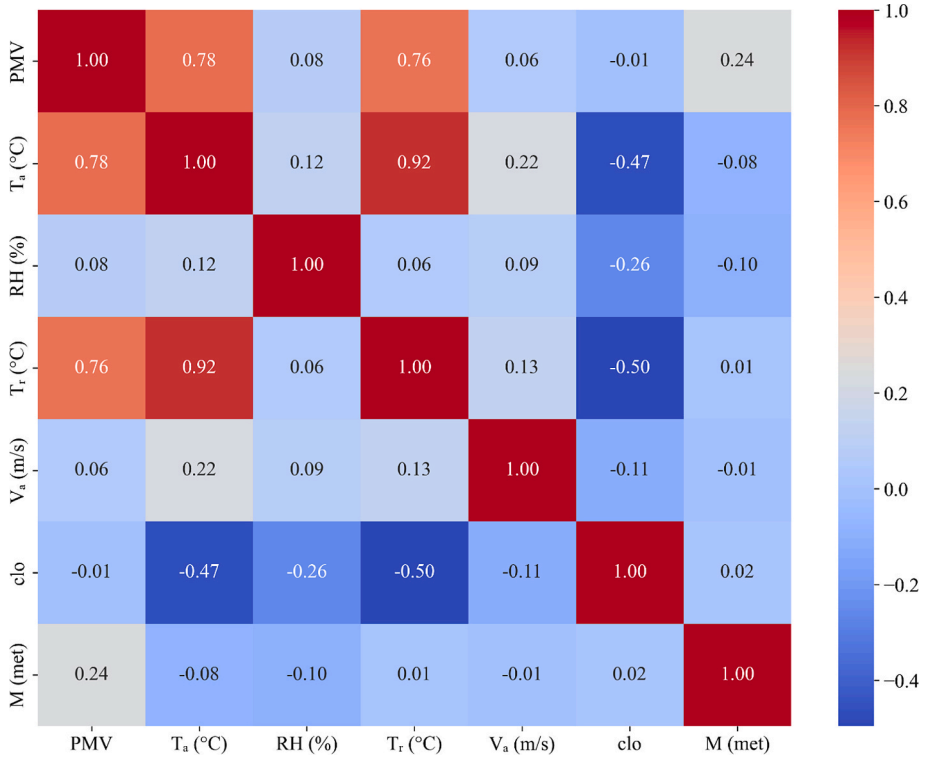


Fig. 2. Correlation matrix Heatmap in python for ASHRAE Database.

analysis.

2.3.1. Handling missing data

In this section, the issue of missing data in the T_r and PMV columns, which contained a considerable number, is tackled. The existence of those missing values poses a risk to the analysis's accuracy and reliability. Therefore, it was crucial to handle these missing values carefully in areas such as thermal conditions, which form the foundation for understanding and predicting comfort, to maintain the integrity of the information within the dataset. Accordingly, the systematic treatment of missing values resulted in making the dataset robust, hence forming a solid basis for accurate and meaningful subsequent analysis.

Missing T_r values were calculated using Equation (1) using the relationship between globe temperature (T_g), T_a and V_a given in the dataset [1,4]. Note that Lamberti et al. [19] also used the same method to calculate T_r. The formula used is [24]:

$$T_r = \left(T_g^4 + 2.5 \times 10^8 \times V_a^{0.6} \times (T_g - T_a)\right)^{0.25} \quad (1)$$

where T_g and T_a represent the globe and air temperatures in Kelvin, respectively, and V_a denotes the air velocity in meters per second. This approach guarantees that the T_r values are deduced in a manner that is scientifically significant, provided that the relevant variables are accessible.

The recalibration of the missing values in the PMV field is done using the available data. The variables important for the recalibration are T_a, T_r, V_a, RH, M, and clo, leaning on basic physical relations governing thermal comfort. For the validation of recalculated PMV, the thermal comfort Python library was applied with the support of recognized standards for the values of thermal comfort. The library was designed to calculate PMV quite accurately through given PMV equations and guidelines. This approach served two purposes in such integration: filling the gaps in the data set and, at the same time, ensuring the integrity and reliability of the PMV data to match with the accepted requirements for thermal comfort analysis.

After cleaning and preprocessing with the imputation of missing values and treatment of outliers, the total records in the final dataset were 55443. The original data set initially had 107,583 records; however, after filtering criteria were put into place and missing values dealt with, along with the recalibration of key variables such as T_r and PMV, the final cleaned dataset consisted of this amount. This is a well-prepared dataset, forming a very solid base for ML models to improve the accuracy and reliability of predictions.

2.3.2. Identification and management of outliers

Outliers in data are points that greatly depart from the dataset's overall distribution. Measurement flaws, data entry problems, or actual variability in the underlying events can all lead to these very extreme numbers. Although occasionally they offer insightful analysis of rare or unusual events, outliers sometimes skew statistical analyses and may compromise the efficacy of ML models. Data preparation thus depends critically on the identification and suitable management of outliers. This work aimed to identify dataset outliers using the Interquartile Range (IQR) method. Calculating the range between the first quartile (Q1) and the third quartile (Q3) and then considering any data point that lies notably beyond this range makes the IQR technique a strong statistical tool that detects outliers. Data points below the value of $Q1 - 1.5 \times IQR$ or above the value of $Q3 + 1.5 \times IQR$ are particularly noted as possible outliers.

Once the outliers have been found and cleaned up, Fig. 3a and 3b show the before and after distributions of outliers in the T_r and V_a variables. Outliers for the T_r variable were identified in Fig. 3a, and the data set exhibited anomalous values such as 150 °C. Similarly, in Fig. 3b, outliers such as 50 m/s for the V_a variable were identified; these values were completely removed from the dataset after the cleaning process.

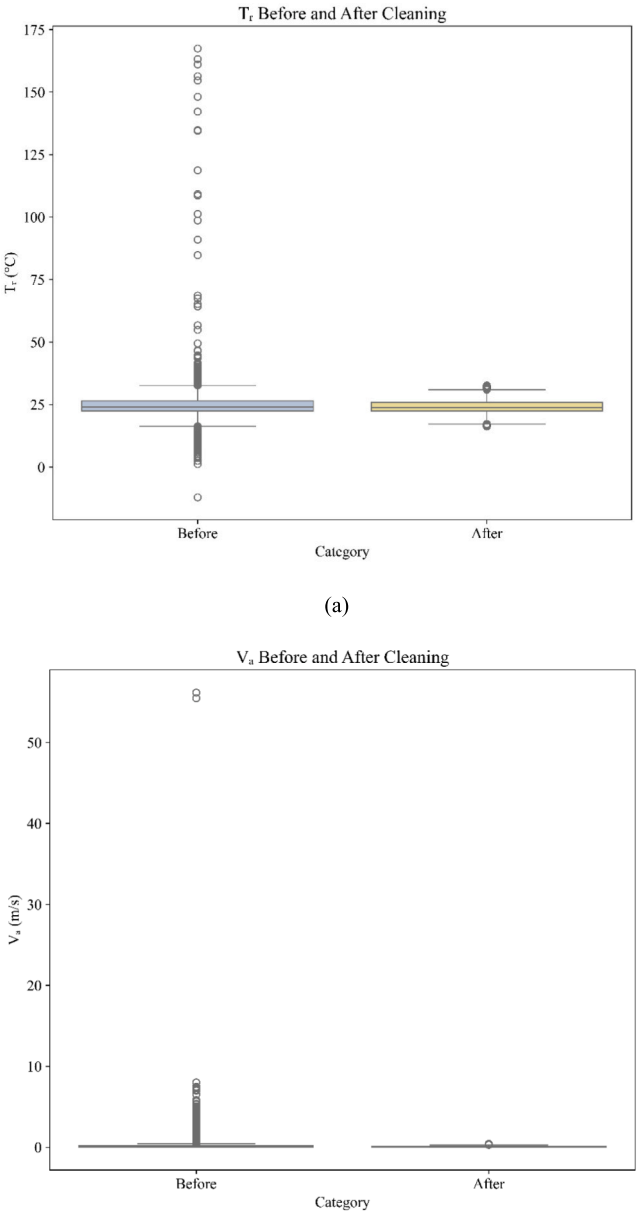


Fig. 3. Outlier detection of the raw ASHRAE Database regarding wrongly stated entries for T_r (a), and V_a (b).

These abnormalities are presumed to be primarily attributable to data entry errors. Eliminating outliers while preserving the statistical integrity of the data set enhances the reliability of the analytical outcomes.

2.3.3. Data limitations

There are many things that can go wrong with data collection, including errors in the data itself or measurements that aren't accurate enough, which can lead to sensing errors. Sometimes, the thing being studied is naturally variable. However, such restrictions can have a significant impact on the outcomes and hence lead to flawed or biased predictions. Because of this, it is important to carefully look over and understand these limitations to make sure that the study can be trusted and that it is appropriate to draw conclusions from it.

This study specifically establishes specific data filtering criteria to focus on standard indoor environmental conditions, thereby ensuring the trustworthiness and usefulness of the analysis. The chosen intervals

for important variables are in accordance with recognized thermal comfort standards, eliminating any outliers or extreme values that may skew the findings.

Table 3 expands extensively on the comprehensive filtering strategy utilized in this study, which is based on clearly defined criteria. In accordance with established norms and guidelines in thermal comfort, each component is confined to certain limits. To evaluate various climatic conditions, such as tropical zones, T_a and T_r are confined to the range of 15 °C to 35 °C. To maintain consistency in data density, relative humidity is constrained between 10 % and 89 % to encompass all climates represented in the ASHRAE Dataset. clo and M are restricted to conventional indoor activities and attire selections. V_a is sustained between 0 m/s and 0.45 m/s to ensure accurate depiction of standard internal air circulation in the data. In addition, PMV is constrained between −3 and 3 on the PMV scale to ensure that the dataset fairly covers the full spectrum of thermal comfort ratings. Finally, thermal comfort variables are shown in Table 3 with their ranges, it should be noted that the values out of the given ranges in the ASHRAE Dataset are omitted as a result of final cleaning process. In order to visually assess the impact of these constraints on the dataset, a histogram showing the distribution of the selected variable ranges is generated and presented in Fig. 4.

According to Fig. 4, for every variable, numbers in each interval represent the concentration of data points, clearly showing how the filtering method affects the dataset. By concentrating on these regulated ranges, the histogram emphasizes the improved quality of the information and guarantees that the following analysis is based on data that fairly reflects the conditions pertinent to thermal comfort. Moreover, the distributions illustrated in Fig. 4 demonstrate the process of data "cleaning" and highlight the focal region inside the filtered data set about the selected intervals. This filtering procedure, however unusual, entails the removal of extreme values from the data set or the exclusion of observations beyond the average range, thereby failing to represent the actual conditions observed in reality. Excluding extremely high M's (e.g., regions with vigorous physical activity) or excessively high/humid conditions (e.g., naturally ventilated structures in tropical areas) from the dataset may diminish the model's efficacy in buildings characterized by significant geographical or functional diversity. Visual data provides a strong basis for thermal comfort prediction analysis and supports the robustness of the filtering criteria, thereby enabling evaluation of the ML models to be built in the upcoming works of the authors of the present investigation. This filtering strategy enables the model to produce more precise predictions in indoor environments classified as "standard" due to data homogeneity and consistency, hence diminishing uncertainty in ML applications. Ultimately, given the data is derived from literature, discrepancies in the calibration of measuring instruments or variations in data-collecting techniques between various research may not be entirely resolved. This may indirectly influence the model's confidence range, particularly for sensitive measures like V_a or T_r. Consequently, it is important to acknowledge that the filtering and removal techniques outlined in the Data Limitations section are essential for consistently analyzing certain situations and for successfully training the foundational ML models; yet, there may be instances that remain unfiltered. Nonetheless, the results of the created model must be meticulously scrutinized for user attributes outside the filtered data or environmental influences. This strategy may be beneficial in energy-efficient workplaces or residential buildings.

