Analysis of outlier detection rules based on the ASHRAE global thermal comfort database

Analysis of outlier detection rules based on the ASHRAE Global Thermal Comfort Database

Shaoxing Zhang [a,b]; Runming Yao [a,b,*]; Chenqiu Du [a], Emmanuel Essah [b], Baizhan Li[a]

[a] Joint International Research Laboratory of Green Buildings and Built Environments (Ministry of Education), Chongqing University, Chongqing 400045, China

[b] School of the Built Environment, University of Reading, UK

Corresponding author* r.yao@cqu.educ.cn; r.yao@reading.ac.uk

Abstract:

ASHRAE Global Thermal Comfort Database has been extensively used for analyzing specific thermal comfort parameters or models, evaluating subjective metrics, and integrating with machine learning algorithms. Outlier detection is regarded as an essential step in data preprocessing, but current publications related to this database paid less attention to the influence of outliers in raw datasets. This study aims to investigate the filter performance of different outlier detection methods. Three stochastic-based approaches have been performed and analyzed based on the example of predicting thermal preference using the Support Vector Machine (SVM) algorithm as a case study to compare the predictions before and after outlier removal. Results show that all three rules can filter some obvious outliers, and the Boxplot rule produces the most moderate filer results, whereas the 3-Sigma rule sometimes fails to detect outliers and the Hampel rule may provide an aggressive solution that causes a false alarm. It has also been discovered that a small reduction in establishing machine learning models can result in less complicated and smoother decision boundaries, which has the potential to provide more energy-efficient and conflict-free solutions.

Keywords: Outlier detection, Thermal preference, ASHRAE global thermal comfort database, Machine learning, Support vector machine

Abbreviations

Af	Tropical rainforest climate in Köppen climate classification
Am	Tropical monsoon climate in Köppen climate classification
Aw	Tropical savanna climate with dry-winter characteristics in Köppen climate classification
BSh	Hot semi-arid climate in Köppen climate classification
BWh	Hot desert climate in Köppen climate classification
Cfa	Humid subtropical climate in Köppen climate classification
Cfb	Oceanic climate in Köppen climate classification
Csb	Warm-summer Mediterranean climate in Köppen climate classification
Cwa	Monsoon-influenced humid subtropical climate in Köppen climate classification
Cwb	Subtropical highland climate or Monsoon-influenced temperate oceanic climate in Köppen climate classification
IQR	Interquartile Range
MAD	Median Absolute Deviation
PMV	Predicted Mean Vote
PPD	Predicted Percentage of Dissatisfied
RBF	Radial Basis Function
SVM	Support Vector Machine

1. Introduction

Providing thermally acceptable indoor environments in buildings can positively promote occupants' satisfaction [1], health [2][3], productivity [4][5], and well-being [6]. Traditional heat-balance based PMV-PPD index [7] tries to transparently explain the interactions between physical environments and human bodies, but it usually provides uniform solutions for different scenarios with little space for model updates and sometimes present poor predicative performance in real practice [8]. On the other hand, the adaptive thermal comfort models focus more on the important adaptive response related to occupants' thermal expectations, physiological acclimation, and behavioral patterns in real buildings. The adaptive approach is more sophisticated and responsive to environmental control algorithms, increasing the opportunities for personalized control and occupant acceptability, reducing energy consumption, and encouraging climatically responsive and environmentally responsible building design

[9].

Data-driven methods, such as support vector machine (SVM) [10][11], random forest [12], decision tree [13], Bayesian approach [14] [15], neural network [16], have been extensively used in developing adaptive thermal comfort models based on the datasets. They are data-sensitive and capable of providing highly customized solutions for specific groups or individuals. During the training process of machine learning algorithms, outliers will skew the results of statistical analyses performed on the dataset, resulting in less effective and useful models [17]. Therefore, the detection of outliers can assist machine learning algorithms in making more rational predictions in buildings.

1.1 Global Thermal Comfort Database II

The ASHRAE Global Thermal Comfort Database II (short name: Comfort Database) is an online and open-source database that includes approximately 81,846 complete data points collected and harmonized from the raw data of 52 field studies from 160 buildings worldwide [18], in addition to the 22,000 records published in Database I under RP-884 project 20 years ago to test the hypothesis of adaptive thermal comfort theory [19]. This dataset was organized with a standard spreadsheet format that contains basic identifiers, instrumental measurements, subjective evaluations, calculated indices, and environmental control. It provides opportunities for scholars to conduct additional analyses benefiting from its large sample size and standardised data format. Since the release of Comfort Database, many research efforts have already been focused on 1) analysis of specific thermal comfort parameters or models, such as testing differences in air and radiant temperatures [20] [21], clothing adjustments in naturally ventilated buildings [22] or classrooms [23], validating [24] or enhancing [25] PMV predictive accuracy, comparing performance of PMV and modified PMV models [26], PMV predictions in mixed-mode buildings [27], building modified SET models [28]; **2) subjective evaluations**, such as potential of extending acceptable temperatures [29], influence of demographic and contextual factors on thermal sensation [30], thermal sensitivity of occupants from different building types or geographic locations [31] [32], identifying key parameters that influence thermal preference [33]; 3) integration with

machine learning algorithms, such as anomaly detection in SVM [34], extracting knowledge for transfer learning[35], comparing predictive accuracies of different machine learning algorithms [36], building predictive models based on SVM [37] or Bayesian inference approach [38] [39].

For a better understanding of Comfort Database related studies, Appendix A summarized their environmental inputs, subjective metrics, contextual factors, data sum, algorithm, and outlier preprocessing methods. The following points are noteworthy.

- **Algorithms:** in most studies, regression-based methods were used concerning the classical adaptive thermal comfort theories. They typically provide linear models for indoor environmental design, but thermal comfort is a complex nonlinear interaction process between the human body and physical environments, and these linear models may fall short of providing an understanding of this process. Thanks to the large data size and better diversity of Comfort Database, many scholars have conducted research using machine learning methods, which often demonstrate higher prediction accuracy than traditional models.

- **Subjective metrics:** thermal sensation vote is the most popular indicator due to its popularity in standards and PMV theory, but other metrics, such as thermal preference or thermal acceptability, have received less attention and can describe occupants' thermal states from different perspectives. When conducting thermal comfort studies in laboratories or real buildings, four subjective thermal comfort metrics are commonly used include thermal sensation, thermal acceptability, thermal satisfaction, and thermal preference: 1) thermal sensation is considered to be the most objective as it associates with physical measurement and PMV index; 2) thermal acceptability turns to be more subjective as people can accept the environments even when they feel uncomfortable; 3) thermal satisfaction usually measures overall assessment during Post Occupancy Evaluation (POE) process; 4) thermal preference direct indicates the preferred adjustment to thermal environments [40]. Among these four metrics, thermal preference is perhaps the most important and direct metric in ambient control because it can indicate to the

HVAC system what type of control action should be taken [12], but the metric of thermal preference has received insufficient attention in Comfort Database related research.

- **Data sum:** In the released ASHRAE dataset [18], the original csv file contains comprehensive information on environmental parameters as well as subjective evaluations of investigated locations in 70 columns. However, some monitoring parameters, such as air velocity at different heights, were rarely collected in practice, and some columns repeated the same information, such as Fahrenheit and Celsius degrees for the same value. Therefore, 42 columns with sufficient data collection and non-overlapping information have been depicted in Fig. 1 to provide a general overview of the ASHRAE dataset while reducing redundancy. The black colour indicates that there is a data point at that specific location, while the white colour indicates that data is missing. Overall, the first few columns of data are fairly complete because the majority of them are basic information, such as publication date, research location, climate, etc. However, many columns to the left have a wide range of blanks. This is due to the different research aims or objectives, contributors to the database gathered various contextual factors to fulfill the adaptive thermal comfort theory. However, combining them into a single format resulted in many irregular missing data points. When other scholars conduct secondary analyses from specific perspectives, the sample size will differ significantly.

- **Outlier processing:** it is reported that the results of regression analysis can be seriously affected by just one or two erroneous data points [41], and an outlier-free dataset also benefits machine learning algorithms by allowing them to train more accurate models [42]. Although outlier detection is a broad topic with numerous technologies and real-world applications, such as fraud detection/diagnosis, loan application, unauthorized access in a computer network, activity monitoring, image/text analysis, motion segmentation, and medical condition monitoring [43], most studies on the Comfort Database did not pay much attention or provide detailed descriptions of outlier processing. To avoid outliers, research [25] and [44]

set a minimum sample size for each analyzed group, but extreme values in normal groups were not addressed. The method of inspection was used in research [45] to filter out unexpected values, but it would require too much labor when the sample size is large. The Boxplot rule was used in research [38] and [26] to select outliers, but a more systematic view of different outlier removal approaches in the thermal comfort domain has yet to be discovered.

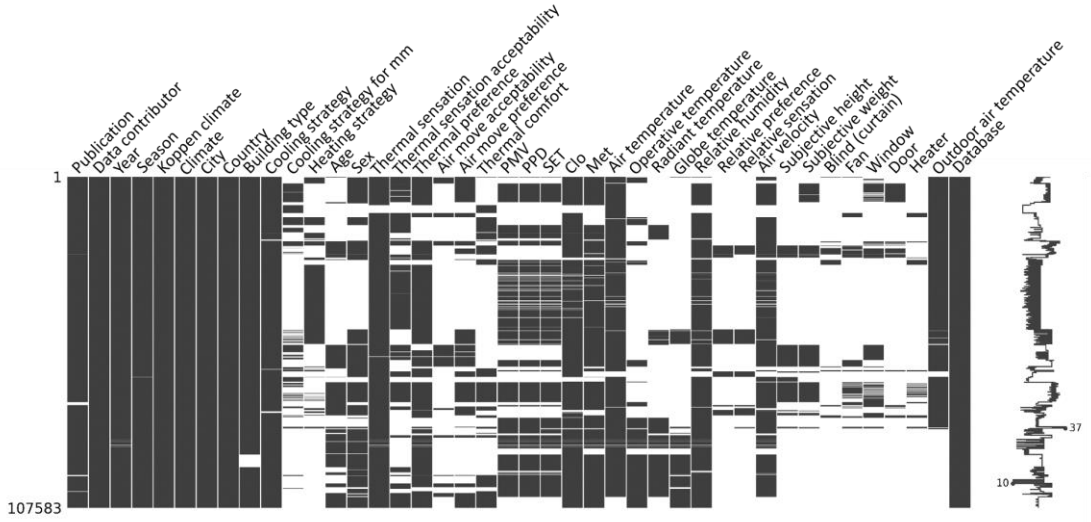


Fig. 1 Visualisation of missing data in Comfort Database (The sparkline at right summarizes the general completeness of the data. The horizontal position of a specific point in this sparkline indicates the number of data points in this row, with the left being less and the right being more. In this case, the minimum and maximum sums are 10 and 37, respectively.)

1.2 Brief review of outlier detection techniques

Outlier detection is defined as the task of identifying patterns in data that differ in some respect from expected behavior. These unexpected patterns are also known as anomalies, discordant observations, exceptions, aberrations, surprises, peculiarities, or contaminants depending on the application domains [46]. The common outlier detection approaches can be categorized as [47]:

- **Stochastic-based:** it calculates the generative probability density function of data. A new observation will be marked as an outlier if its probability density is low in

comparison to the statistical distribution fitted to previous data [48]. This approach is mathematically well-grounded as a "*transparent*" method, but its performance is limited when the sample size is very small.

- **Distance-based:** it assumes that the normal data points have close neighbours, whereas outliers are located far away from those points [49]. Unlike the Stochastic-based approach, it does not require prior knowledge of data distribution, but it suffers from selecting the appropriate distance metrics or cluster width to establish the similarity between data points.

- **Reconstruction-based:** it trains the underlying data using neural networks or principal components analysis (PCA). When new test data is added, the reconstruction error (the distance between it and its representation) will be related to the outlier score [50]. This approach allows for model training flexibility, but its performance is highly dependent on model parameters and may suffer from searching for the best training method.

- **Domain-based:** it will create a boundary based on the structure of the training data, and outliers will be determined according to their proximity to this boundary [51]. This approach is often achieved using support vector machine (SVM) algorithms, and it faces the challenge of selecting suitable kernel functions and tuning hyperparameters for the desired boundary region.

- **Information-theoretic based:** it assumes that outliers are supposed to change the information content of the entire dataset (based on Shannon's information entropy or entropy-related indices), and any subsets with the greatest difference will contain outliers [52]. The drawback of this approach is that it is only sensitive when there are a large number of outliers in the dataset.

These five outlier detection approaches have numerous real-world applications in different domains, such as IT security [53], healthcare [54], industrial monitoring [55], image processing [56], text mining [57], and sensor networks [58]. However, a common understanding of outliers has not been reached, and all of these methods have been used

in various domains based on specific considerations in practice and theory. Therefore, it is difficult to recommend which outlier detection method is always the best due to the availability/dimension/continuity/format of data, the specific application domain, and the wide variety of real-world datasets. Specifically, stochastic-based rules are frequently used in the analysis that can be mathematically described, including probability density function (pdf) [59] or Hidden Markov Model (HMM) [60], while distance-based rules measure data similarity in areas such as climate data [61], network intrusion [62], and protein sequences [63]. Reconstruction-based rules model the underlying data by developing complicated neural network structures, such as LSTM (long short-term memory) [64], or by projecting data into lower dimensional spaces, such as PCA (Principal Components Analysis) [65]. Domain-based rules generate decision boundaries (usually based on SVM) and are applied in various fields including audio recordings [66], text data [67], functional magnetic resonance imaging [68], and identifying patient deterioration in vital signs [69]. Information-theoretic based rules regard entropy as the fundamental concept, such as developing conditional entropy or relative conditional entropy [52], combining mutual information [70], and incorporating it into the Bayesian network frame [71].

In thermal comfort studies, outlier detection is usually involved in data preprocessing to remove the misleading effects of extreme values. The available techniques include stochastic-based methods like the 3-Sigma rule [72], the Boxplot rule [73], and the Hampel rule [74]; distance-based methods such as cook distance [75] and k-nearest neighbour (KNN) [76]; manually inspection [45] or setting fixed ranges [77]; and binning or adjusting variables into specific intervals [78]. The majority of studies in thermal comfort research filtered outliers using stochastic-based methods, and more details can be found in Appendix B. Although outlier detection approaches have been extensive used in many domains, they typically appear only during data preprocessing rather than explaining the impact of these outliers on model establishment in the thermal comfort community. Instead of enumerating and calculating the performance of all outlier detection methods, the aim of this study is to investigate how stochastic-based

outlier removal affects data distribution in the database and examine the filter performance of stochastic-based outlier detection approaches on Comfort Database.

2. Methodology

Despite the fact that there are five major types of outlier detection techniques, stochastic-based approaches are used in more than half of the thermal comfort research publications (Appendix B), with a focus on three rules: the 3-Sigma rule, the Boxplot rule, and the Hampel rule. However, the foundation and performance of these three filtering methods have received less attention in previous published articles. Therefore, this section summarizes the theoretical foundations of these three rules. To investigate how these detection rules perform on the machine learning algorithms, the SVM (Support Vector Machine) was taken to analyze predictive performance and decision boundaries both before and after outlier removal as a case study to demonstrate the shared characteristics. Fig. 2 depicts a schematic overview of the methodology.

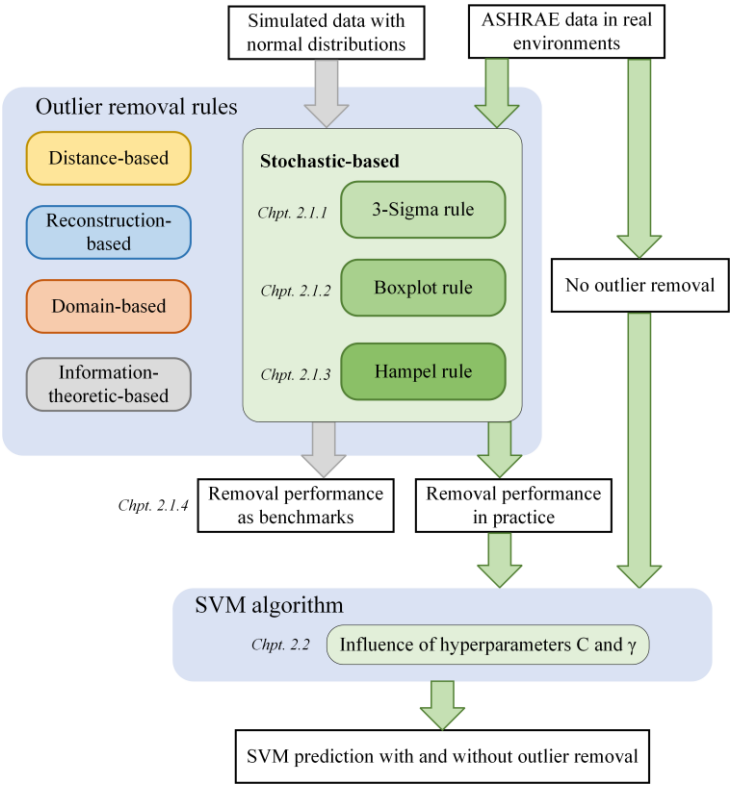


Fig. 2 The schematic overview of the methodology

Within SVM classifiers, "*thermal preference*" was chosen as the predicted label because it can indicate to the built environments what type of control action should be taken [12] but has not been widely discussed in previous publications. The stochastic-based outlier detection approach typically employs the three steps to detect an outlier listed below [48]:

1) Compute a reference value x_0 and a measured variation ζ from the data sequence $\{x_k\}$;

2) Choose a threshold parameter t;

3) Test every data in sequence $\{x_k\}$ to determine whether it is an outlier according to the rule described:

$$|x_k - x_0| > t\,\zeta \tag{1}$$

Eq. (1) intuitively states that if the new data $x_k$lies too far from the reference value x_0, it is recognized as an outlier. The aggressiveness of the detection procedure will be adjusted by the threshold t. If t equals 0, all data different from $x_0$will be outliers; if t is too large, the detection rule will find no outliers.

