Prediction of occupant thermal state via infrared thermography and explainable AI

https://doi.org/10.1016/j.enbuild.2024.114153

Journal of Energy and Building

Prediction of occupant thermal state via infrared thermography and explainable AI

Shaoxing Zhang [a], Runming Yao [a, b, *], Hong Wei [c], 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, Reading, UK

[c] Department of Computer Science, University of Reading, Reading, UK

Abstract

Accurate and real-time assessment of occupant thermal comfort can provide a solid foundation for efficient air conditioning operations. Existing studies already show the feasibility of using contactless technologies for thermal comfort prediction assisted by machine learning algorithms. However, the lack of transparency in machine learning often weakens user trust. This study performs explainable AI analysis to explore the potential of infrared imaging in thermal comfort evaluation. Specifically, an investigation was carried out in a climatic chamber, and infrared cameras were used to collect facial temperature data. Five popular ensemble tree models were employed to construct prediction models, and explainable AI analysis was performed using SHAP (SHapley Additive exPlanations) theory. Results show that combining additional facial information can significantly improve the overall model performance, and certain facial attributes present high contributions based on SHAP values. Combining facial features with explainable AI provides a convincing basis for thermal comfort assessment. The high SHAP values of facial features can also contribute to finding selective occupants with low neutral voting rates, providing evidence for customized cooling or heating from building systems.

Keywords: Thermal comfort, XGBoost, explainable AI, facial thermography

Abbreviations

AdaBoost	Adaptive Boosting
BP	Blood Pressure
DBP	Diastolic Blood Pressure
Env	Features of environmental parameters
GBDT	Gradient Boosting Decision Trees
HVAC	Heating Ventilation and Air Conditioning
LightGBM	Light Gradient Boosting Machine
LIME	Local Interpretable Model-agnostic Explanations
RF	Random Forest
RH	Relative humidity
PMV	Predicted Mean Vote
SBP	Systolic Blood Pressure
SET	Standard Effective Temperature
SHAP	SHapley Additive exPlanations
TSV	Thermal Sensation Vote
T_a	Air temperature
T_{ankle}	Ankle temperature
T_{ear}	Ear temperature
$T_{facials}$	Features of temperature variations from forehead, inner canthus, nose, and nasolabial fold to the cheek temperature
T_g	Globe temperature
T_{wrist}	Wrist temperature
XGBoost	eXtreme Gradient Boosting

1. Introduction

Comfortable and pleasurable indoor thermal environments can benefit occupants in terms of comfort, health, well-being, and productivity. In unfavorable outdoor climates, heating, ventilation, and air conditioning (HVAC) systems are commonly utilized to improve indoor thermal conditions, resulting in significant energy usage. According to the European Commission, heating and cooling energy in buildings and industries account for 50% of the EU's annual energy consumption [1]. Currently, most international (ISO-7730 [2], EN 16798 [3]) and national standards (ASHRAE-55 [4],

CIBSE Guide-A [5], GB/T 50785 [6]) stipulate the classic Predicted Mean Vote (PMV) method as the primary approach for evaluating indoor thermal comfort under HVAC operations. The PMV index effectively combines environmental parameters (such as air temperature, radiant temperature, relative humidity, and air velocity) with subjective factors (including metabolic rate and clothing level) to assess thermal comfort. This approach has demonstrated satisfactory performance in a range of building types [7], such as school buildings [8], mosques and churches [9], hospitals [10], residential apartments [11], office buildings [12], etc.

However, certain dynamic and statistic factors that can potentially affect thermal comfort were not considered in the evaluation process in current PMV approach, such as age, gender, race, acclimation, prior thermal exposure, and food/drink intake [13]. The collections of certain parameters are also expensive and challenging, such as radiant temperature, air velocity, and metabolic rate [14]. Furthermore, because of factors such as device-user distance or user movement [15], the arrangement of environmental measuring devices typically depicts specific measurement areas, frequently failing to capture the actual surroundings near each occupant. Increasing the number of environmental devices to improve the representativeness of environmental parameters will increase data collection costs even further, and related measurement errors and miscalculations of PMV can also result in greater energy consumption for maintaining the indoor thermal environment [16]. To address these challenges, researchers have created thermal comfort models based on real-time physiological parameters, such as occupants' skin temperature of back hand [17], wrist temperature [18], heart rate variability [19], etc.

The majority of these physiological-based solutions necessitate the deployment of additional devices in touch with occupants' skin and the collection of subjective feedback in real-time [20]. These can interfere with users' daily work, perhaps creating the Hawthorne effect [21], in which people change their behaviors to fit the expectations of observers when they feel watched. Furthermore, long-term contact with monitoring devices may not be appropriate for certain vulnerable or special populations, such as infants, burned patients, etc. Therefore, some thermal comfort studies have turned their

attention to non-intrusive methods of monitoring occupant thermal states using images and videos. Infrared thermography is one such technology, which uses infrared cameras to collect skin temperatures from exposed parts of the human body (face [15], hands [22], etc.) or clothing temperature. This contactless method enables the creation of personalized predictive models for each individual while avoiding obvious interference. Many of these studies have utilized machine learning algorithms for model training and prediction.

A major challenge to the widespread adoption of machine learning is its lack of interpretability. Even though extensive research and evidence demonstrate its superior performance, the perception of a model as a full black box with little or no human intermediation may create worries about its **trust** in real-world applications. The Recital 71 from European Union's new General Data Protection Regulation (GDPR) emphasizes the "*right to explanation*" of data subject during the algorithmic decision-making process, which should include the right to "*obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision*" [23] instead of blindly accepting black box models, but making these black box models transparent faces several barriers [24]:

- **Intentional concealment**: corporations and institutions deliberately keep decision-making processes hidden from public scrutiny.
- **Technical literacy gaps**: access to underlying code alone is often insufficient due to limitations in the technical proficiency of general public.
- **Human cognitive limitations**: a mismatch exists between the mathematical optimization in high-dimensional machine learning and the demands of human-scale reasoning and interpretive styles.

The first two issues involve public concerns, making it more challenging for researchers to improve due to their extensive societal, cultural, educational, and awareness-related characteristics. The third issue is technological in nature and could potentially be tackled through advancements in machine learning algorithms, building decision processes that are more explainable and align better with human reasoning modes, etc.

1.1 Related work

Several recent studies employed machine learning algorithms to predict thermal comfort based on facial physiological parameters. Ghahramani et al. [25] used Hidden Markov Models based on infrared facial imaging to predict thermal discomfort with an accuracy of 82.8%. Cosma and Simha [26] analyzed facial and clothing temperatures using principal component analysis (PCA) and found that clothing temperature also plays a positive role in thermal comfort assessment. He et al. [27] studied cheek, nose, and hand temperatures using Random Forest and observed that with an increasing number of input features, the prediction accuracy improved from 83% to 96%. Aryal and Becerik-Gerber [28] investigated the impact of features on model performance using Random Forest, Support Vector Machines, and K-Nearest Neighbor, and discovered that using facial features as additional input for model training will increase prediction accuracy by 3-4%. These studies highlight the potential of thermal imaging for capturing facial features as well as the effectiveness of machine learning in establishing thermal comfort models.

Many research endeavors have also been dedicated to making complex machine learning models interpretable based on explainable AI approaches, with encouraging results [29]. One of these approaches is the Local Explanation Method, which tries to explain the decision-making processes of complicated machine learning models by constructing a simplified model (often linear) for specific prediction instances and offering effective interpretation for the local context. LIME (Local Interpretable Model-agnostic Explanations) [30] and SHAP (SHapley Additive exPlanations) [31] are the two most popular Local Explanation Methods and SHAP can be considered an improved version of LIME to some extent. Several researchers already employed SHAP to study thermal comfort from diverse perspectives. Qiao et al. [32] used SHAP to assess the gender impact on thermal comfort in underground public transportation. Their findings revealed that women tend to be more sensitive to low temperatures (below 24°C), whereas men exhibit greater sensitivity to high temperatures (above 29°C). Lan et al. [33] employed SHAP to evaluate individual differences in thermal comfort among classroom students. According to their findings, overweight and obese

students preferred cooler temperatures. Yang et al. [34] utilized SHAP to investigate the interpretability of a public thermal comfort database. They observed thresholds at which particular feature contributions abruptly shift, implying that the true neutral environment may be a dynamic high-dimensional space formed of specific combinations of features in certain ranges with changing forms, rather than just a concept of temperature boundaries. Baek et al. [35] investigated infrared thermography of seated human subjects (wearing short sleeves). They developed deep convolutional neural networks (CNN) to forecast thermal comfort and visualized SHAP values at pixel levels throughout the CNN prediction. Their visualizations revealed the effects of exposed skin temperature and clothing temperature on predictions, emphasizing the importance of clothing temperature in predicting thermal perception. These SHAP-based studies demonstrated the effectiveness of explainable AI on thermal comfort data and research.

1.2 Existing research gaps

Despite advancements in predicting thermal comfort through infrared thermography and explainable AI, critical research gaps persist, and a coherent connection between these two domains is yet to be fully established. Many infrared thermography studies have successfully constructed high-performing thermal comfort models using physiological parameters and machine learning algorithms, but the inner mechanisms of these black box models remain inadequately explained. Meanwhile, the majority of explainable AI-based thermal comfort studies, although valuable in their approaches, have primarily incorporated non-physiological factors like age, gender, and BMI as supplementary inputs for model development. This falls short of fully capturing the complex dynamics of real-time variations in occupants' physiological responses in practice, which play an important role in representing occupants' thermal states in real-time. These gaps indicate the potential for the strengths of these two approaches to compensate for the weaknesses of each other. By focusing on a more comprehensive understanding of occupants' physiological parameters and their immediate interactions with indoor conditions, there arises an opportunity to advance the field through a novel

approach that integrates infrared thermography and explainable AI. The combination of these two approaches allows us to not only construct accurate prediction models in a contactless way, but also to investigate the fundamental mechanisms of machine learning algorithms on how physiological and environmental variables impact thermal comfort.

1.3 The objectives of this study

Given these considerations, this study aims to develop a contactless infrared thermography method to predict thermal comfort based on machine learning algorithms using the explainable AI to explicate the model training and predicting processes, as illustrated in Fig. 1. During the climatic chamber studies, environmental and physiological parameters have been collected together by data loggers, infrared cameras and other instruments, while only two thermocouples were attached to the individuals' wrist and ankle to minimize interference with the subjects. In order to find the best machine learning model, Random Forest and four different types of boosting trees were evaluated using all features, and the best-performing model was chosen for further investigation on feature selections. The SHAP model, which has been widely used in the thermal comfort and medical domains [36][37], was chosen for the explainable AI analysis to investigate the inner structure of complex machine learning models. In addition, the effects of different TSV mapping scenarios were examined at the end.

The main contributions of this paper include the following three aspects:

(1) Ensemble tree algorithms were employed to evaluate model performance by considering various combinations of input features, including environmental conditions, facial physiological parameters, and other non-facial physiological data. The distinctive role of facial thermography in model training was clarified.

(2) Explainable AI approach was used to quantify the contribution degree of each single sample and its cumulative effect on each feature. We particularly focused on demonstrating the contributions of facial thermography to predicting hot and cold

thermal sensations.

(3) The thresholds of facial temperature variations were identified to indicate thermal discomfort of both the general population and individuals. It validated the potential of explainable AI in addressing individualized thermal comfort prediction, shedding light on personalized heating or cooling strategies within building systems.

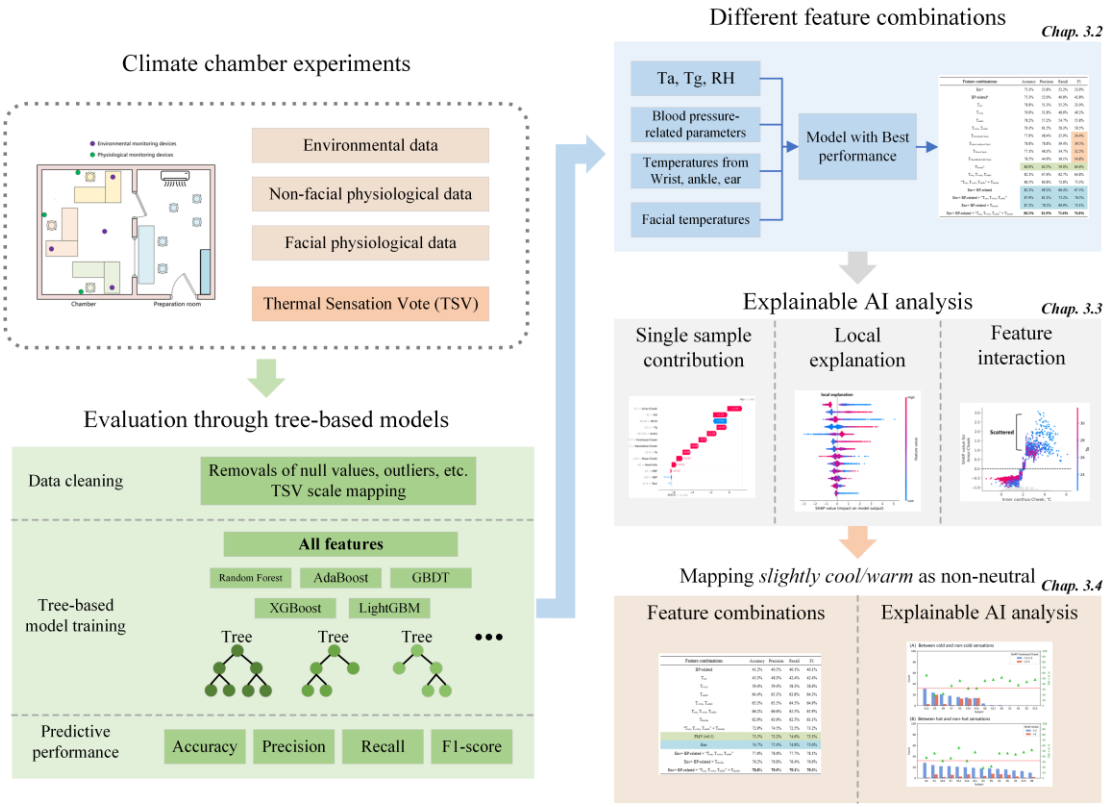

Fig. 1 Schematic view of this study

2. Method

2.1 Experimental settings

2.1.1 Surveys in climate chamber

A two-week series of subject tests were conducted in a climate chamber at Chongqing University in June, 2021. The chamber was operated from 8:00 to 18:30 and each test lasted for 90 minutes. To investigate subjects' thermal responses in moderately cold and hot environments, the indoor conditions were controlled in five scenarios ranging from 22-32°C: (1) constant 22°C; constant 26°C; (3) constant 32°C; (4) increasing from 22°C to 32°C; and (5) decreasing from 32°C to 22. Before the test, all subjects were informed

to wear typical summer clothes at the typical ensembles of 0.36 clo suggested by ASHRAE-55 and not to smoke, drink liquor, or sleep late. During the test, all subjects firstly stayed in the preparation room for 30 minutes where the temperature was kept the same with the initial temperature setting in the 90-minute chamber experiment to avoid the feelings of sudden thermal stimuli. It is important to note that no thermal comfort questionnaires were delivered during this 30-minute preparation period. Instead, our analysis focused on the data collected during the 90-minute experimental session conducted within the climate chamber. In the chamber, the subjects remained sitting posture and carried out light office activities, such as reading or typing (1 to 1.1 met). The questionnaires, facial thermography, blood pressure, and other physiological measurements were performed every 10 minutes (Fig. 2).

