Energy & Buildings 210 (2020) 109776

Contents lists available at ScienceDirect

Energy & Buildings

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

Comparing machine learning algorithms in predicting thermal sensation using ASHRAE Comfort Database II

Maohui Luo[a,b], Jiaqing Xie[c], Yichen Yan[d], Zhihao Ke[e,*], Peiran Yu[f], Zi Wang[a,b], Jingsi Zhang[a,*]

[a] School of Mechanical Engineering, Tongji University, Shanghai, China
[b] Center for the Built Environment, University of California, Berkeley, USA
[c] School of Engineering, The University of Edinburgh, Edinburgh, UK
[d] School of Science, Xi'an University of Technology, Xi'an, China
[e] School of Civil Engineering, University of Leeds, Leeds, UK
[f] School of Electronic and Electrical Engineering, University of Leeds, Leeds, UK

ARTICLE INFO

Article history:
Received 21 October 2019
Revised 12 December 2019
Accepted 8 January 2020
Available online 11 January 2020

Keywords:
Building environment
Thermal comfort
Comfort model
PMV

ABSTRACT

Predicting building occupants' thermal comfort via machine learning (ML) is a hot research topic. Many algorithms and data processing methods have been applied to predict thermal comfort indices in different contexts. But few studies have systematically investigated how different algorithms and data processing methods can influence the prediction accuracy. In this study, we first summarized the recent literature from perspectives of predicted comfort indices, algorithms applied, input features, data sources, sample size, training proportion, predicting accuracy, etc. Then, we applied nine ML algorithms and three data sampling methods to predict the 3-point and 7-point thermal sensation vote (TSV) in ASHRAE Comfort Database II. The results show that with an accuracy of 66.3% and 61.1% for 3-point and 7-point TSV respectively, Random Forest (RF) has the best performance among the tested algorithms. Compared to the Predicted Mean Vote (PMV) model, ML TSV models generally have higher accuracy in TSV prediction. Based on feature importance analysis, the air temperature, humidity, clothing, air velocity, age, and metabolic rate are the top six important features for TSV prediction. The RF algorithm can achieve 63.6% overall accuracy in TSV prediction with the top three features, which is only 2.6% lower than involving 12 input features. Further, this paper addressed other common considerations in ML comfort model establishment such as tuning hyperparameters, splitting of training and testing data, and encoding methods. We also provided Python and R programming codes and packages as appendixes, which can be a good reference for future studies.

1. Introduction

1.1. Thermal comfort in buildings

Climate change and energy consumption are among the major urgent issues we humans have to face in the 21st century. In building heating, ventilation, and air-conditioning (HVAC) field, researchers tried to think differently about what thermal comfort conditions should be provided to the occupants [1–3]. To guide the HVAC practice, intensive efforts have been paid on predicting building occupants' thermal comfort. Comfort models like the heat-balance based predicted mean vote (PMV) [4] and the statistical-based adaptive comfort model [5] have attracted great attention. In return, being compiled into building environment evaluation standards [6–8], these two classic models helped a lot in actual practice. However, the current comfort models still have limitations like lacking self-learning or self-correction capability [9], which largely reduced their accuracies when applying them in different contexts such as mixed-mode buildings [10] and changing metabolic rate scenarios [11]. Taking the PMV model, for instance, its accuracy has been a contested topic since early studies from Humphreys [12] and de Dear et al. [13] where they observed that PMV failed to predict thermally neutral temperature in naturally ventilated (NV) buildings.

[*] Corresponding authors.
E-mail addresses: cn16zk@leeds.ac.uk (Z. Ke), zhang_js@tongji.edu.cn (J. Zhang).

https://doi.org/10.1016/j.enbuild.2020.109776

Nomenclature

Abbreviation of algorithms

ANN	artificial neural network
AB	Adaboost
CART	classification and the regression trees
DT	decision tree
ENN	edited nearest neighbors
ELM	extreme learning machine
GPR	Gaussian process regression
GBM	gradient boosting machine
KNC	K-neighbor classifier
KNN	K nearest neighbors
LR	linear regression
LoR	logistic regression
ML	machine learning
MLP	multi-Layer Perception
NB	Naive Bayes
NN	neural networks
RF	random forest
ROS	random over sampling
SVM	supportive vector machine
SMO	synthetic minority oversampling technique edited nearest neighbors

Thermal comfort related abbreviations

AC	air-conditioned buildings
CLO	clothing insulation (Clo)
MET	metabolic rate (met)
NV	naturally ventilated buildings
PMV	predicted mean vote
PPD	percentage of predicted dissatisfaction
RH	relative humidity (%)
SET	standard effective temperature (°C)
TCV	thermal comfort vote
TPV	thermal preference vote
TSV	thermal sensation vote
T_{air}	indoor air temperature (°C)
T_{op}	operative temperature (°C)
T_{out}	outdoor air temperature (°C)
T_r	mean radiant temperature (°C)
V_{air}	air velocity (m/s)

Other symbols

CO_2.	carbon dioxygen concentration (ppm)
HVAC	heating, ventilation, and air-conditioning
MSE	mean square error
R^2	correlation coefficients (r) square

1.2. The era opened by data-driven technology

The advances in machine learning (ML) opened a new approach for thermal comfort prediction to overcome the challenges faced by existing comfort models. Using ML method to predict building occupants' thermal comfort status can be an alternative approach [14]. During the past years, an increasing amount of publications with such purposes have been published, forming a hot research topic. To give a better glimpse of these studies, Table 1 summarizes related publications during the period of 2016–2019, from perspectives of the predicted comfort indices, algorithms, input parameters, data source, sample size, training proportion, and prediction accuracy achieved. The following observations are noteworthy.

First, a wide range of ML algorithms have been applied for thermal comfort prediction. Algorithms such as random forest (RF), neural networks (NN), support vector machine (SVM), K Nearest Neighbors (KNN), and Gradient Boosting Machine (GBM) are frequently used (see the 'ML algorithms' in Table 1). Compared to the PMV model, almost every ML algorithm was able to achieve better prediction performance (see the 'Key finding' and 'Accuracy' columns in Table 1). Even algorithms as basic as Decision Tree (DT) and Logistic Regression (LoR) can have higher accuracy if their hyperparameters were tuned well. This observation suggests that integrating ML algorithms with thermal comfort data analysis is very promising.

Second, thermal comfort indices such as thermal sensation vote (TSV), thermal comfort vote (TCV), and thermal preference vote (TPV) have been targeted as outputs of ML algorithms. Sometimes, the PMV was also used for better HVAC system control [24,25,31,32]. Besides, as shown in the 'Scale units' column of Table 1, the scale units of these indices are other important considerations. Taking TSV for instance, studies like [16,18,23,26] used the 3-point voting scale, wrapping up the original 7-point TSV into a 3-point scale unit. Meanwhile, studies like [17,20,29] insisted to use the 7-point scale, which is commonly used in thermal comfort studies to reflect more detailed information on the degrees of occupants‘ cold or hot sensation.

Third, the input variables varied remarkably among different studies, but most of them have included the same attributes as the PMV model [4] (e.g. air temperature T_{air}, relative humidity RH, air velocity V_{air}, mean radiant temperature T_r, metabolic rate MET, and clothing insulation CLO). Many studies also included other variables like skin temperatures [15,16,18,22,23,26–30] and outdoor climate conditions [17,20,21]. Some studies even included personal characteristics [20,28], background HVAC information, and occupants' integration behaviors with HVAC system [21].

Forth, the data source and sample size between different studies varied significantly. Because the ML comfort model belongs to the data-driven model, a reliable data source is very critical for the model establishment. As shown in Table 1, both laboratory experiments and field investigations are common approaches for data collection. For example, the ASHRAE-RP884 database which was built on a collection of field investigations has been frequently utilized [17,19,33]. Different data sources can influence the sample size remarkably. Usually, laboratory experiments tended to have a much smaller sample size than field studies or datasets excerpted from an existing database.

1.3. The missing gap

When looking at the 'Accuracy' column in Table 1, it's easy to notice that the model accuracies varied remarkably among studies, sometimes even within the same study it varied between algorithms. This leads us to think about the reasons behind these variances.

The first reason can be the variance of ML algorithms. There are basic algorithms such as DT and LR, and more advanced algorithms like RF, SVM, and GBM. Given the fact that building occupants' thermal comfort status is affected by multiple factors and even some random causes [35], algorithms with capabilities to control high dimensions like RF, SVM usually produced higher accuracy than those without them [21,33]. *The second explanation* can be the scale units of the aimed comfort index. Different voting scale units can largely influence the prediction accuracy because it is much easier to predict a 3-category index than a 7-category index. That is why many previous studies preferred to predict the 3-point vote (see the 'Scale units' column in Table 1). *The third reason* can be the input variables and sample sizes. Usually, the combination of physiological variables (such as skin temperature, heart rate, etc.) and environmental variables (such as T_{air}, V_{air}, RH, etc..) can achieve higher accuracy than only using a single type of input features. Also, enough training samples is another critical aspect

Table 1
Machine learning-based thermal comfort model review. (The literature was searched on Google Scholar with ‘machine learning’ and ‘thermal comfort’. The results were arranged in chronological order. 20 papers between 2016–2019 were identified and their related contents were excerpted.)