2.1 Stochastic-based outlier detection rules

This section presents some fundamentals of three stochastic-based models which have been widely used in thermal comfort studies (details in Appendix B): the 3-Sigma rule, Boxplot rule, and Hampel rule.

2.1.1 3-Sigma rule

The basic idea of the 3-Sigma rule, also known as the extreme studentized deviation (ESD) identifier [79], states that if a data sequence is well approximated to Gaussian random variables, the probability of seeing a value x_i more than three standard deviations away from the mean is only about 0.3%:

$$|x_i - \hat{\mu}| > 3\hat{\sigma} \tag{2}$$

Where $\hat{\mu}$ is the sample mean, and $\hat{\sigma}$ is the sample standard deviation.

However, outliers themselves in the dataset can cause significant errors in estimating the mean and standard deviation values, which are the foundations of this outlier

detection procedure [80]. As a result, the magnitude of the differences between the observed value x_i and mean value $\hat{\mu}$ may be too small, and the scale estimate could be too large, making detecting outliers more difficult.

2.1.2 Boxplot rule

The Boxplot rule was introduced by John Tukey for exploratory data analysis in the 1970s [81]. It detects outliers by utilizing quartile information with box and whiskers plots, which include the lower quartile (Q1, 25th percentile), median (Q2, 50th percentile), upper quartile (Q3, 75th percentile), and interquartile range (IQR=Q3-Q1):

$$x_i > Q_3 + 1.5IQR \ \cup \ x_i < Q_1 - 1.5IQR \tag{3}$$

Boxplot rule can be applied to data with asymmetric distributions in addition to Gaussian distributions due to the specific computation of the 25th and 75th percentiles. It also substitutes the median value for the mean value, removing the potentially misleading effects of the extremely large outliers and making it more outlier-resistant.

2.1.3 Hampel rule

To honour Hampel's contribution in describing the useful characteristics of the median absolute deviation (MAD) scale estimator [82], Davies and Gather [83] proposed the Hampel rule based on the MAD scale estimator:

$$|x_i - \tilde{x}| > \alpha S \tag{4}$$

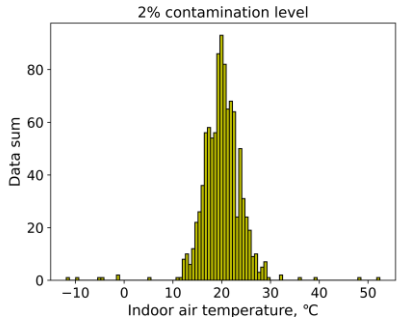
$$S = \frac{1}{0.6745} median\{|x_i - \tilde{x}|\} \tag{5}$$

Where $\tilde{x}$ is the median value, α is the threshold parameter (suggested as 3 in [48]), and S is the MAD scale estimator.

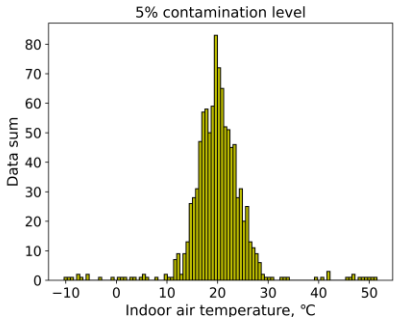
Similar to the symmetry of the 3-sigma rule, the Hampel rule is unable to present the asymmetric property of data distributions. The MAD scale estimator used in the Hampel rule also has lower outlier sensitivities than the mean and standard deviation values.

2.1.4 Performance simulation of three outlier detection rules

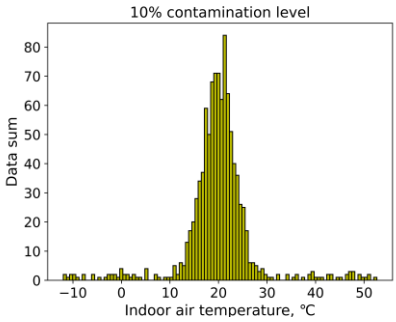
This section intends to compare the performance of outlier detection rules in sections 2.1.1 to 2.1.3. There were 1000 sampling data sets, which followed the normal distribution with a mean value of 20 and a standard deviation value of 3.33. It was generated using the Python package *numpy.random* with a random seed of 42. This seed parameter ensures that the *same* set of random numbers appears each time when the *same* seed is reset [84]. Outliers with different contamination levels were generated beyond the 20-30 boundary, as shown in Fig. 3. The introduction of simulated data here is intended to reflect the filtering performance of the three stochastic-based outlier detection rules when subjected to the same normally distributed data and noise. It serves as a benchmark and a compared reference for the analyses using the Comfort Database in this paper because real-world data typically deviates from a standardised normal distribution.



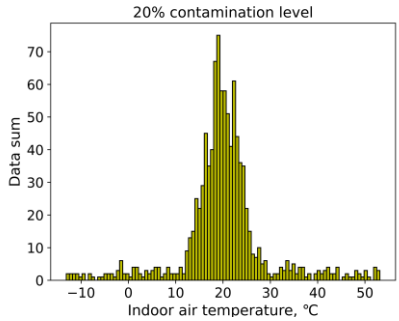
(a) 2% contamination level



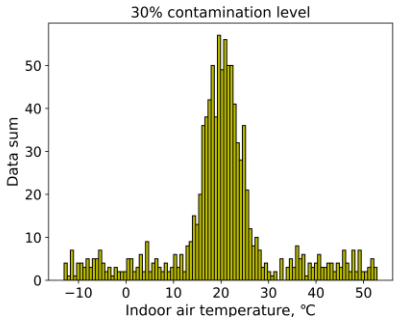
(b) 5% contamination level



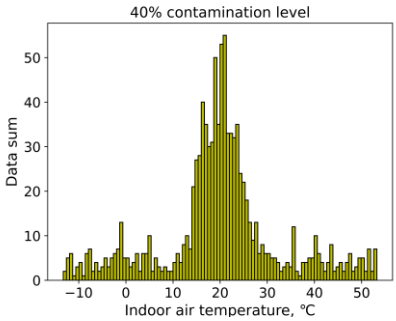
(c) 10% contamination level



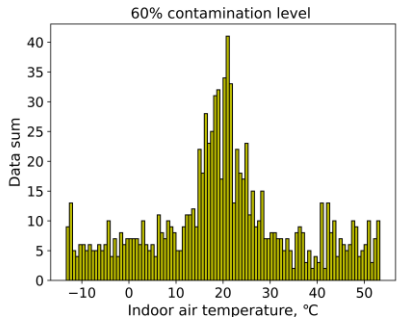
(d) 20% contamination level



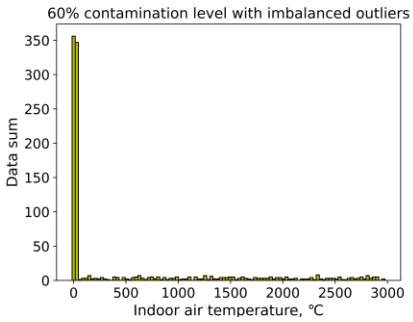
(e) 30% contamination level



(f) 40% contamination level



(g) 60% contamination level



(h) 60% contamination level with very imbalanced outliers

Fig. 3. Generated data points under different contamination levels

Table 1. Predictive boundaries and outlier sums of three outlier detection rules under different contamination levels

	3-sigma upper	**3-sigma lower**	**Boxplot upper**	**Boxplot lower**	**Hampel upper**	**Hampel lower**
2% boundary	32.8	7.0	29.1	10.7	30.09	9.67
2% sum (10 outliers)	4	7	11 (one false positive)	11 (one false positive)	7	8
5% boundary	37.0	2.6	29.6	9.71	30.84	8.74
5% sum (25 outliers)	15	13	21	22	19	21
10% boundary	41.8	-1.8	30.1	10.1	31.1	9.0
10% sum (50 outliers)	21	18	40	43	38	41
20% boundary	47.3	-7.0	31.7	8.1	32.9	6.6
20% sum (100 outliers)	18	14	78	66	72	62
30% boundary	55.5	-15.3	34.5	5.6	36.3	3.9
30% sum (150 outliers)	0	0	112	106	97	91
40% boundary	58.4	-18.4	37.5	2.8	39.3	0.5
40% sum (200 outliers)	0	0	104	108	93	95
60% boundary	67.5	-27.5	53.3	-13.3	57.2	-17.3
60% sum (300 outliers)	0	0	0	0	0	0
imbalanced 60% boundary	2937.1	-2018.6	1419.9	-833.2	75.0	-35.7
imbalanced 60% sum (300 outliers)	2	0	159	0	296	0

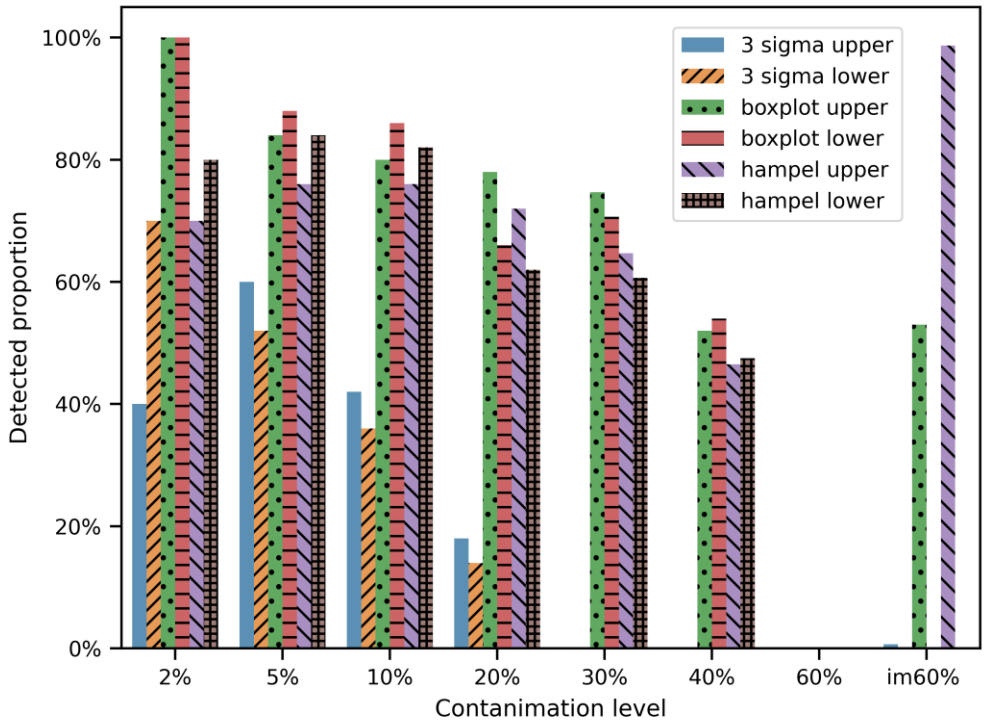


Fig 4. Predictive accuracies of outlier detection rule

Table 1 and Fig. 4 present the results of outlier removal. When the contamination level increases, two median-value based rules (Boxplot and Hampel rules) outperforms the mean-value based rule (3-sigma rule) more. The 3-sigma rule can rarely detect any outliers when the contamination level is over 30%. Pearson [48] also demonstrated mathematically that the 3-sigma rule can fail to detect outliers when contamination levels exceed 10%. For two median-value based rules, the Boxplot rule generally outperforms than Hampel rule. However, in this extremely imbalanced case Fig. 3h, the Hampel rule performs better because it is based on the 2/4 quantile value, whereas the boxplot is based on the 1/4 and 3/4 quantile values. The Hampel rule may be more appropriate when the 2/4 quantile can provide more solid information to represent the data distribution.

2.2 Support vector machine (SVM) algorithm

The support vector machine (SVM) is a supervised algorithm based on the Vapnik-Chervonenkis theory, which attempts to statistically explain the learning process [85].

When compared to other machine learning algorithms, the SVM has the advantage of avoiding over-fitting and local minima, performing well with limited training data, and solving problems with non-linear and high-dimensional patterns. It has been widely used in face detection [86], text categorization [87], image classification [88], bioinformatics [89], protein fold and remote homology detection [90], handwriting recognition [91], generalized predictive control (GPC) [92], and also in the thermal comfort research domain. Megri et al. [93] demonstrated the feasibility of applying SVM to small groups of people, such as the sick, disabled, or elderly. Chaudhuri et al. [94] used skin temperature to assess thermal comfort or discomfort based on SVM and extreme learning machine (ELM) classifiers. The predictive accuracy of their SVM classifier is around 87%, which is 7% higher than the ELM classifier. Dai [95] also investigated the predictive performance of skin temperature in steady-state conditions based on SVM and achieved 90% predictive accuracy. Aryal and Becerik-gerber [96] employed five machine learning algorithms to analyze individual thermal sensations using wrist-worn sensors, thermal cameras, and environmental parameters. Their findings suggested that SVM with a quadratic kernel outperformed other algorithms. However, Zhou et al [37] discovered that using an SVM model to analyze RP-884 data has the benefits of self-learning and self-correction, but it can be unreliable in extreme conditions. Therefore, this paper will regard the SVM algorithm as one case study to discuss how extreme conditions, namely outliers or anomalies, affects the performance of machine learning algorithms.

The SVM algorithm was proposed by Cortes and Vapnik in 1995 [97]. The fundamental goal of the SVM classifier is to create the best hyperplanes possible to distinguish the data vectors with different features. The optimization problem of the *soft-margin* SVM is:

$$minimize\ \frac{1}{2}||w||^2 + C\sum_i \xi_i \tag{6}$$

$$\text{subject to: } \mathrm{y_i}(w^T x_i + b) \geq 1 - \xi_i \tag{7}$$

Where w is the *weight vector*, ξ_i are known as *slack variables* that allow an example to be in the margin ($0 \leq \xi_i \leq 1$) or misclassified ($\xi_i \geq 1$), b is the bias, and parameter C is a regularization factor that determines the relative importance of maximizing margins and minimizing the amount of slack.

After using Lagrange multipliers, the *dual* formulation form of SVM minimization can be expressed in terms of variables α_i:

$$min\frac{1}{2}\sum_{ij} y_i\, \alpha_i y_j \alpha_j K(x_i x_j) - \sum_i \alpha_i \tag{8}$$

Where $K(x_i x_j)$ is a kernel function that converts the original non-linear observations into a higher-dimensional space where they can be separated.

The Gaussian Radial Basis Function (RBF) kernel is expressed as:

$$K(x_i x_j) = \exp\left(-\gamma||x_i - x_j||^2\right) \tag{9}$$

Where parameter γ is one RBF SVM hyperparameter that controls the flexibility of the resulting classifier by defining how far the influence of a training example can reach. To solve a specific RBF SVM classification problem, a pair of parameters C and γ is typically chosen using optimization procedures, such as grid search, random search, simulated annealing, and Bayesian optimization [98]. The SVM algorithm also includes several other kernel functions like linear, polynomial, and sigmoid. Because human interaction with thermal environments is highly non-linear [99], this paper focuses on the RBF kernel due to its greater flexibility in generating decision boundaries. Several studies [100] [101] [102] have also demonstrated that the RBF kernel is more effective at predicting thermal comfort states compared with other kernel functions.

To avoid the imbalance scale effects of data in this study, all SVM model inputs have been standardised (μ=0, σ=1) using the function *StandardScaler* in the Python package *sklearn*. The hyperparameters C and γ in SVM have been tested by the grid search method in the RBF kernel function for the SVM classifier. The best cross-

validation accuracy was obtained with exponentially growing sequences ranging from 10^{-4} to 10^{4}. The data set was randomly divided into 70% and 30% proportions for training samples and testing predictive accuracy.

2.3 Data processing

This research processes the data in Comfort Database with the following procedures:

1) Removing any data that does not contain records of thermal sensation vote, thermal preference, air temperature, relative humidity, air velocity, clothing level, and metabolic rate at the same time, with a focus on the HVAC and NV operation strategy.