Fifteen master students voluntarily joined the experiments with monetary compensation and contributed to the generation of approximately 2,000 thermal response questionnaires. However, due to one participant discontinuing his involvement, his data were excluded from the analysis. Subsequently, we removed null data and outliers based on the boxplot method to ensure data quality [38]. This refinement process yielded a final dataset of 1,697 data points with valid questionnaires and environmental/physiological records from fourteen people (9 males and 5 females), which were used to build machine learning models for predicting thermal comfort and model interpretation.

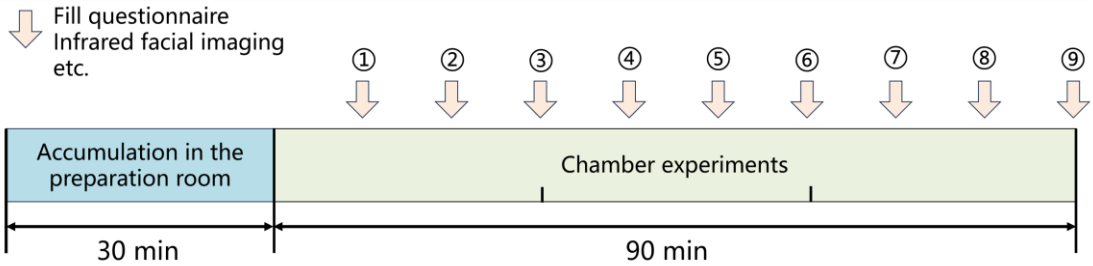

Fig. 2 Testing procedure of the experiment (①-⑨ represent the surveyed moments when occupants are required to complete questionnaires and take physiological measurements/thermography)

2.1.2 Measurements

Table 1 shows the technical specifications of the instruments used in this study. In

general, three types of measurements were monitored and collected during the experiment (Fig. 3):

- Environmental conditions: air temperature and relative humidity were monitored by HOBO devices near each subject. Globe temperature was collected every by a black bulb thermometer placed in the middle of the chamber.
- Non- Facial physiological parameters: wrist temperature and ankle temperature were collected by a HOBO 4-channel thermocouple logger with two thermocouples attached to the skin of each subject's wrist and ankle. Omron blood pressure monitors were used to collect SBP (systolic blood pressure), DBP (diastolic blood pressure), and Heartrate (pulse). The ear temperature was also measured by an ear thermometer.
- Facial physiological parameters: a FLIR thermal camera was used to capture the facial thermal response of forehead, inner canthus, cheek, nose, and nasolabial fold, as shown in Fig. 4. For the sensitivity of thermal camera, temperature drifting is considered to be a common problem, which can occur after device calibration and cause all points on the thermal image to increase or decrease [39]. Following device calibrations, we shot two thermal images of the same subject at the same moment. Temperature drifting of general values reached around 4°C, but their inner fluctuation between each point changed only up to 0.2°C, as shown in Table 2. Therefore, instead of using the absolute value measured by the thermal camera, this study used the variations between pixel points to indicate facial thermal information.

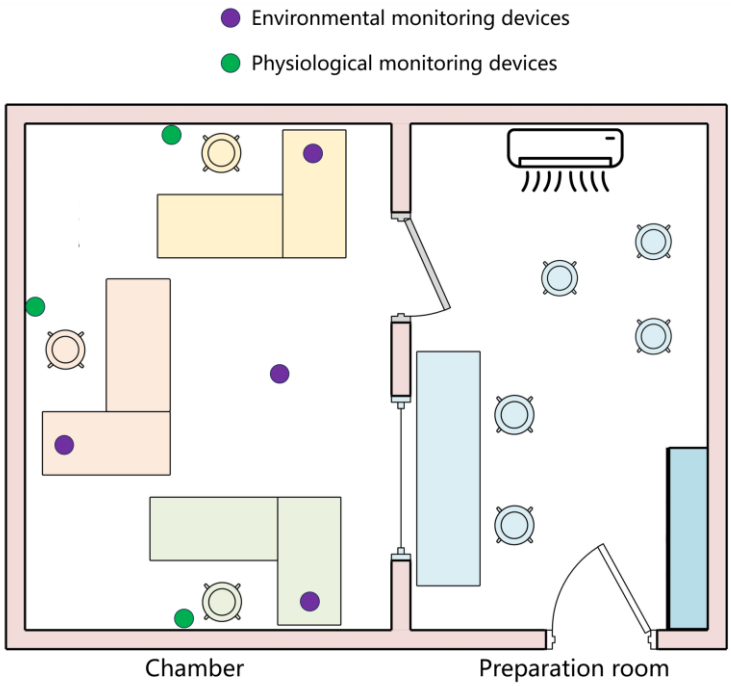

Fig. 3 Layout of the climate chamber.

Table 1. Technical specifications of measuring instruments

Model	Manufacturer	Measuring parameters	Measuring frequency	Range	Accuracy
HOBO UX100-011	Onset	Air temperature	1 second	-20~70℃	±0.21℃ (0~50℃)
		Relative humidity	1 second	1~95%	±2.5% (10~90%)
JTR04	Beijing JT Technology	Globe temperature	10 minutes	10~80℃	±0.2℃ (20~40℃)
HOBO 4-channel thermocouple logger	Onset	Wrist temperature Ankle temperature	1 second	-20~70℃	±0.21℃
HEM-7012	OMRON	SBP (systolic blood pressure) DBP (diastolic blood pressure)	10 minutes	0~299 mmHg	±4 mmHg
		Heartrate (pulse)	10 minutes	40~180 bpm	5%
YHT200 ear thermometer	Yuwell	Ear temperature	10 minutes	34~42.2℃	±0.2℃ (35~42℃) ±0.3℃ (beyond 35~42℃)
FLIR E6-XT thermal camara	FLIR Systems, Inc	Facial temperatures	10 minutes	-20~550℃, 240 × 180 pixels	±2℃ or ±2% of reading,

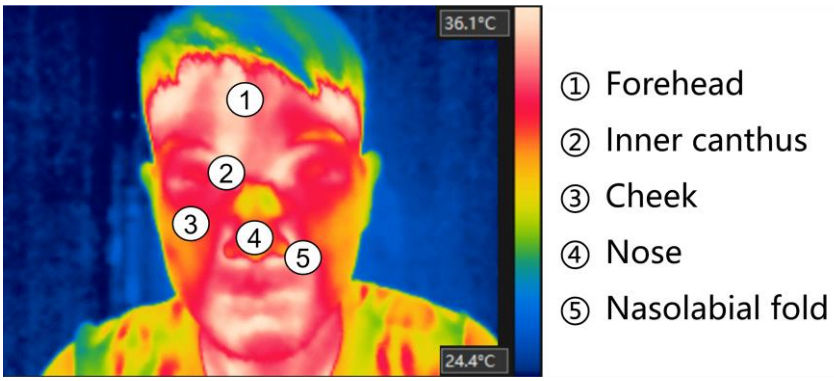

Fig. 4 Extracted temperature points from facial tomography

Table 2. Temperature drifting of one subject at the same moment after device calibration

	Forehead (℃)	Inner canthus (℃)	Cheek (℃)	Nose (℃)	Nasolabial fold (℃)
Record 1	34.2	34.8	34.2	34.1	34.1
Record 2	38.1	38.6	38.3	38.0	38.2
Deviation	-3.9	-3.8	-4.1	-3.9	-4.1

2.1.3 Subjective feedback

Throughout the experiment, each subject was required to provide thermal feedback every 10 minutes. The thermal sensation vote (TSV) was primarily used to assess subjects' thermal responses. The scale ranged from -3 to +3 based on ASHRAE 55-2020 [4]: -3 (cold), -2 (cool), -1 (slightly cool), 0 (neutral), 1 (slightly warm), 2 (warm), and 3 (hot). For the specific model training and establishment in this study, votes of -1, 0, and +1 were combined as *comfort*, while -3 and -2 were considered *cold* and +2 and +3 as *hot*. Therefore, the 7-scale classification problems were reduced to 3-scale problems.

2.2 Ensemble tree models

In recent years, deep learning models have achieved remarkable success in handling complex and unstructured data in various domains [40], including image recognition [41], recommender systems [42], natural language processing [43], etc. On the other hand, tree-based models can consistently outperform typical deep learning models when the data is individually meaningful and lacks strong multi-scale temporal or spatial features [44]. Both deep learning [45] and tree-based approaches [46] made significant contributions to the field of thermal comfort research, allowing for the

development of highly precise models as well as improved comprehension of underlying patterns in data. In this paper, our sample size of 1,697 may not be sufficient for building a deep learning model. Consequently, we used tree-based methods to construct machine learning models for predicting thermal comfort. For classic tree classifier, it often faces the challenges of overfitting when trees are too complex and lacking generalization to unseen data [46]. Therefore, several ensemble approaches were developed to improve the predictive performance and robustness of tree-based models, such as bootstrap sampling (Random Forest) [46], weak learners boosting (AdaBoost) [47], residual minimizing in each interaction (Gradient Boosting Decision Trees) [48], etc. According to the characteristics of tabular-style data in this study, five popular ensemble tree algorithms have been employed to construct machine learning models for predicting thermal comfort, as shown in Table 3.

In this paper, the raw data have been cleaned by removing null values and outliers using the Boxplot rule [38]. Because each ensemble tree model has distinct characteristics and appropriate hyperparameter tuning can improve model performance [49], the grid search method with 5-folder cross validation was used to identify the best parameter combinations. The collected data were divided into training and testing subsets in an 8:2 ratio. In order to ensure reproducibility and consistency in the data splitting process, we specified the parameter "*random_state*" as 42 when utilizing the "*train_test_split*" function in Python. This setting allowed us to reproduce the same splitting results consistently across multiple runs of the code. During the training processes, the *label encoding* method was used to convert text data into numeric data, because this method was found to be an effective way to process thermal comfort data [50]. No data normalization or standardization (scaling to [0,1] range) was performed during the pre-processing procedure as tree-based models are known to be robust to feature scaling [51]. All environmental and physiological parameters were used as inputs for model training, while the 3-scale TSV was the output. To comprehensively evaluate ensemble tree models, four evaluation metrics "*accuracy, precision, recall, and F1 score*" were used for assessing the classification problems, because relying solely on accuracy will lead to *accuracy cheating* [52], especially when the dataset is imbalanced.

Table 3. Features of popular ensemble tree models and applications in thermal comfort studies

Model	Year	Main feature	Key hyperparameters	Strengths	Weaknesses	Applications in thermal comfort studies
Random Forest (RF) [46]	1995	Use random feature selection and bootstrap sampling to construct each decision tree.	Number of estimators Max depth Min samples leaf Criterion: gini, entropy	Good at handling high-dimensional data, outliers, and missing values.	Overfit when dealing with noisy data and highly correlated features.	Gender difference based on wearable sensing (over 90% accuracies) [53] Thermal pleasure based on cutaneous thermoreceptor activity (83% accuracy and 67% F1 score) [54] Thermal state based on infrared thermography (83-96% accuracies) [27]
AdaBoost [47]	1997	Focus on wrong classification and boost the weak learner (tree).	Number of estimators Learning rate	Good accuracy and generalizability on complex classification problems.	Overfit when dealing with noisy data and outliers.	Individual preference based on skin temperature and heating behaviors (84% accuracy) [55] Outdoor thermal comfort based on UTCI index and bike ridership data (75% acceptable predictions) [56] Thermal comfort prediction based on heart rate variability (93.7% accuracy) [57]
Gradient Boosting Decision Trees (GBDT) [48]	2001	Iterate decision trees based on residuals in each round.	Number of estimators Learning rate Max depth Min samples leaf	Good generalizability on large datasets.	Slow training process and high requirements of computing resources.	Air conditioner usage in residential buildings (89.5% accuracy) [58] Impacts of climate change on thermal comfort (72% and 91% accuracies) [59] Gender differences in underground public Transportation (29% and 35% increased accuracies) [32]
eXtreme Gradient Boosting (XGBoost) [60]	2016	Based on GBDT. Embed parallelization, regularization, and greedy algorithm to optimize the training process.	Max depth Subsample Min child weight Gamma	Good accuracy and efficiency on large datasets and complex features.	Sensitive to noise and outliers.	Thermal comfort prediction based on local skin temperatures (72.5% and 78.3% accuracies) [61] Outdoor thermal comfort based on optimized tree algorithms (95.21% accuracy) [59]

Light Gradient Boosting Machine (LightGBM) [62]	2017	Based on GBDT. Exclude data with small gradients and bundle mutually exclusive features to reduce training time.	Number of estimators Learning rate Max depth Subsample	Fast training efficiency on large datasets.	Worse performance on small datasets compared with GBDT.	Individual difference in classrooms (over 88% accuracies) [34] Passengers in high-speed Railway based on electroencephalography (0.1704 RMSE and 0.1261 MAE) [63] Rapid establishment of prediction models (89.3% average F1 score) [64] Cooling load prediction in a commercial building (95.94% accuracy) [65]

2.3 Model interpretation

Machine learning models are becoming increasingly widespread because they can achieve superior performance and even surpass human capacity in many applications, such as the game of GO [66], language translation [67], and protein folding [68]. However, their inner mechanisms remain "*black boxes*", and one critical concern is the *trust* in the reasoning behind their predictions: *if the users do not trust a model or a prediction, they will not use it* [30]. Among all the effective approaches to explain machine learning models, *local feature attribution* is considered a prominent approach. It helps to understand individual predictions by assigning attribution scores to each feature, thereby providing insights into the model's decision-making process and feature importance [69]. Within this methodology, Lundberg and Lee [31] introduced SHAP (SHapley Additive exPlanations) as a powerful tool to interpret the predictions of machine learning models. At its core, SHAP was built upon the Shapley value [70], a concept with a long history in game theory for assigning contributions of players in cooperative games. SHAP adopts the idea of examining different orders of adding inputs to determine the attribution scores for each feature.

According to the Shapley value, φ_i is the local importance of feature *i* [31]:

$$\varphi_i = \sum_{S \subseteq F \backslash \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}\left(x_{S \cup \{i\}}\right) - f_S(x_S) \right] \qquad (1)$$

Where |S| is the size of the subset before adding the feature *i*, |F| is the number of features, $S \subseteq F \backslash \{i\}$ is all possible subsets without the feature *i*, $x_{S \cup \{i\}}$ is the subset S

with feature i added, and S is the subset without feature i. The second part of the equation (1) $[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)]$ represents the marginal contribution, which captures the incremental contribution of a specific player (feature) in the overall game (model). Whereas the first part of equation (1) $\frac{|S|!(|F|-|S|-1)!}{|F|!}$ is the weight for combinations of this occurrence.

