

Using machine learning algorithms to multidimensional analysis of subjective thermal comfort in a library

Citation for published version (APA):

Song, G., Ai, Z., Zhang, G., Peng, Y., Wang, W., & Yan, Y. (2022). Using machine learning algorithms to multidimensional analysis of subjective thermal comfort in a library. *Building and Environment*, 212, Article 108790. <https://doi.org/10.1016/j.buildenv.2022.108790>

Document license:
TAVERNE

DOI:
[10.1016/j.buildenv.2022.108790](https://doi.org/10.1016/j.buildenv.2022.108790)

Document status and date:
Published: 15/03/2022

Document Version:
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.



Using machine learning algorithms to multidimensional analysis of subjective thermal comfort in a library

Ge Song^{a,b}, Zhengtao Ai^{a,b}, Guoqiang Zhang^{a,b,*}, You Peng^c, Wei Wang^d, Yan Yan^d

^a College of Civil Engineering, Hunan University, Changsha, Hunan, 410082, China

^b National Center for International Research Collaboration in Building Safety and Environment, China

^c Urban Planning and Transportation Research Group, Department of the Built Environment, Eindhoven University of Technology, P.O. Box 513, 5600MB, Eindhoven, the Netherlands

^d School of Architecture, Hunan University, Changsha, 410082, China

ARTICLE INFO

Keywords:

Thermal comfort
Machine learning
Principal component analysis
Multidimensional psychological parameters
Public building

ABSTRACT

The thermal comfort in public buildings is affected by multiple psychological and physical factors. A deep understanding of these factors is necessary for intelligent ventilation control and architectural design. In this study, it was quantified the physiological and psychological parameters of the occupants in a library situated in Changsha using a questionnaire. Then, with the use of Principal Component Analysis (PCA), the dimensionality of the database was reduced. We found that the Zero-mean and unit variance projection is better than the Zero-mean projection, and 43-dimensional data can replace more than 90% of the original 61-dimensional data. Machine learning algorithms were used to analyze the results after PCA, which showed that the effect of thermal comfort on emotions is greater than that of emotions on thermal comfort. In addition, the performance of six machine learning algorithms (Linear Regression (LR), Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), Classification and Regression Trees (CART), Gaussian Naive Bayes (NB), and Support Vector Machine (SVM)) were compared. It is found that the predicted results of LDA were more accurate, and other algorithms showed different performances in different cases. These findings can contribute to the study of the subjective and objective feelings of indoor thermal comfort in public buildings, thereby guiding architectural design, intelligent control of ventilation systems, and realizing human-building interaction interfaces.

1. Introduction

Energy consumption in buildings accounts for nearly 40% of global energy consumption [1], especially in the US and UK [2,3]. A large amount of this energy is required to maintain a comfortable indoor thermal environment. Compared with residential and office buildings, public buildings need more energy to maintain the indoor environment. For example, the Heating, Ventilation and Air Conditioning (HVAC) systems in commercial buildings consume 40% of energy [4]. The energy consumption of large-scale public buildings with more than 20,000 m² gross floor area in China had reached 100 billion kWh per year which only constituted less than 4% of the national urban building area while normal public buildings in China had reached 160 billion kWh per year [5]. However, even with the huge energy consumed, people are still dissatisfied with the indoor environment. In addition, and especially important in public buildings on campus, a comfortable indoor

environment is a critical condition to ensure the health and to favor the learning of the students [6–9].

Defining thermal comfort requirements from the occupants is essential for energy saving. The current definition of thermal comfort is mostly based on the environment and human physiological parameters, ignoring the influence of psychological parameters of occupants [10]. Thermal comfort is affected by multiple factors of human psychological and physiological parameters such as individual differences (age, gender, height, weight), thermal inertia, thermal preference, mood, and building decoration [11–14]. Compared with the conventional HVAC systems based on indoor environmental parameters, temperature, humidity, and Carbon Dioxide (CO₂) concentration, the indoor environment control based on the real-time feedback of the occupants has the potential to become more energy-efficient and effective for the indoor environment adjustment of public buildings.

As a powerful analysis method for nonlinear problems, Machine

* Corresponding author. College of Civil Engineering, Hunan University, Changsha, Hunan, 410082, China.

E-mail address: gqzhang@hnu.edu.cn (G. Zhang).

<https://doi.org/10.1016/j.buildenv.2022.108790>




Received 31 August 2021; Received in revised form 23 December 2021; Accepted 10 January 2022

Available online 19 January 2022

0360-1323/© 2022 Elsevier Ltd. All rights reserved.

Table 1

The measurement parameters and instruments used in field study.

Physical quantity	Instrument	Range	Accuracy	
Air temperature and Relative humidity	UX100-003	20 °C–70 °C 15%–95%	0–50 °C: ± 0.21 °C; 25%–85%: $\pm 3.5\%$; <25%/>85%: $\pm 5\%$	
Mean radiant temperature (Globe temperature)	JTR-04	5–120 °C	± 0.5 °C	
CO ₂ concentration	T7001	0–10000 ppm	± 50 ppm $\pm 3\%$	

Learning (ML) is usually used to predict the thermal comfort of occupants [15–23]. Previous studies on thermal comfort are mostly based on Fanger's human thermal balance assessment model [10,24], however, those studies were performed under the static and conditions laboratory studies [25]. Compared with the physiological parameters of the human body, environmental parameters have always been the main factor in analyzing thermal comfort due to their advantages, like a relatively easy-to-obtain data [26]. This has led to excessive ignorance of individual differences in the process of obtaining the thermal requirements of the occupants. The excellent calculation and self-learning capabilities of ML can be fully exploited as a thermal comfort analysis tool. Logistic Regression (LR) [27–30], Linear Discriminant Analysis (LDA) [31–33], K-Nearest Neighbors (KNN) [34–36], Classification and Regression Trees (CART) [37], Gaussian Naive Bayes (NB) [38], and Support Vector Machines (SVM) [23,39] where the six ML algorithms considered for the analysis of multiple problems. These 6 methods are simple representations of linear (LR and LDA) and nonlinear (KNN, CART, NB and SVM) ML algorithms [40].

Previous studies show that the space, culture, seasons, and other factors of public buildings affect thermal comfort by acting on physiology and psychology [41,42]. However, little research has systematically studied the relationship between the attitudes of the occupants in campus building space, emotion perception, and thermal comfort. This paper takes a university library in Changsha (Hunan), and uses machine learning methods to analyze the psychological and physiological state of the subjects from multiple dimensions attributes and the influence of the public building space environment on the comfort of occupants.

1.1. Positive and negative affect schedule (PANAS) and WHO-5 Well-Being Index

PANAS is a psychological measurement scale proposed by David Watson, Lee Anna Clark, and Auke Tellegen in 1988 [43]. The scale was designed with 20 5-point questions based on positive and negative emotions, which is widely used to quantitatively evaluate individual emotions. The WHO-5 Well-Being Index consists of 5 questions, which were developed by the WHO European Regional Office to assess positive and negative well-being [44]. This paper uses these two parts (PANAS and WHO-5 Well-Being Index) to show the psychological factors of human thermal comfort.

1.2. Environmental assessment, adaptability and preferences

To ensure accurate feedback of the occupants, the thermal comfort assessment indicators collected in this study also include environmental assessment, thermal adaptation, and thermal preference [45], including the evaluation of indoor thermal environment, ventilation, air humidity, light, noise, air quality, and interior design [46]. Adaptability indicates whether you can adjust your thermal comfort by changing the current state. Environmental preference expresses the demands for adjustment of the current environment. The specific details will be explained in section 2.

2. Field survey

2.1. Measurement equipment and survey questionnaires

The indoor physical environment measurement and the questionnaire survey of occupants were carried out at the same time. T_a , RH , and

Table 2

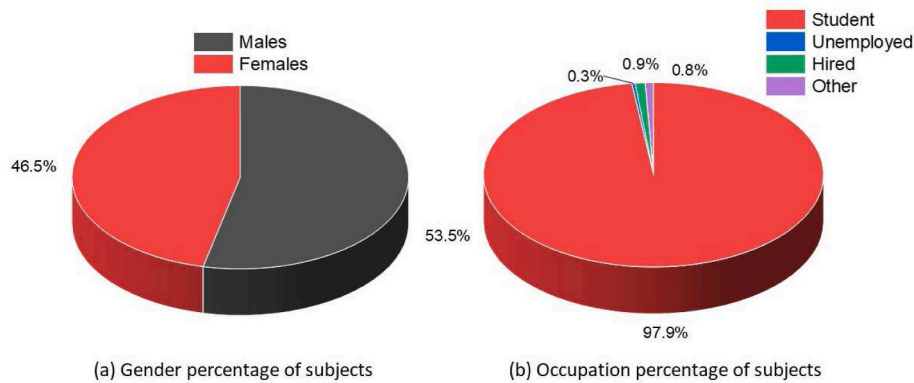
Statistics of the input parameters related to the PMV model.

