BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 09/10 pp92–105
Volume 6, Number 1, March 2023
DOI: 10.26599/BDMA.2022.9020028

Predicted Mean Vote of Subway Car Environment Based on Machine Learning

Kangkang Huang, Shihua Lu*, Xinjun Li, Ke Feng, Weiwei Chen, and Yi Xia

Abstract: The thermal comfort of passengers in the carriage cannot be ignored. Thus, this research aims to establish a prediction model for the thermal comfort of the internal environment of a subway car and find the optimal input combination in establishing the prediction model of the predicted mean vote (PMV) index. Data-driven modeling utilizes data from experiments and questionnaires conducted in Nanjing Metro. Support vector machine (SVM), decision tree (DT), random forest (RF), and logistic regression (LR) were used to build four models. This research aims to select the most appropriate input variables for the predictive model. All possible combinations of 11 input variables were used to determine the most accurate model, with variable selection for each model comprising 102 350 iterations. In the PMV prediction, the RF model was the best when using the correlation coefficients square (R^2) as the evaluation indicator (R^2: 0.7680, mean squared error (*MSE*): 0.2868). The variables include clothing temperature (CT), convective heat transfer coefficient between the surface of the human body and the environment (CHTC), black bulb temperature (BBT), and thermal resistance of clothes (TROC). The RF model with *MSE* as the evaluation index also had the highest accuracy (R^2: 0.7676, *MSE*: 0.2836). The variables include clothing surface area coefficient (CSAC), CT, BBT, and air velocity (AV). The results show that the RF model can efficiently predict the PMV of the subway car environment.

Key words: predicted mean vote; random forest; variable selection; thermal comfort

1 Introduction

The advancement of urbanization has elicited widespread concern regarding urban traffic issues. The underground railway network is an essential part of public transportation, thus, various places are also vigorously opening and building subways[1]. Human thermal comfort in subway cars has become an important issue[2]. Monitoring the thermal comfort of passengers in the subway car environment is urgently needed to ensure their thermal comfort and health[3].

Regarding thermal comfort research, thermal comfort in buildings is popular because people spend more than 80% of their time indoors[4]. Many scholars have studied the internal thermal environments of buildings in different climatic regions. Taking a primary school classroom as an example, some scholars used the ANSYS FLUENT software to perform a series of three-dimensional numerical simulation experiments and numerical studies on indoor fluid flow thermal comfort[5]. Considering the green walls around the buildings as the research object, the best vertical greening mode for indoor thermal comfort in the cold winter and hot summer areas is determined[6]. Some scholars even questioned the building thermal comfort and energy consumption in their research and found that the aforementioned saves energy and improves thermal comfort. Yang et al.[7] proposed a building

• Kangkang Huang, Shihua Lu, Xinjun Li, Ke Feng, Weiwei Chen, and Yi Xia are with the School of Energy and Mechanical Engineering, Nanjing Normal University, Nanjing 210046, China. E-mail: 2402169616@qq.com; lushihua@njnu.edu.cn; lixinjun@nnu.edu.cn; fengke124@gmail.com; chenweiwei@njnu.edu.cn; seuxiayi@163.com.

* To whom correspondence should be addressed.

Manuscript received: 2022-05-12; revised: 2022-07-06; accepted: 2022-07-21

model predictive control system based on adaptive machine learning, reducing energy consumption and improving indoor thermal comfort. Zhu et al.[8] designed rural tourism buildings to achieve the multi-objective optimization of building energy consumption, daylighting, and thermal comfort performance. Zhang and Lin[9] proposed a simulation-rational thermal comfort model, that is, a simulation prediction value voting with a simulation function, and verified the mixed-mode buildings. The average absolute error and robustness of thermal sensation prediction were reduced. Gao et al.[10] used reinforcement learning for energy-efficient thermal comfort control of buildings. The results showed that the aforementioned strategy can improve thermal comfort prediction performance by 14.5%, reduce heating, ventilation, and air-conditioning (HVAC) energy consumption by 4.31%, and improve occupant thermal comfort by 13.6%.

This research focuses on thermal comfort in a subway car environment. Many research methods are used to evaluate the thermal comfort of the subway. For example, some scholars have used the relative warm index to examine the thermal environment of the Tehran Metro[11, 12]. Ampofo et al.[13] proposed an "acceptable" thermal comfort evaluation standard based on the thermal characteristics and humid climate of the London Underground. Sinha and Rajasekar[14] introduced an agent-based modeling method to assess the thermal comfort of a subway station in New Delhi, India. Zhang et al.[15] proposed an innovative subway environmental control system and discussed the energy performance of innovative environmental control systems in five cities representing five climate zones in China. However, complexity science can be surprisingly effective when traditional methods are not optimized for solving problems. These areas include quantitative, real-world, and even predictive models that combine statistical data analysis, modeling work, analytical methods, and laboratory experiments[16].

Machine learning (ML) has been increasingly applied to various fields. Big data are widely generated, collected, analyzed, and utilized in various intelligent systems to support pleasant and comfortable living and working conditions[17]. For example, Greener et al.[18] reviewed the application of ML in biology in recent years and discussed some emerging directions of ML. Meanwhile, Houssein et al.[19] summarized the latest technologies, challenges, and future visions of ML in the quantum field and simultaneously proposed two methods to improve the performance of classical ML. In addition, ML can be used in credit markets to predict default events[20]. Surprisingly, ML is already being used to combat climate change. Rolnick et al.[21] described how ML can be a powerful tool for reducing greenhouse gas emissions and helping societies adapt to a changing climate. Simultaneously, ML is also widely used in construction.

Many types of research on applying ML to thermal comfort are available, among which a popular application is to use ML algorithms to predict personal thermal comfort[22–24]. In addition, Shan and Yang[25] combined ML technology and passive electroencephalogram measurement to explore the real-time thermal comfort state of classified occupants. The ML model can predict thermal comfort in a highly asymmetric and dynamic thermal environment of a car cabin in real-time without relying on CFD simulation[26]. A predictive model is established through ML. The artificial neural network model is superior to the traditional thermal balance based model in predicting the thermal comfort voting and thermal sensation voting (TSV) of residents in natural ventilation houses[27]. Different ML algorithms are used to predict the TSV of the thermal comfort model, and simulation experiments are conducted to evaluate the performance of the proposed machine learning algorithm[28]. Li et al.[29], Liu[30], and Chaudhuri et al.[31] developed personal comfort models trained by ML. Among their models, random forests have high accuracy rates when compared with other ML models, reflecting the superiority of the integrated algorithm. Park and Choi[32] built a multivariate logistic regression model to predict typical window opening and closing patterns and evaluate indoor air quality and energy performance of residential buildings.

None of the above ML studies have addressed the thermal comfort of subway cars. These studies do not involve the influence of different feature combinations in the ML model. Thus, this paper refers to the research scheme of all feature combinations adopted by Deb et al.[33] The current research aims to establish the best predicted mean vote (PMV) index prediction model for the subway environment while finding the best feature combination. This research method will be introduced in Section 2. Section 3 presents the evaluation of the prediction results of the four models through two indicators and discusses the results. Finally, Section 4 shows the main conclusion of this paper.

2 Research Method

Table 1 is the variable naming table.

2.1 Field and questionnaire surveys

The Nanjing government officially opened subways in 2005. Nanjing is the first city to open subways in all districts and counties and the third city to open cross-city subway lines after Shanghai and Guangzhou. The relative humidity, average radiation temperature, air velocity, air temperature, air pressure, and other environmental parameters in the cabin were measured to study the relationship between the environmental parameters of the subway car and the thermal comfort of the human body quantitatively. Simultaneously, the clothing temperature, height, age, weight, clothing, gender, and bare skin temperature of the passenger under investigation were recorded.

Table 1 Nomenclature.