2) Based on the available sample size and outdoor temperature distribution, three representative climate zones were chosen for analysis as shown in Fig. 5 with red marks on the x-axis: hot semi-arid climate (BSh, 3844 records), humid subtropical climate (Cfa, 3074 records), and temperate oceanic climate (Cfb, 3176 records), with a total sum of 10,094.

3) Applying 3-Sigma, Boxplot, and Hampel rules to filter outliers in three instrumental measurements: air temperature, relative humidity, and air velocity.

4) Employing the SVM algorithm with RBF kernel functions to classify the thermal preferences of "*prefer warmer*" and "*prefer cooler*" with training and cross-validation process carried out using the functions *SVC* and *GridSearchCV* in Python package *sklearn*.

5) Comparing the predictive performance of PMV and SVM models for HVAC and NV buildings before and after using outlier removal rules, the PMV values were computed through the function *pmv_ppd_optimized* in Python package *pythermalcomfort* developed by Tartarini and Schiavon [103].

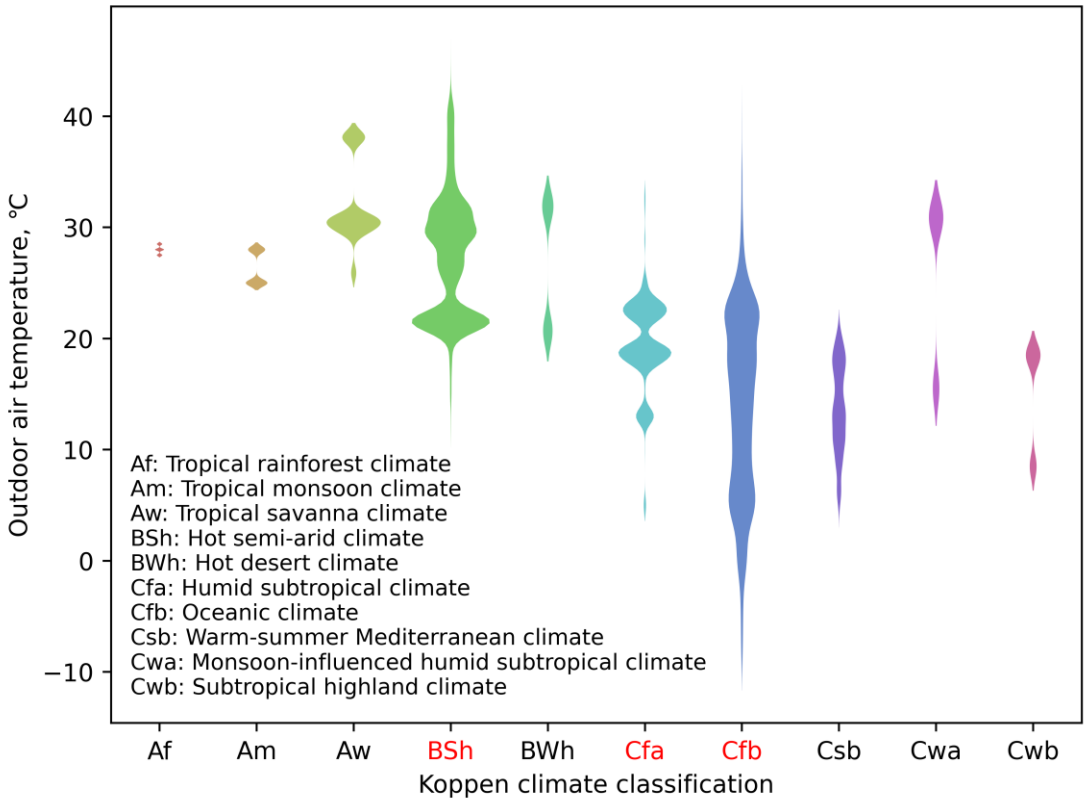


Fig. 5. Outdoor air temperature distribution after removing unqualified data in the Comfort Database

3. Results

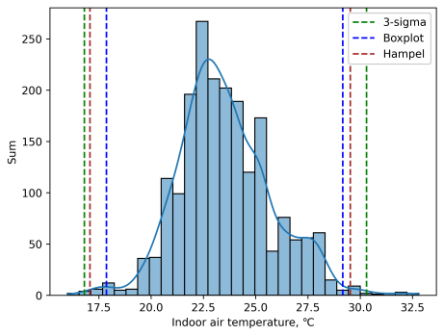
3.1 Outlier detection of three removal rules

3.1.1 Different climate zones

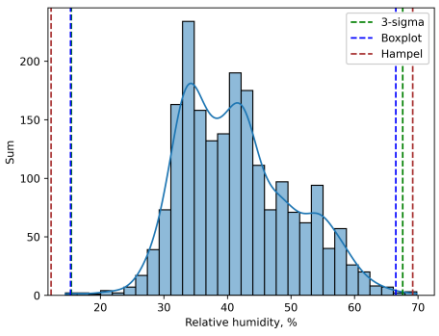
To better visualise the effects of three different outlier removal rules, this paper employs the histogram figure to summarise the distribution of particular parameters such as air temperature, followed by dotted lines on both sides of the histogram to indicate the boundaries provided by different rules. As shown in Fig. 6, the dashed lines on the left and right represent the lower and upper limits generated by the outlier removal rules, respectively. Specifically, the green colour represents the 3-Sigma rule, the blue colour represents the Boxplot rule, and the red colour represents the Hampel rule.

The outlier removal results of the HVAC buildings in different climate zones in the Comfort Database are shown in Fig. 6, where we can see the indoor air temperature

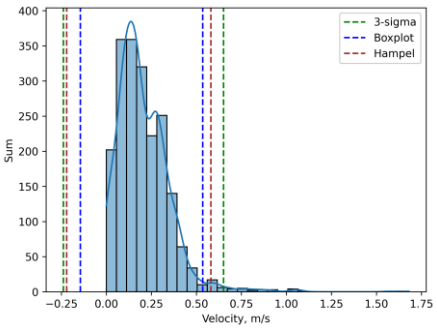
shows a relatively standard normal distribution. The 3-Sigma rule (green dotted lines) provides the greatest degree of tolerance for extreme values. This is consistent with the analysis in section 2.1.4, which shows that the 3-Sigma rule is more likely to fail as contamination levels rise. Two median-value based rules (Boxplot and Hampel) both filter more outliers compared with the 3-Sigma rule. In general, the normal range of the Boxplot rule (blue dotted lines) is tighter than the Hampel rule (brown dotted lines). The relative humidity has double-peak distributions rather than normal distributions in each climate zone. Unlike previous filter cases, all three rules show a clear sum reduction in outlier detection, particularly in the latter two climates (humid subtropical and temperate oceanic) with a more obvious double-peak feature. For air velocity, the 3-Sigma rule draws the largest normal ranges in three climates similar to the results for air temperature. The upper limit of the 3-Sigma rule in temperate oceanic climates is nearly three times that of two median-value-based rules, indicating that the 3-Sigma rule is very susceptible to being affected by extreme outliers and broadening its normal range.



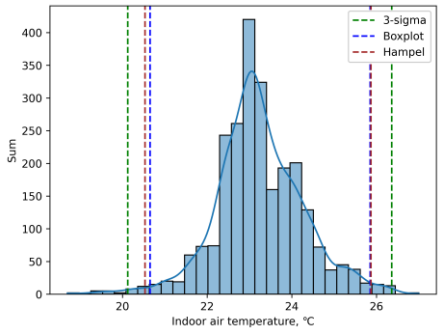
(a) Indoor air temperature in hot semi-arid



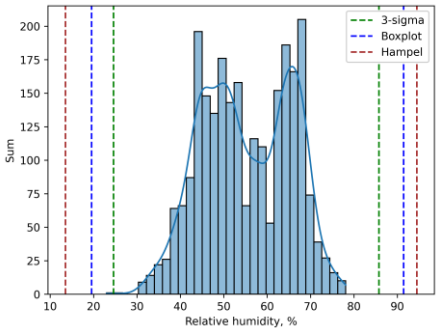
(b) Relative humidity in hot semi-arid



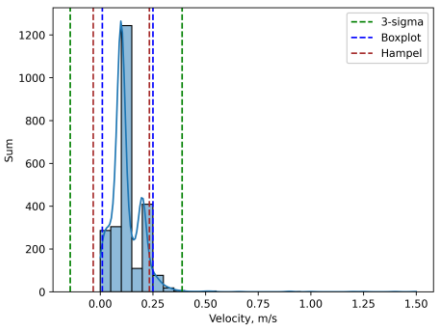
(c) Air velocity in hot semi-arid



(d) Indoor air temperature in a

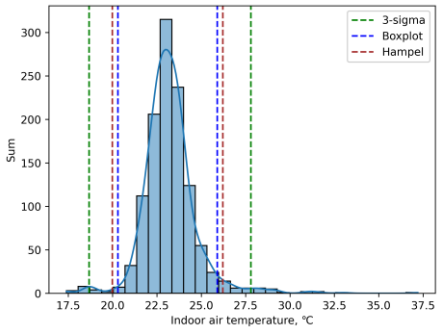


(e) Relative humidity in a humid



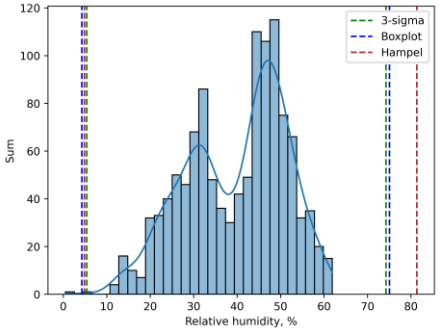
(f) Air velocity in a humid

humid subtropical



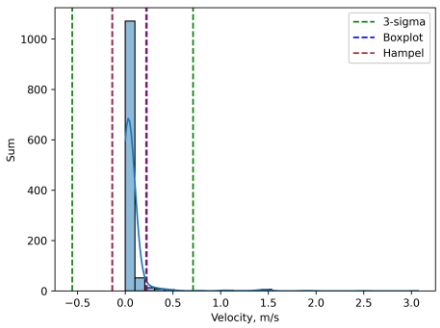
(g) Indoor air temperature in temperate oceanic

subtropical



(h) Relative humidity in temperate oceanic

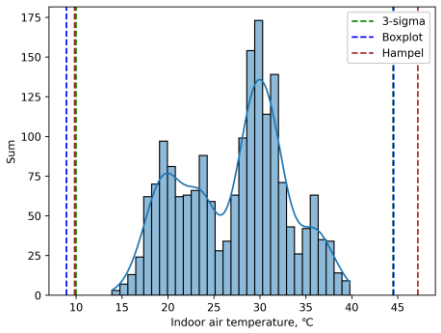
subtropical



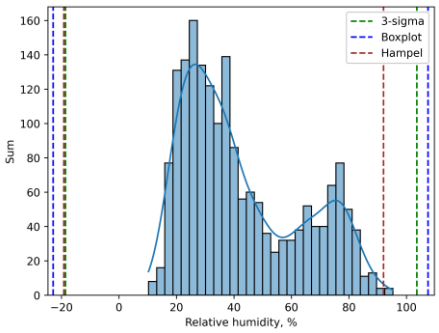
(i) Air velocity in temperate oceanic

Fig. 6. Outlier removal in HVAC buildings in different climate zones

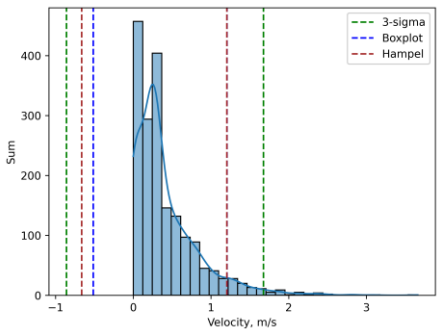
The outlier removal of the NV buildings in different climate zones is shown in Fig. 7. We can see that the indoor temperature in the first two climates (hot-semi-arid and humid subtropical) all have double-peak distributions, while all normal ranges of three rules fail to detect any outliers. For relative humidity in a hot semi-arid climate, the 3-Sigma rule and the Boxplot rule even create upper boundaries that exceed 100% while the Hampel rule computes a more reasonable upper limit around 90% (Fig. 7b). The 3-Sigma rule sets the upper boundary of air velocity in temperate oceanic climate around 3m/s, whereas the maximum allowable air velocity for compensating hot feeling in ASHRAE 55-2020 [104] is only 1.6 m/s.



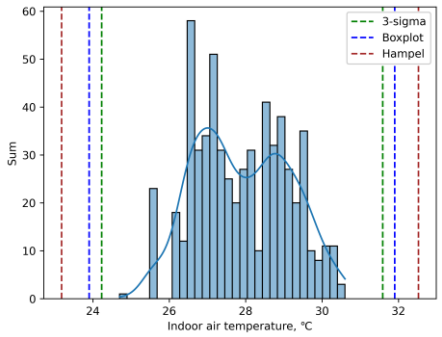
(a) Indoor air temperature in hot semi-arid



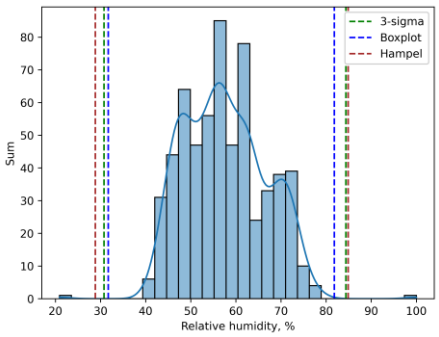
(b) Relative humidity in hot semi-arid



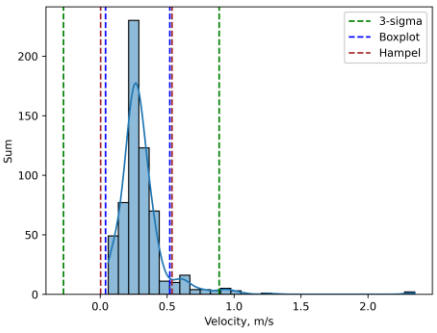
(c) Air velocity in hot semi-arid



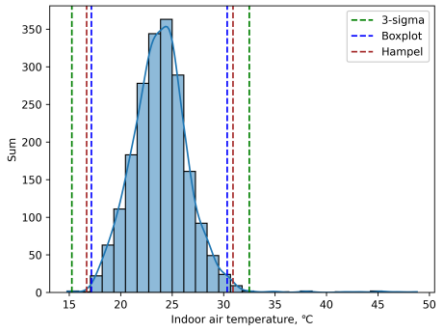
(d) Indoor air temperature in a humid subtropical



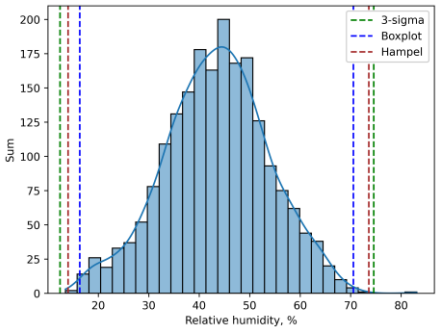
(e) Relative humidity in a humid subtropical



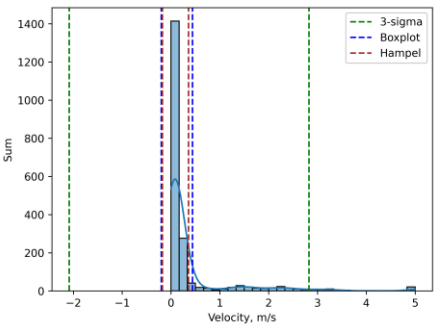
(f) Air velocity in a humid subtropical



(g) Indoor air temperature in temperate oceanic



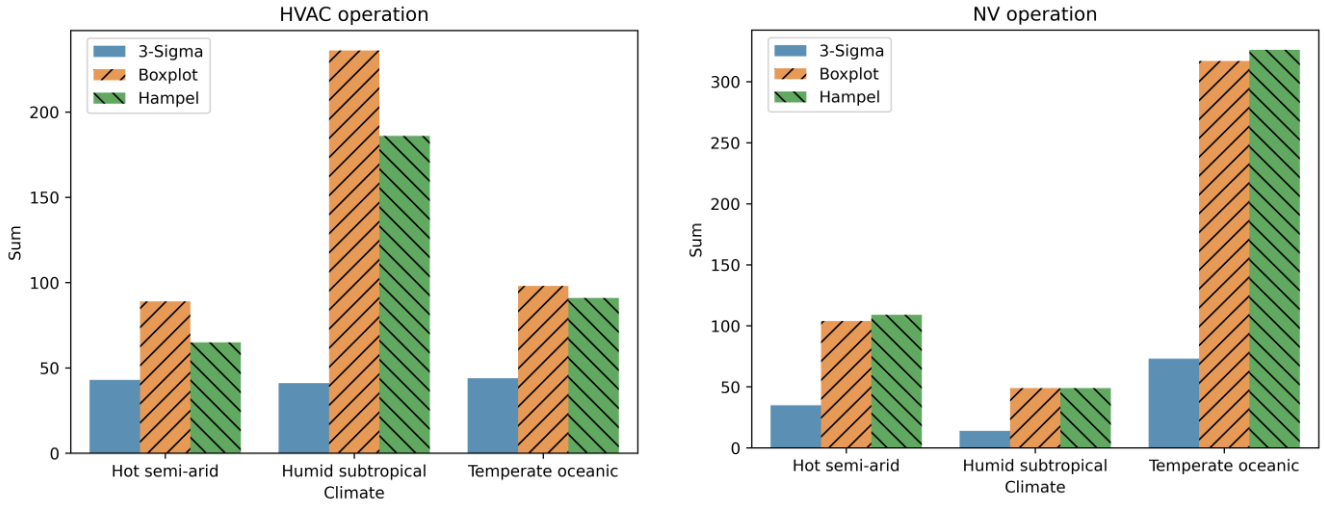
(h) Relative humidity in temperate oceanic



(i) Air velocity in temperate oceanic

Fig. 7. Outlier removal in NV buildings in different climate zones

The total removal sum from three outlier detection rules is shown in Fig. 8, after combining all outliers detected based on air temperature, relative humidity, and air velocity. The two median-value based rules (Boxplot and Hampel) generally remove more outliers than the mean-value based rule (3-Sigma).