	T_a (°C)	RH (%)	CO_2 (l)	MRT (°C)	Age	Height (cm)	Weight (kg)	Clo
mean	20.39	51.95	928.27	19.55	21.43	170.31	61.47	1.15
std	4.60	12.56	330.67	4.54	5.37	8.76	13.66	0.54
min	10.72	28.71	217.9	8.8	10	116	42	0.02
25%	16.33	42.14	623.9	15.125	19	164	51	0.78
50%	22.397	50.07	857.45	21.7	20	170	60	1.15
75%	23.615	59.74	1231.4	22.9	22	177	70	1.50
max	33.287	81.23	2050.7	25.9	57	199	86	4.05

Table 3

6 parts correlation matrix title.

Part1 various environmental and occupants parameters and related to the PMV model									
Ta	RH	CO ₂	MRT	Gender	Age	Height	Weight	Clo	TCV
Part2 Correlation matrix 7 assessments of the indoor environment			indoor thermal sensation at the moment						TES
			indoor relative humidity at the moment						RHS
			indoor ventilation sensation at the moment						VS
			indoor light sensation at the moment						LS
			indoor audio sensation at the moment						AS
			indoor IAQ sensation at the moment						IAQS
			indoor design sensation at the moment						IDS
Part3 Correlation matrix of 4 environmental preferences			hope for thermal environment						HTE
			hope for indoor RH environment						HRH
			hope for indoor ventilation						HV
			hope for indoor light environment						HL
Part4 Correlation matrix of the 5 attributes of the WHO-5 Happiness Index			I have felt cheerful and in good spirits						CS
			I have felt calm and relaxed						CR
			I have felt active and vigorous						AV
			I woke up feeling fresh and rested						FR
			my daily life has been filled with things that interest me						LI
Part5 Correlation matrix of evaluation of the adaptability of 7 indoor spaces			increase comfort from changing location						CL
			increase comfort from changing cloth						CC
			increase comfort from changing environment such as Ta, RH, ventilation, light and noise						CE
			increase comfort from providing facilities						PF
			others activities interfere the comfort						IC
			green plants provided here will increase the comfort						PC
			the purpose of coming here has been satisfied						PS
Part6 Correlation matrix of 20 emotional characteristics of PANS emotional assessment (10 positive and 10 negative)									
Interested	Alert	Excited	Inspired	Strong	Determined	Attentive	Enthusiastic	Active	Proud
Irritable	Upset	Ashamed	Distressed	Nervous	Guilty	Jittery	Hostile	Scared	Afraid

**Fig. 1.** Distribution of subjects by gender and occupation.

MRT have been selected to evaluate the Indoor Thermal Environment (ITE). CO₂ concentration has been selected to evaluate IAQ. The instruments and detailed parameters are shown in Table 1.

The corresponding questionnaire survey was divided into 9 parts: 1. The basic information of the occupants includes which includes gender, age, weight, height, nationality, long-term residence marital status, education level, work status, housing status, and monthly income. 2. The status of the subjects caused by different building functions includes how long the subjects stay, how long they plan to stay, whether they are accompanied, whether air-conditioning is frequently used in their residence, whether they stayed in an air-conditioned room before going to the library, the state of the subjects before the questionnaire (e.g. standing, seated, walking, or practicing sport), the distance from the residence to the library, the frequency of going to the library, and the mean of transportation they used. 3. The thermal resistance of the clothes based on what the subjects were wearing. 4. PANAS emotional assessment. 5. WHO-5 Happiness Index as shown in part 4 of Table 3. 6. The indoor environment assessment as shown in part 2 of Table 3. 7. The adaptability assessment considering the points of part 5 in Table 3. 8.

Environmental preferences as provided in part 2 of Table 3. 9. The general thermal comfort vote (TCV).

2.2. Buildings and subjects

This study was conducted in a library from Hunan University, which is located in Changsha at latitude 8°41'N and longitude 114°15'E, a typical city in hot summer and cold winter zone of China according to the division of five climate zones in China [47], and humid subtropical climate according to the Köppen climate classification [48]. The library has 8 floors with a height of 4.5 m each floor and a total indoor area of 35,000 m². The transition season, autumn in 2020, has been select to avoid too much difference between indoor and outdoor environment parameters. Collected environmental feedback records of 1162 occupants including 622 males and 540 females, aged between 10 and 50 years old, are mostly students or teachers of Hunan University. Table 2 shows a summary of the input parameters related to the PMV model. The distribution of subjects by gender and occupation is shown in Fig. 1 and their status can be found in Fig. 2 and Fig. 3.

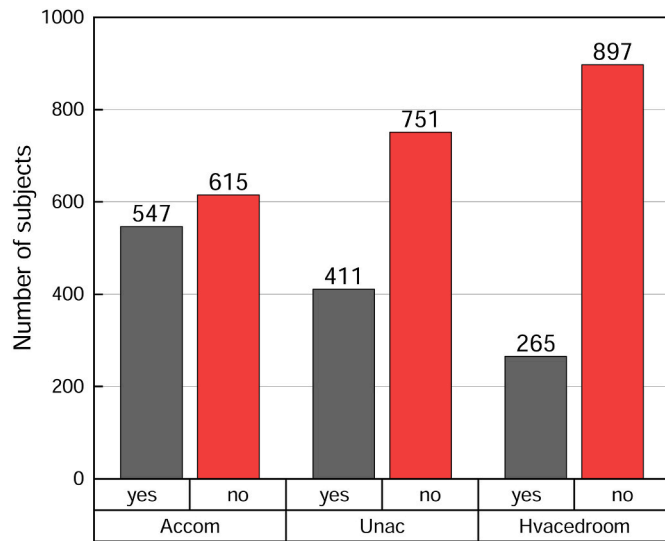


Fig. 2. Subject's status: accompanied by others (Accom), the residence is often air-conditioned (Unac), and stayed in an air-conditioned room before coming (Hvacedroom).

It is mentioned in the most widely used Predicted Mean Vote (PMV) model proposed by Fanger [10] in 1970 that six factors are affecting the thermal comfort of the human body in buildings, and four environmental factors: air temperature (T_a), Relative Humidity (RH), Mean Radiant Temperature (MRT) and air velocity; two personal factors: clothing insulation (Clo) [49] and metabolic rates of activity (MET). Considering that the special time of the COVID-19 epidemic, the Indoor Air Quality (IAQ) affected by the ventilation rate of public buildings need more attention. In this study, CO_2 concentration was selected as an important indicator to evaluate IAQ instead of air velocity. T_a , RH , MRT , and CO_2 are recorded every second to ensure that it can be corresponded to the time of each questionnaire. The measurement instruments were

placed at 1.1 m high in the middle of test room. MRT was calculated by air temperature, globe temperature and air velocity according to the function in ISO 7726 [50]:

$$MRT = \left[(GT + 273)^4 + \frac{1.1 \times 10^8 \times v_a^{0.6}}{\epsilon D^{0.4}} (GT - T_a) \right]^{1/4} - 273 \quad (1)$$

where MRT is the mean radiant temperature ($^{\circ}C$), GT is the globe temperature ($^{\circ}C$), v_a is the air velocity at the level of the globe (m/s), ϵ is the emissivity of the globe (no dimension), D is the diameter of the globe (m), and T_a is air temperature ($^{\circ}C$). Since this experiment use the standard globe, so $D = 0.15$ m, and $\epsilon = 0.95$. In addition, $v_a = 0.2$ m/s was used according to indoor air design parameters of public buildings in Chinese Standard [51].

The Clo is obtained from the clothing condition part of the questionnaire. This part is divided into 8 parts according to the type of clothes, a total of 46 types of Clo , and the final result is the sum of the Clo of the subjects. The metabolic rate is difficult to accurately measure, but it has a great correlation with the subject's gender, height, weight, exercise status, etc. Therefore, we added several features that may have an impact on the metabolic rate in the questionnaire.

