Journal of Construction Engineering, Management & Innovation
2025 8(4):346-379
DOI 10.31462/jcemi.2025.04346379

RESEARCH ARTICLE

Ensemble machine learning algorithms for thermal comfort prediction in HVAC systems of smart buildings

Merve Kuru Erdem[1], Osman Gökalp[2], Gulben Calis[1]

[1] Ege University, Faculty of Engineering, Department of Civil Engineering, İzmir, Türkiye
[2] Izmir Institute of Technology University, Faculty of Engineering, Department of Computer Engineering, İzmir, Türkiye

Article History

Received 12 March 2025
Revised 02 October 2025
Accepted 27 October 2025

Keywords

Thermal comfort prediction
HVAC control
Machine learning
Ensemble learning
Random forest
Gradient boosting

Abstract

Predicting the thermal comfort of building occupants is of paramount importance in the operation of smart buildings, providing a data-driven approach to control Heating, Ventilation, and Air Conditioning (HVAC) systems for managing occupant thermal comfort and energy use, which aligns with modern sustainability and efficiency goals. Recently, ensemble machine learning (ML)-based thermal comfort prediction models have been proposed to provide more accurate estimation of thermal comfort; however, these efforts often lack a systematic and comprehensive evaluation across a wide range of ML models within a single study. To address this gap, this study presents a systematic comparative analysis of four ensemble ML frameworks (bagging, boosting, stacking, and voting) with six basic ML algorithms (Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Decision Tree, Multilayer Perceptron, and Multinomial Naïve Bayes) and six advanced ensemble ML algorithms (Random Forest, Rotation Forest, Extra Trees, Gradient Boosting Classifier, Histogram Gradient Boosting Classifier, and Extreme Gradient Boosting). The analysis is conducted using the widely recognized ASHRAE Global Thermal Comfort Database II, providing both 3-point and 7-point Thermal Sensation Vote (TSV) predictions. Accuracy, precision, recall and F1 metrics are used for evaluation and 10-fold cross validation is applied for further comparison. The results demonstrate the Histogram Gradient Boosting (HGB) algorithm achieved the highest F1 score (0.638) for 7-point TSV prediction whereas the Random Forest (RF) algorithm provided the highest F1 score (0.549) for 7-point TSV prediction. In practice, these findings suggest that integrating RF and HGB models into Building Management Systems or IoT-based HVAC platforms can support real-time adaptive control, helping practitioners to reduce energy use while maintaining occupant comfort.

1. Introduction

Thermal comfort is a subjective measure of how comfortable a person feels in their environment and maintaining a comfortable indoor environment is important for the health and productivity of building occupants [1, 2]. As a result, ensuring thermal comfort within buildings is of great importance, and HVAC systems are predominantly employed to achieve and regulate it. Consequently, a large share of building energy consumption is dedicated to heating and cooling processes [3].

Correspondence Gulben Calis gulben.calis@ege.edu.tr

eISSN 2630-5771

Therefore, accurately predicting thermal comfort has become a crucial step toward designing energy-efficient, occupant-centric building management strategies. However, predicting the thermal comfort of building occupants presents significant challenges attributable to the several factors such as the subjective nature of comfort perception, demographic factors influencing occupant comfort (i.e. age, gender) and limited availability of comprehensive occupant data.

In recent years, accurate prediction of Thermal Sensation Vote (TSV) has become crucial for optimizing HVAC systems in buildings, contributing to enhanced occupant comfort and energy efficiency [4]. Despite several studies investigating the use of ML techniques for TSV prediction, there remains a lack of systematic and comprehensive evaluation of both basic and advanced ensemble ML algorithms using the widely recognized ASHRAE Global Thermal Comfort Database II. This study aims to address this gap by providing a systematic comparative analysis of sixteen ML models, including four ensemble ML frameworks with six basic ML algorithms and six advanced ensemble ML algorithms, using both 3-point and 7-point TSV scales. By adopting a unified dataset and experimental approach, the study offers a robust evaluation of model performance across key metrics, thereby informing future research and practical applications in the field.

To address the drawbacks, the objectives of this study are determined as follows:

1. Generalizability of Results: To predict TSV using the comprehensive ASHRAE Comfort Database II to ensure the generalizability of TSV prediction models across diverse occupant feedback.
2. Evaluation of Ensemble ML Algorithms: To systematically compare the performance of four ensemble ML frameworks (bagging, boosting, stacking, and voting) combined with six basic ML algorithms (Logistic Regression (LR), K-Nearest Neighbors (KNN), Decision Tree (DT), Support Vector Machine (SVM), Multinomial Naive Bayes (MNB), Multilayer Perceptron (MLP)), and six advanced ensemble ML algorithms (Random Forest (RF), Rotation Forest (ROF), Extra Trees (XT), Gradient Boosting Classifier (GB), Histogram Gradient Boosting Classifier (HGB), and Extreme Gradient Boosting (XGB)).
3. Impact of TSV Scale Units: To provide a detailed analysis of how different TSV scale units (3-point and 7-point) impact the prediction performance of various ML models, offering new insights into the selection of scale units in thermal comfort prediction.

By achieving these objectives, this study seeks to identify the most suitable ML algorithm for TSV prediction to be integrated into HVAC systems for improved energy efficiency and thermal comfort in buildings. The remainder of this paper is structured as follows: Section 2 (Literature Review) presents the existing studies and analyzes the research gap in the literature, Section 3 (Methods) outlines the research design, including data description and preprocessing, principles of ensemble ML algorithm, and performance metrics. Section 4 (Results) presents the experimental setup, test performance for the basic, ensemble and advanced ML models as well as further analysis of the results. Section 5 (Discussion) provides a comparative interpretation of the model outcomes, presents the importance of variables, and discusses their practical implications of thermal comfort prediction in HVAC systems of smart buildings. Finally, Section 6 (Conclusion) summarizes the key findings, highlights the practical relevance of TSV prediction in smart buildings, and offers directions for future research in intelligent building operations.

2. Literature Review

There are mainly two traditional models, namely the Predicted Mean Vote (PMV) and the adaptive, that are used for predicting thermal comfort in buildings. The PMV model calculates the thermal comfort of building occupants based on the indoor environmental variables, clothing insulation (CLO), and metabolic rate (MET) [5], however; numerous studies have already identified discrepancies between the PMV and actual thermal

comfort of occupants [6–15]. Broday et al. [8] evaluated the uncertainty of the PMV model and Predicted Percentage of Dissatisfied (PPD) by using Monte Carlo method and concluded that the accuracy of the PMV model is highly affected by the uncertainties in the input data. Zhou et al. [16] found that the PMV model overestimates the actual thermal sensation of occupants at high outdoor temperatures and solar radiation intensities. Cheung et al. [9] assessed the reliability of the PMV model in predicting thermal comfort by utilizing the ASHRAE Global Thermal Comfort Database II. The results show that the PMV model tends to underestimate thermal discomfort in hot and humid climates and to overestimate it in cold and dry climates. Their investigation revealed that the PMV model's ability to predict accurately was limited, with an overall accuracy rate of only 34%. The second main traditional thermal comfort prediction model is the adaptive model, which provides thermal comfort prediction of occupants while considering the thermal preferences of building occupants and the adjustments made by occupants to maintain their comfort requirements. Jiao et al. [17] developed adaptive thermal comfort models based on the principles of the PMV model and validated that the developed models obtained more precise predictions of thermal comfort than the PMV model. Du et al. [7] compared the accuracy of the PMV model and the adaptive-PMV model by using the Chinese Thermal Comfort Database and found that the adaptive-PMV model yielded more precise predictions for thermal comfort than the PMV model. Although, adaptive thermal comfort models are becoming more widely recognized and accepted in building regulations, its implementation constitutes some limitations and challenges including the need for greater consistency, standardization and more data on adaptive behavior in different cultures [18].

To overcome these challenges, ML-based thermal comfort models have been proposed to provide more accurate predictions of thermal comfort, adapt in real time, process complex data, and integrate seamlessly into smart building systems [19]. ML approaches have emerged as promising alternatives to conventional thermal comfort models, such as the Predicted Mean Vote (PMV), which often fail to capture individual and contextual variations in real-world environments. A comprehensive review by Qavidel Fard et al. [20] emphasized that ML algorithms not only outperform PMV in predictive accuracy but also offer flexibility to incorporate diverse inputs, including physiological signals, environmental variables, and behavioral factors. Importantly, their analysis highlighted ensemble techniques and feature engineering as underexplored but critical avenues for robust thermal comfort modeling.

Based on these findings, Yu et al. [21] investigated the trade-off between accuracy and deployability by comparing algorithms across high-precision and low-cost sensing scenarios. Their findings revealed that while sophisticated inputs yield better accuracy, practical systems can still achieve reliable results with fewer variables, underscoring the need for application-specific model design. Similarly, Yang et al. [22] investigated discriminant analysis, decision trees, ensemble algorithms, and KNN using laboratory data enriched with physiological and demographic features. Their results suggested that model success depended less on raw accuracy and more on the ability to capture non-linear, high-dimensional interactions between environmental and personal variables. In another study, Yang et al. [23] extended this approach with broader physiological coverage, showing that algorithmic sensitivity to input selection was as important as the choice of ML method itself. Cen et al. [24] further demonstrated the feasibility of lightweight, wearable-based models, showing that even minimal inputs could provide reliable real-time comfort assessment, paving the way for non-intrusive monitoring.

Other studies have explored the role of algorithm choice and data preprocessing in shaping predictive performance. Zhou et al. [25] found that SVM regression, particularly with radial basis kernels, was effective in modeling nonlinear comfort responses, reinforcing the adaptability of kernel-based approaches. Rehman et al. [26]

showed that deeper networks, such as deep artificial neural networks, excelled at capturing the complex relationships between environmental and physiological variables, while also stressing the importance of matching model complexity to data characteristics. Fayyaz et al. [27] highlighted how imputation, feature selection, and class balancing significantly influenced outcomes, discussing that data preprocessing techniques can be as decisive as algorithm choice. Regarding contextually, field-based studies such as Chai et al. [28] found that adaptation-related factors like climate, gender, and age were critical in naturally ventilated residences, while Sibyan et al. [29] demonstrated that even simple probabilistic classifiers like Naive Bayes could outperform regression when applied in resource-constrained, real-world settings.

Expanding on data diversity, Shan et al. [30] introduced electroencephalogram (EEG) signals into thermal comfort modeling, opening the field to neurophysiological sensing. Tardioli et al. [31] developed a hybrid method integrating Internet of Things (IoT) sensor data with building physics simulations, highlighting the benefits of combining mechanistic and data-driven models. Large-scale benchmarking efforts, such as Luo et al. [32], systematically compared nine algorithms on the ASHRAE Global Thermal Comfort Database II, showing that feature importance analyses revealed consistent drivers like indoor air temperature, relative humidity (RH), and CLO, regardless of algorithm choice. In terms of the role of input dimensionality, some studies [33] demonstrated that expanding feature sets improved accuracy, while others [34] showed that spatial attributes such as occupant distance proximity to windows or AC (air conditioning) units significantly improved robustness. Demographic-focused models [35] also confirmed that personal factors (e.g., age, education, income) mediate comfort perception, underscoring the need to account for social and cultural diversity in predictive frameworks.

Recent work has further extended ML-based comfort prediction into practical building operation and special population contexts. Boutahri and Tilioua [36] combined IoT-based Raspberry Pi sensors with ensemble methods (RF, XGBoost) to simultaneously optimize comfort and HVAC energy use, demonstrating the dual utility of ML for sustainability goals. Ren et al. [37] incorporated physiological signals such as ECG (electrocardiogram), EEG, and skin temperature, showing that advanced resampling techniques like SMOTETomek markedly improved predictive reliability, especially in imbalanced datasets. Li et al. [38] integrated Bayesian optimization and SHAP(SHapley Additive exPlanations)-based feature interpretability into XGBoost models for office buildings, identifying dominant drivers like air velocity (V_{air}), CLO, and mean radiant temperature, thus, offering insights beyond predictive accuracy. Avci [39] explored macro-contextual factors such as climate class, season, building type, and ventilation strategy across 61,000 global observations, demonstrating that while such inputs improved generalizability across settings, they sometimes reduced precision compared to physiological data. Finally, He et al. [40] focused on elderly occupants in climate chamber experiments, revealing that boosting methods like XGBoost not only provided superior accuracy but also captured behavioral adaptations such as fan use, highlighting the importance of tailoring models to vulnerable populations.

In summary, these studies reflect a paradigm shift in thermal comfort prediction research. Rather than competing solely on accuracy scores, the field is moving toward holistic frameworks that integrate environmental, physiological, spatial, demographic, and macro-contextual variables, while also emphasizing interpretability, generalizability, and real-world applicability. In addition, these studies demonstrate the growing sophistication of ML-based thermal comfort models, emphasizing the importance of model selection and evaluation. RF, XGBoost, and KNN frequently emerge as the most effective models, with Artificial Neural Network (ANN), SVM, and Naïve Bayes (NB) also showing competitive results in certain contexts. The analysis of the literature highlights several research gaps as outlined below:

- Most of the studies are based on either the field or lab data, which provide personalized outputs, and, thus, cannot be generalized. On the other hand, datasets such ASHRAE Thermal Comfort Database provide a wide range of feedback from occupants, which enable to obtain universal outputs. However, ASHRAE datasets are rarely used in studies focusing on the performance of ensemble ML frameworks and advanced ensemble ML algorithms.
- Various basic ML algorithms (i.e. SVM, KNN) have been frequently applied for thermal comfort prediction. Although ensemble ML frameworks (e.g., bagging, boosting, stacking, and voting) and advanced ensemble ML algorithms (e.g., Random Forest, Rotation Forest, Extra Trees, Gradient Boosting Classifier, Histogram Gradient Boosting Classifier, Extreme Gradient Boosting) have demonstrated superior performance in other engineering domains such as fault detection [31, 32], energy forecasting [33–36], and engineering planning and design [37], their comparative performance in thermal comfort prediction has not been systematically evaluated using a consistent dataset and experimental framework.
- TSV is the most common output of ML algorithms in thermal comfort prediction models [41–44]. It is seen that most of the studies used 7-point and 3-point TSV scale units, but it is still not clear to which extend the selected scale unit might affect the performance of the prediction models.

3. Methods

In this study, different ensemble ML frameworks that employ several basic ML algorithms and advanced ensemble ML algorithms that utilize specialized learning techniques are applied to ASHRAE Database II to systematically evaluate the performance of these methods in predicting thermal comfort. Fig. 1 presents an overview of the research design, which is elaborated in the following subsections.

3.1. Dataset description and data preprocessing

The ASHRAE Global Thermal Comfort Database II was selected for this study due to its comprehensive scope and global relevance. It is the most extensive publicly available dataset on thermal comfort, containing over 109,033 rows of data collected from multiple projects. This dataset's diversity, encompassing various building types, occupancy patterns, and environmental conditions, ensures a representative basis for developing generalizable thermal comfort prediction models. Unlike other datasets, such as the SCATs database, which focuses primarily on office environments in Europe, or the standalone RP-884 dataset, the ASHRAE Global Thermal Comfort Database II offers unmatched breadth and depth by consolidating data from diverse climates, regions, and populations worldwide. Initiated in 2014 under the leadership of the Center for the Built Environment at the University of California, Berkeley, and the University of Sydney's Indoor Environmental Quality Laboratory, the dataset includes 25,288 rows from the ASHRAE RP-884 database, to provide a robust foundation for thermal comfort research. Furthermore, its widespread adoption in recent studies (e.g., Cheung et al. [9], Bai et al. [45], Han et al. [46]) highlights its reliability and utility in advancing thermal comfort research.

In this study, the dataset obtained from the ASHRAE Global Thermal Comfort Database II (version 2.01) used in Luo et al.'s [32] study was used to compare the performance of ensemble machine learning algorithms in predicting thermal comfort of occupants with the traditional machine learning algorithms. TSV of the occupants in the ASHRAE Global Thermal Comfort Database II (version 2.01) were determined as the output variable. To select features, statistical analyses were performed to check whether the independent variables have a statistically significant effect on thermal sensation.