Reference	Predicted comfort indices	Scale units	ML algorithms	Input variables	Data source	Sample size	Training size	Key finding	Accuracy	Thermal condition
[15]	TCV	3-point TCV • Cold discomfort • Comfort • Warm discomfort	• Gaussian Processing classifier (GPC) • K-neighbor classifier (KNC) • RF classifier SVM classifier	• Local skin temperatures • Clothing surface temperatures • Skin temperature variance • Skin temperature difference between body parts	Lab study with skin temperature and subjective questionnaire	• 288 comfortable • 276 warm discomfort • 44 cold discomfort	24 subjects • 20 for training • 4 for testing	ML-based comfort models have good accuracy and not sensitive to algorithms.	• SVM 87.7% • GP 82.5% • KN 86.0% • RF 87.7%	Transient condition
[16]	TSV	3-point TSV • Cold discomfort • Comfort • Warm discomfort	• Conventional NN • RF	• Physiological parameters like skin temperature, skin conductance, pulse rate, etc.. • Environmental parameters like Tair, Vair, RH, Tr, CLO, etc.	Field study with sensor measurement	800 data samples	• 70% training • 15% validation • 15% testing	ML comfort models have higher accuracy than PMV.	92.9%–93.3%	Steady-state condition
[17]	TSV	7-point TSV • Hot • Warm • Slightly warm • Neutral • Slightly cool • Cool • Cold	• KNN • RF • SVM	• Environmental parameters like Tair, Tout, Vair, RH, Tr, CLO, MET, etc.	ASHRAE RP884	5576 Data samples	• 90% training • 10% testing	ML methods have 5,4%–6% higher accuracy than PMV.	• KNN 49.3% • SVM 48.7% • RF 48.7%	Steady state condition
[18]	TSV	3-point TSV	l RF	• Tair, Vair, CO2, illuminance, health condition, living time in homes	Field study	1040 older people	• 80% training, 20% testing	Extract factors affecting older people’s thermal sensation	• Overall accuracy is 56.6%	Steady-state condition
	TSV	3-point TSV	• RF	• Local skin temperatures in the head, lower arm, upper leg, chest and back	Lab study	18 older people	• 80% training, 20% testing	Identify high-importance local skin temperatures for older people’s thermal sensation	• Overall accuracy is 76.7%	Steady-state condition
[19]	TCV	Continuous TCV	Deep NN	• Tair, Vair, RH, Tr, CLO, Met	ASHRAE RP884	11164 data samples	• 80% training • 20% testing	The model improves thermal comfort prediction accuracy.	MSE is 1.16	Steady-state condition
[20]	TSV	7-point TSV	• Bagging • ANN • SVM	• Tair, Tout, Vair, RH, Tr, CLO, MET, age, gender, building type, degree of thermal environment control	Field study	• 467 for NV • 346 for AC	——	Bagging method has higher TSV prediction accuracy.	R^2 • Bagging 0.50 • ANN 0.49 • SVM 0.45 • PMV 0.41	Steady state condition
[21]	Thermal preference (TP)	3-point scale TP • Warmer • No change • Cooler	• Logistic Regression • (LoR) • DT • GBM • GPC • RF • SVM	• TP and clothing insulation. • Occupant behaviors • Date and Time • Tair, Top, RH, Tout, sky cover, precipitation, etc.	Field study	4743 entries with 67 features	——	Algorithms with capabilities to control high dimensions like RF can produce higher accuracy.	Machine learning models produce a median accuracy of 73% higher than the 51% of PMV.	Steady-state condition

(*continued on next page*)

Table 1 (*continued*)

Reference	Predicted comfort indices	Scale units	ML algorithms	Input variables	Data source	Sample size	Training size	Key finding	Accuracy	Thermal condition
[22]	• TSV • TP	• 5-points TSV • 3-point TP	• Polynomial fit • ANOVA • t-Test	• Facial temperatures	Lab study	• 10 males • 6 females	——	Facial skin temperature can predict one's thermal comfort state.	——	Transient condition
[23]	TSV	3-points TSV	RF	• Hand skin temperature • Change rate of hand skin conductance • Fluctuation of pulse rate • Blood oxygen saturation • Etc.	Lab study	10 males 10 females	• 80% training • 20% validation	RF can predict thermal state with high accuracy by using physiological parameters.	• 93% for male • 94% for female	Transient condition
[24]	PMV	The same as 7-point TSV	SVM	• Tair, Vair, RH, Tr, CLO, Met	Lab study	120 males 75 females	• 793 for training • 18 for testing	SVM can be used for better predicting PMV.	Correlation coefficient between 0.93 and 0.99	Transient condition
[25]	PMV	The same as 7-point TSV	• hybrid SVM • NB • SVM	• Tair, RH, CLO, MET • clock time • duration	3 imaginary subjects	3 imaginary people	• 28–1000 for training • 3000 for testing	SVM has better performance than other algorithms.	72.8–94.6% depends on dataset size	Steady-state condition
[26]	TSV	3-points TSV	• Extreme learning machine (ELM) • SVM	• Skin temperature and its gradient • body surface area • clothing insulation	Lab study	Male: 10, Female: 10	• training 50% • testing 50%	ML thermal sensations models can obtain good accuracy.	65% and 87% for on-normalized and normalized skin temperature, respectively.	Transient condition
[27]	TP	3-point TP	• LR • KNN • RF • SVM	• Tair, RH, CO2 level, Tout, Window states, etc. • Heart rate, skin temperature, MET, CLO	Field study	3 and 7 subjects for single and multiple occupancy case, respectively.	——	The TP model achieves 80% prediction accuracy.	Over 80% for RF.	Steady-state condition
[28]	TSV	5-point TSV	• DT • Stepwise regression	• Local skin temperatures at the forehead, neck, chest, back, arm, belly, waist, wrist. Etc. • Physiological information like body mass index and gender.	Lab study	10 males, 6 females	——	Combining the waist, arm, and wrist skin temperatures can get 95.9% accuracy.	64%–95.9% depending on the number of input skin temperatures.	Transient condition
[29]	TSV	7-point TSV	• Linear Discriminant Analysis • CART • Gaussian NB • KNN • LR • SVM	• Skin temperature, electrodermal activity, heart rate • Top, RH	Field study	5 males 3 females	——	CART has the highest estimation accuracy.	The highest accuracy can be 99.3%.	Transient condition
[30]	TP	3-point TP	RF	• Skin temperatures in the face	Lab study	7 males, 5 females	——	Prediction accuracy can be 85%–92.7 depending on test cases.	85.0–92.7%	Cooling, heating, and steady-state condition

(*continued on next page*)

Table 1 (*continued*)

Reference	Predicted comfort indices	Scale units	ML algorithms	Input variables	Data source	Sample size	Training size	Key finding	Accuracy	Thermal condition
[31]	PMV	The same as 7-point TSV	• Support vector regression • NN • LR	• Tair, RH, Tr	– –	– –	• 90% training • 10% testing	Nonlinear models performed significantly better than linear ones.	– –	Steady-state condition
[32]	PMV	The same as 7-point TSV	• GPR • NN • SVM	• Tair, Vair, RH, Tr, CLO, MET	– –	650 groups data	500 for training 150 for testing	GPR got higher PMV prediction accuracy.	– –	Steady-state condition
[33]	TCV	3-point TCV	• Adaboost • RF • SVM	• Tair, Vair, Tr, CLO, MET, Tout, Age	ASHRAE RP884	Around 12000	– –	SVM can get a prediction accuracy of 76.7% while that of PMV is 35.4%.	• 56.7–76.7% for SVM • 52.1–74.1% for RF • 51.1–61.4% for Adaboost	Steady-state condition
[34]	TSV	7-point TSV	• NN	• Tout, RH. Vair, Tair	Field study	– –	– –	– –	– –	Transient condition

of the data-driven model establishment [21]. *Other factors* such as the sampling method, cross-validation, training, and testing proportion, tuning parameters for each algorithm, evaluation criteria, and so forth, can also influence the ML model's performance.

For better application of ML algorithms in thermal comfort prediction, a systematic comparison examining how the above factors would influence the model performance would be of great value. Moreover, if the original codes of these ML models can be shared publicly, the whole thermal comfort research field can benefit from it because it can avoid unnecessary repetitive coding work for future successors.

1.4. Objectives

To bridge the above missing gap, a systematic comparison was conducted to examine how different algorithms, sampling methods, and other factors would affect the TSV prediction under different voting scale units. The original codes of this work will be provided in Python and R programming so that future researchers can easily access them.

2. Methods

To fulfill the aims, Fig. 1 shows the overall research design. Eight aspects related to ML comfort model development were examined. To implement the study, Fig. 2 shows the flowchart of research procedures. It involves data cleaning, data sampling, model establishment, and model evaluation. It also shows considerations such as the algorithm comparison, TSV voting scale, trading-off the number of input variables and sample size, etc. More explicit descriptions of each aspect are presented as following.

2.1. Supervised learning algorithms and coding languages

Based on the literature summary in Table 1, nine algorithms were frequently used for thermal comfort prediction so that we chose them to do the comparison here (see Fig. 1). Given the limited space, we highlight the main features of each algorithm and introduce the hyperparameters and parameters of each model. Table 2 lists their coding packages in python and R programming, the tuning range for hyperparameters, and the optimal value of each hyperparameter. To obtain optimal hyperparameters, we used the grid searching method by setting the researching range and steps. The original codes were attached as Appendix A and Appendix B.

Linear Regression (LR) is one of the simplest ML regression algorithms. For a high dimensional dataset, linear regression can figure out the intuitive relationship between input and output via coefficient adjustment.

In this study, we took a characteristic function with a_n as the coefficient for each feature and c as a constant. The cost function is the MSE calculated from predicted values and actual TSV. Computer system can tune the coefficients to ensure the global minimum MSE. We adopt 'Scikit-learn Linear_model' package in python and 'lm()' function in R-programming .To optimize the model performance, the coefficients should be higher than 0.7 when selecting features. *Note, the nine algorithms we mentioned do not include LR because it is discarded after comparing classification algorithms.*

Decision Tree (DT) is based on making strategies. The tree structure is outstanding among all nine algorithms and also easy to understand. Since these three algorithms are based on the DT classifier, we do not introduce a lot about DT. DT's set of parameters and hyperparameters are the same as GBM, RF, and Adaboost. Hence the number of trees is approximately 180 while the depth of the tree is around 18.