(a) HVAC results (total sum is 2008, 1836, and 2466 in Hot semi-arid, humid subtropical, and Temperate oceanic)

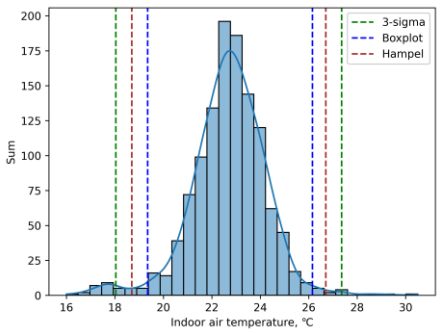
(b) NV results (total sum is 608, 1173, and 2003 in Hot semi-arid, humid subtropical, and Temperate oceanic)

Fig. 8. Removal sum in different climate zones

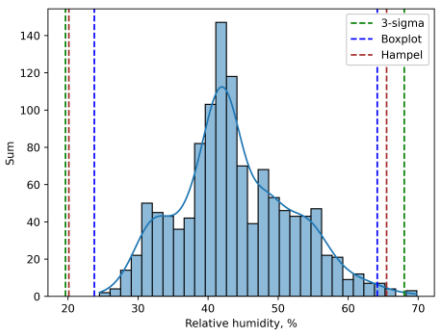
3.1.2 Different countries

Taking into account the available sample size in Comfort Database, data from the HVAC operation buildings in the hot semi-arid climate and the NV buildings in temperate oceanic climate were chosen to be broken down into country levels and further investigated whether there are differences between countries in the same climate.

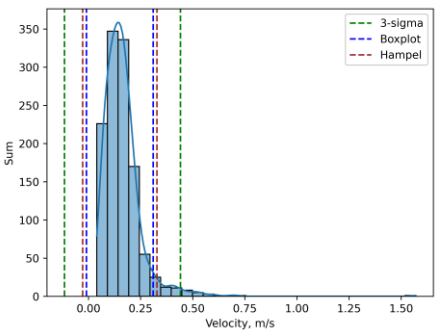
Fig. 9 shows the results after dividing the data from HVAC buildings in hot semi-arid climates by countries Australia and India. The results of air temperature and air velocity removal at the country level show similar trends of removal at the climate level, but relative humidity each presents a single-peak distribution with a different peak value (Fig. 9b and 9e). The three rules can all detect some outliers in humidity at the country level. However, combining them at the climate level leads to a double-peak distribution (Fig. 6b), making it more difficult for these rules to find outliers.



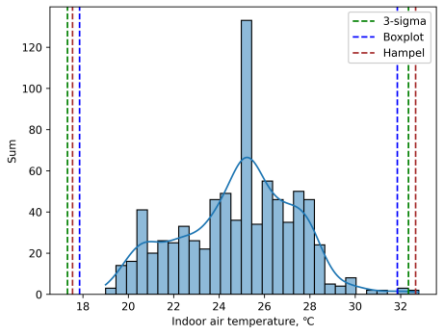
(a) Indoor air temperature in Australia



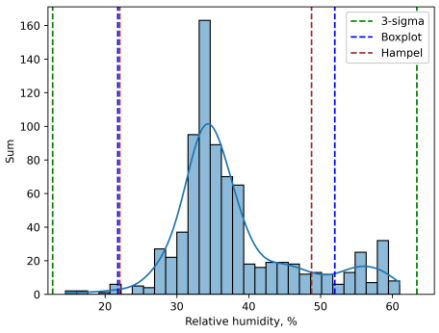
(b) Relative humidity in Australia



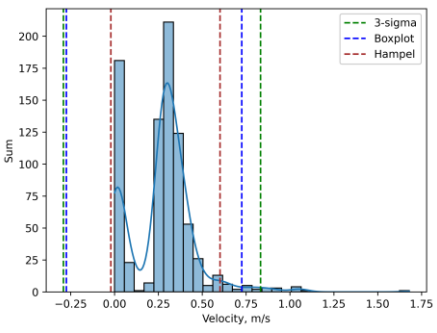
(c) Air velocity in Australia



(d) Indoor air temperature in India



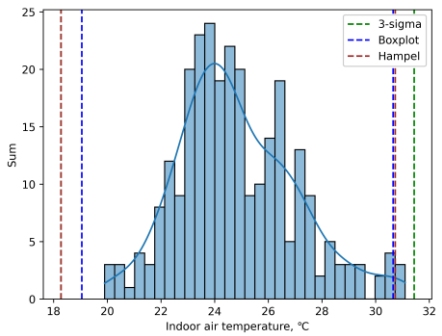
(e) Relative humidity in India



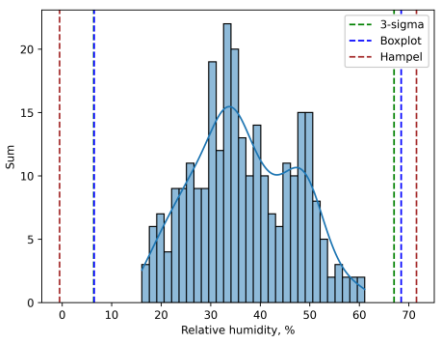
(f) Air velocity in India

Fig. 9. Outlier removal in HVAC buildings in different countries from hot semi-arid (Fig. 6a is divided into Fig. 9a and 9d; Fig 6b is divided into Fig. 9b and 9e; Fig. 6c is divided into Fig. 9c and 9f)

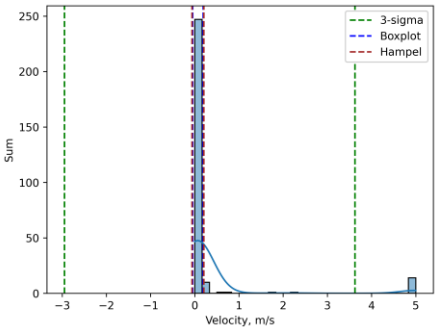
Fig. 10 shows the results of outlier removal of data in France, Germany, and the UK in NV buildings from the temperate oceanic climate. Although all three rules can detect some air temperature outliers, the 3-Sigma rule still offers the most tolerance range. In the case of relative humidity, however, all three rules fail to identify outliers in the France and Germany cases (Fig. 10b and 10e), but identify the upper extreme values in the UK case (80% RH in Fig. 10h). This may be due to the reason why that data of relative humidity in France and Germany cases do not strictly follow the normal distribution with many leaks in the middle range, pushing the 1/4 and 3/4 percentiles further away from middle compared to a more standard normal distribution, such as UK case. Therefore, the IQR value used in the Boxplot rule has been increased and the range will be expended consequently. With similar causality, the Hampel rule's median absolute deviation (MAD) scale will be magnified and lead to a wider boundary range.



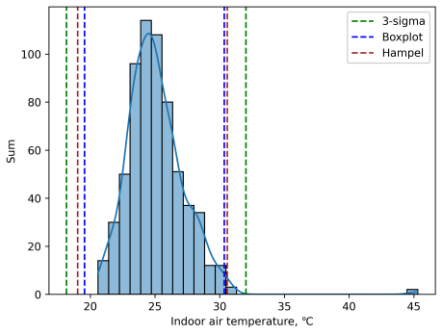
(a) Indoor air temperature in France



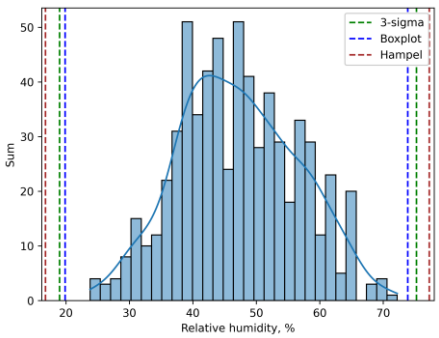
(b) Relative humidity in France



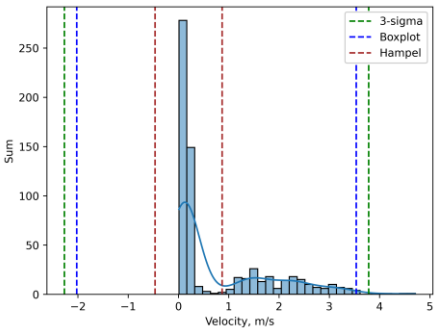
(c) Air velocity in France



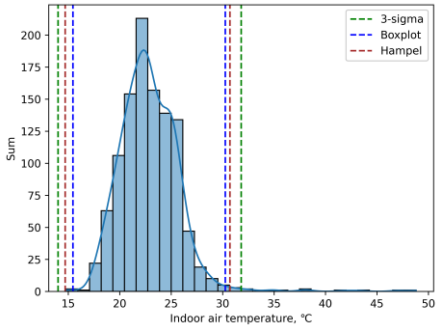
(d) Indoor air temperature in Germany



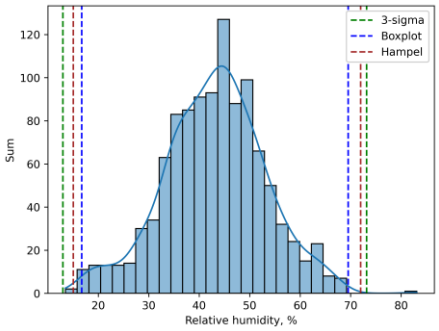
(e) Relative humidity in Germany



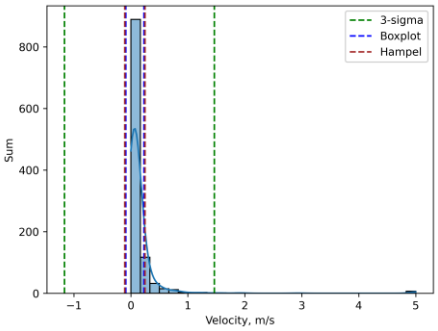
(f) Air velocity in Germany



(g) Indoor air temperature in the UK



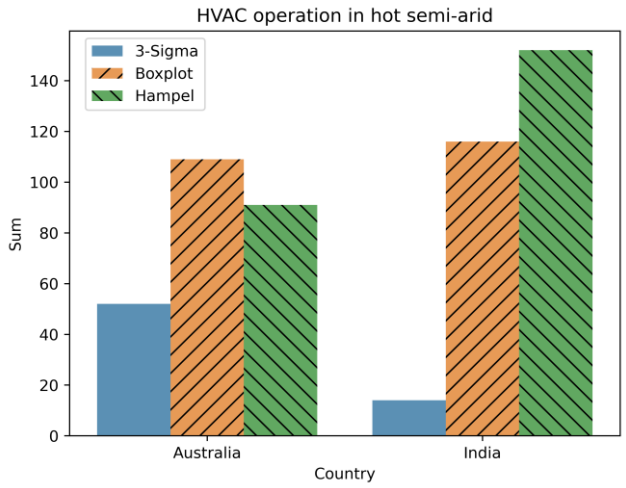
(h) Relative humidity in the UK



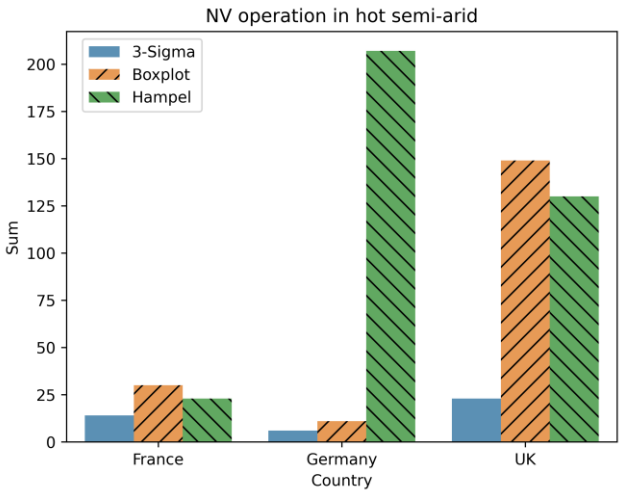
(i) Air velocity in the UK

Fig. 10. Outlier removal in NV buildings in different countries from temperate oceanic climates (Fig. 7g is divided into Fig. 10a, 10d, and 10g; Fig 7h is divided into Fig. 10b, 10e, and 10h; Fig. 7i is divided into Fig. 10c, 10f, and 10i)

The final outlier removal sums are shown in Fig. 11. The 3-Sigma rule still presents the most conservative solution. However, in the Germany case, the Hampel rule removed far more sum of outliers compared with the 3-Sigma and Boxplot rules. The Hampel rule is found to set a very narrow range for air velocity data around 1 m/s, whereas the other two rules set around 3.5 to 4 m/s. This implies that the Hampel rule can effectively remove outliers when the data is extremely imbalanced the mean value can represent the main information in the data, which is consistent with the simulated results in Fig. 3h that the 3-Sigma, Boxplot, and Hampel rules detected 2/300, 159/300, and 296/300 outliers, respectively. If the air velocity data in Fig. 10f between 1 and 3.5 m/s are considered outliers, the Hampel rule may be the only effective one. Otherwise, the Hampel rule may be too aggressive.



(a) HVAC removal results (total sum is 1202 and 806 in Australia and India)



(b) NV removal results (total sum is 275, 643, and 1085 in France, Germany, and the UK)

Fig. 11. Removal sum in different counties

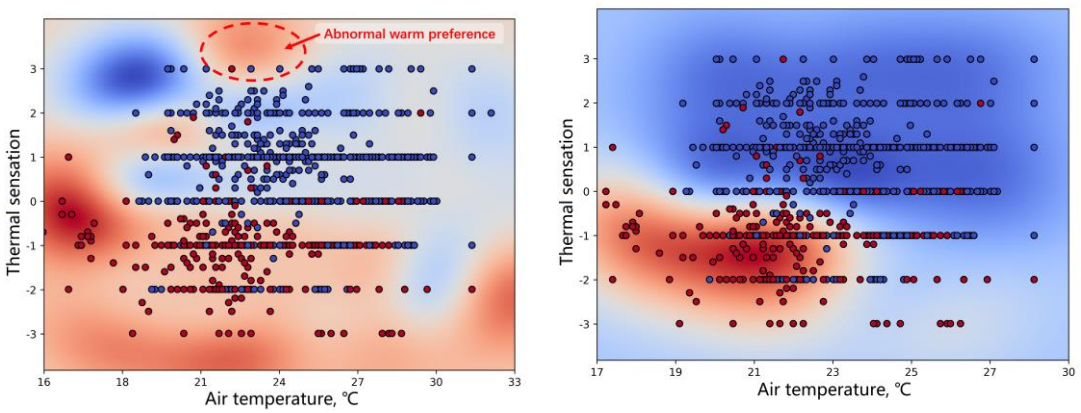
3.2 SVM prediction under three outlier removal rules

Instead of going through each case, this paper focuses on the hot semi-arid (BSh) climate under SVM predictions, which has the highest outdoor air temperature of the three climates, resulting in 652 "*prefer warmer/cooler*" votes in HVAC buildings and 753 "*prefer warmer/cooler*" votes in NV buildings, as shown in Table 2. The results of the outlier detection applied to subjective thermal preference in the Comfort Database are shown in Fig. 11 and Fig. 12. Two-dimensional data "*indoor air temperature*" and "*thermal sensation vote*" have been used to predict thermal preference. Fig. 12, Fig. 13, Table 3, and Table 4 present the SVM classification results of occupants' thermal preferences across different temperature ranges under HVAC and NV operations. Fig. 12a/13a shows the prediction models built without using outlier removal; Fig. 12b/13b, Fig. 12c/13c, and Fig. 12d/13d show the models built with outliers removed based on the 3-Sigma rule, Boxplot rule, and Hampel rule, respectively. The red and blue circles represent actual thermal preferences for warmer and cooler temperatures, while the red and blue mesh boundaries were generated by SVM classification using the best accuracy prediction.