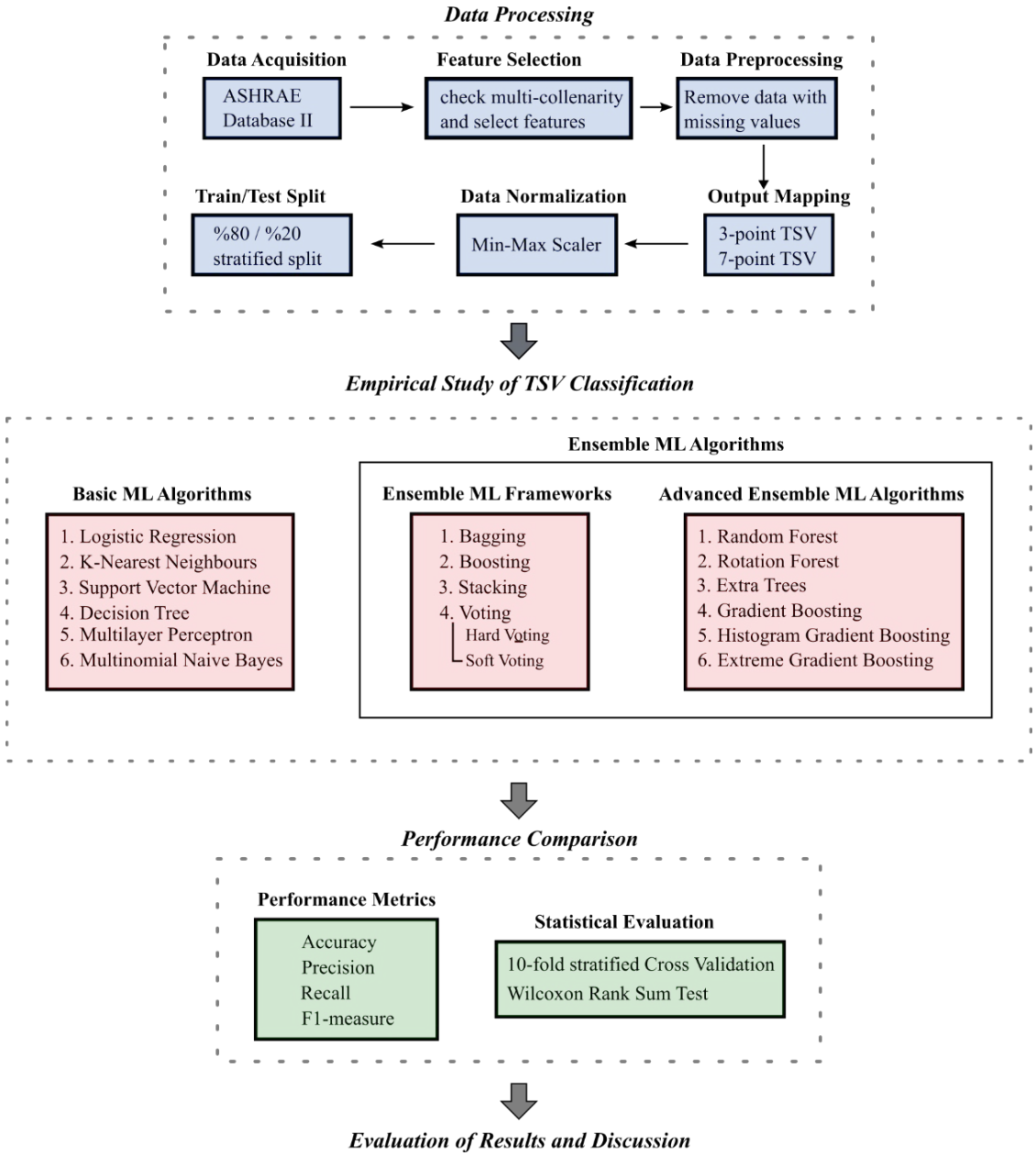

Fig. 1. Research design

Logistic regression analysis was applied to continuous variables such as indoor and outdoor parameters (T_{air}, V_{air}, RH, SET, T_{out}), whereas the chi-square test was used for categorical variables. Furthermore, to evaluate potential multicollinearity, the Pearson correlation analysis was conducted for indoor and outdoor air parameters. Subsequently, 'T_{air}', 'V_{air}', 'RH', 'SET', 'CLO', 'MET', 'Age', 'Sex', 'T_{out}', 'Season', 'Building operation mode', and 'Building type' were selected as input features to achieve higher model performance. Moreover, data rows that do not have data for at least one of these selected variables from the original dataset were omitted to achieve higher model performance. As a result, a new dataset consisting of 10,618 data rows was obtained. Description of the variables is presented in Table 1.

Table 1. Description of the variables

Category	Variables	Description	Unit	Range	Number of samples
Input Variables	T_{air}	Air temperature measured in the occupied zone	ºC	13.4-45.3	10,618
	V_{air}	Air speed in the occupied zone	m/s	0-4.71	
	RH	Relative humidity in the occupied zone	%	14.5-88.8	
	SET	Standard Effective Temperature in Celsius degree	ºC	10.93-38.94	
	CLO	Intrinsic clothing ensemble insulation of the occupant	clo	0.23-2.87	
	MET	Average metabolic rate of the occupant	met	0.7-3.5	
	Age	Age of the occupants	year	16-95	
	Sex	Sex of the occupants		Male, Female	
	T_{out}	Outdoor monthly average temperature when the field study was done	ºC	5.3-38.1	
	Season	Season when the study was done	-	Spring, Summer, Autumn, Winter	
	Building operation mode	Air Conditioned = can be air, radiant, etc. and no operable windows. Naturally Ventilated = no mechanical cooling, but with operable windows. Mixed Mode = mechanical cooling and operable windows (can include concurrent, changeover, or zoned).	-	Air conditioned, Naturally ventilated, Mixed mode	
	Building type	Type of building in which the study was done	-	Classroom, Office, Senior center	
Output Variable	TSV	ASHRAE Thermal sensation vote of occupant	-	from -3 (cold) to +3 (hot)	

To investigate the impact of scale units on the performance, 7-point TSV output was used for generating the wrapped-up 3-point TSV output (Table 2). In order to enhance the predictive performance of the models, min-max normalization is utilized to scale the input data between zero and one due to the presence of varying value ranges.

Table 2. Descriptions of TSV outputs and sample sizes

7-point TSV			3-point TSV		
Categories	Values	Sample size	Categories	Values	Sample size
Cold	(-3) – (-2.6)	40	Cool side	(-0.6) – (-3)	1887
Cool	(-1.6) – (-2.5)	297			
Slightly cool	(-0.6) – (-1.5)	1550			
Neutral	(-0.5) – (+0.5)	5740	Neutral	(-0.5) – (+0.5)	5740
Slightly warm	(+0.6) – (+1.5)	1908	Warm side	(+0.6) – (+3)	2991
Warm	(+1.6) – (+2.5)	751			
Hot	(+3) – (+2.6)	332			

This approach ensures consistency and facilitates effective comparisons across different features, leading to improved model performance. Additionally, it supports compatibility with algorithms such as MNB, which require non-negative input values.

3.2. Principles of ensemble ML frameworks

Ensemble ML frameworks utilize the predictions of multiple models to enhance the accuracy and stability of a base ML algorithm. These base algorithms serve as the fundamental building blocks within the ensemble and can range from decision trees to neural networks. By combining the strengths of different base algorithms, ensembles can achieve improved prediction accuracy, robustness, and generalization, making them a valuable approach in machine learning. In addition, ensemble ML frameworks can tackle the class imbalance issue and provide superb performance in comparison with other solutions [47]. This section introduces the principles of different ensemble ML frameworks.

Bagging (Bootstrap Aggregating) is an ensemble ML framework that combines multiple models, trained on different subsets of the training data, to improve the accuracy and stability of a base ML algorithm [48, 49]. The key idea behind bagging is that by using multiple base models, each trained on a slightly different subset of the training data, ensemble machine learning algorithms help reduce prediction variance and enhance the overall accuracy of the model. The steps for applying bagging ensemble ML framework are as follows: (1) to split the dataset into M random subsets of size n, called bootstrap samples (2) to select the base ML algorithms (i.e decision trees, logistic regression, or neural networks); (3) to train base ML algorithms to make predictions on each bootstrap sample; (4) to combine M base ML algorithms to form an ensemble model by summing up the predictions of each base model.

Boosting is another popular ensemble ML framework that combines multiple weak models to create a stronger model with improved accuracy [50, 51]. The central concept of boosting is to train a series of weak classifiers in succession, with each model correcting the errors of its predecessors, leading to a strong ensemble model, and, thus, the ensemble model can learn more about the difficult examples and improve its overall accuracy. The steps for applying boosting ensemble ML framework are as follows: (1) to train the entire dataset by a basic ML algorithm; (2) to identify the examples where the base ML algorithm fails and to assign a higher weight to examples that were previously misclassified; (3) to train the new ML model on the updated dataset; (4) to repeat steps 2 and 3 for a fixed voting of iterations or until the performance of the ensemble ML model plateaus; (5) to combine the predictions of all the models to form the final ensemble model.

Stacking, or stacked generalization, is an advanced ensemble learning framework that improves predictive accuracy by combining the outputs of multiple base models through a meta-learner [52, 53]. Stacking operates on the principle of training a meta-learner that uses the outputs of multiple base models as input features to enhance overall predictive performance and, thus, combines the strengths of the individual base ML algorithm and makes the final predictions. The steps for applying stacking ensemble ML framework are as follows: (1) to split the dataset into a training set and a holdout set; (2) to select multiple base ML algorithms; (3) to train base ML algorithms on the training set; (4) to make predictions on the holdout set using the trained base ML algorithms and to use these predictions as input to the meta-model; (5) to train the meta-model on the holdout set, using predictions of the base models as input and the true labels as output; (6) to combine the predictions of the base ML algorithms and the meta-model to form the final ensemble model.

Voting is a basic ensemble ML technique that aggregates the predictions of multiple models using a predefined voting rule to determine the final output [54, 55]. The main concept underlying voting is that leveraging the collective predictions of various models can lead to improved accuracy and reduced variability in outcomes. The voting rule can be either a majority vote (for classification

problems) or an average (for regression problems) of the individual model predictions. In the voting of classification problems, two types of voting schemes exist: hard voting and soft voting. The former involves selecting the prediction with the highest count of votes, while the latter entails aggregating the individual probabilities of each prediction across models, followed by the selection of the prediction with the highest overall probability. The steps for applying voting ensemble ML framework are as follows: (1) to split the dataset into a training set and a holdout set; (2) to select the base ML algorithms (i.e decision trees, logistic regression, or neural networks); (3) to train base ML algorithms to make predictions on the holdout set; (4) to combine the predictions of the base ML algorithms using a voting rule to form the final ensemble model.

3.3. Principles of advanced ensemble ML algorithms

In this study, RF, ROF, XT, GB, HGB, and XGB are selected as advanced ensemble ML algorithms since they have been proven to be effective in many applications and have a track record of success [56–58].

RF is a powerful and flexible algorithm that can be used for both regression and classification tasks. It is based on the idea of decision trees which are prone to overfitting. RF overcomes this problem by constructing an ensemble of decision trees and using bagging to reduce the variance of each tree. Subsequently, RF combines the predictions of all the trees, either by averaging them or by taking the majority vote, to make predictions on new data. To further reduce the variance of the trees, RF also introduces a randomization step when constructing each tree. At each node of the tree, instead of using the best split among all possible splits, RF considers only a random subset of the features. This ensures that the trees in the ensemble are diverse and uncorrelated.

ROF combines multiple models to improve the predictive performance of a single model. It is particularly useful when the data contains a large number of features that may be correlated or redundant [59]. The key idea behind ROF is to randomly rotate the feature space before training each model in the ensemble. This means that instead of using the original features as input to the models, ROF first applies a random rotation to the feature space, and then uses the rotated features as input. By doing so, ROF creates a new set of features that are uncorrelated and potentially more informative than the original features. ROF uses a weighted averaging method, where the predictions of each model are weighted by the performance of the model on a validation set. This ensures that the models with the highest accuracy contribute more to the final prediction, while the models with lower accuracy contribute less.

XT (Extremely Randomized Trees) is also based on decision trees and is particularly useful when the data is noisy or the number of features is large, as it can reduce the risk of overfitting and improve the robustness of the model. In XT, the trees are constructed using a variant of the decision tree algorithm, where the splits are selected by a random threshold that is selected uniformly at random within the range of the feature values and not based on a quality criterion. By introducing this additional randomness, XT aims to create more diverse and less correlated decision trees, which can improve the performance of the ensemble. To make predictions on new data, XT combines the predictions of all the trees in the ensemble. XT uses a simple averaging method, where the predictions of all trees are averaged to obtain the final prediction.

GB is based on boosting and uses a series of weak models, typically decision trees, to form a strong model [60]. It is particularly useful when the data contains complex non-linear relationships, as it can capture these relationships through the ensemble of weak models [61]. The core idea of GB is to iteratively add weak learners to the ensemble, each one trained to minimize the residual errors of the overall model. In each iteration, the algorithm fits a new weak model to the residual errors of the current model. To prevent overfitting, GB uses a technique called "shrinkage", where the predictions of each weak model are multiplied by a small

learning rate before adding them to the overall model. This means that each weak model contributes only a small amount to the overall model, and the learning rate controls the trade-off between model complexity and accuracy. To make predictions on new data, GB combines the predictions of all weak models in the ensemble, weighted by the learning rate and the performance of each model on a validation set. By aggregating the weighted outputs of all weak models, the final prediction serves as the optimal approximation of the target variable given the input data.

HGB, also known as Hist Gradient Boosting, is particularly useful when the data contains many continuous features. The key idea behind HGB is to discretize the continuous features into bins and build histograms of these bins, which can be used to efficiently calculate gradients and Hessians for each feature. To construct each weak model in the ensemble, HGB uses a greedy algorithm that selects the best split point for each histogram. The split point is chosen to maximize the reduction in the loss function, which is typically the mean squared error for regression tasks or the log-loss for classification tasks. This allows HGB to construct accurate weak models that capture the complex non-linear relationships between the target variable and the features and. To prevent overfitting, HGB uses several regularization techniques, including maximum depth constraints, minimum number of samples per leaf, and learning rate. The maximum depth and the minimum number of samples per leaf serve as constraints that regulate the complexity of individual weak learners, while the learning rate determines the extent to which each weak model contributes to the final ensemble. Similar to GB, HGB generates predictions for new data by aggregating the outputs of all weak learners, with each contribution weighted by both the learning rate and the model's performance on a validation set. The final prediction is the sum of all weighted predictions, which represents the best possible approximation of the target variable given the data and the ensemble of weak models.

XGB uses decision trees and is particularly useful when the data is high-dimensional, sparse, or noisy. The key idea behind XGB aligns with other gradient boosting algorithms: it iteratively adds weak learners to the ensemble, each one aimed at minimizing the overall prediction error. To construct each weak model in the ensemble, XGB uses several techniques such as parallelization, regularization, feature importance and missing value handling to improve the performance and scalability of the algorithm. In addition, split finding, approximate computing, and weighted quantile sketch techniques are used to improve the accuracy of the decision trees. To make predictions on new data, similar to GB and HGB, XGB also combines the predictions of all weak models in the ensemble, weighted by the learning rate and the performance of each model on a validation set. The final prediction is the sum of all weighted predictions, which represents the best possible approximation of the target variable given the data and the ensemble of weak models.

3.4. Performance metrics

Performance metrics are essential for evaluating the generalization ability of a developed model. While accuracy remains the most frequently reported metric in thermal comfort studies [62], it is insufficient for unbalanced multi-class problems and may even produce erroneous results [23]. Therefore, several performance metrics must be taken into account for evaluation. In this study, four performance metrics—accuracy, precision, recall, and F1-score—are employed to evaluate the effectiveness of the proposed methods. It should be noted that Mean Absolute Percentage Error (MAPE) is more appropriate for regression models, and, thus, it was not included in this study that focuses on classification-based TSV prediction. In the aforementioned metrics, True Positive (TP) refers to a case where the model correctly predicts the presence of the target class. True Negative (TN) indicates a correct prediction of the absence of the target class. False Positive (FP) occurs when the model incorrectly predicts the presence of the target class, while False Negative (FN) refers to an incorrect prediction where the model fails to detect the presence of the target class.