The calculation of the original Shapley value is an NP-hard (nondeterministic polynomial time) subset sum problem, because it needs to consider the combinations and permutations of all possible subsets. For N features, there are 2^N different subsets to be considered, while the number of permutations for each subset is N!. As a result, the overall computational complexity of Shapley value will be O(2^N*N!) and it grows exponentially with the number of features, making it extremely challenging to directly solve in high-dimensional feature spaces. The SHAP method solved this by developing *additive feature attribution methods* based on the idea of *local methods* designed in LIME (Local Interpretable Model-agnostic Explanations) that can create a local approximation of the complex model for a specific input [30]:

$$\xi = \underset{g\epsilon G}{\operatorname{argmin}}\, L(f, g, \pi_x) + \Omega(g) \tag{2}$$

Where ξ is the objective function in LIME, g is the simplified interpretation model (mostly linear) to the original complex model f, G is the family of g, π_x is the proximity that measures locality around input x, $\Omega(g)$ is the complexity penalty for g. It means that LIME tries to find a simple model g that minimizes the two lost terms L and Ω, while L approximates the complex model in the local area and Ω ensures the simplicity of g.

Lundberg and Lee [31] developed the Shapley kernel to identify specific forms of terms in equation (2) that are consistent with the three key properties of the Shapley value in game theory: **1) Local accuracy**: the explainable model produces roughly the same output of the actual model in the local area ($g \approx f$); **2) Missingness**: if one feature is excluded from the model, its attribution is zero ($\varphi = 0$); and **3) Consistency**: if the contribution of a particular feature changes, the attribution in the explanatory model can not change in the opposite direction.

However, SHAP kernel suffers from ignoring feature dependence and correlation, which can lead to biased interpretations [71]. Lundberg et al. [44] later presented TreeExplainer, a popular and improved variant of SHAP kernel for interpreting tree-based models, to address these limitations by explicitly modeling conditional expectation predictions, effectively accounting for feature correlations. It computes exact Shapley values for tree models efficiently by collapsing calculations specific to each leaf in the tree. TreeExplainer also introduces the SHAP interaction value, capturing local interaction effects between features and enhancing model understanding. By incorporating interventional expectations and path coverage information, TreeExplainer enables robust interpretations with correlated features [44], and it reduces the original NP-hard exponential complexity of Shapley value to a manageable O(*TLD*^2) complexity, where *T* is the number of trees, *L* is the number of leaves, and *D* is the maximum depth of any tree. Compared with the other two popular Explainable AI methods LIME [30] and DeepLIFT [72], the SHAP method was proved to achieve higher performance and be more consistent with human intuitions on classification problems [31].

In this paper, thermal sensation votes were collected using the ASHRAE 7-scale, ranging from cold (-3) to hot (+3). Two binary-classification models were trained to interpret tree-based models based on the function *TreeExplainer* in package SHAP [44], as summarized in Table 4. The original SHAP value for binary problems generates a probability ranging from 0 to 1, where positive contributions push the value towards a probability of 1 and negative contributions push it towards 0. However, directly using the probability output as feature contributions will result in undercounting small negative contributions (close to zero). To address this, SHAP introduces the log odds function, which is logit(p)=log(p/(1-p)) and it maps the probability range [0,1] to a symmetric range (-∞, +∞). This transformation ensures a fair and balanced account of contributions to both positive and negative sides.

Table 4. thermal sensation scales in ASHRAE-55 and classification scales in this paper

Criteria for categorizing sensations	Cold	Cool	Slightly cool	Neutral	Slightly warm	Warm	Hot
ASHRAE 55 7-scale TSV	-3	-2	-1	0	1	2	3

Relaxed neutral conditions	3-scale TSV	Cold	Cold	Neutral	Neutral	Neutral	Hot	Hot
	Cold and non-cold sensations	1	1	0	0	0	0	0
	Hot and non-hot sensations	0	0	0	0	0	1	1
Stringent neutral conditions	3-scale TSV	Cold	Cold	Cold	Neutral	Hot	Hot	Hot
	Cold and non-cold sensations	1	1	1	0	0	0	0
	Hot and non-hot sensations	0	0	0	0	1	1	1

3. Results

3.1 Data overview

The data collection took 2 weeks and each test included a 30-minute acclimation stage and a 90-minute experiment stage. A total of 1697 valid sensation votes were used to build ensemble tree models after removing null values and outliers. Tables 5 and 6 present statistical data from environmental and physiological measurements, respectively. The measured air temperatures were well controlled, corresponding to chamber settings of 22℃, 26℃, and 32℃: three constant conditions with mean values of 22.29℃, 25.90℃, and 31.78℃ were observed, while uniform temperature distributions were observed under two transient conditions from 22℃ to 32℃ with mean air temperature around 27℃ (Fig. 5). The globe temperature presents similar trend with air temperature. For three blood pressure related parameters, heartrate increases in hotter environments, while SBP (systolic blood pressure) and DBP (diastolic blood pressure) decrease.

Table 5. Statistical information of environmental and non-facial physiological measurements

Chamber settings	Ta (℃)	Tg (℃)	RH (%)	SBP (mmHg)	DBP (mmHg)	Heartrate (bpm)	T_ear (℃)	T_wrist (℃)	T_ankle (℃)
22℃	22.19±0.17	22.42±0.6	66.03±4.48	107.31±12.06	68.74±10.26	72.18±8.68	36.97±0.27	30.46±1.53	27.16±1.65
26℃	25.90±0.14	25.99±0.10	59.24±2.24	103.85±13.49	65.02±10.59	74.88±10.42	37.10±0.26	32.60±0.78	29.68±1.36
32℃	31.78±0.11	31.59±0.56	58.08±0.70	99.29±12.27	60.19±8.64	80.07±10.63	37.38±0.24	34.59±0.58	33.32±0.51
22-32℃	27.14±2.72	26.74±2.47	58.60±2.96	100.95±10.46	62.88±7.84	75.44±10.30	37.15±0.23	32.36±1.91	30.01±1.79
32-22℃	27.22±2.53	27.53±2.33	60.53±3.30	102.01±10.02	62.86±7.66	76.87±10.97	37.20±0.23	32.79±1.31	30.55±1.79
Total	26.93±3.27	26.93±3.11	60.23±3.97	102.34±11.57	63.62±9.06	75.95±10.57	37.16±0.27	32.56±1.81	30.17±2.31

Table 6. Statistical information of facial physiological measurements and TSV

Chamber settings	Forehead-Cheek (℃)	Inner-Cheek (℃)	Nose-Cheek (℃)	Nasolabial-Cheek (℃)	TSV	Sum
22℃	2.39±1.25	3.07±1.29	-0.19±2.34	2.40±1.09	-1.24±0.69	245
26℃	1.39±0.81	1.85±0.77	0.70±1.19	1.61±0.77	-0.21±0.64	243

32℃	0.16±0.51	0.71±0.7	0.45±0.99	0.24±0.60	1.87±0.83	243
22-32℃	1.42±0.95	1.89±1.06	0.95±1.37	1.65±0.99	0.26±0.92	489
32-22℃	1.07±0.85	1.64±0.94	0.51±1.33	1.35±0.85	0.20±1.01	477
Total	1.28±1.09	1.82±1.17	0.55±1.51	1.46±1.07	0.19±1.22	1697

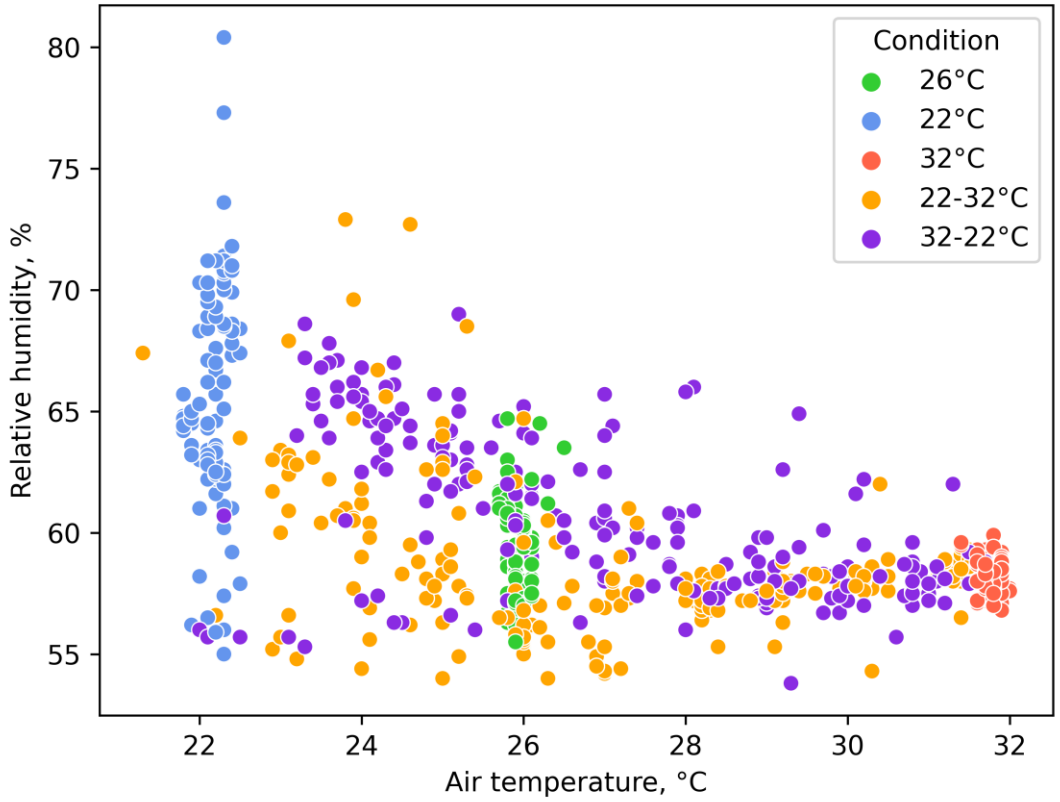

Fig. 5 Distributions of air temperature and relative humidity under five experimental scenarios

Fig. 6 illustrates the distributions of ear, wrist, and ankle temperatures measured by an ear thermometer and attached thermocouples, and classified into three different thermal sensation categories. From cold to hot sensations, the ear temperatures remain relatively stable, fluctuating around ±0.2℃. In contrast, wrist and ankle temperatures show a significant increase, with ankle temperatures consistently lower than wrist temperatures

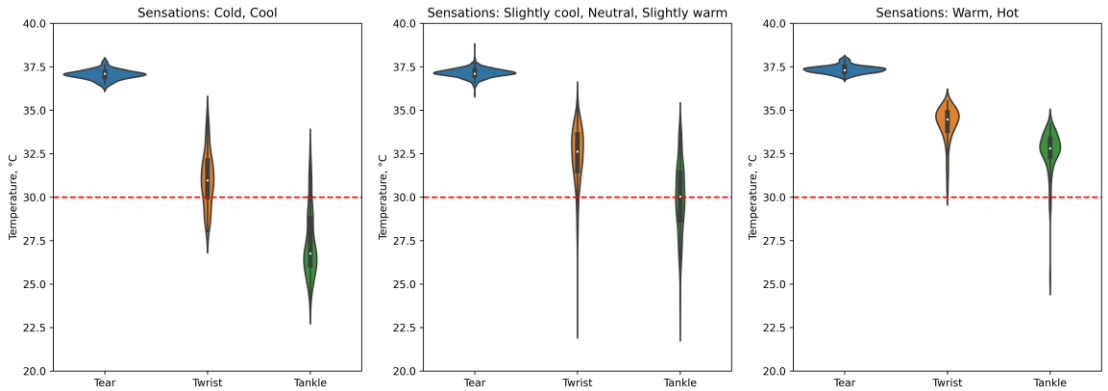

Fig. 6 Distributions of ear, wrist, and ankle temperatures under different thermal sensations

To avoid the influence of temperature drifting caused by infrared camera calibration, this study utilized temperature variations between measure pixels to indicate facial thermal information. The cheek temperature was chosen as the baseline due to its generally lower values. Fig. 7 depicts the facial temperature variations. For hot sensations, the temperature differences between measured points and cheek are close to zero (red dotted line), indicating a relatively uniform distribution of facial temperature. Under neutral sensations, many of the four measured points show increasing temperature differences with the cheek. However, during cold sensations, all measured points show significant differences except for the nose, suggesting obviously cooling in both nose and cheek regions, while the other three areas maintain relatively higher temperatures, especially the inner canthus (orange area). Fig. 8 illustrates the linear relationships between the four facial variations at different air temperatures. The forehead, inner canthus, and nasolabial fold temperatures exhibit similar negative gradients with slower rates of temperature decrease compared to the baseline temperature of the cheek, as opposed to the nose (which shows a positive gradient). As the temperatures decrease from 32°C to 22°C, some nose temperatures show values even 6°C lower than the cheek temperature, resulting in a higher overall decreasing rate of temperature than the cheek.

