Energy & Buildings 357 (2026) 117175

Contents lists available at ScienceDirect

Energy & Buildings

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

Upgrading UTCI through supervised learning using the RUROS dataset

José-Antonio Rodríguez-Gallego [a], Eduardo Diz-Mellado [b], Marialena Nikolopoulou [c], Tomás Chacón-Rebollo [d], Carlos Rivera-Gómez [e], Carmen Galán-Marín [e,*]

[a] Departamento de Matemática Aplicada I, Escuela Técnica Superior de Ingeniería Informática, Universidad de Sevilla, Avda. Reina Mercedes, 41012 Seville, Spain
[b] Departamento de Máquinas y Motores Térmicos, Escuela Superior de Ingeniería, Universidad de Cádiz, Avda. Universidad de Cádiz, 10, 11510 Puerto Real, Spain
[c] School of Art and Architecture, University of Kent, Canterbury, UK
[d] Departamento de Ecuaciones Diferenciales y Análisis Numérico, Facultad de Matemáticas, Universidad de Sevilla, Avda. Reina Mercedes, 41012 Seville, Spain
[e] Departamento de Construcciones Arquitectónicas 1, Escuela Técnica Superior de Arquitectura, Universidad de Sevilla, Avda. Reina Mercedes, 2, 41012 Seville, Spain

ARTICLE INFO

Keywords:
OTC classification
Explainable machine learning
RUROS dataset
Outdoor thermal comfort
UTCI
Naïve Bayes

ABSTRACT

Outdoor thermal comfort (OTC) plays a key role in climate-resilient urban planning, but widely used indices such as the Universal Thermal Climate Index (UTCI) are constrained by their complex thermophysiological formulations and limited interpretability. This paper proposes an explainable Machine Learning (ML) framework for binary OTC classification based on the RUROS dataset, which contains 6,079 valid questionnaire responses from seven European cities. By substituting thermoregulatory simulations with a data-driven methodology, this study offers a transparent tool that clarifies how environmental variables shape human thermal perception. We assessed 36 distinct workflows in seven cities, comparing inherently interpretable models such as Logistic Regression and Decision Trees against standard baselines. By validating on cities that were held out during training, we confirmed that the model generalizes well to varied climatic conditions. Our final recommendation, a Naïve Bayes model with Downsampling (NBD), relies on air temperature, wind speed, and relative humidity to estimate comfort probabilities. The NBD model outperforms the UTCI, reaching a balanced accuracy of 0.611 versus 0.585 for the index. Statistical verification using Wilcoxon signed-rank tests and bootstrap confidence intervals shows that this advantage is both consistent and meaningful. In addition to higher predictive accuracy, NBD provides a probabilistic framework that enables urban planners to map heat-risk areas in terms of likelihood rather than relying on fixed thresholds. Together, these results underscore the promise of explainable ML for delivering more flexible and dependable evaluations to support well-being in public urban environments.

1. Introduction

Thermal comfort is a critical aspect of human well-being, influencing health, productivity, and overall quality of life [1,2]. It refers to the state in which individuals feel neither too hot nor too cold, experiencing thermal neutrality, a concept which is particularly relevant in urban planning, architecture, and occupational health, where maintaining optimal thermal conditions can improve comfort, reduce energy consumption, and enhance performance [3,4].

1.1. Research context

In the last few years, thermal comfort has evolved from being an important issue to a critical one for human well-being [5]. One of the most evident effects of this evolution is the exacerbation of the phenomenon known as Urban Heat Island (UHI) [6], marking the effect the thermal conditions can have over high-density centers, such as cities. This problem is especially relevant when considering the deadly effect it can have on children and elderly people [7,8].

Therefore, the analysis and development of classification and prediction models for outdoor thermal comfort is key, providing reliable tools for an appropriate assessment and treatment of this problem. Since their introduction, comfort indices have addressed this task, providing such an assessment under variable circumstances [9]. Examples of frequently used indices include the UTCI [10] and PET [11], which allow users to understand climatic adaptation from an urban and personal perspective. The usage of these indices has been further expanded with machine learning models [12], combining model-based approaches, the indices, with data-driven models.

* Corresponding author.
E-mail address: cgalan@us.es (C. Galán-Marín).

https://doi.org/10.1016/j.enbuild.2026.117175
Received 4 August 2025; Received in revised form 9 February 2026; Accepted 14 February 2026
Available online 19 February 2026

1.2. Literature review

The complexity of thermal comfort arises from the interplay of multiple environmental and personal factors. Air temperature, relative humidity, wind speed, and mean radiant temperature (MRT) all influence perceived comfort, but so do individual characteristics such as metabolism, clothing insulation, and personal acclimatization[13]. Psychological and behavioral aspects further add to this complexity, as individuals adapt to different thermal conditions based on expectations and past experiences [14]. Given these intricate interactions, appropriate assessing thermal comfort requires accurate tools, which can be divided into two main groups: indices, mathematical models that provide a fast and reliable estimation of the human response under a set of conditions; and the advanced modelling techniques, such as Machine Learning (ML).

1.2.1. Outdoor thermal comfort indices

The evaluation of OTC is typically based on quantitative indices that transform environmental variables into interpretable metrics, representing human thermal stress. The need for specific indices rises from the profound differences with conventional indoor comfort models [13,14]. Within this framework, OTC indices seek to combine the main microclimatic elements —usually air temperature, mean radiant temperature, wind speed, and humidity— with models of human heat exchange to approximate perceived OTC [11,15]. These based indices are conceptually distinct from subjective, survey-derived metrics such as the Thermal Sensation Vote (TSV) or Thermal Comfort Vote (TCV), which quantify reported perception rather than simulated physiological response [16,17]. Although both viewpoints are informative, index-centered approaches offer a standardized basis for comparison across locations and for model development, explaining their predominant use in OTC research.