Abbreviation	Full name
AV	Air velocity
AT	Air temperature around the human body
BBT	Black bulb temperature
CHTC	Convective heat transfer coefficient between the surface of the human body and the environment
CSAC	Clothing surface area coefficient
CT	Clothing temperature
DT	Decision tree
LR	Logistic regression
MR	Metabolic rate
MRT	Mean radiant temperature
MSE	Mean squared error
PMV	Predicted mean vote
R^2	Correlation coefficients square
RF	Random forest
RH	Relative humidity
ST	Skin temperature
SVM	Support vector machines
TROC	Thermal resistance of clothes
TSV	Thermal sensation voting

The XIMA ST9450 thermal imaging camera was used in this study to obtain the clothing temperature (CT) and skin temperature (ST) of passengers. The thermal image is shown in Fig. 1. The XIMA AR866A thermal anemometer was also utilized to measure the air temperature (AT) around the human body and air velocity (AV). A black bulb thermometer was used to measure the black bulb temperature (BBT) and air relative humidity (RH) around the investigated passengers. The mean radiation temperature (MRT) in the cabin was also obtained. The empty box barometer can record the air pressure changes in the carriage. Using the classic Harris-Benedict[34] equation to find the metabolic rate (MR), the mechanical efficiency of the work performed by the human body is approximately 0 when riding the subway. The RH is known from $\varphi = P/P_{sat}$ (P is the partial pressure of water vapor, φ is the value of RH, and P_{sat} refers to the saturated water vapor partial pressure), where P_{sat} is determined in accordance with Eq. (1) (t_a refers to the air temperature). The clothing table was examined to determine the thermal resistance of clothes (TROC) while estimating the clothing surface area coefficient (CSAC) of passengers. The convective heat transfer coefficient between the surface of the human body and the environment (CHTC) was finally obtained by using the convective heat transfer formula proposed by De Dear et al.[35]

$$P_{sat} = \exp\left(23.299\,02 - \frac{3890.939}{t_a + 230.3980}\right) \tag{1}$$

In addition to the abovementioned measurements, a questionnaire survey was also conducted on the occupants. The questionnaire records the primary information of the passengers. The statistical data reveal

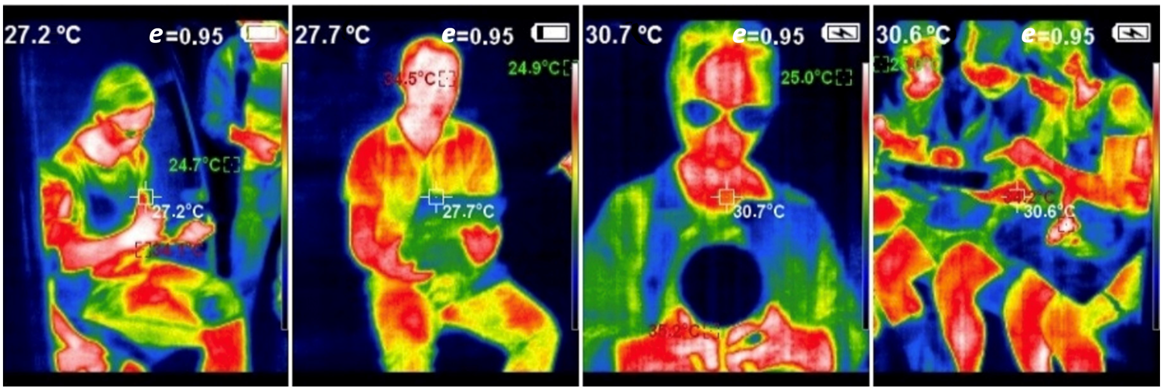


Fig. 1 Thermal image of the human surface inside a subway car.

that the majority of passengers are young and middle-aged, of which 45.1% are male passengers and 54.9% are female passengers. The existing thermal comfort model can be calculated and evaluated with the data obtained above. The main steps of this research scheme are shown in Fig. 2.

2.2 Evaluation index and characteristic data

Several methods are available for predicting thermal comfort in a subway environment. TSV is a widely used standard in comfort research[31], and another is the predicted mean vote (PMV) model proposed by Fanger[36]. The PMV index was selected in this study to evaluate the thermal comfort environment of the subway, and the ASHRAE standard 55[37] was adopted to obtain the PMV value through the relevant formula[38]. Table 2 shows the detailed data distribution.

Table 3 shows the 11 categorized relevant feature data from the experimental data. This study uses the 11 features to develop four models to predict the PMV index in the environmental thermal comfort of subway cars, aiming to find the optimal combination of eigenvalues.

2.3 Prediction models

2.3.1 Decision tree

The decision tree (DT) is a tree structure (can be binary or non-binary). DT includes the ID3 algorithm[39], C4.5 algorithm[40], and CART algorithm[41], which are the core technologies for classification and prediction[42]. The ID3 algorithm mainly aims to construct a DT recursively and select features at each node of the DT by applying the information gain criterion. The C4.5 algorithm is similar to the ID3 algorithm. Meanwhile, C4.5 uses a confidence gain ratio to select features during generation. The full name of the CART algorithm is the classification and regression tree model. As implied by its name, the CART algorithm can be used for classification and regression. The final generated DT is a binary tree. The right branch takes the value “No” and the left one takes the value “Yes”. This study uses the Gini coefficient CART algorithm for modeling.

Table 2 Information on the PMV parameters in this study.

Classification	PMV
Mean	0.939
Standard deviation	0.933
Maximum	4.486
Minimum	−2.359

Table 3 Variables used for model building.

No.	Variable
1	Metabolic rate (W/m^2)
2	Air temperature around the human body (℃)
3	Clothing surface area coefficient
4	Clothing temperature (℃)
5	Mean radiant temperature (℃)
6	Convective heat transfer coefficient between the surface of the human body and the environment (W·m^{-2} · ℃)
7	Relative humidity (%)
8	Black bulb temperature (℃)
9	Air velocity (m/s)
10	Thermal resistance of clothes (clo)
11	Skin temperature (℃)

2.3.2 Random forest

A random forest (RF) comprises multiple DTs, each of which is different. A part of the samples from the training data are randomly selected when building a DT, and all data features are not used; only some features are randomly selected for training. Each tree uses different samples and features and results in varying training results. The output classes of RFs are determined by

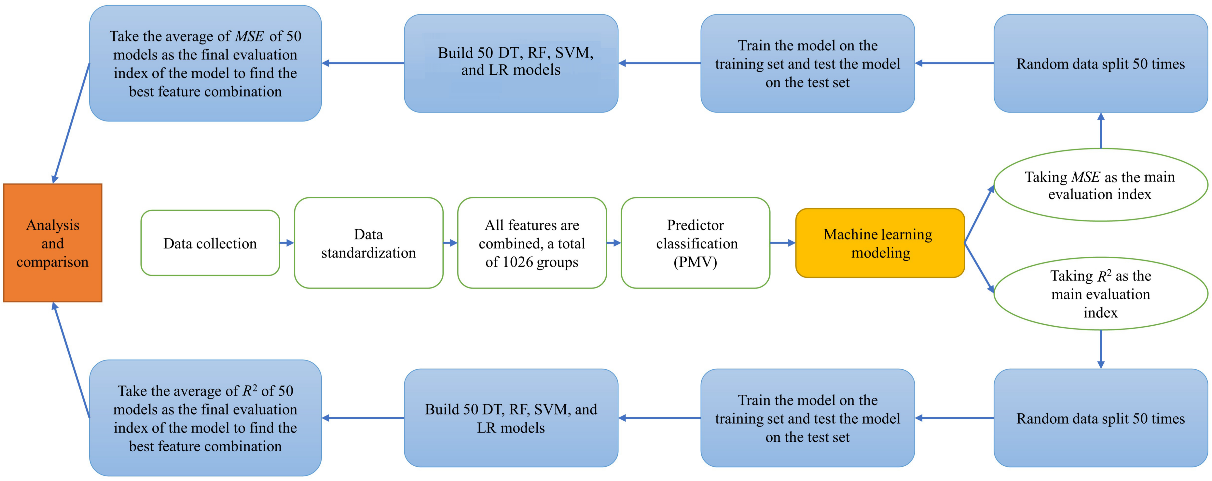

Take the average of *MSE* of 50 models as the final evaluation index of the model to find the best feature combination
Build 50 DT, RF, SVM, and LR models
Train the model on the training set and test the model on the test set
Random data split 50 times
Analysis and comparison
Data collection
Data standardization
All features are combined, a total of 1026 groups
Predictor classification (PMV)
Machine learning modeling
Taking *MSE* as the main evaluation index
Taking R^2 as the main evaluation index
Take the average of R^2 of 50 models as the final evaluation index of the model to find the best feature combination
Build 50 DT, RF, SVM, and LR models
Train the model on the training set and test the model on the test set
Random data split 50 times

Fig. 2 Overall flowchart of the research scheme.

the mode of the class output by individual trees by combining the bootstrap resampling method with the DT algorithm[43,44]. Suppose T(assumed number) DTs are generated. Thus, the original training set contains m samples and the number of features is n. The entire process is then presented as follows.

(1) Randomly sample m times with replacement from the original dataset containing m samples to obtain m samples (duplicate samples will be available).