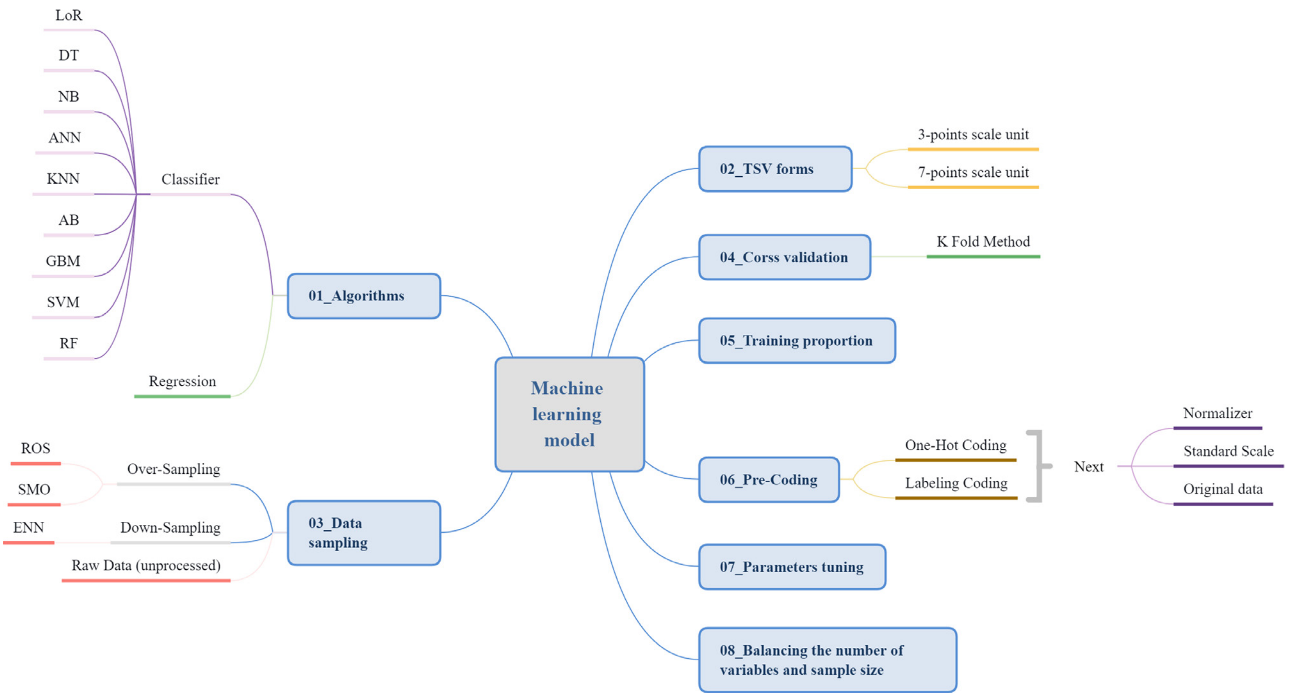

Fig. 1. Overall research design.

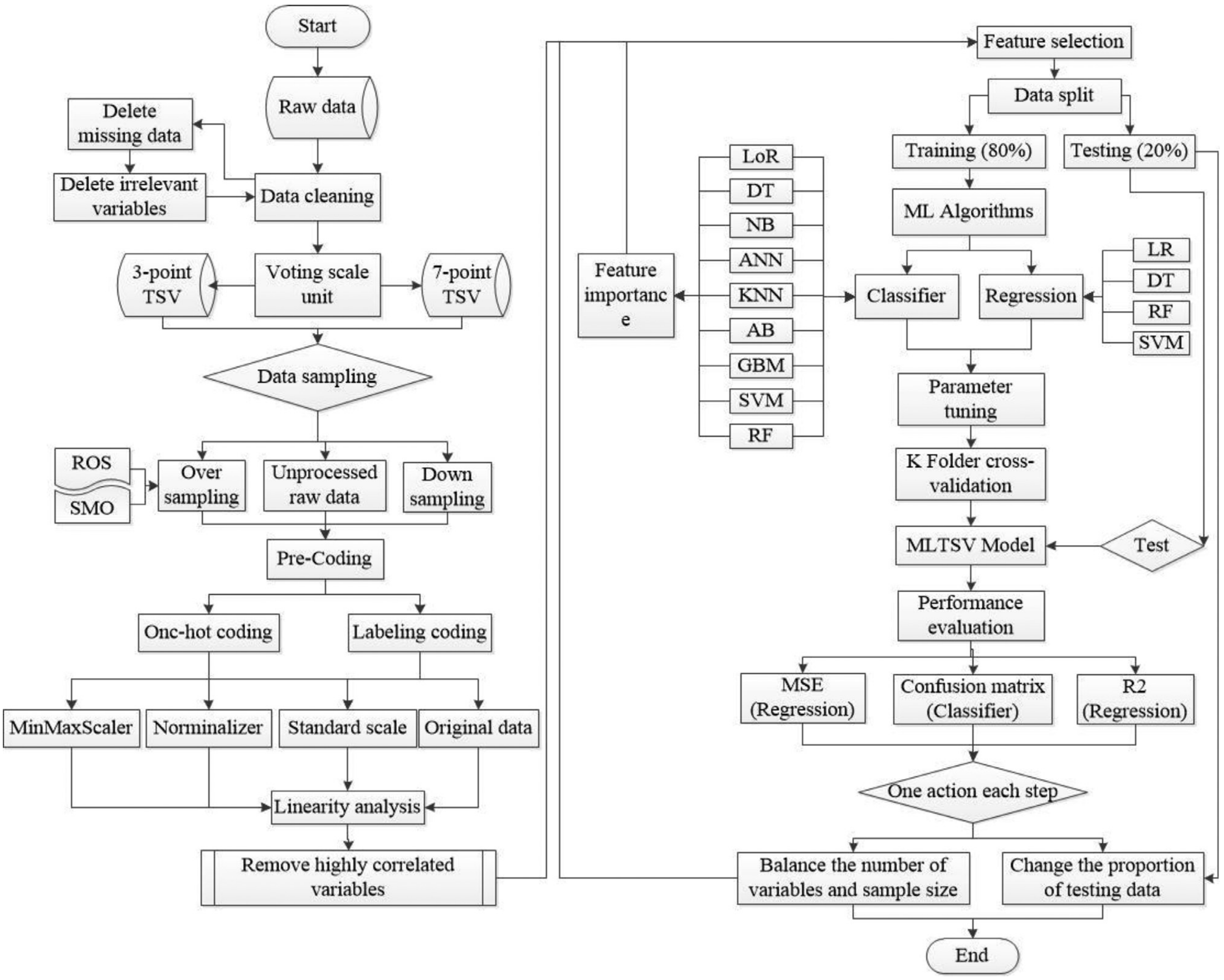

Fig. 2. Flowchart of the study.

Table 2
R and Python packages and key parameters for different algorithms.

Algorithms	Package used		Hyperparameters (Tuning range)	Optimal hyperparameter value	Symbol in R	Symbol in Python
	R	**Python**				
Random Forest (RF)	Random Forest	Scikit-learn Ensemble RandomForestClassifier, RandomForestRegressor	Number of trees (150–200)	180	Ntree	n_estimators
			Depth of the tree (14–20)	18	mtry	max_depth
Support Vector Machine (SVM)	e1071	Scikit-learn SVR, Scikit-learn SVC	Cost (0–2)	0.51	Type, kernel, gamma, cost	C,gamma
			Gamma (0–2)	0.84		
Artificial Neural Network (ANN)	AMORE, neuralnet, nnet,RSNN	Scikit-learn Neural_network MLPClassifier	Hidden layer sizes (10–20)	13	n.neurons, learning.rate.global, hidden.layer, method, hidden, stepmax, algorithm, linear.output, softmax, decay, maxit	hidden_layer_sizes,max_iter, random_state
Gradient Boosting Machine (GBM)	gbm	Scikit-learn Ensemble GradientBoostingClassifier	Number of trees (150–200)	180	Weights, shrinkage, n.cores	learning_rate, n_estimators, max_depth, max_features, random_state
			Depth of the tree (10–20)	18		
Adaboost (AB)	adabag	Scikit-learn Ensemble AdaBoostClassifier	Number of trees (150–200)	180	coeflearn	n_estimators,learning_rate, max_depth
			Depth of the tree (10–20)	10		
Naive Bayes (NB)	e1071, klaR	Scikit-learn Naive_Bayes, GaussianNB	Variable distribution	Gaussian	Laplace, newdata	GaussianNB, MultinomialNB, BernoulliNB
K-Nearest Neighbors (KNN)	Kknn, class	Scikit-learn Neighbors, KNeighborsClassifier	K numbers (20–40)	28	k, distance, kernel	n_neighbors
Linear Regression (LR)	embodied function lm()	Scikit-learn Linear_model, Linear Regression	Coefficients (>0.7)	>0.7	Coefficients, residuals, terms	——
Logistic Regression (LoR)	embodied function glm()	Scikit-learn Linear_model, LogisticRegressionCV	Cost (0–1)	0.36	Familymethod, Hess	Penalty, solver

Logistic Regression (LoR) is mainly suitable for binary classification. The sigmoid function is the main function that turns the output into the range of 0–1. By setting the thresholds, the outputs will be transferred into 0 or 1, so that it can be treated as Bernoulli distribution. The hyperparameter, in this case, is the cost, while the parameter is the threshold. To optimize the model performance, the cost was tuned in the range 0–1. The optimal cost is approximately 0.36.

Support Vector Machine (SVM) aims to find its significant support vectors to make the classification groups as far as possible by counting the distance of a maximum-margin hyperplane between each category. It can be a linear or nonlinear classifier depending on the kernel function.

In this study, we choose the Gaussian kernel as the kernel function. The 'cost' and 'gamma' of Gaussian kernel are the hyperparameters that we set before training the SVM model. Through pilot trials, we found that when 'gamma' and 'cost' were in the range of 0–2, the model got better performance. The optimal gamma is approximately 0.84 while the optimal cost is approximately 0.51. We also found that other parameters of the internal model (such as n_estimator) are less important and they can be automatically tuned by the computer system. We adopt 'Scikit-learn SVR' and 'Scikit learn SVC' package in python and 'e1071' package in R-programming with initialized hyperparameters of 'kernel', 'gamma' and 'cost'.

Artificial Neural Network (ANN) belongs to the neural network. We choose Multi-Layer Perception (MLP) as our main approach since it's the simplest among forwarding propagation methods. Except for input and output layers, MLP also has hidden layers in which each neuron relates to the input layer and output layer.

In this study, we set the hyperparameter 'hidden layers' as 2 for better performance and avoid overfitting. The 'relu' activation function was used. Other parameters like the 'weight' and 'thresholds' were automatically tuned by computer system. We adopt 'Scikit-learn MLPClassifier' package in python and 'AMORE', 'neuralnet', 'nnet' and'RSNN' package in R-programming with the initialized hyperparameters of 'hidden_layer_sizes'. The tested neurons range was 10–20, while 13 neurons got the global maximum accuracy.