3. Data analysis

3.1. Correlation matrix

Since the selected public building is a university library, the subjects are mostly students, and the attributes such as marital status, education level, housing situation and monthly income distribution of these subjects are relatively concentrated, which can be ignored to save computing resources because of the small effect on comfort prediction. A total of 60 features related to thermal comfort are required. The 60-dimensional features are selected to consider the thermal satisfaction of the environment from the individual differences, emotions, environmental preferences, and other aspects of occupants as much as possible. The correlation between the features should be first analyzed because of the big quantity of attributes. All correlation analyses are divided into 6

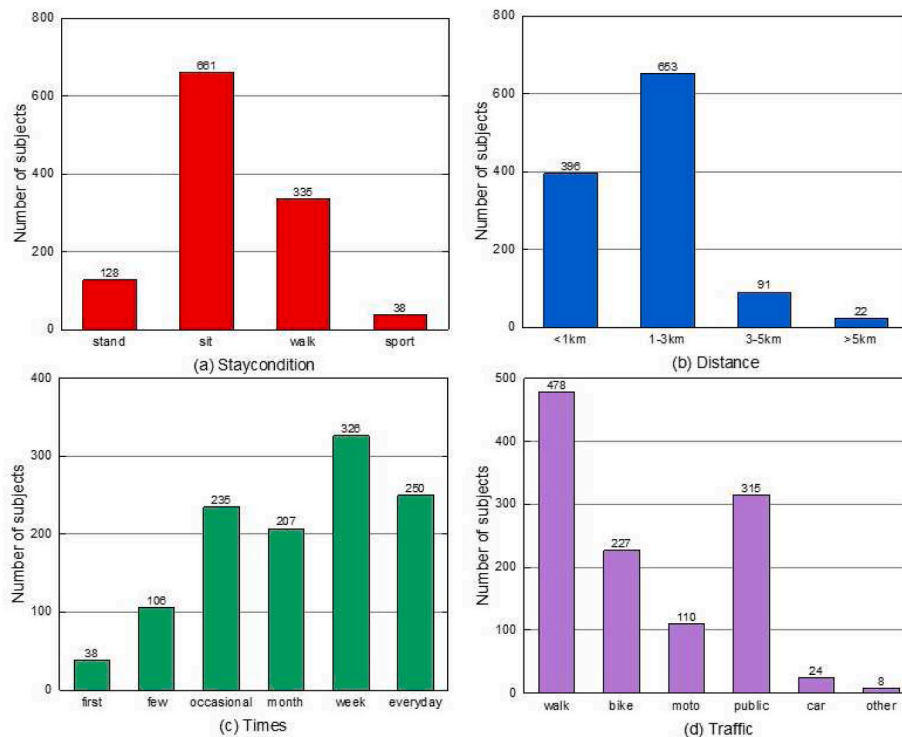


Fig. 3. (a) The stay condition of the subjects, (b) the distance from the residence to library, (c) how often they to the library, and (d) the mean of transport they use.

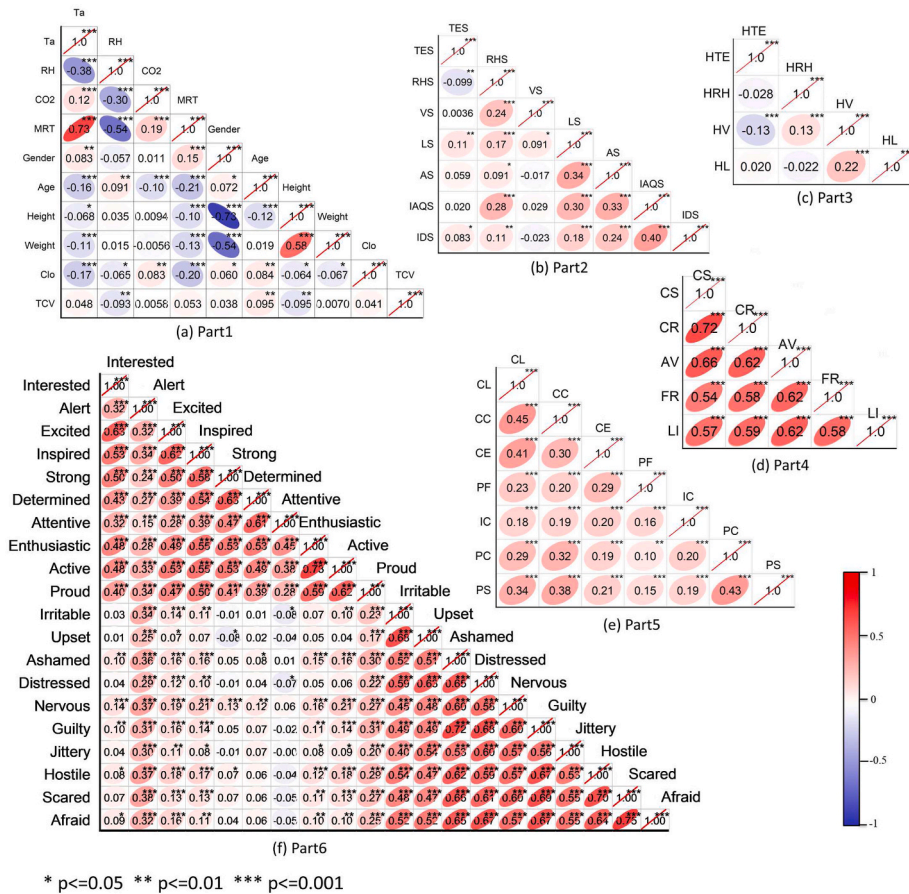


Fig. 4. The correlation analysis results and significant mark (* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$) of (a) Part1, (b) Part2, (c) Part3, (d) Part4, (e) Part5, and (f) Part 6.

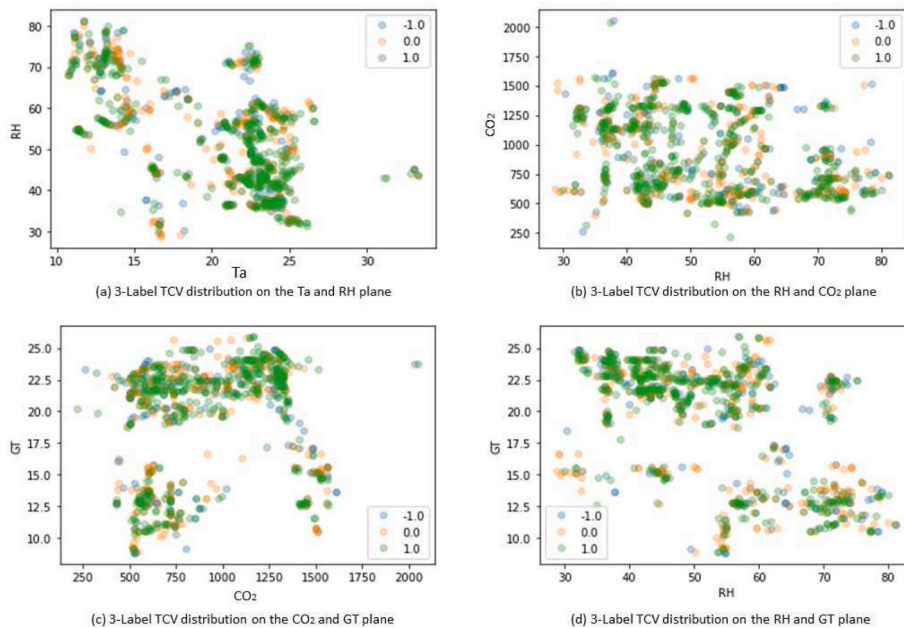


Fig. 5. 3-Label TCV distribution on the different 2-dimensional environmental parameters planes.

parts as shown in Table 3, and the results as shown in Fig. 4 and Fig. 5.

The correlation between the TCV and the environmental parameters, and the correlations between each environmental parameter are small.