(2) Train a DT using the sampled data.

(3) Repeat Steps (1) and (2) for a total of T times to obtain T-trained DTs.

(4) Use the voting method (classification tree) or simple average method (regression tree) to generate the final result from the prediction results of T DTs.

2.3.3 Logistic regression

Logistic regression[45] (LR) is a powerful and efficient method that analyzes the effect of a set of independent variables on binary outcomes by quantifying the unique contribution of each independent variable. LR is the preferred method for binary classification tasks. This method outputs a discrete binary result between 0 and 1. The result is generally either 1 or 0. People use the original LR to solve the two-class problem, continuously improve the algorithm, and obtain an LR method to solve the multiclass problem. Equation (2) is the classification equation.

$$f(x) = \frac{1}{1 + \mathrm{e}^{-x}} \tag{2}$$

where $f(x)$ refers to the classification function, x refers to the value of the abscissa independent variable, and e refers to a fixed value constant.

2.3.4 Support vector machine

Support vector machine (SVM) can handle massive data classification problems[46]. Especially before the emergence of deep learning, the SVM was considered the most successful ML algorithm in recent years. The SVM is a binary classification model. Given a training set $D = \{(x_1,y_1,), (x_2,y_2,), \ldots, (x_m,y_m,)\}$ ($(x_m,y_m,)$ refers to the coordinate point), classification learning aims to search a hyperplane S: $w^{\mathrm{T}}x + b = 0$ (w^{T}represents a vector, x refers to the value of the abscissa independent variable, and b refers to an unfixed value constant), thereby dividing samples of different categories. Different classes of samples in the sample space of D are distinguished. From simple to complex, SVM models include linear SVM in linearly separable cases, linear SVM, and nonlinear SVM.

When the data are linearly separable, the SVM attempts to find the hyperplane of hard margin maximization because the classification result produced by such a hyperplane is the most robust, and the linear classifier learned from this is called a linearly separable SVM; when the data are approximately linearly separable, classifiers can also be learned by maximizing soft margins (called linear SVM); when the data are linearly inseparable, kernel methods can be used. A nonlinear SVM and soft margin maximization are learned.

2.3.5 Evaluation indicators

This study uses the correlation coefficients square (R^2) and mean squared error (*MSE*) to evaluate the quality of the model, as in Eqs. (3) and (4). Among these, R^2 and *MSE* are evaluation methods. A large R^2 value results in a small *MSE* value, thus improving the model. Theoretically, a high R^2 value induces a low *MSE* value; however, an opposite situation may emerge. To this end, this research will use the two evaluation methods for illustration.

$$R^2 = \left(\frac{\sum (X_i - X_m)(Y_i - Y_m)}{\sqrt{\sum (X_i - X_m) \sum (Y_i - Y_m)}} \right)^2 \tag{3}$$

$$MSE = \frac{1}{N} \sum_{i=1}^{n} (X_i - Y_i)^2 \tag{4}$$

where X_i and Y_i refer to the experimental values, X_m and Y_m are the average values, and N refers to the number of test data.

2.4 Variable selections

Each combination of the 11 variables is used to build a predictive model. R^2 and *MSE* are utilized to test the results of each prediction model. The dataset is split into a 7:3 ratio into training and testing sets. Furthermore, this study developed 50 models for each combination and took the average as the final value considering the randomness of dataset partitioning. Figure 3 shows an example of selecting three variables for all iterations. The feature data were normalized before building the model. The model building steps are presented below.

(1) Identify the permutations and combinations of 11 variables.

(2) Develop 50 DT/RF/LR/SVM models for each combination. 30% of data are used for validation and testing and 70% of data are used for training.

(3) Randomly divide the training and test sets 50 times and develop 50 models.

(4) Calculate the mean R^2 and *MSE* values for 50

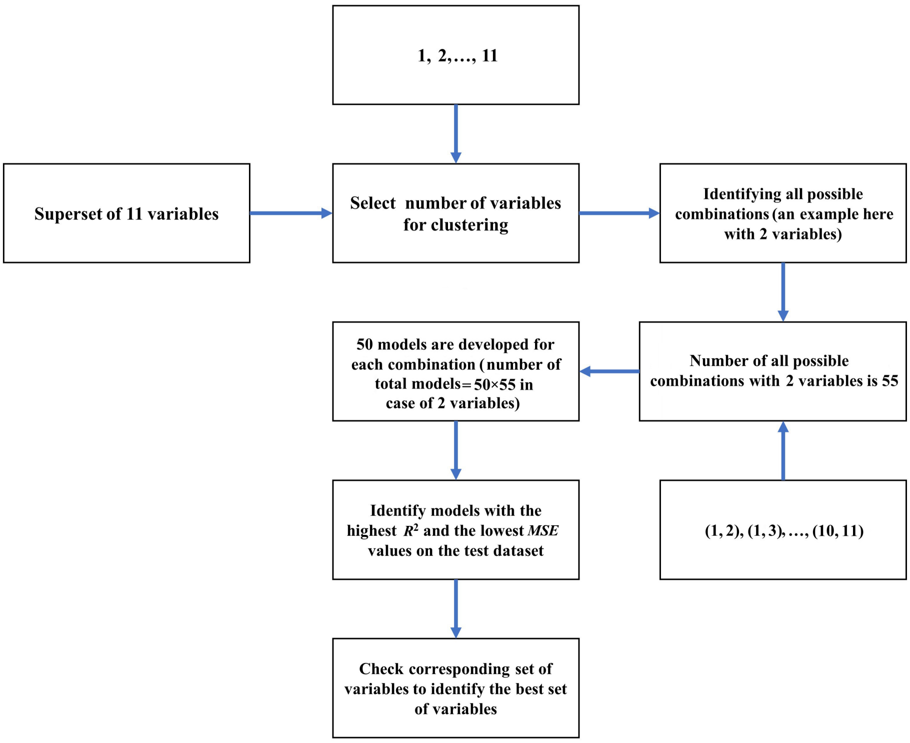

1, 2,..., 11
Superset of 11 variables
Select number of variables for clustering
Identifying all possible combinations (an example here with 2 variables)
Number of all possible combinations with 2 variables is 55
50 models are developed for each combination (number of total models= 50×55 in case of 2 variables)
(1, 2), (1, 3),..., (10, 11)
Identify models with the highest R^2 and the lowest *MSE* values on the test dataset
Check corresponding set of variables to identify the best set of variables

Fig. 3 Flowchart for developing a predictive model with a combination of variables.

models for all variable combinations.

(5) Select the combination with the highest average R^2/*MSE* for further evaluation.

3 Result and Discussion

3.1 Result of variable selection

3.1.1 Take R^2 as the evaluation index

The best combination of 2047 feature combinations when R^2 is used as the evaluation index will be introduced in this section. Tables 4–7 intuitively demonstrate the predictions of all feature combinations of the four models.

Table 4 R^2 and *MSE* values of the best-combined variables for the test dataset in the DT model (R^2 as the evaluation indicator).

Number of variables	Number of all combinations	Best number	R^2	*MSE*
1	11	4	0.4653	0.9495
2	55	31	0.5588	0.6115
3	165	125	0.5709	0.5870
4	330	280	0.5727	0.5921
5	462	226	0.5785	0.5901
6	462	58	0.5715	0.5922
7	**330**	**281**	**0.5798**	**0.5772**
8	165	31	0.5736	0.6024
9	55	1	0.5692	0.5973
10	11	1	0.5615	0.6047
11	1	1	0.5635	0.62
Total	2047	–	–	–

Table 5 R^2 and MSE values of the best-combined variables for the test dataset in the RF model (R^2 as the evaluation indicator).

Number of variables	Number of all combinations	Best number	R^2	*MSE*
1	11	4	0.4608	0.9776
2	55	31	0.7145	0.3477
3	165	117	0.7455	0.3268
4	**330**	**281**	**0.7680**	**0.2868**
5	462	226	0.7485	0.3119
6	462	274	0.7292	0.3440
7	330	274	0.7146	0.3638
8	165	153	0.7035	0.3857
9	55	20	0.7222	0.3572
10	11	11	0.7110	0.3628
11	1	1	0.6994	0.3855
Total	2047	–	–	–

In the DT model, the model has the best prediction accuracy when the input has seven eigenvalues. Figures 4a and 5a indicate that the 281st result corresponding to the combinations has the best R^2 value. The variables corresponding to this combination were AT, CT, MRT, BBT, AV, TROC, and ST.