Accuracy, defined as the ratio of correctly classified samples to the total number of samples, is a widely used performance metric. However, in cases of imbalanced data, it may fail to reflect the true performance of the model. Its formal expression is provided in Eq. (1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision, or positive predictive value, measures how many of the instances predicted as positive are actually true positives. It reflects the model's ability to prevent false positive classifications. The formal expression of precision is given in Eq. (2)

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall, also referred to as sensitivity, is the proportion of true positive predictions among all actual positive instances. In other words, it reflects the model's ability to correctly identify all positive cases. The formal definition of recall is presented in Eq. (3).

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

F1 score combines precision and recall using their harmonic mean, offering a single metric that balances both concerns. It is particularly useful in scenarios involving imbalanced data, where positive classes are more critical. The equation for the F1 score is shown in Eq. (4).

$$F1\ score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (4)$$

For the metrics accuracy, precision, recall, and F1, a value closer to 1 indicates strong predictive performance, while a value closer to 0 indicates poor predictive performance.

4. Results

4.1. Experimental setting

During the experimental study, all the methods were implemented in Python 3.1 using libraries Scikit-learn (version 1.1.2), rotation-forest (version 1.0), and xgboost (version 1.7.4). If it is not explicitly specified, default parameters in the software libraries were used. To customize parameters, the 7-point TSV case was selected to conduct parameter analysis since the results obtained for this case could also be representative for the 3-point case. Default values of parameters were changed to specific ones after conducting preliminary experiments. As a result, the default values of following parameters were changed (Fig. 2):

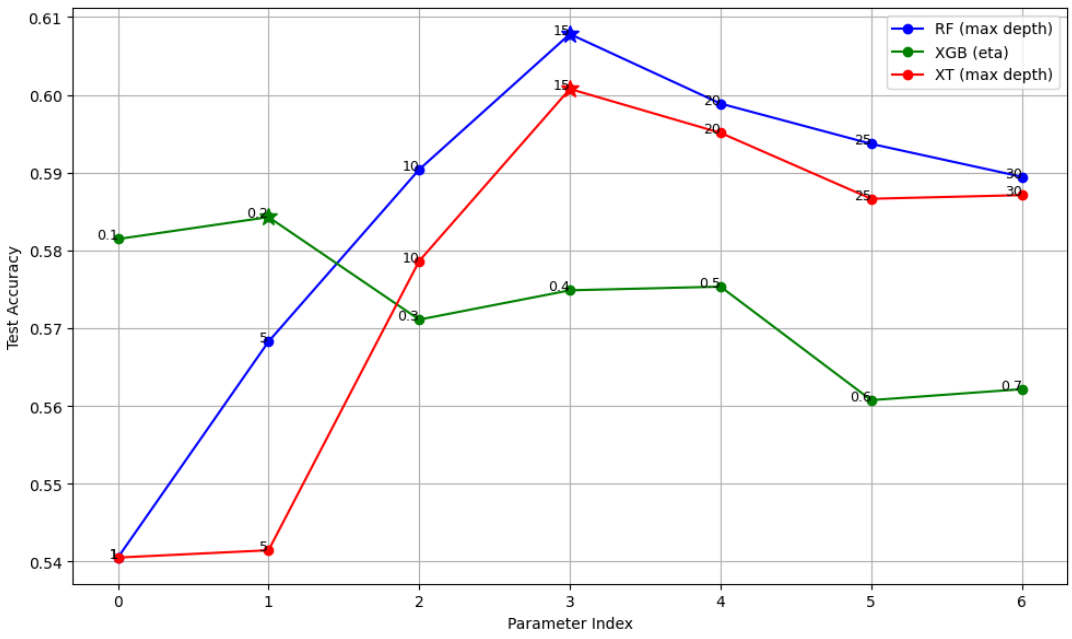

Fig. 2. Parameter analysis of RF, XT, and XGB for 7-point TSV prediction

- Maximum depth parameter of the RF algorithm was changed to 15.
- Maximum depth parameter of the XT algorithm was changed to 15.
- Eta parameter of the XGB algorithm was changed to 0.2.
- To prevent convergence issues in the Multi-layer Perceptron and Logistic Regression algorithms, the max iteration parameter was changed to 1000 for both algorithms.

To ensure the replicability of the findings, the main hyper-parameters used in this study is listed in Table 3. Data were split into training (80%) and testing (20%) using the stratified strategy to maintain the class ratio in this imbalanced dataset. As supported by previous studies [32, 45, 63–65], the 80/20 train-test split ratio provides an effective balance between training and evaluation, offering enough data to build the model while preserving a meaningful amount for testing.

Precision, recall, and F1 score can originally be applied on binary classification problems. However, since TSV prediction in this work includes multiple target classes, a weighted average approach was applied to calculate these metrics. Weights are determined according to the support, or the number of test samples of each class label.

4.2. Results of basic ML algorithms

In this section, the performance of 6 basic ML classification algorithms (LR, KNN, SVM, DT, MLP, and MNB) were measured.

Fig. 3 presents different performance metrics calculated for 3-point and 7-point TSV prediction. The results show that the classification for 3-point TSV generally achieved higher performance metric scores than the 7-point classification for TSV. The F1 score ranges of 3-point and 7-point TSV prediction–are 0.379 - 0.614–and 0.379 - 0.516, respectively. Based on all performance metrics except for F1 score, MLP outperforms other basic ML algorithms for both 3-point and 7-point TSV prediction whereas KNN performs better than other ML algorithms according to F1 score. Based on all performance metrics, MNB algorithm performs significantly worse than the other algorithms for 3-point TSV prediction. For 7-point TSV prediction, MNB is the worst performing algorithm according to F1 and precision performance metrics whereas DT is the worst performing algorithm according to accuracy and recall performance metrics.

Table 3. Main hyper-parameters used in the study

Classifier	C	Max iteration	Kernel	Max depth	n_Estimators	n_Neighbors	eta	Final Estimator
LR	1	1000	-	-	-	-	-	-
KNN	-	-	-	-	-	5	-	-
SVM	1	-	RBF	-	-	-	-	-
DT	-	-	-	None	-	-	-	-
MLP	-	1000	-	-	-	-	-	-
MNB	-	-	-	-	-	-	-	-
Bagging	-	-	-	-	10	-	-	-
Boosting (AdaBoost)	-	-	-	-	50	-	-	-
Stacking	-	-	-	-	-	-	-	LR
RF	-	-	-	15	100	-	-	-
ROF	-	-	-	4	10	-	-	-
XT	-	-	-	15	100	-	-	-
GB	-	-	-	3	100	-	-	-
HGB	-	100	-	-	-	-	-	-
XGB	-	-	-	-	100	-	0.2	-

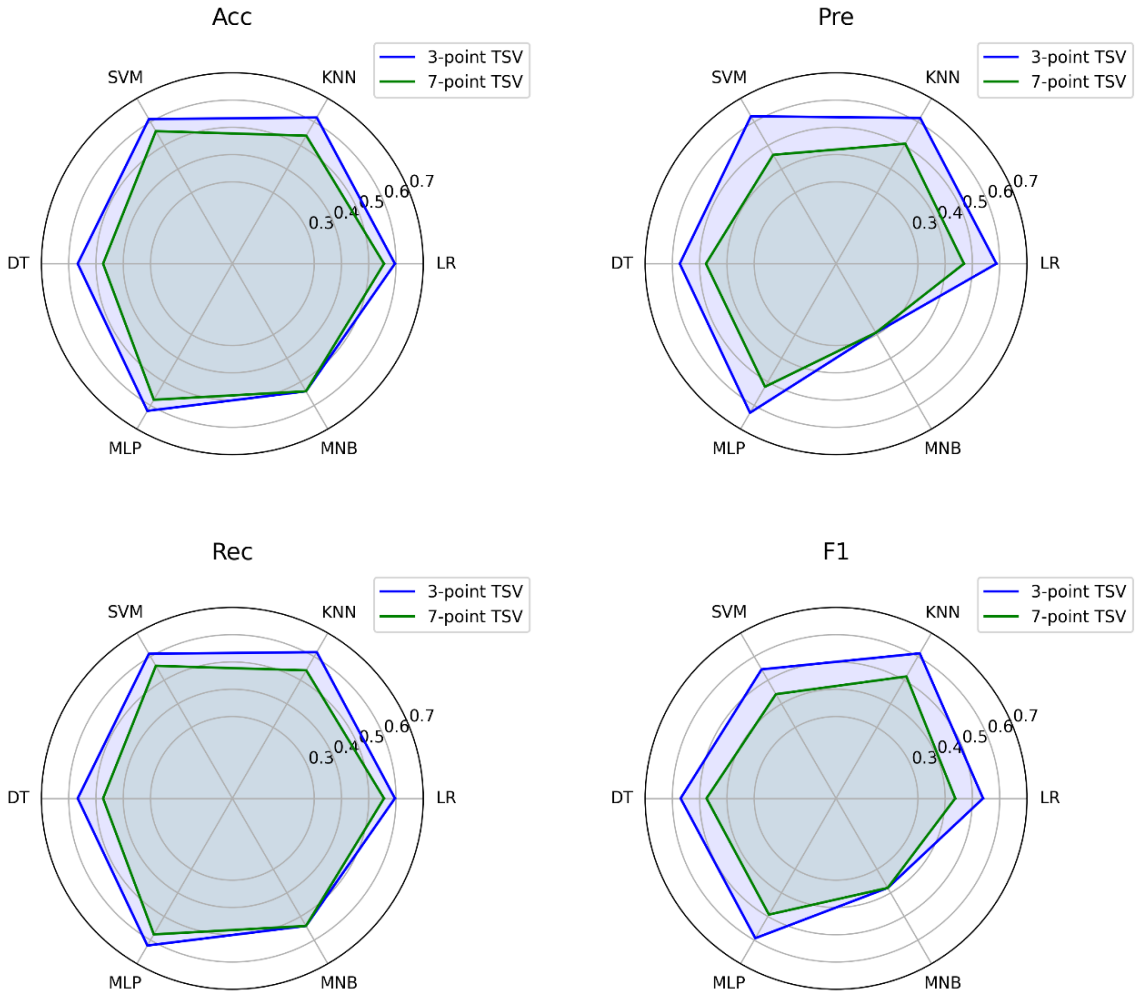

Fig. 3. Performance of basic ML algorithms

4.3. Results of ensemble ML algorithms

4.3.1. Bagging ensemble ML algorithms

Fig. 4 demonstrates the results of using basic ML algorithms as base estimators in bagging ensemble framework. The results show that the utilization of the bagging ensemble method substantially enhanced the performance of the DT algorithm more than other ML algorithms for both 3-point and 7-point TSV prediction. For example, F1 scores of the individual application of DT for 3-point was 0.570, which increased to 0.636 for the bagging ensemble DT algorithm. In addition, F1 scores of the individual application of DT for 7-point was 0.475, which increased to 0.549 for the bagging ensemble DT algorithm. The rest of the bagging applications of ML algorithms performed similar to that of their respective individual applications.

4.3.2. Boosting ensemble ML algorithms

This section provides the results of boosting ensemble framework using AdaBoost algorithm. Since sample weighting support is necessary for base estimators, only 4 of the basic ML algorithms, namely LR, SVM, DT, and MNB, were used. As it is shown in Fig. 5, the boosting ensemble DT algorithm for the 3-point TSV prediction achieves the best performance in terms of all the performance metrics, with a maximum F1 score of 0.565.

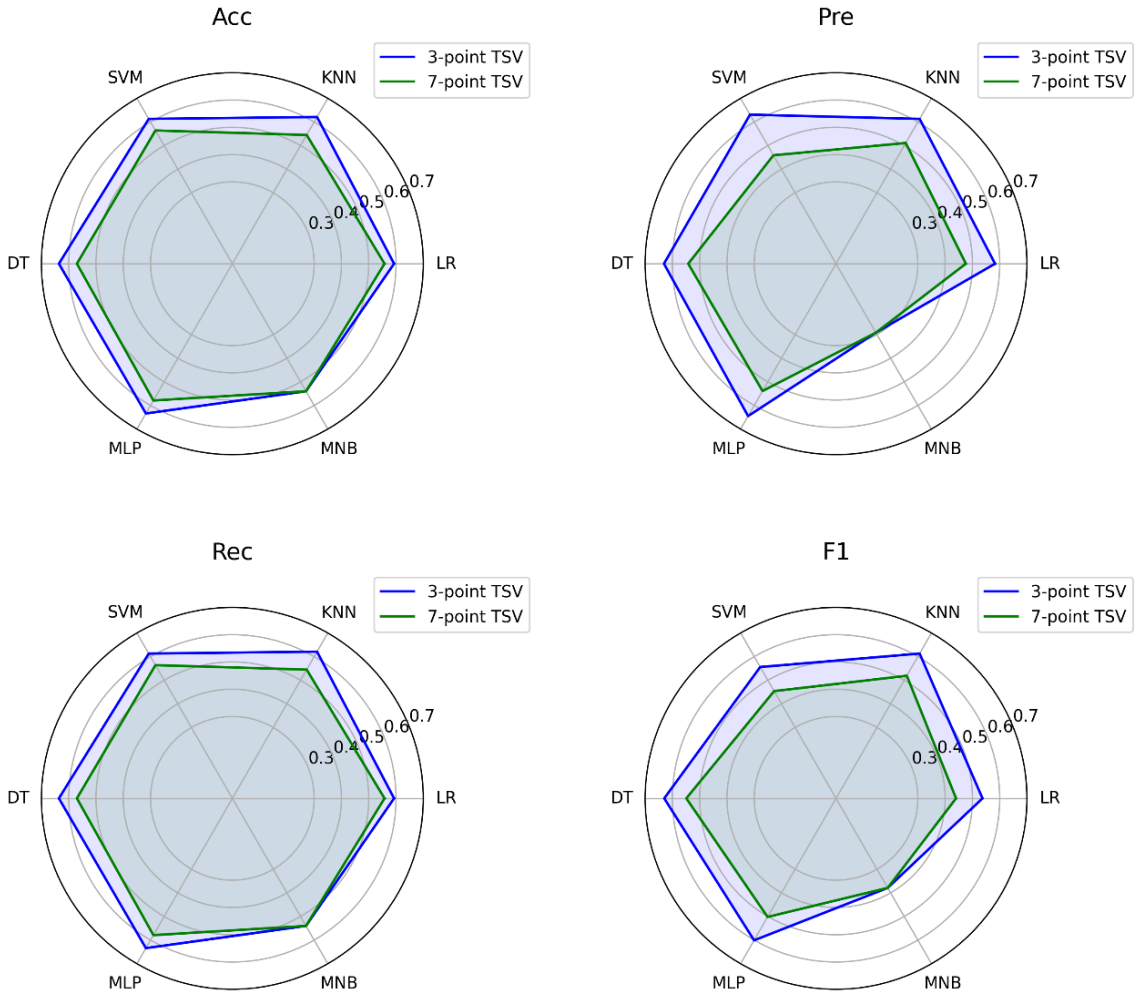

Fig. 4. Performance of bagging ensemble ML algorithms

However, a significant performance loss was observed in all the boosting ensemble algorithms except for boosting ensemble MNB algorithm for the 7-point TSV prediction compared to their respective results obtained for basic ML algorithms. Although the SVM algorithm may not be the top-performing choice in a boosting framework, it demonstrates better preservation of metrics when transitioning from a 3-point to a 7-point prediction scale unit.

4.3.3. Stacking ensemble ML algorithms

In stacking ensemble design, DT was excluded since stacking should take advantage of strong predictors unlike bagging and boosting ML frameworks in which weak learners such as DT are preferable. As a result, the combinations of ML algorithms presented in Table 4 are created for evaluating the performance of stacking ensembles. It should be noted that algorithms were grouped into combinations of 3 and 5 in order to avoid potential ties when making predictions or voting in the ensemble.

The results of stacking ensembles are shown in Fig. 6. Although the methods show similar performance, the combinations of C1, C2, C3, C7, C8, C9, C11 are generally more successful than other combinations for both 3-point and 7-point TSV prediction. Overall performance of stacking ensemble algorithms for 3-point TSV prediction is higher than that of 7-point TSV prediction since 3-point TSV is easier to be predicted.