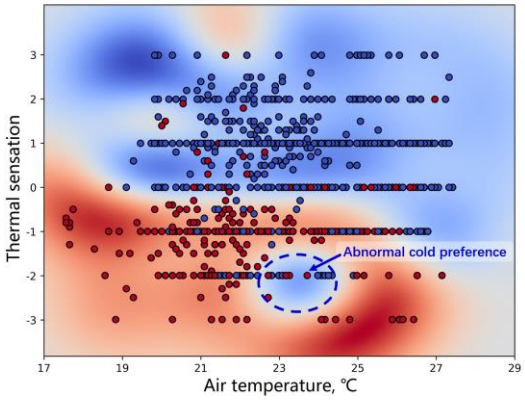
Table 2. Data sum of voting prefer warmer or cooler in HVAC and NV buildings

Climate	HVAC			NV		
	Total in HVAC	Prefer warmer or cooler	Percentage of prefer changing	Total in NV	Prefer warmer or cooler	Percent of prefer changing
Hot semi-arid (BSh)	2008	652	32.5%	1836	753	41.0%
Humid subtropical (Cfa)	2466	953	38.6%	608	428	70.4%
Temperate oceanic (Cfb)	1173	424	36.1%	2003	655	32.7%

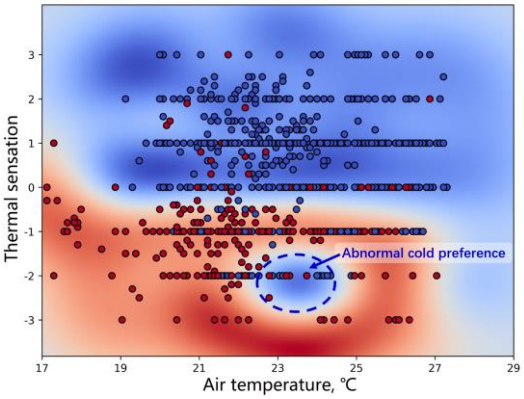
The SVM algorithms achieved relatively high predictive accuracy before and after outlier removal, ranging from 79.7% to 83.6% in Fig. 12. Before outlier removal, the SVM achieved 82.7% accuracy in Fig. 12a, indicating that the SVM can adequately adapt to extreme values and generate corresponding boundaries for specific groups of data. In Fig. 12a, one obvious predictive boundary with red colour (marked with a red dotted circle) in the top middle predicts that occupants in HVAC buildings will prefer warmer even when voting for the hottest sensation (+3), which is contradictory. However, the predictive performance of thermal preference has been improved after excluding potential outliers: the 3-Sigma and Hampel rule completely eliminated this abnormal warm preference (Figs. 12b and 12d), while the Boxplot rule weakens them (Fig. 12c). The Boxplot and Hampel removal rules both classify abnormal blue regions (prefer cooler when feeling cool) in the bottom middle that is surrounded by red regions (marked with blue dotted circles in Figs. 12c and 12d).



(a) No filter (accuracy: 82.7%)



(b) 3-Sigma rule (accuracy: 79.7%)



(c) Boxplot rule (accuracy: 81.8%)

(d) Hampel rule (accuracy: 83.6%)

Fig. 12. SVM models of thermal preference in HVAC buildings (red contour: prefer warmer; blue contour: prefer cooler)

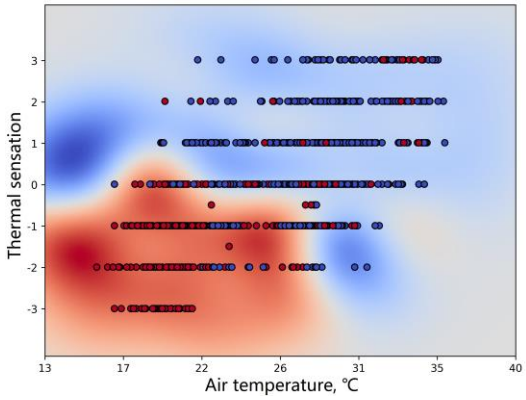
Table 3 shows the hyperparameters obtained by a grid search to achieve the best accuracy performance for SVM classification through cross-validation. The parameters γ remain constant for all of these four cases, but the parameters C related to the Boxplot and Hampel rules are both 10, which is lower than no outlier removal (100) and higher than the 3-Sigma rule (0.1). The lower value of parameter C will encourage the SVM classifier to use a larger margin, resulting in a simpler decision or boundary [98]. As parameters γ in all cases remain constant (γ = 1), the differences of parameter C are in Table. 3 are well supported by the SVM classification results in Fig. 12: the no filter rule (Fig. 12a) with the highest C value of 100 presents the most complicated boundaries, dividing most regions into preferred warmer categories with an obvious anomaly at the top; the 3-Sigma rule (Fig. 12b) with the lowest C value of 0.1 shows simplest boundaries; the C values of Boxplot and Hampel rules are both 10 in the median, which also has moderate degree of contours (Figs. 12c and 12d).

Table 3. Best combinations of SVM hyperparameters in HVAC buildings

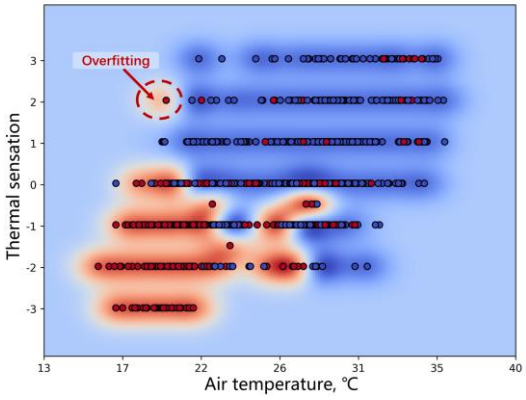
Parameter	No filter	3-Sigma	Boxplot	Hampel
C	100	0.1	10	10
γ	1	1	1	1

Similarly, Fig. 13 illustrates SVM classification results in NV buildings. When

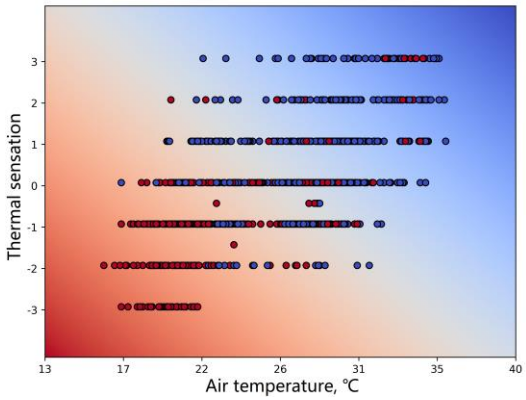
compared to the case without filtering outliers (Fig. 13a), using the 3-Sigma rule for outlier removal results in overfitting, which causes the SVM classifier to try to remember the features of the training data rather than generating meaningful patterns (Fig. 13b). This could be caused by the limitation of a grid search method in determining the best pair of hyperparameters C and γ to achieve best predictive accuracy. In Fig. 13c, the SVM results with the Boxplot removal rule show the most general contours but the lowest predictive accuracy. This is due to the extremely low hyperparameter value of 0.001, which allows for a broad decision region while tolerating higher bias [105]. In Fig. 13d, the SVM classifier with the Hampel removal rule has the most complex boundaries and the highest hyperparameter C of 1000, with two unexpected blue predictive regions among the red regions at the bottom (marked with blue dotted circles). These complex boundaries will also make it difficult to link occupants' real-time feedback to building operation decision-making. For example, if the indoor temperature is around 31 °C, as shown in the green background in Fig. 13d, the SVM classifier will recommend *seven* different types of strategies for building operation, which is complicated and less energy efficient.



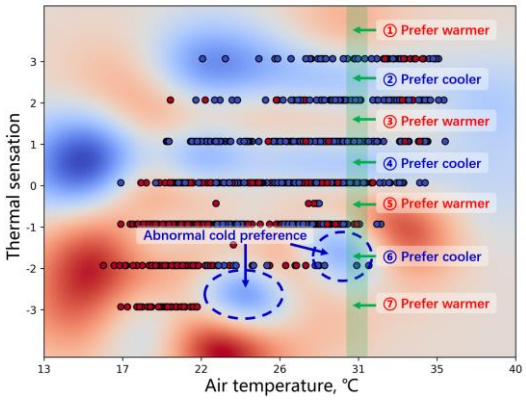
(a) No filter (accuracy: 83.5%)



(b) 3-Sigma rule (accuracy: 84.3%)



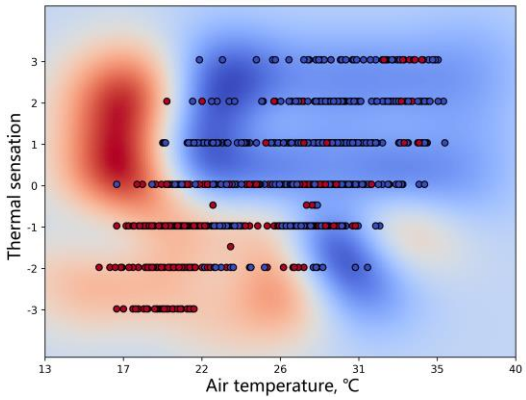
(c) Boxplot rule (accuracy: 80%)



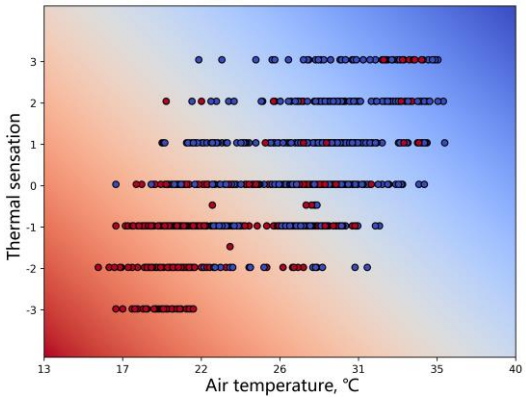
(d) Hampel rule (accuracy: 83.6%)

Fig. 13. SVM models of thermal preference in NV buildings (red contour: prefer warmer; blue contour: prefer cooler)

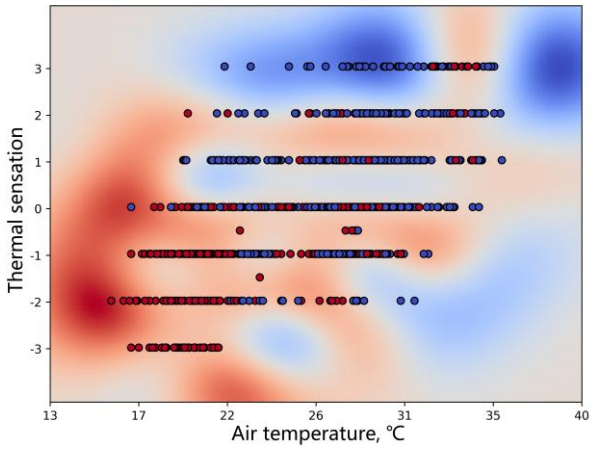
Fig. 14 depicts the SVM classification results using the data under the 3-Sigma removal rule with hyperparameters from the no filter, Boxplot, and Hampel rules in Table 4. The overfitting phenomenon has been greatly reduced by sacrificing the predictive accuracy of 0.3% to 2.4%. As a result of the preceding examples in Fig. 13b and Fig. 14 using the same data under the 3-Sigma removal rule, it is suggested that the hyperparameter setting plays a significant role in influencing the outlines of the decision boundary, while this hyperparameter setting is affected by the data points, validation method, and search resolution/step.



(a) Hyperparameters in no filter case (accuracy: 84.0%)



(b) Hyperparameters in Boxplot rule case (accuracy: 81.9%)



(c) Hyperparameters in Hampel rule case (accuracy: 83.4%)

Fig. 14. SVM models with data under the 3-Sigma removal rule with different pairs of hyperparameters C and γ

Table 4. Best combinations of SVM hyperparameters in NV buildings

Parameter	No filter	3-Sigma	Boxplot	Hampel
C	10	1	1	1000
γ	1	10	0.001	1

3.3 Trade-off between predictive accuracy and complexity of decision boundary

Because machine learning algorithms typically use the black box method to find the best hyperparameters under the constraints of accuracy or other evaluation approaches, overfitting can occur in some cases, making the established machine learning model difficult to extend to other new datasets. This can be illustrated in Fig. 13b, the SVM classifier generates one overfitting cluster with only one data point that ensures overall predictive accuracy highest. Furthermore, when a single evaluation criterion (accuracy or confusion matrix) is used as the sole criterion, the decision boundary could become too complicated to provide practical value at the application level. This is shown in Fig. 13d that although the accuracy is second best just below the overfitting case of Fig. 13b, it generates too many clusters in certain intervals. For example, in the temperature range of 26 to 30°C, seven different clusters have to be fulfilled to maintain the expected predictive accuracy, which will increase the difficulty for building design or operational systems to generate effective solutions for satisfying occupants' thermal preferences in

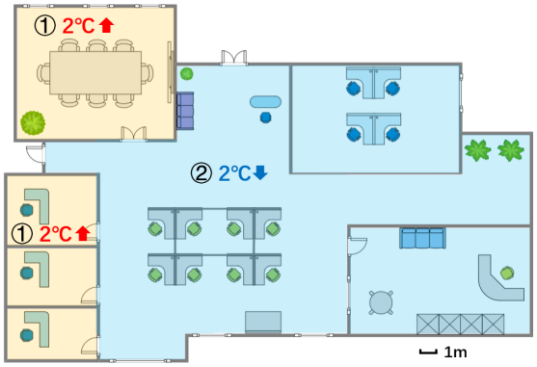
practice.

To better illustrate this scenario, Figs 15a, and 15b are used to simulate two hypothetical office cases that correspond to the cluster results in Figs 13d and 13c. The red arrows and numbers reflect the warm preference and expected temperature adjustment, while the blue ones represent the cool preference. It will be difficult for buildings to provide suitable and personalized operation solutions when people with different preferences clustered by machine learning algorithms are in the same shared space. One example in Fig. 15a is that there is not enough space for each cluster to generate an isolated room due to the high cluster sum, and people in zones 4 and 5 with different preferences must share the public environment, and preference conflict will be difficult to eliminate if personal comfort systems (PCS) [106] are not available. This problem is partially solved in Fig. 15b: because of the low cluster sum, it is easier for people with similar thermal preferences to congregate in one area, reducing the number of customized thermal zones and relieving preference conflict. Relevant information about thermal preferences can be obtained by either offline investigation or online monitoring, such as gathering historical thermal responses [107] [108], adjusting temperature setpoints from occupants [109], uploading real-time thermal feedback through smartphones [110], utilizing physiological signals like heart rate [111], facial temperature [112], wrist temperature [113], etc.

When a building is required to provide too many customized environments based on various subjective feedbacks, heat exchanges between rooms will occur through internal walls. To compensate for these heat exchanges and maintain a stable level of personalized temperatures, more energy will be used for heating or cooling in each room. In Fig. 15a, for example, people in zone 5 expect temperatures to rise by 2 ℃, while those in zone 6 expect temperatures to fall by 1 ℃. Assume zones 5 and 6 both meet people's thermal requirements and ignore climate or other influential factors, namely that the temperature in zone 5 is 3 °C higher than it is in zone 6, and heat will constantly flow through the internal walls between zones 5 and 6. To compensate, the building must implement a less energy-efficient strategy that provides more heating in

zone 5 and more cooling in zone 6. However, in the case shown in Fig. 15b, fewer clusters are required, allowing the building to maintain stable thermal environments with fewer fluctuations. As a result, less heat loss will be exchanged between different rooms via internal walls, which is beneficial for lowering total energy loads.

(a) Seven clusters with a predictive accuracy of 83.6% under Hampel filtering



(b) Two clusters with a predictive accuracy of 80% under Boxplot filtering

Fig. 15. Hypothetical cases in the office corresponding to cluster results under SVM classification

Employing different filtering rules will lead to different data distributions, which affects the selection of the best pair of hyperparameters in the SVM classifiers model, resulting in changes in prediction accuracy and decision boundary, but overall predictive accuracy remains high (80-84.3%). Fig. 16 depicts the number of clusters in different temperature ranges of the four SVM classifiers from Fig. 13 in section 3.2. If predictive accuracy is the only concern, the cases under 3-Sigma and Hampel rules are more likely to be accepted. However, based on the specific decision boundary, it is observed that the 3-Sigma case has the overfitting problem in the temperature range of 17-21°C, while

the Hampel case suggests seven clusters in the temperature range of 26-30°C. These two cases in practice could have similar unfavourable situations discussed in Fig. 15a. Therefore, results from the Boxplot rule will present the most competitive solution because fewer clusters are required, and extreme values have been removed.