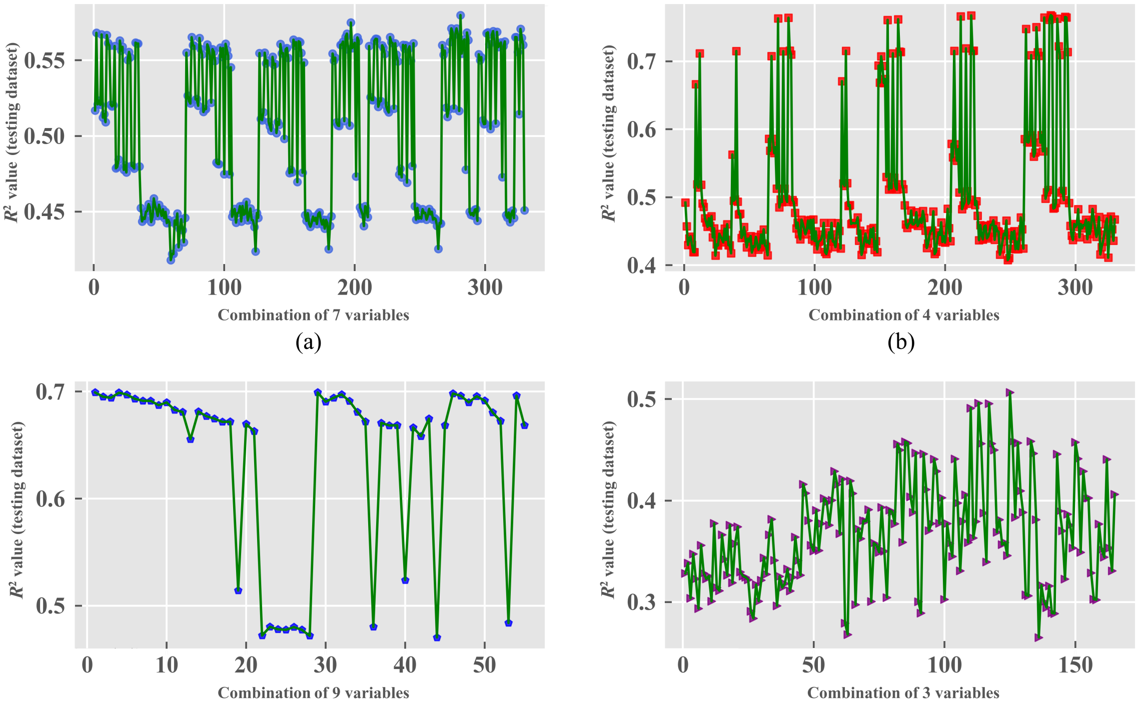
Four features comprise the best RF model. Figures 4b and 5b reveal that the 281st result corresponding to the combinations has the best R^2 value. The detected

Table 6 R^2 and *MSE* values of the best-combined variables for the test dataset in the LR model (R^2 as the evaluation indicator).

Number of variables	Number of all combinations	Best number	R^2	*MSE*
1	11	5	0.4613	0.8807
2	55	31	0.6369	0.4556
3	165	117	0.6647	0.4086
4	330	149	0.6802	0.3784
5	462	269	0.6851	0.3771
6	462	325	0.6925	0.3653
7	330	270	0.6980	0.3616
8	165	60	0.6988	0.3626
9	**55**	**1**	**0.6991**	**0.3616**
10	11	1	0.6989	0.3615
11	1	1	0.6971	0.3649
Total	2047	–	–	–

Table 7 R^2 and *MSE* values of the best-combined variables for the test dataset in the SVM model (R^2 as the evaluation indicator).

Number of variables	Number of all combinations	Best number	R^2	*MSE*
1	11	4	0.4620	0.9247
2	55	31	0.4801	1.0445
3	**165**	**125**	**0.5065**	**0.9295**
4	330	280	0.4911	0.8955
5	462	356	0.4715	0.9338
6	462	405	0.4438	1.0310
7	330	218	0.4058	1.2419
8	165	124	0.3759	1.3868
9	55	1	0.3735	1.4660
10	11	9	0.3746	1.4198
11	1	1	0.3707	1.4062
Total	2047	–	–	–



(a) (b) (c) (d)

Fig. 4 PMV test R^2. (R^2 was taken as the evaluation criteria. (a) R^2 values of the DT model when seven feature combinations are input. (b) R^2 values of the RF model when four feature combinations are input. (c) R^2 values of the LR model when nine feature combinations are input. (d) R^2 values of the SVM model when three feature combinations are input.)

variables included CT, CHTC, BBT, and TROC.

In the LR model, the model has the best prediction accuracy when the input is nine eigenvalues. Figures 4c and 5c show that the first result corresponding to the combinations has the best R^2 value. The variables corresponding to this combination were MR, AT, CSAC, CT, MRT, CHTC, RH, BBT, and AV.

Three features comprise the best SVM model. Figures 4d and 5d indicate that the 125th result corresponding to the combinations has the best R^2 value. The variables were CT, BBT, and AV.

3.1.2 Take *MSE* as the evaluation index

The best combination of 2047 feature combinations when *MSE* is used as the evaluation index will be presented in this subsection. Tables 8–11 intuitively reveal the predictions of all feature combinations of the four models.

In the DT model, the model has the best prediction

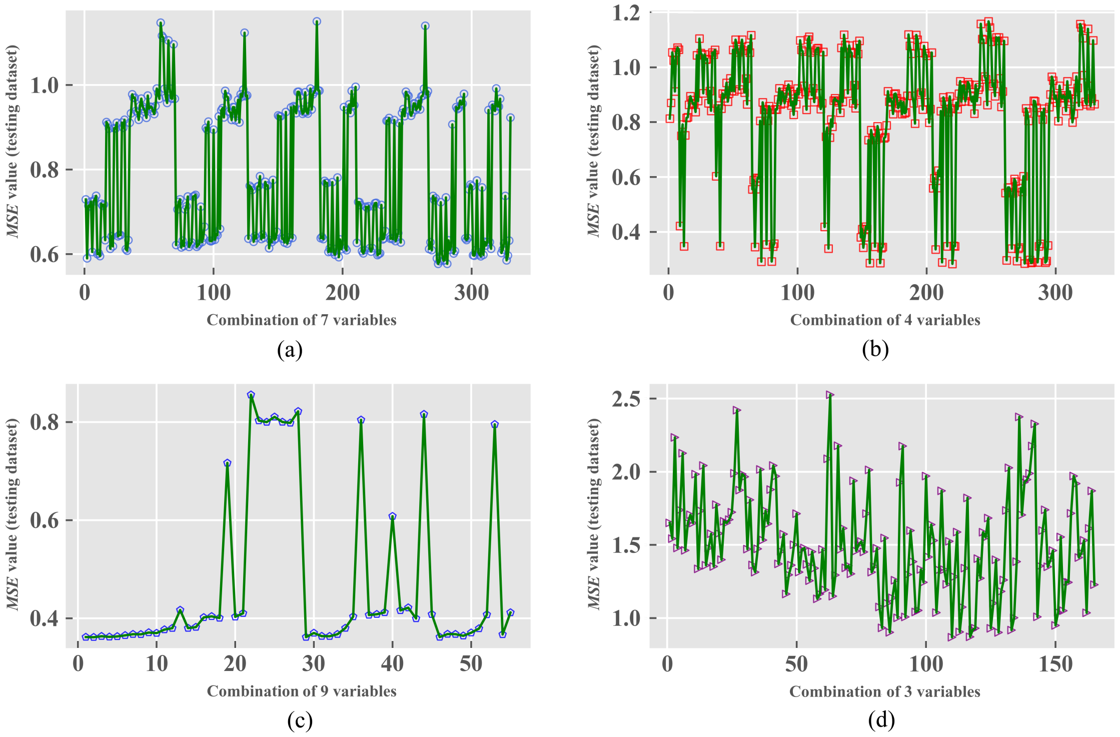


Fig. 5 PMV test *MSE*. (R^2 was taken as the evaluation criteria. (a) *MSE* values of the DT model when seven feature combinations are input. (b) *MSE* values of the RF model when four feature combinations are input. (c) *MSE* values of the LR model when nine feature combinations are input. (d) *MSE* values of the SVM model when three feature combinations are input.)

Table 8 R^2 and *MSE* values of the best-combined variables for the test dataset in the DT model (*MSE* as the evaluation indicator).