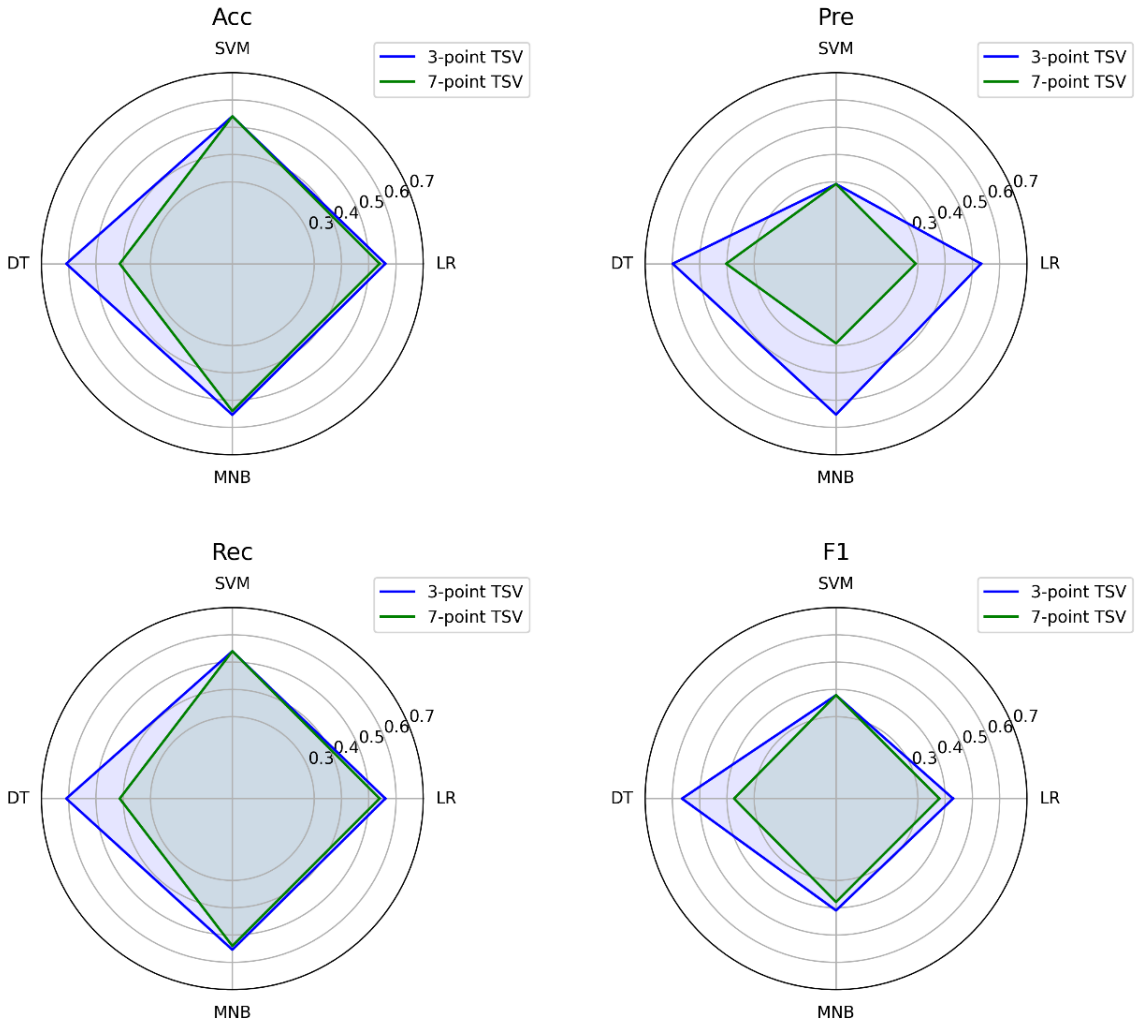

Fig. 5. Performance of boosting ensemble ML algorithms

Table 4. Different combinations of basic ML algorithms that construct stacking ensembles

Combination	Ensemble
C1	LR, KNN, SVM
C2	LR, KNN, MLP
C3	LR, KNN, MNB
C4	LR, SVM, MLP
C5	LR, SVM, MNB
C6	LR, MLP, MNB
C7	KNN, SVM, MLP
C8	KNN, SVM, MNB
C9	KNN, MLP, MNB
C10	SVM, MLP, MNB
C11	LR, MLP, MNB, KNN, SVM

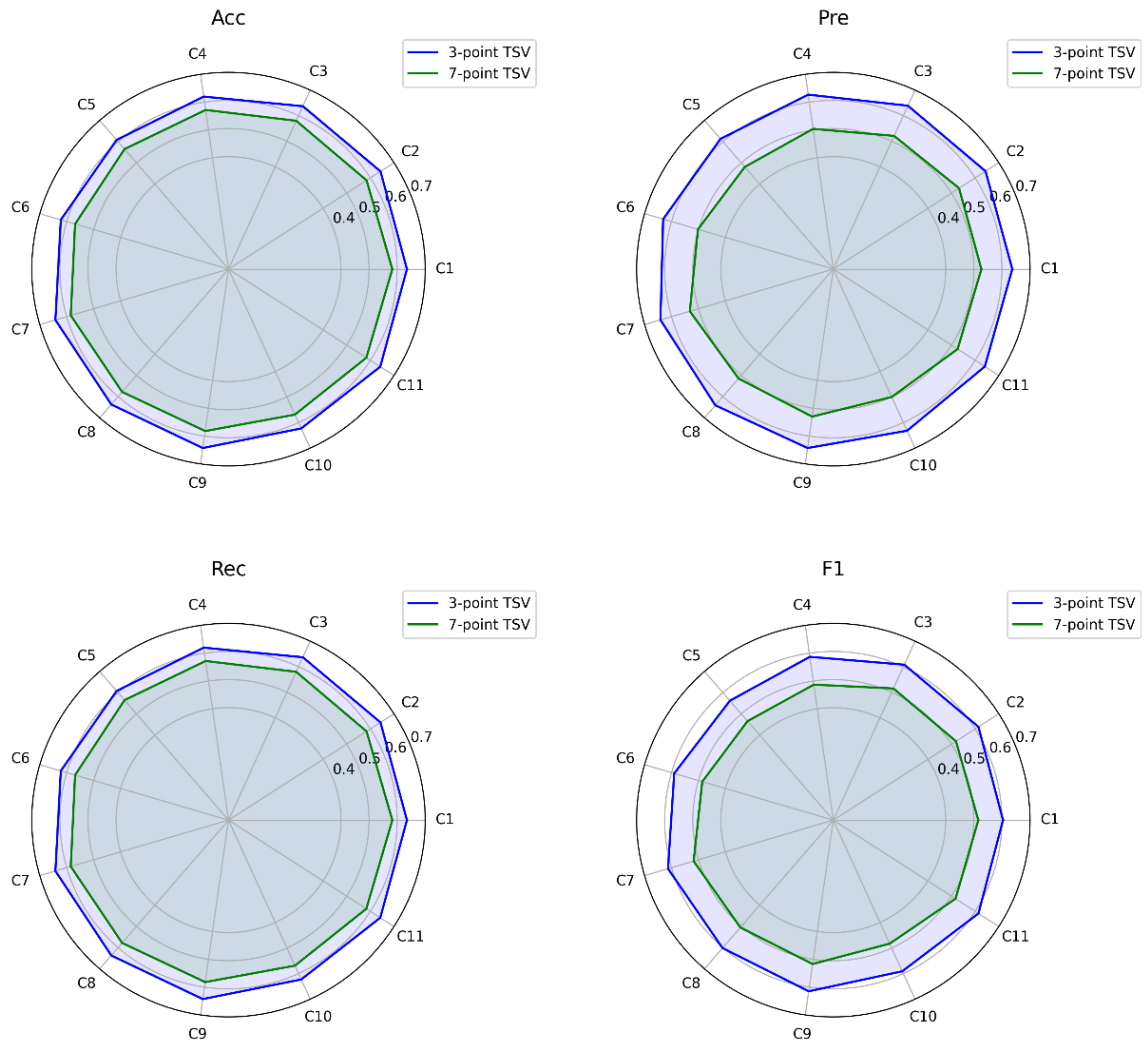

Fig. 6. Performance of stacking ensemble ML algorithms

It can also be observed that the impact of changing TSV scale units on performance metrics is more significant for precision and recall metrics compared to accuracy and F1 metrics.

4.3.4. Voting ensemble ML algorithms

In a voting ensemble framework, each of the well-performing constituent algorithms makes a prediction on the output classes by combining their individual predictions. Therefore, like the approach followed in Section 3.3.2, DT was excluded in voting ensemble design and combinations listed in Table 4 were used again for evaluating the performance.

Fig. 7 shows the performance of hard voting framework, in which the majority rule is applied and the output class with highest number of predictions is selected. In terms of F1 score, C2, C7, and C9 combinations perform better than the other combinations for both 3-point and 7-point TSV predictions. Overall performance of voting ensemble ML algorithms for 3-point TSV prediction is higher than that of 7-point TSV prediction. However, the performance loss in terms of precision when transitioning from 3-point to 7-point TSV prediction is relatively low for the C2, C7, and C9 algorithms compared to other combinations.

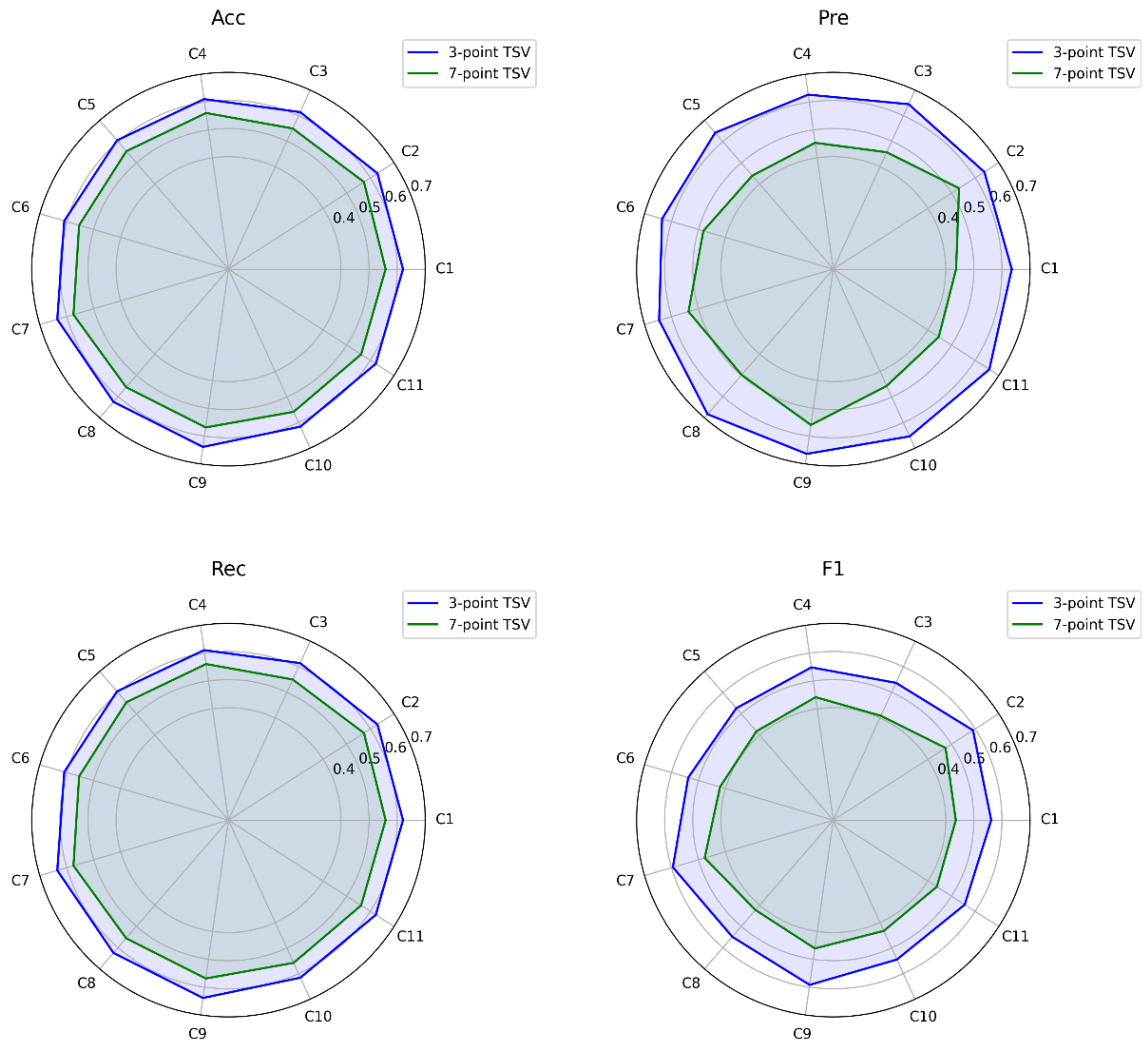

Fig. 7. Performance of hard voting ensemble ML algorithms

Fig. 8 demonstrates the performance of soft voting framework in which the output class with the highest total prediction probability is selected. Similar to hard voting, C2, C7, and C9 outperforms others in terms of F1 score for both 3-point and 7-point TSV predictions. Especially for 7-point TSV prediction, overall performance of soft voting framework is better than that of hard voting in terms of F1 and precision metrics.

4.4. Results of advanced ensemble algorithms

This section measures the performance of advanced ensemble algorithms including RF, ROF, XT, GB, HGB, and XGB. The results in Fig. 9 suggest that the overall performance of advanced ensemble ML algorithms are superior to previous ensemble frameworks presented in this study. Among the algorithms, RF, XT and XGB show the best overall performance considering both 3-point and 7-point TSV prediction. Across all performance metrics, RF achieved the highest accuracy in 7-point TSV prediction, outperforming other advanced ensemble models. However, different performance metrics yielded different results for the 3-point TSV prediction, in which XT is the best performing algorithm based on accuracy (0.66), precision (0.665) and recall (0.66) scores whereas HGB is the best performing one according to the F1 score (0.638).

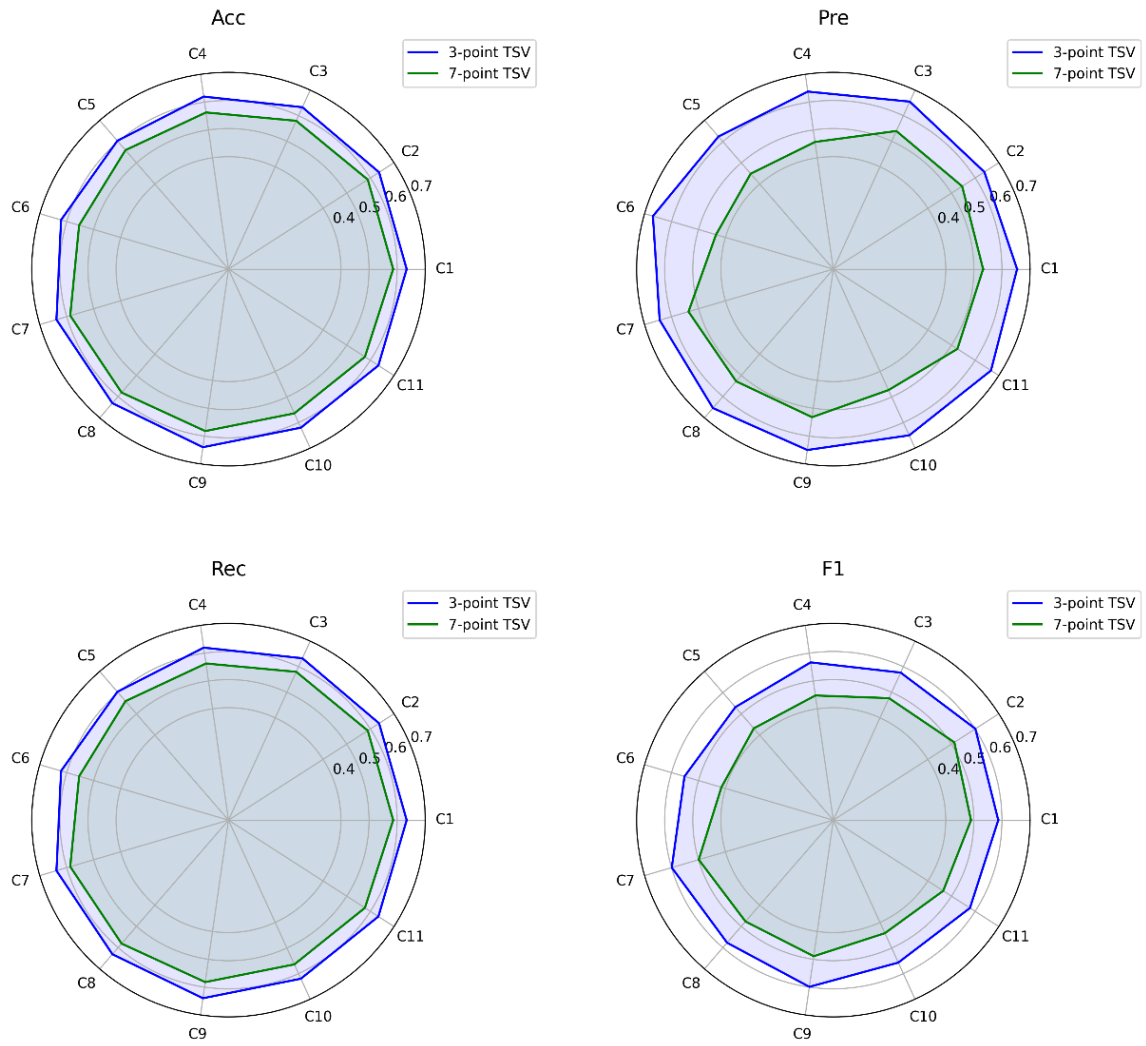

Fig. 8. Performance of soft voting ensemble ML algorithms

4.5. Further analysis of the results

This section provides further analysis of findings obtained from ML algorithms that were experimented with in this study. Although 4 different performance metrics are used, F1 score is selected as the primary performance metric for discussing the results since F1 score places equal emphasis on precision and recall, making it a good approach to evaluating results, especially when interpreting imbalanced data where accuracy can be misleading [66].

