Building and Environment 247 (2024) 111053

Contents lists available at ScienceDirect

Building and Environment

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

Calibrating subjective data biases and model predictive uncertainties in machine learning-based thermal perception predictions

Ruoxin Xiong [a], Ying Shi [a], Haoming Jing [b], Wei Liang [c], Yorie Nakahira [b], Pingbo Tang [a,*]

[a] *Department of Civil and Environmental Engineering, Carnegie Mellon University, Pittsburgh, 15213, PA, USA*
[b] *Department Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, 15213, PA, USA*
[c] *School of Architecture, Carnegie Mellon University, Pittsburgh, 15213, PA, USA*

ARTICLE INFO

Keywords:
Thermal perceptions
Subjective bias
Model calibration
Predictive uncertainties
Machine learning

ABSTRACT

Heating, Ventilation, and Air Conditioning (HVAC) systems in large-scale buildings often struggle to ensure satisfactory thermal comfort for diverse occupants while minimizing energy waste. Achieving this goal requires developing reliable prediction models that capture the changing and varied occupant thermal perceptions in different spaces. Despite their widespread use, many machine learning (ML) based prediction models suffer from subjective data biases and model predictive uncertainties, causing inaccurate estimation for occupant needs and leading to suboptimal building controls. The authors propose a data-model integration method that identifies and calibrates the inherent uncertainties of existing ML models in both data and model dimensions, ensuring reliable thermal perception predictions. This method introduces the Multidimensional Association Rule Mining (M-ARM) algorithm to identify biased human responses by exploring interrelationships among four perception metrics: thermal sensation, comfort, acceptability, and preference. Our method reveals significant performance enhancements in seven ML models, enhancing the F1-score by up to 5.53%. By leveraging reliability diagrams and Expected Calibration Error (ECE) scores, we also expose the models' vulnerability to miscalibration and the need for calibrated predictions. We further evaluate six calibration techniques (e.g., Platt Scaling and Isotonic Calibration) on these models and uncover their potential to enhance prediction reliability performance, highlighting a reduction of up to 80.66% in ECE scores. The authors also investigated the impacts of dataset sizes, classifiers, and calibration methods on the proposed method. Our research offers insight into creating robust data-driven strategies for thermal perception predictions, ultimately contributing to optimized occupant comfort and energy efficiency in buildings.

1. Introduction

Maintaining thermal comfort is essential for the well-being and productivity of individuals in shared workplaces. Despite the prevalent ASHRAE standard that requires delivering satisfactory thermal conditions for a minimum of 80% of occupants [1], heating, ventilation, and air conditioning (HVAC) systems, especially in large-scale commercial facilities often fall short of achieving these thermal comfort targets. For instance, a ten-year survey involving 52,980 occupants across 351 commercial buildings discovered that 43% of respondents expressed dissatisfaction with the temperature in their workspace [2]. This inadequacy can be attributed to the reliance on fixed temperature setpoints by HVAC systems, which often results in over-conditioning or under-conditioning of specific zones, ultimately leading to occupant discomfort [3]. Moreover, the discrepancy between fixed setpoints and occupants' thermal preferences contributes to energy waste [4]. Recent studies have shown that incorporating suitable comfort models for HVAC control can yield energy savings of more than 10% [5,6]. These findings emphasize the importance of developing comfort-driven HVAC systems that improve occupant thermal comfort while reducing energy consumption [7,8].

Reliable predictions of occupant thermal perceptions are crucial for implementing such comfort-driven HVAC controls. These prediction models should reveal the relationship between the ambient thermal environment (e.g., temperature and humidity), individual factors (e.g., clothing level and metabolic rate), and occupants' thermal perceptions. Two well-established models often used in building environment assessment standards are the Predicted Mean Vote-Percentage of Dissatisfied (PMV-PPD) model and the Adaptive Comfort model [1]. Despite their prevalence, these models have inherent drawbacks, most notably the absence of self-learning or self-correction abilities, which curtails

[*] Corresponding author.
E-mail addresses: ruoxinx@andrew.cmu.edu (R. Xiong), yingshi@andrew.cmu.edu (Y. Shi), haomingj@andrew.cmu.edu (H. Jing), weiliang@andrew.cmu.edu (W. Liang), ynakahir@andrew.cmu.edu (Y. Nakahira), ptang@andrew.cmu.edu (P. Tang).

https://doi.org/10.1016/j.buildenv.2023.111053
Received 9 August 2023; Received in revised form 13 November 2023; Accepted 16 November 2023
Available online 18 November 2023

their accuracy across diverse scenarios [9]. For example, the PMV-PDD model achieves an approximate accuracy of 34% when tested on the ASHRAE Global Thermal Comfort Database II, currently the most comprehensive database for thermal comfort [10]. To address these limitations, recent studies have explored various machine learning (ML) models, such as the Support Vector Machine (SVM) and Decision Tree (DT), to capture the complex relationships between the physical environment and occupants' subjective thermal perceptions [9,11]. ML models' ability to recognize non-linear interactions between variables from extensive data sets can significantly enhance the accuracy of thermal comfort predictions. Furthermore, their inherent capacity to adapt to changing conditions and self-correct over time makes them more suitable for practical applications [12,13].

Despite a growing interest in ML-based thermal comfort predictions, several significant research gaps remain. One gap lies in identifying the subjective bias on thermal comfort perceptions and understanding their impacts on prediction accuracy [14,15]. Distinct from objective measurements of the physical environment, thermal comfort labels are largely subjective. Consequently, these labels are susceptible to varying interpretations of "comfortable" thermal conditions among individuals or distinct groups [16,17]. On the other hand, subjective bias may arise from imprecise evaluations or flawed data documentation practices [12]. Typically, thermal perception labels are obtained from occupants through thermal environment surveys, which frequently pose questions such as, "What is your general thermal sensation?" Yet, the simplicity of a single-question scale often fails to ensure the accurate measurement of subjective perceptions [16]. Such subjective biases can introduce noise and uncertainty into ML models that depend on subjective labels for model training and validation, potentially resulting in suboptimal control strategies for building energy management systems [18,19]. Therefore, developing an approach to identify subjective data biases on thermal comfort perceptions is essential to successfully apply ML-based thermal comfort models in designing occupant-centric building HVAC controls.

Another significant gap is the insufficient evaluation and calibration of predictive uncertainties in existing thermal perception prediction models. Inherent uncertainties, such as model parameters, structure, and data quality, may result in these models producing uncertain predictions [20,21]. These uncertain predictions could, in turn, guide inappropriate HVAC control actions, causing both occupant discomfort and energy waste. In the context of practical applications for building HVAC systems, it becomes imperative to establish a well-calibrated prediction model, which can provide an indication of the likelihood of its predictions being either accurate or fallible [22]. The calibration of predictive uncertainties serves two fundamental purposes: (1) it improves the transparency of the decision-making process and fosters trust in ML models [23], and (2) it promotes informed HVAC control decisions by exploiting the model's predictive uncertainties and misprediction costs, thus managing desired control performance associated with varying conditions [24–26]. Despite these considerations, most existing studies still rely on the models' error rate (or accuracy) as the primary metric for their selection and deployment in real-world scenarios [12]. Unfortunately, high prediction accuracy does not necessarily assure commendable reliability performance [27]. Hence, there exists a need for a comprehensive evaluation of the reliability performance of existing ML-based prediction models for thermal perceptions.

This study addresses the above-mentioned limitations by developing a data-model integration method that improves the reliability of ML-based approaches for optimizing building occupant comfort. Considering the potential interrelationships among subjective perception metrics (such as thermal sensation, thermal comfort, thermal acceptability, and thermal preference), this study proposes a multidimensional association rule mining (M-ARM) method to capture biased human responses to thermal environments with quantitative performance. Based on calibrated subjective biases in thermal perceptions, the authors examine the impacts of these biases on the prediction accuracy of existing ML-based methods. The study also evaluates the reliability performance of these models using the reliability diagrams and Expected Calibration Error (ECE) [28]. Our analysis results underscore the miscalibration issues prevalent within current thermal perception models. To bolster the reliability of these models, this study examines the calibration performance of six prevalent model calibration methods in terms of ML methods and dataset sizes. The outcomes of this study provide insights into the advancement of ML-based strategies, with the aim of achieving reliable thermal perception predictions. These predictions can aid in reducing energy consumption and enhancing occupant well-being within buildings.

The organization of this paper is as follows. Section 2 reviews relevant research studies on uncertainties in ML-based thermal perception predictions. Section 3 introduces the proposed data-model integration framework designed to calibrate subjective data biases and model predictive uncertainties. Subsequently, Section 4 introduces the implementation details and evaluation metrics in this study. Section 5 evaluates the proposed framework using extensive experiments, analyzes the effect of the dataset size, and examines the impacts of subjective bias on the model reliability. Section 6 discusses the study's limitations and suggests potential directions for future research. Finally, Section 7 provides a summary of the research findings.

2. Literature review

This section reviews the related literature, focusing on three primary aspects: (1) uncertainty sources in ML-based thermal perception predictions, (2) subjective data biases in occupant thermal perceptions, and (3) prediction uncertainties of ML models in thermal perception predictions.

2.1. Uncertainty sources in ML-based thermal perception predictions

Thermal perceptions represent the personal state of satisfaction with the thermal environment and are assessed by subjective evaluation [1]. Maintaining a suitable level of thermal comfort for building occupants is a principal objective for HVAC design engineers. Achieving this goal requires the development of reliable thermal comfort models that can predict diverse occupant needs across different spaces.

Thermal perception prediction is typically formulated as a classification or regression task. The PMV-PPD model, for instance, is widely recognized and implemented across several standards, such as ANSI/ASHRAE Standard 55 [1], serving as an industry benchmark to determine acceptable thermal conditions in indoor environments. However, these methods may not effectively capture individual thermal comfort variances, which could result in inaccurate assessments of personalized thermal requirements and lead to suboptimal energy management within buildings. Recent studies have employed various ML algorithms [9,12], such as Logistic Regression (LR), DT, SVM, K-Nearest Neighbors (KNN), Naïve Bayes (NB), Multi-layer Perceptron (MLP), and Ensemble learning algorithms, including Gradient Boosting Machine (GBM), Adaptive Boosting (AdaBoost), and Random Forest (RF), for predicting thermal perceptions. However, the application of ML methods introduces prediction uncertainties pertaining to their outputs, which need to be addressed as the decision-making of these systems could potentially impact human well-being and building energy. To facilitate informed decision-making in uncertain scenarios and potential safety implications, these models should provide a guaranteed functionality and estimate the probability of their predictions falling outside the desired range.

Fig. 1 presents the uncertainty sources involved in characterizing the dynamics of the Human–Building–Environment (HBE) systems, which encompasses data biases and model uncertainties. In real-world scenarios, ML methods require extensive data for modeling the relationship between physical characteristics (e.g., air temperature and

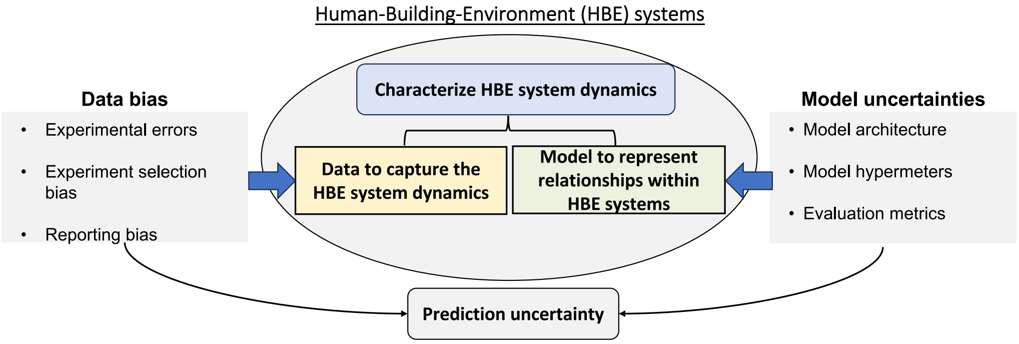

Human-Building-Environment (HBE) systems
Data bias
• Experimental errors
• Experiment selection bias
• Reporting bias
Characterize HBE system dynamics
Data to capture the HBE system dynamics
Model to represent relationships within HBE systems
Model uncertainties
• Model architecture
• Model hypermeters
• Evaluation metrics
Prediction uncertainty

Fig. 1. Uncertainty sources for characterizing the Human–Building–Environment (HBE) system dynamics.

humidity), personal factors (e.g., metabolic rate and clothing insulation), and their associated thermal perceptions [9,12,29]. However, all collected data types, encompassing sensor-based and human inputs, have inherent accuracy limitations and could be subject to numerous data quality and measurement reliability issues, including experiment errors, experimental selection bias, and reporting bias [16,30]. Selection bias can manifest during the delineation of the study population. Once this population is identified, selection bias may arise if the recruitment and enrollment criteria inherently vary between distinct study groups. Experimental errors typically relate to procedural mishaps, such as variations in ambient conditions or experimental sequences, observed during experimental trials. Reporting bias becomes evident post-experiment when participants offer inconsistent feedback or when researchers omit to present precise results. Consequently, the inherent uncertainty in an ML model's predictions is closely tied to the quality of its input data [31].