Ensemble Learning concentrates on aggregating small and weak models into strong and large models. Even if a few of the sub-classifiers perform bad, other classifiers can fix the gap, thus leading to a better generalization. Generally, it includes Bootstrap Bagging, Boosting, and Stacking methods.

Random Forest (RF) is a representative of Bootstrap Bagging. By put-back sampling features and datasets, the number of N models is measured based on DT. Computer system won't present the best DT model or prune a sub-tree but it will calculate the average vote of each DT model and avoid the problem of a single model. The hyperparameters of RF are tree numbers and tree depth. Parameters are inner thresholds that the system generates to split branches.

We adopt 'Scikit-learn Ensemble' package in python and 'Random Forest' package in R-programming. To optimize the model performance, the number of trees and the depth of trees were tuned in the range of 150–200 and 10–20, respectively. Approximately 180 trees and 18 tree depth can generate a model with global maximum accuracy.

Adaboost(Ab) is the abbreviation of Adaptive Boosting. In the Ab model, the weight of features is refreshed each turn and depends on the accuracy of an input that can be classified correctly. The weight of the features will decrease if the input adapts to the current classifier perfectly, so that a weaker classifier can be improved as a stronger classifier. Because we want to see whether Ensemble Learning performs better than general classifiers, we chose DT classifier as the base classifier. The hyperparameters of Ab are tree numbers and tree depth in this case. Parameters are thresholds of tree nodes.

We adopt 'Scikit-learn Ensemble' package in python and 'adabag' package in R-programming. To optimize the model performance, the number of trees and the depth of trees were tuned in the range of 150–200 and 10–20, respectively. When the accuracy of the model reaches the global maximum, the number of trees is approximately 180 while the depth of the tree is around 18. It is almost the same result as RF has shown.

Gradient Boosting Machine (GBM) is another boosting method. The difference between GBM and Adaboost is that GBM owns a specific cost function based on gradient descent algorithm. With DT as base classifier, its hyperparameters are tree numbers and tree depth. Parameters of the model are thresholds of tree nodes.

We adopt 'Scikit-learn Ensemble' package in python and 'gbm' package in R-programming. To optimize the model performance, the number of trees and the depth of trees were tuned in the range of 150–200 and 10–20 respectively. When the accuracy of the model reaches the global maximum, the number of trees is approximately 180 while the depth of the tree is around 18.

Naïve Bayes (NB) is a simple Bayes algorithm. It assumes that all features are independent of each other. The input should be distributed when applying NB. Three widely used distributions are 'Gaussian', 'Multinomial', and 'Bernoulli', among which the Bernoulli distribution requires input a boolean number. They are also the hyperparameters of the NB model. We adopt 'Scikit-learn Naive_Bayes' package in python and 'e1071', 'klaR' package in R-programming. The Gaussian distribution shows the best performance among three input distributions in this study.

K-Nearest Neighbors (KNN) is a simple supervised learning method. The K, which is the number of nearest neighbors, is a hyperparameter of the KNN model. Assuming the nearest neighbors are classified correctly, KNN counts the number of each group in K neighbors and identify the largest number of the group. It is closely connected with distance theory to tune the data representation. We adopt 'Scikit-learn Neighbors' package in python and 'Kknn', 'class' package in R-programming. To optimize the model performance, K was tuned in the range of 20–40. When K was tuned to 28 approximately, the model got the best performance.

2.2. Data source and TSV forms

To date, there are several publicly available thermal comfort databases including the ASHRAE RP-884 database with 25,616 data samples developed in the 1990s [36], the Smart Control and Thermal Comfort database (SCATs) with 27284 data samples developed in 2000s [37], and the latest ASHRAE Global Thermal Database (marked as the 'Database II' in following sections) with larger sample size of 81,968 [38]. Researchers like Lu [17], Gao [19], Zhou [9] and Farhan [33] have applied ML algorithms to the RP-884 database. Given the timeliness and the integrality of the data, we choose the latest Database II to do the comparison.

Table 3 presents a description of the variables in Database II. TSV was chosen as the targeted comfort index. To examine how its scale units would influence the prediction accuracy, both the 7-point voting and the wrapped-up 3-point voting were predicted. Table 4 shows the sample size of each voting category. 12 variables including the '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. In total, 10618 samples with all the above variables were qualified.

2.3. Data pre-coding

As shown in Table 3, the unit and data distribution vary with input features. There are large numeric numbers (such as T_{air} and T_{out}), small numeric numbers (such as the RH and CLO) and categorical variables (such as sex and season). It is essential to

Table 3
Variable descriptions in Database II

Category	Variables	Unit	Range	Number of samples	Number of missing samples
Comfort indices	***Thermal sensation vote (TSV)***		7-point scale units	79,179	2789
	Thermal preference vote (TP)		Warmer, No change, Cooler	62,195	19,773
	Thermal comfort vote (TCV)		7-point scale units	26,067	55,901
	Thermal sensation acceptability		Acceptable, Unacceptable	53,123	28,845
	Air movement acceptability		Acceptable, Unacceptable	12,741	69,227
	Air movement preference vote		More, No change, Less	36,166	45,802
	Humidity preference vote		More humid, No change, Drier	12,243	69,725
	Humidity sensation vote		7-point scale units	12,540	69,428
	PMV		7-point scale units	48,189	33,779
	PPD	%	5–99	48,189	33,779
Indoor environment	$\boldsymbol{T_{air}}$	°C	0.6–63.2	78,112	3856
	RH	%	0–99.7%	72,907	9061
	$\boldsymbol{V_{air}}$	m/s	0–56.17	70,829	11,139
	T_{op}	°C	6.0–39.4	14,481	67,487
	T_r	°C	4.4–148.1	13,147	68,821
	SET	°C	6.6–61.5	47,974	33,994
Personal factors	***CLO***	clo	0–2.89	74,380	7588
	Met	met	0.7–5	66,968	15,000
	Age	year	6–95	24,863	57,105
	Sex		Male, Female	42,922	39,046
Outdoor environment	$\boldsymbol{T_{out}}$	°C	-18.4–45.1	79,338	2630
	Season		Spring, Summer, Autumn, Winter	81,742	226
	Climate		Koppen climate classification [39]	81,968	0
Building information	***Building operation mode***		AC, NV, Unknown	26,051	55,917
	Building type		Classroom, Others, Office, Multifamily housing, Senior center	81,968	0

Table 4
TSV categories

7-point TSV (NA:2789)		Wrapped 3-point TSV			
Categories	Values	Sample size	Scale units	Values	Sample size
Hot	2.5–3	332	Warm side	0.5–3	2991
Warm	1.5–2.4	751			
Slightly warm	0.5–1.4	1908			
Neutral	-0.5–0.4	5740	Neutral	-0.5–0.4	5740
Slightly cool	-1.5—0.6	1550	Cool side	-3—0.6	1887
Cool	-2.5—1.6	297			
Cold	-3—2.6	40			

unify their formats before analyzing their impact on TSV prediction. There are many pre-coding methods to do this work. We did a comparison to find the proper method for thermal comfort datasets.

For different data formats, we compared the one-hot encoding and label encoding, which are two common methods to deal with textual data. Label encoding transforms the character labels into numerical labels. We used simple integer numbers to label non-numeric features such as ‘Season’, ‘Building’, ‘Mode’ and ‘Age’. The main difference between one-hot encoding and label-encoding is the input dimension. For example, if one sample has value in ‘Autumn’, one-hot encoding will set its ‘Season’ column as ‘1’ and other columns as ‘0’. They will be fixed in a matrix-vector and transformed into an array. One-hot encoding enlarges dimensions while label-encoding maintains the original size of the input dimension.

For different data units, we compared the MinMaxScaler, StandardScaler, and Normalizer to figure out which one is better. These three data preprocessing methods are used to measure the input data at the same scale (level). MinMaxScaler sets a range (normally 0–1) and zooms in or out numeric inputs into this range. StandardScaler treats numeric inputs around 0 and sets the variance as 1, so that the processed data will be more concentrated than unprocessed data. Normalizer transforms input to standard norm equals to 1.

2.4. Data sampling

Given that the raw data of each variable are usually ununiformly distributed, it is necessary to choose the proper sampling method before implementing ML algorithm to avoid unbalanced prediction. For categories with a small sample size that can not reflect important information of its nature, the over-sampling method is often used to amplify the sample size. Contrarily, the down-sampling method is often used to cut down the categories with a large sample size.

To see how different sampling methods would affect the model performance, we compared the following data balancing methods: Random Over Sampling (ROS), Synthetic Minority Oversampling Technique Edited Nearest Neighbors (SMO), Edited Nearest Neighbors (ENN), and the unprocessed Raw Data. More random data can be generated by using ROS. ENN is the method to cut down the number of data amount in dominant features. SMO combines oversampling with down-sampling methods. For sections not about tuning data sampling methods, we adopted the unprocessed raw data distribution.

2.5. K-fold cross-validation method

K-Fold cross-validation is widely used to modify the algorithm parameters during the model establishment. Without the cross-validation, the prediction result would be useless because too much training data can result in overfitting and it can lead to bad generalization. In order to obtain better performance, we varied the K value in k-fold cross-validation within a range of (5–100), trying to find the proper range of K. For sections not about tuning the K values, it was set at 20 for the cross-validation, and this value complies with the results from Section 3.4 and Appendix C.

2.6. Training proportion

The 80% training proportion rule is favored by many researchers (see the ‘Training size’ column in Table 1). But the best training and testing ratio may vary with different datasets. To find the proper training proportion for TSV prediction via indoor environmental parameters and personal characteristics, we tested a wide range of training proportions (0–100%). In sections not about tuning training/testing proportion, we maintained 80% training proportion.

2.7. Model evaluation

Many criteria can be used to evaluate the performance of ML TSV models, such as the stability, robustness, computing cost, accuracy, etc. Given the sample sizes of thermal comfort studies are relatively small when compared to current big data problems [40], we mainly focused on accuracy while ignoring the computing cost. When comparing the accuracy metrics, it is the mean accuracy of each TSV category. For example, for the 3-point TSV, the reported accuracy is the average accuracy of the three TSV categories.

For regression models, we used the Mean Square Error (MSE) and R^2 for evaluation. For classification models, we adopted the prediction accuracy. Additionally, the R^2 and MSE can also reflect the robustness and veracity of the classification model. A better model requires higher R^2, lower MSE, and higher prediction accuracy.