Table 5 and Table 6 present the best algorithms in each experiment group and list their performance based on F1 score for 3-point and 7-point TSV predictions, respectively. Note that weighted calculation in multiclass classification to deal with the imbalance can result in F1 scores that are not between precision and recall.

Tables 5 and 6 suggest that advanced and bagging are the two prominent ensemble categories, achieving higher F1 scores than the others. Besides, the advanced ensemble ML algorithm algorithms, namely HGB and RF, outperform the bagging framework when all performance metrics are considered. Another notable finding from the results is that DT should be prioritized when designing a bagging ensemble algorithm.

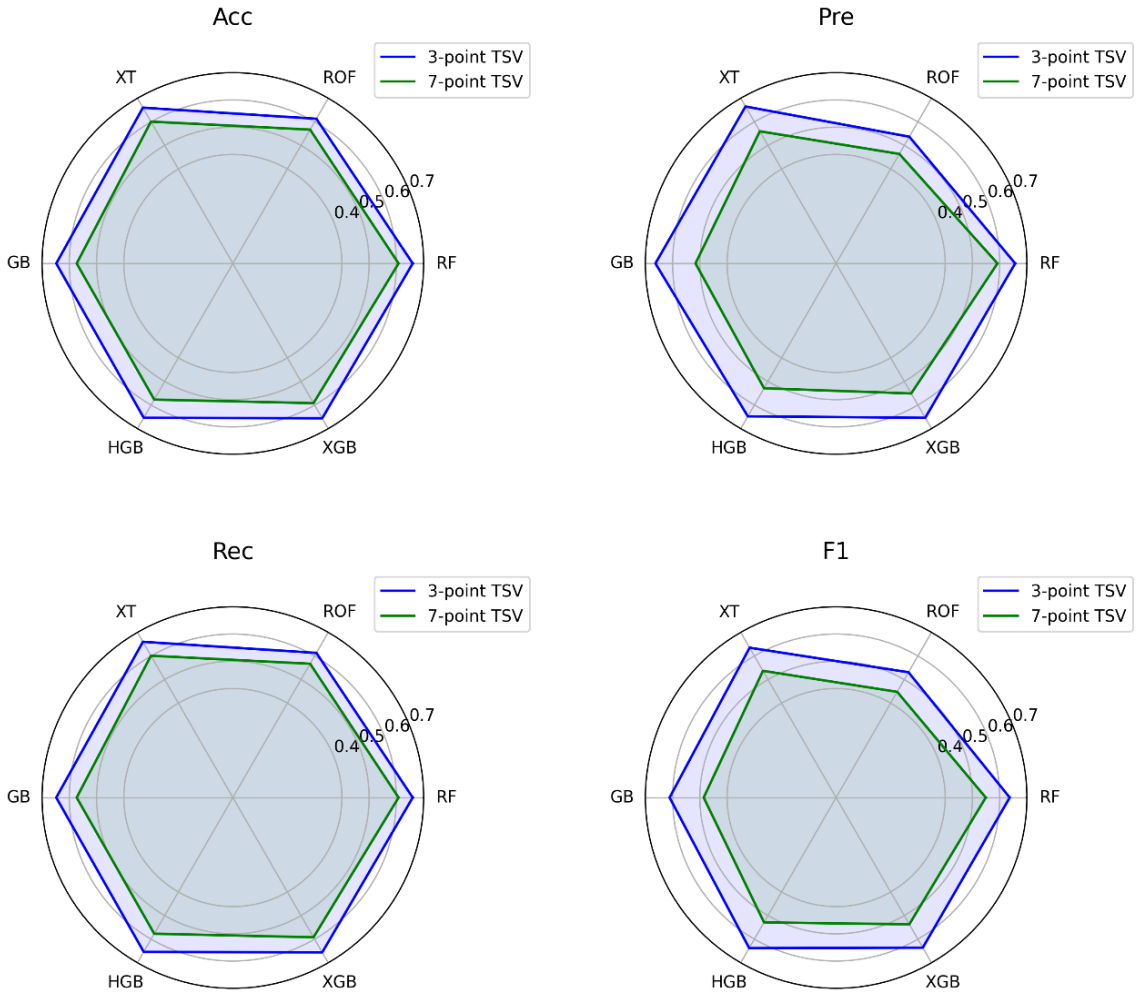

Fig. 9. Performance of advanced ensemble ML algorithms

Table 5. Best performing ML algorithms for 3-point TSV

Category	Best performing algorithm	Accuracy	Precision	Recall	F1
Basic ML	KNN	0.619	0.617	0.619	0.614
Bagging	DT	0.636	0.631	0.636	0.630
Boosting (Adaboost)	DT	0.609	0.599	0.609	0.565
Stacking	C9 (KNN, MLP, MNB)	0.643	0.643	0.643	0.615
Voting (hard)	C7 (KNN, SVM, MLP)	0.635	0.646	0.635	0.595
Voting (soft)	C2 (LR, KNN, MLP)	0.637	0.639	0.637	0.601
Advanced	HGB	0.654	0.648	0.654	0.638

DTs can split the data based on the most informative features, which can help to separate the minority class from the majority class. However, it may be prone to overfitting if the training data is biased or unrepresentative of the population. In the context of bagging, which involves training multiple models, the problem of overfitting can be mitigated, and the decision tree algorithm can benefit from splitting on the most informative features. On the other hand, the results show that the boosting framework (Adaboost) exhibits relatively poor performance on F1 metric.

Table 6. Best performing ML algorithms for 7-point TSV

Category	Best performing algorithm	Accuracy	Precision	Recall	F1
Basic ML	KNN	0.542	0.508	0.542	0.516
Bagging	DT	0.570	0.542	0.570	0.549
Boosting (Adaboost)	LR	0.540	0.292	0.540	0.379
Stacking	C2 (LR, KNN, MLP)	0.584	0.532	0.584	0.519
Voting (hard)	C7 (KNN, SVM, MLP)	0.575	0.537	0.575	0.477
Voting (soft)	C2 (LR, KNN, MLP)	0.589	0.545	0.589	0.512
Advanced	RF	0.608	0.592	0.608	0.549

This may be because the minority class is extremely underrepresented in the dataset. This can result in lower weight being assigned to the minority class instances, making it more difficult for the algorithm to learn their characteristics and leading to poorer performance on the minority class. There is no notable difference between stacking framework and basic ML in terms of F1 scores. Therefore, it can be said that stacking is not suitable when dealing with this kind of imbalanced dataset.

Regarding voting frameworks, soft voting framework achieves better than the hard voting. In TSV prediction, taking into account the probabilities predicted by each model can lead to a more accurate classification. Soft voting can provide more information to the ensemble by considering the confidence levels of each model's predictions, which can improve the overall performance of the ensemble. Nevertheless, it is evident from the results that voting schemes in general did not show good performance in this study.

Among basic ML algorithms, KNN shows the best F1 scores for both 3- and 7-point TSV classifications. When ensemble approaches are compared to basic ML on this dataset, it appears that the accuracy, precision, and recall scores of the former category are better than those of the latter. On the other hand, in terms of F1 scores, which provides more accurate evaluation on imbalanced data, only the bagging and advanced ensemble outperform basic ML. It seems that, as a non-parametric algorithm, KNN makes no assumptions about the underlying distribution of the data, and, therefore, can be effective in detecting patterns in the minority class compared to boosting and voting. This is because the minority class may be ignored by the voting and boosting ensembles in this kind of imbalanced dataset.

To further compare the F1 score performance of the algorithms and obtain more accurate results, each of the selected algorithms were run using stratified 10-fold cross validation (CV) technique on the whole dataset. After collecting 10 different F1 scores per algorithm, the results with box-plots are shown in Fig. 10, where the abbreviations Bag, Bst, Stk, HV and SV were used for bagging, boosting, stacking, hard voting and soft voting, respectively. For both 3- and 7-point TSV prediction cases, the results from box plots suggest that advanced (HGB and RF) and bagging (DT) categories outperform others. Boosting, stacking, and voting ensemble frameworks cannot exceed the performance of basic ML algorithms such as KNN.

To investigate the differences between each pair of algorithms, the Wilcoxon rank sum test was applied to the data obtained from the 10-fold cross-validation experiment. Table 7 and Table 8 list the results of the test in terms of p-values. Wilcoxon rank sum, or Mann-Whitney U, is a non-parametric statistical test that is used to compare two independent samples to determine if they come from the same population or not. In this study, populations are the F1 scores obtained per algorithm.

The null hypothesis indicates that the performance of two algorithms is the same. The threshold for statistical significance is set to 0.05. If the calculated p-value is less than 0.05, the null hypothesis is rejected and that there is evidence of a significant difference between the two algorithms.

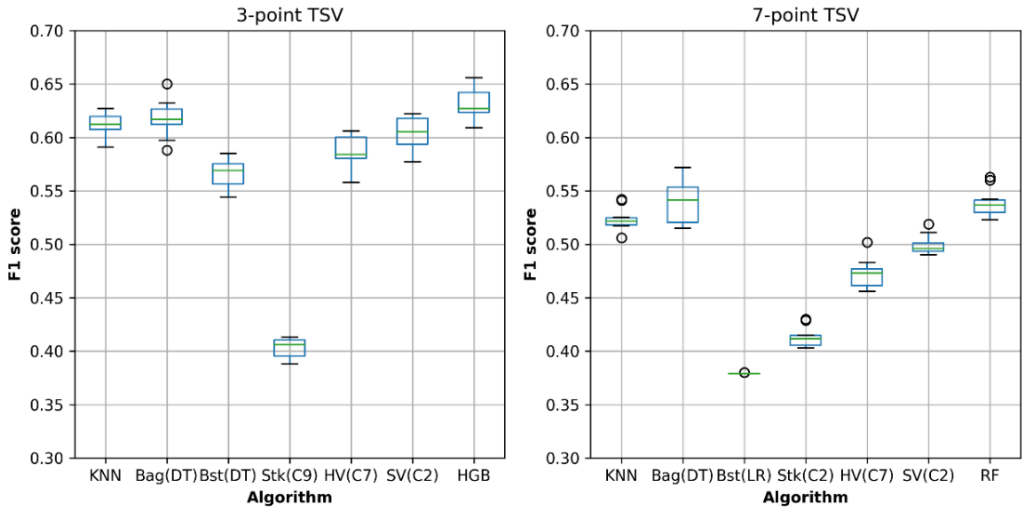

Fig. 10. Box-plots of selected algorithms on 10-fold CV

Table 7. Results (p-values) of Wilcoxon rank sum test for 3-point TSV

	KNN	Bag(DT)	Bst(DT)	Stk(C9)	HV(C7)	SV(C2)	HGB
KNN	-	4.06E-01	*1.57E-04*	*1.57E-04*	*8.81E-04*	2.26E-01	*7.28E-03*
Bag(DT)	-	-	*1.57E-04*	*1.57E-04*	*1.15E-03*	1.12E-01	1.31E-01
Bst(DT)	-	-	-	*1.57E-04*	*1.40E-02*	*3.81E-04*	*1.57E-04*
Stk(C9)	-	-	-	-	*1.57E-04*	*1.57E-04*	*1.57E-04*
HV(C7)	-	-	-	-	-	*4.94E-02*	*1.57E-04*
SV(C2)	-	-	-	-	-	-	*1.50E-03*
HGB	-	-	-	-	-	-	-

Italic numbers indicate that the result is statistically significant at the 0.05 level

Table 8. Results (p-values) of Wilcoxon rank sum test for 7-point TSV

	KNN	Bag(DT)	Bst(LR)	Stk(C2)	HV(C7)	SV(C2)	RF
KNN	-	1.51E-01	*1.57E-04*	*1.57E-04*	*1.57E-04*	*5.83E-04*	*1.26E-02*
Bag(DT)	-	-	*1.57E-04*	*1.57E-04*	*1.57E-04*	*3.81E-04*	9.10E-01
Bst(LR)	-	-	-	*1.57E-04*	*1.57E-04*	*1.57E-04*	*1.57E-04*
Stk(C2)	-	-	-	-	*1.57E-04*	*1.57E-04*	*1.57E-04*
HV(C7)	-	-	-	-	-	*1.31E-03*	*1.57E-04*
SV(C2)	-	-	-	-	-	-	*1.57E-04*
RF	-	-	-	-	-	-	-

Italic numbers indicate that the result is statistically significant at the 0.05 level

The results indicate that the superiority of the HGB algorithm for 3-point TSV is statistically significant, except for the bagging ensemble approach. Regarding the 7-point TSV, while the bagging ensemble exhibited a higher average F1 score compared to the basic ML approach, this difference was not statistically significant according to Table 7. However, RF stands out as the algorithm that differs from the others, except for bagging.

To gain further insight into these performance differences across classes, the confusion matrices and class distributions associated with both classification tasks were examined. The training

and testing confusion matrices for the 3-point and 7-point TSV classification tasks are presented in Figs. 11 and 12, respectively. Label 0 indicates the coldest thermal sensation, with higher labels corresponding to progressively warmer sensations; the label descriptions and class distributions are presented in Table 2.

In the 3-point classification, the model achieves strong performance across all classes, particularly on the dominant neutral class. Misclassifications primarily occur between neighboring categories (e.g., cool side and neutral), which is acceptable given the subjective and continuous nature of thermal sensation. The test set performance shows a similar trend, indicating good generalization. For the 7-point classification, the model successfully predicts major classes like neutral and slightly warm, while performance decreases on rare extremes such as cold and hot. Errors are largely between adjacent categories (e.g., slightly cool and neutral), suggesting that the model preserves ordinal consistency even in a more complex setting.