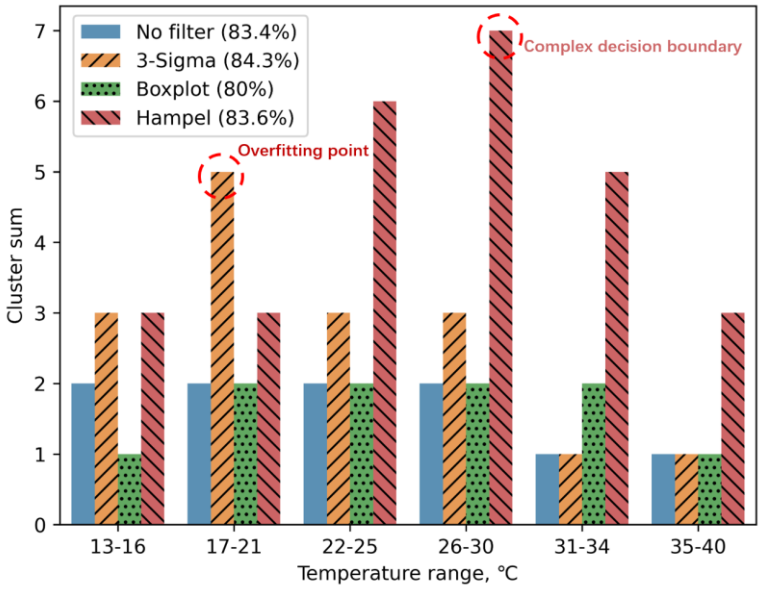


Fig. 16. Cluster sum of four different pairs of hyperparameter in SVM classifier in different temperature ranges

4. Discussion

4.1 Filtering performance of three outlier removal rules

In summary, the filtering results of simulated cases (Fig. 3 and Fig. 4) and data from public datasets in various climates (from Fig. 6 to Fig. 8) and countries (from Fig. 9 to Fig. 10) revealed:

- *3-Sigma rule*: it is influenced by the mean value and standard deviation of the dataset and is easily distracted by extreme values. When the contamination level exceeds 20% in simulated cases, it fails to detect any outliers. It is typically the most conservative strategy and filters out the fewest outliers, such as setting the upper

limit of air velocity as high as 3 m/s in Fig. 6i.

- *Boxplot rule*: it is related to the 1/4, 2/4, and 3/4 percentiles of data distribution and can resist a certain level of extreme value influence. When the analyzed data is close to the standard normal distribution in public datasets, it filters the outliers with the highest sum (Figs. 6a, 6c, 6d, 6g, 7g, 7h, 8a, 9a, 9d, 9g, and 9h).
- *Hampel rule*: it is calculated using the 2/4 percentile and the median absolute deviation (MAD). It performs similarly to the Boxplot rule with extreme value resistance, but occasionally provides too aggressive solutions, such as suggesting air velocity below 1 m/s during natural ventilation in Fig. 10f.
- *Data distribution*: When the data distribution does not conform to a strict normal distribution, the above three removal rules may all fail to detect any outliers, such as cases with two peaks (Figs. 6c, 6h, 7a, 7d, 10b) and leaks in the middle (Fig. 10e). The three rules all detect few outliers of relative humidity in HVAC buildings from hot semi-arid climates with double peak distribution, but when data is broken down from climate level (Fig. 6b) to country level (Figs. 9b and 9e), these detection rules perform better on selecting extreme values. Therefore, the activity of combining data from different sources may result in new data distributions that are unsuitable for outlier detection using the three rules described above.

Even though the above three rules have different theoretical foundations and performance, they can all detect some obvious outliers, such as the nearly 100% humidity level in Fig. 7e. The outlier removal principle in these three rules is based on the hypothesis that if some observations deviate too far from one reference value (mean or median), they are labelled as outliers [48]. However, it is debatable whether these extreme values are simply carelessness errors or have a hidden meaning. Hughes et al. [114] conducted winter surveys in the living and bed rooms of the 65+ elderly and discovered that certain indoor air temperatures are far below the boxplot lower limits. These extremely low temperatures would have the most dangerous consequences for the elderly's thermal comfort and respiratory illness, and are worth further investigation.

Furthermore, general outlier removal rules may oversimplify the study conditions and risk removing valid data, such as regarding unique thermal preferences as abnormal voting patterns [34]. In general, the usage of three outlier removal rules involved in this paper has made certain assumptions and simplifications to eliminate data that deviates far from the expected value in the dataset. More research on irregular situations or anomaly detection is still needed.

4.2 Hyperparameters C and γ in SVM algorithms

The hyperparameters C and γ in the RBF kernel were determined using the *grid search* method with *5-fold cross-validation* to achieve the best *accuracy* scoring. Outlier removal using 3-Sigma, Boxplot, and Hampel rules will influence this hyperparameter determination and thus affect the training process of the SVM algorithm.

- *Hyperparameter C*: when C is large, a big penalty is assigned to margin errors, causing the hyperplane to be close to the data points and usually presenting complex decision boundaries (Figs. 12a and 13d, C = 100 and 1000). These complex decision boundaries will necessitate facility managers or HVAC systems in buildings to run more personalized scenarios to meet different people's preferences and increase the difficulty of operation.

- *Hyperparameters γ*: when γ is large, more local vectors are allowed to participate in the boundary decision process, which increases the risk of overfitting (Fig. 13b, γ = 10). However, if γ is too small, the decision boundaries may be nearly linear and contribute less to personalizing the comfort model (Fig. 13c, γ = 0.001). The small γ value may also sacrifice model accuracy as fewer local vectors are involved in the boundary decision process.

- *Outlier removal rules*: After removing outliers, the predictive accuracy of SVM models does not vary significantly and remains relatively high at around 80% to 84%. However, removing extreme data points may prevent the SVM algorithm from producing unreasonable classifications, such as people voting +3 prefer still warmer at the top of Fig. 12a. The Hampel rule removed nearly 1/3 of the dataset

simply by considering the parameter air velocity (Fig. 10f), indicating that the Hampel rule may be too aggressive in cases where median value cannot represent the main information of the entire data distribution.

Traditional thermal comfort models, such as PMV, have been criticized for their poor predictive accuracy in real-world buildings (34% in [24]). On the contrary, many studies have used SVM algorithms to predict thermal comfort with high accuracy (76.7% in [115], 87% in [94], 89% in [116], and 90% in [95]). This paper discovered that several pairs of tuning hyperparameters in RBF kernel functions can all ensure overall predictive accuracy with only minor fluctuations (79.7% to 84.3%), but the specific pair of hyperparameters may lead to the SVM algorithm generating more user-friendly operation strategies for real-world buildings with only a minor reduction in accuracy (less individualized heating or cooling areas that still meet the majority of people's preferences). Therefore, it is questionable whether the extremely high accuracy or other evaluation indices should be the only dominant benchmark when the overall performance of machine learning algorithms is already very high.

4.3 Limitation and future work

This paper focuses on stochastic-based outlier detection methods, the performance of which is heavily influenced by the amount and distribution of data. Its limitations and future research directions could be:

1. To predict thermal preference, the training process of SVM models only considers the two representative dimensions: thermal sensation vote and indoor air temperature. More dimensional representations can assist models in making more comprehensive decisions.

2. The optimal hyperparameter matching obtained in this paper has room for improvement due to the limitation of the search method and search step. Although the RBF kernel function is chosen in this paper due to its nonlinear capability and widespread acceptance in thermal comfort research, other kernel functions, such as linear kernel [117], may also show good predictive outcomes in specific situations.

The confusion matrix here only discusses the accuracy, and combination with other indicators, such as recall and F1-score, could lead to a more reasonable evaluation of machine learning models.

3. The performance of outlier removal rules was evaluated using the SVM algorithm as a case study due to the aim and length limitations of this paper. It is anticipated that these outlier removal methods will be extended to other machine learning algorithms in the future to achieve more generalised outcomes.

4. The amount of data considered in this paper is limited due to the data sum and features stored in the original Comfort Database and the criteria for selecting data in this paper, resulting in selecting data from three climate zones. It is suggested to conduct more performance evaluations at a broader range of climate data in the future.

5. Although the stochastic-based methods discussed in this paper are mathematically well-grounded, prior knowledge and a specific data sum are required to accurately estimate the data distribution during the implementation stage, which is often costly or even unavailable in thermal comfort practice. Therefore, it would be valuable to explore distance-based or other unsupervised outlier detection methods that can overcome the limitations of stochastic-based methods.

5. Conclusions

Outliers in thermal comfort data can significantly skew the performance of machine learning algorithms. This study compares the effectiveness of three stochastic-based outlier detection techniques on data preprocessing namely 3-sigma, Boxplot upper, and Hampel, and their predictive performance based on the Comfort Database. The following findings and suggestions emerge:

- The characteristics of data distribution influence the performance of statistical-based outlier detection methods. When the data distribution is bimodal, all three

stochastic-based methods are prone to miss outliers. The risk of meta-analysis, such as analyses based on the Comfort Database that combine data from many individual scholars' studies, is that the inherent distribution of the original data will be transferred to a new distribution with completely fresh statistical features, such as combining two groups of data with Gaussian distribution with different mean values results in a bimodal distribution. This may increase the difficulty for stochastic-based methods to eliminate outliers.

- The 3-Sigma rule tolerates extreme values that may fail to mark outliers, whereas the Hampel rule can be too aggressive, resulting in false alarms. In general, the Boxplot rule provides the best performance and robustness in data preprocessing. Some anomalies in SVM classifications will disappear after outlier removal. The SVM classifiers can maintain a high level of predictive accuracy both before and after outlier removal.

- Specific outlier detection approaches will lead to different hyperparameters in machine learning models, resulting in changes to the classification boundary and model prediction accuracy. A small trade-off between cluster boundary and predictive accuracy can lead to more energy-efficient strategies and fewer thermal preference conflicts. Therefore, it is debatable whether predictive accuracy should always be the only benchmark for evaluating the effectiveness of a machine learning model for thermal comfort research in buildings.

- For future applications with approximately normally distributed data, the Boxplot rule is the most recommended one in most cases based on the filtering performance of the simulation and the Comfort Database in this paper. The 3-Sigma rule is recommended if it is expected to filter out fewer extreme values or if the proportion of outliers in the known monitoring data is very low. The Hampel rule can be used when the distribution of monitoring data is asymmetrical or an extreme filtering effect is expected. In general, having a fixed mode to achieve the desired filtering results in all situations is difficult due to the ever-changing nature of data in reality, as well as differences in specific analyses and requirements. Meanwhile, the

extreme values that were filtered out may be outliers, but they may also contain special patterns or potentials that should be investigated further.

Acknowledgement

The Chongqing University team appreciates the grants support from the National Key R&D Program of China [Grant No: 2022YFC3801504] and the Natural Science Foundation of Chongqing, China (Grant No. cstc2021ycjh-bgzxm0156). Mr Shaoxing Zhang acknowledges the financial support from the program of the China Scholarship Council (No. 202006050214).

Appendix

Appendix A. Thermal comfort research using Comfort Database

Reference	Environmental parameters	Subjective metrics	Contextual factors	Data sum	Algorithm
Hu et al. [35]	• PMV inputs	• TSV	-	11,164	• Transfer learning
Vellei et al. [118]	• PMV inputs	• TSV	• Time of day	21,000	• Regression
Cheung et al. [24]	• PMV	• TSV • TCV	• Building types • Climate • Operation strategy	56,771	• Regression
Li et al. [25]	• PMV	• TSV	• Age • Building type • Climate • Gender • Season	17,841	• Correspondence analysis • Regression
Li et al. [29]	• PMV • Air temperature	• TSV • TCV • TS	• Building type • Continent • Season	94,145	• Statistical analysis

		• TPV			
Zhang and de Dear [44]	• PMV inputs • Outdoor air temperature	• TSV	• Building type • Gender • Operation strategy • Season	18,966	• Regression
Wang et al. [22]	• PMV inputs • Outdoor air temperature	• TSV	• Building type • Continent • Gender • Operation strategy • Season	11,717	• Correspondence analysis • Regression
Gaffoor et al. [27]	• Air temperature • Radiant temperature • Operative temperature • PMV • Outdoor air temperature	• TSV	• Building type • Season	1,121	• Regression
Forgiarini et al. [31]	• Air temperature • Clothing level • Relative humidity • Outdoor air temperature	• TSV	• Building type • Continent	63,377	• Regression
Wang et al. [32]	• Operative temperature	• TSV	• Age • Building type	57,908	• Regression

			• Climate • Country • Gender • Height and weight • Operation strategy		
Rupp et al. [120]	• Clothing level • PMV • Air temperature	• TSV	• Building type • Season • Operation strategy	58,954	• Regression
Luo et al. [36]	• Air temperature • SET • Relative humidity • Clothing level • Air velocity • Metabolic rate • Outdoor air temperature	• TSV	• Age • Building type • Gender • Operation strategy • Season	10,619	• Decision tree • Gradient Boosting Machine • K nearest neighbors • Linear regression • Neural network • Random forest • Support vector machine
Zhou et al. [37]	• PMV inputs • Outdoor air temperature	• TSV	• Operation strategy	20,954	• Neural network • Support vector machine
Schweiker [45]	• PMV inputs	• TSV	• Building type	57,084	• Regression

	• Outdoor air temperature • Air temperature • Operative temperature • Relative humidity • Air velocity • Outdoor air temperature		• Operation strategy		
Ma et al. [38]	• PMV inputs • Outdoor air temperature	• TSV • TPV	• Age • Control behavior • Weight	78,113	• Bayesian neural network
Yao et al. [26]		• TSV	• Climate • Country	7,837	• Regression

Note: TSV = Thermal sensation vote; TCV = Thermal acceptability vote; TS = Thermal satisfaction; TPV = Thermal preference vote; PMV inputs: air temperature, radiant temperature, relative humidity, air velocity, clothing level, and metabolic rate; SET: standard effective temperature.

Appendix B. Thermal comfort research with outlier removal

Reference	Theme	Sample size	Removing method	Analyzed parameters	Main finding
Li et al. [29]	Acceptable temperature ranges in real buildings	62,444 responses from a public dataset	Fixed range (10^{th} and 90^{th} quantiles)	·Neutral temperature	People's acceptable temperature range in real buildings is wider than the standard recommendation.
Thapa and Indraganti [121]	Adaptive thermal comfort model in	2,608 responses from 436 subjects	3-sigma	-	The comfortable temperature range in India is wider than published reports

	India				
Thapa et al. [122]	Adaptive thermal comfort model in India	444 responses from 34 subjects	3-sigma	·Clothing insulation	People in India found cooler temperature comfortable compared with the standard suggestion.
Thapa et al. [72]	Cold and cloudy climates in Indian	2,608 responses from 436 subjects	3-sigma	·Clothing insulation ·Comfort temperature	People's thermal sensations are less sensitive than PMV prediction
Li et al. [123]	Real-time monitor using cameras	12 subjects	3-sigma	·Non-facial pixels	Thermal comfort can be indicated by pixels from the ears, nose, and cheeks.
Li et al. [124]	Auto-track using cameras	16 subjects	Median filter	·Noisy pixels	Distribution of facial temperature can be related to thermal comfort state.
Chen et al. [125]	CFD analysis in kitchen	20 group data for simulation	6-sigma	·Particulate concentration	Exhaust volume around 11-14m^3 in kitchen is good for air quality and thermal comfort in kitchen
Hawila et al. [126]	Glass façade design	-	Turkey's (boxplot) and Grubbs' test	·Thermal sensation vote	Proposed façade design can ensure a comfortable environment with PMV ranging from -0.381 to 0.107.
Hurtado et al. [127]	Demand flexibility on thermal storage and	60 building energy simulations	Boxplot	·Ramp rate ·Power capacity	Buildings in a hot climate can offer higher flexibility potential compared

	comfort management				with a cold climate
Elnaklah et al. [73]	Green building performance	120 subjects	Boxplot	·Occupant satisfaction ·Air temperature ·Relative humidity ·CO2 (ppm)	Thermal comfort in surveyed green buildings has been improved, but not air quality, visual and acoustic comfort.
Noda et al. [128]	School children in air-conditioning classroom	97 subjects	Boxplot	·Thermal sensation vote	In air-conditioned classrooms, 34.01% of children reported cold discomfort and 30.93% reported hot discomfort.
Gautam et al. [129]	Thermal history influence between local and migrant people	395 subjects	Boxplot	·Indoor air temperature ·Globe temperature ·Radiant temperature ·Operative temperature ·Outdoor air temperature	When compared to locals, migrants reported more sweating and a lower preferred temperature.
Craenendonck et al. [130]	Cold discomfort caused by construction joints	56 subjects	Boxplot	·Thermal sensation vote	Small-area radiant asymmetry has limited effects on thermal sensation.
Su et al. [131]	Asymmetric radiant environment	66 subjects	Boxplot	-	Non-uniform environments can affect thermal

	at different exposure distances				acceptability.
Liu et al. [132]	urban spatial characteristics on outdoor thermal comfort	1,870 responses	Boxplot	·PET ·OUT-SET* ·UTCI	"Compact high-rise + scattered trees" appears to be a preferred urban design strategy in Shenzhen, China.
Rewitz and Müller [133]	Physiological responses in transient environments	48 subjects	Boxplot	·Thermal sensation vote ·Skin temperature	BMI may be the most influential factor in physiological response.
Zhang et al. [134]	Footwarmers in office	2,774 responses	Boxplot	·Thermal acceptability	Compared with central heating, a personal heating system can ensure both comfort and energy savings.
Lipczynska et al. [135]	Productivity under ceiling fans assisted HVAC environments	15 subjects	Boxplot	·Outdoor air temperature ·Outdoor relative humidity ·Solar radiation ·Rainfall	The shared control of the ceiling fan can raise the HVAC setpoint from 23°C to 26°C.
Zhao et al. [136]	Non-uniform thermal radiation in a street canyon	2,226 valid pedestrian routes	Boxplot	·f300 (frequency of the sunlight and shadow areas in a 300m long pedestrian route)	The chest, back, and head are most vulnerable body parts in non-uniform radiation environments.