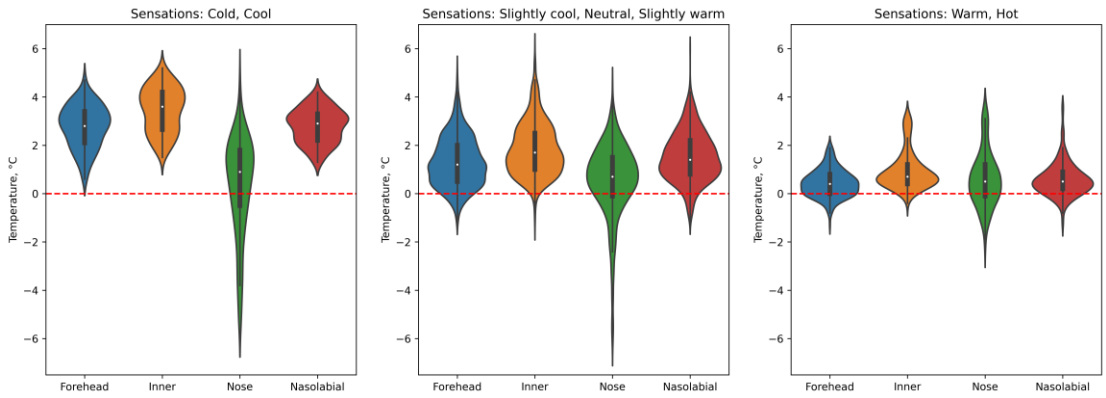

Fig. 7 Temperature variations of forehead, inner canthus, nose, and nasolabial fold from check

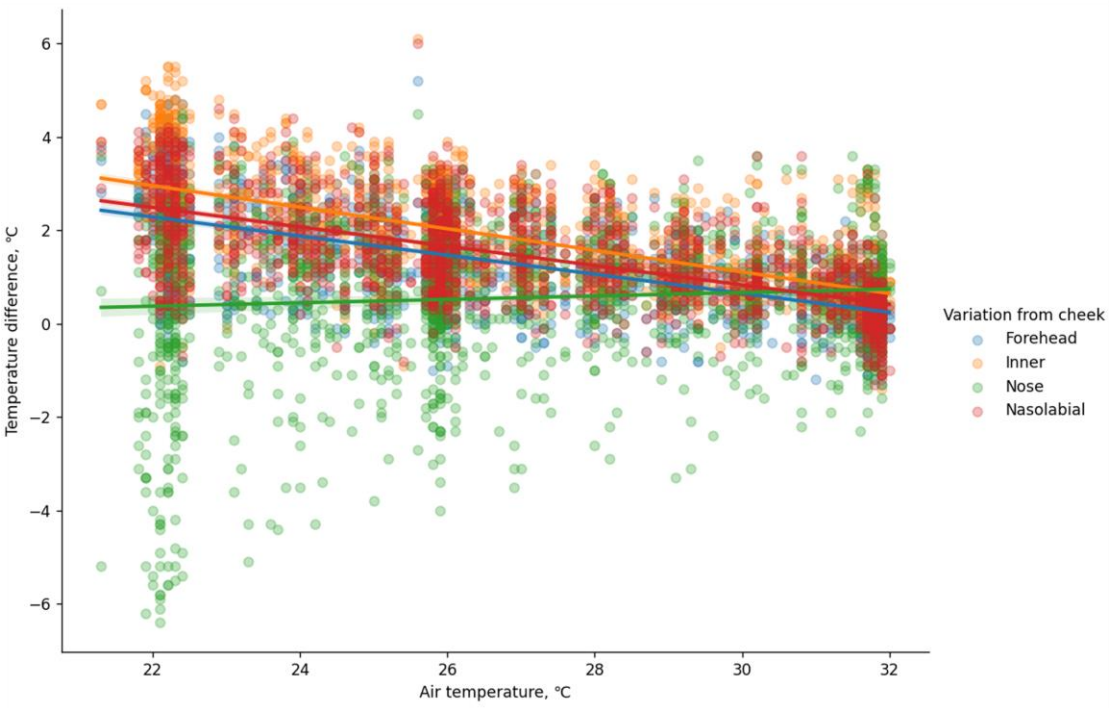

Fig. 8 Distributions and linear regressions of facial temperature variations under different air temperatures

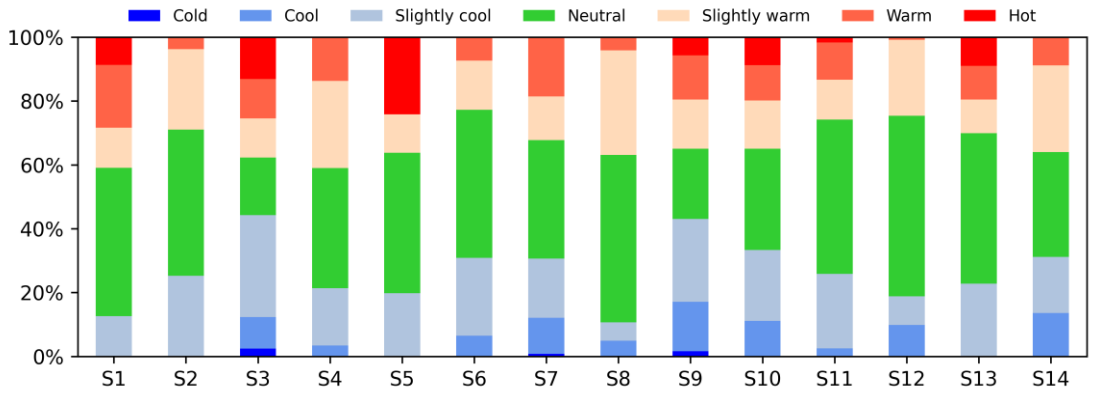

Fig. 9 Thermal sensation votes of investigated fourteen subjects

Fig. 9 shows the distribution of thermal sensation votes of investigated subjects based on ASHRAE-55 7-scale. The overall TSV distribution appears to be relatively balanced, with the majority of individuals voting neutral (green), except for subjects S3 and S9. These two subjects are also the only ones who vote cold (-3). In addition, the entire group has a higher prevalence of combined votes for "*cool*" and "*warm*" sensations.

3.2 Model performance

After mapping the ASHRAE 55 7-scale TSV to the 3-scale TSV used in this paper

(slightly cool and slightly cool as neutral, cool as cold, and warm as hot), we employed five popular ensemble tree algorithms to train the machine learning models for predicting the TSV. The entire dataset of 1,697 samples was divided into a training and testing set at an 8:2 ratio, resulting in 1,357 samples for training and 340 samples for testing. Table 7 shows the best combinations of hyperparameters obtained through the grid-search method. The corresponding predictive performance of these five tree-based models is presented in Table 8 and compared with classic PMV predictions. All tree-based models achieved accuracy above 85%, except for AdaBoost which achieved 76%. XGBoost obtained the highest accuracy, precision, and F1 score, while PMV had the highest recall. Therefore, the XGBoost algorithm was further examined with different combinations of input features, as shown in Table 9.

When using individual Env. (T_a, T_g, RH), BP-related (systolic blood pressure, diastolic blood pressure, and heartrate), ear temperature, wrist temperature, or ankle temperature for prediction, the accuracy and F1-score both remain below 80% and 60%, respectively. When using the individual facial feature alone, XGBoost achieves an accuracy of 77-78%, but the overall F1-score is quite low, consistently less than 40%. This indicates that although the model can predict the correct labels reasonably well, its general ability to correctly identify both the actual positive samples (recall) and the predicted positive samples (precision) is poor. However, when combining all facial information, the prediction performance surpasses that of other individual features (green background in Table 9). As the number of features continues to increase, the prediction performance further improves. Moreover, by adding features "$T_{ear,}$ $T_{wrist,}$ T_{ankle}" and facial features to the basic combination (Env. + BP-related), the accuracy improves by 5%, and other metrics also show significant improvements from 3.5% in recall to 11.9% in precision (blue backgrounds in Table 9). The magnitudes of improvement for both approaches are similar, but the facial features' improvement is slightly lower, ranging from 0.8% to 3.4%, compared to the addition of "$T_{ear,}$ $T_{wrist,}$ T_{ankle}" features across four evaluation metrics. The XGBoost model achieves the best performance when all features are used.

Table 7. Optimal parameters for model training based on grid search

Model	Number of estimators [50, 100, 150, 200]	Learning rate [0.1, 0.25, 0.5, 0.75, 1.0]	Max depth [3, 5, 10, 15, 20, 25]	Min samples leaf [1, 2, 5, 10]	Subsample [0.6, 0.8, 1.0]	Other hyperparameters
RF	100	-	15	2	-	Criterion: entropy from ['gini', 'entropy']
AdaBoost	100	0.25	-	-	-	Algorithm: SAMME from [SAMME, SAMME.R]
GBDT	200	0.1	3	1	-	-
XGBoost	-	-	5	-	0.8	Min child weight: 5 from [1, 5, 10] Gamma: 1 from [0.5, 1, 1.5, 2, 5]
LightGBM	50	0.1	20	-	0.6	-

Table 8. Performance metrics of different ensemble tree models using all features and PMV

	Accuracy	Precision	Recall	F1	Training duration
RF	85.6%	75.3%	63.7%	67.5%	4.7 minutes
AdaBoost	76.2%	59.5%	73.7%	63.6%	2.6 minutes
GBDT	86.2%	79.0%	69.8%	73.1%	52.1 minutes
XGBoost	**88.2%**	**81.9%**	73.6%	**76.8%**	19.5 minutes
LightGBM	87.6%	79.8%	73.3%	75.8%	11.3 minutes
PMV (±1)	66.6%	56.3%	**78.0%**	58.7%	0.64 seconds

Note: Three PMV inputs were simplified in the chamber environments: air velocity = 0.1 m/s, metabolic rate = 1 met (reading activity in office), clothing level = 0.46 clo (consists of 0.36 clo for typical summer ensemble and an additional 0.1 clo for a sitting chair), and mean radiant temperature = f (air temperature, globe temperature, air velocity) in equation (9) according to ISO 7726-1998 [73]. Precision and recall focus on capturing true positive and false negative samples in classification problems, and F1-score provides a balanced evaluation of both precision and recall.

Table 9. Performance metrics of XGBoost models using different combinations of features

Feature combinations	Accuracy	Precision	Recall	F1
Env[a]	75.3%	53.8%	52.2%	52.9%
BP-related[b]	75.3%	52.0%	40.8%	42.8%
T_{ear}	78.8%	51.3%	35.2%	33.0%

T_{wrist}	76.8%	51.8%	48.6%	49.2%
T_{ankle}	78.2%	57.2%	54.7%	55.8%
T_{wrist}, T_{ankle}	79.1%	61.5%	58.3%	59.5%
$T_{Forehead-Check}$	77.9%	48.4%	35.9%	34.4%
$T_{Inner\ canthus-Check}$	78.8%	78.8%	39.4%	39.5%
$T_{Nose-Check}$	77.1%	48.3%	34.7%	32.5%
$T_{Nasolabial\ fold-Check}$	78.5%	44.9%	36.1%	34.8%
$T_{facials}$[c]	80.9%	62.5%	59.0%	60.6%
T_{ear}, T_{wrist}, T_{ankle}	82.1%	67.6%	62.7%	64.8%
"T_{ear}, T_{wrist}, T_{ankle}" + $T_{facials}$	86.5%	80.8%	72.8%	75.3%
Env+ BP-related	82.1%	69.2%	66.4%	67.5%
Env+ BP-related + "T_{ear}, T_{wrist}, T_{ankle}"	87.9%	81.1%	73.2%	76.5%
Env+ BP-related + $T_{facials}$	87.1%	78.5%	69.9%	73.1%
Env+ BP-related + "T_{ear}, T_{wrist}, T_{ankle}" + $T_{facials}$	**88.2%**	**81.9%**	**73.6%**	**76.8%**

Not: [a]Env represents three physical parameters: air temperature, global temperature, and relative humidity; [b]BP-related represents three blood pressure related parameters: systolic blood pressure, diastolic blood pressure, and heartrate; [c]$T_{facials}$ represents four temperature variations from forehead, inner canthus, nose, and nasolabial fold to the cheek temperature.

3.3 Feature interpretation

3.3.1 Contribution of a single sample to the total ranking in SHAP

We use the SHAP (SHapley Additive exPlanations) values proposed by Lundberg and Lee [44] to explain the contributions of training samples and features in the chamber experiments. Specifically, we employed XGBoost to train two binary classification models: "*non-cold vs. cold*" and "*non-hot vs. hot*". The SHAP values were computed for each sample using all features in the dataset. A positive SHAP value indicates that the model's output is closer to 1 (cold or hot label), while a negative value suggests the output is closer to 0 (non-cold or non-hot label). This approach allowed us to gain insights into the model's decision process and understand how individual samples contribute to the classification outcomes, as shown in Fig. 10. The left thermal images depict the same individual voting cold and hot, while the right waterfall figures illustrate the decision process. The red arrows in the waterfall plot represent positive contributions to the model's output towards 1 (feeling cold or hot), while the blue arrows represent negative contributions towards 0. E[F(X)] denotes the baseline value, which is the average output of the training set. Starting from this baseline, each

additional feature of the sample leads to an expected change in the output. After computing all features and sorting them based on their absolute magnitudes, we obtain the final SHAP value f(x). In Fig. 10, both E[F(X)] baseline values are less than 0, while the final SHAP values are greater than 0. This indicates that the majority of the votes in both models are classified as non-cold or non-hot, and the two selected samples in the figure contribute to the explanation process by having a voting output of cold or hot. For these two specific votes, the SHAP value exhibits significantly positive contributions of four facial features for the cold voting, with contribution values ranging from 0.69 to 1.59, particularly the inner canthus. However, when feeling hot, these facial features show relatively smaller contributions, with values ranging from -0.18 to 0.5.

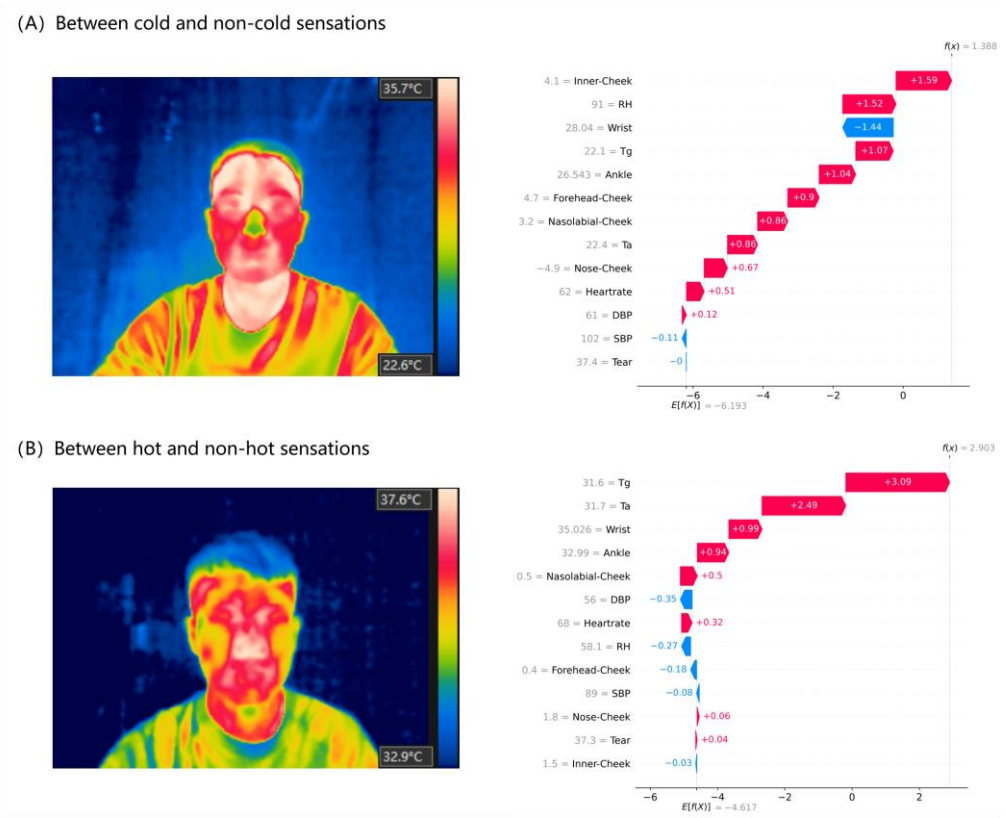

Fig. 10 Decision process of SHAP values with cold and hot voting

3.3.2 Importance ranking and local explanation

After computing the SHAP contributions of each feature for every sample, their absolute values are averaged to obtain the global feature importance ranking, and original SHAP values are concluded in the beeswarm plot to display an information-dense summary, as shown in Fig. 11. In the beeswarm plot, the y-axis order represents

the importance ranking, the position of each dot on the x-axis displays the corresponding SHAP value, and the color intensity of each point indicates the magnitude of the associated feature value. For instance, in the "*Ta*" row of Fig. 11 (a), the concentrated red dots on the left suggest the model's tendency to predict non-cold sensations during **high** air temperatures. Conversely, the presence of blue dots with a long tail on the right indicates a stronger possibility to predict cold sensations when air temperatures are **low**, and this long tail feature suggests that feature Ta can affect each individual differently. When the SHAP value is 0, it indicates that the corresponding feature contributes equally to the model's binary classification on the output of 1 or 0, meaning it does not provide a distinct impact on the model prediction. On the other hand, when the SHAP value deviates significantly from 0, it implies an increased contribution to the model prediction.