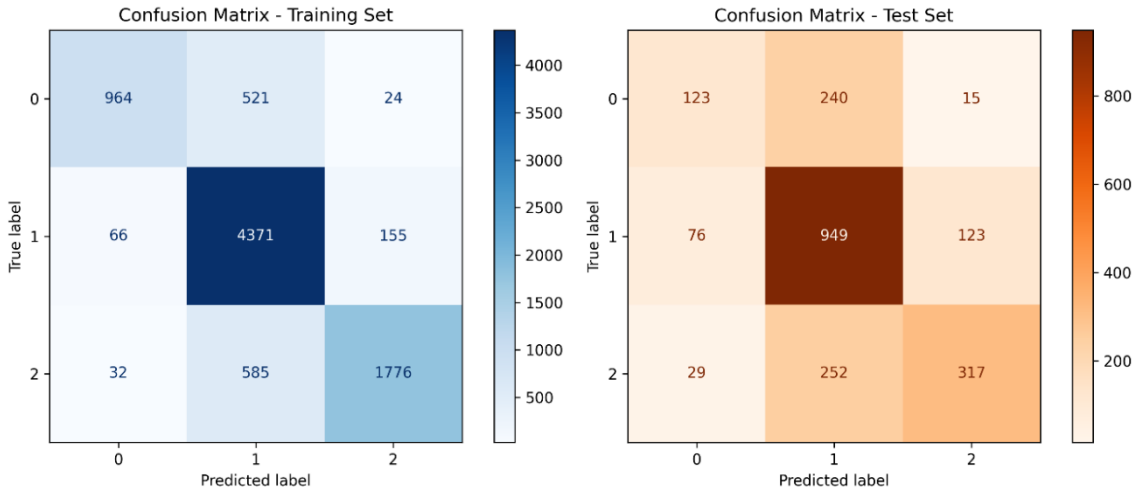

Fig. 11. Confusion matrices for 3-point TSV prediction

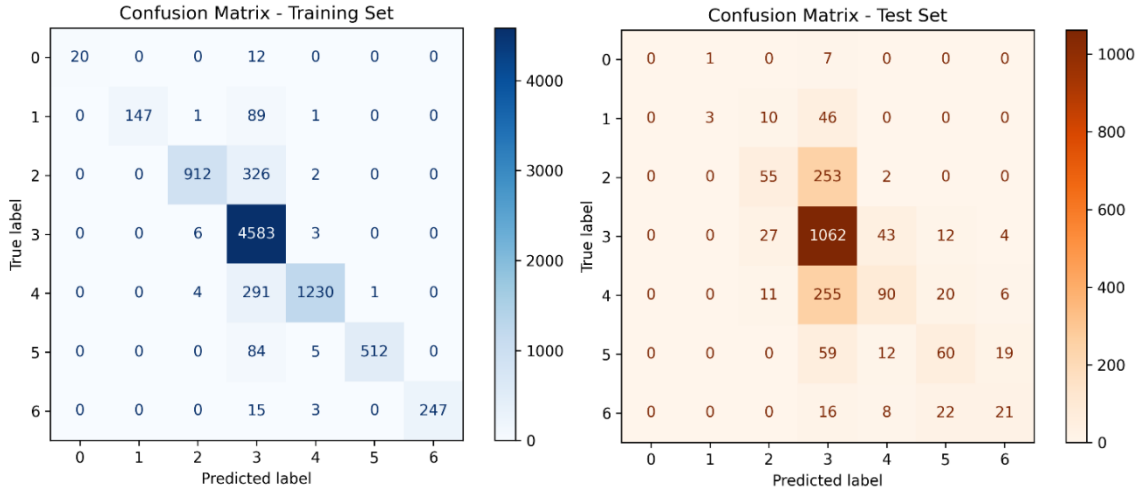

Fig. 12. Confusion matrices for 7-point TSV prediction

These results also reflect the class imbalance present in the dataset, which leads the model to favor frequently occurring classes. Nonetheless, most misclassifications occur between adjacent categories, indicating that the model preserves the ordinal structure of thermal sensation. Although accuracy varies across classes (particularly in the more challenging 7-point classification) the model demonstrates consistent generalization and structure-aware behavior. This suggests it serves as a strong baseline for thermal sensation prediction.

Overall, the results show that, except for AdaBoost, ensemble learning algorithms can improve the overall classification performance in terms of accuracy, precision, and recall metrics. Among the ensemble categories, advanced and bagging algorithms achieve higher F1 scores than others, with advanced algorithms (HGB and RF) outperforming the bagging framework in nearly all performance metrics. Additionally, DT is found to be the best choice for designing a bagging ensemble algorithm. The poor performance of the boosting framework (AdaBoost) on the weighted F1 metric suggests that this may be due to the extreme underrepresentation of the minority class in the dataset. Stacking is found to be inferior to other ensemble approaches when dealing with this imbalanced TSV dataset. The soft voting framework is found to be superior to hard voting, and taking into account the probabilities predicted by each model can lead to more accurate classification in TSV prediction. However, voting frameworks in general do not perform well in this study. When compared to basic ML, the accuracy, precision, and recall scores of the ensemble approaches are better, while only bagging and advanced ensembles outperform basic ML in terms of F1 scores. KNN was found to be the best F1 scorer for both 3-point and 7-point TSV prediction among the basic ML algorithms, likely because it makes no assumptions about the underlying distribution of the data and is therefore effective in detecting patterns in the minority class compared to boosting and voting ensembles, which may ignore the minority class in imbalanced datasets. Moreover, the results of the statistical test indicate that the advanced ensemble category comprising HGB and RF algorithms significantly outperform other methods at the 0.05 significance level. However, statistical analysis indicates no significant difference in performance between the bagging ensemble and the basic ML approach. Together, based on the experimental results, RF and HGB are the most suitable options for developing TSV classification applications using machine learning due to their superior performance over other methods.

Regarding the comparison of TSV scale units, the performance of ML algorithms for 3-point TSV prediction is higher than that of 7-point TSV prediction since 3-point classification is relatively easier than 7-point classification. Among all basic algorithms experimented in this study, KNN shows the best performance for both 3-point and 7-point TSV prediction with F1 scores of 0.614 and 0.516, respectively. Among ensemble ML frameworks, boosting (Adaboost) ensemble algorithms provide low performance scores for both 3-point and 7-point TSV prediction. Advanced ensemble ML algorithms outperform both basic ML algorithms and other ensemble learning frameworks. In particular, HGB with F1 score of 0.638, is the best performing advanced ensemble ML algorithm for 3-point TSV prediction whereas, RF with F1 score of 0.549, achieves the best performance for 7-point TSV prediction. It should be noted that the classification accuracy of 0.66 for 3-point and 0.61 for 7-point TSV may initially appear to be relatively low. However, it is important to consider that these values can still be deemed successful when compared to the respective random choice accuracies of 0.33 for 3-point predictions and 0.14 for 7-point predictions. Furthermore, the absence of superior findings in the existing literature serves as an indicator that we have encountered avoidable bias within the dataset.

4.6. Importance analysis of variables

In this study, feature importance values derived from the fitted classifier were also analyzed to identify the variables that contributed most to the prediction of thermal sensation. The analysis

obtained via RF represents the relative contribution of each variable to reducing classification uncertainty, with higher values indicating stronger influence. The results (Fig. 13) show that environmental variables were the dominant predictors of thermal sensation. Indoor air temperature (0.1724) and SET (0.1692) emerged as the two most influential features, followed by relative humidity (0.1514) and air velocity (0.1334). Personal variables such as clothing insulation (0.1206), age (0.0847), and metabolic rate (0.0525) also contributed meaningfully. By contrast, contextual variables such as sex (0.0264), building type (0.0079), and season (0.0124) had only minor influence. These findings are consistent with thermal comfort theory, where thermal sensation is primarily shaped by indoor climate conditions and individual physiological variables, while demographic or contextual variables play a secondary role. This analysis clarifies how the model made its predictions and strengthens confidence in the plausibility of the results.

5. Discussion

5.1. Performance of ML algorithms

The results of this study demonstrate that ensemble learning algorithms, except for AdaBoost, significantly improve overall classification performance in terms of accuracy, precision, and recall metrics. Specifically, advanced ensemble ML algorithms (HGB and RF) and bagging frameworks achieve higher F1 scores than other methods. This observation aligns with existing literature. For example, in a recent study, ML algorithms including RF and SVM predicted 3-point TSV with 60 – 66.2% accuracy and 7-point TSV with 50 – 61.1% accuracy. The results showed that RF obtained the best performance, aligning with the findings where RF achieves high F1 scores for both 3-point and 7-point TSV predictions [32]. Furthermore, studies applying RF in different settings confirmed its strong performance. For example, in one study focusing on personalized data-driven models using the ASHRAE Global Thermal Comfort Database II, RF achieved over 70% accuracy for thermal preference votes [33]. Another study using spatial parameters found that RF outperformed other methods, improving prediction accuracy by 38.6% when variables like distance to windows and AC units were included [34]. Similarly, studies investigating demographic predictors also identified RF and KNN as the best-performing algorithms for both thermal sensation and thermal satisfaction [35].

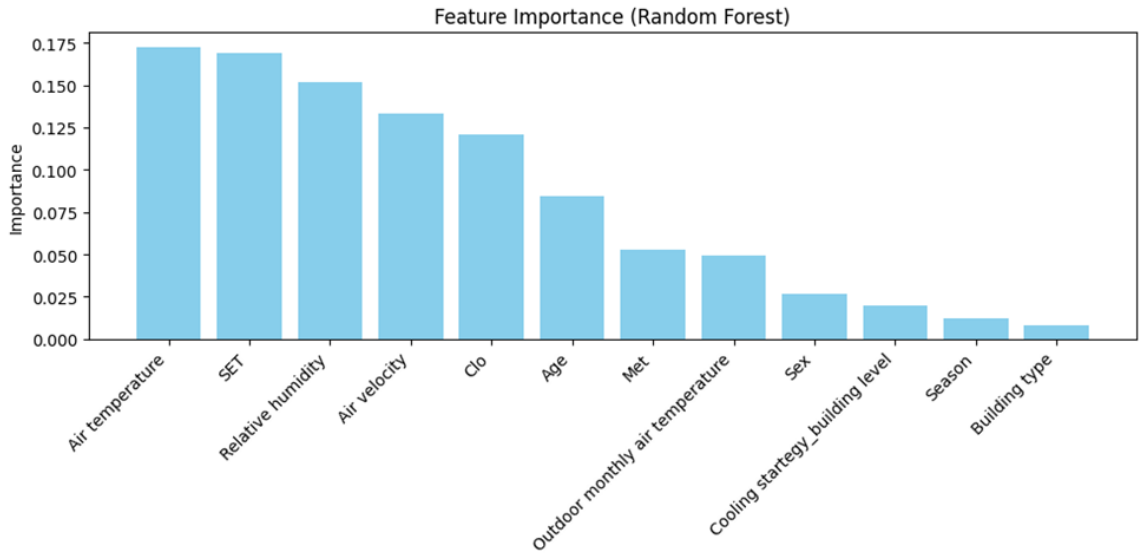

Fig. 13. Feature importance values of input variables for the RF model

Another study found that ensemble learning models (DCF and RF) performed better for predicting thermal preferences than traditional models such as DT, LR and NB [45]. Similarly, RF is highlighted as a top performer among ensemble algorithms. Additionally, Deep Forest achieved the best accuracy of 82%, supporting the observation that tree-based models (like RF) excel in predicting occupant thermal preference [46]. The boosting framework (AdaBoost) performed poorly on the weighted F1 metric, likely due to the underrepresentation of the minority class. This observation is consistent with the finding of AdaBoost's lower performance in a recent study [67] in which the KNN algorithm outperformed other algorithms, including boosting methods, in predicting TSV. Stacking was determined to be less effective compared to other ensemble methods in the context of addressing an imbalanced TSV dataset. Additionally, while soft voting outperformed hard voting, its overall performance was unsatisfactory. This result is consistent with the existing literature [22, 68, 69], which indicates that stacking and voting frameworks are not typically recognized as leading performers in such scenarios. KNN was the best F1 scorer for both 3-point and 7-point TSV prediction among basic ML algorithms. This finding is supported by multiple studies where KNN outperformed several algorithms, including traditional and ensemble methods, particularly for 3-point and 7-point TSV predictions [22, 23, 45, 67, 70]. Statistical tests showed that advanced ensemble algorithms (HGB and RF) significantly outperformed other methods at the 0.05 significance level, while no significant difference was found between bagging ensembles and basic ML approaches. This observation is supported by several studies [24, 31, 32, 45, 71, 72] in which RF and advanced ensemble methods generally achieved higher accuracy and F1 scores. In summary, the findings align with existing literature, particularly regarding the superiority of advanced ensemble algorithms like RF and HGB in predicting thermal preferences. The consistent observation across multiple studies that ensemble methods outperform traditional algorithms underscores the robustness of the results. However, the study also highlights specific limitations, such as the poor performance of boosting frameworks like AdaBoost and the relative underperformance of voting methods, which are crucial considerations for future research.

5.2. Importance of variables in TSV prediction

In this study, the feature importance analysis obtained from RF showed that indoor air temperature and SET were the two most influential predictors, followed by relative humidity and air velocity; among personal variables, clothing insulation, age, and metabolic rate had moderate contributions, while contextual variables such as sex, building type, and season exhibited only minor influence. These findings are highly consistent with previous research. For instance, Luo et al. [32], using the ASHRAE Global Thermal Comfort Database II, demonstrated that the most influential features in RF models were indoor air temperature, SET, RH, air velocity, clothing insulation, age, and metabolic rate, while outdoor temperature, season, operation mode, sex, and building type played a secondary role, which align almost exactly with the findings of this study.

In naturally ventilated residential buildings, Chai et al. [11] confirmed that both indoor and outdoor environmental variables and personal variables such as CLO and MET significantly affected thermal comfort votes, while adaptive control measures (e.g., window opening) had a limited effect on thermal comfort vote but higher impact on TSV. This finding supports the importance of environmental and personal variables whereas contextual and adaptive variables play a situationally dependent secondary role.

A systematic review by Mamani et al. [73] further underlined that the most frequently used variables in thermal comfort modeling are environmental variables including air temperature, relative humidity, mean radiant temperature, and air velocity while clothing insulation and metabolic rate remain fundamental. More recent studies also highlight the growing role of physiological signals (e.g., skin temperature, heart rate, EEG), indicating

a trend toward individualized comfort modeling. Similarly, Yu et al. [22], in a climate chamber experiment, compared professional (high-precision physiological sensing) and practical (low-cost inputs) setups, finding that skin temperatures were strong predictors, while age gained greater importance under low-cost and practical scenarios. The fact that age emerged as a moderately important factor even without physiological inputs further confirms that demographic differences still have explanatory power.

Using a large dataset of 17,814 samples, Haghirad et al. [33] showed that expanding feature sets beyond classical PMV variables improved predictive performance, but also observed that climate, cooling strategies, age, and sex were weaker than core environmental and personal variables, supporting this study's findings which also show that contextual and demographic variables play a secondary role .

Demographic-focused research by Kocaman et al. [35] found that sex, age, and thermal history were significant for TSV and satisfaction (TSa), whereas education, income, and occupation were not significant for TSV, which aligns with this study's finding of low importance for sex in thermal sensation prediction.

In conclusion, both theoretical and empirical findings consistently show that indoor air temperature, SET, relative humidity, and air velocity are dominant determinants of thermal sensation; CLO and MET are the key personal variables capturing inter-individual differences; age plays a moderate role; and contextual and demographic variables such as sex, building type, or season generally exert only secondary influence. The results of this study are thus both physically plausible and well aligned with existing literature, reinforcing confidence in the interpretability and credibility of the proposed model.

5.3. Importance of TSV scale units

In relation to the TSV scale units, the 3-point TSV prediction demonstrated higher performance compared to the 7-point TSV prediction, attributable to the relative simplicity of 3-point classification. This observation aligns with existing literature [32, 74], which shows that accuracy metrics for 3-point and 7-point TSV predictions following similar trends, with simpler classifications achieving superior performance. The findings of this study show that the classification accuracy of 0.66 for 3-point and 0.61 for 7-point TSV are successful compared to random choice accuracies, which are 0.33 and 0.14, respectively. Although existing literature indicates that the accuracy of 7-point TSV prediction models using the ASHRAE Thermal Comfort dataset varies between 40% and 60% [32, 45, 46, 75], they also report that ASHRAE Global Database poses a significant challenge due to its imbalanced dataset nature, causing a relatively low accuracy performance. The findings of this study indicate that ensemble frameworks generally enhance performance metrics for TSV prediction on the ASHRAE Comfort Database II, although some exceptions exist. Notably, in terms of F1 scores, advanced ensemble algorithms like RF and HGB are preferable choices. It can be concluded that ensemble learning serves as a valuable approach to address this problem. Given the strong dependence of machine learning algorithm performance on the underlying data, careful selection among various alternatives becomes essential.

Overall, the results show clear differences in performance between basic ML algorithms and ensemble methods for TSV prediction. Among the basic algorithms, KNN performed strongly in handling imbalanced data for 3-point and 7-point TSV classification, due to its non-parametric nature, which effectively detects patterns in minority classes. However, its ability to capture complex relationships is limited compared to ensemble methods. Advanced ensemble techniques, such as RF and HGB, significantly outperformed basic algorithms by combining multiple models to achieve higher scores in performance metrics. While basic ML algorithms offer simplicity and computational efficiency, advanced ensemble methods provide greater predictive power and robustness, particularly for addressing the complexities of TSV prediction data.