MRT has a greater correlation with Ta (0.73) and RH (−0.54). It is also higher between gender and height (−0.73), gender and weight (−0.54), and between height and weight (0.58). The adaptability and preference

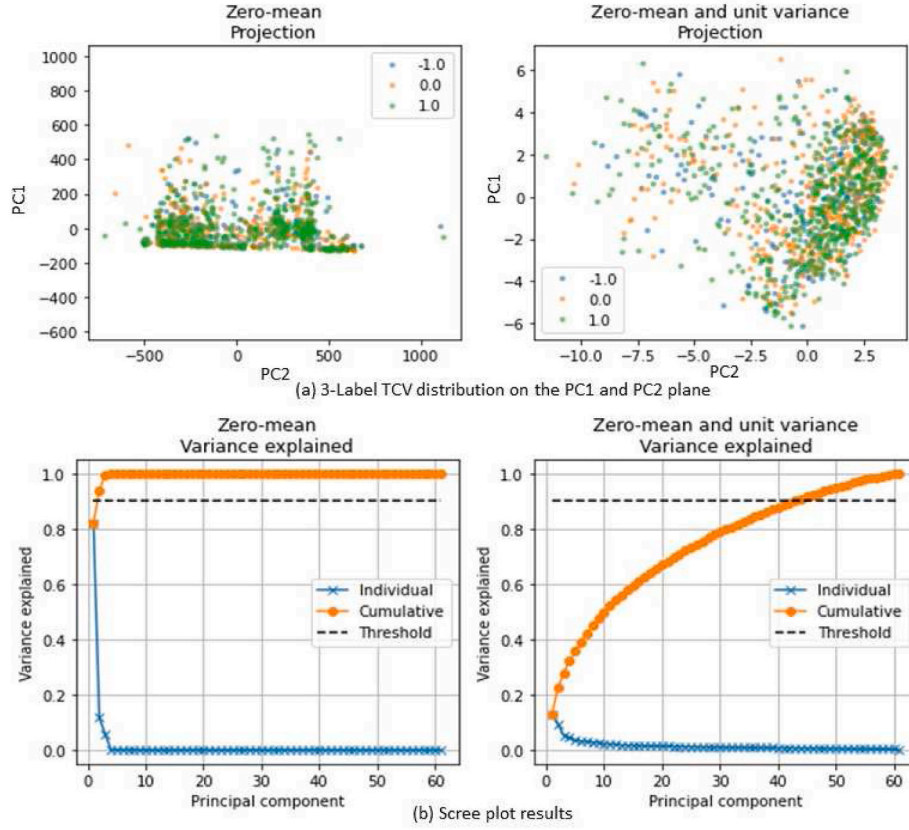


Fig. 6. PCA results using zero-mean and zero-mean and unit variance respectively.

for thermal, humid, ventilated, light, and acoustic environments are also less correlated. Interestingly, 7 indicators of environmental adaptability are positively correlated. Among the 20 PANAS sentiment evaluation indicators, there is a strong correlation between different positive sentiments as well as between different negative sentiments. However, the correlation between a positive and a negative sentiment from the set is not significant. Similarly, there is also a strong positive correlation between each of the WHO-5.

Therefore, dimensionality reduction can be performed by reducing the characteristics of the indicators with higher correlation if the dimensionality of the influencing factor is too high to analyze. The attributes with higher correlation are mostly emotional evaluation indicators and happiness indicators, the better way to dimensionality reduce is to transfer the 20-D emotional matrix and 5-D happiness matrix. This process assumes that the occupant's TCV result is affected by emotions, but in practice, it is difficult to clearly define whether the comfort vote result is affected by emotions or emotions vote affected by the comfort.

3.2. Principal component analysis (PCA)

Another method that could be used to solve this problem, the principal component analysis (PCA), which could transform 60-D raw data to observe each attribute. PCA is the most effective method for data dimensionality reduction, and it is very convenient to observe the classification of data before using machine learning.

Adding the 61st dimension attribute TCV, the entire data is a 1162×61 matrix. The attributes matrix was created in Python 3.8, and then, ordered by classification label. Since 60-D attributes are difficult to observe in a distribution of data through two-dimensional visualization images, as shown in Fig. 5, PCA was used.

The first step of PCA is dataset standardization, each attribute that needs to be ensured has a mean of zero and variance of one:

$$Y = X - 1\mu \quad (2)$$

Where μ is a (row) vector containing the mean value of each attribute and 1 is a column vector of ones in all entries) and then calculating the Singular Value Decomposition (SVD) of the zero mean data.

$$Y = USV^T \quad (3)$$

$$\rho_m = \frac{S_{mm}^2}{\sum_{m'=1}^n S_{m'm'}^2} \quad (4)$$

Where ρ_m is how much of the variation in the data each PCA component accounts for, and S_{mm}^2 is the squared singular values.

The results of PCA are shown in Fig. 6, where the coefficients have been plotted as vectors in the principal component space to interpret the principal directions, and the PC represented by PC1 and PC2 have exceeded 90% in the zero-mean figure, while in the Zero-mean and unit variance figure, the combined principal components of PC1-PC43 exceed 90%. However, the dataset show more disperses in Zero-mean and unit variance figure projection.

It can be seen that although the visibility of the data processed by PCA is greatly improved, the classification trend according to TCV is still not obvious. In the same way, it can be observed the classification of data based on gender, uncomfortability, etc.

It's easy to observe from the gender and uncomfortable binary classification distribution on PC1 and PC2 plane, that the data classification status of the questionnaire survey is not very satisfactory, as shown in Fig. 7. It seems that the randomness of the subjects is too high since it is difficult to control the occupants of a public building. Furthermore, natural ventilation, which is used too much in the transition season, also increases the control difficulty of the indoor environment. The excess of attributes considered make a strong non-

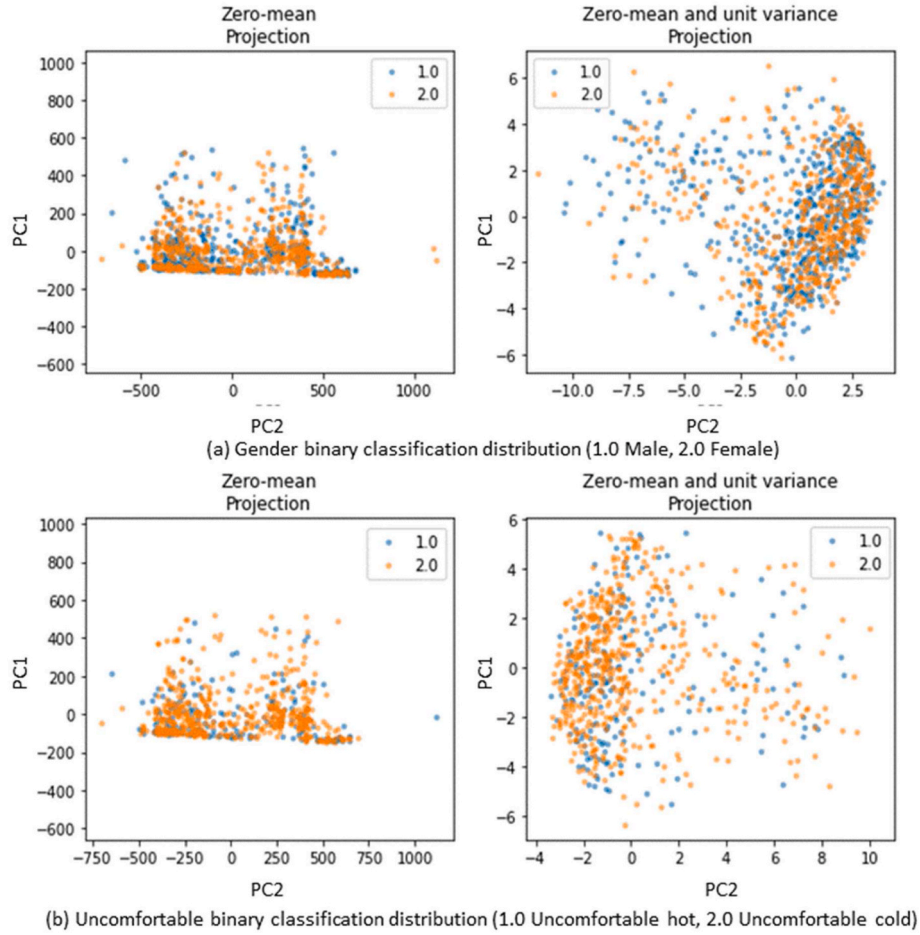


Fig. 7. Gender and uncomfortable binary classification distribution on PC1 and PC2 plane.

linearity, which will become a challenge to machine learning algorithms. However, in practice, it is very common to not get very ideal data. How to optimize the algorithm and solve problems in such an unsatisfactory datasets is the biggest challenge in machine learning research.