Table 3
Elimination of the raw ASHRAE Database outliers to obtain the modified ASHRAE Database considering specific ranges of each variable for filtering the data process.

Variable	Range	Average value	Justifications
PMV	−3 − +3	0.06	According to ASHRAE 55 and ISO 7730, it fluctuates between −3 (extreme cold) and + 3 (extreme hot).
T_a	15–35 °C	24.19	This range was chosen considering the different climatic situations counting tropical climates
RH	10–89 %	46,73	This range was chosen considering the different climatic situations counting tropical climates
T_r	15–35 °C	24.31	This range was chosen considering the different climatic situations counting tropical climates
V_a	0–0.45 m/s	0.09	The interval where the data set is dense was preferred.
clo	0.08–1.2	0.65	A clo value of 0.5 is often assigned to lightweight apparel, while a value of about 1.0 is designated for winter garments.
M	0.8–3	1.21	According to ASHRAE 55, M varies depending on physical activity. It can be 0.8 M for sitting and 3.0 M for moderate activity.

These carefully selected filtering parameters help to concentrate the research on data points reflecting appropriate and regulated indoor environmental conditions. This work proposes to augment the reliability and applicability of the subsequently developed ML models by imposing these limitations. The subsequent sections will rigorously assess the impact of these limitations on the model's efficacy and generalizability, elucidating how the filtered dataset influences the outcomes and application of the ML approach in predicting thermal comfort.

2.3.4. An analysis of data filtering and its impact on the distribution of variables

This part involves a thorough analysis of the data filtering process and its subsequent impact on the distribution of essential variables in the dataset. The aforementioned filtering procedures were applied to refine the dataset by confining the variables to ranges characteristic of standard indoor ambient conditions. The methodology is crucial for focusing the analysis on the most pertinent contexts for assessing thermal comfort. It eliminates outliers and extreme conditions that could distort the results or create bias. Comparing the dataset before and after these filtering criteria will help us assess variable distribution changes.

2.3.4.1. Visualizing relationships between variables: Scatter plot matrix of the original dataset. Scatter plots provide the basic tool for graphically examining pairs of variables' associations in EDA. Based on the values of two variables, a scatter plot lines its points, each an observation from the dataset, on the pertinent axes. Scatter plots, which use a comparison between one variable and another, allow researchers to identify possible patterns, correlations, or anomalies in the data [25]. To completely grasp the links among the variables in the first dataset, a scatter plot matrix was developed. The scatter plots in this matrix give a graphic picture of the interactions among the variables, showing every conceivable pair of variables. The scatter plot matrix allows us to identify potential data outliers, groupings, and linear or non-linear associations. Such visual analysis allows one to identify links and relationships that may later influence the modeling process. The scatter plot matrix, which guides our research, reveals information on the original dataset's distribution and structure, as illustrated in Fig. 5. This visualization assists the creation of predictive models and helps to find important links and patterns, guiding the use of data filtering criteria.

The scatter plot matrix reveals various problems in the original data that might influence the analysis. First of all, some variables like "V_a" and "M" exhibit highly skewed distributions, implying that they are not well represented on their entire range. This lack of representation may limit the ability to extrapolate models developed from this database. Moreover, there are evident prospects for outliers, particularly when it comes to relationships between "T_a" and "T_r." Proper management of these outliers can distort the study and reduce the accuracy of prediction models. Other variable pairs also show weak or no linear correlation,

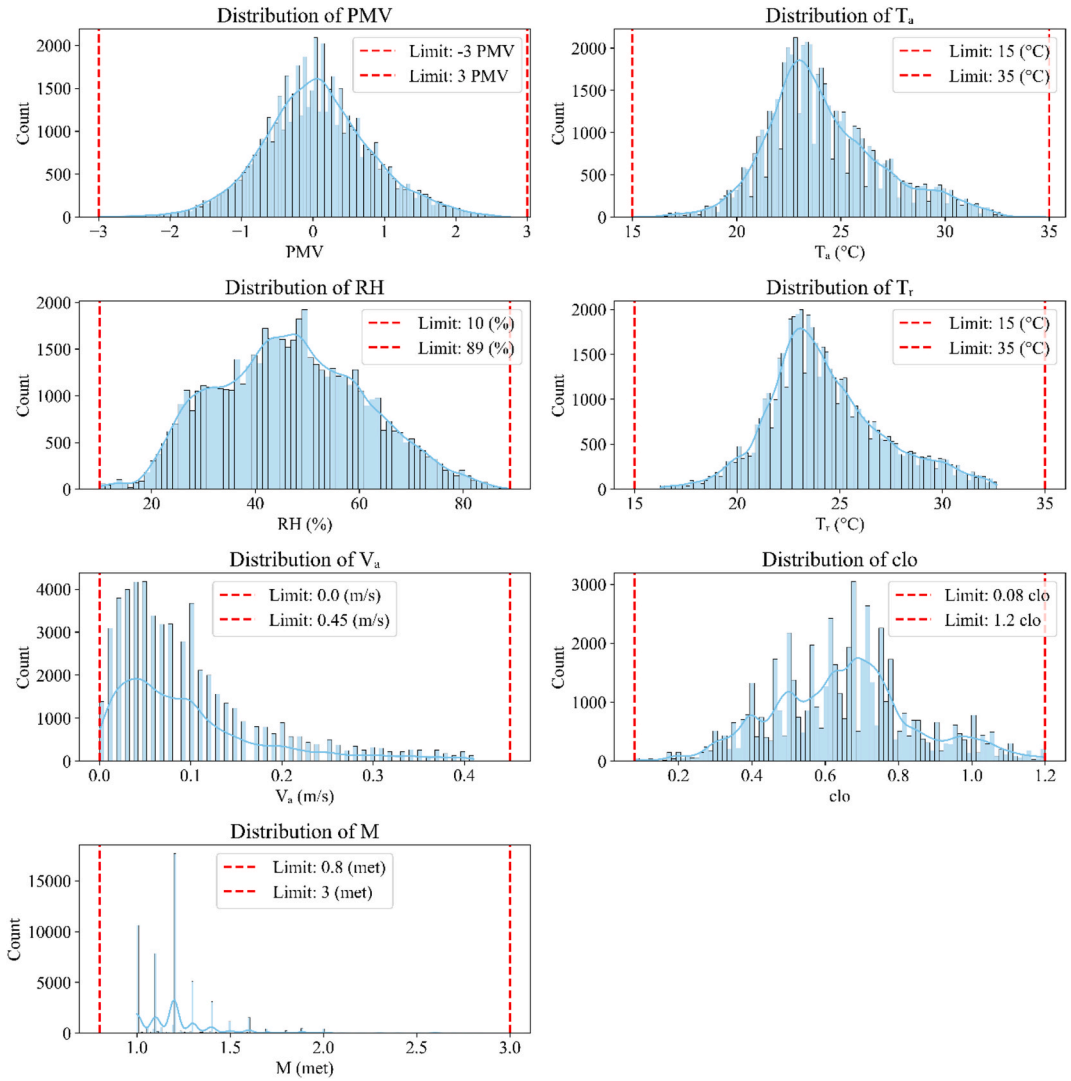


Fig. 4. Histograms of the raw ASHRAE Database regarding eliminations for the key variables in thermal comfort and PMV values after data cleaning considering specified limits.

indicating that linear modeling techniques may have difficulty accurately representing any underlying patterns pertaining to the data involved. Furthermore, the strong positive relationship between "T_a" and "T_r" causes concerns over multicollinearity, which can make it harder to understand regression models by increasing variability and hiding which variables are more important. As a result, comprehensive data preprocessing is required to address such issues and improve the reliability of subsequent analysis.

2.3.4.2. Scatter plot matrix of the filtered data. After applying filter criteria, Fig. 6 displays the pairplot. The purpose of this criteria was to eliminate data points that fall outside the normal ranges for indoor environmental conditions, as defined by standards such as ASHRAE and ISO. The resulting distributions demonstrate greater limitations because variables such as T_a, RH, T_r, V_a, clo, and M have narrower ranges. The consequent improved dataset more accurately represents common interior situations, minimizing the impact of unusual and extreme values.

According to Fig. 6, the scatter plot matrix of the filtered dataset shows some really significant changes over the original one. First, the elimination of outliers and extreme values has reflected the concentration of variables such as "T_a" and "T_r." This concentration indicates that the data now more accurately matches typical indoor conditions. There are also clear links between variables, like the linear correlation between "T_a" and "T_r," which shows that the filtering has reduced unnecessary variation while keeping important correlations.

The filtering process has successfully narrowed the dataset to the cases most pertinent for evaluating thermal comfort. Eliminating data points that are outside the standard comfort range enhances the analysis's rigor, guaranteeing that the derived insights are relevant to actual interior circumstances. The filtered data exhibits more uniform distributions, hence diminishing the probability of distorted conclusions resulting from extreme outliers. Implementing this step is essential for improving the precision and dependability of any future modeling or statistical analysis conducted on the data.

The visual comparison between Figs. 5 and 6 highlights the importance of data filters in environmental studies. The pairplots demonstrate how filtering can effectively isolate the underlying relationships within the dataset, thereby improving the representativeness of the data for typical indoor situations and improving the overall quality of the analysis.

Fig. 7 illustrates the preprocessing workflow of the data analysis employed in this study. The process begins with the loading of the dataset as shown in Fig. 1, followed by EDA to identify key trends and potential issues, such as missing data as shown in Table 2 and outliers as shown in Fig. 3. The next step is to handle missing data, ensuring no gaps

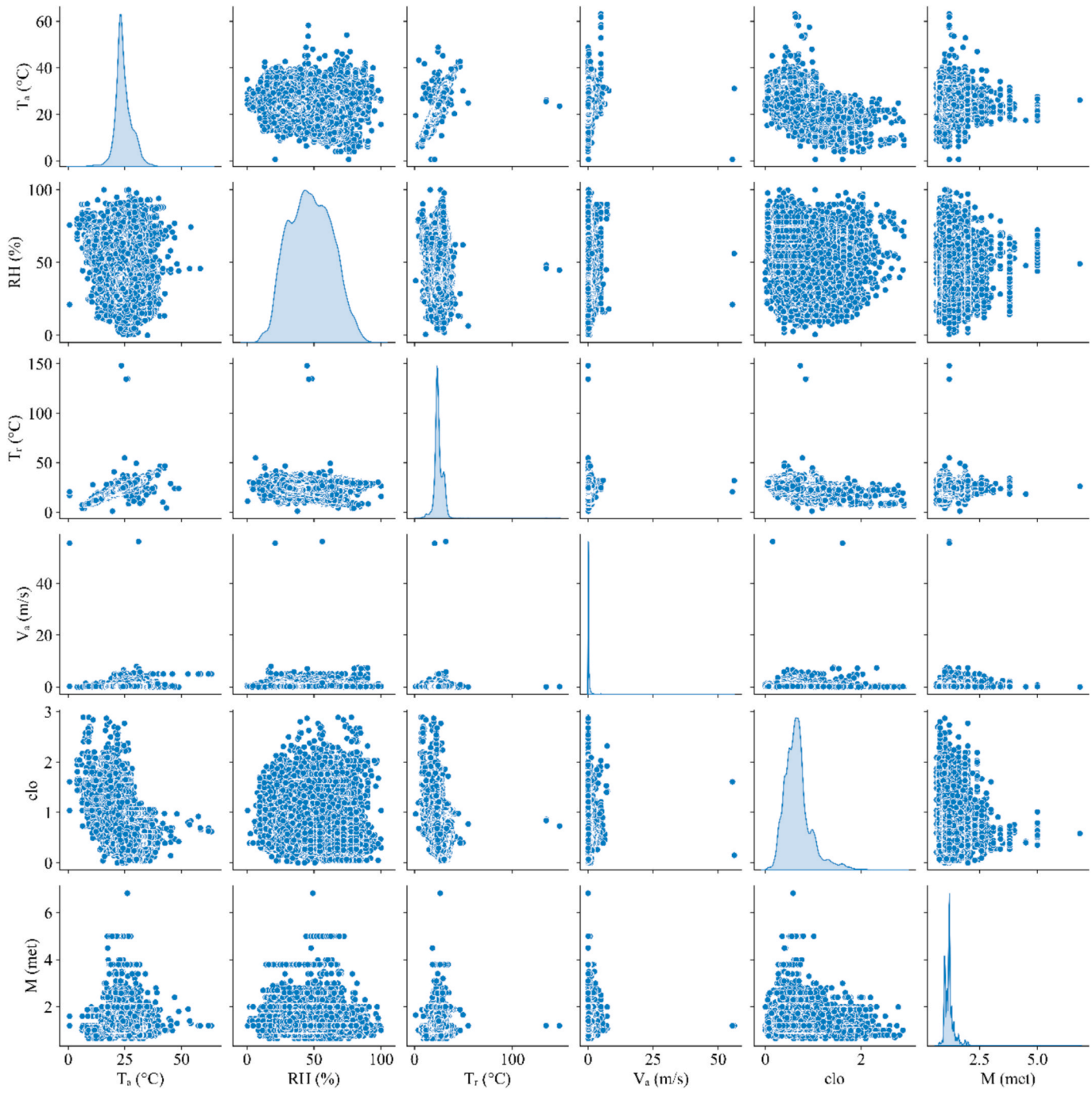


Fig. 5. Pairplots using the ASHRAE Database for the relation of key variables with each other before applying filtering criteria.