Lau and Choi [137]	Aesthetic and acoustic quality on outdoor thermal comfort	1,917 responses	Boxplot	·UTCI ·Air temperature	A quiet and beautiful outdoor environment can increase people's thermal tolerance.
Aryal and Becerik-Gerber [74]	Personalized comfort model based on wearable devices or thermal imaging	20 subjects	Hampel	·Air temperature ·Wrist skin temperature	Data from wearable devices or the thermal cameras can provide 3%-5% additional prediction accuracy for machine learning algorithms.
Wang et al. [34]	Anomaly detection in thermal votes	11,000 responses	K-Nearest Neighbor	·Thermal sensation votes ·Thermal comfort votes	Occupants' strange votes will lead to bias in operation of building systems
Schweiker [45]	Combining adaptive and heat balance model	57,084 responses	Manual inspection	·Improving ATHB model	Outdoor climates, building types, and cooling strategies can be considered in improved model
Amasyali and El-Gohary [75]	Occupant-behavior on energy consumption and comfort	12 subjects	Cook distance	·Cooling and lighting energy consumption ·Occupant-behavior data	Optimizing occupant behavior can result in 11-22% energy savings and increased thermal comfort.
Manu et al. [77]	Building adaptive comfort model	6,330 responses	Fixed range	· Indoor temperature	Building adaptive model for mixed-mode building in India, and PMV will over-predict sensation.
Qiao et al.	Machine	628	Adjust	·Generated data	Adjusting ventilation

[138]	learning model for underground space	responses	sampling interval	from variational autoencoder (VAE)	is more effective for thermal comfort in underground space compared with adjusting temperature and humidity.
Tse and Jones [139]	Transitional space in buildings	1,316 responses	Binning values	·Clothing value ·Thermal sensation vote	People in transient places prefer to adjust themselves to restore comfort rather than making changes to building systems.
Khoshbakht et al. [140]	Control strategies in mixed-mode buildings	1,001 responses	Binning values	·Thermal sensation vote ·Excluded unwell or sick participants	The adaptive thermal comfort model is more suitable in semi-control or manual operation rather than fully automated operation.
Deuble and de Dear [78]	Thermal expectations in mixed-mode buildings	1,359 responses from 60 subjects	Binning values	·Thermal sensation vote	The adaptive comfort model is more applicable in mixed-mode buildings than PMV-PPD models.
Ma et al. [141]	Elderly's thermal perceptions in an urban park	1417 responses	Binning values	·PET	Except outdoor microclimate, park use by the elderly is also affected by facility form and space function.

References

[1] M. Frontczak *et al.*, "Quantitative relationships between occupant satisfaction and satisfaction aspects of indoor environmental quality and building design," *Indoor Air*, vol. 22, no. 2, pp. 119–131, 2012.

[2] W. P, W. DP, S. J, C. G, and F. PO, "The effects of outdoor air supply rate in an office on perceived air quality, sick building syndrome (SBS) symptoms and productivity," *Indoor Air*, vol. 10, no. 4, pp. 222–236, 2000.

[3] F. Yuan *et al.*, "Thermal comfort in hospital buildings–A literature review," *J. Build. Eng.*, vol. 45, p. 103463, 2022.

[4] K. W. Tham and H. C. Willem, "Room air temperature affects occupants' physiology, perceptions and mental alertness," *Build. Environ.*, vol. 45, no. 1, pp. 40–44, 2010.

[5] E. Kükrer and N. Eskin, "Effect of design and operational strategies on thermal comfort and productivity in a multipurpose school building," *J. Build. Eng.*, vol. 44, p. 102697, 2021.

[6] S. Altomonte *et al.*, "Ten questions concerning well-being in the built environment," *Build. Environ.*, vol. 180, p. 106949, 2020.

[7] P. Fanger, *Thermal comfort, Analysis and Applications in Environmental Engineering*. Copenhagen: Danish Technical Press, 1970.

[8] M. A. Humphreys and J. Fergus Nicol, "The validity of ISO-PMV for predicting comfort votes in every-day thermal environments," *Energy Build.*, vol. 34, no. 6, pp. 667–684, 2002.

[9] R. de Dear and G. Brager, "Developing an Adaptive Model of Thermal Comfort and Preference," *ASHRAE Trans.*, vol. 1041, pp. 1–18, 1998.

[10] L. Jiang and R. Yao, "Modelling personal thermal sensations using C-Support Vector Classification (C-SVC) algorithm," *Build. Environ.*, vol. 99, pp. 98–

106, 2016.

[11] Y. Wu and B. Cao, “Recognition and prediction of individual thermal comfort requirement based on local skin temperature,” *J. Build. Eng.*, vol. 49, p. 104025, 2022.

[12] 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 Build.*, vol. 166, pp. 391–406, 2018.

[13] S. S. Shetty, D. Chinh, M. Gupta, and S. K. Panda, “Learning desk fan usage preferences for personalised thermal comfort in shared offices using tree-based methods,” *Build. Environ.*, vol. 149, pp. 546–560, 2019.

[14] J. Langevin, J. Wen, and P. L. Gurian, “Modeling thermal comfort holistically: Bayesian estimation of thermal sensation, acceptability, and preference distributions for office building occupants,” *Build. Environ.*, vol. 69, pp. 206–226, 2013.

[15] S. Lee, I. Bilionis, P. Karava, and A. Tzempelikos, “A Bayesian approach for probabilistic classification and inference of occupant thermal preferences in office buildings,” *Build. Environ.*, vol. 118, pp. 323–343, 2017.

[16] W. Liu, Z. Lian, and B. Zhao, “A neural network evaluation model for individual thermal comfort,” *Energy Build.*, vol. 39, no. 10, pp. 1115–1122, 2007.

[17] W. Li, J. J. Squiers, E. W. Sellke, W. Fan, J. M. Dimaio, and J. E. Thatcher, “Outlier detection and removal improves accuracy of machine learning approach to multispectral burn diagnostic imaging,” *J. Biomed. Opt.*, vol. 20, no. 12, p. 121305, 2015.

[18] V. Földváry *et al.*, “Development of the ASHRAE Global Thermal Comfort Database II,” *Build. Environ.*, vol. 142, pp. 502–512, 2018.

[19] R. J. De, “A global database of thermal comfort field experiments,” in *Ashrae Winter Meeting*, 1998.

[20] M. Dawe, P. Raftery, J. Woolley, S. Schiavon, and F. Bauman, “Comparison of mean radiant and air temperatures in mechanically-conditioned commercial buildings from over 200,000 field and laboratory measurements,” *Energy Build.*, vol. 206, 2020.

[21] A. Forsthoff, P. Mehnert, and H. Neffgen, “Comparison of laboratory studies with predictions of the required sweat rate index (ISO 7933) for climates with moderate to high thermal radiation,” *Appl. Ergon.*, vol. 32, no. 3, pp. 299–303, 2001.

[22] L. Wang, J. Kim, J. Xiong, and H. Yin, “Optimal clothing insulation in naturally ventilated buildings,” *Build. Environ.*, vol. 154, no. 19, pp. 200–210, 2019.

[23] A. Nakagawa, H. Ikeda, Y. Maeda, and T. Nakaya, “A survey of high school students’ clothing in classroom,” *J. Build. Eng.*, vol. 32, p. 101469, 2020.

[24] T. Cheung, S. Schiavon, T. Parkinson, P. Li, and G. Brager, “Analysis of the accuracy on PMV – PPD model using the ASHRAE Global Thermal Comfort Database II,” *Build. Environ.*, vol. 153, pp. 205–217, 2019.

[25] Y. Li *et al.*, “Development of an adaptation table to enhance the accuracy of the predicted mean vote model,” *Build. Environ.*, vol. 168, p. 106504, 2020.

[26] R. Yao *et al.*, “Evolution and performance analysis of adaptive thermal comfort models – A comprehensive literature review,” *Build. Environ.*, vol. 217, p. 109020, 2022.

[27] M. A. Gaffoor, M. Eftekhari, and X. Luo, “Evaluation of thermal comfort in mixed-mode buildings in temperate oceanic climates using American Society of Heating, Refrigeration, and Air Conditioning Engineers Comfort Database II,” *Build. Serv. Eng. Res. Technol.*, p. 01436244211044670, 2022.

[28] W. Ji, Y. Zhu, and B. Cao, “Development of the Predicted Thermal Sensation (PTS) model using the ASHRAE Global Thermal Comfort Database,” *Energy Build.*, vol. 211, p. 109780, 2020.

[29] P. Li, T. Parkinson, G. Brager, S. Schiavon, T. C. T. Cheung, and T. Froese, “A data-driven approach to defining acceptable temperature ranges in buildings,” *Build. Environ.*, vol. 153, pp. 302–312, 2019.

[30] F. Zhang and R. de Dear, “Impacts of demographic, contextual and interaction effects on thermal sensation—Evidence from a global database,” *Build. Environ.*, vol. 162, p. 106286, 2019.

[31] R. Forgiarini, T. Parkinson, J. Kim, and R. De Dear, “The impact of occupant’s thermal sensitivity on adaptive thermal comfort model,” *Build. Environ.*, vol. 207, p. 108517, 2022.

[32] Z. Wang *et al.*, “Revisiting individual and group differences in thermal comfort based on ASHRAE database,” *Energy Build.*, vol. 219, p. 110017, 2020.

[33] Y. Bai, K. Liu, and Y. Wang, “Comparative analysis of thermal preference prediction performance in different conditions using ensemble learning models based on ASHRAE Comfort Database II,” *Build. Environ.*, vol. 223, no. 13, p. 109462, 2022.

[34] Z. Wang, T. Parkinson, P. Li, B. Lin, and T. Hong, “The Squeaky wheel: Machine learning for anomaly detection in subjective thermal comfort votes,” *Build. Environ.*, vol. 151, pp. 219–227, 2019.

[35] W. Hu, Y. Luo, Z. Lu, and Y. Wen, “Heterogeneous transfer learning for thermal comfort modeling,” *BuildSys 2019 - Proc. 6th ACM Int. Conf. Syst. Energy-Efficient Build. Cities, Transp.*, pp. 61–70, 2019.

[36] M. Luo *et al.*, “Comparing machine learning algorithms in predicting thermal sensation using ASHRAE Comfort Database II,” *Energy Build.*, vol. 210, p. 109776, 2020.

[37] X. Zhou *et al.*, “Data-driven thermal comfort model via support vector machine algorithms: Insights from ASHRAE RP-884 database,” *Energy Build.*, vol. 211, p. 109795, 2020.

[38] N. Ma, L. Chen, J. Hu, P. Perdikaris, and W. W. Braham, “Adaptive behavior and different thermal experiences of real people: A Bayesian neural network approach to thermal preference prediction and classification,” *Build. Environ.*, vol. 198, p. 107875, 2021.

[39] Z. Wang and T. Hong, “Learning occupants’ indoor comfort temperature through a Bayesian inference approach for office buildings in United States,” *Renew. Sustain. Energy Rev.*, vol. 119, p. 109593, 2020.

[40] Z. Wang, J. Wang, Y. He, Y. Liu, B. Lin, and T. Hong, “Dimension analysis of subjective thermal comfort metrics based on ASHRAE Global Thermal Comfort Database using machine learning,” *J. Build. Eng.*, vol. 29, p. 101120, 2020.

[41] J. P. Stevens, “Outliers and influential data points in regression analysis,” *Psychol. Bull.*, vol. 95, no. 2, pp. 334–344, 1984.

[42] R. Domingues, M. Filippone, P. Michiardi, and J. Zouaoui, “A comparative evaluation of outlier detection algorithms : Experiments and analyses,” *Pattern Recognit.*, vol. 74, pp. 406–421, 2018.

[43] R. Items, W. Rose, W. Rose, T. If, and W. Rose, “A survey of outlier detection methodologies,” *Artif. Intell. Rev.*, vol. 22, no. 2, pp. 85–126, 2004.

[44] F. Zhang and R. De Dear, “Impacts of demographic , contextual and interaction effects on thermal sensation — Evidence from a global database,” *Build. Environ.*, vol. 162, p. 106286, 2019.

[45] M. Schweiker, “Combining adaptive and heat balance models for thermal sensation prediction: A new approach towards a theory and data-driven adaptive thermal heat balance model,” *Indoor Air*, vol. 32, no. 3, p. e13018,

2022.

[46] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly Detection: A Survey,” *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, 2009.

[47] M. A. F. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, “A review of novelty detection,” *Signal Processing*, vol. 99, pp. 215–249, 2014.

[48] R. K. Pearson, *Mining imperfect data: Dealing with contamination and incomplete records*. Society for Industrial and Applied Mathematics, 2005.

[49] H. V, K. I, and F. P, “Outlier Detection Using k-Nearest Neighbour Graph,” *Proc. 17th Int. Conf. Pattern Recognition, 2004. ICPR 2004.*, vol. 3, pp. 430–433, 2004.

[50] M. Markou and S. Singh, “Novelty detection: a review — part 2: neural network based approaches,” *Signal Processing*, vol. 83, no. 12, pp. 2499–2521, 2003.

[51] B. Scholkopf, R. Williamson, A. Smola, J. Shawe-taylor, and J. Platt, “Support Vector Method for Novelty Detection,” *Adv. Neural Inf. Process. Syst.*, vol. 12, no. 3, pp. 582–588, 2000.

[52] W. Lee and D. Xiang, “Information-theoretic measures for anomaly detection,” *Proc. 2001 IEEE Symp. Secur. Privacy. S&P 2001*, pp. 130–143, 2001.

[53] A. Lakhina, M. Crovella, and C. Diot, “Mining anomalies using traffic feature distributions,” *ACM SIGCOMM Comput. Commun. Rev.*, vol. 35, no. 4, pp. 217–228, 2005.

[54] L. Tarassenko, A. Hann, D. Young, J. R. Hospital, H. Way, and O. Ox, “Integrated monitoring and analysis for early warning of patient deterioration,” *Br. J. Anaesth.*, vol. 97, no. 1, pp. 64–68, 2006.

[55] D. A. Clifton, P. R. Bannister, and L. Tarassenko, “A framework for novelty detection in jet engine vibration data,” *Key Eng. Mater.*, vol. 347, pp. 305–310,

2007.

[56] R. Ramezani, P. Angelov, and X. Zhou, “A fast approach to novelty detection in video streams using recursive density estimation,” *2008 4th Int. IEEE Conf. Intell. Syst.*, vol. 2, pp. 14–22, 2008.

[57] S. Basu, M. Bilenko, and R. J. Mooney, “A probabilistic framework for semi-supervised clustering,” *Proc. tenth ACM SIGKDD Int. Conf. Knowl. Discov. data Min.*, pp. 59–68, 2004.

[58] Y. Zhang, N. Meratnia, and P. Havinga, “Outlier detection techniques for wireless sensor networks: A survey,” *IEEE Commun. Surv. tutorials*, vol. 12, no. 2, pp. 159–170, 2010.

[59] D. A. Clifton, S. Hugueny, and L. Tarassenko, “Novelty Detection with Multivariate Extreme Value Statistics,” *J. Signal Process. Syst.*, vol. 65, pp. 371–389, 2009.

[60] D. Yeung and Y. Ding, “Host-based intrusion detection using dynamic and static behavioral models,” *Pattern Recognit.*, vol. 36, no. 1, pp. 229–243, 2003.

[61] S. Chawla and P. Sun, “SLOM: a new measure for local spatial outliers,” *Knowl. Inf. Syst.*, vol. 9, pp. 412–429, 2006.

[62] A. Nagaraja, U. M. A. Boregowda, and K. Khatatneh, “Similarity Based Feature Transformation for Network Anomaly Detection,” *IEEE Access*, vol. 8, pp. 39184–39196, 2020.

[63] P. Sun, S. Chawla, and B. Arunasalam, “Mining for Outliers in Sequential Databases,” *Proc. 2006 SIAM Int. Conf. Data Min.*, pp. 94–105.