Number of variables	Number of all combinations	Best number	R^2	*MSE*
1	11	8	0.4588	0.9046
2	55	31	0.5588	0.6115
3	165	57	0.5676	0.5865
4	330	151	0.5670	0.5815
5	**462**	**423**	**0.5722**	**0.5722**
6	462	323	0.5672	0.5865
7	330	274	0.5710	0.5772
8	165	131	0.5678	0.5829
9	55	1	0.5693	0.5973
10	11	1	0.5616	0.6048
11	1	1	0.5635	0.6200
Total	2047	–	–	–

Table 9 R^2 and *MSE* values of the best-combined variables for the test dataset in the RF model (*MSE* as the evaluation indicator).

Number of variables	Number of all combinations	Best number	R^2	*MSE*
1	11	8	0.4525	0.9389
2	55	31	0.7145	0.3477
3	165	125	0.7444	0.3246
4	**330**	**220**	**0.7676**	**0.2836**
5	462	287	0.7468	0.3113
6	462	323	0.7209	0.3361
7	330	281	0.7122	0.3604
8	165	123	0.6973	0.3797
9	55	47	0.7169	0.3498
10	11	11	0.7111	0.3628
11	1	1	0.6994	0.3856
Total	2047	–	–	–

accuracy when the input has five eigenvalues. Figures 6a and 7a reveal that the 423rd result corresponding to the combinations has the best R^2 value. The variables corresponding to this combination were CT, MRT, BBT, AV, and TROC.

Four features comprise the best RF model. Figures 6b and 7b show that the 220th result corresponding to the combinations has the best R^2 value. The variables were CSAC, CT, BBT, and AV.

In the LR model, the model has the best prediction accuracy when the input is eight eigenvalues. Figures 6c and 7c shows that the 58th result corresponding to the combinations has the best R^2 value. The variables corresponding to this combination number are MR, AT, CT, MRT, CHTC, RH, BBT, and TROC.

Three features comprise the best SVM model, as shown in Figs. 6d and 7d. The 110th result corresponding to the combination had the best R^2 value. The variables

Table 10 R^2 and *MSE* values of the best-combined variables for the test dataset in the LR model (*MSE* as the evaluation indicator).

Number of variables	Number of all combinations	Best number	R^2	*MSE*
1	11	4	0.4579	0.8747
2	55	31	0.6370	0.4556
3	165	117	0.6648	0.4087
4	330	149	0.6803	0.3784
5	462	267	0.6843	0.3722
6	462	325	0.6925	0.3654
7	330	1	0.6955	0.3616
8	**165**	**58**	**0.6954**	**0.3601**
9	55	2	0.6952	0.3613
10	11	1	0.6989	0.3616
11	1	1	0.6972	0.3650
Total	2047	–	–	–

Table 11 R^2 and *MSE* values of the best-combined variables for the test dataset in the SVM model (*MSE* as the evaluation indicator).

Number of variables	Number of all combinations	Best number	R^2	*MSE*
1	11	5	0.4558	0.9076
2	55	32	0.4641	0.8898
3	**165**	**110**	**0.4910**	**0.8687**
4	330	263	0.4863	0.8881
5	462	339	0.4649	0.9315
6	462	386	0.4427	1.0105
7	330	218	0.4059	1.2419
8	165	124	0.3759	1.3869
9	55	31	0.3721	1.4284
10	11	3	0.3701	1.4129
11	1	1	0.3708	1.4063
Total	2047	–	–	–

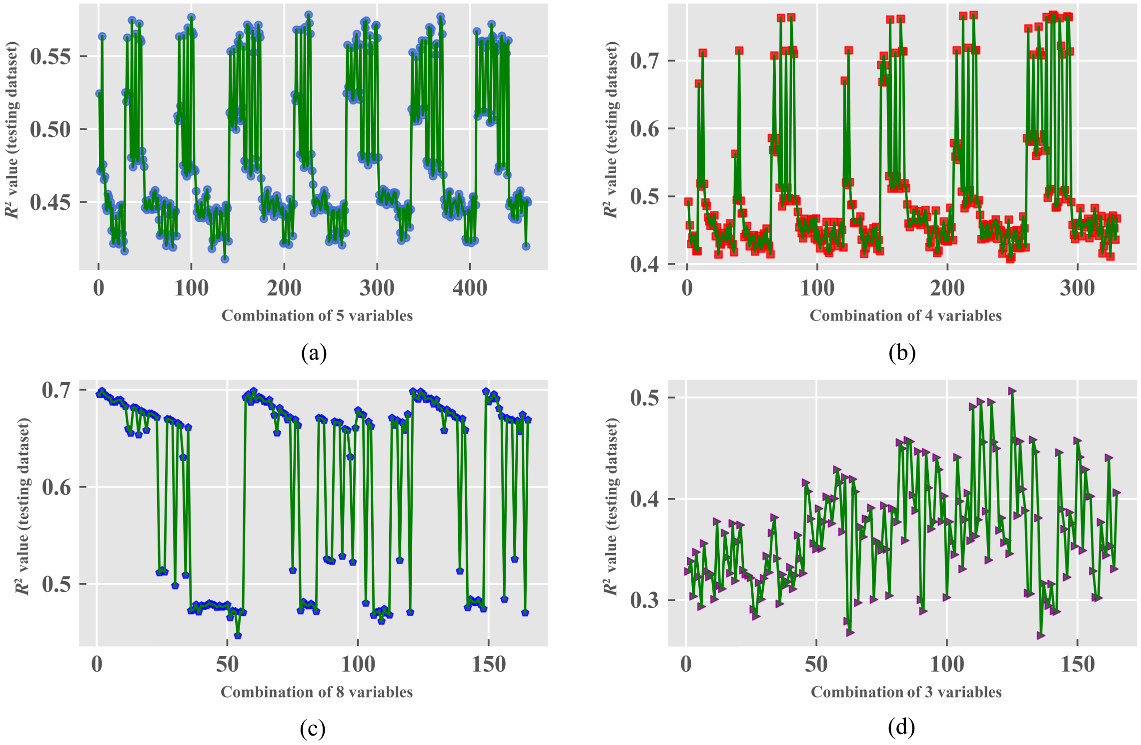


Fig. 6 PMV test R^2. (*MSE* was taken as evaluation criteria. (a) R^2 values of the DT model when five feature combinations are input. (b) R^2 values of the RF model when four feature combinations are input. (c) R^2 values of the LR model when eight feature combinations are input. (d) R^2 values of the SVM model when three feature combinations are input.)

are CT, MRT, and CHTC.