remain that could compromise the analysis. Subsequently, the outlier identification phase is executed to identify and address any extreme numbers that may distort the results. The data undergoes transformation processes, including normalization or encoding, to prepare it for analysis as shown in Fig. 4. Finally, filters are applied to narrow down the dataset to relevant scenarios, leading to a dataset that is ready for robust modeling as shown in Fig. 6. This structured approach, as depicted in the flowchart Fig. 7, ensures that the data is meticulously prepared, reducing the risk of inaccuracies in the subsequent model-building phase. SHAP method was applied to boost the explainability of the model. The RF model was trained first; subsequently, the SHAP values were computed, and an individual analysis of each input's effect on the PMV estimate was examined. This displayed more clearly the interactions inside the model and the relative significance of the variables.

3. Results and discussion

It is practical that the low-parameter PMV approach can be used in crowded and large indoor spaces, like conference centers, multipurpose halls, and other similarly large venues, even though it has some limitations due to limited data. Individuals congregating in such settings frequently exhibit differing metabolic rates and varying degrees of clothing insulation efficacy. Nonetheless, significant data may be obtained without a comprehensive measurement infrastructure, since the proposed model can forecast PMV using only a limited number of essential parameters (Ta, clo, or M). Rapid and rational comfort evaluations are therefore attainable by collecting data on the ambient temperature of the space and the general level of attire in the audience. Subsequent studies may seek to improve the model by collecting more

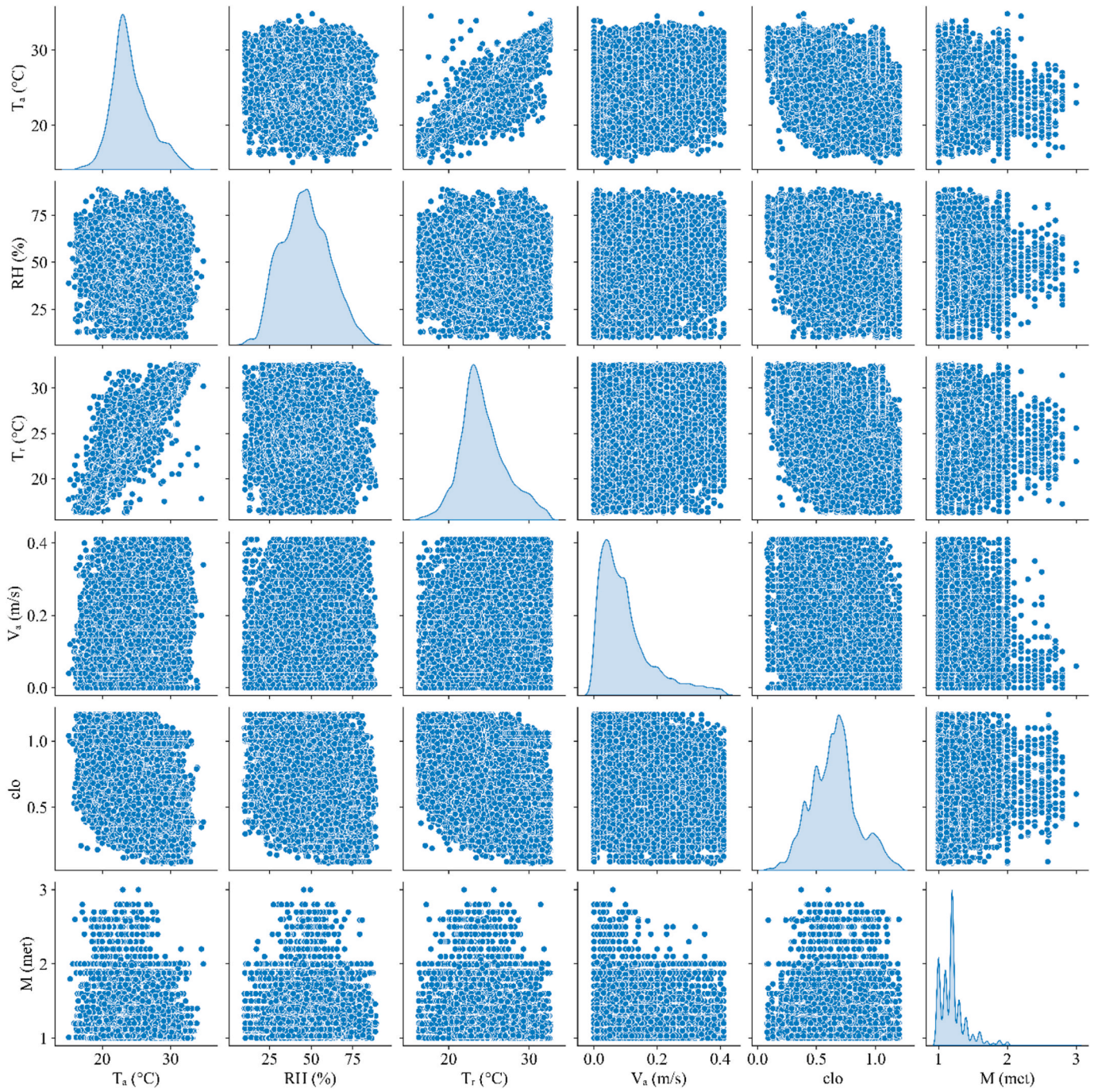


Fig. 6. Pairplots using the ASHRAE Database for the relation of key variables with each other after applying filtering criteria.

comprehensive data, such as additional variables like activity type, sitting time, and demographic age profiles, and tailoring it to broader contexts. The main aim of the current work was to provide details on database evaluation and its consistency through preprocessing stages, marking it as a pioneering study in the literature. The user can apply the proposed methodology in this work to a specific case if they write codes in Python.

3.1. Feature selection

Feature Selection is essential in data preparation, as it seeks to identify the most pertinent traits that significantly aid in predicting the target variable. Reducing the number of features will help us to simplify the model, improve interpretability, lower overfitting, and lower computational cost. This work combines Feature importance, SelectKBest, SHAP, PDP, and P-box analyses to ensure a robust and comprehensive feature selection procedure. Fig. 8 represents the feature selection process that follows data preprocessing. Each methodology successively contributes to finding the most important variables for the model, as given by the individual pathways leading to the final feature set. Together these techniques give assurance that the selected features are both statistically robust and interpretable, and they become the basis on which the final model is built.

Feature importance is an efficient method for figuring the significance of a feature for assessing the significance of a feature by examining its ability to reduce node impurity across all trees in the forest. SelectKBest analysis provides a straightforward method for identifying the best predictive features by ranking attributes in relation to the target

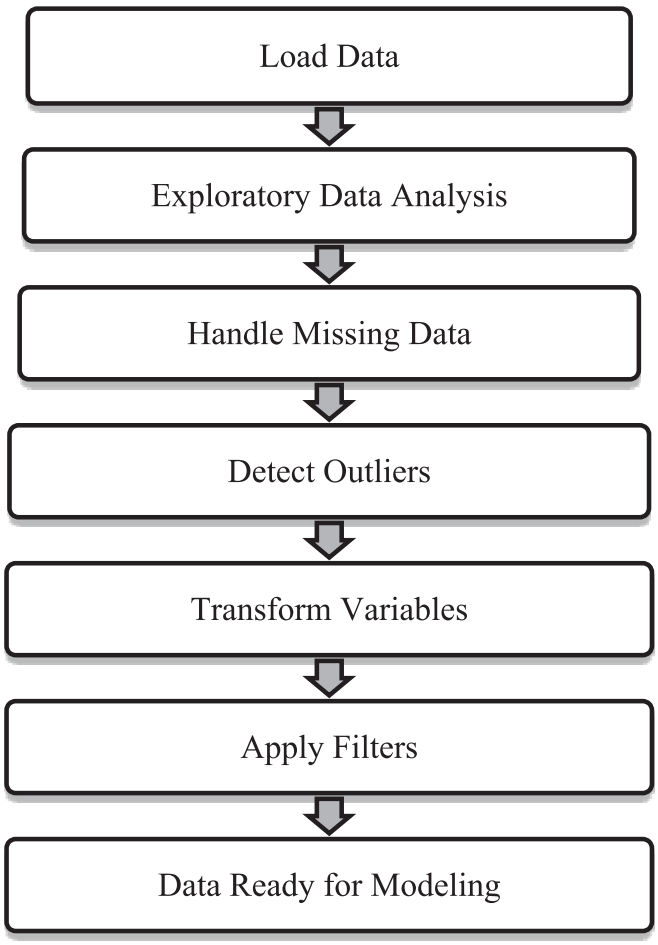


Fig. 7. Preprocessing workflow of the data analysis.

variable according to statistical significance. Originating from cooperative game theory, SHAP values offer a comprehensive, instance-specific analysis of the influence of each feature on model predictions, guaranteeing that our selection encompasses both general and nuanced feature significance. The P-box analysis shows how the relative value of features changes, which is not always evident. This would facilitate the process of determining which traits contribute consistently across several different data sets.

The following techniques have been selected for their individual advantages: SelectKBest, for its statistical soundness; RF, for its resistance and group evaluation possibilities; SHAP, for its good understandability; and P-box Plot, to investigate reliability and variability. All of these techniques combined provide a thorough way of feature selection and ways to improve generality, understandability, and model performance. The number of features in our ML study had to be reduced to boost the model's efficacy, prevent overfitting, and improve interpretability. Applying different techniques for feature selection can achieve this goal. The comparison and analysis of various approaches revealed the most important independent variables for determining the PMV in indoor situations. For this purpose, Table 4 presents a summary of feature selection processes based on the feature selection methods and thermal comfort parameters of the PMV formula by Fanger [6]. The asterisks in this table indicate the most effective features, which are derived outputs of the analyses. The following subsections below include the detailed explanations on all outputs in Table 4.

In this study, RF and SVM models were preferred for PMV estimation. PMV is a variable that is affected by the combination of environmental and individual factors and has complex and nonlinear relationships. Correlation analysis reveals that T_r, T_a, clo, M, RH and V_a variables show

Table 4
Summary of feature selection processes.