4. Machine learning models

Logistic regression (LR) is a linear regression model used in machine learning, which aims at solving a classification problem. It is used to deal with the regression problem of the dependent variable as a categorical variable [27,52,53].

Linear Discriminant Analysis (LDA) is a method to realize the classification of two or more object features, and is applied in the fields of data statistics, pattern recognition, and machine learning [31].

K-Nearest Neighbors (KNN) algorithm predicts new data points by searching the entire training set of the K most similar instances (neighbors) and summarizing the output variables of those K instances [54]. The Euclidean distance between two points $x_1 (x_{11}, x_{12}, \dots, x_{1n})$ and $x_2 (x_{21}, x_{22}, \dots, x_{2n})$ in n-dimensional space is:

$$d_{12} = \sqrt{\sum_{k=1}^n (x_{1,k} - x_{2,k})^2} \quad (5)$$

And the Chebyshev distance:

$$d_{12} = \max(|x_{1i} - x_{2i}|), \quad i (i = 1..n) \quad (6)$$

Classification and Regression Trees (CART) is a binary tree structure model, which can be used to solve classification problems or regression

problems [55]. In the process of building a tree, impurity is used to measure the quality of node selection. The higher the purity, the better the classification effect. This is measured by the Gini Impurity,

$$Gini = 1 - \sum_{i=1}^n p_i^2 \quad (7)$$

Where p_i represents the probability value in the discrete probability distribution.

The Gaussian Naive Bayes (NB) is a classification method based on Bayes' theorem and the assumption of independence of characteristic conditions [56]. The probability density function of the normal distribution is:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (8)$$

The basic model of Support Vector Machine (SVM) is to find the best separation hyperplane in the feature space to maximize the interval between positive and negative samples on the training set [57].

The expression of the hyperplane that can correctly divide the positive and negative samples is defined as:

$$\omega^T x + \gamma = 0 \quad (9)$$

Where ω is the normal vector of the hyperplane, the parameter $\frac{\gamma}{\|\omega\|}$ determines the offset from the origin to the hyperplane along the normal vector ω .

Table 4
7-Label TCV proportion.

Label	-3	-2	-1	0	1	2	3
Proportion	0.2%	1.4%	12.4%	29.8%	37.3%	16.5%	2.4%

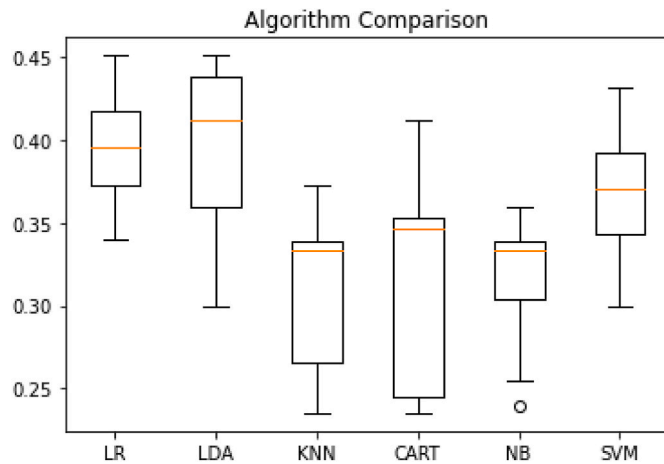


Fig. 8. 7-Label TCV prediction result algorithm comparison box plot.

5. Results

The 10-fold cross validation was chosen to estimate model accuracy. The random seed was set via the random state argument to fixed number to ensure that the same splits of the training dataset in each algorithm. The hyperparameters of all the models in this study were optimized experimentally using this training data. For LR, elastic net regression for penalty, LIBLINEAR solver, and binary multiclass classification were chosen. For LDA, singular value decomposition with an absolute threshold of 10^{-4} was used. For KNN, 5 neighbors with a uniform weighting and a leaf size of 30 were chosen. For CART, it was used the criterion Gini impurity with a maximum depth of the tree of 10, a minimum number of samples required to split the internal nodes of 5, and a minimum number of samples required to be at a leaf node of 2. For NB, the default parameters of the scikit-learn package in Python were used. Finally, for SVM, the regularization parameter was increased to 2.

5.1. 7-Label prediction

As shown in Table 4, in the original data TCV is divided into 7 levels, due to the transitional season of hot summer and cold winter in China, it can be seen that the Neutral 0 and Slightly Warm 1 accounts for 29.8% and 37.3%, more than half of the overall data.

We used six machine learning algorithms LR, LDA, KNN, CART, NB, SVM to predict TCV through other 60-dimension attributes. Fig. 8 shows the box plot distribution of the accuracy of the prediction results. Table 5 shows the predicted result of the mean and standard deviation of five machine learning algorithms. Table 6 shows the precision, recall, f1-score, support, micro avg, macro avg, and weighted avg of the predicted results. Precision: The correct prediction is positive, which accounts for the proportion of all positive predictions. Recall: the correct prediction is positive, the proportion of all is positive. f1-score: the harmonic average of precision and recall. support (the number of

samples in each classification or the total number of samples in the test set). micro avg (micro average): the average of all data results. macro avg: The average value of all tag results. weighted avg (weighted average): the weighted average of all tag results. Fig. 9 shows the confusion matrix of the 7-label prediction results. Since very few samples of -3 were obtained, the validation is not randomly assigned, and the prediction result becomes 6 labels. The x-axis and y-axis represent the true label and the predicted label, respectively, and 1, 2, 3, 4, 5, 6 represent -2 level, -1 level, 0 level, 1 level, 2 level, 3 level, respectively. It can be seen from the calculated prediction results that the accuracy is very low, and it cannot meet the expected accuracy requirements at all. The confusion matrix in Fig. 9 shows that due to the uneven distribution of the sample size, the prediction accuracy in the 0 and 1 level is higher. In the 7-label classification prediction calculation, LR showed the best result, and KNN, CART, NB all showed poor predictions.

5.2. 3-Label prediction and 2-label prediction

Since the prediction accuracy of 7 labels is very low, the analysis may be caused by uneven samples. Usually one of the ways to improve the accuracy of machine learning algorithms is to reduce the number of predicted labels [58]. This section categorizes all cool sensations (<0) as

Table 6
7-Label precision, recall, f1-score, support, micro avg, macro avg and weighted avg of the predicted results.

	precision	recall	f1-score	support
-2	0	0	0	2
-1	0.17	0.06	0.09	16
0	0.29	0.27	0.28	33
1	0.4	0.7	0.51	50
2	0.5	0.05	0.08	22
3	0	0	0	4
accuracy			0.36	127
Macro avg		0.23	0.18	0.16
Weighted avg		0.34	0.36	0.3

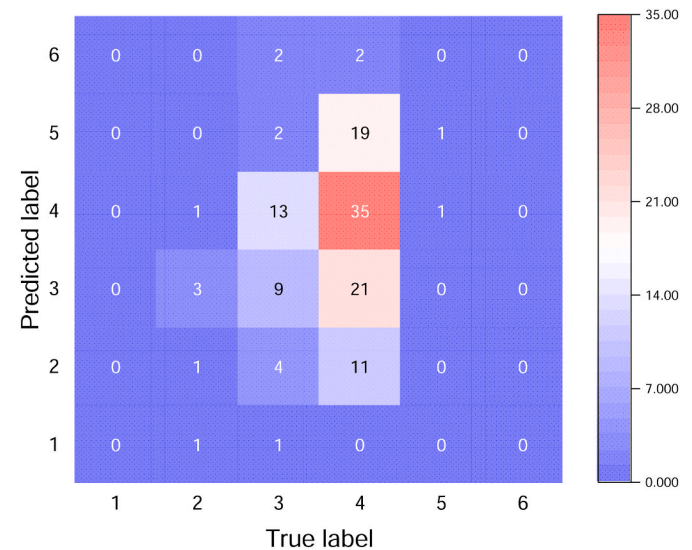


Fig. 9. 7-Label TCV prediction result confusion matrix.

Table 5
7-Label predict result of mean and standard deviation of 6 machine learning algorithms.

	LR	LDA	KNN	CART	NB	SVM
Mean	0.393098	0.394980	0.310275	0.316196	0.316157	0.369373
Std	0.035904	0.052527	0.047499	0.065987	0.038142	0.037582

Table 7
3-Label and 2-Label TCV proportion.