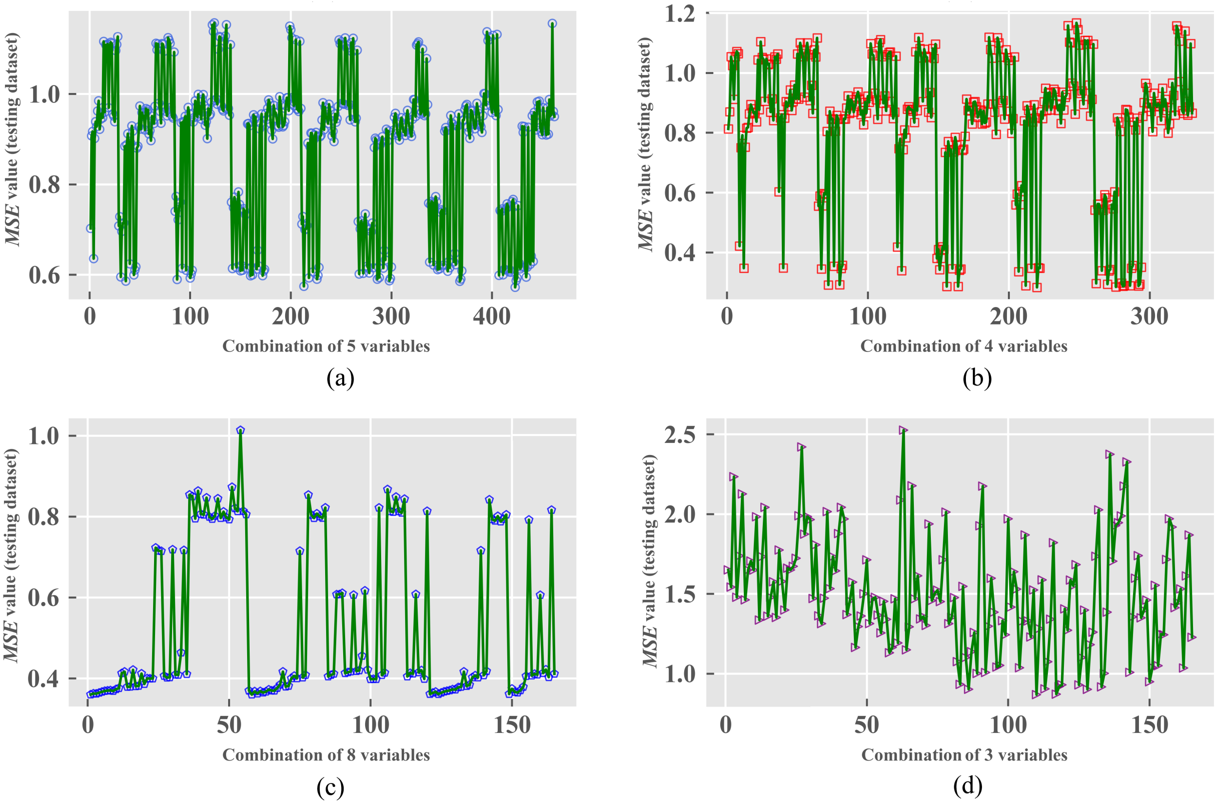
3.2 Predicted results by the model

3.2.1 DT model

When AT, CT, MRT, BBT, AV, TROC, and ST are inputted as features, the prediction effect of the DT model is shown in Fig. 8a (R^2: 0.5798, *MSE*: 0.5772). Moreover, when CT, MRT, BBT, AV, and TROC are used as feature input, the prediction effect of the DT model is shown in Fig. 9a (R^2: 0.5722, *MSE*: 0.5772). The prediction effect of the former is better than that of the latter.

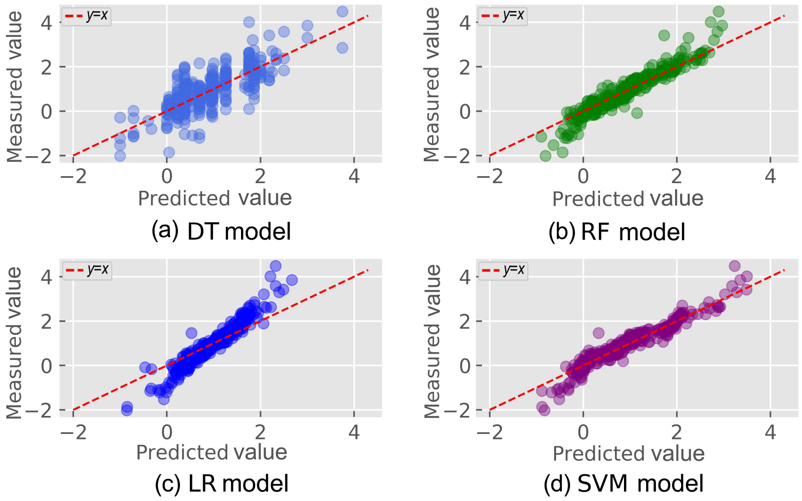
3.2.2 RF model

In the RF model, when CT, CHTC, BBT, and TROC are inputted as features, the prediction effect of the RF model is shown in Fig. 8b (R^2: 0.7680, *MSE*: 0.2868). Furthermore, when CSAC, CT, BBT, and AV are used as feature input, the prediction effect of the RF model



(a) (b) (c) (d)

Fig. 7 PMV test *MSE*. (*MSE* was taken as the evaluation criteria. (a) *MSE* values of the DT model when five feature combinations are input. (b) *MSE* values of the RF model when four feature combinations are input. (c) *MSE* values of the LR model when eight feature combinations are input. (d) *MSE* values of the SVM model when three feature combinations are input.)



(a) DT model (b) RF model (c) LR model (d) SVM model

Fig. 8 Distribution plot of measured and predicted values for the best model when R^2 is used as the evaluation indicator.

is shown in Fig. 9b (R^2: 0.7676, *MSE*: 0.2836). The prediction effect of the former is only slightly different from that of the latter.

3.2.3 LR model

When MR, AT, CSAC, CT, MRT, CHTC, RH, BBT, and AV are inputted as features, the prediction effect of the LR model is shown in Fig. 8c (R^2: 0.6991, *MSE*: 0.3616). Moreover, when MR, AT, CT, MRT, CHTC, RH, BBT, and TROC are used as feature input, the prediction effect of the LR model is shown in Fig. 9c (R^2: 0.6954, *MSE*: 0.3601). The prediction effect of the former is also only slightly different from that of the latter.

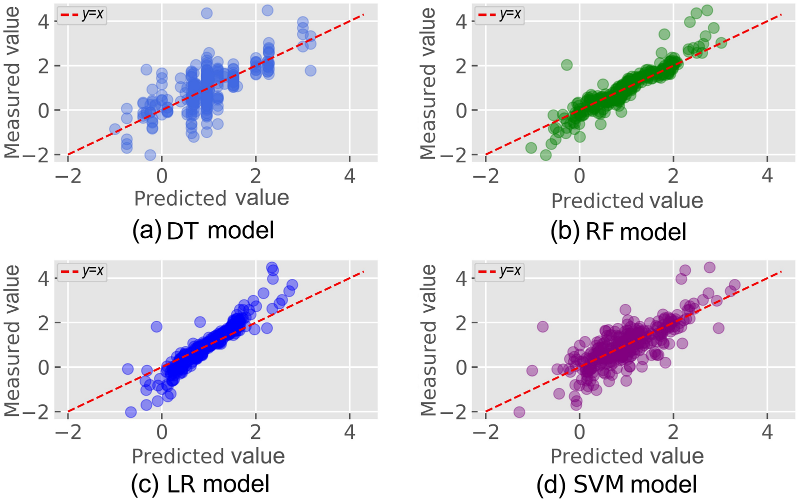


Fig. 9 Distribution plot of measured and predicted values for the best model when *MSE* is used as the evaluation indicator.

3.2.4 SVM model

In the SVM model, when CT, BBT, and AV are inputted as features, the prediction effect of the SVM model is shown in Fig. 8d (R^2: 0.5065, *MSE*: 0.9295). Furthermore, when CT, MRT, and CHTC are used as feature input, the prediction effect of the SVM model is shown in Fig. 9d (R^2: 0.4910, *MSE*: 0.8687). The prediction effect of the former is only slightly different from that of the latter.

The choice of features affects the prediction performance of the ML model. All the features are arranged and combined in this research: A total of 11 arrangements is available and each arrangement has different combination types. The R^2 and *MSE* values of the four models for the best combination of predictions in each arrangement are summarized to observe the prediction effect of the four models intuitively. Figure 10 shows that R^2 is used as the evaluation standard, wherein the R^2 value of the RF model is at the highest value in Fig. 10a and the *MSE* value is at the lowest value in Fig. 10b. In Fig. 11, using *MSE* as the evaluation standards, the RF model also has the lowest *MSE* value and the highest R^2 value. The above results can illustrate the considerable superiority of the RF model.

4 Conclusion

This study obtained relevant data through experiments and investigations, applied ML to the subway car, and built a variable selector based on the exploration of all possible subsets of 11 variables. This research developed an algorithm to select the most accurate set of variables as part of the variable selection process. The conclusions of this research are presented as follows.

(1) This research first uses field measurements and

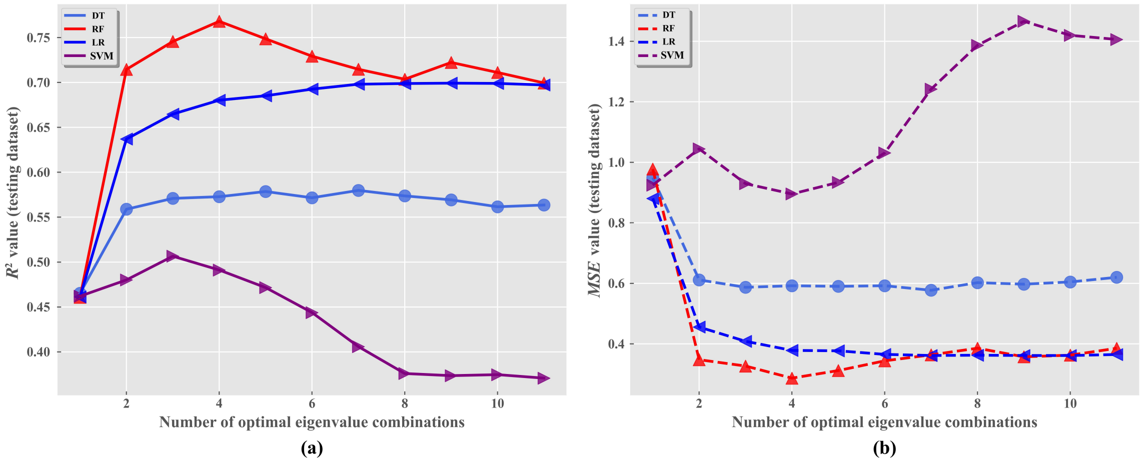


Fig. 10 PMV test R^2 and *MSE* of four models when R^2 is used as the evaluation indicator (The *X*-axis is the feature number. The left *Y*-axis is the highest R^2 in the number of feature combinations, and the right *Y*-axis is the *MSE* corresponding to the highest R^2 in the number of feature combinations).