RF Model						
Feature Selection Methods / Thermal Comfort Parameters	T_a (°C)	RH (%)	T_r (°C)	V_a (m/s)	clo	M (met)
Feature Importance (Fig. 9a)	x		x*		x	x
SelectKBest (Fig. 10)	x*		x*		x	x
SHAP (Fig. 11a)	x*		x		x	x
P-box (Fig. 12)	x*		x*			
PDP (Fig. 14a)	x		x			
SVM Model						
Feature Selection Methods / Thermal Comfort Parameters	T_a (°C)	RH (%)	T_r (°C)	V_a (m/s)	clo	M (met)
Feature Importance (Fig. 9b)	x*		x		x	x
SelectKBest	–	–	–	–	–	–
SHAP (Fig. 11b)	x		x*		x	x
P-box (Fig. 13)	x*		x*			
PDP (Fig. 14b)	x		x			

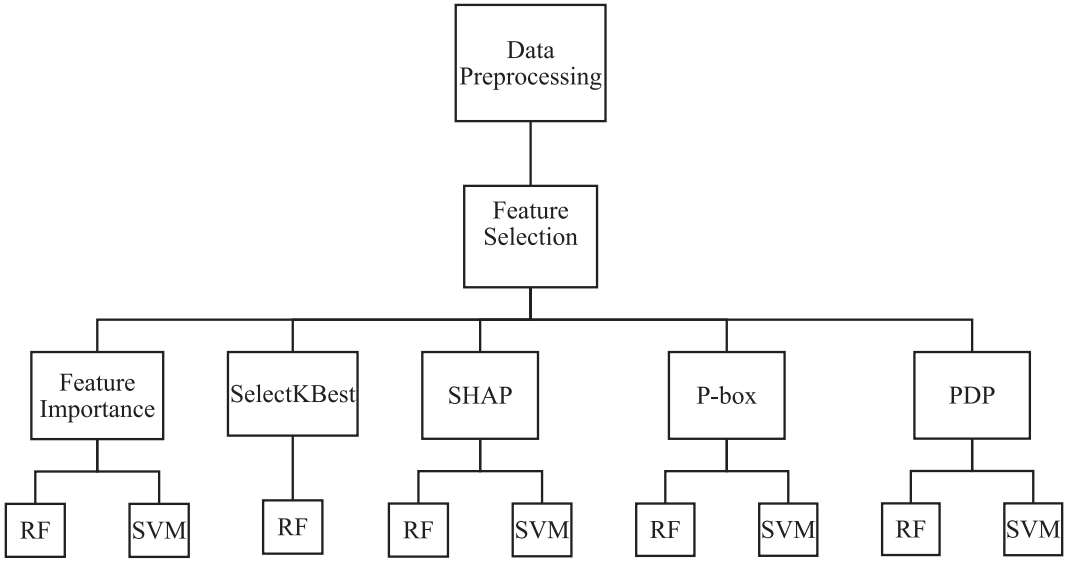


Fig. 8. The feature selection processes that follow data preprocessing.

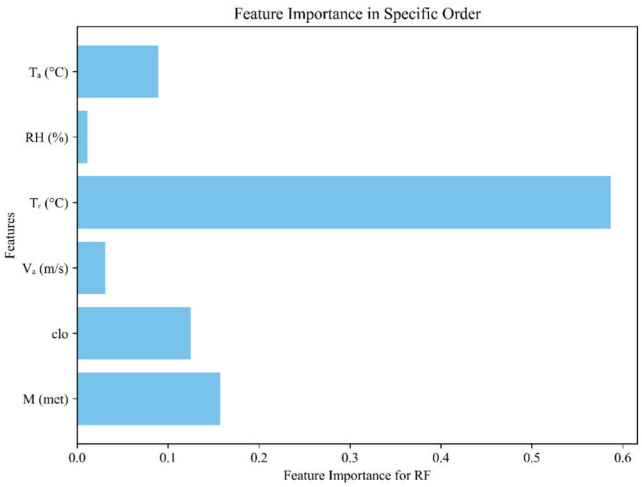
strong relationships with PMV. Therefore, models with more flexible and powerful learning capacity should be used. RF is an ensemble learning algorithm composing of a high number of decision trees and is quite successful in modeling complex, nonlinear data relationships [26]. One of the most important advantages of this model is that it increases model explainability by determining variable importance ranking [27]. RF is resistant to overfitting, provides flexibility in managing missing data, and generates more robust predictions in the presence of noisy input. Conversely, the SVM model demonstrates enhanced efficacy, particularly in high-dimensional datasets and scenarios including nonlinear interactions [28]. SVM can flexibly model nonlinear decision boundaries through kernel methods and thus offers an effective alternative in complex estimation problems such as PMV. The high generalization performance, especially in small and medium-sized data sets, has led to the use of SVM in this study.

3.1.1. Feature importance

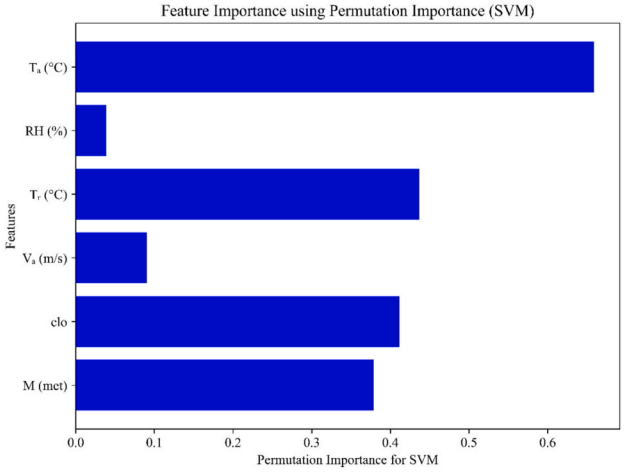
RF is a way to learn as a group that makes a forest of trees by figuring out how important each feature is while modeling a target variable [29]. During training, several trees are constructed, and one feature is utilized to split at a node while considering its contribution to reducing node impurity. To determine an overall rating, the significance of all attributes throughout the trees is added. Fig. 9a presents the feature importance scores derived from the RF method. As shown by Fig. 9a, the most important variable was T_r, with a feature importance of 0.6. The result is the strongest evidence for T_r's important impact on PMV prediction. The importance of T_a, clo and M in this model is further supported. Nevertheless, the effect of factors such as RH and V_a is minimal, suggesting that they have a limited influence on the model's predictive accuracy. Fig. 9b illustrates the thermal comfort parameter significance ratings obtained from the SVM methodology. Fig. 9b shows that the most important variable is T_a with a feature importance of 0.6. The result is the most convincing evidence of the significant effect of T_a on PMV prediction. By this analysis, the importance of T_r, clo and M is further supported with a score of 0.4. However, the effect of features such as RH and V_a is negligible, indicating their limited impact on the model's prediction accuracy. The results indicate that the feature significance derived by SVM and RF are almost the same and generally confirm one another.

3.1.2. SelectKBest analysis

To further validate and cross-check the results obtained from the Feature importance, the SelectKBest approach [30] was also applied. This assesses the statistical significance of the represented features in the univariate testing of the target variable. The scores of the features in the SelectKBest approach are shown in Fig. 10. The bar chart displays feature importance scores derived from the SelectKBest analysis. These scores depict the extent to which each feature contributes to predicting the target variable. The graph indicates that T_a and T_r have more influence on the predictions than the other variables with a score of about 70000. This indicates that T_a and T_r are significant enough to enable accurate individual predictions of the target variable. M and clo also show average relevance, staying within the class boundaries and following T_a and T_r. Conversely, RH and V_a had very low scores. This indicates that they do not have a strong one-to-one relationship with the dependent variable under study. These findings suggest that temperature-associated factors (T_a and T_r) are crucial predictors of the response variable due to their direct impact on individual comfort levels in temperature situations. Besides satisfactorily trimming away unnecessary features, SelectKBest does well to prioritize to those features that have large effects alone. This information provides us with practical suggestions for future improvements to our model. Results from SelectKBest are in concordance with the previous RF findings: T_r and clo appear to be the first two ranked features. Most interestingly, after T_r, T_a shows a high score, thus proving the relationship between T_a and PMV. M's relative importance is only slightly less than that of T_r and clo, but it



(a)



(b)

Fig. 9. Feature importance analyses as one of the investigated feature selection processes for RF (a), and SVM (b) ML models implementing the modified ASHRAE Database.

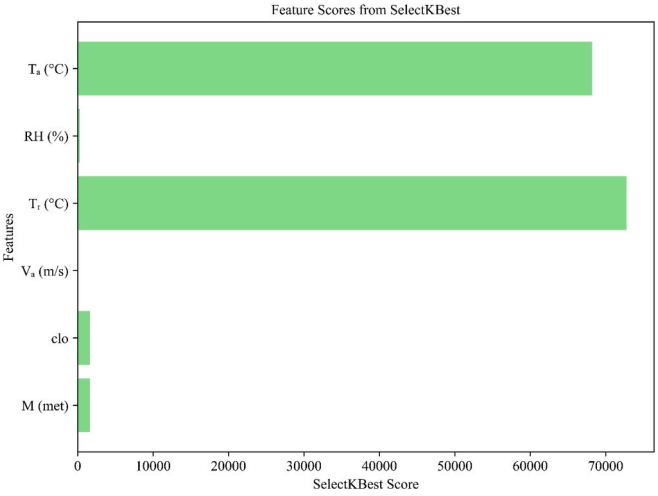


Fig. 10. SelectKBest scores for RF implementing the modified ASHRAE Database as one of the investigated feature selection processes.

still stands as a top predictor for this method.

3.1.3. SHAP analysis

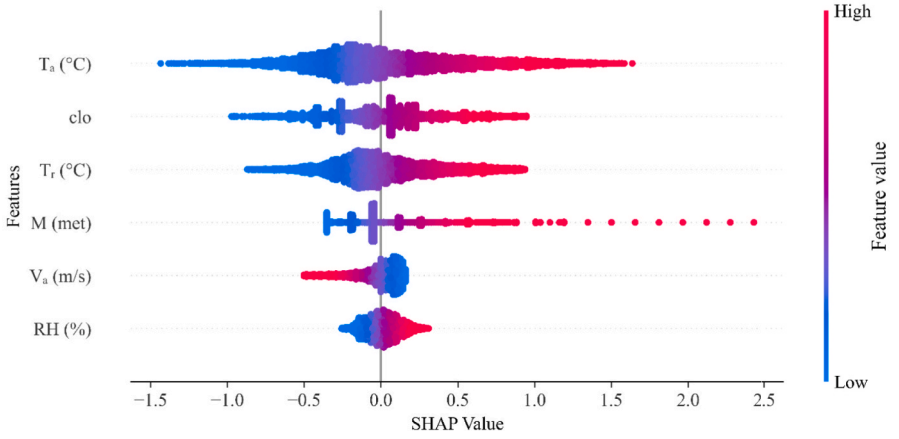
After the preliminary selection of features using Feature importance and validation by SelectKBest, the influence of all selected features was further proved by the SHapley Additive exPlanations method. SHAP is a powerful tool that provides detailed insights about how each feature affects to the final model estimation. The model achieves its objective by allocating each feature's contribution to each prediction, making it one of the most interpretable methods for determining feature importance [31].