In Fig. 11 (a), when classifying between "*cold*" and "*non-cold*" votes, the air temperature exhibits rank one and no points with SHAP = 0, and a concentration of red points is observed on the negative side. This suggests that air temperature positively or negatively contributes to all samples, with higher temperatures tending to lead to a "*non-cold*" vote. In contrast to "*hot*" and "*non-hot*" votes in Fig. 11 (b), air temperature also remains ranked one and extends further towards both the negative and positive sides. A cluster of blue points on the negative side indicates that air temperature plays a larger role in this judgment, with the leftmost concentration of blue points indicating that lower temperatures tend to result in "*non-hot*" votes. The right long tail of the air temperature in the beeswarm plot indicates that the feature's contribution varies across individual samples and some samples rely strongly on this feature. The SHAP results based on game theory are consistent with human intuition and previous experimental findings, demonstrating the relationship between temperature and cold/hot sensations, as well as individual differences within the same thermal environments.

The top-contributing physiological factors are all facial-related features, with the feature "*inner canthus-cheek*" being prominent in cold sensation evaluation, and the feature "*nose-cheek*" in hot sensation evaluation. Both features exhibit similar local explanations, with red long tails on the right and blue clusters on the left, indicating that

larger facial temperature differences (higher red values) positively contribute to specific users' cold/hot judgments under certain thermal environments.

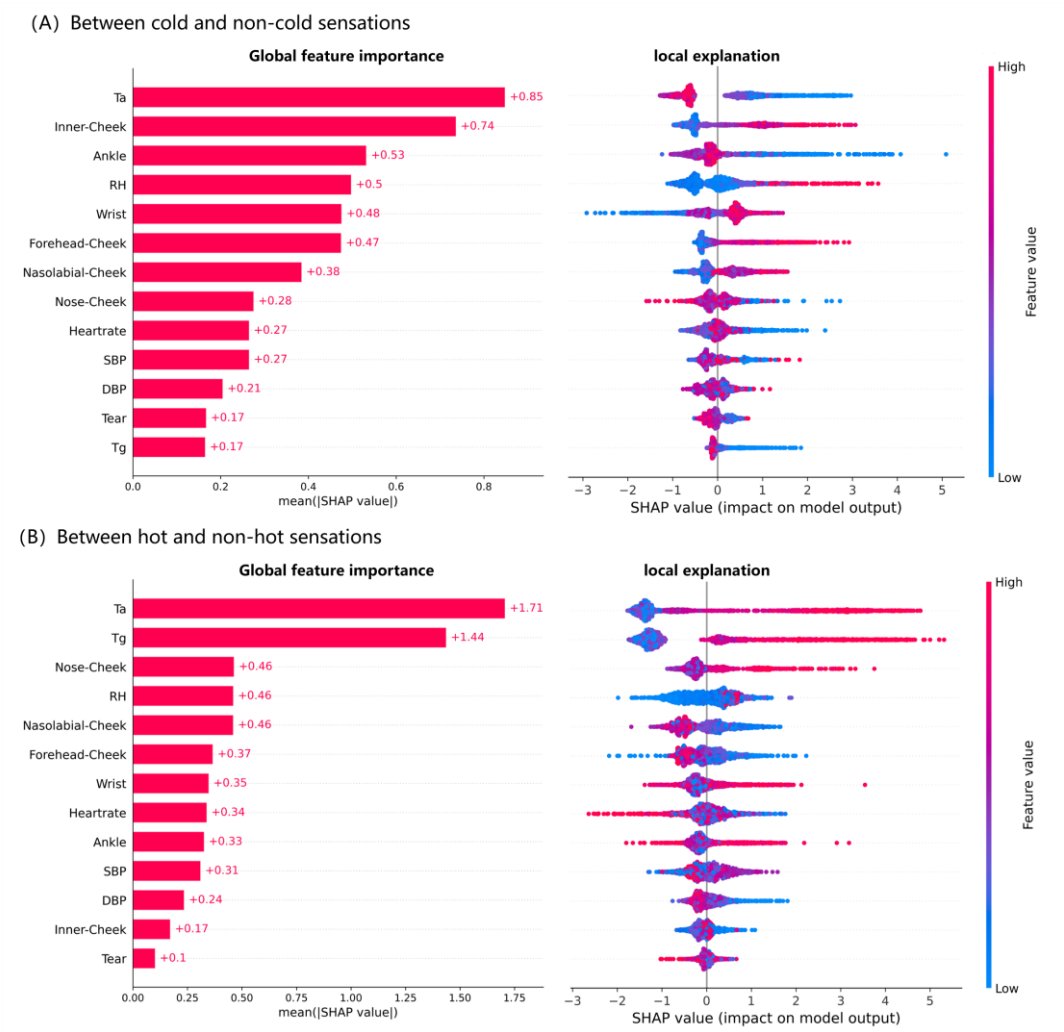

Fig .11 Importance ranking based on the global feature and local explanation

3.3.3 Feature interactions

Fig. 12 examines the impact of each feature at various feature values on the prediction for all samples, focusing specifically on the rank of one environmental feature and the rank one physiological feature used during the training of two models. Points above the black dotted line (SHAP = 0) indicate positive contributions to the prediction output of "*1*" (voting cold or hot). The fill color of each point is determined by the feature that exhibits the highest local interaction effect with the considered feature. This interaction effect is computed using the original "*Shapley interaction index*" from game theory, which allocates credit not only to each individual player but also to all possible combinations of players [44].

When evaluating hot and cold sensations, the feature "*air temperature*" exhibits fewer points concentrated around SHAP = 0, indicating its significant impact on model prediction. Similarly, the feature "*inner canthus-cheek*" in cold sensation evaluation shows scattered points with positive SHAP values, some of which are close to 3, resembling the pattern observed for "*air temperature*". However, a larger number of points are distributed at lower levels, with SHAP values ranging from 0.5 to 1.5, suggesting a relatively lower contribution to the model's prediction compared to "*air temperature*". As for the feature "*nose-cheek*" in hot sensation evaluation, it shares similarities with the results for "*inner canthus-cheek*", where some points show substantial positive contributions. Nevertheless, a significant concentration of points around SHAP = 0 indicates that during training, many samples assign limited importance to this feature.

The interaction effects of each feature exhibit distinct patterns. Specifically, the interaction between "*inner canthus-cheek*" and "*air temperature*" shows a concentration of high values (red points) around SHAP = 0, indicating a limited contribution of "*inner canthus-cheek*" to the evaluation of cold sensation when air temperature is high. In the case of the interaction between "*nose-cheek*" and "*globe temperature*", a clear separation of filled globe temperature points is observed when "*nose-cheek*" is greater than 2℃ (right side of the plot). This indicates that a higher "*nose-cheek*" temperature provides a stronger indication for hot sensation judgments under high globe temperatures, compared to low globe temperatures.

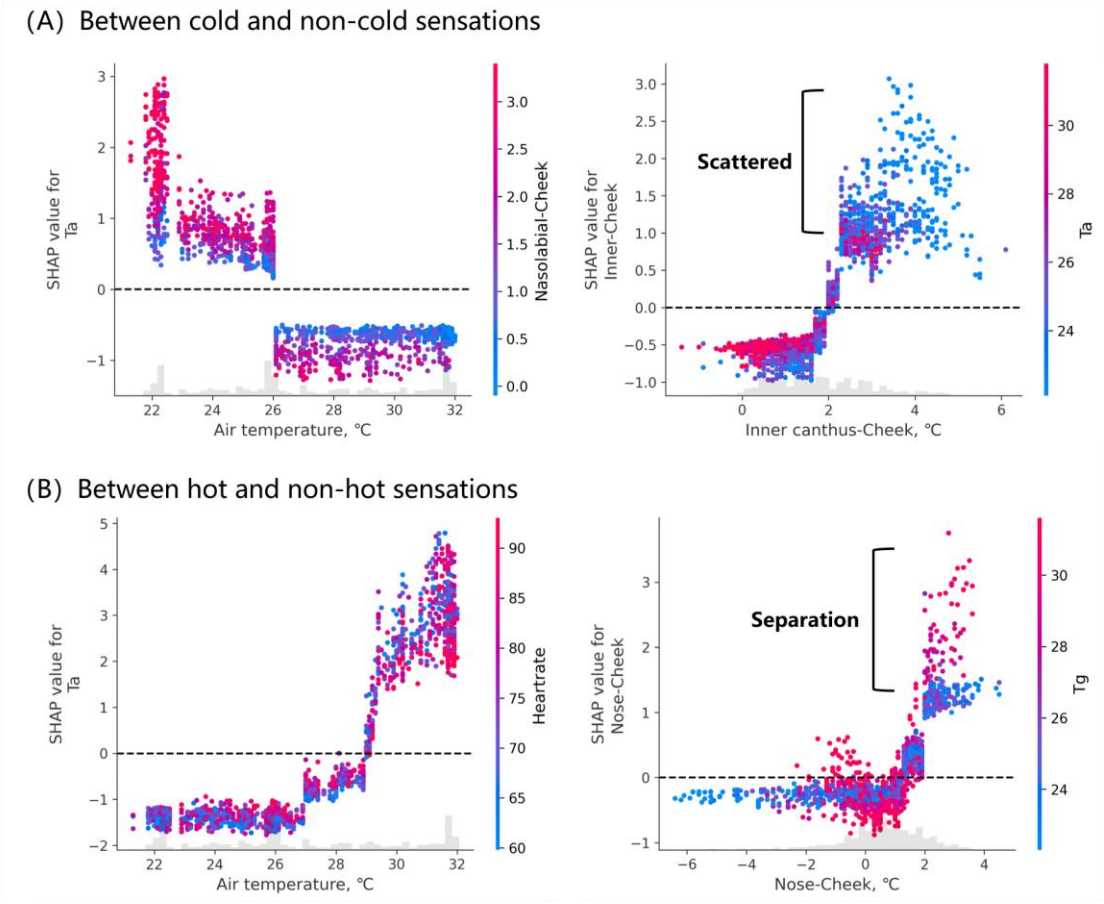

Fig. 12 SHAP dependence plot of key environmental and physiological features

In Fig. 12, the interaction plots for features "*inner canthus-cheek*" and "*nose-cheek*" feature exhibit discrete clusters of scattered points and separated points, which were further divided based on SHAP value=1.5 and the voting count of different subjects, resulting in Fig. 13. It can be observed that facial features exhibit strong contributions for the majority of subjects, while some subjects obtain weaker contributions. Subjects with lower contributions from facial features tend to be less selective (or *picky*) in their thermal environments (higher green triangles). For example, S5 and S13 maintain relatively stable facial temperatures across experimental conditions and have a high proportion of TSV=0, accounting for 44% and 48%, respectively, and the counts of their high facial temperature variation are close to 0. On the other hand, subjects with more significant facial reactions tend to be more selective in thermal environments, such as S3 and S9, as evidenced by their high counts above SHAP=1.5 and low proportions of TSV=0, which are 19% and 22%, respectively. For subjects S12 and S7, they also show high facial temperature variation counts, but the proportion of high SHAP values (>1.5)

is relatively low, which may result in a higher number of TSV=0 votes (56% and 37%, respectively).

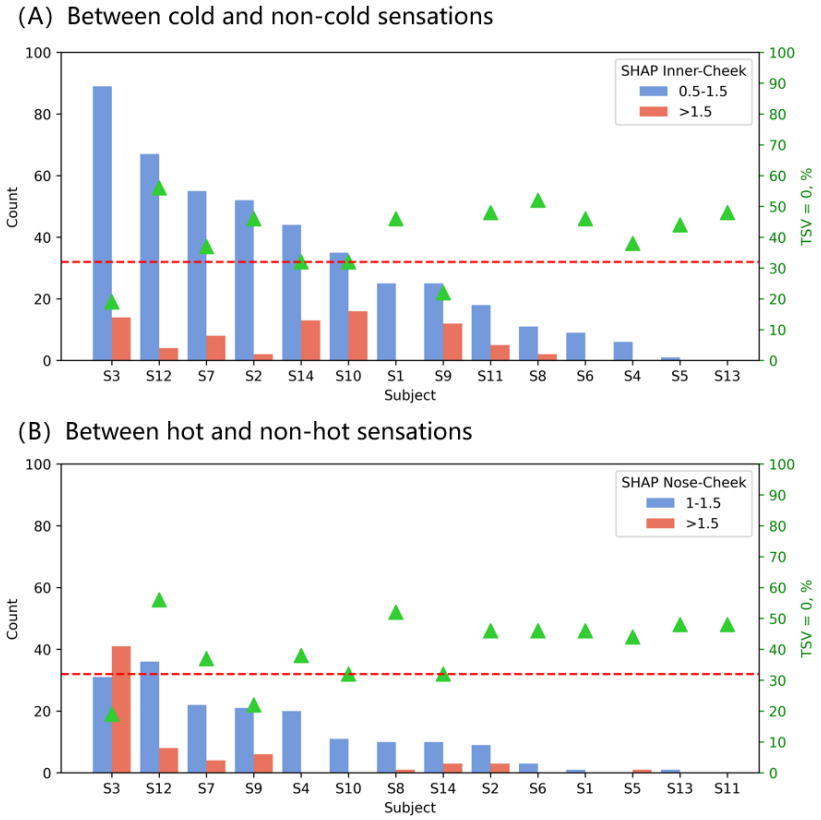

Fig. 13 Counts of high SHAP values for physiological features in interpreting cold and hot sensations in relaxed neutral conditions corresponding to the proportion of TSV=0 in green triangles ("*inner canthus-cheek*" > 2°C, and "*nose-cheek*" > 1.7°C).

3.4 Mapping slightly cool and slightly warm to non-comfort sensations as stringent neutral conditions

Previous analysis classified the votes for "*slightly cool*" and "*slightly warm*" as neutral. Several standards require air-conditioned environments to maintain a narrower comfort range, such as ISO 7730 and EN 16798, which define the category A comfort zone as being within ±0.2 PMV. In this section, we reassigned the votes for "*slightly cool*" (-1) and "*slightly warm*" (+1) as uncomfortable, classifying -3, -2, -1 as cold sensations and +1, +2, +3 as hot sensations, with 0 remaining neutral. Following that, we used the XGBoost algorithm for the training process and output prediction to assess whether significant differences exist in the results.