Label	-1	0	1
3-Label Proportion	13.5%	29.8%	56.7%
2-Label Proportion	/	48.7%	51.3%

-1, all warm sensation (>0) as +1, the neutral (0) unchanged, and again use these 5 machine algorithms to predict TCV. The 3-Label Proportion in Table 7 3-Label and 2-Label TCV proportion shows the proportion of each label, which is more even than that of 7 labels, but level 1 still accounts for too much. Therefore, this part merges the 0-level and -1 level and removes some 1 level values, so that the classification is relatively uniform, as shown in 2-Label Proportion of Table 7. From the box plot of the algorithm accuracy comparison results in Fig. 10 and the confusion matrix in Fig. 11, it is obvious that the prediction result of 2-labels is better than that of 3-labels, and in the binary classification, the comparison of the results of each algorithm is also more significant. In the 3-labels, the SVM algorithm shows the highest accuracy and the smallest variance respectively. This may be because other algorithms are mostly used for the prediction of two classification problems. After changing to the two classification problem, the results of LR, LDA, and

NB has been significantly improved. The 3-Label and 2-Label predict results of the mean and standard deviation of five machine learning algorithms are shown in Table 8. The 3-Label and 2-Label precision, recall, f1-score, support, micro avg, macro avg, and weighted avg of the predicted results are shown in Table 9.

The cost of results quality improvement of 3-label and 2-label is ignoring more nuanced useful information that can distinguish different degrees of the cold and hot sensation of the occupants, which in practice implies whether to choose higher TCV prediction result accuracy or more levels in the responses of the occupants. When the request of the thermal comfort system is just to turn on/off the cooling or heating, it is better to choose higher TCV prediction. However, when the request is to know the cooling/heating power, more details about the responds of the occupants are needed (even though more computational resources are needed).

5.3. Binary classification prediction of emotion and happiness

To explore the influence of factors such as environment and comfort level on emotion and happiness, the results of emotion and happiness are used as prediction labels for the machine learning algorithms. It was

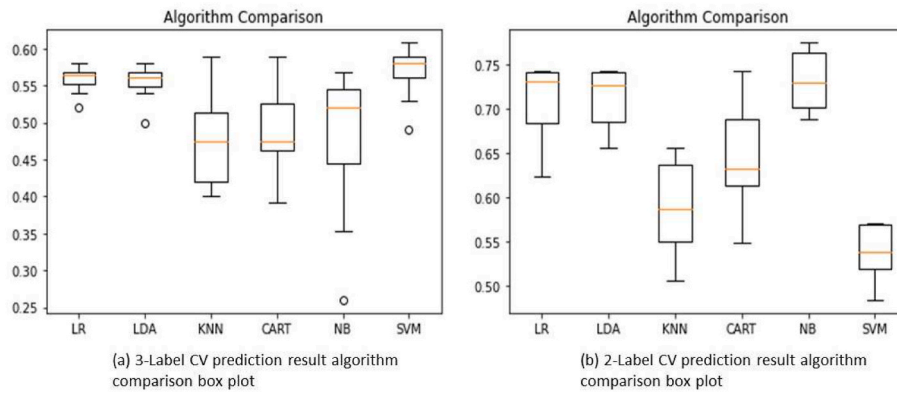


Fig. 10. 3-Label and 2-Label TCV prediction result algorithm comparison accuracy box plot.

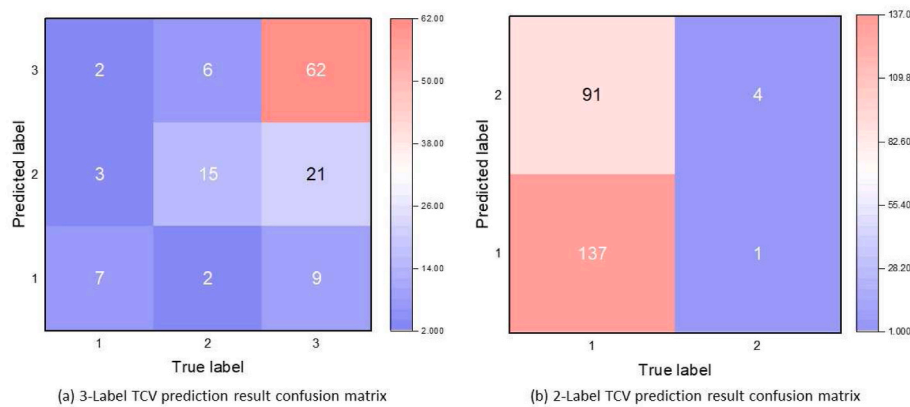


Fig. 11. 3-Label and 2-Label TCV prediction result confusion matrix.

Table 8
3-Label and 2-Label predict result of mean and standard deviation of six machine learning algorithms.

		LR	LDA	KNN	CART	NB	SVM
3-Label	Mean	0.55835	0.55439	0.47498	0.48682	0.4749	0.56839
	Std	0.01671	0.02133	0.06023	0.05724	0.0945	0.03375
2-Label	Mean	0.7094	0.7137	0.58662	0.64263	0.73089	0.53824
	Std	0.03914	0.03062	0.05321	0.05478	0.03155	0.02896

Table 9

3-Label and 2-Label precision, recall, f1-score, support, micro avg, macro avg and weighted avg of the predicted results.

	3-Label	precision	recall	f1-score	support
3-Label	−1	0.29	0.11	0.16	18
	0	0.33	0.13	0.19	39
	1	0.59	0.89	0.71	70
	accuracy			0.54	127
	macro avg	0.4	0.38	0.35	127
	weighted avg	0.47	0.54	0.47	127
2-Label	0	0.6	0.99	0.75	138
	1	0.8	0.04	0.08	95
	accuracy			0.61	233
	macro avg	0.7	0.52	0.41	233
	weighted avg	0.68	0.61	0.48	233

used PCA to reduce the dimensions of 10-dimensional positive emotions, 10-dimensional negative emotions, 5-dimensional happiness index, and 25-dimensional emotions and the sum of happiness, as shown in Fig. 12. Two principal components are selected, and the proportions of PC1 and PC2 after PCA processing are shown in Table 10. The PCA results of positive emotions and negative emotions are similar. The WHO-5 happiness index has the best PCA results, while the sum of emotions and happiness has the worst PCA results. This may be because the higher correlation features are more suitable for PCA for dimensionality reduction. We only selected the results of PC1 for binary classification in this part. The PC1 results less than 0 are classified as 0 labels, and greater than 0 are classified as 1 label.

When the prediction object becomes the emotion and happiness of the binary classification, the quality of the prediction result improves significantly, as shown in Fig. 13. It can be seen that the prediction results on emotion and happiness of Fig. 13(d), are the best quality prediction results, and the happiness prediction results are the worst prediction quality. These results show the opposite trend with the PCA result, which shows that in this case, the influence of the dimensionality

on the result of the machine learning algorithm is greater than the proportion of the principal component of the PCA result. Attributes with less dimensionality and higher correlation will have a higher advantage in the PCA results. PC1 is more representative of the overall data. However, compared to the PCA quality, reducing the dimensionality is more important in the application of machine learning algorithms.

6. Conclusions and future research

In this paper, machine learning algorithms were used to multi-dimensionally analyze the comfort of the occupants in a library to know the subjective indicators of thermal comfort, which is usually overlooked in the research of thermal comfort of public buildings. We found that in thermal comfort research, psychological factors mostly defaulted as independent variables. However, the impact of thermal comfort on subjective feelings is greater than that of subjective feelings on thermal comfort.

Machine learning algorithms show computational advantages in the application of multi-dimensional evaluation of thermal comfort. In contrast, the SVM algorithm has more advantages in solving multivariate classification problems, and its advantages in binary classification problems have declined. LDA has the best application effect due to the sufficient number of samples in this experiment. Since the CNRT algorithm is good at processing progressive questionnaires, CNRT and KNN did not have a high computational advantage in this case. NB performs better when dealing with binary classification problems, but it performs

Table 10

Proportion of PC1 and PC2.