Prediction uncertainty of ML-based thermal perception models primarily arises from two main sources: aleatoric and epistemic uncertainties [32,33]. Aleatoric (data) uncertainty pertains to the inherent noise found in observations, while epistemic uncertainty is attributed to insufficient knowledge needed to accurately define model parameters. Aleatoric uncertainty, inherent to data variability, necessitates methodological refinements such as enhanced data quality and strategic feature optimization [34]. Conversely, epistemic (model) uncertainty, indicative of model-specific insufficiencies, demands advanced modeling architectures, rigorous model calibration, increased dataset size, and the assimilation of domain-specific insights [20]. These strategies fortify model robustness and elevate predictive performance in analytical frameworks.

Unfortunately, most existing studies focus on evaluating the model's discriminative ability (i.e., the capacity to generate accurate predictions) [12]. Metrics typically employed for this purpose include accuracy, precision, recall, mean squared error (MSE), and mean absolute error (MAE). Limited studies have thoroughly examined the reliability performance of these ML-based thermal perception prediction models. Model reliability refers to the correlation between a model's predicted probability (ranging from 0% to 100%) and the observed probability, assessed using metrics such as calibration diagrams and ECE [28]. The high prediction accuracy does not guarantee high model reliability performance [35]. Some ML models may display poor calibration, distorting predicted probabilities and leading to incorrect control actions [36], ultimately causing occupant discomfort and energy inefficiency.

2.2. Subjective data biases in thermal perceptions

Unlike objective measurements of the physical environment, self-reported thermal perceptions are largely subjective, making the associated biases harder to detect and analyze [19,37]. These subjective thermal perceptions of varying conditions, serving as labels, are crucial for training and evaluation purposes in ML-based methods. Typically, a diverse group of participants is recruited to report their perceptions and satisfaction levels regarding their thermal environment. The most commonly employed thermal perception metrics include thermal sensation (7-point scale), thermal preference (3-point scale), thermal acceptability (2-point scale), and thermal comfort (6-point scale). Table 1 provides an overview of these four thermal comfort responses, which are frequently combined to explore different aspects of occupant perceptions.

Subjective evaluations of indoor environments are inherently affected by inter- and intra-individual differences, such as biological, psychological, and background factors [14,38]. A notable study by Schweiker et al. [17] examined subjective thermal assessments across 26 countries, finding that contextual differences, such as climate and language, influenced thermal comfort perceptions. This work brought into question the reliability of these scales as predictive measures of occupant comfort. Furthermore, subjective data biases could result from imprecise evaluations, non-representative population sampling, or inaccurate data documentation [12,16], which may, in turn, lead to inaccurate predictions of thermal perceptions among building occupants.

Detecting subjective bias poses a considerable challenge, given that subjective comfort votes do not offer ground truth information. Several studies have explored distance-based and stochastic-based techniques to identify and mitigate subjective bias in thermal perceptions [19]. For example, Zhang et al. [39] applied three stochastic approaches, including the 3-Sigma, Boxplot, and Hampel rules, to filter anomalous data. Although these techniques have demonstrated a certain degree of efficacy, they often do not sufficiently accommodate intricate and subjective aspects of thermal perceptions. These methods assume that similar conditions prompt similar thermal responses and identify outliers predicated on input features, simplifying the inherent variability and complexity associated with subjective perceptions. Furthermore, existing methods do not sufficiently account for the intertwined nature of various thermal perception metrics, instead considering each metric independently. These approaches fail to capture the multifaceted, interconnected nature of thermal comfort perception, thus limiting their ability to fully understand and predict this multidimensional phenomenon. To address these limitations, this study examines the association patterns within subjective perceptions while intentionally avoiding modeling relationships between input features and labels. By incorporating potential interrelationships among subjective perceptions, this study aims to provide an in-depth understanding of subjective thermal evaluations and robust detection mechanisms for potential response biases.

2.3. Uncertainty quantification and calibration for ML-based thermal perception prediction models

Recent research has utilized a range of ML algorithms for predicting thermal perceptions, including LR, DT, SVM, KNN, NB, and MLP, as well as ensemble techniques like AdaBoost and RF [9,12]. Uncertainty

Table 1
Summary of subjective thermal perception metrics in thermal comfort surveys.

Thermal metrics	Subjective rating scale
Thermal sensation	hot (+3), warm (+2), slightly warm (+1), neutral (0), slightly cool (−1), cool (−2), cold (−3)
Thermal preference	cooler, no change, warmer
Thermal acceptability	unacceptable (0), acceptable (1)
Thermal comfort	very uncomfortable (1), uncomfortable (2), slightly uncomfortable (3), slightly comfortable (4), comfortable (5), very comfortable (6)

quantification (UQ) refers to estimating the uncertainty level in those models' predictions. This can include both data and model uncertainty, with the latter arising from the model's lack of knowledge. Prevalent UQ techniques for ML models include [36]:

- Ensemble methods: These techniques harness the diversity within the ensemble to gauge uncertainty by employing multiple learners. An exemplar is the RF classifier, where uncertainty is inferred from the distribution of tree votes.
- Bayesian approaches: These methodologies combine prior beliefs regarding model parameters with the likelihood function derived from observed data, providing a posterior distribution that reflects model uncertainty.
- Probabilistic models: These inherently yield probabilistic results. For instance, LR employs the sigmoid function to estimate the probability of a sample belonging to a particular class. MLP structures generate probabilities based on activation functions like sigmoid (binary tasks) or softmax (multiclass tasks) in their output layer.

In real-world HVAC control systems, calibrated model uncertainties are crucial in enhancing model interpretability and fostering trust among users, given that humans are inherently intuitive towards probabilities [23]. In applications, ML methods are expected to be capable of indicating the level of confidence in their predictions for uncertainty-informed building thermal regulation [22]. This means that the probability associated with the predicted perceptions should directly correlate with the likelihood of their correctness. Further, the integration of predictive uncertainties can assist in optimizing the control actions of HVAC systems by quantifying uncertainties and refraining from decision-making in the face of significant uncertainties [24,25]. By understanding and addressing predictive uncertainties, HVAC systems can make more informed decisions for thermal controls and ultimately improve thermal comfort for occupants while reducing energy consumption. For example, Chao et al. [24] introduced a hybrid control strategy that determined the HVAC reference temperature point by fusing the setpoint determined by the model and the setpoint recommended by the experts based on the prediction confidence level. This strategy showed more efficiency than the conventional controllers in terms of both comfort level and energy-saving. Maasoumy et al. [25] characterized the impact of model uncertainty on model-based controllers and developed a methodology for selecting a controller type (i.e., Robust Model Predictive Control (RMPC), Model Predictive Controllers (MPC), and Rule Based Control (RBC)) as a function of building model uncertainty.

Unfortunately, algorithmic predictions may not always produce valid probability estimates aligned with the underlying true probabilities, limiting their utility as reliable uncertainty quantifiers [36,40]. Such discrepancies often arise from model miscalibration [40]. Given the prevalent application of ML-based models in thermal perception predictions, this study examines model reliability in quantifying prediction uncertainties and seeks to recalibrate uncertainties in instances of model prediction deficiencies.

Calibration involves techniques applied after the primary model training (post-hoc) to assess and improve how well a model's predicted probabilities align with the true event probabilities and to make corrections for miscalibrated models. These methods adjust the model's predictions, thereby enhancing calibration accuracy. Previous studies have developed several calibration techniques for binary classifiers, including Platt scaling [41], isotonic calibration [35], and beta calibration [42]. Extensions of the above approaches include Bayesian Binning into Quantiles (BBQ) [43], which performs Bayesian averaging of multiple calibration maps obtained with equal-frequency binning. For multiclass calibration, the problem has been approached by fragmenting it into one-vs-all binary calibration tasks [35], one for each class. Recently, native multiclass calibration methods were also introduced, including temperature scaling [28], which can be seen as a multiclass extension of the Platt scaling. In light of the growing utilization of ML models for thermal perception prediction, the authors will evaluate the efficacy of state-of-the-art calibration methods for these ML-based models. By incorporating these calibration techniques, our goal is to increase the reliability performance of model predictions and ultimately provide trustworthy thermal perception predictions for practical applications.

3. Methodology

This section introduces the proposed framework that calibrates subjective data biases and model predictive uncertainties, as illustrated in Fig. 2. Unlike the existing ML workflows for thermal perception predictions (marked by the dashed arrow in Fig. 2), the authors present a data-model integration method (highlighted by the solid arrow in Fig. 2) that aims to identify and calibrate the inherent uncertainties of ML models across data and model dimensions. Specifically, the developed M-ARM algorithm aims to identify and understand biased human responses by exploring anomalous association patterns among inter-related subjective perception metrics: thermal sensation, thermal comfort, thermal acceptability, and thermal preference. Using the reliability diagrams and ECE score, this study assesses the reliability performance of existing ML-based thermal perception prediction models. Furthermore, the authors incorporate state-of-the-art calibration methods to reduce prediction uncertainties. Our overarching objective is to provide more reliable thermal perception predictions, which can effectively guide HVAC control practices.

3.1. Data collection and preprocessing

The ASHRAE Global Thermal Comfort Database II (Comfort Database II) [44] serves as our benchmark for exploring the impact of subjective data biases and model predictive uncertainties on ML-based thermal perception prediction models. The selection of this database is primarily due to its considerable sample size and standardized data format. The Comfort Database II is a culmination of systematic data collection and harmonization from thermal comfort field studies conducted globally over the past two decades. It encapsulates objective measurements of indoor environments alongside their corresponding "right-here-right-now" subjective evaluations, sourced from 160 buildings worldwide [44]. Including additional data from the original ASHRAE RP-884 database [45], the Comfort Database II consists of 109,033 entries. Within the database are four subjective thermal metrics: thermal sensation (7-point), thermal comfort (6-point), thermal preference (3-point), and thermal acceptability (2-point).

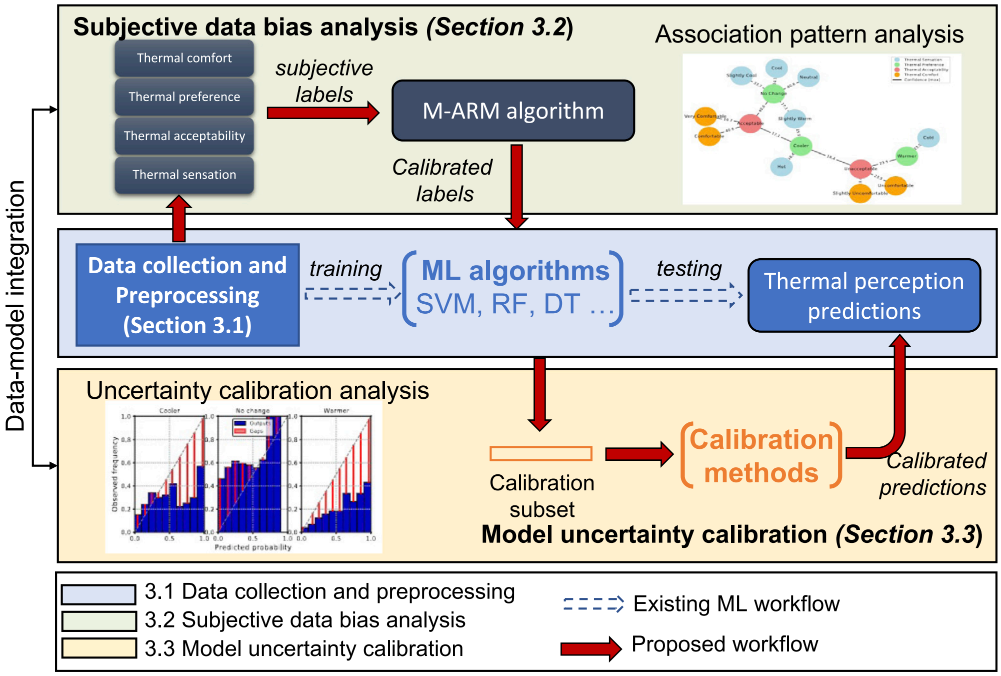


Fig. 2. Overview of the proposed data-model integration framework for supporting reliable thermal perception predictions.