Table 10 shows the predictive performance of XGBoost models. Compared to previous

results in Table 8, the overall predictive performance shows significant reductions, which could be attributed to significant individual differences in perceiving the narrow range of thermal neutrality that increases the prediction difficulty. Four evaluation metrics (accuracy, precision, recall, and F1-score) fall below 70% when only physiological features are considered. However, when all physiological features are combined, the performance exceeds 70%. In contrast, PMV (green backgrounds) and the XGBoost model with environmental variables (blue backgrounds) demonstrate superior predictive performance with all evaluation metrics exceeding 70%. When physiological features are added to the environmental feature-based XGBoost model, performance gradually improves, with the potential to raise evaluation metrics by around 4%.

Table 11 further categorizes precision, recall, and F1-score by each vote, revealing that apart from blood pressure-related features, other feature combinations generally exhibit the best predictive performance for the hot vote, with the highest precision, recall, and F1-score. Conversely, they exhibit the poorest predictive performance for the neutral vote. This suggests that predicting sensations of hot and cold can be easier, while predicting neutral sensations is often more challenging due to individual differences in temperature preferences within the same moderate environment.

Table 10. Performance metrics of XGBoost models using different combinations of features when mapping votes slightly cool and slightly warm to neutral

Feature combinations	Accuracy	Precision	Recall	F1
BP-related	41.2%	40.5%	40.1%	40.1%
T_{ear}	45.3%	48.3%	42.4%	42.4%
T_{wrist}	59.4%	59.4%	58.3%	58.6%
T_{ankle}	64.4%	65.1%	63.8%	64.3%
T_{wrist}, T_{ankle}	65.3%	65.5%	64.5%	64.9%
T_{ear}, T_{wrist}, T_{ankle}	66.2%	66.6%	65.5%	65.9%
$T_{facials}$	62.9%	63.9%	62.5%	63.1%
"T_{ear}, T_{wrist}, T_{ankle}" + $T_{facials}$	72.9%	74.5%	72.5%	73.2%
PMV (±0.5)	73.5%	73.2%	74.6%	73.3%
Env	74.7%	75.3%	74.8%	75.0%
Env+ BP-related + "T_{ear}, T_{wrist}, T_{ankle}"	77.6%	78.6%	77.7%	78.1%
Env+ BP-related + $T_{facials}$	76.2%	76.8%	76.4%	76.6%
Env+ BP-related + "T_{ear}, T_{wrist}, T_{ankle}" + $T_{facials}$	**78.8**%	**79.5**%	**79.1**%	**79.3**%

Table 11. Performance metrics of XGBoost models using different combinations of features when mapping votes slightly cool and slightly warm to neutral for each vote

Feature combinations	Vote	Accuracy	Precision	Recall	F1
BP-related	Cold	41.2%	39.5%	35.6%	37.4%
	Neutral		**44.0%**	**51.1%**	**47.3%**
	Hot		38.0%	33.6%	35.7%
T_{ear}, T_{wrist}, T_{ankle}	Cold	66.2%	63.9%	58.9%	61.3%
	Neutral		62.3%	68.6%	65.3%
	Hot		**73.6%**	**69.0%**	**71.2%**
$T_{facials}$	Cold	62.9%	65.4%	56.7%	60.7%
	Neutral		57.0%	62.8%	59.7%
	Hot		**69.4%**	**68.1%**	**68.8%**
PMV (±0.5)	Cold	73.5%	71.5%	72.4%	72.0%
	Neutral		72.3%	57.8%	64.2%
	Hot		**75.8%**	**93.4%**	**83.7%**
Env	Cold	74.7%	76.5%	72.2%	74.3%
	Neutral		69.3%	70.8%	70.0%
	Hot		**80.0%**	**81.4%**	**80.7%**
Env+ BP-related + "T_{ear}, T_{wrist}, T_{ankle}" + $T_{facials}$	Cold	78.8%	81.6%	78.9%	80.2%
	Neutral		73.0%	75.2%	74.1%
	Hot		**83.9%**	**83.2%**	**83.6%**

Fig. 14 depicts the feature contributions based on SHAP values. In its two middle subplots, "*air temperature*" still remains the most influential feature compared to the previous results, extending even further and forming corresponding clusters. This implies that more samples tend to assign higher contributions to "*air temperature*" regardless of whether the output is 1 (cold or hot) or 0 (non-cold or non-hot) during the training process.

However, the most influential physiological parameter has shifted in both models. In assessing cold sensations, the feature "*forehead-cheek*" has now taken the place of the previously prominent feature "*inner canthus-cheek*", which was previously ranked 2. Consequently, "*inner canthus-cheek*" has been relegated to the bottom of the importance ranking. On the other hand, for the evaluation of hot sensations, the most influential physiological feature is "*ankle*". The overall importance ranking of facial features shows a decline, with some being surpassed by blood pressure features. The dependence plots for the feature "*air temperature*" in Figs. 15 (a) and 15 (b) exhibit smooth S-shaped curves. These curves indicate significant contributions from data

points located at extreme air temperature values, while the near-linear change of SHAP values for intermediate air temperatures demonstrates a gradual change of impact on feature contribution.

In the right subplot in Fig. 14 (a), a sudden truncation occurs when "*forehead-cheek*" values are approximately below 0.2℃. On the right side of truncation, the linear relationship suddenly ends, and a significant proportion of data points with high ankle temperatures tend to converge towards SHAP=0, indicating that feature "*forehead-cheek*" becomes less useful in making reliable judgments of feature contribution. In the right subplot of Fig. 14 (b), when evaluating hot sensations, the feature "*ankle*" exhibits higher contributions when the RH is low. This observation can be attributed to the correlation between low RH and high air temperature.

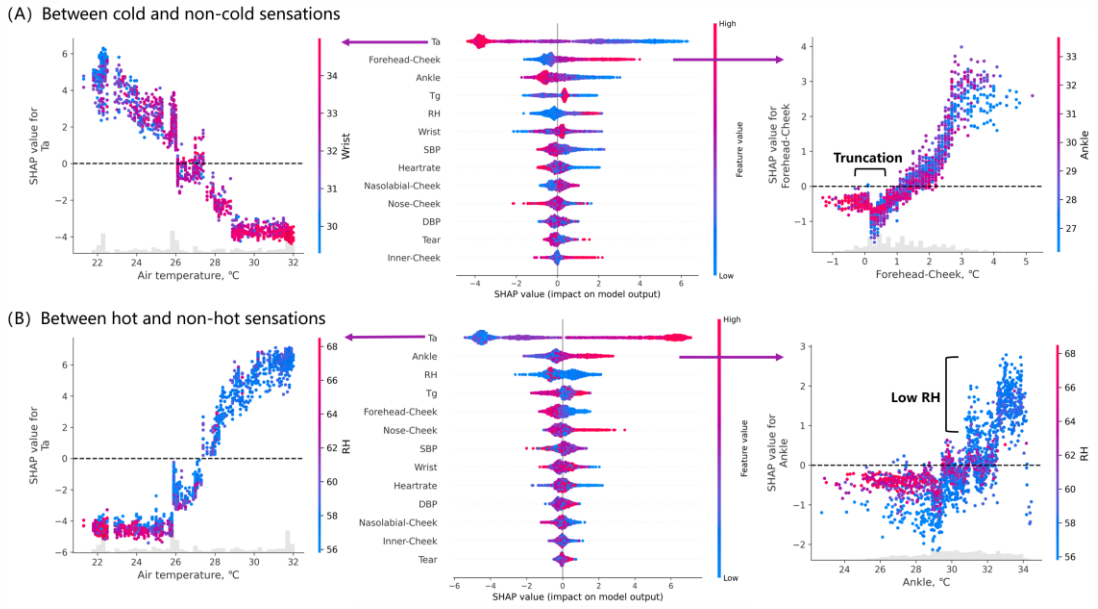

Fig .14 Importance ranking based on local explanation and dependence plot of top two features filled with their most interactive feature

Fig. 15 shows the high SHAP value counts of the features "*forehead-cheek*" and "*ankle*" under stringent neutral conditions, with the green triangles reflecting the proportion of users voting for TSV=0. According to Fig. 15 (a), some subjects who are more selective about their environments have a higher frequency of high SHAP values for the facial feature "*forehead-cheek*", such as S3 and S9. Meanwhile, subjects with low or zero counts of high facial feature contributions exhibit higher proportions of TSV=0, as seen on the right side of Fig. 15 (a). In contrast, Fig. 15 (b) shows the results for the feature

"*ankle*" where no comparable pattern was detected, demonstrating that all subjects' physiological responses related to the ankle are relatively similar. Therefore, compared to the feature "*ankle*", the SHAP values of the facial feature "*forehead-cheek*" are more effective in distinguishing between people who are more or less selective about thermal conditions, thus providing a better representation of individual differences.

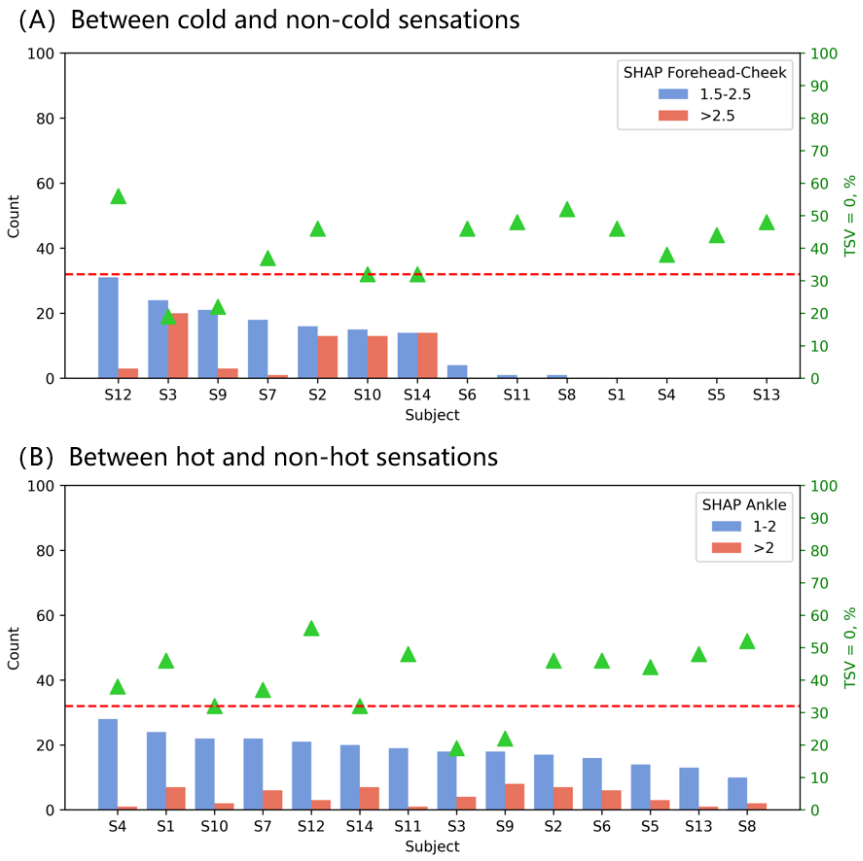

Fig. 15 Counts of high SHAP values for physiological features in interpreting cold and hot sensations in stringent neutral conditions corresponding to the proportion of TSV=0 in green triangles ("*forehead-cheek*" > 2.3°C, and "*ankle*" > 30.75°C).

Currently widely adopted Fanger's PMV and Gagge' SET models are both based on statistical modeling of the "*standard person*", assuming uniform physiological parameters such as BMI and baseline core body temperature. However, in reality, every individual has unique biological characteristics in terms of physiological regulation and skin type [74]. These characteristics are influenced by demographic factors (such as age, gender, and race) and physiological factors (such as metabolism, and hormone levels), which affect their perception of temperature and thermal comfort, resulting in variations in temperature responses and skin properties. Obermeyer et al. [75], after

statistical analysis of 243,506 core body temperature data points from 35,488 patients, excluding extreme core temperature situations like emergencies, found that core body temperature decreases by approximately 0.021°C with each additional decade of age. There are also differences in core body temperature between races, with African-American women having a higher temperature than white men by 0.052°C.

Research also indicates a decline in core body temperature among some Americans (0.03°C per decade) [76] and tropical populations (declined by 0.05°C/year over 16 years) [77] compared to previous levels. Therefore, individuals' baseline core body temperatures are correlated with many factors, and there is a risk of misjudgment when using a completely uniform physiological standard for modeling (such as the "*standard person*" used in PMV and SET models established about 5 decades ago) to assess individual health or thermal comfort. Our experimental results further indicate differences in facial responses and thermal preferences among individuals within the same race and similar age groups. Further analysis and understanding of these differences will contribute to the construction of more accurate and theoretically solid thermal comfort models.

4. Discussion

4.1 Performance of tree-based models and the PMV index

This study trained random forests and four popular boosting tree models (AdaBoost, GBDT, XGBoost, and LightGBM) to predict the thermal comfort of subjects in a well-controlled climate chamber. After conducting a grid search for hyperparameter tuning, XGBoost, which demonstrated the best performance, was selected for further exploration based on different feature combinations, and its results were compared with the PMV index. Overall, XGBoost models achieved better performance when physiological features were added as extra inputs alongside environmental features (4-24% performance improvements). This emphasizes the data-driven nature of machine learning algorithms and their benefits of integrating extra feature dimensions for better predictions.

The inclusion of facial features in the XGBoost training has shown positive effects.

When mapping "*slightly cool*" and "*slightly warm*" as comfortable (relaxed neutral conditions), the XGBoost model using only facial information achieved better predictive performance (80.9% accuracy) compared to features of wrist and ankle (76.8% and 78.2% accuracies). However, when categorizing "*slightly cool*" and "*slightly warm*" as uncomfortable (stringent neutral conditions), all physiological feature-based XGBoost models performed worse compared with only using environmental parameters (74.7% accuracy). Among them, the XGBoost model using only facial features achieved an accuracy of 62.9%, outperforming the wrist feature (59.4% accuracy) but falling behind the ankle feature (64.4% accuracy). The precision, recall, and F1-score metrics also demonstrated consistent patterns with the accuracy results. This implies that the labeling of "*slightly cool*" and "*slightly warm*" can significantly impact XGBoost's predictive performance, indicating that each individual's requirements for extremely neutral environments (TSV=0) vary greatly. This variation brings uncertainty to the training process of machine learning algorithms. When the requirements for neutral environments are relaxed, individual physiological parameters outperform environmental parameters in XGBoost training. Therefore, in stringent neutral conditions, environmental factors could play a more important role in predicting thermal comfort, while under relaxed neutral conditions, physiological parameters become more significant.