	Positive emotions	Negative emotions	WHO-5 happiness	Emotions and happiness
PC1	52.31%	59.78%	67.89%	38.71%
PC2	10.94%	8.31%	10.91%	15.03%

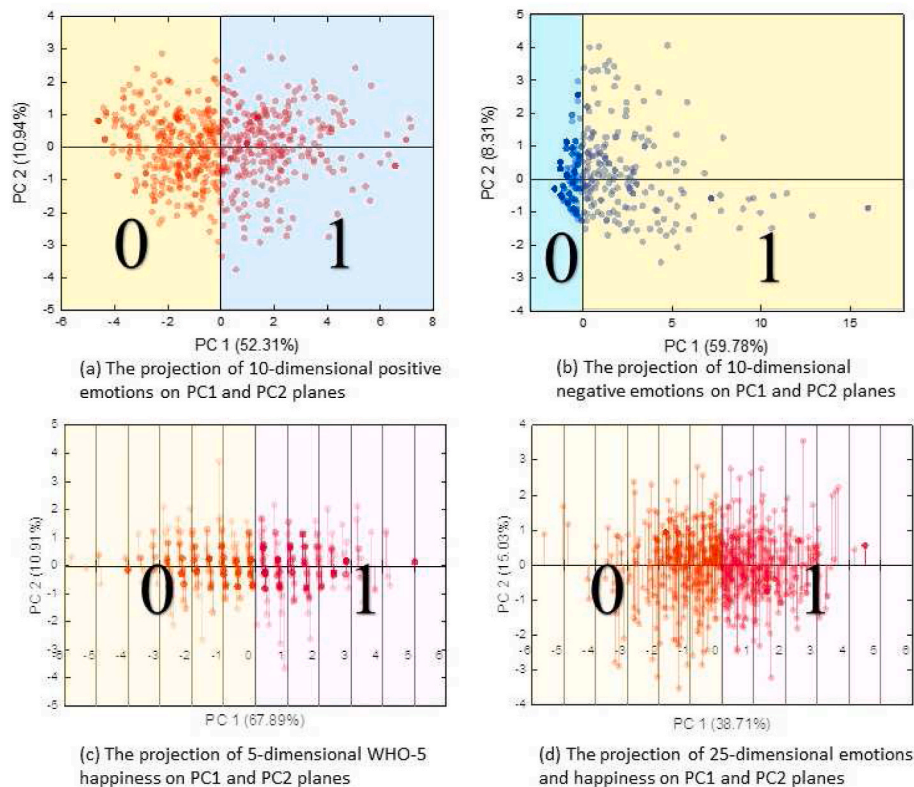


Fig. 12. Positive emotion, negative emotion, happiness index PCA binary classification.

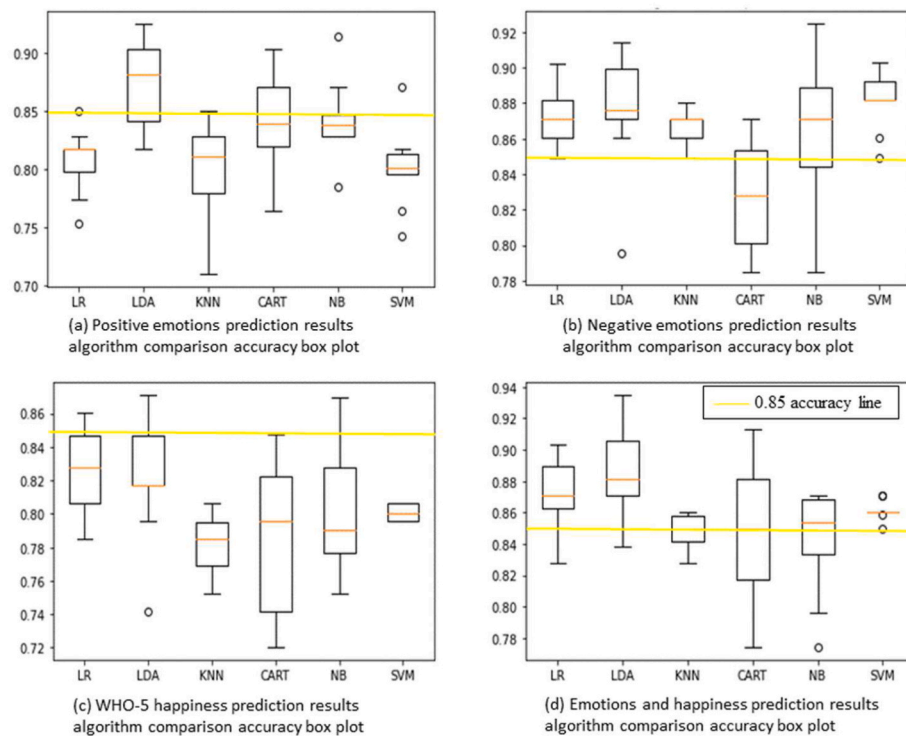


Fig. 13. Prediction result algorithm comparison accuracy box plot.

poorly when dealing with multivariate problems because the attributes are not completely independent of each other. The LR algorithm works well when predicting thermal comfort, but it lacks accuracy when predicting emotion and happiness, which proves that the sample distribution of thermal comfort is more even than that of the emotion and happiness index.

PCA has a significant effect when applied to data dimensionality reduction, which makes the classification of the data can be observed more clearly. In this study, the Zero-mean and unit variance projection is better than the Zero-mean projection, and the data after projection is more categorizable. From the Zero-mean and unit variance results, it can be seen that through PCA, the accuracy is not reduced. In the case of degree, the 61-dimensional problem can be reduced to at least 43-dimensional, and the principal component of 43-dimensional can represent more than 90% of the information.

Finally, the experimental data were only collected in a public building with natural ventilation in the transitional seasons in southern China in hot summer and cold winter due to the limited experimental conditions. The occupants feel hot and the data collection results are not evenly distributed, which limits the application effects of machine learning algorithms. Subsequent research can perform an analysis of public buildings in different seasons throughout the year. Considering the high randomness of natural ventilation patterns in public buildings, the questionnaire hopes to cover as much as possible the physiological and psychological dimensions that affect thermal comfort, which increases the difficulty of collecting the database. In terms of attribute analysis, this research mainly considers dimensionality reduction and thermal comfort evaluation analysis from the perspective of ML algorithms. It mainly regards emotion and happiness index as psychological impact and other factors as physiological impact. Follow-up research can be based on the conclusion of this article and focuses on the analysis of the mutual influence between various attributes and the degree of influence on the physical and psychological aspects of the occupants.

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.

References

- [1] International Energy Agency, "Buildings A source of enormous untapped efficiency potential," Int Energy Agency, [Online]. Available: <https://www.iea.org/topics/buildings>. [Accessed 20 June 2020].
- [2] L. Pérez-Lombard, J. Ortiz, C. Pout, A review on buildings energy consumption information, *Energy Build.* 40 (2008) 394–398, 2008/01/01/.
- [3] Z. Wang, B. Lin, Y. Zhu, Modeling and measurement study on an intermittent heating system of a residence in Cambridgeshire, *Build. Environ.* 92 (2015) 380–386, 2015/10/01/.
- [4] A. Mirakhorli, B. Dong, Occupancy behavior based model predictive control for building indoor climate—a critical review, *Energy Build.* 129 (2016) 499–513, 2016/10/01/.
- [5] W.G. Cai, Y. Wu, Y. Zhong, H. Ren, China building energy consumption: situation, challenges and corresponding measures, *Energy Pol.* 37 (2009) 2054–2059, 2009/06/01/.
- [6] S. Nagpal, J. Hanson, C. Reinhart, A framework for using calibrated campus-wide building energy models for continuous planning and greenhouse gas emissions reduction tracking, *Appl. Energy* 241 (2019) 82–97, 2019/05/01/.
- [7] P. Ferreira, A. Ruano, S. Silva, E. Conceicao, Neural networks based predictive control for thermal comfort and energy savings in public buildings, *Energy Build.* 55 (2012) 238–251.
- [8] N. Wong, H. Feriadi, P. Lim, K. Tham, C. Sekhar, K. Cheong, Thermal comfort evaluation of naturally ventilated public housing in Singapore, *Build. Environ.* 37 (2002) 1267–1277.
- [9] W. Yang, G. Zhang, Thermal comfort in naturally ventilated and air-conditioned buildings in humid subtropical climate zone in China, *Int. J. Biometeorol.* 52 (2008) 385–398.
- [10] P.O. Fanger, Thermal comfort. Analysis and applications in environmental engineering, *Thermal Comfort. Anal. Appl. Environ. Eng.* (1970).
- [11] K. Katić, R. Li, B. Kingma, W. Zeiler, Modelling hand skin temperature in relation to body composition, *J. Therm. Biol.* 69 (2017) 139–148, 2017/10/01/.
- [12] F. Auffenberg, S. Stein, A. Rogers, A Personalised Thermal Comfort Model Using a Bayesian Network, 2015.
- [13] H. Wang, L. Liu, Experimental investigation about effect of emotion state on people's thermal comfort, *Energy Build.* 211 (2020) 109789.
- [14] C.K.C. Lam, J. Hang, D. Zhang, Q. Wang, M. Ren, C. Huang, Effects of short-term physiological and psychological adaptation on summer thermal comfort of outdoor exercising people in China, *Build. Environ.* 198 (2021) 107877.