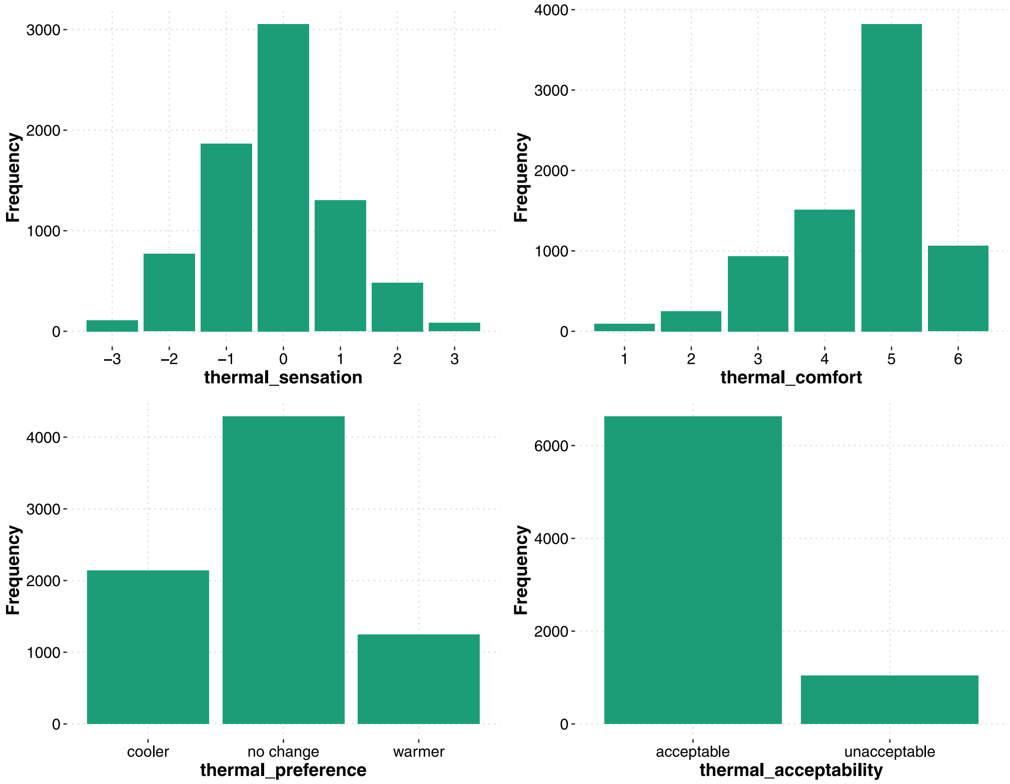


Fig. 3. Distributions of the four subjective perception metrics in the ASHRAE Global Thermal Comfort Database II post data preprocessing. The subjective perception metrics include thermal sensation (7-point scale), thermal comfort (6-point scale), thermal preference (3-point scale), and thermal acceptability (2-point scale).

To refine the data in the Comfort Database II, this study excluded any data entries that lacked records of the aforementioned six input features and four subjective labels. Following these data cleaning procedures, the processed database contains around 7,600 data records. Fig. 3 presents the post-processing distributions of the four subjective perception metrics in the Comfort Database II. It reveals a notable deficiency in certain perception classes within the dataset, such as the +3 (hot) votes. In alignment with previous research [46], this study discarded label categories that contain fewer than 100 data records from the Comfort Database II to maintain robustness and validity in data analysis.

In accordance with previous ML-based research [12], the authors focus on six input features, specifically clothing insulation (clo), metabolic rate (met), air temperature (ta), relative humidity (rh), mean radiant

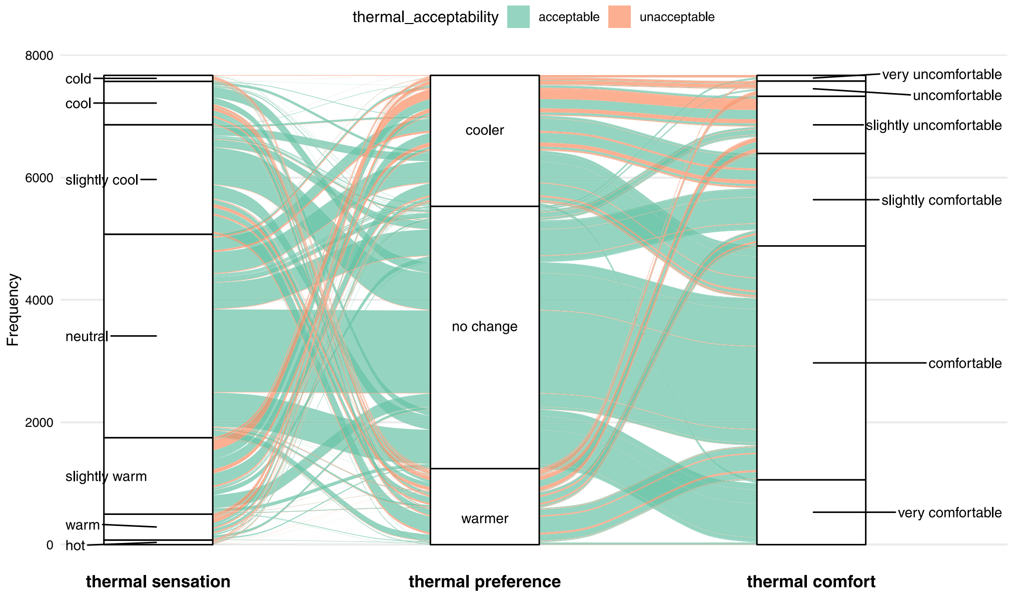


Fig. 4. Illustrations of the relationship between the four subjective thermal metrics in the Comfort Database II. Thermal comfort: from cold (−3) to hot (+3). Thermal preference: cooler, no change, and warmer. Thermal comfort: from very uncomfortable (1) to very comfortable (6). Thermal acceptability: acceptable and unacceptable.

temperature (tr), air velocity (vel). Moreover, this study has integrated four subjective perception labels, including thermal sensation (TSV), thermal preference (TPV), thermal acceptability (TAV), and thermal comfort (TCV). The thermal perception prediction models can be defined as a mapping f:

$$y = f(ta, rh, tr, vel, clo, met) \tag{1}$$

where y is the subjective thermal perceptions.

3.2. Detection of subjective biases based on the multidimensional association rule mining

Obtaining subjective perceptions from the occupants using a single-question scale can hardly guarantee an accurate measurement. However, potential interconnections among thermal perception metrics offer opportunities for cross-validation, thus enhancing our understanding of human responses to thermal environments. For example, thermal sensation reflects the subjective perception of the thermal environment and directly influences thermal comfort. A neutral thermal sensation generally indicates a state of high thermal comfort, while deviations from neutral may lead to discomfort. Moreover, thermal comfort influences thermal acceptability, as occupants who are comfortable are more likely to consider the environment acceptable.

To effectively detect subjective biases in thermal perceptions, this study examines the M-ARM method that simultaneously considers the intricate interrelationships of multiple subjective metrics. Our primary assumption is that frequent association patterns, which represent common thermal perceptions among occupants, are indicative of typical responses in a given thermal environment. Consequently, data points that deviate significantly from these patterns are hypothesized to potentially represent subjective biases in data collection from the statistical point of view. However, this methodology carries inherent limitations. Principally, the proposed method may inadvertently exclude valid but non-typical or “hard-to-predict” data points. These rare responses, while potentially representing actual thermal experiences, may not conform to the most frequently observed patterns and, as a result, could be filtered out. This poses a risk of narrowing the diversity and richness of our dataset, potentially overlooking some unique thermal perception experiences. Additionally, there is a possibility that the enhanced performance of model predictions, observed post-filtering, could be attributed to the removal of these non-typical data points rather than a true reflection of overall trends in thermal perceptions. This raises concerns about the method’s ability to capture the full characteristics of thermal experiences, potentially skewing results towards more diverse responses. Thus, considering these limitations is essential to interpreting the results, particularly in terms of the range and diversity of thermal perceptions captured in our analysis.

The proposed M-ARM method consists of three main steps: generating candidate association rules, computing support and confidence measures, and filtering anomalous association rules for thermal perceptions. To mitigate the risk of losing insightful data, our approach includes a supplementary analysis of filtered data points, preserving the comprehensiveness of the dataset.

3.2.1. Generating candidate association rules for thermal perceptions

The correlation outcomes illustrated in Fig. 4 expose potential subjective biases intertwined with thermal perceptions. For instance, a group of occupants demonstrates a rather paradoxical array of responses: they report feeling “cold” (thermal sensation), show a preference for “cooler” conditions (thermal preference), and express “very comfortable” (thermal comfort), all while finding these concurrent experiences “acceptable” (thermal acceptability). Such inconsistency and conflicting thermal responses infuse a degree of noise and uncertainty into ML models, which could obscure clear pattern recognition and predictive accuracy.

The generation of candidate rules involves combining multiple subjective perception metrics as antecedents, with a target metric serving as the consequent. This step aims to uncover potential associations between the distinct thermal perception metrics. An association rule R_i is a logical expression structured as follows:

$$R_i : A_1 \wedge A_2 \wedge \cdots \wedge A_n \Rightarrow C \tag{2}$$

where $A_j (j \in 1, 2, \ldots, n)$ represents the antecedent metrics, and C denotes the consequent metric.

In this study, each metric is denoted by a pair in the format `<attribute, value>`. For instance, thermal sensation can be represented as `<sensation, neutral>`. To compute the anomalous association rules, we designate the targeted metrics as the consequent metrics C, while the other three serve as antecedent metrics A. As an illustration, when the goal is to predict thermal sensation, a multidimensional association rule R_i can be expressed as: `<preference, no changes>` ∧ `<acceptability, acceptable>` ∧ `<comfort, comfortable>`⇒ `<sensation, neutral>`.

3.2.2. Computing support and confidence measures for association rules

This study employs one-hot encoding to transform multilevel metrics into binary problems [47]. The transformation facilitates the interpretation of relationships between distinct levels of categorical data. Each distinct value of a quantitative metric is mapped to a boolean attribute within the pair <`attribute`, `value`>, e.g., <`no changes`, 1 >.

Support (Sup) and confidence ($Conf$) measures for each candidate rule are computed to quantify their significance. The Sup of a rule represents the proportion of perception votes that include both the antecedent and the consequent of the rule. Rules with extremely low Sup might be considered anomalous due to their infrequency and the potential absence of significant patterns in the data. The Sup of a rule ($A \Rightarrow C$) is defined as follows:

$$Sup(A \Rightarrow C) = P(A \cap C) \tag{3}$$

Confidence ($Conf$) is the conditional probability $P(C|A)$ of a subjective vote containing C and A. Rules with very low $Conf$ may be regarded as anomalous because they imply a weak association between the antecedent and the consequent, possibly due to noise or data errors. The $Conf$ of a rule ($A \Rightarrow C$) is defined as follows:

$$Conf(A \Rightarrow C) = \frac{P(A \cap C)}{P(A)} \tag{4}$$

3.2.3. Filtering anomalous association rules for thermal perceptions

To effectively generate association rules, this study employs the Apriori algorithm [48], a widely used method in ARM problems. The authors define a very low Sup threshold (i.e., 0.0001) to avoid missing rare rules [49].