This study also proves that PMV can achieve satisfactory results in 3-class classification problems when operating in stringently controlled conditions. Its predictive performance was found to be similar to that of fine-tuned XGBoost when using only environmental parameters as inputs. However, when XGBoost incorporated additional physiological factors, PMV's performance fell behind.

4.2 Contribution of facial thermography using SHAP value

Although machine learning has made great advances in several domains, including thermal comfort research, the interpretability of these models remains a critical challenge in practical deployments. The interpretability degree of a certain model can significantly influence people's trust in using it. Traditional tree-based models can provide feature importance rankings, but they are unable to quantify the individual

contributions of each sample to model training, identify interactions between samples, or find feature threshold values that would have the greatest impact on model training. However, the SHAP-based explainable AI has overcome some of these challenges. It can provide a local perspective to explain machine learning models by mapping players in a cooperative game to specific features and player allocation scores to feature contributions. This allows us to better understand how each unique sample and feature impacts the training and decision-making processes of data-driven machine learning algorithms.

For the SHAP interpretation during XGBoost training process, facial features showed significant positive contributions, with their importance generally higher than other physiological parameters, such as wrist, ankle, and blood pressure-related parameters. Under relaxed neutral conditions, the feature "*inner canthus-cheek*" ranked second, closely following the feature "*air temperature*". Its pronounced right-skewed red tails in the SHAP local explanation plot in Fig. 11 (a) suggest that high "*inner canthus-cheek*" variation values can provide effective evidence for cold sensation judgments. However, in the SHAP interaction plot of Fig. 12 (a), it can be observed that these high "*inner canthus-cheek*" differences were more dispersed, indicating individual differences exist within these points. Moreover, its strongest interaction feature "*air temperature*" indicates that lower temperatures are more likely to result in higher "*inner canthus-cheek*" variations, which will push the model towards predicting cold sensations. Although the feature "*inner canthus-cheek*" may not present high contributions for all samples, its overall predictive contributions remain promising, as evidenced by a mean absolute SHAP value of 0.74. This value is slightly lower than the mean absolute SHAP value of rank one feature "*air temperature*", which is 0.85. Similar notable feature contributions were also observed for "*nose-cheek*" variation and "*forehead-cheek*" variation in the SHAP analysis, obtaining ranks 3 and 2, respectively.

After further dividing the high SHAP values of the feature "*inner canthus-cheek*" by a threshold of SHAP=1.5 in Fig. 13 (a) when "*inner canthus-cheek*" values are greater than 2℃, it is clear that this feature reflects individual differences. Subjects with consistently low "*inner canthus-cheek*" variations tend to be less selective about

thermal environments (subjects 5 & 13, 44-48% of voting TSV=0), while subjects with high facial reactions may be more selective (subjects 3 & 9, 19-22% of voting TSV=0). Interestingly, even for subjects with intense facial reactions, if the proportion of SHAP>1.5 values remains low, they appear to be more accepting of their surroundings (subjects 7 & 12, 37-56% of voting TSV=0). Therefore, the SHAP-based local explanations and interaction effects can provide a more comprehensive understanding compared to feature importance rankings. This will allow us to gain insights not only into the overall importance of features but also into the specific contributions of individual samples and their distribution. Furthermore, these analyses shed light on the interactions between different features, offering a more in-depth knowledge of their mutual influences.

In general, it's essential to highlight the unique capabilities of machine learning models in filling the gap left by traditional modeling approaches: **1) Individualized predictions:** machine learning models can provide individualized predictions of thermal sensation by considering a broader range of physiological parameters and environmental variables, including facial thermography data, as demonstrated in our study; **2) Adaptive learning:** unlike static models such as PMV, machine learning models can continuously learn and adapt to new data, allowing them to evolve and improve their predictive performance over time; **3) Interpretability:** while interpretability remains a challenge in machine learning, techniques such as SHAP-based explainable AI, as shown in our analysis, offer insights into the underlying factors driving thermal comfort predictions. This will enhance transparency and trust in the decision-making process of AI models.

By leveraging machine learning techniques, we can bridge the gap between traditional static models and the dynamic, individualized nature of thermal comfort assessment, thereby enhancing the applicability and accuracy of our predictions.

4.3 Relationship between facial interpretation and physiological basis

This section further explores the physiological basis of top-ranking facial features in SHAP local explanation: "*inner canthus-cheek*", "*nose-cheek*", and "*forehead-cheek*".

4.3.1 Inner canthus

The inner canthus is considered the warmest facial region that closely reflects the core temperature of human body, as it receives an abundant blood supply from the lacrimal branch of the ophthalmic artery [78]. Furthermore, being located within a well-protected facial recess, the inner canthus experiences less heat loss from radiation and convection compared with other regions on the face. This will contribute to the relative stability of its temperature in individuals who are not experiencing fever. Pascoe and Fisher [79] investigated the core body temperature of 22 university students. They discovered that as the ambient temperature ranged from 15.5℃ to 26.6℃, the temperature of the inner canthus increased by only 1.2℃, rising from 35.7℃ to 36.9℃. Therefore, the feature "*inner canthus-cheek*" can be considered as a representation of cheek temperature to a certain extent. Previous research in thermal comfort has already demonstrated that cheek temperature is highly indicative for predicting thermal comfort [80] and shows significant correlations with thermal sensations [81]. The SHAP interpretation in this study further elucidates that cheek temperature plays a more significant role in predicting cold sensations, especially under broader neutral conditions during the training process of machine learning models when "*inner canthus-cheek*" variations are beyond 2℃. However, when neutral conditions are more stringent, its contributions become limited compared to environmental parameters.

4.3.2 Nose

The nose is typically the coldest and most temperature-sensitive area on the face, because of its high surface area to volume ratio [82], the avascular nature of its cartilaginous component [83], and the influence of inhaled air before it is warmed by the nasal mucosa in the nasal cavity [84]. Ghahramani et al. [15] exposed individuals to cold and heat stress in office environments. They observed that nose temperature (31.70 ± 2.33℃ and 34.78 ± 1.66℃) generally exhibited lower mean values and higher standard deviations compared to forehead (34.06 ± 0.58℃ and 35.53 ± 0.58℃) and cheek (33.27 ± 1.18℃ and 35.31 ± 0.71℃). Because the maximum heat exposure in their study was 29°C, there was no occurrence of the nose temperature exceeding the cheek temperature, which happened in this study and could have been caused by the

more extensive exposure to higher temperatures at 32℃. Tejedor et al. [85] discovered a high correlation (95.14%) between nose temperature and skin temperature in the elderly, making it a potential thermal comfort indicator. However, nose temperature is also considered to be linked to emotions. For example, in infants under one-year old, nose temperature can drop by 2 ℃ within 2 minutes after laughing [86], whereas in adults, it tends to rise after experiencing feelings of happiness or positive emotions [87]. These differences are believed to be indications of the body's development at various periods of life [86]. To control for these confounding factors, the participants in this study were instructed to do typical office work to avoid direct emotional changes caused by entertainment or other factors. Our SHAP interaction analysis reveals that when the feature "*nose-cheek*" exceeds 1.7 ℃, its contribution grows dramatically, especially in the *picky* users S3 and S9.

4.3.3 Forehead

The forehead generally achieves high mean temperatures due to its proximity to the brain, allowing conductive and convective heat transfer that helps regulate brain temperature [88]. Additionally, the forehead is well vascularized and has a uniform surface area with a thin layer of subcutaneous fat [89]. It could be used for fever detection [90], as well as remote sensing of heart and respiration rates [91]. The forehead is also identified important body part for thermosensitivity [92]. Parkinson et al. [54] measured cutaneous thermoreceptor activity on the forehead in dynamic thermal environments and found that under the "*Cool front fan/High speed*" condition, the receptor impulses on the forehead were significantly higher compared to other body regions. This observation aligns with the result of this study, where the SHAP values of the forehead feature exhibit high contributions in indicating cold sensation. Choi and Yeom [93] investigated personalized thermal comfort modeling using seven different body parts and discovered that the forehead and arms had the strongest correlation with thermal sensation, particularly in males. Pavlin et al. [94] designed an embedded mechatronic device primarily based on forehead temperature collected by infrared cameras, providing a non-invasive solution for building automation or Industry 4.0 applications. The findings of this study indicate that when the "*forehead-cheek*"

variation is greater than 2.3℃, the computation of SHAP values has the potential to distinguish potential users who are more selective or less selective towards the environment. This could provide a more precise decision basis for related non-invasive solutions for more personalized and refined controls.

5. Conclusion

This study proposes a contactless method for estimating occupant thermal state by combining facial infrared thermography, environmental variables, and physiological parameters. Five ensemble tree algorithms were examined using all of the features collected in the chamber experiment, and the best-performing XGBoost model was chosen for further feature selection and explainable AI analysis. The novelty of this research lies in transitioning machine learning models for thermal comfort research from "*black-box*" to "*gray-box*" by conducting explainable AI analysis on the contribution of each feature and specific sample within the high-performance machine learning models, with a particular emphasis on non-contact facial-related infrared features. This will contribute to increasing the trust in non-contact intelligent assessment of human thermal comfort in buildings, thereby enhancing the credibility and reliability of AI model deployments. The main conclusions are:

(1) The approach used to map the TSV labels has a substantial impact on the training and predictive performance of tree-based models for classification problems. When categorizing "*slightly cool/warm*" as comfortable, using one single facial feature within the XGBoost model produces acceptable accuracies of 77.1-78.8% but poor F1-scores of 32.5-39.5%. By combining all facial features, the accuracy and F1-score of XGBoost were increased to 80.9% and 60.9%, respectively, which outperform the predictions obtained from wrist and ankle temperatures, as well as the PMV index. These findings demonstrate that incorporating more facial features can significantly enhance model performance. When "*slightly cool/warm*" is categorized as uncomfortable, the PMV index demonstrates preferable predictive performance (73.5% accuracy), slightly behind the fine-tuned XGBoost model utilizing solely environmental parameters (74.7% accuracy), but is superior to that

of multiple XGBoost models utilizing only physiological features (45.3-62.9% accuracies). In both mapping scenarios, the incremental introduction of facial features for the XGBoost training exhibits a progressive enhancement of model performance.

(2) The SHAP-based explainable AI analysis reveals a consistent distinction of air temperature as the foremost contributing factor, followed by temperature variations in specific facial areas (inner canthus, nose, forehead, and cheek areas) and ankle temperature. Elevated SHAP values become pronounced when the features "*inner canthus-cheek*", "*nose-cheek*", "*forehead-cheek*", and "*ankle*" exceed 2°C, 1.7°C, 2.3°C, and 30.75°C, respectively. Noticeably, high facial SHAP values can contribute to distinguishing individual differences and filtering selective occupants, whereas ankle SHAP values can not.

(3) Among the facial features extracted from infrared thermography, the "*inner canthus-cheek*" and "*forehead-cheek*" show significant local contributions in assessing cold discomfort (rank 2 mean SHAP values), while the "*nose-cheek*" shows remarkable local contributions in assessing hot discomfort (rank 2 mean SHAP values). This alignment correlates to documented trends in the reaction of facial organs to temperature fluctuations observed in the medical literature. Given the obstruction issue with the inner canthus caused by glasses, it is suggested to incorporate the forehead, nose, and cheek temperatures for evaluating occupants' thermal state in practical applications. These facial features can significantly enhance the predictive performance of AI models, allowing them to accurately predict the energy needed by HVAC systems.

(4) Although this paper demonstrates the potential of using the SHAP method to indicate thresholds in facial features and assess individual differences, thereby offering the possibility of individualized cooling or heating from building systems, it is important to note that an excessive and indiscriminate extrapolation of big data and AI solutions could also create socio-ethical quandaries, potentially resulting in discriminatory or equity-related concerns in spaces like buildings. Further exploration is still needed for greater human involvement in AI solutions.

Acknowledgements

The Chongqing University team appreciates the grants support from the National Natural Science Foundation of China (Grant No. 52278090), the Ministry of Science and Technology of the People's Republic of China (Grant No. 2022YFC3801504), and the Natural Science Foundation of Chongqing, and China (Grant No. cstc2021ycjh-bgzxm0156).

Reference

[1] T. Fleiter *et al.*, "Mapping and analyses of the current and future (2020-2030) heating/cooling fuel deployment (fossil/renewables). Work package 1: Final energy consumption for the year 2012," 2016.

[2] International Standard Organization, "ISO 7730 Ergonomics of the Thermal Environment—Analytical Determination and Interpretation of Thermal Comfort Using Calculation of the PMV and PPD Indices and Local Thermal Comfort Criteria." 2005.

[3] E. UNI, "EN 16798-1:2019 Energy Performance of Buildings-Ventilation for Buildings-Part 1: Indoor Environmental Input Parameters for Design and Assessment of Energy Performance of Buildings Addressing Indoor Air Quality." Thermal Environment, Lighting and Acoustics 16798.1, Brussels, Belgium, 2019.

[4] ASHRAE, "Thermal Environmental Conditions for Human Occupancy, ANSI/ASHRAE Standard 55-2020." Atlanta, 2020.

[5] CIBSE, "CIBSE Guide A: Environmental design. 8th edition." London. http://www.cibse.org/getattachment/Knowledge/CIBSE-Guide/CIBSE-Guide-A-Environmental-Design-NEW-2015/Guide-A-presentation.pdf.aspx Accessed 3 November 2019, 2015.

[6] MOHURD, *Evaluation standard for indoor thermal environment in civil buildings (GB/T 50785-2012)*. Ministry of Housing and Urban-Rural Development (MOHURD), Beijing, China, 2012.

[7] J. van Hoof, "Forty years of Fanger's model of thermal comfort: comfort for

all?," *Indoor Air*, vol. 18, no. 3, pp. 182–201, 2008.

[8] A. A.-W. Hawila, A. Merabtine, M. Chemkhi, R. Bennacer, and N. Troussier, "An analysis of the impact of PMV-based thermal comfort control during heating period: A case study of highly glazed room," *J. Build. Eng.*, vol. 20, pp. 353–366, 2018.

[9] A. Yüksel, M. Arıcı, M. Kraj, M. Civan, and H. Karabay, "A review on thermal comfort, indoor air quality and energy consumption in temples," *J. Build. Eng.*, vol. 35, p. 102013, 2021.

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

[11] S. M. Abdollahzadeh, S. Heidari, and A. Einifar, "Evaluating thermal comfort and neutral temperature in residential apartments in hot and dry climate: A case study in Shiraz, Iran," *J. Build. Eng.*, vol. 76, p. 107161, 2023.

[12] P. Jafarpur and U. Berardi, "Effects of climate changes on building energy demand and thermal comfort in Canadian office buildings adopting different temperature setpoints," *J. Build. Eng.*, vol. 42, p. 102725, 2021.

[13] R. Yao, B. Li, and J. Liu, "A theoretical adaptive model of thermal comfort – Adaptive Predicted Mean Vote (aPMV)," *Build. Environ.*, vol. 44, no. 10, pp. 2089–2096, 2009.