[64] M. Xia, J. Sun, and Q. Chen, “Outlier Reconstruction Based Distribution System State Estimation Using Equivalent Model of Long Short-term Memory and Metropolis-Hastings Sampling,” *J. Mod. Power Syst. Clean Energy*, vol. 10, no. 6, pp. 1625–1636, 2021.

[65] L. Zhang, T. Tan, Y. Gong, and W. Yang, “Fingerprint Database Reconstruction Based on Robust PCA for Indoor Localization,” *Sensors*, vol. 19, no. 11, p. 2537, 2019.

[66] A. Rabaoui, H. Kadri, and N. Ellouze, “New approaches based on One-Class SVMS for impulsive sounds recognition tasks,” *2008 IEEE Work. Mach. Learn. Signal Process.*, pp. 285–290, 2008.

[67] S. F. Hussain, “A novel robust kernel for classifying high-dimensional data using Support Vector Machines,” *Expert Syst. Appl.*, vol. 131, pp. 116–131, 2019.

[68] D. R. Hardoon, U. K. So, and L. M. Manevitz, “fMRI Analysis via One-class Machine Learning Techniques,” *IJCAI’05 Proc. 19th Int. Jt. Conf. Artif. Intell.*, pp. 1604–1605, 2005.

[69] L. Clifton, D. A. Clifton, P. J. Watkinson, and L. Tarassenko, “Identification of Patient Deterioration in Vital-Sign Data using One-Class Support Vector Machines,” *Proc. Fed. Conf. Comput. Sci. Inf. Syst.*, pp. 125–131, 2011.

[70] F. Ye, H. Zheng, C. Huang, and Y. Zhang, “Deep unsupervised image anomaly detection: An information theoretic framework,” *2021 IEEE Int. Conf. Image Process.*, pp. 1609–1613, 2021.

[71] L. Itti and P. Baldi, “Bayesian surprise attracts human attention,” *Vision Res.*, vol. 49, no. 10, pp. 1295–1306, 2009.

[72] S. Thapa, A. Kumar, G. Kumar, and M. Indraganti, “Adaptive thermal comfort in the different buildings of Darjeeling Hills in eastern India – Effect of difference in elevation,” *Energy Build.*, vol. 173, pp. 649–677, 2018.

[73] R. Elnaklah, I. Walker, and S. Natarajan, “Moving to a green building: Indoor environment quality, thermal comfort and health,” *Build. Environ.*, vol. 191, no. December 2020, p. 107592, 2021.

[74] A. Aryal and B. Becerik-Gerber, “A comparative study of predicting individual thermal sensation and satisfaction using wrist-worn temperature sensor, thermal camera and ambient temperature sensor,” *Build. Environ.*, vol. 160, p. 106223, 2019.

[75] K. Amasyali and N. M. El-gohary, “Real data-driven occupant-behavior optimization for reduced energy consumption and improved comfort,” *Appl. Energy*, vol. 302, p. 117276, 2021.

[76] Z. Wang, T. Parkinson, P. Li, B. Lin, and T. Hong, “The Squeaky wheel: Machine learning for anomaly detection in subjective thermal comfort votes,” *Build. Environ.*, vol. 151, pp. 219–227, 2019.

[77] S. Manu, Y. Shukla, R. Rawal, L. E. Thomas, and R. de Dear, “Field studies of thermal comfort across multiple climate zones for the subcontinent: India Model for Adaptive Comfort (IMAC),” *Build. Environ.*, vol. 98, pp. 55–70, 2016.

[78] M. P. Deuble and R. J. de Dear, “Mixed-mode buildings: A double standard in occupants’ comfort expectations,” *Build. Environ.*, vol. 54, pp. 53–60, 2012.

[79] A. R. Martel, “Revised Upper Percentage Points of the Extreme Studentized Deviate from the Sample Mean,” *Publ. Astron. Soc. Pacific*, vol. 127, no. 949, p. 258, 1956.

[80] Y. Zhao, B. Lehman, R. Ball, J. Mosesian, and J. De Palma, “Outlier Detection Rules for Fault Detection in Solar Photovoltaic Arrays,” *2013 Twenty-Eighth Annu. IEEE Appl. Power Electron. Conf. Expo.*, pp. 2913–2920, 2013.

[81] J. W. Tukey, *Exploratory data analysis*. Addison-Wesley, 1977.

[82] F. R. Hampel, “The breakdown points of the mean combined with some rejection rules,” *Technometrics*, vol. 27, no. 2, pp. 95–107, 1985.

[83] L. Davies and U. Gather, “The Identification of Multiple Outliers,” *J. Am. Stat.*

Assoc., vol. 88, no. 423, pp. 782–792, 1993.

[84] W. G. Cochran, “Random sampling (numpy.random),” *The SciPy community*. [Online]. Available: https://numpy.org/doc/1.16/reference/routines.random.html. [Accessed: 01-Dec-2022].

[85] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 1999.

[86] E. Osuna and R. Freund, “Training support vector machines: an application to face detection,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. pattern Recognit.*, pp. 130–136, 1997.

[87] T. Joachims, “Text categorization with support vector machines: Learning with many relevant features,” *Eur. Conf. Mach. Learn. Springer, Berlin, Heidelb.*, pp. 137–142, 1998.

[88] X. Li, L. Wcing, and E. Sung, “Multilabel SVM active learning for image classification,” *2004 Int. Conf. Image Process.*, vol. 4, pp. 2207–2210, 2004.

[89] J. Bedo, C. Sanderson, and A. Kowalczyk, “An efficient alternative to svm based recursive feature elimination with applications in natural language processing and bioinformatics,” *Australas. Jt. Conf. Artif. Intell. Springer, Berlin, Heidelb.*, pp. 170–180, 2006.

[90] H. M. Muda, P. Saad, and R. M. Othman, “Remote protein homology detection and fold recognition using two-layer support vector machine classifiers,” *Comput. Biol. Med.*, vol. 41, no. 8, pp. 687–699, 2011.

[91] C. Bahlmann, B. Haasdonk, H. Burkhardt, and A. Freiburg, “Online handwriting recognition with support vector machines-a kernel approach,” *Proc. eighth Int. Work. Front. Handwrit. Recognit.*, pp. 49–54, 2002.

[92] L.-J. LI, H.-Y. SU, and J. CH, “Generalized predictive control with online least

squares support vector machines," *Acta Autom. Sin.*, vol. 33, no. 11, pp. 1182–1188, 2007.

[93] A. C. Megri, I. El Naqa, and F. Haghighat, "A learning machine approach for predicting thermal comfort indices," *Int. J. Vent.*, vol. 3, no. 4, pp. 363–376, 2005.

[94] T. Chaudhuri, D. Zhai, Y. C. Soh, H. Li, and L. Xie, "Thermal comfort prediction using normalized skin temperature in a uniform built environment," *Energy Build.*, vol. 159, pp. 426–440, 2018.

[95] C. Dai, H. Zhang, E. Arens, and Z. Lian, "Machine learning approaches to predict thermal demands using skin temperatures: Steady-state conditions," *Build. Environ.*, vol. 114, pp. 1–10, 2017.

[96] A. Aryal and B. Becerik-gerber, "A comparative study of predicting individual thermal sensation and satisfaction using wrist-worn temperature sensor, thermal camera and ambient temperature sensor," *Build. Environ.*, vol. 160, p. 106223, 2019.

[97] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[98] J. Wainer and P. Fonseca, "How to tune the RBF SVM hyperparameters? An empirical evaluation of 18 search algorithms," *Artif. Intell. Rev.*, vol. 54, no. 6, pp. 4771–4797, 2021.

[99] H. Hensel and K. Schafer, *"Thermoreception and temperature regulation in man." Recent advances in medical thermology.* Springer, Boston, MA, 1984.

[100] T. Chaudhuri, D. Zhai, Y. C. Soh, H. Li, L. Xie, and X. Ou, "Convolutional neural network and kernel methods for occupant thermal state detection using wearable technology," *2018 Int. Jt. Conf. Neural Networks (IJCNN). IEEE*, pp. 1–8, 2018.

[101] T. Chaudhuri, D. Zhai, Y. Chai, H. Li, and L. Xie, “Thermal comfort prediction using normalized skin temperature in a uniform built environment,” *Energy Build.*, vol. 159, pp. 426–440, 2018.

[102] J. Ngarambe, G. Y. Yun, and M. Santamouris, “The use of artificial intelligence (AI) methods in the prediction of thermal comfort in buildings: energy implications of AI-based thermal comfort controls,” *Energy Build.*, vol. 211, p. 109807, 2020.

[103] F. Tartarini and S. Schiavon, “SoftwareX pythermalcomfort: A Python package for thermal comfort research,” *SoftwareX*, vol. 12, p. 100578, 2020.

[104] ASHRAE, “Thermal Environmental Conditions for Human Occupancy, ANSI/ASHRAE Standard 55-2020.” Atlanta, 2020.

[105] A. Ben-Hur and J. Weston, *“A user’s guide to support vector machines.”In Data mining techniques for the life sciences*. Humana Press, pp. 223-239, 2010.

[106] J. Kim, S. Schiavon, and G. Brager, “Personal comfort models – A new paradigm in thermal comfort for occupant-centric environmental control,” *Build. Environ.*, vol. 132, pp. 114–124, 2018.

[107] H. Liu, Y. Wu, B. Li, Y. Cheng, and R. Yao, “Seasonal variation of thermal sensations in residential buildings in the Hot Summer and Cold Winter zone of China,” *Energy Build.*, vol. 140, pp. 9–18, 2017.

[108] H. B. Rijal, H. Yoshida, and N. Umemiya, “Seasonal and regional differences in neutral temperatures in Nepalese traditional vernacular houses,” *Build. Environ.*, vol. 45, no. 12, pp. 2743–2753, 2010.

[109] Y. Peng, Z. Nagy, and A. Schlüter, “Temperature-preference learning with neural networks for occupant-centric building indoor climate controls,” *Build. Environ.*, vol. 154, pp. 296–308, 2019.

[110] V. L. Erickson and A. E. Cerpa, “Thermovote: Participatory sensing for

efficient building HVAC conditioning," *BuildSys 2012 - Proc. 4th ACM Work. Embed. Syst. Energy Effic. Build.*, no. May, pp. 9–16, 2012.

[111] Y. Zhai *et al.*, "Transient human thermophysiological and comfort responses indoors after simulated summer commutes," *Build. Environ.*, vol. 157, pp. 257–267, 2019.

[112] M. Burzo *et al.*, "Multimodal Sensing of Thermal Discomfort for Adaptive Energy Saving in Buildings," *Build. Environ.*, vol. 46, no. 12, pp. 2529–2541, 2004.

[113] J. Choi and V. Loftness, "Investigation of human body skin temperatures as a bio-signal to indicate overall thermal sensations," *Build. Environ.*, vol. 58, pp. 258–269, 2012.

[114] C. Hughes, S. Natarajan, C. Liu, W. J. Chung, and M. Herrera, "Winter thermal comfort and health in the elderly," *Energy Policy*, vol. 134, p. 110954, 2019.

[115] A. A. Farhan, K. Pattipati, B. Wang, and P. Luh, "Predicting individual thermal comfort using machine learning algorithms," *IEEE Int. Conf. Autom. Sci. Eng.*, vol. 2015-Octob, pp. 708–713, 2015.

[116] L. Jiang and R. Yao, "Modelling personal thermal sensations using C-Support Vector Classification (C-SVC) algorithm," *Build. Environ.*, vol. 99, pp. 98–106, 2016.

[117] C. Dai, H. Zhang, E. Arens, and Z. Lian, "Machine learning approaches to predict thermal demands using skin temperatures: Steady-state conditions," *Build. Environ.*, vol. 114, pp. 1–10, 2017.

[118] M. Vellei, W. O. Brien, and S. Martinez, "Some evidence of a time-varying thermal perception," *Indoor Built Environ.*, vol. 31, no. 3, pp. 788–806, 2022.

[119] S. W. Raudenbush and A. S. Bryk, *Hierarchical linear models: Applications and data analysis methods*. SAGE Publications, 2002.

[120] R. F. Rupp, O. B. Kazanci, and J. Toftum, “Investigating current trends in clothing insulation using a global thermal comfort database,” *Energy Build.*, vol. 252, p. 111431, 2021.

[121] S. Thapa and M. Indraganti, “Evaluation of thermal comfort in two neighboring climatic zones in Eastern India — an adaptive approach,” *Energy Build.*, vol. 213, p. 109767, 2020.

[122] S. Thapa, A. Kr, and G. Kr, “Thermal comfort in naturally ventilated office buildings in cold and cloudy climate of Darjeeling, India – An adaptive approach,” *Energy Build.*, vol. 160, pp. 44–60, 2018.

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

[124] D. Li, C. C. Menassa, and V. R. Kamat, “Robust non-intrusive interpretation of occupant thermal comfort in built environments with low-cost networked thermal cameras,” *Appl. Energy*, vol. 251, p. 113336, 2019.

[125] Z. Chen, J. Xin, and P. Liu, “Air quality and thermal comfort analysis of kitchen environment with CFD simulation and experimental calibration,” *Build. Environ.*, vol. 172, p. 106691, 2020.

[126] A. A. Hawila, A. Merabtine, N. Troussier, and R. Bennacer, “Combined use of dynamic building simulation and metamodeling to optimize glass facades for thermal comfort,” *Build. Environ.*, vol. 157, pp. 47–63, 2019.

[127] L. A. Hurtado, J. D. Rhodes, P. H. Nguyen, I. G. Kamphuis, and M. E. Webber, “Quantifying demand flexibility based on structural thermal storage and comfort management of non-residential buildings: A comparison between hot and cold climate zones,” *Appl. Energy*, vol. 195, pp. 1047–1054, 2017.

[128] L. Noda, A. V. P. Lima, J. F. Souza, S. Leder, and L. M. Quirino, “Thermal and visual comfort of schoolchildren in air-conditioned classrooms in hot and

humid climates," *Build. Environ.*, vol. 182, p. 107156, 2020.

[129] B. Gautam, H. B. Rijal, H. Imagawa, G. Kayo, and M. Shukuya, "Investigation on adaptive thermal comfort considering the thermal history of local and migrant peoples living in sub-tropical climate of Nepal," *Build. Environ.*, vol. 185, p. 107237, 2020.

[130] S. Van Craenendonck, L. Lauriks, C. Vuye, and J. Kampen, "Local effects on thermal comfort: Experimental investigation of small-area radiant cooling and low-speed draft caused by improperly retro fitted construction joints," *Build. Environ.*, vol. 147, pp. 188–198, 2019.

[131] X. Su, Z. Wang, Y. Xu, and N. Liu, "Thermal comfort under asymmetric cold radiant environment at different exposure distances," *Build. Environ.*, vol. 178, p. 106961, 2020.

[132] L. Liu *et al.*, "Quantitative effects of urban spatial characteristics on outdoor thermal comfort based on the LCZ scheme," *Build. Environ.*, vol. 143, pp. 443–460, 2018.

[133] K. Rewitz and D. Müller, "Influence of gender, age and BMI on human physiological response and thermal sensation for transient indoor environments with displacement ventilation," *Build. Environ.*, vol. 219, p. 109045, 2022.

[134] H. Zhang *et al.*, "Using footwarmers in offices for thermal comfort and energy savings," *Energy Build.*, vol. 104, pp. 233–243, 2015.

[135] A. Lipczynska, S. Schiavon, and L. T. Graham, "Thermal comfort and self-reported productivity in an office with ceiling fans in the tropics," *Build. Environ.*, vol. 135, pp. 202–212, 2018.

[136] H. Zhao, G. Xu, Y. Shi, J. Li, and Y. Zhang, "The characteristics of dynamic and non-uniform thermal radiation experienced by pedestrians in a street canyon," *Build. Environ.*, vol. 222, p. 109361, 2022.

[137] K. K. Lau and C. Y. Choi, “The influence of perceived aesthetic and acoustic quality on outdoor thermal comfort in urban environment,” *Build. Environ.*, vol. 206, p. 108333, 2021.

[138] R. Qiao, X. Li, S. Gao, and X. Ma, “Improvement of thermal comfort for underground space: Data enhancement using variational autoencoder,” *Build. Environ.*, vol. 207, p. 108457, 2022.

[139] M. Y. Jason and P. Jones, “Evaluation of thermal comfort in building transitional spaces-Field studies in Cardiff, UK,” *Build. Environ.*, vol. 156, pp. 191–202, 2019.

[140] M. Khoshbakht, Z. Gou, and F. Zhang, “A pilot study of thermal comfort in subtropical mixed-mode higher education office buildings with different change-over control strategies,” *Energy Build.*, vol. 196, pp. 194–205, 2019.

[141] X. Ma, Y. Tian, M. Du, B. Hong, and B. Lin, “How to design comfortable open spaces for the elderly ? Implications of their thermal perceptions in an urban park,” *Sci. Total Environ.*, vol. 768, p. 144985, 2021.