Since frequent association patterns represent the common thermal perceptions of occupants, data points diverging from these patterns may indicate potential data biases. For instance, a reported thermal response that aligns with more frequent patterns is less likely to be considered an outlier, as it echoes the consistent perceptions of occupants in the dataset. Based on the predefined $Conf$ threshold, the authors filter out inconsistent and contradictory association patterns, retaining only those that meet these criteria as common association rules.

Additionally, this study used the Chi-square test [50] to determine if the consequent metrics are independent of the antecedent metrics at the selected significance level. For association rules, the null hypothesis typically posits that the items in the rule are independent. If $\chi^2(A \Rightarrow C) > t_\alpha(n)$, it suggests that the observed co-occurrence of the rule might be the potential subjective bias. The implementation of the proposed method is described in Algorithm 1.

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{5}$$

where O_{ij} is the observed frequency, and E_{ij} is the expected frequency for each rule.

3.3. Evaluating and calibrating model predictive uncertainty of thermal perception predictions

For any input instances $\mathbf{x} \in \mathcal{X}$, the probabilistic ML classifier outputs a class label $i \in \{1, \dots, k\}$ and an associated probability (also known as confidence) c, where $c = \max(f(\mathbf{x}))$ and $i = \arg\max f(\mathbf{x})$. Here, the confidence c for a given input $\mathbf{x}$ is the maximum value of the function $f(\mathbf{x})$. If $f(\mathbf{x})$ produces a probability distribution over classes (i.e., $f(\mathbf{x}) = [p_1, p_2, \dots, p_k]$ where each p_i is the probability of class i), then c is indeed the maximum probability among all classes. The classifier assigns to the input instance $\mathbf{x}$ the class label i with the highest probability.

A classifier is considered confidence-calibrated if the predicted probability matches the observed accuracy for the most likely class to be predicted. For example, given 100 predictions, each with a predicted probability of 0.8, we expect that 80 should be correctly classified. More formally, a probabilistic classifier is confidence-calibrated if for any confidence level $c \in [0, 1]$

Algorithm 1 Detect anomalous associated thermal perception patterns with Chi-square filtering

```
Input: Comfort Database II $D$, minimum support threshold $min\_sup$,
    confidence threshold $conf$, significance level $\alpha$
Output: Anomalous associated thermal perception patterns
Initialize candidate set $C_1$ by scanning $D$ and counting support for individual metrics
$k \leftarrow 1$
while new associated thermal perception patterns are found do
    Generate associated thermal perception patterns $L_k$ from candidate set $C_k$, where $Sup(X) \geq min\_sup$ for all $X \in L_k$
    Generate candidate set $C_{k+1}$ from $L_k$ using the Apriori property [48]:
        Join:  Pairwise join $L_k$ with itself to generate $(k+1)$-item candidate patterns
        Prune: Remove candidates with any $(k+1)-1$ subset not in $L_k$
    Calculate the support and confidence of each candidate in $C_{k+1}$ by scanning the Comfort Database II $D$
    Increment $k$: $k \leftarrow k+1$
end while
for each association rule $A \Rightarrow C$ in $\bigcup_{i=1}^{k} L_i$ do
    Calculate the observed frequency $O_{ij}$ and the expected frequency $E_{ij}$
    Calculate $\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$
    if $\chi^2(A \Rightarrow C) > t_\alpha(n)$ then
        Mark the association rule $A \Rightarrow C$ for further examination
    end if
end for
return Anomalous thermal perception patterns from $\bigcup_{i=1}^{k} L_i$
```

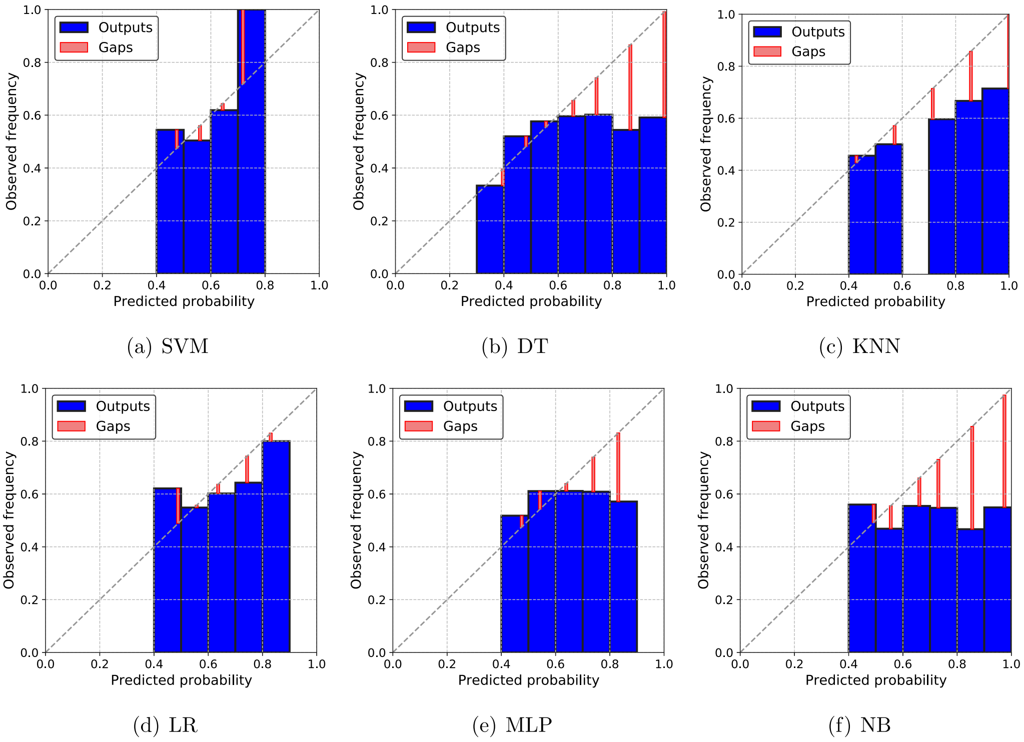
$$P(Y = i | f_i(\mathbf{x}) = c) = c \tag{6}$$

Here, if the classifier gives a probability c to class i for input $\mathbf{x}$, then the actual likelihood that $\mathbf{x}$ belongs to class i should indeed be c. So, for example, if the classifier is 70% confident that $\mathbf{x}$ belongs to class i, then, over many such predictions, around 70% of them should be correct.

3.3.1. Assessing model calibration performance for existing ML-based thermal perception prediction models

Reliability diagrams can assess model calibration performance by categorizing model predictions into bins based on the confidence score associated with each predicted class. Within each bin, the average confidence and accuracy are computed and plotted. Ideally, a well-calibrated model will exhibit points close to the diagonal line in the reliability diagram, indicating equal accuracy and confidence.

Fig. 5 presents the reliability diagrams of various ML models for predicting occupant thermal perceptions using the Comfort Database II, with predicted labels and confidence associated with the highest-scoring class. Deviations from the perfect diagonal line, denoted as “gaps”, indicate miscalibration. Larger gaps correspond to a more significant miscalibration. Overconfidence arises when the model’s prediction confidence surpasses actual accuracy (i.e., the red bar falls below the diagonal), leading to more false positives. Conversely, underconfidence occurs when the model’s prediction confidence is below actual accuracy (i.e., the red bar goes above the diagonal), resulting in more false negatives. The results reveal that most ML-based thermal perception prediction models experience some degree of miscalibration and are prone to poor calibration. Uncalibrated predicted probabilities of occupant thermal needs might guide incorrect HVAC control actions, causing occupant discomfort and energy waste.



(a) SVM (b) DT (c) KNN (d) LR (e) MLP (f) NB

Fig. 5. Confidence-reliability diagrams for un-calibrated ML methods of thermal preference predictions on the dataset. Note that bins corresponding to confidences less than 1/10 will be empty. SVM: Support Vector Machines; DT: Decision Trees; KNN: K-Nearest Neighbors; LR: Logistic Regression; MLP: Multi-Layer Perceptron; NB: Naïve Bayes. For more detailed implementation information, refer to Section 4.

3.3.2. Calibrating model predictive uncertainties

The observed miscalibration performance of ML methods (see Fig. 5) highlights the need to improve their prediction reliability for guiding reliable HVAC control actions.

Classifier calibration aims to adjust the classifier's output to more closely reflect the true probability of correctness. This *post-hoc* process typically involves transforming an uncalibrated probabilistic classifier using a hold-out validation dataset to learn a calibration map $u : \mathbb{R}^k \rightarrow \mathbb{R}^k$ [42]. The calibration methods investigated in this study include Platt scaling [41], isotonic calibration [35], Beta calibration [42], Bayesian Binning into Quantiles (BBQ) [43], Temperature scaling (*Temp.*) [51], and Gaussian Process calibration (*GPcalib*) [21].

Given the multi-class nature of certain thermal perception metrics (e.g., thermal preference, thermal sensation, and thermal comfort), this study uses a one-vs-all approach [35] to extend binary calibration methods (e.g., Platt scaling [41] and isotonic calibration [35]) to multiclass problems. The calibrated probabilities for each class are predicted separately, and post-processing is performed to normalize their predictions. The calibration processes for each method are as follows:

- Platt scaling [41] fits a sigmoid function to the model scores obtained from the calibration set. The label predicted by the underlying model is treated as the positive class, whereas all other labels are treated as the negative class. Given an uncalibrated probability estimation score c, the predictive function is:

$$u(c; w, b) = (1 + \exp(-wc - b))^{-1} \tag{7}$$

  where $w, b \in \mathbb{R}$ are scaling parameters optimized via maximum likelihood in the validation set.
- Isotonic regression [52] fits the piecewise isotonic (monotonically increasing) function I to transform uncalibrated outputs:

$$u(c; I) = I(c) + \epsilon \tag{8}$$

  where ϵ is the model bias. This function aims to minimize the square loss between the predicted and observed probability.
- Beta calibration [42] defines a family of calibration maps based on the likelihood ratio between two Beta distributions. The predictive function is expressed as:

$$u(c; w_1, w_2, b) = (1 + \exp(-w_1 \cdot \ln c - w_2 \cdot \ln(1 - c) - b))^{-1} \tag{9}$$

  where $w_1, w_2, b \in \mathbb{R}$ are the scaling parameters.
- Bayesian Binning into Quantiles (BBQ) [43] incorporates multiple binning models and uses a Bayesian score function S to weigh the accuracy in each bin. The predictive function is expressed as:

$$u(c; S) = \frac{S(m_i) \cdot P_{m_i}(c)}{\sum_{j=1}^{N} S(m_j)} \tag{10}$$

  where N represents the total number of binning models, and $P_{m_i}(c)$ is the estimated probability by the binning model m_i for the uncalibrated classifier output c.
- Temperature scaling (*Temp.*) [51], an extension of Platt scaling, utilizes a single scalar parameter, known as "temperature" ($t \in \mathbb{R}$), for all classes. Let $\mathbf{p} = f(\mathbf{x})$ represent the predicted probability vector for a classifier f. The predictive function is formulated as:

$$u_j(\mathbf{p}; t) = \frac{\exp(-t \cdot \text{logit} p_j)}{\sum_{j=1}^{k} \exp(-t \cdot \text{logit} p_j)}, \forall j \in 1, \dots, k \tag{11}$$

- Gaussian Process calibration (*GPcalib*) [21] models the relationship between predicted and true class probabilities through a Gaussian process. This method defines a Gaussian process prior to the latent function as $g(p_j) = \mathcal{GP}(p_j; \mu, K)$, with μ as the mean function and K as the kernel. The predictive function is given by:

$$u_j(\mathbf{p}; \mu, K) = \frac{\exp(g(p_j))}{\sum_{j=1}^{k} \exp(g(p_j))}, \forall j \in 1, \dots, k \tag{12}$$

The calibration performance, visualized by the bin distributions in the reliability diagram), is subject to several influential factors:

- Classifier characteristics: A classifier might generate probabilities concentrated around specific values. For example, a classifier

Table 2
Parameter sets of grid search for the ML algorithms.