2.8. Research implementation

Given factors such as ML algorithms, data sampling, cross-validation and training/testing proportion can affect model performance, control variable method was adopted. For example, when analyzing the effects of k-fold cross-validation, we fixed other variables such as TSV form and training/testing proportion. Similarly, when tuning the training/testing proportion, we used the 20-fold cross-validation. The 80-20 training/testing proportion and 20-fold cross-validation were chosen because they are empirical values from literature and achieved good performance in our pilot study.

First, we cleaned the raw data in Database II, determined 12 input features, and targeted at TSV prediction. Rather than replacing the missing values with synthetic data, all the samples with one or more missing values were deleted to ensure the quality of datasets. *Then,* the qualified 10619 samples were treated as two different datasets: the 3-point and 7-point TSV cases (see Section 3.2). To balance the TSV distribution on each voting category, we compared three common balancing technologies plus the unprocessed raw data (see Section 3.3). *Simultaneously,* two precoding methods (one-hot-coding and labeling coding) were used to uniform the forms of variables. The normalizer and standard scale were adopted to improve the generalization ability of ML model (see Section 3.6). *After that,* linearity analysis was used to remove highly correlated variables (see Fig. 12). *After feature selection,* the datasets were split into training and testing data (see Section 3.5). 9 classifier algorithms were applied for further comparison (see Section 3.1). The key parameters and hyperparameters were tuned to optimize the model performance for each algorithm (see sect5ion 3.7). The K value in cross-validation and training data proportions were varied to examine their influences (see Section 3.4). *Furthermore,* the trade-off between the number of input features and the sample size was discussed (see Section 4.1).

3. Results

3.1. Effects of algorithms

Regression model. Fig. 3 shows the accuracy of different ML regression models. Among the four commonly used algorithms, the RF regression achieved the highest R^2 value (0.30) and the lowest MSE value (0.32), which means it has better TSV prediction than the other three. However, even for the RF regression, its R^2 value was not high enough (R^2 >0.8) to conclude a good correlation between the actual and predicted TSV, which is because of the disperse distribution of raw TSV data. The correlation coefficient can be very low without averaging the TSV in a certain temperature bin. This phenomenon has been observed in previous thermal comfort studies such as [41,42]. For this reason, we only focus on classifier models in following sections.

Classifier model. Fig. 4 compares the performance of nine algorithms. The box ranges were resulted from parameter tuning and will be discussed in Section 3.7. The black circles beyond the box represent outliers that didn’t obey the majority distribution principle. Each box range reflects the robustness and stability of the algorithm. The narrower the box, the more stable the model performance is.

Algorithms of RF, GBM, Adaboost, KNN, and ANN produced over 60% accuracy, with the best performance (66.2%) coming from the RF algorithm. Compared to these five algorithms, the SVM and DT’s accuracy was relatively lower, ranging from 56% to 59%. The NB model has the lowest accuracy and depends on which distribution method was used. By using Gaussian distribution, polynomial distribution, or Bernoulli distribution, the NB model can get an accuracy of 49.8%, 52.5%, and 54.6%, respectively. However, it’s worthy to mention that the classic PMV model only got an overall accuracy of 42.5% for 3-point TSV prediction, which is lower than all the nine tested ML algorithms.

3.2. Effects of TSV voting scales

Fig. 5 compares the accuracy of 3-point and 7-point TSV prediction. Not surprisingly, the 3-category prediction has constantly higher accuracy than the 7-category ones. The accuracy range of these two predictions is 60–66.2% (excluding NB and ANN) and 52–57% (excluding NB), respectively.

Although using the original 7-point TSV without wrapping it into 3-point scale, ML algorithms can still achieve pretty high accuracy, higher than that of 3-point PMV prediction (which is 42.5%). The 3-point TSV is easier to be predicted, but it also sacrifices much nuanced useful information that can distinguish different degrees of building occupants’ cold/hot sensation. In real practice, one should determine whether to pursue higher TSV prediction accuracy or to capture more detailed information on occupants’ thermal response. Both pursuits have practical meaning. An example scenario is the air conditioning system control, where we only need to give a warmer or cooler commander to the system. In this context, a 3-point TSV can fulfill the requirement and ML algorithms like RF, GBM can do this job very well.. Another application scenario is the temperature design where we need to know the temperature demand in 1 or 0.5 °C precise. For this case, a continuous TSV scale would be better because 1 scale unit TSV change on a 7-point voting scale is equivalent to 3°C air temperature change [43]. In this context, the classifier ML algorithms may not suitable.

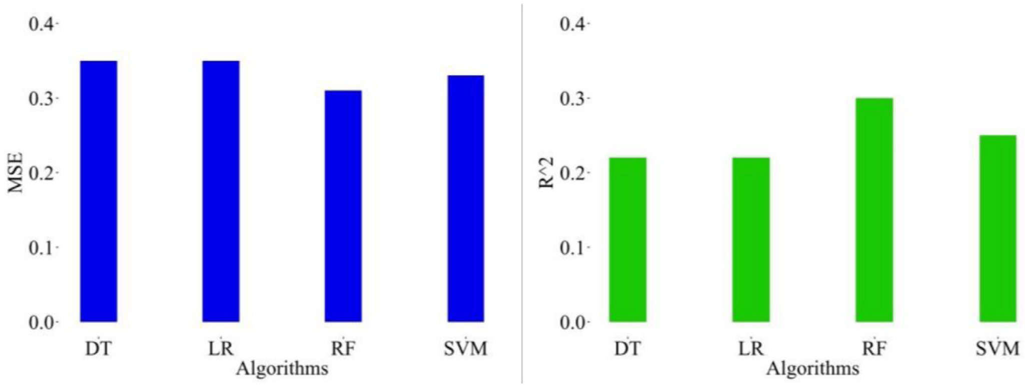

Fig. 3. The accuracy of regression models. Note, the results are from 3-point TSV prediction.

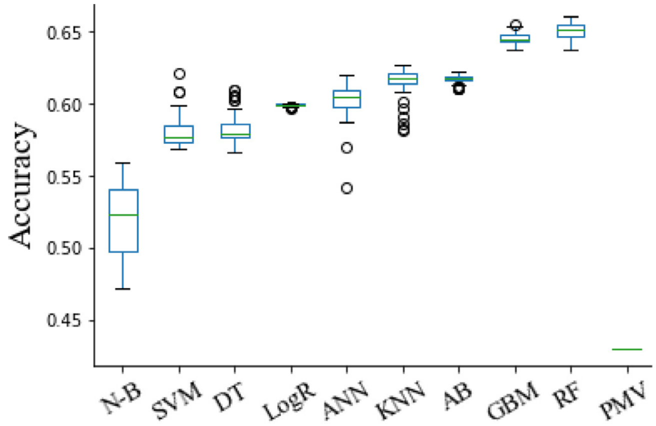

Fig. 4. The accuracy of classifier models. Note, the results are from 3-point TSV prediction. And the PMV accuracy is from binned group results.

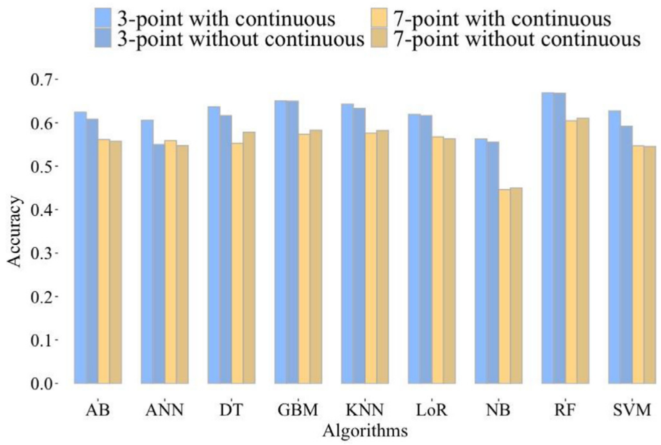

Fig. 5. The accuracy for 3-point TSV and 7-point TSV prediction. (The legend 'With continuous' means including TSV votes on the continuous scale, otherwise it excluded those votes.)

3.3. Effects of sampling method

Fig. 6 orders the prediction performance of each algorithm combining with different sampling (data balancing) methods. Among the combinations, the RF algorithm with unprocessed raw data gets the highest accuracy of 66.2% for 3-point TSV prediction and 61.1% for 7-point TSV prediction. If we group the results by different algorithms, there is a general trend that 'unprocessed Raw Data > ROS > SMO > ENN'. But for some combinations, such as the RF (ENN), SVM (ENN), and KNN (ENN), they achieved higher accuracy for 7-point TSV prediction than that of 3-point TSV prediction. That means proper data balancing technology indeed has the potentiality to remarkably enhance the ML model's performance. Also, it is worth to note that PMV only got 31.2% and 43% accuracy for the 7-category and 3-category TSV, respectively, which is lower than most of the ML combinations.

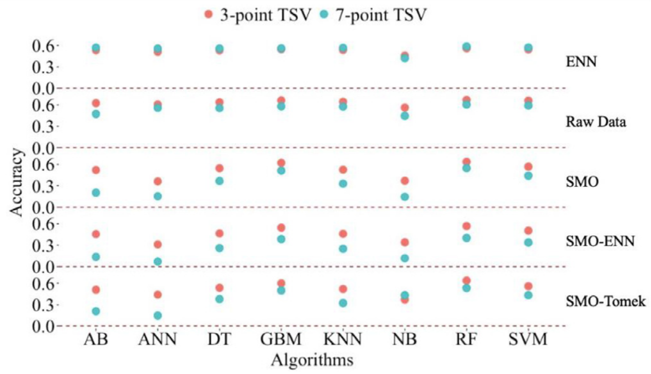

Fig. 6. The accuracy of different sampling methods and algorithm combinations.