[14] S. Nižetić, N. Pivac, V. Zanki, and A. M. Papadopoulos, "Application of smart wearable sensors in office buildings for modelling of occupants' metabolic responses," *Energy Build.*, vol. 226, no. July 2018, 2020.

[15] A. Ghahramani, G. Castro, B. Becerik-Gerber, and X. Yu, "Infrared thermography of human face for monitoring thermoregulation performance and estimating personal thermal comfort," *Build. Environ.*, vol. 109, pp. 1–11, 2016.

[16] F. R. D'Ambrosio Alfano, B. W. Olesen, B. I. Palella, and G. Riccio, "Thermal comfort: Design and assessment for energy saving," *Energy Build.*, vol. 81, no. 2014, pp. 326–336, 2014.

[17] L. Arakawa, V. Soebarto, and T. Williamson, "Performance evaluation of

personal thermal comfort models for older people based on skin temperature , health perception , behavioural and environmental variables," *J. Build. Eng.*, vol. 51, no. March, p. 104357, 2022.

[18] J. H. 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.

[19] W. Liu, Z. Lian, and Y. Liu, "Heart rate variability at different thermal comfort levels," *Eur. J. Appl. Physiol.*, vol. 103, no. 3, pp. 361–366, 2008.

[20] S. Li, X. Zhang, Y. Li, W. Gao, F. Xiao, and Y. Xu, "A comprehensive review of impact assessment of indoor thermal environment on work and cognitive performance-Combined physiological measurements and machine learning," *J. Build. Eng.*, vol. 71, p. 106417, 2023.

[21] M. Favero, I. Sartori, and S. Carlucci, "Human thermal comfort under dynamic conditions: An experimental study," *Build. Environ.*, vol. 204, p. 108144, 2021.

[22] X. Cheng, B. Yang, T. Olofsson, G. Liu, and H. Li, "A pilot study of online non-invasive measuring technology based on video magnification to determine skin temperature," *Build. Environ.*, vol. 121, pp. 1–10, 2017.

[23] T. W. Kim and B. R. Routledge, "Why a right to an explanation of algorithmic decision-making should exist: A trust-based approach," *Bus. Ethics Q.*, vol. 32, no. 1, pp. 75–102, 2022.

[24] J. Burrell, "How the machine 'thinks': Understanding opacity in machine learning algorithms," *Big data Soc.*, vol. 3, no. 1, pp. 1–12, 2016.

[25] A. Ghahramani, G. Castro, S. A. Karvigh, and B. Becerik-Gerber, "Towards unsupervised learning of thermal comfort using infrared thermography," *Appl. Energy*, vol. 211, pp. 41–49, 2018.

[26] A. C. Cosma and R. Simha, "Thermal comfort modeling in transient conditions using real-time local body temperature extraction with a thermographic camera," *Build. Environ.*, vol. 143, pp. 36–47, 2018.

[27] Y. He *et al.*, "Smart detection of indoor occupant thermal state via infrared thermography, computer vision, and machine learning," *Build. Environ.*, vol.

228, p. 109811, 2023.

[28] 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.

[29] A. Holzinger, A. Saranti, C. Molnar, P. Biecek, and W. Samek, *“Explainable AI methods-a brief overview.” In International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers (pp. 13-38)*. Cham: Springer International Publishing, 2020.

[30] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why should i trust you?’ Explaining the predictions of any classifier,” *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. data Min.*, pp. 1135–1144, 2016.

[31] S. M. Lundberg and S. Lee, “A unified approach to interpreting model predictions,” *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 1–10, 2017.

[32] R. Qiao, Z. Wu, S. Gao, Q. Jiang, and X. Liu, “Towards inclusive underground public transportation: Gender differences on thermal comfort,” *Build. Environ.*, vol. 242, p. 110569, 2023.

[33] Y. Yang, Y. Yuan, Z. Han, and G. Liu, “Interpretability analysis for thermal sensation machine learning models: An exploration based on the SHAP approach,” *Indoor Air*, vol. 32, no. 2, p. e12984, 2022.

[34] H. Lan, H. Cynthia, and Z. Gou, “A machine learning led investigation to understand individual difference and the human-environment interactive effect on classroom thermal comfort Area Under the Receiver Operating Characteristic,” *Build. Environ.*, vol. 236, p. 110259, 2023.

[35] J. Baek, D. Yoon, H. Park, D. Minh, and S. Chang, “Vision-based personal thermal comfort prediction based on half-body thermal distribution,” *Build. Environ.*, vol. 228, p. 109877, 2023.

[36] S. M. Lundberg *et al.*, “Explainable machine-learning predictions for the prevention of hypoxaemia during surgery,” *Nat. Biomed. Eng.*, vol. 2, no. 10, pp. 749–760, 2018.

[37] H. Chen, A. B. Dincer, S. Lundberg, M. Kaeberlein, and S. Lee, “Interpretable machine learning prediction of all-cause mortality,” *Commun. Med.*, vol. 2, p. 125, 2022.

[38] S. Zhang, R. Yao, C. Du, E. Essah, and B. Li, “Analysis of outlier detection rules based on the ASHRAE global thermal comfort database,” *Build. Environ.*, vol. 234, p. 110155, 2023.

[39] E. F. J. Ring and K. Ammer, *“The technique of infrared imaging in medicine.”In Infrared Imaging: A casebook in clinical medicine (pp. 7-14)*. Bristol, UK: IoP Publishing, 2015.

[40] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, pp. 436–444, 2015.

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[42] L. Lü, M. Medo, C. Ho, Y. Zhang, and Z. Zhang, “Recommender systems,” *Phys. Rep.*, vol. 519, no. 1, pp. 1–49, 2012.

[43] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *arXiv Prepr. arXiv*, p. 1810.04805, 2018.

[44] S. M. Lundberg *et al.*, “From local explanations to global understanding with explainable AI for trees,” *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, 2020.

[45] N. Gao, W. Shao, M. Saiedur, J. Zhai, K. David, and F. D. Salim, “Transfer learning for thermal comfort prediction in multiple cities,” *Build. Environ.*, vol. 195, p. 107725, 2021.

[46] T. K. Ho, “Random Decision Forests,” *Proc. 3rd Int. Conf. Doc. Anal. Recognit.*, vol. 1, pp. 278–282, 1995.

[47] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, 1997.

[48] J. H. Friedman, “Greedy function approximation: a gradient boosting

machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001.

[49] R. G. Mantovani, A. L. D. Rossi, J. Vanschoren, B. Bischl, and A. C. P. L. F. Carvalho, "To tune or not to tune: recommending when to adjust SVM hyper-parameters via meta-learning," *2015 Int. Jt. Conf. neural networks*, pp. 1–8, 2015.

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

[51] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. New York: springer, 2009.

[52] S. Liu, S. Schiavon, H. P. Das, M. Jin, and C. J. Spanos, "Personal thermal comfort models with wearable sensors," *Build. Environ.*, vol. 162, p. 106281, 2019.

[53] 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.

[54] P. T *et al.*, "Predicting thermal pleasure experienced in dynamic environments from simulated cutaneous thermoreceptor activity," *Indoor Air*, vol. 31, no. 6, pp. 2266–2280, 2021.

[55] K. Kati, R. Li, and W. Zeiler, "Machine learning algorithms applied to a prediction of personal overall thermal comfort using skin temperatures and occupants' heating behavior," *Appl. Ergon.*, vol. 85, p. 103078, 2020.

[56] E. Young, P. Kastner, T. Dogan, A. Chokhachian, S. Mokhtar, and C. Reinhart, "Modeling outdoor thermal comfort along cycling routes at varying levels of physical accuracy to predict bike ridership in Cambridge , MA," *Build. Environ.*, vol. 208, p. 108577, 2022.

[57] K. N. Nkurikiyeyezu, Y. Suzuki, and G. F. Lopez, "Heart rate variability as a predictive biomarker of thermal comfort," *J. Ambient Intell. Humaniz.*

Comput., vol. 9, no. 5, pp. 1465–1477, 2018.

[58] H. Liu, H. Sun, H. Mo, and J. Liu, "Analysis and modeling of air conditioner usage behavior in residential buildings using monitoring data during hot and humid season," *Energy Build.*, vol. 250, p. 111297, 2021.

[59] A. Rysanek, R. Nuttall, and J. Mccarty, "Forecasting the impact of climate change on thermal comfort using a weighted ensemble of supervised learning models," *Build. Environ.*, vol. 190, p. 107522, 2021.

[60] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," *Proc. 22nd acm sigkdd Int. Conf. Knowl. Discov. data Min.*, pp. 785–794, 2016.

[61] 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.

[62] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 1–9, 2017.

[63] Y. Peng *et al.*, "Passenger overall comfort in high-speed railway environments based on EEG: Assessment and degradation mechanism," *Build. Environ.*, vol. 210, p. 108711, 2022.

[64] J. Wu, C. Shan, J. Hu, J. Sun, and A. Zhang, "Rapid establishment method of a personalized thermal comfort prediction model," *2019 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, pp. 3383–3386, 2019.

[65] R. Zhao *et al.*, "Building cooling load prediction based on lightgbm," *IFAC-PapersOnLine*, vol. 55, no. 11, pp. 114–119, 2022.

[66] D. Silver *et al.*, "Article Mastering the game of Go without human knowledge," *Nat. Publ. Gr.*, vol. 550, no. 7676, pp. 354–359, 2017.

[67] Y. Wu *et al.*, "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation," *arXiv Prepr. arXiv*, p. 1609.08144, 2016.

[68] O. Ronneberger *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. May, 2021.

[69] H. Chen, I. C. Covert, S. M. Lundberg, and S. Lee, "Algorithms to estimate

Shapley value feature attributions," *Nat. Mach. Intell.*, vol. 5, pp. 590–601, 2023.

[70] L. S. Shapley, "A value for n-person games," *Contrib. to Theory Games*, vol. 2, pp. 307–317, 1953.

[71] D. Janzing, L. Minorics, and P. Blobaum, "Feature relevance quantification in explainable AI: A causal problem," *Int. Conf. Artif. Intell. Stat. PMLR*, vol. 108, pp. 2907–2916, 2020.

[72] A. Shrikumar, P. Greenside, A. Y. Shcherbina, and A. Kundaje, "Not just a black box: Learning important features through propagating activation differences," *arXiv Prepr. arXiv*, p. 1605.01713, 2016.

[73] International Standard Organization, "ISO 7726: Ergonomics of the thermal environment - instruments for measuring physical quantities." 1998.

[74] A. Shajkofci, "Correction of human forehead temperature variations measured by non-contact infrared thermometer," *IEEE Sens. J.*, vol. 22, no. 17, pp. 16750–16755, 2021.

[75] Z. Obermeyer, J. K. Samra, and S. Mullainathan, "Individual differences in normal body temperature: longitudinal big data analysis of patient records," *BMJ*, vol. 359, p. j5468, 2017.

[76] M. Protsiv, C. Ley, J. Lankester, T. Hastie, and J. Parsonnet, "Decreasing human body temperature in the United States since the industrial revolution," *Elife*, vol. 9, p. e49555, 2020.

[77] M. Gurven *et al.*, "Rapidly declining body temperature in a tropical human population," *Sci. Adv.*, vol. 6, no. 44, p. eabc6599, 2020.

[78] D. D. Pascoe and G. Fisher, "Comparison of measuring sites for the assessment of body temperature," *Thermol. Int.*, vol. 19, no. 1, pp. 35–42, 2009.

[79] S. Erdogmus and F. Govsa, "Arterial features of inner canthus region: confirming the safety for the flap design," *J. Craniofac. Surg.*, vol. 17, no. 5, pp. 864–868, 2006.

[80] D. Li, C. C. Menassa, and V. R. K. Department, "Non-intrusive interpretation of human thermal comfort through analysis of facial infrared thermography,"

Energy Build., vol. 176, pp. 246–261, 2018.

[81] B. Salehi, A. Hamid, and M. Maerefat, "Intelligent models to predict the indoor thermal sensation and thermal demand in steady state based on occupants' skin temperature," *Build. Environ.*, vol. 169, p. 106579, 2020.

[82] D. Gavhed, T. Mäkinen, I. Holmér, and H. Rintamäki, "Face temperature and cardiorespiratory responses to wind in thermoneutral and cool subjects exposed to -10 C," *Eur. J. Appl. Physiol.*, vol. 83, pp. 449–456, 2000.

[83] M. S. Reuther, K. K. Briggs, B. L. Schumacher, K. Masuda, R. L. Sah, and D. Watson, "In vivo oxygen tension in human septal cartilage increases with age," *Laryngoscope*, vol. 122, no. 11, pp. 2407–2410, 2012.

[84] S. Yu, X. Sun, and Y. Liu, "Numerical analysis of the relationship between nasal structure and its function," *Sci. World J.*, vol. 2014, p. 581975, 2014.

[85] B. Tejedor, M. Casals, M. Gangolells, M. Macarulla, and N. Forcada, "Human comfort modelling for elderly people by infrared thermography: Evaluating the thermoregulation system responses in an indoor environment during winter," *Build. Environ.*, vol. 186, p. 107354, 2020.

[86] R. Nakanishi and K. Imai-matsumura, "Facial skin temperature decreases in infants with joyful expression," *Infant Behav. Dev.*, vol. 31, no. 1, pp. 137–144, 2008.

[87] E. Salazar-López *et al.*, "The mental and subjective skin: Emotion, empathy, feelings and thermography," *Conscious. Cogn.*, vol. 34, pp. 149–162, 2015.

[88] H. Wang, M. Kim, K. P. Normoyle, and D. Llano, "Thermal regulation of the brain—an anatomical and physiological review for clinical neuroscientists," *Front. Neurosci.*, vol. 9, pp. 1–6, 2016.

[89] S. Ariyaratnam and J. P. Rood, "Measurement of facial skin temperature," *J. Dent.*, vol. 18, no. 5, pp. 250–253, 1990.

[90] E. F. J. Ring, A. Jung, J. Zuber, P. Rutowski, B. Kalicki, and U. Bajwa, "Detecting fever in Polish children by infrared thermography," *Proc. 9th Int. Conf. Quant. Infrared Thermogr.*, vol. 2, no. 5, pp. 35–38, 2008.

[91] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic

imaging using ambient light," *Opt. Express*, vol. 16, no. 26, pp. 21434–21445, 2008.

[92] E. Arens, H. Zhang, and C. Huizenga, "Partial- and whole-body thermal sensation and comfort — Part I : Uniform environmental conditions," *J. Therm. Biol.*, vol. 31, no. 1–2, pp. 53–59, 2006.

[93] J. H. Choi and D. Yeom, "Study of data-driven thermal sensation prediction model as a function of local body skin temperatures in a built environment," *Build. Environ.*, vol. 121, pp. 130–147, 2017.

[94] B. Pavlin, G. Carabin, G. Pernigotto, A. Gasparella, and R. Vidoni, "An embedded mechatronic device for real-time monitoring and prediction of occupants' thermal comfort.," *ASME Int. Mech. Eng. Congr. Expo.*, vol. 52118, p. V08AT10A052, 2018.