5.4. Practical implications

Thermal comfort prediction models have a wide range of practical applications in smart buildings, where they significantly enhance the functionality and efficiency of building management systems. One of the most direct applications is in the optimization of HVAC systems. Thermal comfort models can predict the need for heating, cooling, or ventilation in different zones of a building, enabling the HVAC system to adjust its output in real-time thereby reducing energy consumption while maintaining comfort. In addition, these models can enable smart buildings to adjust temperature settings automatically based on predictive data about external weather conditions, the number of occupants, and their thermal comfort preferences [76]. As a result, thermal comfort prediction models integrated into building automation systems can be beneficial for real-time control and optimization of HVAC operations. Thermal comfort prediction models can also be integrated into occupancy sensors to adjust the temperature in real-time based on the number of people in a room, which means that energy is not wasted heating or cooling unoccupied or partially occupied spaces [77–79]. As another example, the system can dim lights in cooler areas to reduce heat emission or increase natural lighting to warm up a space, enabling a more holistic approach to smart building management [80–82]. Subsequently, the implementation of closed-loop control strategies that dynamically adjust building systems can maintain desired comfort levels while minimizing energy consumption. During the design and construction phases, thermal comfort prediction models can help architects and engineers make informed decisions about building orientation, materials, and systems, ensuring that the new building is capable of maintaining thermal comfort efficiently [83]. By maintaining an optimal thermal environment, these models contribute to the health and productivity of the building's occupants. In essence, thermal comfort prediction models are pivotal in the operation of smart buildings, providing a data-driven approach to managing comfort and energy use, which aligns with modern sustainability and efficiency goals [84–87].

Recently, efforts to control HVAC consumption systems in buildings are progressing at full speed [88–92]. The primary goal while managing HVAC systems is to ensure the maximum thermal comfort for the occupants whilst keeping the building's energy consumption at a minimum. Achieving this goal necessitates the most accurate prediction of the occupants' thermal comfort. To this end, data collected from the building, its surroundings, and its occupants through Building Management Systems (BMS), IoT technologies, and surveys are utilized to execute a thermal comfort prediction model. Within this context, the importance of thermal comfort prediction models cannot be overstated. It serves as a bridge between energy efficiency and occupant comfort, two critical aspects of modern building design and operation. By accurately predicting thermal comfort, building managers can make informed decisions about adjusting the HVAC settings, thereby avoiding unnecessary energy use while ensuring that the occupants' comfort is not compromised. This not only leads to significant energy savings but also enhances the well-being and productivity of the occupants, as they experience an environment that is tailored to their comfort needs. In summary, the thermal comfort prediction model is a cornerstone of sustainable building management. It exemplifies how cutting-edge technology can be harnessed to create buildings that are not only energy-efficient but also provide a comfortable and conducive environment for their occupants. The ongoing advancements in this field are paving the way for smarter, more responsive buildings that are in tune with the needs of their occupants while also contributing to the broader goals of energy conservation and environmental sustainability. Fig. 14 illustrates the integration of occupant thermal comfort prediction models with HVAC systems in buildings. It includes a detailed diagram showing the flow of information between data collection (i.e via sensors, BMS), prediction models, and the HVAC system.

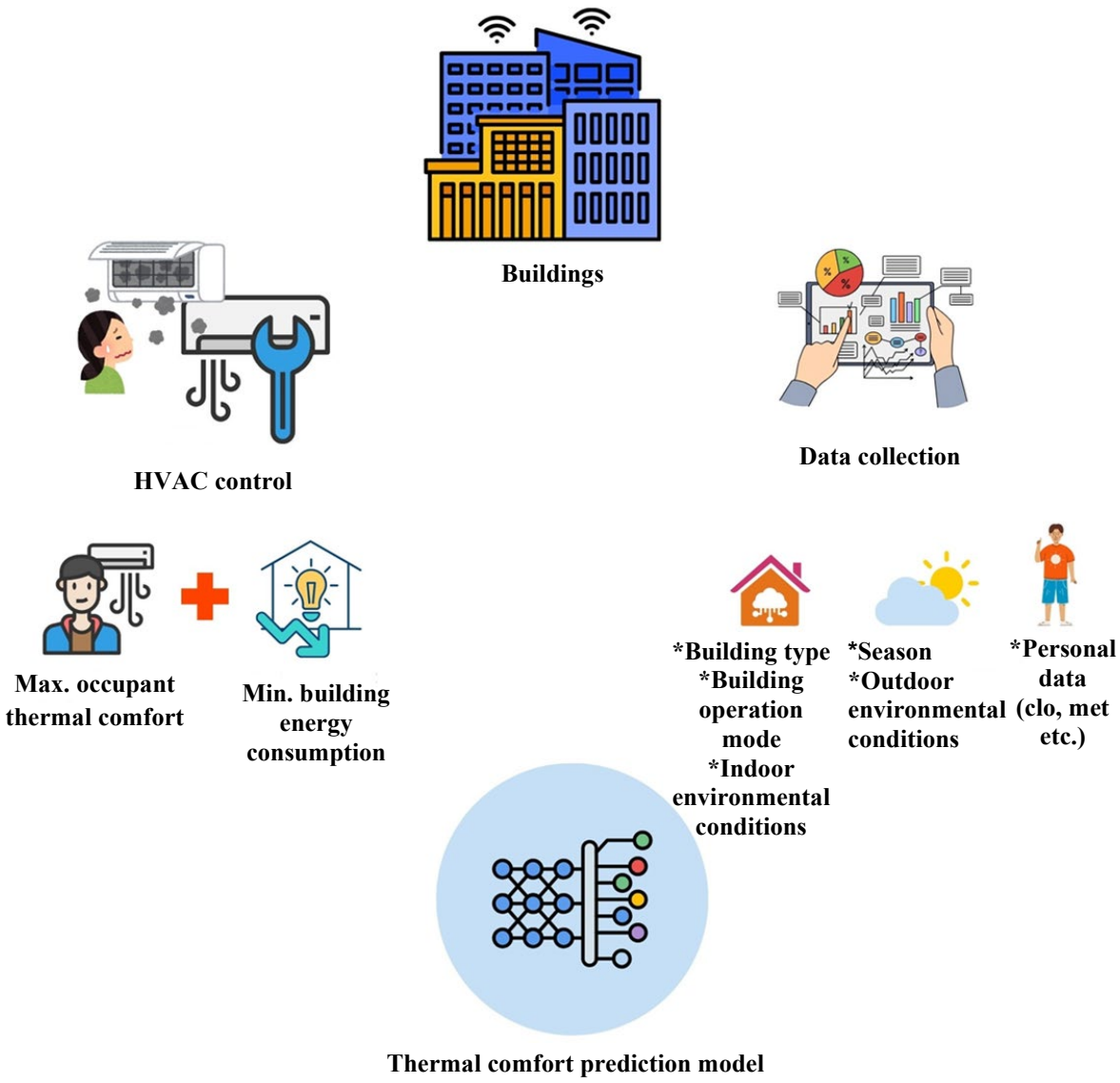

Fig. 14. Diagram for integrating thermal comfort prediction models with HVAC systems in smart buildings

6. Conclusion

In this study, four ensemble ML frameworks (bagging, boosting, stacking, and voting), six basic ML algorithms (LR, KNN, SVM, DT, MLP, MNB), and six advanced ensemble ML algorithms (RF, ROF, XT, GB, HGB, XGB) were systematically compared for 3-point and 7-point TSV prediction using the ASHRAE Comfort Database II. Our comprehensive analysis provided valuable insights into the relative performance of these models under different scenarios.

For 3-point TSV prediction, the F1 score ranged between 0.379 and 0.614, while for 7-point TSV prediction, it ranged between 0.379 and 0.516. Among basic ML algorithms, KNN consistently achieved the highest F1 scores, while MNB showed the lowest performance. Bagging ensemble frameworks significantly improved DT's performance, though gains were limited for other models. Boosting frameworks showed variable results, with the boosting DT performing best for 3-point TSV, while SVM retained its performance under boosting. Stacking frameworks improved

performance, particularly in 3-point TSV. Soft voting ensembles also outperformed hard voting, especially for 7-point TSV. Overall, advanced ensemble algorithms outperformed basic algorithms, with XT, HGB, and RF achieving the best performances across different metrics. Notably, HGB and RF significantly outperformed all other methods at the 0.05 significance level.

Practical implications are particularly significant. Integrating the best-performing models (HGB and RF) into BMS and HVAC control platforms can enable real-time prediction of occupants' thermal comfort, allowing for dynamic adjustment of heating, cooling, and ventilation in different zones of a building. This would reduce unnecessary energy use while maintaining comfort, contributing to both sustainability and occupant well-being. In addition, practitioners such as facility managers and building owners can embed ensemble ML models in smart sensors and IoT-based monitoring systems to provide adaptive, closed-loop HVAC control strategies. These models can guide practitioners to optimize setpoints, integrate occupant feedback, and balance comfort with energy savings. Policymakers and standards bodies may also leverage such models to update building codes, moving toward performance-based regulations that incorporate data-driven comfort predictions.

In summary, the findings highlight that advanced ensemble ML algorithms not only enhance prediction accuracy but also provide actionable tools for practitioners to improve building performance. Future research should focus on testing these models in real-world building environments, expanding datasets with diverse demographic and contextual features, and exploring hybrid frameworks that combine physics-based and data-driven approaches. Such efforts will help bridge the gap between theoretical performance and large-scale deployment, ensuring that ML-driven comfort models become a cornerstone of sustainable and human-centered building management.

Declaration

Funding

This research received no external funding.

Author Contributions

M. Kuru Erdem: Conceptualization, Methodology, Resources, Data Analysis, Writing-Original draft. O. Gökalp: Conceptualization, Methodology, Resources, Data Analysis, Writing-Original draft, Writing- Review & Editing. G. Calis: Conceptualization, Project administration, Supervision, Writing- Review & Editing.

Acknowledgments

Not applicable.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Ethics Committee Permission

Not applicable.

Conflict of Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

[1] Kaushik A, Arif M, Tumula P, Ebohon OJ (2020) Effect of thermal comfort on occupant productivity in office buildings: Response surface analysis. Build Environ 180:107021. https://doi.org/10.1016/j.buildenv.2020.107021

[2] Jiang Y, Luo Z, Wang Z, Lin B (2019) Review of thermal comfort infused with the latest big data and modeling progresses in public health. Build Environ 164:106336. https://doi.org/10.1016/j.buildenv.2019.106336

[3] Aldin SS, Sözer H (2022) Comparing the accuracy of ANN and ANFIS models for predicting the thermal data. J Constr Eng Manag Innov 5:119–139. https://doi.org/10.31462/jcemi.2022.02119139

[4] Birgönül Z (2024) A user-centric analysis and survey-based approach for early design decision-making criteria of the 'Symbiotic Data Platform': Integration of BIM with IoT for occupants' thermal comfort. J Constr Eng Manag Innov 7:56–76. https://doi.org/10.31462/jcemi.2024.01056076

[5] ISO 7730 (2005) 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. 52

[6] Zheng Z, Zhang Y, Mao Y, Yang Y, Fu C, Fang Z (2021) Analysis of SET* and PMV to evaluate thermal comfort in prefab construction site offices: Case study in South China. Case Stud Therm Eng 26. https://doi.org/10.1016/j.csite.2021.101137

[7] Du H, Lian Z, Lai D, Duanmu L, Zhai Y, Cao B, Zhang Y, Zhou X, Wang Z, Zhang X, Hou Z (2022) Evaluation of the accuracy of PMV and its several revised models using the Chinese thermal comfort Database. Energy Build 271:112334. https://doi.org/10.1016/j.enbuild.2022.112334

[8] Broday EE, Ruivo CR, Gameiro da Silva M (2021) The use of Monte Carlo method to assess the uncertainty of thermal comfort indices PMV and PPD: Benefits of using a measuring set with an operative temperature probe. J Build Eng 35:101961. https://doi.org/10.1016/j.jobe.2020.101961

[9] Cheung T, Brager G, Parkinson T, Li P, Brager G (2019) Analysis of the accuracy on PMV – PPD model using the ASHRAE Global Thermal Comfort Database II. Build Environ 153:205–217. https://doi.org/10.1016/j.buildenv.2019.01.055

[10] Kiki G, Kouchadé C, Houngan A, Zannou-Tchoko SJ, André P (2020) Evaluation of thermal comfort in an office building in the humid tropical climate of Benin. Build Environ 185:. https://doi.org/10.1016/j.buildenv.2020.107277

[11] Chai Q, Wang H, Zhai Y, Yang L (2020) Using machine learning algorithms to predict occupants' thermal comfort in naturally ventilated residential buildings. Energy Build 217:109937. https://doi.org/10.1016/j.enbuild.2020.109937

[12] Kim J, Tartarini F, Parkinson T, Cooper P, De Dear R (2019) Thermal comfort in a mixed-mode building: Are occupants more adaptive? Energy Build 203:109436. https://doi.org/10.1016/j.enbuild.2019.109436

[13] de Dear R, Xiong J, Kim J, Cao B (2020) A review of adaptive thermal comfort research since 1998. Energy Build 214:109893. https://doi.org/10.1016/j.enbuild.2020.109893

[14] Calis G, Kuru M (2017) Assessing user thermal sensation in the Aegean region against standards. Sustain Cities Soc 29:77–85. https://doi.org/10.1016/j.scs.2016.11.013

[15] Calis G, Kuru M (2019) Statistical significance of gender and age on thermal comfort : A case study in Turkey. Proc Inst Civ Eng Eng Sustain 172:40–51. https://doi.org/10.1680/jensu.17.00003

[16] Zhou X, Lai D, Chen Q (2021) Evaluation of thermal sensation models for predicting thermal comfort in dynamic outdoor and indoor environments. Energy Build 238:110847. https://doi.org/10.1016/j.enbuild.2021.110847

[17] Jiao Y, Yu H, Yu Y, Wang Z, Wei Q (2020) Adaptive thermal comfort models for homes for older people in Shanghai, China. Energy Build 215:109918. https://doi.org/10.1016/j.enbuild.2020.109918

[18] Carlucci S, Bai L, de Dear R, Yang L (2018) Review of adaptive thermal comfort models in built environmental regulatory documents. Build Environ 137:73–89. https://doi.org/10.1016/j.buildenv.2018.03.053

[19] Qin H, Wang X (2022) A multi-discipline predictive intelligent control method for maintaining the thermal comfort on indoor environment. Appl Soft Comput 116:108299. https://doi.org/10.1016/j.asoc.2021.108299

[20] Qavidel Fard Z, Zomorodian ZS, Korsavi SS (2022) Application of machine learning in thermal comfort studies: A review of methods, performance and challenges. Energy Build 256:111771. https://doi.org/10.1016/j.enbuild.2021.111771

[21] Akman G, Yorur B, Boyaci AI, Chiu MC (2023) Assessing innovation capabilities of manufacturing companies by combination of unsupervised and supervised machine learning approaches. Appl Soft Comput 147:110735. https://doi.org/10.1016/j.asoc.2023.110735

[22] Yu C, Li B, Wu Y, et al (2022) Performances of machine learning algorithms for individual thermal comfort prediction based on data from professional and practical settings. J Build Eng 61:105278. https://doi.org/10.1016/j.jobe.2022.105278

[23] Yang B, Li X, Liu Y, Chen L, Guo R, Wang F, Yan K (2022) Comparison of models for predicting winter individual thermal comfort based on machine learning algorithms. Build Environ 215:108970. https://doi.org/10.1016/j.buildenv.2022.108970

[24] Cen C, Cheng S, Wong NH (2022) Physiological sensing of personal thermal comfort with wearable devices in fan-assisted cooling environments in the tropics. Build Environ 225:109622. https://doi.org/10.1016/j.buildenv.2022.109622

[25] Zhou X, Xu L, Zhang J, Niu B, Luo M, Zhou G, Zhang X (2020) Data-driven thermal comfort model via support vector machine algorithms: Insights from ASHRAE RP-884 database. Energy Build 211:109795. https://doi.org/10.1016/j.enbuild.2020.109795

[26] Rehman SU, Javed AR, Khan MU, Nazar Awan M, Farukh A, Hussien A (2022) PersonalisedComfort: A personalised thermal comfort model to predict thermal sensation votes for smart building residents. Enterp Inf Syst 16(7):1852316. https://doi.org/10.1080/17517575.2020.1852316