Algorithms	Parameter sets	Parameter range
SVM	Regularization parameter Kernel Kernel coefficient	0.1, 1, 10 Linear, RBF scale, 1/n_features
MLP	Hidden layer sizes Alpha	50, 100 0.0001, 0.001
RF	Estimators Max depth Min samples split	100, 500, 1000 none, 10, 30 2, 5, 10
DT	Criterion Max depth Min samples split	Gini, Entropy, Log loss none, 10, 30 2, 5, 10
KNN	Neighbors Weights Algorithms	3, 5, 7 Uniform, distance BallTree, KDTree, brute-force search
Gaussian NB	Variance of smoothing	$1e{-}9$
LR	Regularization parameter Penalty Solver	0.1, 1, 10 L1, L2, both L1 and L2, none “newton-cg”, “lbfgs”, “sag”, “saga”

predominantly predicting with extreme confidences (e.g., close to 0 or 1) may leave several intermediate bins unoccupied.

- Calibration method characteristics: Calibration methods may exhibit varying decisiveness levels. Some calibration methods might be more conservative, adjusting probabilities closer to a midpoint. In contrast, other methods might be more assertive, pushing predictions towards the extremes. This results in distinct bin distribution patterns in their corresponding reliability diagrams.
- Data characteristics: The nature of calibration data, such as dataset size and distribution, affects the calibration performance.

4. Experimental setup

This study employs seven widely used ML algorithms: SVM, MLP, RF, DT, KNN, NB, and LR, for predicting thermal perceptions. A 5-fold nested cross-validation approach is employed, wherein the outer loop manages dataset splits for testing, and the inner loop conducts a grid search for hyperparameter tuning on the training data. Table 2 presents the parameter sets and ranges utilized for grid search optimization of the ML algorithms.

As the actual thermal responses are unknown, this study utilizes an indirect indicator to evaluate the effectiveness of the proposed method. Following previous studies [19], the authors compare the prediction accuracy performance before and after filtering biased data. Upon the removal of biased data, this study anticipates enhanced accuracy performance from the ML-based classifiers.

Due to the imbalanced classes in the Comfort Database II (as illustrated in Fig. 3), this study chose the weighted F1 score to evaluate the proposed algorithm’s effectiveness by following previous studies [53, 54]. The F1 score for class i is calculated as follows:

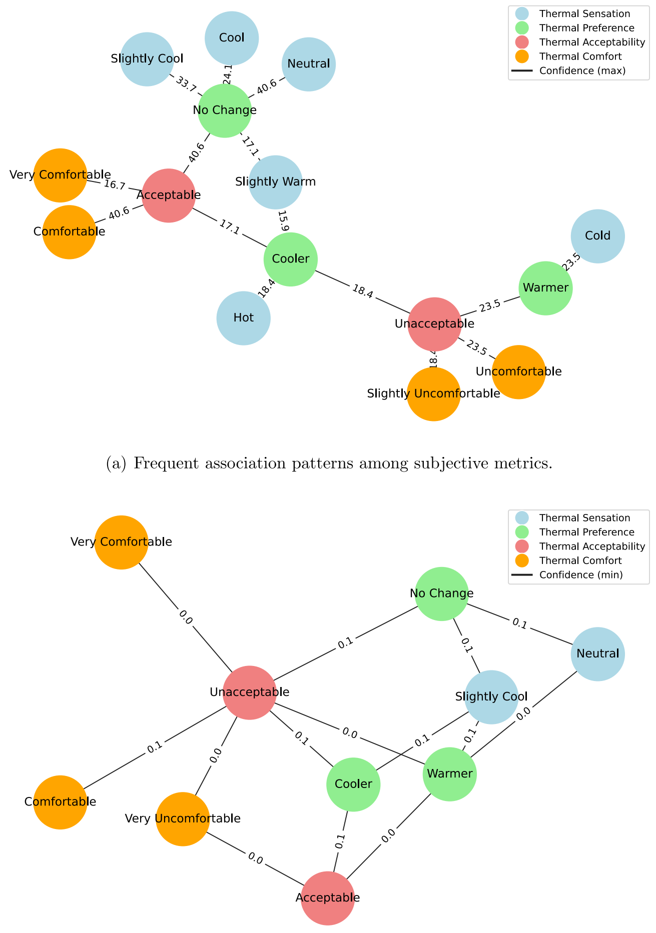
$$\mathrm{F1}_i = \frac{2 \times \mathrm{Precision}_i \times \mathrm{Recall}_i}{\mathrm{Precision}_i + \mathrm{Recall}_i} \tag{13}$$

where $\mathrm{Precision}_i = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FP}_i}$ and $\mathrm{Recall}_i = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i}$. Here, TP_i, FP_i, and FN_i represent the number of True Positives, False Positives, and False Negatives for class i, respectively.

The weighted F1 score is the weighted average of F1 scores across all classes, computed as follows:

$$\text{Weighted F1} = \sum_{i=1}^{k} w_i \times \mathrm{F1}_i \tag{14}$$

where w_i is the weight assigned to class i, computed as the proportion of samples in class i.



(a) Frequent association patterns among subjective metrics.

(b) Infrequent association patterns among subjective metrics.

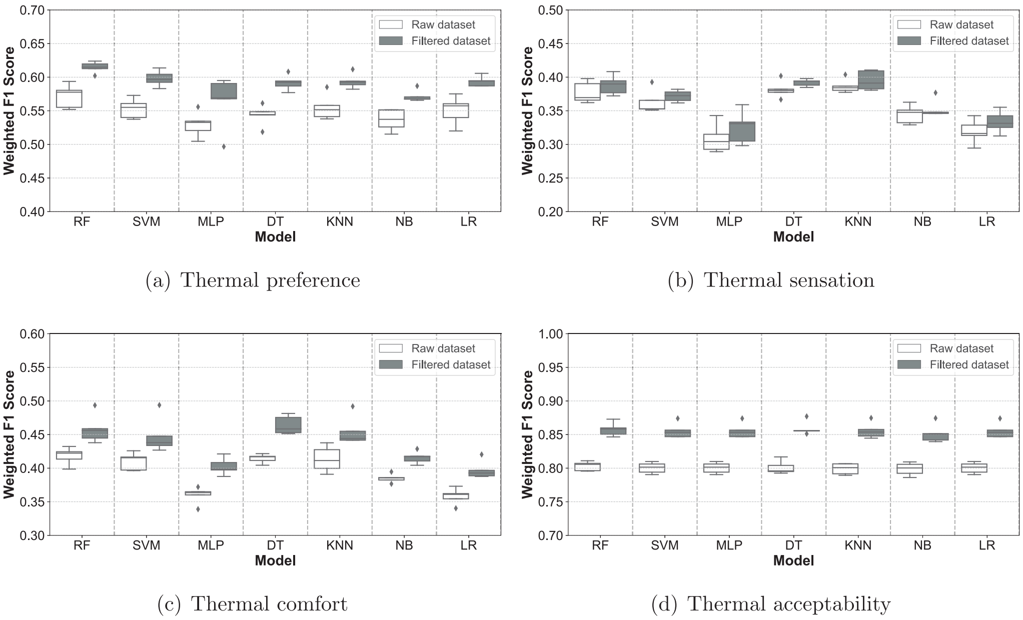
Fig. 6. Illustrations of the top-10 frequent and infrequent association patterns between four subjective perception metrics (thermal sensation–thermal preference–thermal acceptability–thermal comfort). The nodes show the categorical attributes of four subjective metrics. The edge shows the weights between these nodes. Edge weights $= Conf \times 100$.

5. Results and discussions

This section introduces the implementation and assessment of the proposed method. The authors conduct extensive experiments designed to (1) analyze potential subjective data biases within the Comfort Database II using the developed M-ARM algorithm, (2) investigate the impact of these subjective biases on thermal perception predictions, and (3) evaluate the effectiveness of state-of-the-art calibration methods in enhancing reliability performance of existing ML-based thermal perception prediction models.

5.1. Impacts of subjective biases on thermal perception predictions

This study uncovers potential instances of subjective biases within the Comfort Database II through the proposed M-ARM approach. Fig. 6 presents the ten most and least frequent association rules, with circles of different colors denoting the four thermal perception metrics: thermal sensation, preference, acceptance, and comfort. Fig. 6(a) displays the top-10 most frequent association rules, with the values along each edge indicating the maximum confidence ($Conf$) between two metrics. For example, the association rule with the $Conf$ of 40.6% is <‘‘comfortable’’, ‘‘acceptable’’, ‘‘no change’’, ‘‘neutral’’>. In contrast, Fig. 6(b) shows the top-10 least frequent association rules, with values along each edge indicating the minimal



(a) Thermal preference

(b) Thermal sensation

(c) Thermal comfort

(d) Thermal acceptability

Fig. 7. The impacts of subjective bias on the weighted F1-score performance of ML-based thermal perception prediction models.

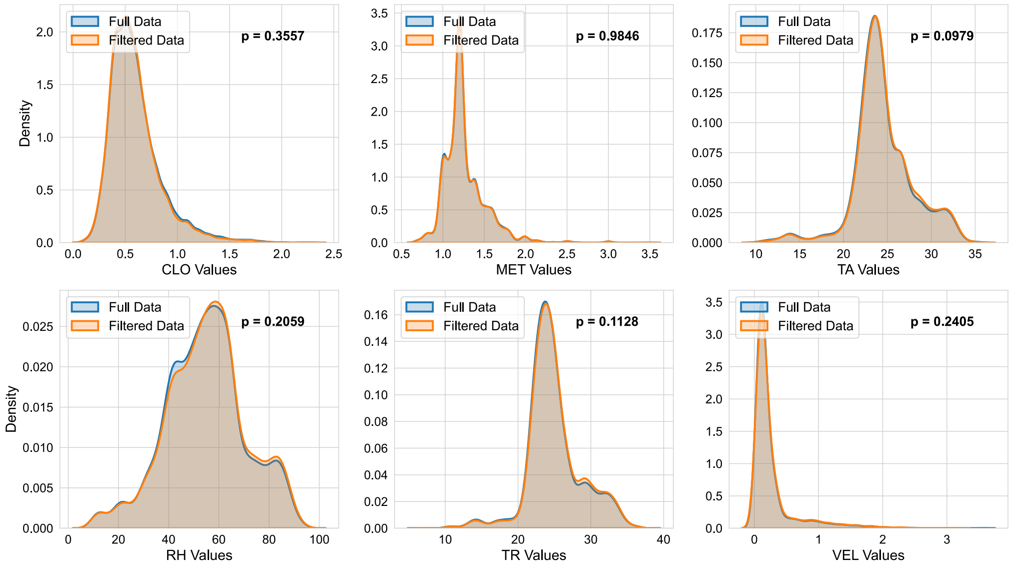


Fig. 8. Density distributions of six influential factors in raw and filtered datasets. The factors include clothing insulation (*clo*), metabolic rate (*met*), air temperature (*ta*), relative humidity (*rh*), mean radiant temperature (*tr*), and air velocity (*vel*). The p-values are derived from the Kolmogorov–Smirnov test [55] on raw and filtered datasets.

confidence between two metrics. The M-ARM algorithm detects several associated responses with extremely low *Conf*, considering these as instances of subjective biases. For example, the occupants who simultaneously chose <`‘‘very uncomfortable’’`, `‘‘acceptable’’`, `‘‘warmer’’`, `‘‘neutral’’`> are associated with significantly low *Conf* values (i.e., 0.03%).

This study then investigates the benefits of implementing confidence thresholds. Initially, the model evaluates the test set without a *Conf* threshold, after which a 5% *Conf* threshold is applied. The impacts of this strategy are analyzed by observing changes in model prediction performance (measured by weighted F1 score) across four aspects of thermal perceptions: thermal preferences, sensations, comfort, and acceptability.

As seen in Fig. 7, all ML methodologies show improvements in their discrimination performance following the filtering of identified biased labels. More specifically, the RF model performs better in predicting thermal preferences, achieving a weighted F1 score of 61.47%. This score represents an average improvement of 4.29% after filtering biased data. On the other hand, the KNN model yields the highest weighted F1 score for thermal sensation prediction, at 39.49%, marking an average improvement of 1.11% after filtering biased data. Furthermore, in predicting thermal comfort, the DT model excels, with a weighted F1 score of 46.39%, indicating an average improvement of 4.01% following the removal of biased data. Similarly, for thermal acceptability prediction, the DT model again achieves the highest weighted F1 score at 85.90%, along with an average improvement of 5.53% after filtering biased data. These results underline the effectiveness of the M-ARM approach in examining the underlying anomalous associations in the data and provide statistical support to the strategy of removing data characterized by subjective biases.