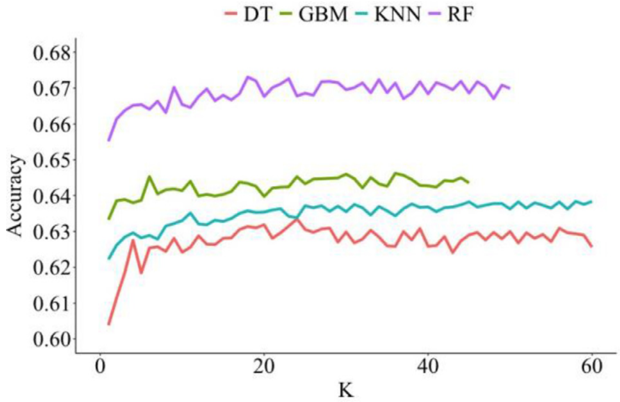

Fig. 7. The effects of K-fold method.

3.4. Effects of K fold methods

Fig. 7 shows the accuracy for the DT, GBM, KNN, and RF algorithms with different K values in K-fold cross-validation. With the increasing K values, the model performance improved gradually. Although there exist some fluctuations in the tendency line, the overall increasing trends are apparent. Generally, one can expect

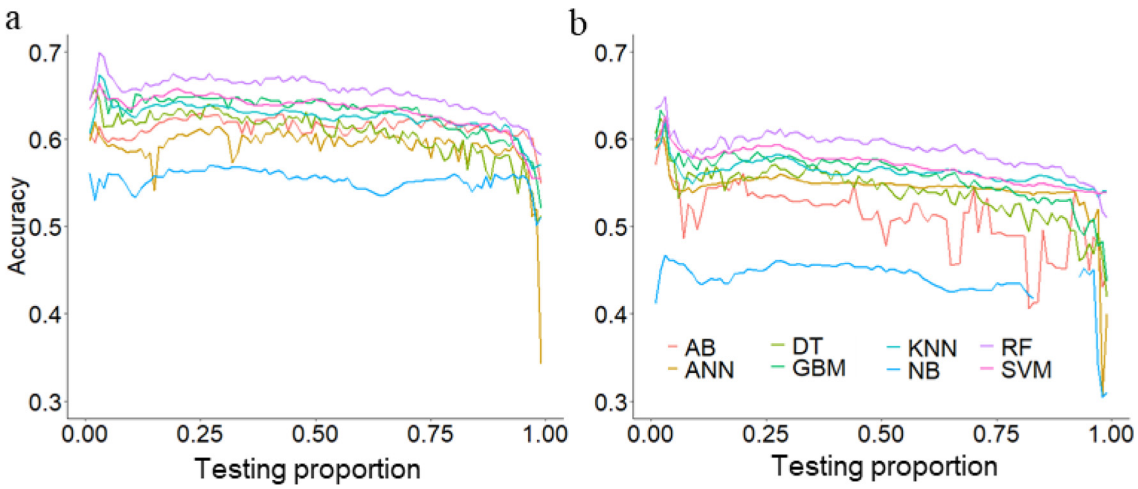

Fig. 8. Accuracy changes with testing proportions: (a) 3-point TSV prediction, (b) 7-point TSV prediction.

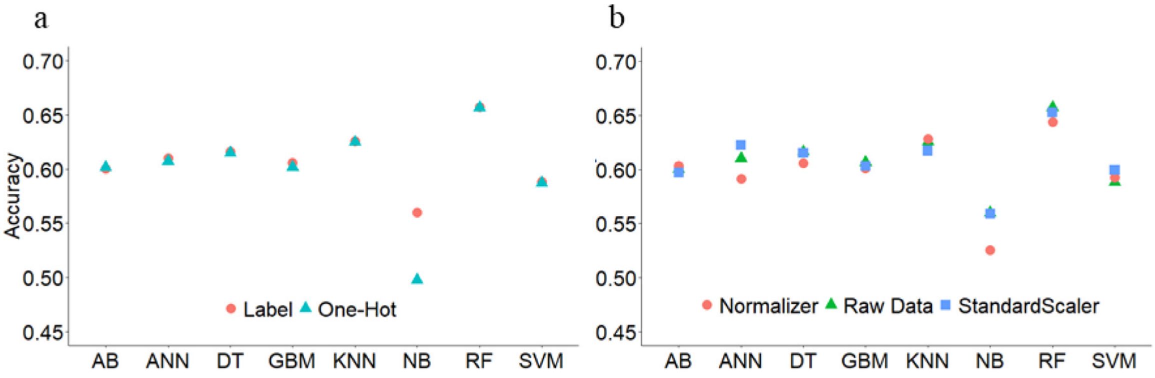

Fig. 9. Accuracy comparison of different pre-coding methods on 3-point TSV prediction: (a) encoding methods, (b) continuous feature processing method.

that once the K value reached 10 or larger, prediction accuracies tend to be stabilized. Together with the discussion in Appendix C, it is recommended to use K values equal or around 20 for the ML comfort model's cross-validation.

3.5. Effects of training proportion

Many studies (see the 'Training size' column in Table 1) adopted the 80% training data rule. Fig. 8 shows how different training proportions would affect TSV prediction accuracy. When looking at the figure, only testing size larger than 10% should be considered, because the prediction could be distorted by random factors if the testing data is less than 10%. Generally, the algorithms can reach peak accuracy when its testing size is around 20–30% (or 70–80% training proportion) for both 3-point and 7-point TSV predictions. With too many testing data (or too little training data), the accuracies decrease continuously.

3.6. Effects of pre-coding

For ML algorithms, encoding methods on discrete and continuous variables will significantly influence the model performance. Fig. 9 (a) shows how different encoding methods affect ML algorithms' prediction accuracy. The label encoding always (for the seven selected algorithms) achieves higher accuracy than the one-hot encoding method.

To process the different variable ranges and different units, the performance of three methods (standard scaler, raw data, and normalizer) varied with ML algorithms. From the perspective of average accuracy, the raw data and standard scaler got accuracies around 60%, while the normalizer method had slightly lower accuracy.

3.7. Effects of hyperparameter tuning

Fig. 10 shows the prediction accuracy and model stability when tuning the parameters in GBM, RF, and SVM algorithms, respectively. *For the GBM model* (Fig. 10 (a), the 3-point TSV prediction accuracy varied within the range of 55.7–66.7% when the number of trees changed from 10 to 20 and the tree depth changed from 1–15. Once the tree depth was larger than 5, its accuracies can be maintained above 63%. *For the RF model* (Fig. 10 (b)), its accuracy varied within the range of 64.0–66.0% when the number of trees changed from 100 to 200 and the tree depth changed from 10–19. The best accuracy was achieved when the tree depth was in the range of 10–20. Once the tree depth was determined, the number of trees causes little difference in the accuracy when it is in the range of 100–200. *For the SVM model* (Fig. 10 (c) and (d)), the 'gamma' has less effect on the model performance than the 'cost'. When gamma and cost increase from 1 to 10 and 1 to 4 respectively, the accuracy decreases from 61.0% to 57.7%. However, this doesn't mean the smaller the 'gamma' and 'cost' the better. Fig. 10(d) makes up the missing of small 'gamma' and 'cost' with a smaller step of 0.01. The maximum accuracy occurred when the 'gamma' was 0.84 and the 'cost' was 0.51. It is suggested that both the range and the tuning steps matter when tuning the parameters.

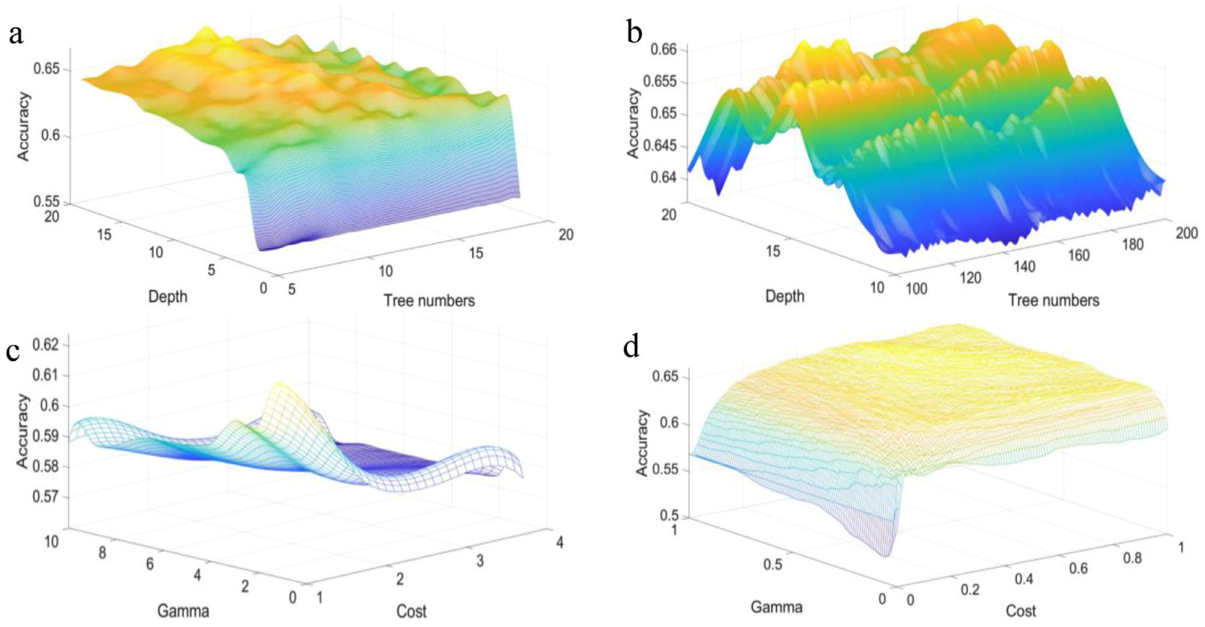

Fig. 10. The accuracy changes for parameter tuning. (a) GBM tuning; (b) RF tuning; (c)SVM gamma changed from 0 to 10 with a 0.1 step, cost changed from 0 to 4 with a 0.1 step; d) SVM gamma and cost changed from 0 to 1 with a 0.01 step. (Note, the grid search is used to tune the hyperparameters. The k number for validation is fixed at 20 and only performance on test data is shown.)