[27] Fayyaz M, Farhan AA, Javed AR (2022) Thermal comfort model for HVAC buildings using machine learning. Arab J Sci Eng 47:2045–2060. https://doi.org/10.1007/s13369-021-06156-8

[28] Chai Q, Wang H, Zhai Y, Yang L (2020) Using machine learning algorithms to predict occupants' thermal comfort in naturally ventilated residential buildings. Energy Build 217:109937. https://doi.org/10.1016/j.enbuild.2020.109937

[29] Sibyan H, Svajlenka J, Hermawan H, Faqih N, Arrizqi AN (2022) Thermal comfort prediction accuracy with machine learning between regression analysis and naïve bayes classifier. Sustain 14: (23):15663. https://doi.org/10.3390/su142315663

[30] Shan X, Yang EH (2020) Supervised machine learning of thermal comfort under different indoor temperatures using EEG measurements. Energy Build 225:110305. https://doi.org/10.1016/j.enbuild.2020.110305

[31] Tardioli G, Filho R, Bernaud P, Ntimos D (2022) An innovative modelling approach based on building physics and machine learning for the prediction of indoor thermal comfort in an office building. Buildings 12:(4):475. https://doi.org/10.3390/buildings12040475

[32] Luo M, Xie J, Yan Y, Ke Z, Yu P, Wang Z, Zhang J (2020) Comparing machine learning algorithms in predicting thermal sensation using ASHRAE Comfort Database II. Energy Build 210:109776. https://doi.org/10.1016/j.enbuild.2020.109776

[33] Haghirad M, Heidari S, Hosseini H (2024) Advancing personal thermal comfort prediction: A data-driven framework integrating environmental and occupant dynamics using machine learning. Build Environ 262:111799. https://doi.org/10.1016/j.buildenv.2024.111799

[34] Alam N, Zaki SA, Ahmad SA, Singh MK, Azizan A, Othman NA (2024) Machine learning approach for predicting personal thermal comfort in air conditioning offices in Malaysia. Build Environ 266:112083. https://doi.org/10.1016/j.buildenv.2024.112083

[35] Kocaman E, Erdem MK, Calis G (2024) Machine learning thermal comfort prediction models based on occupant demographic characteristics. J Therm Biol 123:103884. https://doi.org/10.1016/j.jtherbio.2024.103884

[36] Boutahri Y, Tilioua A (2024) Machine learning-based predictive model for thermal comfort and energy optimization in smart buildings. Results Eng 22:102148. https://doi.org/10.1016/j.rineng.2024.102148

[37] Ren J, Zhang R, Cao X, Kong X (2024) Experimental study on the physiological parameters of occupants under different temperatures and prediction of their thermal comfort using machine learning algorithms. J Build Eng 84:108676. https://doi.org/10.1016/j.jobe.2024.108676

[38] Li Y, Gao F, Yu J, Fei T (2025) Machine learning based thermal comfort prediction in office spaces: Integrating SMOTE and SHAP methods. Energy Build 329:115267. https://doi.org/10.1016/j.enbuild.2024.115267

[39] Avci AB (2025) Machine learning-based prediction of thermal comfort: exploring building types, climate, ventilation strategies, and seasonal variations. Build Res Inf 3218:1–18. https://doi.org/10.1080/09613218.2025.2462932

[40] He M, Liu H, Zhou S, Yao Y, Kosonen R, Wu Y, Li B (2025) Machine learning-based assessment of thermal comfort for the elderly in warm environments: Combining the XGBoost algorithm and human body exergy analysis. Int J Therm Sci 209:109519. https://doi.org/10.1016/j.ijthermalsci.2024.109519

[41] Lu S, Wang W, Lin C, Hameen EC (2019) Data-driven simulation of a thermal comfort-based temperature set-point control with ASHRAE

RP884. Build Environ 156:137–146. https://doi.org/10.1016/j.buildenv.2019.03.010

[42] Wu Z, Li N, Peng J, Cui H, Liu P, Li H, Li X (2018) Using an ensemble machine learning methodology-Bagging to predict occupants' thermal comfort in buildings. Energy Build 173:117–127. https://doi.org/10.1016/j.enbuild.2018.05.031

[43] Lou H, Ou D (2019) A comparative field study of indoor environmental quality in two types of open-plan offices: Open-plan administrative offices and open-plan research offices. Build Environ 148:394–404. https://doi.org/10.1016/j.buildenv.2018.11.022

[44] Kim J, Zhou Y, Schiavon S, Raftery P, Brager G (2018) Personal comfort models: Predicting individuals' thermal preference using occupant heating and cooling behavior and machine learning. Build Environ 129:96–106. https://doi.org/10.1016/j.buildenv.2017.12.011

[45] Bai Y, Liu K, Wang Y (2022) Comparative analysis of thermal preference prediction performance in different conditions using ensemble learning models based on ASHRAE Comfort Database II. Build Environ 223:109462. https://doi.org/10.1016/j.buildenv.2022.109462

[46] Han X, Hu Z, Li C, Wu J, Li C, Sun B (2023) Prediction of human thermal comfort preference based on supervised learning. J Therm Biol 112:103484. https://doi.org/10.1016/j.jtherbio.2023.103484

[47] Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F (2018) Learning from Imbalanced Data Sets. Springer International Publishing, Cham.

[48] Zhu B, Qian C, vanden Broucke S, Xiao J, Li Y (2023) A bagging-based selective ensemble model for churn prediction on imbalanced data. Expert Syst Appl 227:120223. https://doi.org/10.1016/j.eswa.2023.120223

[49] Louk MHL, Tama BA (2023) Dual-IDS: A bagging-based gradient boosting decision tree model for network anomaly intrusion detection system. Expert Syst Appl 213:119030. https://doi.org/10.1016/j.eswa.2022.119030

[50] Menor-Flores M, Vega-Rodríguez MA (2023) Boosting-based ensemble of global network aligners for PPI network alignment. Expert Syst Appl 230:120671. https://doi.org/10.1016/j.eswa.2023.120671

[51] Chacón H, Koppisetti V, Hardage D, Choo KK, Rad P (2023) Forecasting call center arrivals using temporal memory networks and gradient boosting algorithm. Expert Syst Appl 224:119983. https://doi.org/10.1016/j.eswa.2023.119983

[52] Xie Y, Sun W, Ren M, Chen S, Huang Z, Pan X (2023) Stacking ensemble learning models for daily runoff prediction using 1D and 2D CNNs. Expert Syst Appl 217:119469. https://doi.org/10.1016/j.eswa.2022.119469

[53] Chung D, Yun J, Lee J, Jeon Y (2023) Predictive model of employee attrition based on stacking ensemble learning. Expert Syst Appl 215:119364. https://doi.org/10.1016/j.eswa.2022.119364

[54] Tavana P, Akraminia M, Koochari A, Bagherifard A (2023) An efficient ensemble method for detecting spinal curvature type using deep transfer learning and soft voting classifier. Expert Syst Appl 213:119290. https://doi.org/10.1016/j.eswa.2022.119290

[55] Zhang W, Yang D, Zhang S (2021) A new hybrid ensemble model with voting-based outlier detection and balanced sampling for credit scoring. Expert Syst Appl 174:114744. https://doi.org/10.1016/j.eswa.2021.114744

[56] Bentéjac C, Csörgő A, Martínez-Muñoz G (2021) A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review 54(3):1937–67.

[57] Caruana R, Niculescu-Mizil A (2006) An empirical comparison of supervised learning algorithms. In: ACM Int Conf Proceeding Ser 148:161–168. https://doi.org/10.1145/1143844.1143865

[58] Ibrahim Y, Okafor E, Yahaya B, Yusuf SM, Abubakar ZM, Bagaye UY (2021) Comparative study of ensemble learning techniques for text classification. In: 1st Int Conf Multidiscip Eng Appl Sci ICMEAS 2021 1–5. https://doi.org/10.1109/ICMEAS52683.2021.9692306

[59] Khamar M, Eftekhari M (2018) Multi-manifold based rotation forest for classification. Appl Soft Comput J 68:626–635. https://doi.org/10.1016/j.asoc.2018.04.026

[60] Wang Y, Wang S, Sima X, Song Y, Cui S, Wang D (2023) Expanded feature space-based gradient boosting ensemble learning for risk prediction of type 2 diabetes complications. Appl Soft Comput 144:110451. https://doi.org/10.1016/j.asoc.2023.110451

[61] Koc K (2023) Determining the Short-term susceptibility of construction workers to occupational accidents using stochastic gradient

boosting. J Constr Eng Manag Innov 6:1–15. https://doi.org/10.31462/jcemi.2023.01001015

[62] Aryal A, Becerik-Gerber B (2020) Thermal comfort modeling when personalized comfort systems are in use: Comparison of sensing and learning methods. Build Environ 185:107316. https://doi.org/10.1016/j.buildenv.2020.107316

[63] Wang Z, Yu H, Luo M, et al (2019) Predicting older people's thermal sensation in building environment through a machine learning approach: Modelling, interpretation, and application. Build Environ 161:106231. https://doi.org/10.1016/j.buildenv.2019.106231

[64] Chaudhuri T, Zhai D, Soh YC, Li H, Xie L (2018) Random forest based thermal comfort prediction from gender-specific physiological parameters using wearable sensing technology. Energy Build 166:391–406. https://doi.org/10.1016/j.enbuild.2018.02.035

[65] Gao G, Li J, Wen Y (2020) DeepComfort: Energy-efficient thermal comfort control in buildings via reinforcement learning. IEEE Internet Things J 7:8472–8484. https://doi.org/10.1109/JIOT.2020.2992117

[66] Farhadpour S, Warner TA, Maxwell AE (2024) Selecting and interpreting multiclass loss and accuracy assessment metrics for classifications with class imbalance: Guidance and best practices. Remote Sens 16:1–22. https://doi.org/10.3390/rs16030533

[67] Hossain T, Nkurikiyeyezu K, Kawasaki Y, Lopez G (2022) Toward the prediction of environmental thermal comfort sensation using wearables. 312–324. https://doi.org/10.3233/aise220058

[68] Salamone F, Bellazzi A, Belussi L, Damato G, Danza L, Dell'Aquila F, Ghellere M, Megale V, Meroni I, Vitaletti W (2020) A machine learning approach for personal thermal comfort perception evaluation: Experimental campaign under real and virtual scenarios. In: E3S Web Conf 197:04001. https://doi.org/10.1051/e3sconf/202019704001

[69] Wang J, Li Q, Zhu G, Kong W, Peng H, Wei M (2024) Recognition and prediction of elderly thermal sensation based on outdoor facial skin temperature. Build Environ 253:111326. https://doi.org/10.1016/j.buildenv.2024.111326

[70] Mohamed Salleh FH, Saripuddin MB, Bin Omar R (2020) Predicting thermal comfort of HVAC building using 6 thermal factors. In: 8th Int Conf Inf Technol Multimedia ICIMU 2020 170–176. https://doi.org/10.1109/ICIMU49871.2020.9243466

[71] Zhan J, He W (2022) Evaluation and prediction of elderly thermal comfort at varying ambient temperatures based on electroencephalogram signals and machine learning. In: Proc of 15th Int Congr Image Signal Process Biomed Eng Informatics CISP-BMEI 3–7. https://doi.org/10.1109/CISP-BMEI56279.2022.9980300

[72] Almadhor A, Wechtaisong C, Tariq U, Kryvinska N, Al Hejaili A, Mohammad UG, Alanazi M (2024) Fine-tuned extra tree classifier for thermal comfort sensation prediction. Comput Syst Sci Eng 48:200–216. https://doi.org/10.32604/csse.2023.039546

[73] Mamani T, Herrera RF, Rivera FM La, Atencio E (2022) Variables that affect thermal comfort and its measuring instruments: A systematic review. Sustain 14(3):1773. https://doi.org/10.3390/su14031773

[74] Čulić A, Nižetić S, Gambiroža JČ, Šolić P (2025) Progress in data-driven thermal comfort analysis and modeling. Energy Build 336(2025):115599.

[75] Deghim F, Banihashemi F, Koth S, Lang W (2022) A data-driven approach for predicting occupant thermal comfort in offices. In: Proc of 33. Forum Bauinformatik 257–264.

[76] Mtibaa F, Nguyen KK, Azam M, Papachristou A, Venne JS, Cheriet M (2020) LSTM-based indoor air temperature prediction framework for HVAC systems in smart buildings. Neural Comput Appl 32:17569–17585. https://doi.org/10.1007/s00521-020-04926-3

[77] Duan W, Wang Y, Li J, Zheng Y, Ning C, Duan P (2021) Real-time surveillance-video-based personalized thermal comfort recognition. Energy Build 244:110989. https://doi.org/10.1016/j.enbuild.2021.110989

[78] Kumar TMS, Kurian CP (2022) Real-time data based thermal comfort prediction leading to temperature setpoint control. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-022-03754-8

[79] Jung S, Jeoung J, Hong T (2022) Occupant-centered real-time control of indoor temperature using deep learning algorithms. Build Environ 208:108633. https://doi.org/10.1016/j.buildenv.2021.108633

[80] Higuera J, Hertog W, Perálvarez M, Carreras J (2014) Hybrid smart lighting and climate control system for buildings. In: IET Conf on Future

Intelligent Cities. https://doi.org/10.1049/ic.2014.0047

[81] Ruggiero S, Iannantuono M, Fotopoulou A, Papadaki D, Assimakopoulos MN, De Masi RF, Vanoli GP, Ferrante A (2022) Multi-objective optimization for cooling and interior natural lighting in buildings for sustainable renovation. Sustain 14:(13):8001. https://doi.org/10.3390/su14138001

[82] Serrano W (2022) iBuilding: Artificial intelligence in intelligent buildings. Neural Comput Appl 34:875–897. https://doi.org/10.1007/s00521-021-05967-y

[83] Hong T, Malik J, Krelling A, O'brien W, Sun K, Lamberts R, Wei M (2023) Ten questions concerning thermal resilience of buildings and occupants for climate adaptation. Build Environ 244:110806. https://doi.org/10.1016/j.buildenv.2023.110806

[84] Albatayneh A, Alterman D, Page A, Moghtaderi B (2018) The impact of the thermal comfort models on the prediction of building energy consumption. Sustain 10:(10):3609. https://doi.org/10.3390/su10103609

[85] Chen H, Dai S, Meng F (2023) Smart building thermal management: A data-driven approach based on dynamic and consensus clustering. Sustain 15:15489. https://doi.org/10.3390/su152115489

[86] Azzi A, Tabaa M, Chegari B, Hachimi H (2024) Balancing sustainability and comfort: A holistic study of building control strategies that meet the global standards for efficiency and thermal comfort. Sustain 16:(5):2154. https://doi.org/10.3390/su16052154

[87] Cantemir E, Kandemir O (2024) Use of artificial neural networks in architecture: Determining the architectural style of a building with a convolutional neural networks. Neural Comput Appl 36:6195–6207. https://doi.org/10.1007/s00521-023-09395-y

[88] Nguyen AT, Pham DH, Oo BL, Santamouris M, Ahn Y, Lim BT (2024) Modelling building HVAC control strategies using a deep reinforcement learning approach. Energy Build 310:114065. https://doi.org/10.1016/j.enbuild.2024.114065

[89] Du Y, Zandi H, Kotevska O, Kurte K, Munk J, Amasyali K, Mckee E, Li F (2021) Intelligent multi-zone residential HVAC control strategy based on deep reinforcement learning. Appl Energy 281:116117. https://doi.org/10.1016/j.apenergy.2020.116117

[90] Deng Z, Wang X, Dong B (2023) Quantum computing for future real-time building HVAC controls. Appl Energy 334:120621. https://doi.org/10.1016/j.apenergy.2022.120621

[91] Michailidis P, Michailidis I, Vamvakas D, Kosmatopoulos E (2023) Model-free HVAC control in buildings: A review. Energies 16:1–45. https://doi.org/10.3390/en16207124

[92] Bogatu DI, Shinoda J, Aguilera JJ, Olesen BW, Watanabe F, Kaneko Y, Kazanci OB (2023) Human physiology for personal thermal comfort-based HVAC control – A review. Build Environ 240:. https://doi.org/10.1016/j.buildenv.2023.110418