Fig. 8 shows the density distributions of six influential factors, including clothing insulation, metabolic rate, air temperature, relative humidity, mean radiant temperature, and air velocity, in raw and filtered datasets. The p-values derived from the Kolmogorov–Smirnov

Table 3
Model reliability performance (measured by Average ECE) of ML-based methods for thermal preference predictions using different calibration methods. The lowest ECE per ML model is highlighted in bold.

Models	Uncal.	Platt	Isotonic	Beta	BBQ	Temp.	GPcalib
RF	.0753	.0709	.0699	.0748	**.0644**	.0765	.0749
SVM	.0619	**.0557**	.0676	.0626	.0625	.0624	.0578
MLP	.0875	.0647	.0724	.0722	**.0643**	.0779	.0728
DT	.1788	.0720	**.0650**	.0716	.0789	.1149	.1795
KNN	.1056	.0504	.0485	.0744	.0816	**.0451**	.1072
NB	.1996	.0672	.0552	.0617	**.0386**	.0803	.0856
LR	.0745	.0700	.0684	.0748	**.0610**	.0850	.0751

test [55] offer a quantitative measure of the difference between the two distributions. The results indicate that the proposed method did not change the distributions of model inputs.

5.2. Assessment of model calibration performance on thermal perception predictions

This subsection investigates the effectiveness of state-of-the-art calibration methods in refining model predictive uncertainties. Consistent with previous research [21,43], this study uses the ECE score as our primary metric for evaluating the calibration performance of the classifiers, where lower ECE values signify superior performance. The ECE is derived by partitioning predictions into M fixed bins (each bin having a size of $1/M$) and subsequently determining the absolute difference between real and predicted probabilities for each bin, as expressed as follows:

$$ECE = \sum_{m=1}^{M} \Delta(m) \cdot |y_m - p_m| \quad (15)$$

where $\Delta(m)$ is the empirical fraction of all instances falling within bin m, while y_m and p_m are the actual and predicted probability in the bin m.

Of the four thermal perception metrics discussed, thermal preference serves as a direct indicator for the preferred control adjustments of the HVAC system in response to thermal environments [18]. Consistent with prior research [39], this study evaluates and compares the calibration performance of various methods, using the prediction of thermal preference as an example in the following analysis.

Table 3 presents the average ECE scores (with bins $M = 10$) both pre- and post-calibration for predicting thermal preference using diverse methods, including Platt scaling (Platt), Isotonic regression (Isotonic), Beta calibration (Beta), Bayesian Binning into Quantiles (BBQ), Temperature scaling (*Temp.*), and Gaussian Process calibration (*GPcalib*). The results show that most ML models exhibit some degree of miscalibration, with ECE generally falling within the range of 6% to 20%. Among uncalibrated models (*Uncal.*), SVM outperforms others with an ECE score of 0.0619, while NB displays the least satisfactory performance with a score of 0.1996.

Further analysis indicates that employing calibration methods can effectively enhance the reliability performance of various ML models in predicting thermal perceptions. For example, the BBQ calibration yields the lowest ECE score for RF (0.0644), MLP (0.0643), and LR (0.0610). Platt calibration proves the most efficient for SVM, resulting in an ECE score of 0.0557. Isotonic calibration demonstrates the best performance for DT with a score of 0.0650, while *Temp.* method is most effective for KNN, with an ECE score of 0.0451. The most optimal combination comprises the NB model and BBQ calibration, attaining the minimum ECE score of 0.0386, thereby indicating an improvement of 0.1610. Conversely, the DT model combined with the GPcalib calibration exhibits the highest ECE score of 0.1795, signaling inadequate calibration performance.

Fig. 9 displays class-specific reliability diagrams for both calibrated and uncalibrated test sets. The results show that uncalibrated models tend to be overconfident (i.e., *predicted probability* > *observed frequency*) when predicting “cooler” and “warmer” classes, while under-confident (i.e., *predicted probability* < *observed frequency*) for the “no change” class. Calibration methods, such as isotonic regression, Temperature scaling, and BBQ, deliver more reliable confidence estimates, leading to better-calibrated bins. Such refined calibration enhances model prediction transparency, thereby bolstering confidence in ML models [23]. Furthermore, these refined uncertainties can greatly aid HVAC control decisions. By considering class priors (like warm or cold) and weighing misclassification repercussions (such as energy use or user discomfort), calibrated models can adeptly navigate objectives related to user comfort, energy efficiency, or air quality [24–26].

Table 4
ECE performance of classifiers and calibration methods across different dataset sizes for thermal preference prediction. For each classifier, we show the mean performance of all calibration methods using different dataset sizes. For each calibration method, we show the mean performance of all classifiers using different dataset sizes. The lowest ECE per classifiers and calibration methods on different dataset sizes is highlighted in bold.

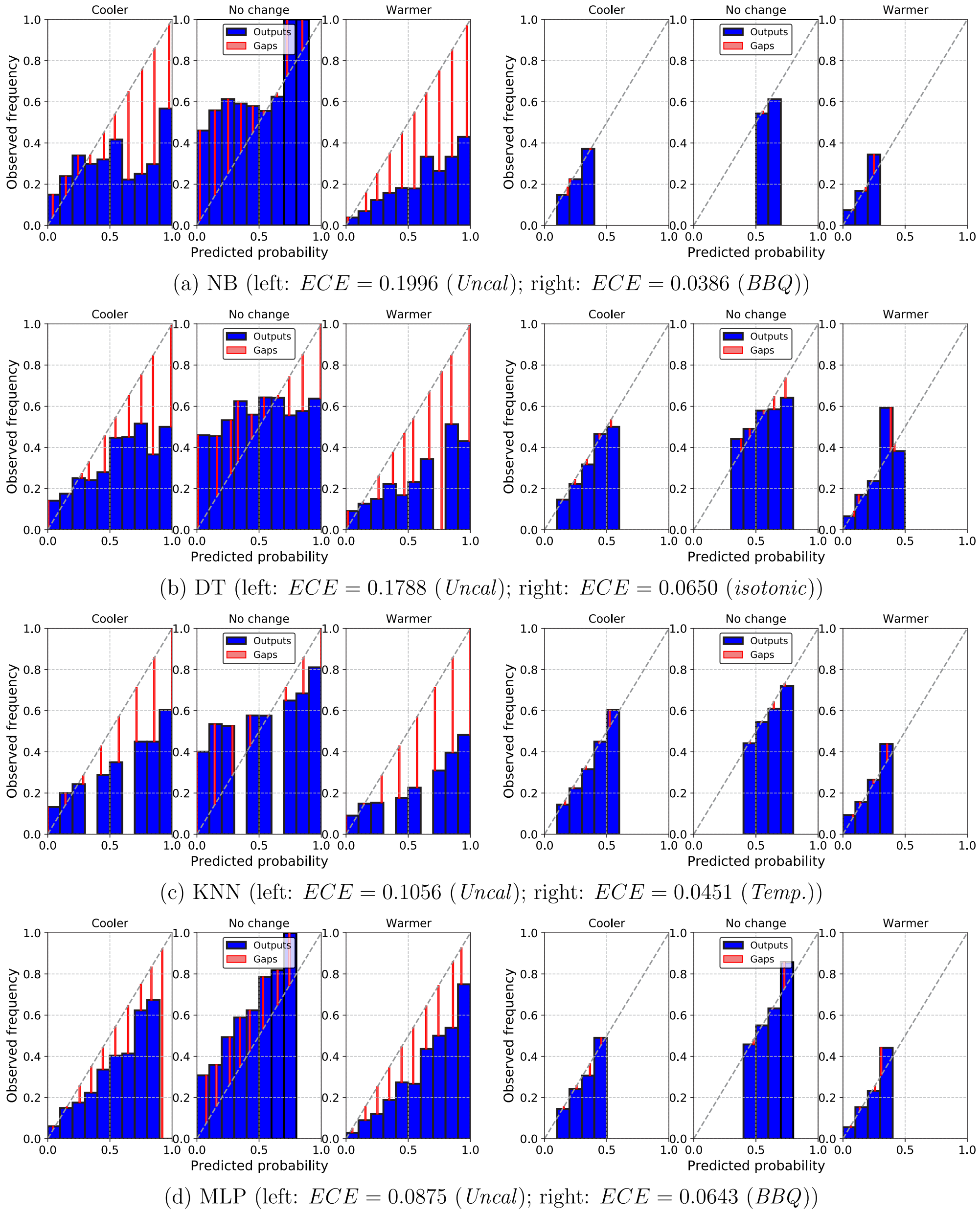
Methods	10%	40%	70%	100%
RF	.0872	.0796	.0745	.0726
SVM	**.0720**	**.0675**	**.0640**	**.0615**
MLP	.0783	.0735	.0727	.0731
DT	.1475	.1198	.1171	.1086
KNN	.0775	.0734	.0736	.0732
NB	.1050	.1006	.0982	.0840
LR	.0824	.0753	.0722	.0726
NoCal	.1361	.1187	.1181	.1119
Platt	.0736	.0655	.0668	.0644
Isotonic	.0913	.0729	.0686	**.0639**
Beta	.0726	.0665	**.0648**	.0703
BBQ	**.0639**	**.0651**	.0683	.0648
Temp.	.0852	.0834	.0788	.0774
GPcalib	.1273	.1172	.1108	.0932

5.3. Effects of dataset size

The authors first explore the interplay between dataset size and the biased data based on predefined $Conf$ thresholds, as demonstrated in Fig. 10. Subsets of the dataset, comprising 10%, 40%, 70%, and 100% of the total data, were used to evaluate this relationship. The results reveal that the prediction performance of ML-based models generally improves as dataset size increases. This observation supports the idea that larger datasets offer more comprehensive information for training ML algorithms, allowing them to learn more accurate and robust representations of the underlying patterns within the data.

Moreover, the observed improvement in the prediction performance of ML models suggests that the elimination of biased data, despite potentially reducing the volume of training data, does not critically diminish the overall information content within the dataset. More specifically, the best-performing algorithm is RF, with a weighted F1 score of 61.89% when utilizing 100% of the filtered dataset. These findings highlight the importance of removing biased data in the context of thermal perception datasets for improving model predicting performance.

This study also investigates the effects of dataset sizes on the calibration performance of different calibration methods. As presented in Table 4, the results indicate that the average ECE performance of all calibration methods and ML-based classifiers tends to decrease as the dataset size increases, implying that larger datasets generally result in better model calibration performance. Among the classifiers, SVM consistently exhibits the best calibration performance across all dataset sizes, outperforming other classifiers in terms of lower average ECE values. KNN and MLP follow with relatively better performance compared to the other classifiers. On the other hand, DT and NB classifiers display relatively poorer calibration performance, as evidenced by their higher



(a) NB (left: $ECE = 0.1996$ ($Uncal$); right: $ECE = 0.0386$ (BBQ))

(b) DT (left: $ECE = 0.1788$ ($Uncal$); right: $ECE = 0.0650$ ($isotonic$))

(c) KNN (left: $ECE = 0.1056$ ($Uncal$); right: $ECE = 0.0451$ ($Temp.$))

(d) MLP (left: $ECE = 0.0875$ ($Uncal$); right: $ECE = 0.0643$ (BBQ))

Fig. 9. Class-wise reliability diagrams for ML-based thermal preference prediction methods pre- and post-calibration.

average ECE values across all dataset sizes. RF and LR classifiers present moderate calibration performance, with average ECE values between the top-performing (SVM, RF, MLP) and lower-performing (DT, NB) classifiers.

Regarding calibration methods, the BBQ method achieves the lowest ECE for smaller dataset sizes (10% and 40%), suggesting its robustness even when working with limited data. However, as the dataset size increased to 70% and 100%, the Beta calibration method showed improved performance, demonstrating the lowest ECE among the calibration methods at 70%. On the contrary, the *GPcalib* displays higher ECE values across all dataset sizes, partially due to its inferior calibration performance on the DT (see Table 3). Platt, Isotonic, and *Temp.* calibration methods demonstrate moderate calibration performance. These findings emphasize the importance of selecting appropriate methods based on specific classifiers and dataset characteristics to achieve optimal calibration performance.