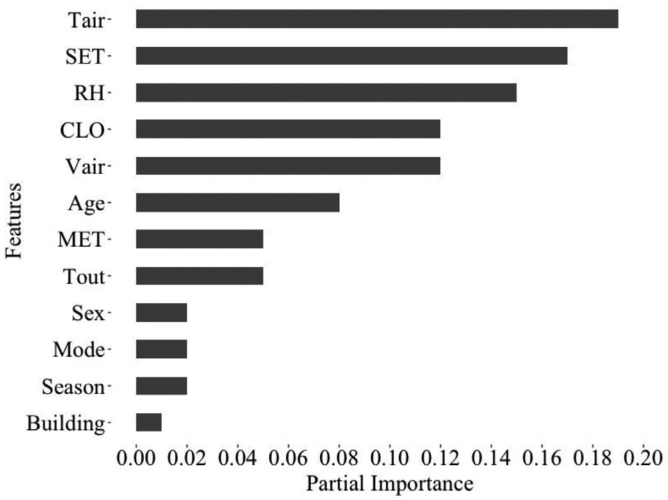

Fig. 11. Feature importance analysis in RF algorithm. The larger the importance coefficient, the more important the feature is.

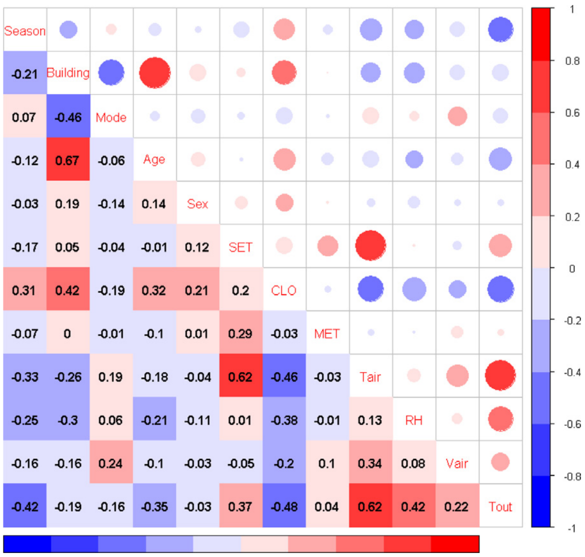

Fig. 12. The correlation coefficients between different variables.

4. Discussion

4.1. Balancing the numbers of input features and the number of data samples

When establishing ML comfort models, it is common to compromise the number of input variables and the sample size. Normally, increasing the input variables and the sample size can improve the model performance (see Table 1) but that means more sensors are needed for data collection. To better understand this issue, we reduced the input variables and varied the sample size within the RF algorithm to see how different combinations can affect the model performance.

Fig. 11 shows the feature importance of the 12 variables we identified for the RF TSV model. The top 7 significant variables include the T_{air}, SET, RH, CLO, V_{air}, Age, and MET, while other variables like T_{out}, Season, Mode, Sex, and Building are less important. Among the top important variables, except the 'Age' and 'SET' are new factors, the rest five are the same with PMV inputs. As the 'SET' itself is a comprehensive thermal comfort index like PMV and it is highly correlated with T_{air} and other variables (see Fig. 12), it was first deleted before removing the other less important variables. Also, one should note that the radiant temperature T_r has been shown as an important attribute to thermal comfort [44,45], but the Database II has limited samples with T_r, making it difficult to investigate its impact. Moreover, it worthy to be noted that 'Sex' was not among the top features determining occupants' thermal sensations. There are many studies debating on this topic, with contradictory results. Some reported that females are thermally more sensitive [46] while others found that there is no sex difference [47]. The present data-driven approach supports the opinion

Table 5
3-point TSV prediction accuracies for different input variables and sample sizes in the RF algorithm with 200 trees and 20 depth.

No.	Input features	Fixed sample size (10618)	Varying sample size				
		Raw Data	Sample size*	Raw Data	ROS	SMO	ENN
12	T_{air}, SET, RH, CLO, V_{air}, Age, MET, T_{out}, Season, Mode, Sex, Building	0.662	10619	0.662	0.639	0.637	0.524
11	T_{air}, RH, CLO, V_{air}, Age, MET, T_{out}, Season, Mode, Sex, Building	0.679	19662	0.639	0.637	0.523	0.508
10	T_{air}, RH, CLO, V_{air}, Age, MET, T_{out}, Season, Mode, Sex	0.668	19182	0.650	0.645	0.642	0.511
9	T_{air}, RH, CLO, V_{air}, Age, MET, T_{out}, Season, Mode	0.664	20816	0.650	0.650	0.637	0.533
8	T_{air}, RH, CLO, V_{air}, Age, MET, T_{out}, Season	0.667	20816	0.651	0.651	0.636	0.538
7	T_{air}, RH, CLO, V_{air}, Age, MET, T_{out},	0.663	20897	0.660	0.640	0.634	0.513
6	T_{air}, RH, CLO, V_{air}, Age, MET	0.657	20915	0.646	0.632	0.627	0.486
5	T_{air}, RH, CLO, V_{air}, Age	0.642	22334	0.645	0.624	0.607	0.482
4	T_{air}, RH, CLO, V_{air}	0.629	64193	0.528	0.500	0.442	0.347
3	T_{air}, RH, CLO	0.636	67504	0.507	0.490	0.439	0.329
2	T_{air}, RH	0.619	71937	0.501	0.484	0.433	0.335
1	T_{air}	0.602	75366	0.480	0.422	0.389	0.281

* Note, when varying the sample size, it was not changed casually, instead, it was determined by the number of samples with all the input features available (see the 'Input features' column in Table 5). Reducing input variables can increase the qualified samples.

that gender is not a vital factor for thermal sensation prediction, but this may change when predicting TCV or TPV.

Table 5 shows the accuracy changes when reducing input variables with and without varying the sample sizes. If the sample size was fixed at 10618, the model accuracy decreased slightly from 66.2% to 60.2% when the input variables decreased from 12 to 1. Even only use the T_{air}, the TSV prediction accuracy can still be as high as 60.2%. If the sample size varied (see the 'Varying sample size' column in Table 3), the accuracy decreased rapidly from 66.2% to 48%, and this drop can't be compensated by the sampling method. This suggests that larger sample size does not always guarantee higher prediction accuracy, especially for datasets with enough samples, too much training may cause overfitting issues. Another potential reason for this can be data quality. The 10618 samples with all the 12 variables are high-quality data in Database II. Adding other inferior quality samples can increase the difficulty of TSV prediction. However, it should be noted that no matter the sample size change or not, if the top five important variables (T_{air}, RH, CLO, V_{air}, Age) were included, the RF model can achieve over 64% accuracy.

4.2. Other considerations for data-driven comfort model

According to the above results, all the tested ML classifiers can achieve higher accuracy in TSV prediction than the PMV, which only got 31.2% and 43% for 7-point and 3-point TSV respectively (see Fig. 7). The optimized ML algorithms can predict 3-point TSV at 60–66.2% accuracy and 7-point TSV at 50–61.1% accuracy. These results suggest the prospect of predicting building occupants' thermal comfort status through the ML approach, but before getting to that state, the following considerations are worth to be further investigated.

The first one is the targeted comfort indices. The current study mainly investigated the TSV prediction. Other comfort indices such as TCV and TPV also need investigation. We applied the nine algorithms for TCV and TPV prediction. Averagely, the accuracy was in a range of 60–70% for 3-point TCV and 65–80% for TPV prediction. This observation together with the literature summary in Table 1 collectively supports a conclusion that integrating the ML method with building occupants' thermal comfort data analysis is promising. *The second issue is the input variables.* The current study used the indoor environment parameters plus building and climate information to predict TSV and got pretty good results (see Fig. 11 and Table 5). Although we discussed the influence of the number of inputs and sample size, the type of variables was largely limited by the database itself. According to some literature in Table 1, diversify input sources, such as combining the physiological measurements and indoor environment parameters, can help thermal comfort prediction. In the future, studies comparing different data sources, especially those involving different types of variables, are needed to be produced. *The third concern is about group prediction versus individual prediction.* The group prediction uses ML algorithms to predict the average of a group of people. The study described in Zhou et al. [9] is such an example. The individual prediction tries to use the ML algorithm to predict individual (or personal) thermal comfort [48]. Kim et al. [21] utilized building occupants' interaction behaviors as an input variable to predict individual heating or cooling demand. Dai et al. [49] took local skin temperatures to predict personal thermal sensation. All these studies provided empirical evidence to build more advanced thermal comfort models via data-driven approaches.

4.3. Limitation and suggestions for future data-driven comfort models

Although we tried to do a systematic comparison for ML comfort model development, there are too many aspects worthy of further investigation. For example, the current analysis (Fig. 12) suggests factors like building type, building operation mode, and climate conditions are not among the top factors, but these factors can indeed influence occupants' thermal perception [9,10]. In futures studies, data-driven comfort models in different contexts are worthy to be investigated. Moreover, this study only compared limited algorithms (based on the literature review in Table 1) while there exist many other more sophisticated and more advanced algorithms. In future studies, it is worthy to explore how other data analytics such as deep learning and reinforced learning work for thermal comfort prediction. Finally, the original codes of different ML algorithms have been provided in python and R programming (see Appendix A and B respectively). It is encouraged for future ML comfort models to share their codes and data publicly.

5. Conclusions

In this study, a comparison of machine learning algorithms for thermal sensation prediction was conducted. The following findings are noteworthy.

(1) Among the nine classifier algorithms, those with capabilities to control high dimensions (such as the RF, ANN, GBM, etc.) can achieve relatively better TSV prediction accuracy than other algorithms without them. But even the most basic ML algorithms like NB and DT can get better TSV prediction than the conventional PMV.

(2) ML algorithms can predict the 3-point TSV with an accuracy of 60–66.2% while predicting the 7-point TSV at 50–61.1% accuracy. The 3-point scale unit is relatively easier to be predicted

but it also sacrifices nuanced information helping to distinguish the degrees of building occupants' cold/hot sensation. In real practice, one should determine whether to pursue higher TSV prediction accuracy or to capture more detailed information on occupants' thermal response.

(3) Among the input variables in the Database II, T_{air}, RH, CLO, V_{air}, Age, and MET are the top six important inputs for RF TSV model. Except for the 'Age', the rest five are the same with PMV inputs. Using these top six features, the RF algorithm can produce an accuracy of around 66% with or without varying sample size, which is very close to the 66.2% accuracy with 12 input features.