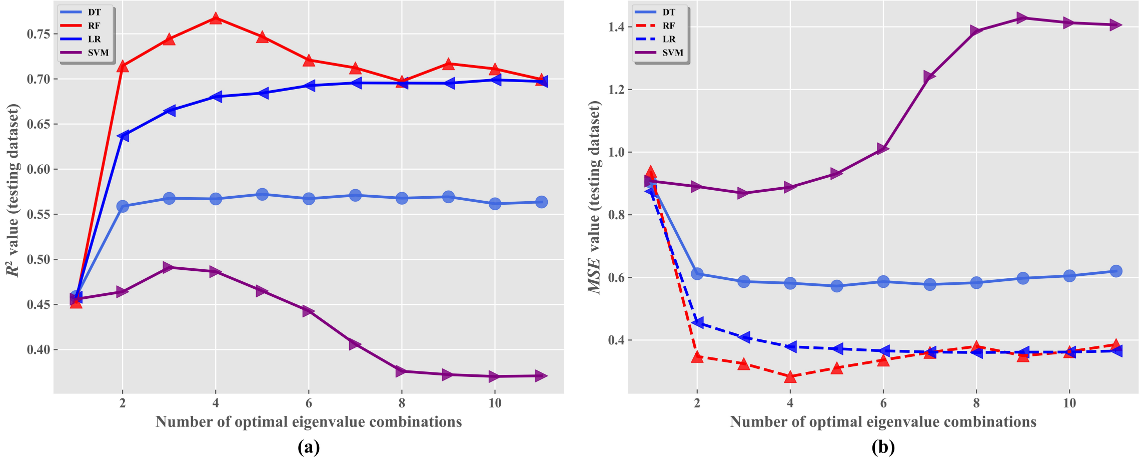


Fig. 11 PMV test R^2 and *MSE* of four models when *MSE* is used as the evaluation (The X-axis is the feature number. The left Y-axis is the R^2 corresponding to the lowest *MSE* in the number of feature combinations, and the right Y-axis is the lowest *MSE* in the number of feature combinations).

questionnaire surveys to obtain relevant data on the thermal environment of subway cars in Nanjing and the thermal comfort of passengers and examine the relationship between relevant parameters and PMV.

(2) After normalizing the data, this research arranges and combines all the features as feature input and establishes four ML models through data driving to make predictions. R^2 and *MSE* are used as evaluation indicators to assess the prediction results.

(3) When using R^2 as the evaluation standard, the four models in the prediction are generally the best in the RF, and the overall arrangement of the four advantages is RF, LR, DT, and SVM. The RF model is optimal when the number of input eigenvalues is 4. The four eigenvalues are CT, CHTC, BBT, and TROC (R^2: 0.7680, *MSE*: 0.2868).

(4) When using *MSE* as the evaluation standard, the four models in the prediction are generally the best in the RF, and the overall arrangement of the four advantages is RF, LR, DT, and SVM. The RF model is optimal when the number of input eigenvalues is 4. The four eigenvalues are CSAC, CT, BBT, and AV (R^2: 0.7676, and *MSE*: 0.2836).

This research can provide references for relevant personnel to investigate future subway thermal comfort prediction and energy conservation.

References

[1] A. K. Y. Ng and J. J. Wang, Transport development in China, *Research in Transportation Economics*, vol. 35, no. 1, pp. 1&2, 2012.

[2] S. Pan, Y. Liu, L. Xie, X. Wang, Y. Yuan, and X. Jia, A thermal comfort field study on subway passengers during air-conditioning season in Beijing, *Sustainable Cities and Society*, vol. 61, p. 102218, 2020.

[3] M. Marzouk and A. Abdelaty, Monitoring thermal comfort in subways using building information modeling, *Energy and Buildings*, vol. 84, pp. 252–257, 2014.

[4] R. Zhao, S. Sun, and R. Ding, Conditioning strategies of indoor thermal environment in warm climates, *Energy and Buildings*, vol. 36, no. 12, pp. 1281–1286, 2004.

[5] Y. Xia, W. Lin, W. Gao, T. Liu, Q. Li, and A. Li, Experimental and numerical studies on indoor thermal comfort in fluid flow: A case study on primary school classrooms, *Case Studies in Thermal Engineering*, vol. 19, p. 100619, 2020.

[6] J. Li, B. Zheng, X. Chen, Z. Qi, K. B. Bedra, J. Zheng, Z. Li, and L. Liu, Study on a full-year improvement of indoor thermal comfort by different vertical greening patterns, *Journal of Building Engineering*, vol. 35, p. 101969, 2021.

[7] S. Yang, M. P. Wan, W. Chen, B. F. Ng, and S. Dubey, Model predictive control with adaptive machine-learning-based model for building energy efficiency and comfort optimization, *Applied Energy*, vol. 271, p. 115147, 2020.

[8] L. Zhu, B. Wang, and Y. Sun, Multi-objective optimization for energy consumption, daylighting and thermal comfort performance of rural tourism buildings in north China, *Building and Environment*, vol. 176, p. 106841, 2020.

[9] S. Zhang and Z. Lin, Adaptive-rational thermal comfort model: Adaptive predicted mean vote with variable adaptive coefficient, *Indoor Air*, vol. 30, no. 5, pp. 1052–1062, 2020.

[10] G. Gao, J. Li, and Y. Wen, Deepcomfort: Energy-efficient thermal comfort control in buildings via reinforcement learning, *IEEE Internet of Things Journal*, vol. 7, no. 9, pp. 8472–8484, 2020.

[11] A. Alahmer, M. Omar, A. R. Mayyas, and A. Qattawi, Analysis of vehicular cabins' thermal sensation and comfort state, under relative humidity and temperature control, using Berkeley and Fanger models, *Building and Environment*, vol. 48, pp. 146–163, 2012.

[12] M. Abbaspour, M. Jafari, N. Mansouri, F. Moattar, N.

Nouri, and M. Allahyari, Thermal comfort evaluation in Tehran metro using relative warmth index, *International Journal of Environmental Science & Technology*, vol. 5, no. 3, pp. 297–304, 2008.

[13] F. Ampofo, G. Maidment, and J. Missenden, Underground railway environment in the UK part 1: Review of thermal comfort, *Applied Thermal Engineering*, vol. 24, nos. 5&6, pp. 611–631, 2004.

[14] K. Sinha and E. Rajasekar, Thermal comfort evaluation of an underground metro station in New Delhi using agent-based modelling, *Building and Environment*, vol. 177, p. 106924, 2020.

[15] H. Zhang, T. Cui, M. Liu, W. Zheng, C. Zhu, S. You, and Y. Zhang, Energy performance investigation of an innovative environmental control system in subway station, *Building and Environment*, vol. 126, pp. 68–81, 2017.

[16] D. Helbing, D. Brockmann, T. Chadefaux, K. Donnay, U. Blanke, O. Woolley-Meza, M. Moussaid, A. Johansson, J. Krause, S. Schutte, et al., Saving human lives: What complexity science and information systems can contribute, *Journal of Statistical Physics*, vol. 158, no. 3, pp. 735–781, 2015.

[17] M. Jusup, P. Holme, K. Kanazawa, M. Takayasu, I. Romić Z. Wang, S. Geček, T. Lipić, B. Podobnik, L. Wang, et al., Social physics, *Physics Reports*, vol. 948, pp. 1–148, 2022.

[18] J. G. Greener, S. M. Kandathil, L. Moffat, and D. T. Jones, A guide to machine learning for biologists, *Nature Reviews Molecular Cell Biology*, vol. 23, no. 1, pp. 40–55, 2022.