The SHAP summary plot for RF in Fig. 11a shows the effects of different features on the PMV predictions with a graphic design. The SHAP value, which describes the contribution of any given feature on the magnitude of the model's output, is displayed on the horizontal axis. Positive SHAP values have a positive effect on PMV, as well as vice versa. Fig. 11a provides the following key observations: T_a is found to be the most important factor, and its SHAP values range widely from positive to negative. This figure shows how strongly T_a depends on and changes based on PMV predictions. Significant effects are also observed in the clo, T_r, and M. clo has a predominantly positive effect on PMV. The result reflects the fact that the higher the level of clo, the higher the PMV value, making the perceived environment warmer. The correlation between T_a and T_r, together with T_a's superior practicality and ease of measurement compared to T_r, aligns with previous results from both Feature importance and SelectKBest analyses, hence endorsing the prioritization of T_a in the final feature set. Furthermore, the variables V_a and RH appear to exert minimal influence on PMV. Fig. 11b presents the SHAP summary plot for SVM and highlights the following critical points: Similar to the RF model, T_r was found to be the most influential factor, with SHAP values ranging from positive to negative over a wide range. This result confirmed the strong influence of T_r on PMV estimates. Moreover, the analysis revealed significant effects on clo, placing M third and Ta fourth after clo. However, V_a and RH were found to be the least influential variables, supporting the previous analysis results. The results confirmed the previous analyses.

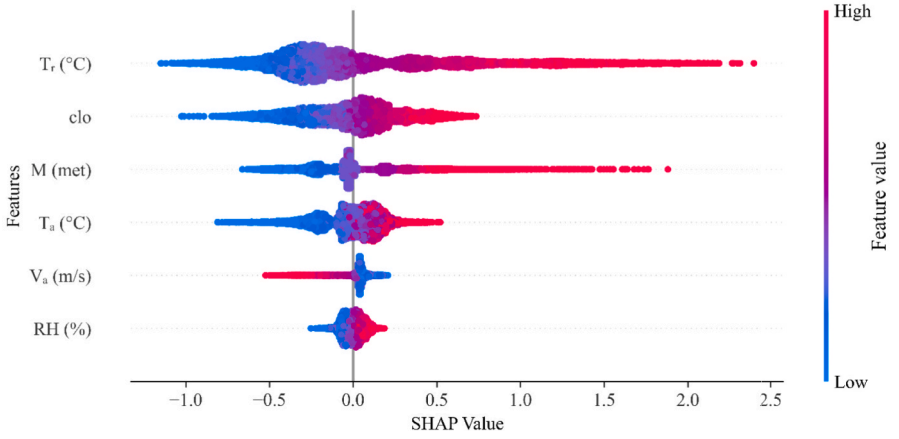
3.1.4. P-box analysis

P-boxes help show how uncertain and varied a piece of information is with the aid of a graph, particularly for showing a variable's range and distribution in different contexts. When it comes to assessing the impact of certain attributes upon a certain variable while considering uncertainty, P-box plots offer a strong approach because they provide both upper and lower limits for potential values. Once again, this approach works well when one needs to understand how far apart or extreme values may be, thus making it an effective partner for any other kind of statistical analysis [32].

To complement the feature importance, SelectKBest and SHAP



(a)



(b)

Fig. 11. SHAP summary plots for variables' impacts on PMV via RF (a), and SVM (b) as one of the investigated ML models implementing the modified ASHRAE Database.

analysis, P-box plots were generated to explore the relationship between key environmental features and the PMV. Fig. 12 shows the P-box plots for RF, which show how the PMV is spread out across different ranges of important factors, such as T_a, T_r, and other environmental variables. As shown in Fig. 12, there is a clear trend that higher values of T_a and T_r lead to higher PMV values. This fits well with the results of the correlation and SHAP analyses. These graphs describe the overall view of how PMV varies not only with the central tendencies but also with the entire distribution and variations. Hence, it helps in identifying potential outliers. The P-box plots unveil quite important facts in that the relationship between T_a, T_r, and PMV is linear and strong. Besides, a number of other variables, for example clo and M, show rather complicated, almost nonlinear relationships with PMV. These are quite important findings to make sure that chosen features bring out a good representation of the dynamics that affect thermal comfort. Thus, the previously undertaken feature selection is strongly substantiated, and a robust framework for P-box analysis is provided to gain better insight into the interaction between environmental conditions and thermal comfort as modeled by PMV.

The P-box plots unveil critical facts in that the relationship between T_a, T_r, and PMV is linear and strong. Besides, several other variables, for example, clo and M, show rather complicated, almost nonlinear relationships with PMV. These are quite important findings to make sure that chosen features bring out a good representation of the dynamics that affect thermal comfort. Thus, the feature selection conducted earlier is strongly supported, and a robust framework for P-box analysis is provided to enhance understanding of the interaction between environmental conditions and thermal comfort, as modeled by PMV. P-box analysis was conducted for the SVM model illustrated in Fig. 13, and the results corresponded closely with those of the RF model. Similarly, Tr

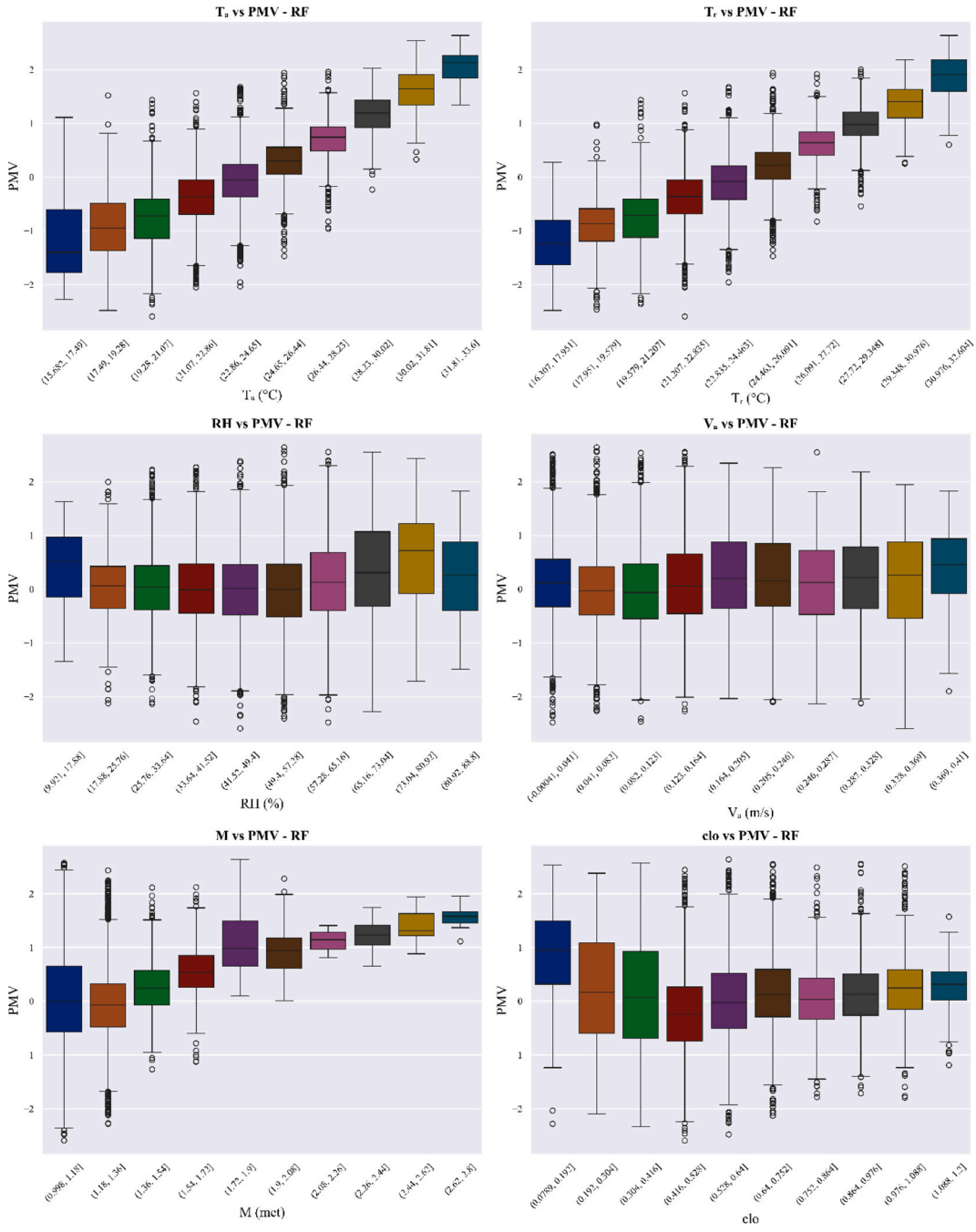


Fig. 12. P-box plots, processing the modified ASHRAE Database, for the relationship between PMV and key environmental features using RF model.

and Ta values increase alongside PMV, with other components exhibiting analogous trends. This evidence demonstrates the validity of the RF model analysis.

3.1.5. Partial dependence plots

Commonly used ML tools, PDP, use one feature to predict an outcome while maintaining the rest constant. The projected outcomes for different cases show how other variables affect the outcome measures through an interactive plot of predicted values resulting from just one input [33]. This technique is particularly useful when interpreting complex models, such as ensemble methods, because it allows one variable to be understood at a time, isolated from other influences, making its behavior more obvious.

The final analysis involved creating what are known as PDPs to better visualize the influence of selected features on PMV values. PDP allows us to graphically display the marginal effect of a feature on the predicted outcome while holding all other features at some value. Fig. 14a contrasts the PDPs of T_a and T_r. The PDP in Fig. 14a shows that both T_a and T_r together exert a positive influence on the PMV, although T_r is slightly stronger. This result implies that T_r may have a more significant impact on thermal comfort than the other two variables. There are almost linear relationships between the two plots, which means that if either T_a or T_r goes up, the PMV value will also go up, giving the impression of more warmth than there really is. But even though these graphs show clear effects of T_a or T_r on PMV, the choice to make the models simpler is still critical. Given their similar contributions to PMV and the marginally greater effect of T_r, it may be practical to use either of them in the modeling process. This study favored the use of T_a as an independent variable due to its purported independence from other parameters, including T_g. Therefore, selecting T_a necessitates avoiding