5.4. Effects of subjective data biases on model calibration performance

This subsection analyzes the effects of subjective biases on model reliability by comparing the ECE performance of various ML methods using raw and filtered datasets. The filtered dataset is generated by removing instances identified as subjective data biases.

Table 5 presents the ECE scores for the seven ML-based models. The results reveal that the ECE performance varies between models using raw and filtered datasets. Notably, the KNN model demonstrates a significant decrease in ECE score, from 0.1056 in the raw dataset to 0.0887 in the filtered dataset, signifying a considerable improvement in the model's reliability after removing instances influenced by subjective biases. This indicates that the KNN model is sensitive to biased data and may produce unreliable predictions when trained on biased datasets. Similarly, the ECE scores for SVM, MLP, DT, NB, and LR models decrease after filtering, with values changing from 0.0619 to 0.0599,

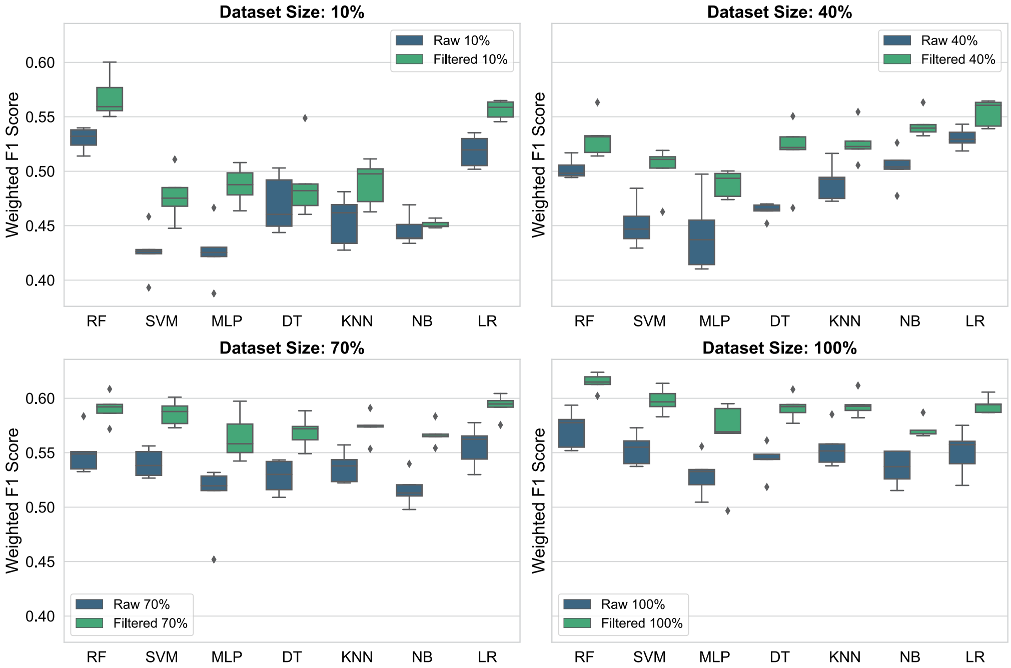


Fig. 10. Effects of subjective bias on thermal preference predictions using different training dataset sizes (10%, 40%, 70%, and 100%).

Table 5
ECE performance (*NoCal*) of various ML methods for predicting thermal preference on the raw and filtered datasets.

Dataset	Models						
	RF	SVM	MLP	DT	KNN	NB	LR
Raw	.0753	.0619	.0875	.1788	.1056	.1996	.0745
Filtered	.0857↑	.0599↓	.0756↓	.1754↓	.0887↓	.1989↓	.0660↓

Note: ↑ indicates the ECE increases after filtering. ↓ indicates the ECE score decreases after filtering.

0.0875 to 0.0756, 0.1788 to 0.1754, 0.1996 to 0.1989, and 0.0745 to 0.0660, respectively. These decreases suggest that the models become more reliable after removing instances affected by subjective biases, although the improvement is less significant than the one observed in the NB model. However, the RF models exhibit increased ECE scores post-filtering, with values changing from 0.0753 to 0.0857. This suggests that the RF models become less reliable after eliminating subjective biases, possibly because of the loss of beneficial information.

In conclusion, our analysis demonstrates that subjective biases can significantly affect model reliability in predicting thermal perceptions. The impact depends on the specific ML method employed, and some models are more sensitive to biased data than others. This highlights the importance of addressing subjective biases in data collection and preprocessing, particularly when working with highly sensitive models to noise and bias. Furthermore, the results emphasize the need to carefully choose the most appropriate ML methods when working with potentially biased datasets to ensure the reliability and validity of the model results.

6. Limitations and future work

This study presents several limitations that need to be addressed in future research. First, our investigation relies exclusively on data from the ASHRAE Global Thermal Comfort Database II. Despite being the most comprehensive thermal comfort database currently available, numerous field studies are inevitably not included in this database. As such, future research intends to verify the effectiveness of the proposed method using additional field datasets.

The second limitation pertains to our method to identify subjective biases in thermal perception responses. We underscore the necessity of accounting for these biases in developing ML-driven models for predicting thermal perceptions. However, due to the absence of ground-truth benchmarks, subjective data biases, characterized by their rarity or deviation from common patterns, may represent unique yet genuine thermal experiences. Their exclusion poses a risk of oversimplifying the thermal perception landscape and could lead to a loss of valuable insights, underscoring a gap in the comprehensiveness of our dataset. Furthermore, we must consider the possibility that the observed enhancement in the performance of our ML models may be, in part, a consequence of excluding non-typical or “hard-to-predict” data points. This consideration is particularly crucial in the absence of established ground-truth benchmarks for thermal perceptions, which leaves open questions regarding the optimal strategies for collecting and interpreting subjective data with high measurement validity and reliability. One way to mitigate these risks is to implement a hybrid approach that combines the frequent association pattern methodology with a secondary analysis of the filtered-out data points. This dual analysis allows for identifying non-typical yet significant thermal experiences that may offer valuable insights into less common but still important occupant responses. The current studies are limited to analyzing six relevant factors for the filtered data points. Future research should incorporate a broader array of features to maintain the dataset’s quality while preserving its completeness. In light of these limitations, our study aims to spotlight the imperative of ensuring subjective data reliability in thermal perception collection and modeling processes rather than providing prescriptive solutions.

Finally, this study systematically evaluated the calibration performance of existing ML-based thermal perception models using ECE scores and reliability diagrams. The results highlighted model miscalibration in the current models. By employing the latest calibration methods for seven ML-based models, our analysis demonstrates that calibration methods can effectively enhance the model’s reliability performance in predicting thermal perceptions. However, the optimal calibration method may differ depending on the underlying models and dataset characteristics. Future research could explore potential improvements in calibration techniques to further improve model reliability performance. Additionally, considering the uncertainties in calibrated model predictions, subsequent studies should aim to develop an intelligent control algorithm. This algorithm would consider the misclassification costs tied to model predictions to ensure guaranteed HVAC system performance. For instance, in instances of low prediction confidence, the algorithm could either seek feedback from occupants [56] or adopt more conservative control measures [25]. However, a detailed implementation of this control algorithm falls outside the scope of the present discussion.

7. Conclusions

Occupant thermal comfort is vital in building designs and operations, affecting energy efficiency and occupant well-being. However, ML-based thermal perception prediction encounters issues with subjective biases and model predictive uncertainties. Using the reliability diagrams and ECE scores, the authors reveal that most ML-based thermal comfort prediction models exhibit some miscalibration and tend to suffer from poor calibration. This study introduces a data-model integration method that calibrates the prediction uncertainties of ML-based thermal perception models. The method presents the M-ARM algorithm to detect potentially biased human responses by leveraging the interrelationship between subjective metrics such as thermal sensation, comfort, acceptance, and preference. Using the ASHRAE Comfort Database II, this study examines the effects of subject bias and the effectiveness of the proposed M-ARM algorithm. The results show that the seven ML approaches exhibit an improved discrimination performance (measured by the weighted F1 score), and the proposed method improves the prediction performance of thermal sensation, thermal comfort, thermal preference, and thermal acceptability by up to 1.01%, 4.01%, 4.11%, and 5.53%, respectively. To further refine model calibration performance, the authors applied six calibration methods, Platt, isotonic, Beta, BBQ, *Temp.*, and *GPcalib*, to the ML models. The results demonstrate that these calibration methods can improve the reliability of various ML approaches in predicting thermal preferences, with a maximum ECE improvement of 0.1610 (80.66% reduction in the uncalibrated ECE score). The proposed framework delivers more reliable thermal perception predictions by integrating the M-ARM algorithm to address subjective biases and applying state-of-the-art calibration methods for model predictive uncertainties. Furthermore, the presented study investigated the impacts of dataset size on the calibration performance of various methods. The BBQ method performed best with smaller datasets (10% and 40%), indicating its robustness with data shortage. However, as the dataset size increased (70% and 100%), Beta calibration demonstrated superior performance, indicating its scalability for larger datasets. This will ultimately guide the development of uncertainty-informed control strategies for optimizing occupant thermal comfort in buildings. Future studies will validate the effectiveness of the proposed method using more field datasets and explore potential improvements in calibration techniques to further enhance model reliability performance.

CRediT authorship contribution statement

Ruoxin Xiong: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Ying Shi:** Writing – review & editing, Investigation, Data curation, Conceptualization. **Haoming Jing:** Writing – review & editing, Validation, Methodology, Conceptualization. **Wei Liang:** Writing – review & editing, Methodology, Data curation. **Yorie Nakahira:** Writing – review & editing, Supervision, Project administration, Conceptualization. **Pingbo Tang:** Writing – review & editing, Supervision, Resources, Project administration, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Pingbo Tang reports financial support was provided by Manufacturing Futures Institute (MFI). Pingbo Tang reports financial support was provided by CMU Autonomous Technologies for Livability And Sustainability (ATLAS) Planning Grant for Engineering Research Center.

Data availability

Data will be made available on request.

Acknowledgments

This work is supported by the CMU Autonomous Technologies for Livability And Sustainability (ATLAS) Planning Grant for Engineering Research Centers and the Manufacturing Futures Institute, made possible by the Richard King Mellon Foundation. The supports are gratefully acknowledged.

References

[1] ASHRAE, ANSI/ASHRAE Standard 55: Thermal environmental conditions for human occupancy, Tech. rep., The American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE), 2011.

[2] C. Karmann, S. Schiavon, E. Arens, Percentage of commercial buildings showing at least 80% occupant satisfied with their thermal comfort, in: Proceedings of 10th Windsor Conference, 2018, pp. 48–54.

[3] A. Aryal, B. Becerik-Gerber, Energy consequences of comfort-driven temperature setpoints in office buildings, Energy Build. 177 (2018) 33–46, http://dx.doi.org/10.1016/j.enbuild.2018.08.013.

[4] H. Zhang, A. Tzempelikos, X. Liu, S. Lee, F. Cappelletti, A. Gasparella, The impact of personal preference-based thermal control on energy use and thermal comfort: Field implementation, Energy Build. 284 (2023) 112848, http://dx.doi.org/10.1016/j.enbuild.2023.112848.

[5] W. Jung, F. Jazizadeh, Human-in-the-loop HVAC operations: A quantitative review on occupancy, comfort, and energy-efficiency dimensions, Appl. Energy 239 (2019) 1471–1508, http://dx.doi.org/10.1016/j.apenergy.2019.01.070.

[6] E. Ono, K. Mihara, K.P. Lam, A. Chong, The effects of a mismatch between thermal comfort modeling and HVAC controls from an occupancy perspective, Build. Environ. 220 (2022) 109255, http://dx.doi.org/10.1016/j.buildenv.2022.109255.

[7] W. O'Brien, A. Wagner, M. Schweiker, A. Mahdavi, J. Day, M.B. Kjærgaard, S. Carlucci, B. Dong, F. Tahmasebi, D. Yan, et al., Introducing IEA EBC annex 79: Key challenges and opportunities in the field of occupant-centric building design and operation, Build. Environ. 178 (2020) 106738, http://dx.doi.org/10.1016/j.buildenv.2020.106738.

[8] Z. Nagy, B. Gunay, C. Miller, J. Hahn, M.M. Ouf, S. Lee, B.W. Hobson, T. Abuimara, K. Bandurski, M. André, et al., Ten questions concerning occupant-centric control and operations, Build. Environ. 242 (2023) 110518, http://dx.doi.org/10.1016/j.buildenv.2023.110518.