- [15] K. Liu, T. Nie, W. Liu, Y. Liu, D. Lai, A machine learning approach to predict outdoor thermal comfort using local skin temperatures, *Sustain. Cities Soc.* 59 (2020) 102216, 2020/08/01/.
- [16] Z. Wang, J. Wang, Y. He, Y. Liu, B. Lin, T. Hong, Dimension analysis of subjective thermal comfort metrics based on ASHRAE Global Thermal Comfort Database using machine learning, *J. Build. Eng.* 29 (2020) 101120, 2020/05/01/.
- [17] S. Yang, M.P. Wan, W. Chen, B.F. Ng, S. Dubey, Model predictive control with adaptive machine-learning-based model for building energy efficiency and comfort optimization, *Appl. Energy* 271 (2020) 115147, 2020/08/01/.
- [18] E. Foda, K. Sirén, Design strategy for maximizing the energy-efficiency of a localized floor-heating system using a thermal manikin with human thermoregulatory control, *Energy Build.* 51 (2012) 111–121, 2012/08/01/.
- [19] A. Ghahramani, C. Tang, B. Becerik-Gerber, An online learning approach for quantifying personalized thermal comfort via adaptive stochastic modeling, *Build. Environ.* 92 (2015) 86–96, 2015/10/01/.
- [20] L. Jiang, R. Yao, Modelling personal thermal sensations using C-Support Vector Classification (C-SVC) algorithm, *Build. Environ.* 99 (2016) 98–106, 2016/04/01/.
- [21] P. Bermejo, L. Redondo, L. de la Ossa, D. Rodríguez, J. Flores, C. Urea, et al., Design and simulation of a thermal comfort adaptive system based on fuzzy logic and on-line learning, *Energy Build.* 49 (2012) 367–379, 2012/06/01/.
- [22] W. Liu, Z. Lian, B. Zhao, A neural network evaluation model for individual thermal comfort, *Energy Build.* 39 (2007) 1115–1122, 2007/10/01/.
- [23] A.C. Megri, I. El Naqa, Prediction of the thermal comfort indices using improved support vector machine classifiers and nonlinear kernel functions, *Indoor Built Environ.* 25 (2016) 6–16.
- [24] J.L. Hensen, Literature review on thermal comfort in transient conditions, *Build. Environ.* 25 (1990) 309–316.
- [25] R.J. De Dear, G.S. Brager, Thermal comfort in naturally ventilated buildings: revisions to ASHRAE Standard 55, *Energy Build.* 34 (2002) 549–561.
- [26] J. Van Hoof, Forty years of Fanger's model of thermal comfort: comfort for all? *Indoor Air* 18 (2008) 182–201.
- [27] D.G. Kleinbaum, K. Dietz, M. Gail, M. Klein, M. Klein, *Logistic Regression*, Springer, 2002.
- [28] S. Menard, *Applied Logistic Regression Analysis*, vol. 106, Sage, 2002.
- [29] D.W. Hosmer Jr., S. Lemeshow, R.X. Sturdivant, *Applied Logistic Regression*, vol. 398, John Wiley & Sons, 2013.
- [30] J.M. Hilbe, *Logistic Regression Models*, Chapman and hall/CRC, 2009.
- [31] S. Balakrishnama, A. Ganapathiraju, *Linear Discriminant Analysis-A Brief Tutorial*, vol. 18, Institute for Signal and information Processing, 1998, pp. 1–8.
- [32] P. Xanthopoulos, P.M. Pardalos, T.B. Trafalis, *Linear discriminant analysis*, in: *Robust Data Mining*, Springer, 2013, pp. 27–33.
- [33] S. Ioffe, Probabilistic linear discriminant analysis, in: *European Conference on Computer Vision*, 2006, pp. 531–542.
- [34] O. Kramer, K-nearest neighbors, in: *Dimensionality Reduction with Unsupervised Nearest Neighbors*, Springer, 2013, pp. 13–23.
- [35] P. Horton, K. Nakai, Better Prediction of Protein Cellular Localization Sites with the it K Nearest Neighbors Classifier, *Ismb*, 1997, pp. 147–152.
- [36] J. Laaksonen, E. Oja, Classification with learning k-nearest neighbors, in: *Proceedings of International Conference on Neural Networks*, vol. 96, ICNN, 1996, pp. 1480–1483.
- [37] W.Y. Loh, *Classification and regression trees*, Wiley interdisciplinary reviews: Data Min. Knowl. Discov. 1 (2011) 14–23.
- [38] W. Lou, X. Wang, F. Chen, Y. Chen, B. Jiang, H. Zhang, Sequence based prediction of DNA-binding proteins based on hybrid feature selection using random forest and Gaussian naive Bayes, *PLoS One* 9 (2014) e86703.
- [39] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, B. Scholkopf, Support vector machines, *IEEE Intell. Syst. Their Appl.* 13 (1998) 18–28.
- [40] M. Mohri, A. Rostamizadeh, A. Talwalkar, *Foundations of Machine Learning*, MIT press, 2018.
- [41] J.F. Nicol, M.A. Humphreys, Adaptive thermal comfort and sustainable thermal standards for buildings, *Energy Build.* 34 (2002) 563–572, 2002/07/01/.
- [42] I. Knez, S. Thorsson, Thermal, emotional and perceptual evaluations of a park: cross-cultural and environmental attitude comparisons, *Build. Environ.* 43 (2008) 1483–1490.
- [43] D. Watson, L.A. Clark, A. Tellegen, Development and validation of brief measures of positive and negative affect: the PANAS scales, *J. Pers. Soc. Psychol.* 54 (1988) 1063.
- [44] C.W. Topp, S.D. Østergaard, S. Søndergaard, P. Bech, The WHO-5 Well-Being Index: a systematic review of the literature, *Psychother. Psychosom.* 84 (2015) 167–176.
- [45] 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, 2018/03/15/.
- [46] Y. Peng, T. Feng, H. Timmermans, A path analysis of outdoor comfort in urban public spaces, *Build. Environ.* 148 (2019) 459–467, 2019/01/15/.
- [47] *Code for Thermal Design of Civil Building* (GB 50176-93, China Planning Press, 1993.
- [48] R. Geiger, *Klassifikation der klimate nach W. Köppen, Landolt-Börnstein-Zahlenwerte und Funktionen aus Physik, Chemie, Astronomie, Geophysik und Technik* 3 (1954) 603–607.
- [49] E.A. McCullough, B.W. Jones, J. Huck, A comprehensive data base for estimating clothing insulation, *Ashrae Trans* 91 (1985) 29–47.
- [50] E. ISO, 7726, *Ergonomics of the Thermal Environment-Instruments for Measuring Physical Quantities* (ISO 7726: 1998), 1998.
- [51] *Design Code for Heating Ventilation and Air Conditioning of Civil Buildings*, 2012, p. GB50736, 2012.
- [52] D.W. Hosmer, S. Lemeshow, R.X. Sturdivant, *Applied Logistic Regression*, Wiley, New York, 2000.
- [53] S. Sperandei, Understanding logistic regression analysis, *Biochem. Med.* 24 (2014) 12–18.
- [54] S.A. Dudani, The distance-weighted k-nearest-neighbor rule, *IEEE Transac. Syst. Man Cybernet.* (1976) 325–327.
- [55] R. Timofeev, *Classification and Regression Trees (CART) Theory and Applications*, Humboldt University, Berlin, 2004, pp. 1–40.
- [56] M. Ontivero-Ortega, A. Lage-Castellanos, G. Valente, R. Goebel, M. Valdes-Sosa, Fast Gaussian Naïve Bayes for searchlight classification analysis, *Neuroimage* 163 (2017) 471–479.
- [57] W.S. Noble, What is a support vector machine? *Nat. Biotechnol.* 24 (2006) 1565–1567.
- [58] S.B. Kotsiantis, I. Zaharakis, P. Pintelas, Supervised machine learning: a review of classification techniques, *Emerg. Artificial Intell. Appl. Comput. Eng.* 160 (2007) 3–24.