(4) When developing a machine-learning-based data-driven thermal sensation model, there are many other factors should be considered. For example, the K-fold cross-validation with K values equal or around 20 and 20–30% testing proportion is recommended. Also, for large datasets like ASHRAE Comfort Database II, the raw data distribution may achieve higher performance than other balancing methods.

Declaration of competing interest

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

CRediT authorship contribution statement

Maohui Luo: Conceptualization, Visualization, Validation, Writing - original draft, Writing - review & editing. **Jiaqing Xie:** Data curation, Writing - original draft. **Yichen Yan:** Data curation, Writing - original draft. **Zhihao Ke:** Data curation, Writing - original draft. **Peiran Yu:** Data curation. **Zi Wang:** Validation, Writing - review & editing. **Jingsi Zhang:** Conceptualization, Visualization, Writing - review & editing.

Acknowledgment

This study was funded by the National Natural Science Foundation of China (51908414), and China Postdoctoral Science Foundation (2017M620442 and 2019M651584). The authors would like to thank Prof. Edward Arens and Hui Zhang at the University of California, Berkeley for their constructive comments.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.enbuild.2019.109753.

References

[1] Y. Zhu, Q. Ouyang, B. Cao, et al., Dynamic thermal environment and thermal comfort, Indoor Air. 26 (1) (2016) 125–137.

[2] H. Zhang, E. Arens, Y. Zhai, A review of the corrective power of personal comfort systems in non-neutral ambient environments, Build. Environ. 91 (2015) 15–41.

[3] F. Zhang, R. de Dear, P. Hancock, Effects of moderate thermal environments on cognitive performance: a multidisciplinary review, Appl. Energy 236 (2019) 760–777.

[4] P.O. Fanger, Thermal Comfort, Danish Technical Press, Copenhagen, 1970.

[5] R. de Dear, G. Brager, The adaptive model of thermal comfort and energy conservation in the built environment, Int. J. Biometeorol. 45 (2001) 100–108.

[6] Technical Committee ISO/TC 159 and Technical Committee CEN/TC 122. ISO 7730, in: 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, European Committee for Standardization, UK, 2005, p. 2006.

[7] ASHRAE, Thermal Environmental Conditions for Human Occupancy, ASHRAE Standard 55, American Society of Heating, Refrigerating and Air-Conditioning Engineers, Atlanta, Georgia, 2013.

[8] CEN, Indoor environmental input parameters for design and assessment of energy performance of buildings: addressing indoor air quality, Thermal Environment, Lighting and Acoustics, European Committee for Standardization, Brussels, 2007 EN 15251.

[9] X. Zhou, L. Xu, J. Zhang, et al. Data-driven Thermal Comfort model via Support Vector Machine Algorithms: Insights from ASHRAE RP-884 Database. Manuscript submitted to Energy and Buildings.

[10] M. Luo, B. Cao, J. Damiens, et al., Evaluating thermal comfort in mixed-mode buildings: a field study in a subtropical climate, Build. Environ. 88 (2015) 46–54.

[11] Y. Zhai, S. Zhao, L. Yang, et al., Transient human thermophysiological and comfort responses indoors after simulated summer commutes, Build. Environ. 157 (2019) 257–267.

[12] M. Humphreys, The Dependence of comfortable temperatures upon indoor and outdoor climates, Stud. Environ. Sci. 10 (1981) 229–250.

[13] G. Brager, R. de Dear, Thermal adaptation in the built environment: a literature review, Build. Environ. 27 (1) (1998) 83–96.

[14] J. Kim, S. Schiavon, G. Brager, Personal comfort models–a new paradigm in thermal comfort for occupant-centric environmental control, Build. Environ. 132 (2018) 114–124.

[15] A. Cosma, R. Simba, Machine learning method for real-time non-invasive prediction of individual thermal preference in transient conditions, Build. Environ. 148 (2019) 372–383.

[16] T. Chaudhuri, Y. Soh, H. Li, L. Xie, A feedforward neural network based indoor–climate control framework for thermal comfort and energy saving in buildings, Appl. Energy 248 (2019) 44–53.

[17] S. Lu, W. Wang, C. Lin, E. Hameen, Data-driven simulation of a thermal comfort-based temperature set-point control with ASHRAE RP884, Build. Environ. 156 (2019) 137–146.

[18] Z. Wang, H. Yu, M. Luo, et al., Predicting older people's thermal sensation in building environment through a machine learning approach: Modelling, interpretation, and application, Build. Environ. (2019) 161.

[19] G. Gao, J. Li, Y. Wen. Energy-efficient thermal comfort control in smart buildings via deep reinforcement learning. arXiv:1901.04693v1.

[20] Z. Wu, N. Li, J. Peng, H. Cui, P. Liu, H. Li, X. Li, Using an ensemble machine learning methodology-Bagging to predict occupants' thermal comfort in buildings, Energy Build. 173 (2018) 117–127.

[21] J. Kim, Y. Zhou, S. Schiavon, P. Raftery, G. Brager, Personal comfort models: predicting individuals' thermal preference using occupant heating and cooling behavior and machine learning, Build. Environ. 129 (2018) 96–106.

[22] D. Li, C. Menassa, V. Kamat, Robust non-intrusive interpretation of occupant thermal comfort in built environments with low-cost networked thermal cameras, Appl. Energy. 251 (2019) 113336.

[23] T. Chaudhuri, D. Zhai, Y. Soh, H. Li, L. Xie, Random forest based thermal comfort prediction from gender-specific physiological parameters using wearable sensing technology, Energy Build. 166 (2018) 391–406.

[24] A. Megri, I. Naqa, Prediction of the thermal comfort indices using improved support vector machine classifiers and nonlinear kernel functions, Indoor Built Environ. 25 (2014) 6–16.

[25] B. Peng, S. Hsieh. Data-driven thermal comfort prediction with support vector machine. Proceeding of the 12th International Manufacturing Science and Engineering Conference. 4-8 June 2017, Los Angles, USA. Volume 3.

[26] T. Chaudhuri, D. Zhai, Y. Soh, H. Li, L. Xie, Thermal comfort prediction using normalized skin temperature in a uniform built environment, Energy Build. 159 (2018) 426–440.

[27] D. Li, C. Menassa, V. Kamat, Personalized human comfort in indoor building environments under diverse conditioning modes, Build. Environ. 126 (2017) 304–317.

[28] J. Choi, D. Yeom, Study of data-driven thermal sensation prediction model as a function of local body skin temperatures in a built environment, Build. Environ. 121 (2017) 130–147.

[29] F. Salamone, L. Belussi, C. Currò, L. Danza, et al., Integrated method for personal thermal comfort assessment and optimization through users' feedback, IoT Mach. Learn. A Case Study Sens. 18 (2018) 1602.

[30] D. Li, C. Menassa, V. Kamat, Non-intrusive interpretation of human thermal comfort through analysis of facial infrared thermography, Energy Build. 176 (2018) 246–261.

[31] W. Zhang, F. Liu, R. Fan, Improved thermal comfort modeling for smart buildings: a data analytics study, Int. J. Electr. Power Energy Syst. 103 (2018) 634–643.

[32] S. Bin, W. Yan, Application of Gaussian process regression to prediction of thermal comfort index, Proceeding of the 11th International Conference on Electronic Measurement and Instruments, 2013 16-19 August.

[33] A. Farhan, K. Pattipati, B. Wang, P. Luh, Predicting individual thermal comfort using machine learning algorithms, Proceeding of International Conference on Automation Science and Engineering, 2015 08 October.

[34] M. Collotta, A. Messineo, G. Nicolosi, G. Pau, A dynamic fuzzy controller to meet thermal comfort by using neural network forecasted parameters as the input, Energies 7 (2014) 4727–4756.

[35] Z. Wang, R. de Dear, M. Luo, et al., Individual difference in thermal comfort: a literature review, Build. Environ. 138 (2018) 181–193.

[36] R. de Dear, A global database of thermal comfort field experiments, ASHRAE Trans. 104 (1998) 1141–1152.

[37] K.J. McCartney, J.F. Nicol, Developing an adaptive control algorithm for Europe: results of the SCATs Project, Energy Build. 34 (6) (2002) 623–635.

[38] V. Foldvary, T. Cheung, et al., Development of the ASHRAE global thermal comfort database II, Build. Environ. 142 (2018) 502–512.

[39] M. Kottek, Jurgen Grieser, C. Beck, et al., World map of the Koppen-Geiger climate classification updated, Mateorologische Zeitschrift 15 (3) (2006) 259–263.

[40] C. Chen, C. Zhang, Data-intensive applications, challenges, techniques and technologies: a survey on Big Data, Inf. Sci. 275 (2014) 314–347.

[41] M. Indraganti, R. Ooka, H. Rijal, Thermal comfort in offices in summer: Findings from a field study under the 'setsuden' conditions in Tokyo, Japan, Build. Environ. 61 (2013) 114–132.

[42]] Y. Zhang, H. Chen, Q. Meng, Thermal comfort in buildings with split air–conditioners in hot-humid area of China, Build. Environ. 64 (2013) 213–224.

[43] Maohui Luo, Edward Arens, Hui Zhang, et al., Thermal comfort evaluated for combinations of energy-efficient personal heating and cooling devices, Build. Environ. 143 (2018) 206–216.

[44] C. Karmann, S. Schiavon, F. Bauman, Thermal comfort in buildings using radiant vs. all-air systems: a critical literature review, Build. Environ. 111 (2016) 123–131.

[45] X. Zhou, Y. Liu, M. Luo, Thermal comfort under radiant asymmetries of floor cooling system in 2h and 8h exposure durations, Energy Build. 188 (2019) 98–110.

[46] N. Gerrett, Y. Ouzzahra, S. Coleby, et al., Thermal sensitivity to warmth during rest and exercise: a sex comparison, Eur. J. Appl. Physiol. 114 (2014) 1451–1462.

[47] D. Meh, M. Denišlič, Quantitative assessment of thermal and pain sensitivity, J. Neurosci. 127 (1994) 164–169.

[48] J. Kim, S. Schiavon, G. Brager, Personal comfort models–A new paradigm in thermal comfort for occupant-centric environmental control, Build. Environ. 132 (2018) 114–124.

[49] C. Dai, H. Zhang, E. Arens, Machine learning approaches to predict thermal demands using skin temperature: steady-state conditions, Build. Environ. 114 (2017) 1–10.