[19] E. H. Houssein, Z. Abohashima, M. Elhoseny, and W. M. Mohamed, Machine learning in the quantum realm: The state-of-the-art, challenges, and future vision, *Expert Systems with Applications*, vol. 194, p. 116512, 2022.

[20] A. Fuster, P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther, Predictably unequal? The effects of machine learning on credit markets, *The Journal of Finance*, vol. 77, no. 1, pp. 5–47, 2022.

[21] D. Rolnick, P. L. Donti, L. H. Kaack, K. Kochanski, A. Lacoste, K. Sankaran, A. S. Ross, N. Milojevic-Dupont, N. Jaques, A. Waldman-Brown, et al., Tackling climate change with machine learning, *ACM Computing Surveys* (*CSUR*), vol. 55, no. 2, pp. 1–96, 2022.

[22] Z. Q. Fard, Z. S. Zomorodian, and S. S. Korsavi, Application of machine learning in thermal comfort studies: A review of methods, performance and challenges, *Energy and Buildings*, vol. 256, p. 111771, 2021.

[23] Z. Hasan and N. Roy, Trending machine learning models in cyber-physical building environment: A survey, *Wiley Interdisciplinary Reviews*: *Data Mining and Knowledge Discovery*, vol. 11, no. 5, p. e1422, 2021.

[24] Y. Feng, S. Liu, J. Wang, J. Yang, Y. -L. Jao, and N. Wang, Data-driven personal thermal comfort prediction: A literature review, *Renewable and Sustainable Energy Reviews*, vol. 161, p. 112357, 2022.

[25] X. Shan and E. -H. Yang, Supervised machine learning of thermal comfort under different indoor temperatures using EEG measurements, *Energy and Buildings*, vol. 225, p. 110305, 2020.

[26] A. Warey, S. Kaushik, B. Khalighi, M. Cruse, and G. Venkatesan, Data-driven prediction of vehicle cabin thermal comfort: Using machine learning and high-fidelity simulation results, *International Journal of Heat and Mass Transfer*, vol. 148, p. 119083, 2020.

[27] Q. Chai, H. Wang, Y. Zhai, and L. Yang, Using machine learning algorithms to predict occupants' thermal comfort in naturally ventilated residential buildings, *Energy and Buildings*, vol. 217, p. 109937, 2020.

[28] M. Abdulgader and F. Lashhab, Energy-efficient thermal comfort control in smart buildings, in *Proc. 2021 IEEE 11th Annual Computing and Communication Workshop and Conference* (*CCWC*), NV, USA, 2021, pp. 22–26.

[29] D. Li, C. C. Menassa, and V. R. Kamat, Non-intrusive interpretation of human thermal comfort through analysis of facial infrared thermography, *Energy and Buildings*, vol. 176, pp. 246–261, 2018.

[30] S. Liu, Personal thermal comfort models based on physiological parameters measured by wearable sensors, *Building Efficiency and Sustainability in the Tropics*, https://escholarship.org/uc/item/3qk6d6tv, 2018.

[31] T. Chaudhuri, D. Zhai, Y. C. Soh, H. Li, and L. Xie, Random forest based thermal comfort prediction from gender-specific physiological parameters using wearable sensing technology, *Energy and Buildings*, vol. 166, pp. 391–406, 2018.

[32] J. Park and C. -S. Choi, Modeling occupant behavior of the manual control of windows in residential buildings, *Indoor Air*, vol. 29, no. 2, pp. 242–251, 2019.

[33] C. Deb, S. E. Lee, and M. Santamouris, Using artificial neural networks to assess HVAC related energy saving in retrofitted office buildings, *Solar Energy*, vol. 163, pp. 32–44, 2018.

[34] D. C. Frankenfield, E. R. Muth, and W. A. Rowe, The Harris-Benedict studies of human basal metabolism: History and limitations, *Journal of the American Dietetic Association*, vol. 98, no. 4, pp. 439–445, 1998.

[35] R. J. De Dear, E. Arens, Z. Hui, and M. Oguro, Convective and radiative heat transfer coefficients for individual human body segments, *International Journal of Biometeorology*, vol. 40, no. 3, pp. 141–156, 1997.

[36] P. O. Fanger, Thermal comfort: Analysis and applications in environmental engineering, *Copenhagen*: *Danish Technical Press*, vol. 45, no. 1, p. 244, 1970.

[37] Thermal Environmental Conditions for Human Occupancy, ASHRAE 55 ADD A-2014, 2014-06-28.

[38] W. Zhang, F. Liu, and R. Fan, Improved thermal comfort modeling for smart buildings: A data analytics study, *International Journal of Electrical Power & Energy Systems*, vol. 103, pp. 634–643, 2018.

[39] A. Rana and R. Pandey, A review of popular decision tree algorithms in data mining, *Asian Journal of Multidimensional Research*, vol. 10, no. 10, pp. 230–237, 2021.

[40] Y. Freund, Boosting a weak learning algorithm by majority, *Information and Computation*, vol. 121, no. 2, pp. 256–285, 1995.

[41] P. Bindra, M. Kshirsagar, C. Ryan, G. Vaidya, K. K. Gupt, and V. Kshirsagar, Insights into the advancements of artificial intelligence and machine learning, the present state of art, and future prospects: Seven decades of digital revolution, in *Smart Computing Techniques and Applications*, S. C. Satapathy, V. Bhateja, M. N. Favorskaya, and T. Adilakshmi, eds. Singapore: Springer, 2021, pp. 609–621.

[42] B. Charbuty and A. M. Abdulazeez, Classification based on decision tree algorithm for machine learning, *Journal of Applied Science and Technology Trends*, vol. 2, no. 1, pp. 20–28, 2021.

[43] F. C. Eugenio, T. L. Badin, P. Fernandes, C. L. Mallmann, C. Schons, M. S. Schuh, R. S. Pereira, R. A. Fantinel, and S. D. P. D. Silva, Remotely piloted aircraft systems (RPAS) and machine learning: A review in the context of forest science, *International Journal of Remote Sensing*, vol. 42, no. 21, pp. 8207–8235, 2021.

[44] L. Breiman, Random forest, *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[45] J. Goyal and R. R. Sinha, Software defect-based prediction using logistic regression: Review and challenges, in *Second International Conference on Sustainable Technologies for Computational Intelligence*, A. K. Luhach, R. C. Poonia, X. -Z. Gao, and D. S. Jat, eds. Singapore: Springer, 2022, pp. 233–248.

[46] H. Ibrahim, S. A. Anwar, and M. I. Ahmad, Classification of imbalanced data using support vector machine and rough set theory: A review, *Journal of Physics: Conference Series*, vol. 1878, no. 1, p. 012054, 2021.

Kangkang Huang received the bachelor degree from Jiangxi University of Science and Technology in 2020. He is currently pursuing the master degree at the School of Energy and Mechanical Engineering, Nanjing Normal University. His main research direction is the application of data mining and machine learning algorithms in building thermal comfort and energy consumption.

Shihua Lu received the bachelor degree from Yangzhou University in 1996, the master degree from Nanjing University of Aeronautics and Astronautics in 2007, and the PhD degree from Nanjing University of Aeronautics and Astronautics in 2012. He is currently an associate professor at the School of Energy and Mechanical Engineering, Nanjing Normal University. His main research direction is the application of artificial intelligence in the HVAC and building energy saving and intelligent control.

Xinjun Li received the bachelor, master, and PhD degrees from Nanjing University of Aeronautics and Astronautics in 2011, 2014, and 2018, respectively. He is currently a lecturer at the School of Energy and Mechanical Engineering, Nanjing Normal University. His main research directions are active chip cooling, active flow control, enhanced heat transfer of aero-engines, and artificial intelligence.

Ke Feng received the bachelor degree from Nanjing University of Technology Pujiang College in 2020. He is currently pursuing the master degree at the School of Energy and Mechanical Engineering, Nanjing Normal University. His main research direction is the application of model predictive control and machine learning algorithms in construction.

Weiwei Chen received the bachelor, master, and PhD degrees from Nanjing University of Aeronautics and Astronautics in 2009, 2012, and 2016, respectively. He is currently a lecturer at the School of Energy and Mechanical Engineering, Nanjing Normal University. His main research directions are supercritical pressure fluid heat transfer, aircraft environmental control, and artificial intelligence.

Yi Xia received the bachelor, master, and PhD degrees from Southeast University in 2001, 2004, and 2015, respectively. He is currently an associate professor at the School of Energy and Mechanical Engineering, Nanjing Normal University. His main research direction is artificial intelligence.