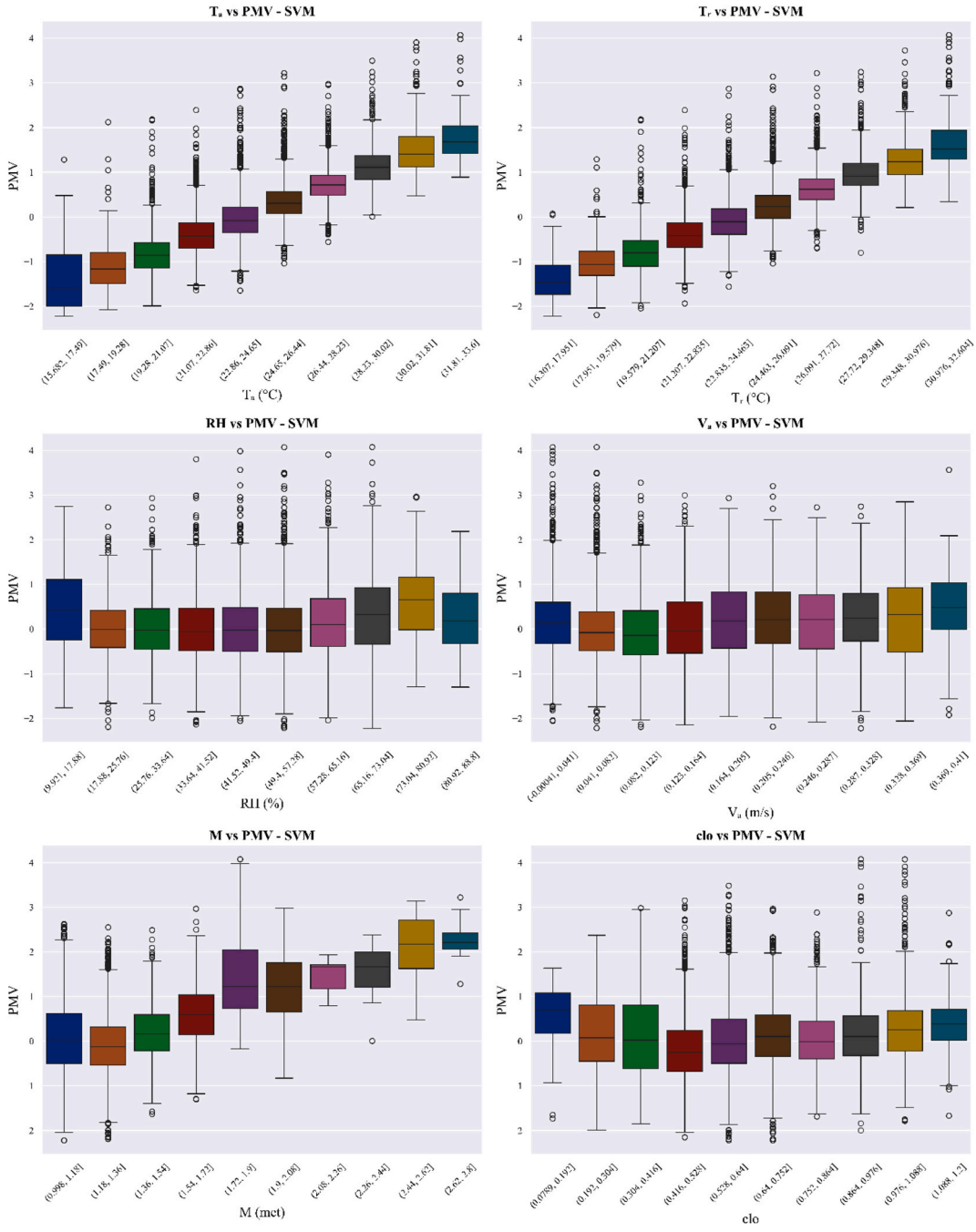
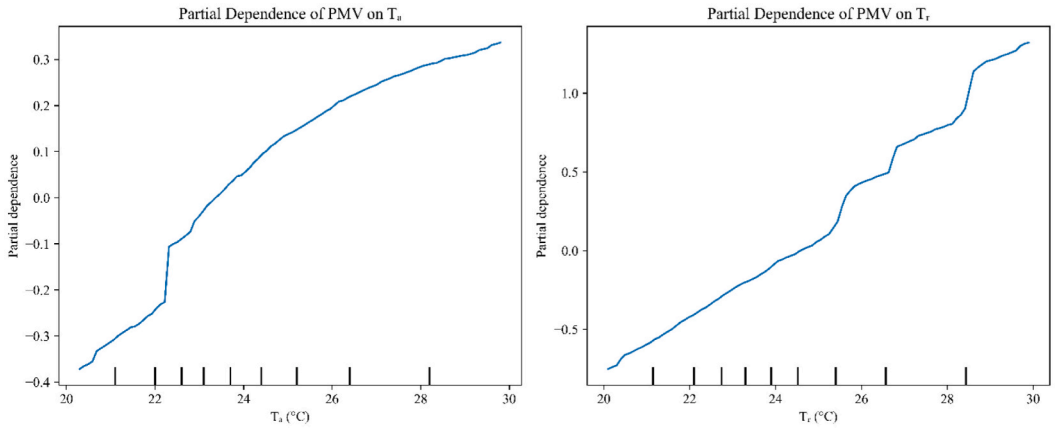
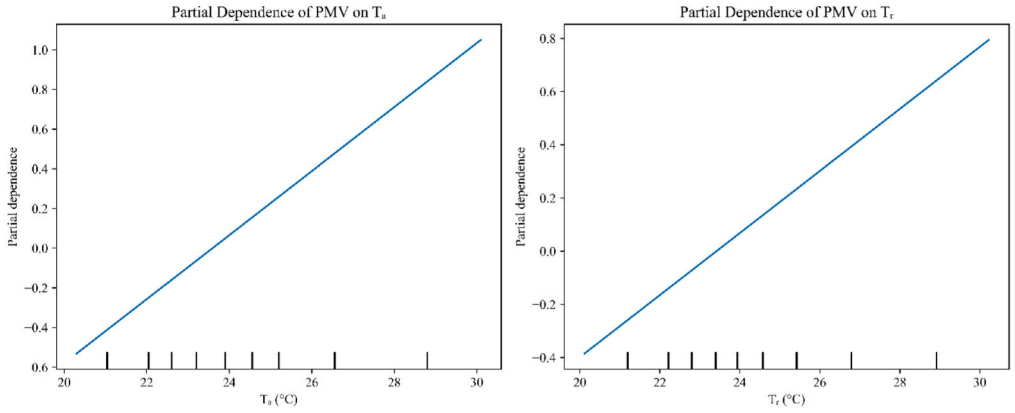


Fig. 13. P-box plots, processing the modified ASHRAE Database, for the relationship between PMV and key environmental features using SVM model.



(a)



(b)

Fig. 14. Partial dependence of PMV values of the modified ASHRAE Database on T_a and T_r using RF (a) and SVM (b) models.

more polarized forms, making the predictions robust yet manageable. As shown in Fig. 14b, the RF model outperforms the SVM model when predicting PMV. The RF model effectively shows the relation between variables such as T_a, T_r, and PMV, while the SVM model only shows these relationships in a linear manner. The RF model clearly shows the rapid increase in PMV after 26 °C, while the SVM model ignores this change and assumes a linear increase. The PDP plot of T_a and T_r shows that RF provides a more adaptable method to estimate complex variables such as PMV.

3.2. *Final features' validation*

The ASHRAE Global Thermal Comfort Database is included in this study; it encompasses almost all geographical regions and building types in the world; therefore, we can assume that it reflects all meteorological circumstances and design scenarios, approximately. The dataset's depth and variety, encompassing several buildings, seasons, and both interior and outside temperature values, enhance the model's adaptability to diverse settings with 107,583 data entries. Specifically, during the data preparation and filtering stages, any outliers were removed and the variables were normalized to set standard ranges in order to make the model more general. Five main factors contribute to the overfitting issue in ML applications: small datasets, complex models, noisy environments, excessive features, and excessive training time. The only thing that could have caused this study to be too good was using too many features. This problem was fixed by cutting the PMV formula from six parameters to three using the ASHRAE Global Thermal Comfort Database. The Chinese thermal comfort dataset [34], developed by seven institutes, consists of 41,977 data sets from 49 cities spanning five temperature zones over a span of two decades. Moreover, to validate this process, the Chinese Thermal Comfort Dataset has been analyzed by the same code written in Python, and the same results on the reduction issue were achieved as shown in Figs. 15, 16, 17, and 18. Finally, while feature importance, the SelectKBest, SHAP, P-box, and PDP analyses were used for feature selection, a comparative analysis of feature selection scores across multiple datasets improves the robustness of the findings. The study has cross-validation and sensitivity analysis on different training subsets.

The consistent results, confirmed by PDP and SHAP, showed that T_a, clo, and M were the most important factors in ML analyses. These are thermal comfort parameters used in HVAC systems to predict PMV. These results were derived from both importance scores: Feature importance and SelectKBest. These variables showed not only the highest scores but also the most appreciable and consistent effect on PMV in all the analyses performed. In this way, a large multimethod validation process assures the final model is highly statistically robust and very interpretable.

Based on the evidence from Feature Importance, SelectKBest, SHAP, P-Box, and PDP analyses, T_a, clo, and M were chosen as the most important features for PMV. All these features received the highest importance scores across all five analyses and exhibited significant effects on PMV. The statistical justification and the visual validation of the selection ensure that the final model is both accurate and

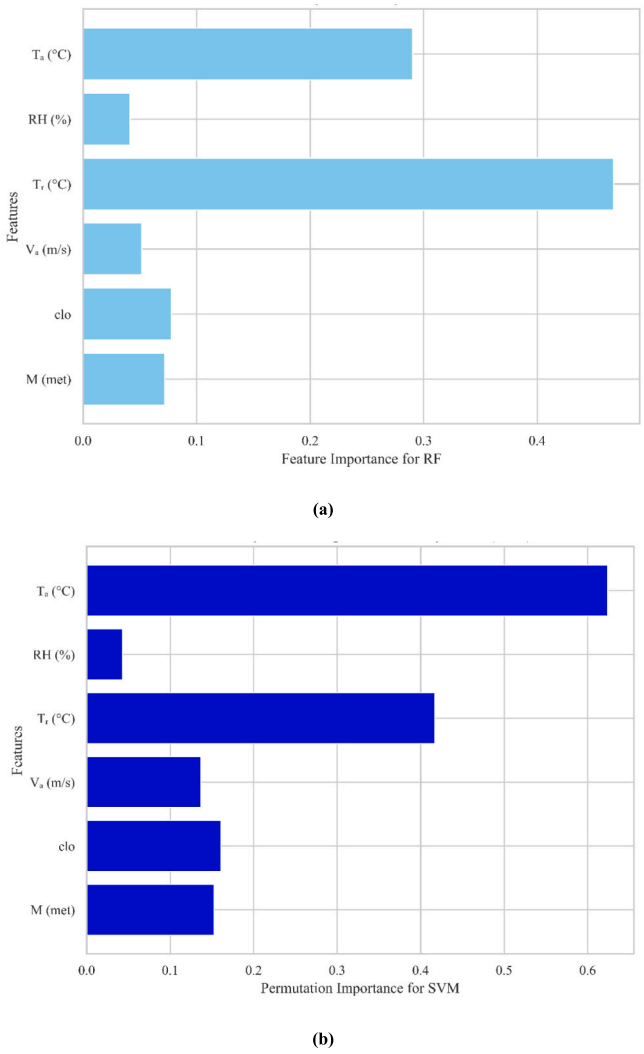


Fig. 15. Feature importance analyses as one of the investigated feature selection processes for RF (a), and SVM (b) ML models implementing the Chinese Database.

comprehensible. Therefore, this feature selection process has been thorough in singling out T_a, clo, and M as key variables driving PMV. The choice made makes the model more accurate while keeping it simple. This builds on what was already done and makes it possible to do a more in-depth study of thermal comfort in indoor settings.

Using Fanger's PMV parameters [6], the selection of final features was T_a, clo, and M after an extensive feature selection approach that employed Feature Importance, SelectKBest, and SHAP analysis as well as validation with PDP. At first, both T_a and T_r were thought to be important variables because they had high significance ranking scores and showed similar trends across all methods. However, it turned out that they had similar effects on predicting PMV, with neither playing a major role according to the PDP. Despite this, both factors continue to significantly contribute to the overall prediction accuracy of the model. Since T_r often depends on other factors like T_g and V_a, which complicate the model more than it should be, T_a was chosen instead to avoid any possible dependencies from these aforementioned factors. Thus, clo and M are the final selected variables that maintain model fidelity while explaining the main drivers of thermal comfort without making it too complex. Therefore, the resultant model is simple yet accurate, making it applicable to indoor environmental investigations. This method can help with energy management and comfort by quickly and easily figuring out the PMV value. This is especially helpful in places where people don't interact with each other much, like nursing homes, nurseries, or elder care facilities.