[9] M. Luo, J. Xie, Y. Yan, Z. Ke, P. Yu, Z. Wang, J. Zhang, Comparing machine learning algorithms in predicting thermal sensation using ASHRAE Comfort Database II, Energy Build. 210 (2020) 109776, http://dx.doi.org/10.1016/j.enbuild.2020.109776.

[10] T. Cheung, S. Schiavon, T. Parkinson, P. Li, G. Brager, Analysis of the accuracy on PMV–PPD model using the ASHRAE Global Thermal Comfort Database II, Build. Environ. 153 (2019) 205–217, http://dx.doi.org/10.1016/j.buildenv.2019.01.055.

[11] Z. Wu, N. Li, J. Peng, H. Cui, P. Liu, H. Li, X. Li, Using an ensemble machine learning methodology-bagging to predict occupants' thermal comfort in buildings, Energy Build. 173 (2018) 117–127, http://dx.doi.org/10.1016/j.enbuild.2018.05.031.

[12] Z.Q. Fard, Z.S. Zomorodian, S.S. Korsavi, Application of machine learning in thermal comfort studies: A review of methods, performance and challenges, Energy Build. 256 (2022) 111771, http://dx.doi.org/10.1016/j.enbuild.2021.111771.

[13] Y. Feng, S. Liu, J. Wang, J. Yang, Y.-L. Jao, N. Wang, Data-driven personal thermal comfort prediction: A literature review, Renew. Sustain. Energy Rev. 161 (2022) 112357, http://dx.doi.org/10.1016/j.rser.2022.112357.

[14] D. Shipworth, G. Huebner, M. Schweiker, B. Kingma, Diversity in thermal sensation: Drivers of variance and methodological artefacts, in: Proceedings of 9th Windsor Conference, 2016, pp. 56–72.

[15] X. Fuchs, S. Becker, K. Schakib-Ekbatan, M. Schweiker, Subgroups holding different conceptions of scales rate room temperatures differently, Build. Environ. 128 (2018) 236–247, http://dx.doi.org/10.1016/j.buildenv.2017.11.034.

[16] J. Wang, Z. Wang, R. de Dear, M. Luo, A. Ghahramani, B. Lin, The uncertainty of subjective thermal comfort measurement, Energy Build. 181 (2018) 38–49, http://dx.doi.org/10.1016/j.enbuild.2018.09.041.

[17] M. Schweiker, M. André, F. Al-Atrash, H. Al-Khatri, R.R. Alprianti, H. Alsaad, R. Amin, E. Ampatzi, A.Y. Arsano, E. Azar, et al., Evaluating assumptions of scales for subjective assessment of thermal environments – Do laypersons perceive them the way, we researchers believe? Energy Build. 211 (2020) 109761, http://dx.doi.org/10.1016/j.enbuild.2020.109761.

[18] Z. Wang, J. Wang, Y. He, Y. Liu, B. Lin, T. Hong, Dimension analysis of subjective thermal comfort metrics based on ASHRAE Global Thermal Comfort Database using machine learning, J. Build. Eng. 29 (2020) 101120, http://dx.doi.org/10.1016/j.jobe.2019.101120.

[19] Z. Wang, T. Parkinson, P. Li, B. Lin, T. Hong, The squeaky wheel: Machine learning for anomaly detection in subjective thermal comfort votes, Build. Environ. 151 (2019) 219–227, http://dx.doi.org/10.1016/j.buildenv.2019.01.050.

[20] C. Moore, J. Doherty, Role of the calibration process in reducing model predictive error, Water Resour. Res. 41 (5) (2005) W05020, http://dx.doi.org/10.1029/2004WR003501.

[21] J. Wenger, H. Kjellström, R. Triebel, Non-parametric calibration for classification, in: International Conference on Artificial Intelligence and Statistics, 2020, pp. 178–190.

[22] N. Ma, L. Chen, J. Hu, P. Perdikaris, W.W. Braham, Adaptive behavior and different thermal experiences of real people: A Bayesian neural network approach to thermal preference prediction and classification, Build. Environ. 198 (2021) 107875, http://dx.doi.org/10.1016/j.buildenv.2021.107875.

[23] Y. Zhang, Q.V. Liao, R.K. Bellamy, Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 295–305, http://dx.doi.org/10.1145/3351095.3372852.

[24] T.-Y. Chao, M.-H. Nguyen, C.-C. Huang, C.-C. Liang, C.-W. Chung, Online self-learning for smart HVAC control, in: 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), IEEE, 2019, pp. 4324–4330, http://dx.doi.org/10.1109/SMC.2019.8914027.

[25] M. Maasoumy, M. Razmara, M. Shahbakhti, A.S. Vincentelli, Selecting building predictive control based on model uncertainty, in: 2014 American Control Conference, IEEE, 2014, pp. 404–411, http://dx.doi.org/10.1109/ACC.2014.6858875.

[26] T. Silva Filho, H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull, P. Flach, Classifier calibration: A survey on how to assess and improve predicted class probabilities, Mach. Learn. (2023) 1–50, http://dx.doi.org/10.1007/s10994-023-06336-7.

[27] S. Zhao, M. Kim, R. Sahoo, T. Ma, S. Ermon, Calibrating predictions to decisions: A novel approach to multi-class calibration, in: Advances in Neural Information Processing Systems, Vol. 34, 2021, pp. 22313–22324.

[28] C. Guo, G. Pleiss, Y. Sun, K.Q. Weinberger, On calibration of modern neural networks, in: Proceedings of International Conference on Machine Learning, 2017, pp. 1321–1330.

[29] W. Liang, R. Xiong, P. Liu, P. Tang, E. Cochran, Improving post-occupancy evaluation engagement using social robots, in: Proceedings of the 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, 2022, pp. 159–167, http://dx.doi.org/10.1145/3563357.3564071.

[30] Y. Wang, P. Tang, K. Liu, J. Cai, R. Ren, J.J. Lin, H. Cai, J. Zhang, N. El-Gohary, M. Berges, et al., Characterizing data sharing in civil infrastructure engineering: Current practice, future vision, barriers, and promotion strategies, J. Comput. Civ. Eng. 37 (2) (2023) 04023001, http://dx.doi.org/10.1061/JCCEE5.CPENG-5077.

[31] C.J. Pannucci, E.G. Wilkins, Identifying and avoiding bias in research, Plast. Reconstr. Surg. 126 (2) (2010) 619–625, http://dx.doi.org/10.1097/PRS.0b013e3181de24bc.

[32] E. Hüllermeier, W. Waegeman, Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods, Mach. Learn. 110 (2021) 457–506, http://dx.doi.org/10.1007/s10994-021-05946-3.

[33] S. Lee, P. Karava, A. Tzempelikos, I. Bilionis, Inference of thermal preference profiles for personalized thermal environments with actual building occupants, Build. Environ. 148 (2019) 714–729, http://dx.doi.org/10.1016/j.buildenv.2018.10.027.

[34] L. Chen, A. Ermis, F. Meng, Y. Zhang, Meta-learning of personalized thermal comfort model and fast identification of the best personalized thermal environmental conditions, Build. Environ. 235 (2023) 110201, http://dx.doi.org/10.1016/j.buildenv.2023.110201.

[35] B. Zadrozny, C. Elkan, Transforming classifier scores into accurate multiclass probability estimates, in: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002, pp. 694–699, http://dx.doi.org/10.1145/775047.775151.

[36] L.H. Mervin, S. Johansson, E. Semenova, K.A. Giblin, O. Engkvist, Uncertainty quantification in drug design, Drug Discov. Today 26 (2) (2021) 474–489, http://dx.doi.org/10.1016/j.drudis.2020.11.027.

[37] R. Rosenman, V. Tennekoon, L.G. Hill, Measuring bias in self-reported data, Int. J. Behav. Healthc. Res. 2 (4) (2011) 320–332, http://dx.doi.org/10.1504/IJBHR.2011.043414.

[38] Z. Wang, R. de Dear, M. Luo, B. Lin, Y. He, A. Ghahramani, Y. Zhu, Individual difference in thermal comfort: A literature review, Build. Environ. 138 (2018) 181–193, http://dx.doi.org/10.1016/j.buildenv.2018.04.040.

[39] S. Zhang, R. Yao, C. Du, E. Essah, B. Li, Analysis of outlier detection rules based on the ASHRAE Global Thermal Comfort Database, Build. Environ. 234 (2023) 110155, http://dx.doi.org/10.1016/j.buildenv.2023.110155.

[40] V. Kuleshov, N. Fenner, S. Ermon, Accurate uncertainties for deep learning using calibrated regression, in: Proceedings of the 35th International Conference on Machine Learning, 2018, pp. 2796–2804.

[41] J. Platt, Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, in: Advances in Large Margin Classifiers, Vol. 10, Cambridge, MA, 1999, pp. 61–74.

[42] M. Kull, T.S. Filho, P. Flach, Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers, in: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Vol. 54, 2017, pp. 623–631.

[43] M.P. Naeini, G. Cooper, M. Hauskrecht, Obtaining well calibrated probabilities using bayesian binning, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2015, pp. 2901–2907, http://dx.doi.org/10.1609/aaai.v29i1.9602.

[44] V.F. Ličina, T. Cheung, H. Zhang, R. De Dear, T. Parkinson, E. Arens, C. Chun, S. Schiavon, M. Luo, G. Brager, et al., Development of the ASHRAE Global Thermal Comfort Database II, Build. Environ. 142 (2018) 502–512, http://dx.doi.org/10.1016/j.buildenv.2018.06.022.

[45] R.J. De Dear, A global database of thermal comfort field experiments, ASHRAE Trans. 104 (1998) 1141.

[46] W. Hu, Y. Luo, Z. Lu, Y. Wen, Heterogeneous transfer learning for thermal comfort modeling, in: Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, 2019, pp. 61–70, http://dx.doi.org/10.1145/3360322.3360843.

[47] Q. Lin, C. Gao, Discovering categorical main and interaction effects based on association rule mining, IEEE Trans. Knowl. Data Eng. 35 (2021) 1379–1390, http://dx.doi.org/10.1109/TKDE.2021.3087343.

[48] R. Agarwal, R. Srikant, Fast algorithms for mining association rules, in: Proceedings of the 20th International Conference on Very Large Data Bases (VLDB), 1994, pp. 487–499.

[49] C. Rudin, B. Letham, A. Salleb-Aouissi, E. Kogan, D. Madigan, Sequential event prediction with association rules, in: Proceedings of the 24th Annual Conference on Learning Theory, Vol. 19, 2011, pp. 615–634.

[50] K. Pearson, X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, Lond. Edinb. Dublin Philos. Mag. J. Sci. 50 (302) (1900) 157–175, http://dx.doi.org/10.1080/14786440009463897.

[51] M. Kull, M. Perello Nieto, M. Kängsepp, T. Silva Filho, H. Song, P. Flach, Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration, in: Advances in Neural Information Processing Systems, Vol. 32, 2019, pp. 12316–12326.

[52] B. Zadrozny, C.P. Elkan, Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers, in: Proceedings of International Conference on Machine Learning, 2001, pp. 609–616.

[53] N. Gao, W. Shao, M.S. Rahaman, J. Zhai, K. David, F.D. Salim, Transfer learning for thermal comfort prediction in multiple cities, Build. Environ. 195 (2021) 107725, http://dx.doi.org/10.1016/j.buildenv.2021.107725.

[54] A. Chennapragada, D. Periyakoil, H.P. Das, C.J. Spanos, Time series-based deep learning model for personal thermal comfort prediction, in: Proceedings of the 13th ACM International Conference on Future Energy Systems, 2022, pp. 552–555, http://dx.doi.org/10.1145/3538637.3539617.

[55] F.J. Massey, The Kolmogorov-Smirnov test for goodness of fit, J. Amer. Statist. Assoc. 46 (253) (1951) 68–78, http://dx.doi.org/10.1080/01621459.1951.10500769.

[56] S. Lee, P. Karava, A. Tzempelikos, I. Bilionis, A smart and less intrusive feedback request algorithm towards human-centered HVAC operation, Build. Environ. 184 (2020) 107190, http://dx.doi.org/10.1016/j.buildenv.2020.107190.