All the works that were summed up in the introduction use ML techniques and the ASHRAE Global Thermal Comfort Database to accurately find TSV, TPV, TP, TCV, and PMV. None of them has used Feature importance, the SelectKBest, SHAP, P-box, and PDP analyses combined with RF and SVM ML models to reduce the six thermal comfort parameters to three and determine the PMV, respectively. In addition, there are two ML methods used for comparison in this study. The SVM method achieves 70 % accuracy, while the RF method achieves 85 %. The reason behind this difference is due to the fact that RF outperforms SVM in accuracy due to its robustness, ability to capture input diversity, and improved model stability. RF's tree-based design provides greater tolerance for variable uncertainty after computation, even if it exhibits larger errors than SVM's estimates. Studies have shown that, although it is difficult to determine the ideal parameter combination within the available dataset, RF effectively encapsulates the variability in the data and improves model stability [35]. RF's exceptional accuracy is attributed to these key elements. According to our literature review as shown in Table 1, apart from Park and Woo's [21] study with 89.7 % accuracy, 25,261 reduced data entries, and three thermal comfort parameters demanding expansive thermal comfort measurement equipment, there is no study near to our 85 % accuracy obtained for 55,443 data entries with three cost-effective thermal comfort parameters using the ASHRAE Database. It should be noted that there are some attempts having the same aim using a specific database whose data entries are much less than ASHRAE ones. As it is known that ML methods give

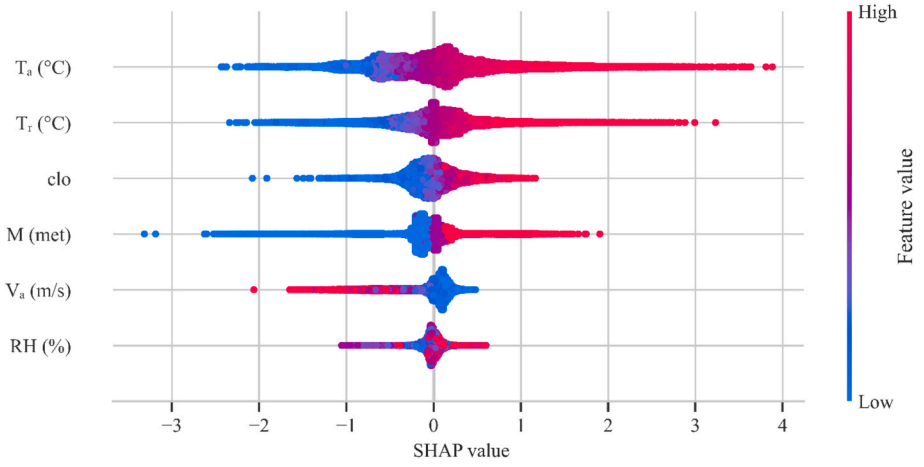


Fig. 16. SHAP summary plots for variables' impacts on PMV via RF as one of the investigated ML models implementing the Chinese Database.

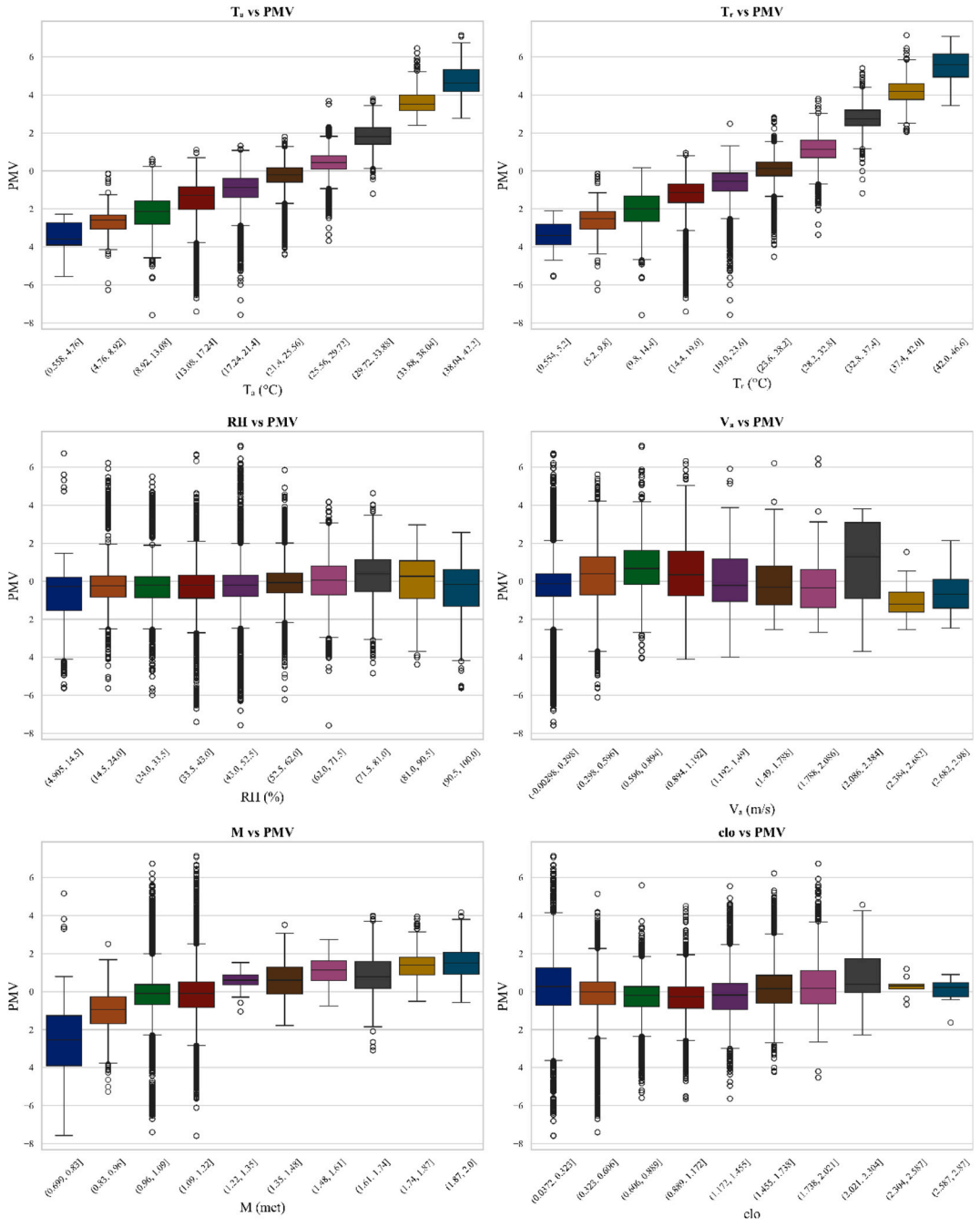
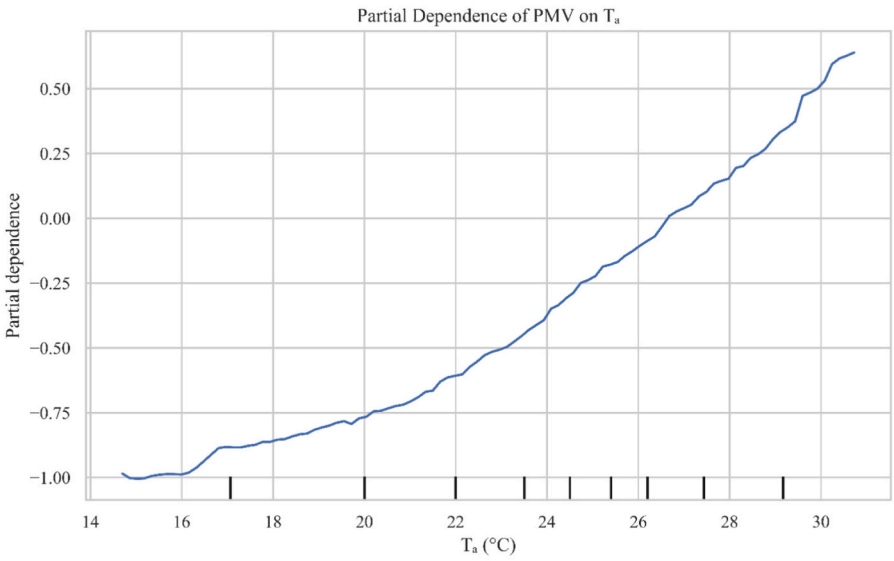


Fig. 17. P-box plots, processing the Chinese Database, for the relationship between PMV and key environmental features using RF model.

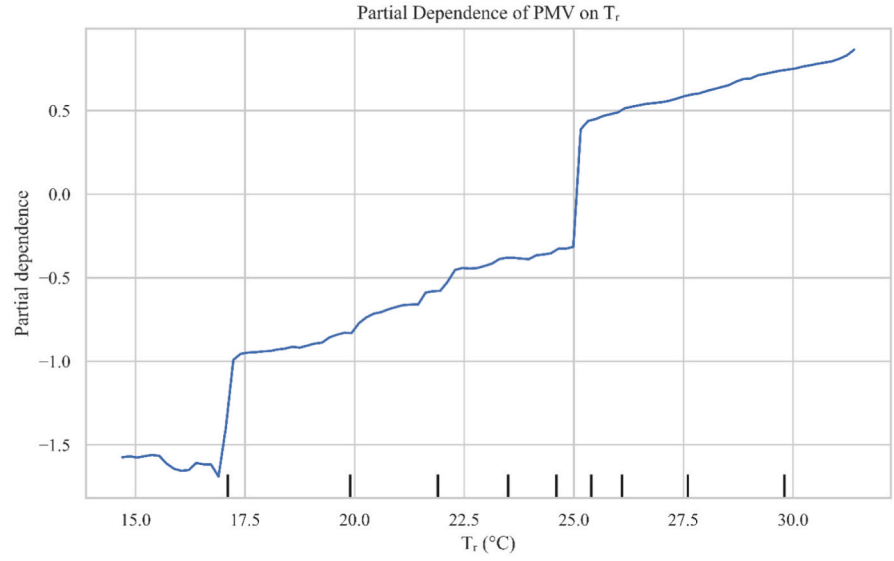
reliable results if they have many diverse data points. Therefore, the Chinese Thermal Comfort Dataset, having 41,977 entries, is preferred to validate our Feature importance, the SelectKBest, SHAP, P-box, and PDP analyses, which are the bases of developed ML models, on the thermal comfort parameters' reduction from six to three.

4. Conclusion

The ASHRAE Global Thermal Comfort Database provides a significant advance in improving thermal comfort prediction through ML. This numerical investigation provides the most comprehensive data cleaning procedure in the literature, which enhances the reliability of thermal comfort indices and paves the way for smarter HVAC systems. With 55,443 observations reduced from 107583, the evaluated database emphasizes the interaction between environmental factors and comfort. As smart building technologies evolve, integrating ML will boost energy efficiency and occupant well-being. The database also supports ongoing research and model validation. As the science progresses, it's crucial to consider thermal comfort's complexity, particularly for vulnerable populations, to design inclusive environments. The stringent data cleaning and standardization procedures enhance the reliability of the thermal comfort indices, laying the groundwork for advanced and flexible HVAC systems. Furthermore, when Feature importance, SelectKBest, SHAP, and PDP analyses were put together, they showed that T_a, clo, and M were the most important factors in predicting PMV. These features ensure that the final model is statistically robust, interpreted, and simple, enabling better analysis of indoor thermal comfort while avoiding unnecessary complexity. ML techniques must be integrated into smart building technologies as they advance to create systems that can adapt dynamically to real-time situations, eventually



(a)



(b)

Fig. 18. Partial dependence of PMV values of the Chinese Database on T_a (a) and T_r (b) using RF model.

enhancing occupant health and well-being.

This study is essential as it can minimize the number of input variables needed for calculations, hence decreasing the quantity of measuring equipment required, based on the literature utilizing ASHRAE databases and ML techniques. To the best of authors' knowledge, no research approaches our 85 % accuracy achieved with 55,443 data entries utilizing three cost-effective thermal comfort parameters from the ASHRAE Database obtained by affordable measurement devices. The correlation matrix showed that the variables influencing thermal comfort were highly interrelated. In particular, the T_a and T_r had a high correlation coefficient of 0.92, showing that there is a strong relationship between these two variables. Moreover, T_a and T_r had strong correlations with PMV, with 0.78 and 0.76 correlation coefficients, respectively, showing that the temperature-related variables are the critical variables determining thermal comfort. On the other hand, clo demonstrated a moderate negative correlation with T_a (−0.47), which reflects how people change their clo behavior as temperatures increase and advance energy efficiency. In conclusion, the methodology demonstrated in this study indicates that PMV may be forecasted with fewer variables, providing benefits in usability and cost-effectiveness. The following key findings emerged from the analyses:

a. The Feature importance method indicated T_r as the most significant characteristic, with a feature importance score of 0.5. Moreover, clo and M were identified as critical attributes, although RH and V_a showed minimal impact on the model's predictive accuracy.
b. The SelectKBest analysis identified T_a and T_r as the most significant variables, indicating considerable statistical relevance and highlighting their critical importance in the model. In contrast, M and clo demonstrated lower scores, indicating reduced individual contributions and suggesting that their impact on the target variable is less direct or more subtle.
c. The SHAP analysis indicated that T_a was the most significant feature, demonstrating considerable dependencies and variability in PMV predictions. clo and M significantly influenced PMV, whereas T_r had a comparatively subdued effect in relation to T_a, clo, and M. V_a and

RH exhibited minimal influence, warranting their omission from the final model to enhance simplicity.

d. P-box analysis revealed a linear and strong relationship between T_a, T_r, and PMV, whereas clo and M exhibited more complex, nonlinear relationships. Similarly, PDP demonstrated that both T_a and T_r had a positive linear effect on PMV, with T_r exerting a slightly stronger influence. Given T_r's dependencies, T_a was selected as the more independent predictor.
e. Specifically, incorporating ML models with data from the ASHRAE Global Thermal Comfort Database identifies T_a, clo, and M as the most important PMV predictors. SHAP and PDP analyses developed based on the RF model, which is both robust and interpretable, are the most successful ones with an accuracy of 85 % to support indoor thermal comfort predictions in comparison to the SVM model with 70 % accuracy under the same circumstances.
f. The Chinese Thermal Comfort Dataset, with 41,977 entries, validates our feature significance analyses, including Feature Importance, SelectKBest, SHAP, P-box, and PDP, which support the developed ML models to reduce thermal comfort parameters from six to three.
g. Decision-makers like building managers will therefore have more faith in the model, and heating–cooling techniques will be more precisely molded in field applications.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors would like to acknowledge that this paper is submitted in partial fulfillment of the requirements for PhD degree at Yildiz Technical University. This study has been financially supported by Yildiz Technical University Scientific Research Projects Coordination Department, Project Number: FBA-2024-6401. The last author, Ahmet Selim Dalkılıç would like to acknowledge the support from by European Union's Research and Innovation Program Horizon Europe under the Marie Skłodowska-Curie grant agreement (No 101130406) and UKRI Engineering and Physical Sciences Research Council (EP/Y036662/1).

Data availability

Data will be made available on request.

References

[1] R.J. de Dear, A global database of thermal comfort field experiments, in: ASHRAE Transactions: Symposia, Vol. 104, American Society of Heating, Refrigeration and Air Conditioning Engineers, Inc., Atlanta, 1998: pp. 1141–1152. https://www.proquest.com/scholarly-journals/global-database-thermal-comfort-field-experiments/docview/192542601/se-2?accountid=17384 (accessed November 15, 2024).
[2] ISO, I, 7730: 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, Management 3 (605) (2005) e615.
[3] ASHRAE, A., ASHRAE Standard 55: Thermal environmental conditions for human occupancy, American Society of Heating, Refrigerating and Air-Conditioning Engineers, Atlanta, GA, USA, 2017.
[4] V. Földváry Ličina, T. Cheung, H. Zhang, R. de Dear, T. Parkinson, E. Arens, C. Chun, S. Schiavon, M. Luo, G. Brager, P. Li, S. Kaam, M.A. Adebamowo, M.M. Andamon, F. Babich, C. Bouden, H. Bukovianska, C. Candido, B. Cao, S. Carlucci, D.K.W. Cheong, J.-H. Choi, M. Cook, P. Cropper, M. Deuble, S. Heidari, M. Indraganti, Q. Jin, H. Kim, J. Kim, K. Konis, M.K. Singh, A. Kwok, R. Lamberts, D. Loveday, J. Langevin, S. Manu, C. Moosmann, F. Nicol, R. Ooka, N.A. Oseland, L. Pagliano, D. Petráš, R. Rawal, R. Romero, H.B. Rijal, C. Sekhar, M. Schweiker, F. Tartarini, S. Tanabe, K.W. Tham, D. Teli, J. Toftum, L. Toledo, K. Tsuzuki, R. De Vecchi, A. Wagner, Z. Wang, H. Wallbaum, L. Webb, L. Yang, Y. Zhu, Y. Zhai, Y. Zhang, X. Zhou, Development of the ASHRAE Global Thermal Comfort Database II, Build Environ 142 (2018) 502–512. Doi: 10.1016/j.buildenv.2018.06.022.
[5] A.A. Farhan, K. Pattipati, B. Wang, P. Luh, Predicting individual thermal comfort using machine learning algorithms, in: In: 2015 IEEE International Conference on Automation Science and Engineering (CASE), 2015, pp. 708–713, https://doi.org/10.1109/CoASE.2015.7294164.
[6] P.O. Fanger, Thermal comfort. Analysis and applications in environmental engineering, 1970.
[7] S. Lu, W. Wang, C. Lin, E.C. Hameen, Data-driven simulation of a thermal comfort-based temperature set-point control with ASHRAE RP884, Build Environ 156 (2019) 137–146, https://doi.org/10.1016/j.buildenv.2019.03.010.
[8] T. Parkinson, R. de Dear, G. Brager, Nudging the adaptive thermal comfort model, Energy Build 206 (2020) 109559, https://doi.org/10.1016/j.enbuild.2019.109559.
[9] Z. Wang, J. Wang, Y. He, Y. Liu, B. Lin, T. Hong, Dimension analysis of subjective thermal comfort metrics based on ASHRAE Global Thermal Comfort Database using machine learning, J. Build. Eng. 29 (2020) 101120, https://doi.org/10.1016/j.jobe.2019.101120.
[10] M. Luo, J. Xie, Y. Yan, Z. Ke, P. Yu, Z. Wang, J. Zhang, Comparing machine learning algorithms in predicting thermal sensation using ASHRAE Comfort Database II, Energy Build 210 (2020) 109776, https://doi.org/10.1016/j.enbuild.2020.109776.
[11] X. Zhou, L. Xu, J. Zhang, B. Niu, M. Luo, G. Zhou, X. Zhang, Data-driven thermal comfort model via support vector machine algorithms: Insights from ASHRAE RP-884 database, Energy Build 211 (2020) 109795, https://doi.org/10.1016/j.enbuild.2020.109795.
[12] G. Gao, J. Li, Y. Wen, DeepComfort: Energy-Efficient Thermal Comfort Control in Buildings Via Reinforcement Learning, IEEE Internet Things J 7 (2020) 8472–8484, https://doi.org/10.1109/JIOT.2020.2992117.
[13] N. Ma, L. Chen, J. Hu, P. Perdikaris, W.W. Braham, Adaptive behavior and different thermal experiences of real people: A Bayesian neural network approach to thermal preference prediction and classification, Build Environ 198 (2021) 107875, https://doi.org/10.1016/j.buildenv.2021.107875.
[14] H. Park, D.Y. Park, Prediction of individual thermal comfort based on ensemble transfer learning method using wearable and environmental sensors, Build Environ 207 (2022) 108492, https://doi.org/10.1016/j.buildenv.2021.108492.
[15] B. Lala, H. Rizk, S.M. Kala, A. Hagishima, Multi-Task Learning for Concurrent Prediction of Thermal Comfort, Sensation and Preference in Winters, Buildings 12 (2022) 750, https://doi.org/10.3390/buildings12060750.
[16] Y. Bai, K. Liu, Y. Wang, Comparative analysis of thermal preference prediction performance in different conditions using ensemble learning models based on ASHRAE Comfort Database II, Build. Environ. 223 (2022) 109462, https://doi.org/10.1016/j.buildenv.2022.109462.
[17] H. Lan, H. (Cynthia) Hou, Z. Gou, A machine learning led investigation to understand individual difference and the human-environment interactive effect on classroom thermal comfort, Build Environ 236 (2023) 110259. Doi: 10.1016/j.buildenv.2023.110259.
[18] X. Feng, E. Bin Zainudin, H.W. Wong, K.J. Tseng, A hybrid ensemble learning approach for indoor thermal comfort predictions utilizing the ASHRAE RP-884 database, Energy Build. 290 (2023) 113083, https://doi.org/10.1016/j.enbuild.2023.113083.
[19] G. Lamberti, R. Boghetti, J.H. Kämpf, F. Fantozzi, F. Leccese, G. Salvadori, Development and comparison of adaptive data-driven models for thermal comfort assessment and control, Total Environ. Res. Themes 8 (2023) 100083, https://doi.org/10.1016/j.totert.2023.100083.
[20] L. Yang, F. Wang, S. Zhao, S. Gao, H. Yan, Z. Sun, Z. Lian, L. Duanmu, Y. Zhang, X. Zhou, B. Cao, Z. Wang, Y. Zhai, Comparative analysis of indoor thermal environment characteristics and occupants' adaptability: Insights from ASHRAE RP-884 and the Chinese thermal comfort database, Energy Build 309 (2024) 114033, https://doi.org/10.1016/j.enbuild.2024.114033.
[21] K.Y. Park, D.O. Woo, PMV Dimension Reduction Utilizing Feature Selection Method: Comparison Study on Machine Learning Models, *Energies* 16 (5) (2023) 2419.
[22] F. Hartwig, B.E. Dearing, Exploratory Data Analysis, No. 16, Sage, 1979.
[23] C.F. Dormann, J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J.R.G. Marquéz, B. Gruber, B. Lafourcade, P.J. Leitão, T. Münkemüller, C. McClean, P.E. Osborne, B. Reineking, B. Schröder, A.K. Skidmore, D. Zurell, S. Lautenbach, Collinearity: a review of methods to deal with it and a simulation study evaluating their performance, Ecography 36 (2013) 27–46, https://doi.org/10.1111/j.1600-0587.2012.07348.x.
[24] ISO 7726:1998 Ergonomics of the thermal environment — Instruments for measuring physical quantities, 1998. https://www.iso.org/standard/14562.html (accessed November 15, 2024).
[25] W.S. Cleveland, A Model for Studying Display Methods of Statistical Graphics, J. Comput. Graph. Stat. 2 (1993) 323, https://doi.org/10.2307/1390686.
[26] L. Breiman, Random Forests. Machine Learning 45 (2001) 5–32.
[27] G. Louppe, Understanding random forests: From theory to practice (Doctoral dissertation, Universite de Liege (Belgium)), 2014.
[28] A.J. Smola, B. Schölkopf, A tutorial on support vector regression, Stat. Comput. 14 (2004) 199–222.
[29] L. Breiman, Random Forests, Mach Learn 45 (2001) 5–32, https://doi.org/10.1023/A:1010933404324.
[30] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer New York, New York, NY, 2009. Doi: 10.1007/978-0-387-84858-7.
[31] S.M. Lundberg, S.-I. Lee, A Unified Approach to Interpreting Model Predictions, in: I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Adv Neural Inf Process Syst, Curran Associates, Inc., 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[32] S. Ferson, W. Troy Tucker, Sensitivity analysis using probability bounding, Reliab. Eng. Syst. Saf. 91 (2006) 1435–1442, https://doi.org/10.1016/j.ress.2005.11.052.
[33] J.H. Friedman, Greedy Function Approximation: A Gradient Boosting Machine, Annals Statistics 29 (2001), https://doi.org/10.1214/aos/1013203451.
[34] L. Yang, S. Zhao, Y. Zhai, S. Gao, F. Wang, Z. Lian, R. de Dear, The Chinese thermal comfort dataset, Sci. Data 10 (1) (2023) 662.
[35] B.F. Huang, P.C. Boutros, The parameter sensitivity of random forests, BMC Bioinf. 17 (2016) 1–13.