The categorization of OTC indices is commonly guided by modeling philosophy and primary application domain [15]. Early empirical or statistical indices are based on observed correlations and have limited physiological foundation, whereas energy-balance formulations—most prominently the Predicted Mean Vote (PMV)—represent steady-state heat transfer and were originally designed for controlled indoor environments [18,19]. More recent thermophysiological or heat-balance indices build on this framework to address outdoor contexts. PET (Physiological Equivalent Temperature) translates a given outdoor exposure into an equivalent indoor thermal condition using the Munich Energy-balance Model, which facilitates intuitive interpretation but depends critically on precise estimation of radiative fluxes [11,20]. SET (Standard Effective Temperature) uses a two-node thermophysiological representation to quantify equivalent comfort under standardized reference conditions, making it applicable to both indoor and semi-outdoor settings [21,22]. The Universal Thermal Climate Index (UTCI) is a more recent development explicitly aimed at outdoor evaluation, integrating detailed thermophysiological modeling with dimensionality reduction techniques to yield a computationally efficient index based on four climatic parameters [10,23]. Taken together, these indices vary in their foundational assumptions, their responsiveness to mean radiant temperature, and their appropriateness for non-steady, outdoor thermal environments.

These indices involve trade-offs between how realistic, interpretable, and practical they are. Indices originally developed for indoor environments, such as PMV, benefit from formal standardization, yet their validity decreases under dynamic outdoor conditions, where variations in solar radiation and wind play a dominant role in shaping thermal sensation [24,25]. PET and SET offer a more detailed representation of human thermophysiology but are highly sensitive to the accurate specification of environmental inputs, especially radiative fluxes [26]. UTCI has gained prominence as a *de facto* reference index in OTC research because it is explicitly designed for outdoor use, has been empirically tested in a wide range of climatic settings, and requires a relatively small number of meteorological variables, which facilitates consistent comparison across studies [9,23]. However, the intricate modeling framework from which UTCI is derived, together with its generalized treatment of inter-individual differences, can make the index less transparent and harder to interpret [27,28]. These issues, as a result, motivate the exploration of complementary data-driven methods, retaining physical meaning while improving interpretability and flexibility.

1.2.2. The UTCI index

The Universal Thermal Climate Index (UTCI) [10] is a widely used metric for evaluating Outdoor Thermal Comfort (OTC) by integrating both environmental and physiological factors [23]. As it only requires the use of four climatic variables, it has been widely used for the assessment of human thermal comfort across different climatic conditions, as well as by meteorologists [9,29,30]. UTCI has become the *de facto* tool for addressing OTC, as compared to other outdoor indices, such as PET [11,31], as research indicates that thermophysiological models similar to UTCI provide better alignment with actual human thermal responses in outdoor settings, and on most occasions it provides a more accurate result, being an index only intended for outdoor environments [24,25,32]. For instance, Tseliou et al. [25] discovered that indices based on energy balance slightly surpassed entirely empirical methods, and Cureau et al. [32] showed that sophisticated thermoregulatory modeling significantly improves the precision of estimating human thermal stress outdoors. This has been the case even compared to modified versions of other indices, such as in the case of the modified bioheat equation for the PET index [20,33] or for the SET index [22].

Beyond objective thermal indices, studies of outdoor thermal comfort frequently rely on subjective indicators such as the Thermal Sensation Vote (TSV) and Thermal Comfort Vote (TCV), which reflect individuals' perceived thermal experiences using ordered categorical scales. While these indicators offer valuable insight into human-centered thermal perception, they also increase modeling complexity. In this work, we employ a binary comfort formulation to ensure compatibility with inherently interpretable classifiers and to yield straightforward probabilistic predictions. Extending the proposed approach to multinomial TSV or TCV classifications represents an important avenue for future research.

UTCI was developed [10]using as foundations Fiala's biothermal model [34,35], which was itself created atop Pennes' bioheat equation, built over a sample made out of only men [36]. Therefore, even if it considers internally non-climatic factors such as height, weight, age or clothing, both the simplification that arises from obtaining the UTCI from Fiala's model using a Principal Component Analysis (PCA), as well as the original bias from the dataset, resonate in its limitations from accurately predicting thermal comfort in women [18,24] and in non-American people [25,37]. As a result, UTCI is indeed a suitable option compared to other indices [38], but it admits improvements addressing these population sectors, as seen in the cited references.

In addition to index-based approaches such as UTCI, simulation-driven frameworks have also been proposed to represent microclimatic conditions and human thermal comfort at fine spatial resolution. SOLWEIG [39], for example, employs a detailed physiological framework to simulate individual thermoregulatory responses in outdoor environments, whereas ENVI-met [40] is a three-dimensional computational fluid dynamics model that evaluates urban microclimates by incorporating airflow dynamics, radiative exchange, and surface–atmosphere interactions. Although these models offer detailed insights into localized thermal environments and effectively complement index-based approaches, their high computational demands and model complexity limit their suitability for large-scale or comparative analyses. By using UTCI as a widely recognized and standardized reference, this study enables the evaluation of machine learning models while preserving methodological simplicity and ensuring reproducibility.

1.2.3. Machine learning methods

Even if the UTCI index is the most widely used classification tool for outdoor thermal comfort [9], in recent years many researchers have been using Machine Learning (ML) methods to improve the accuracy of not only the UTCI [41], but also other indices while providing a comparison [42], such as the case for the PET [20,33] and the SET indices [22]. For the UTCI, both Linear Regression models [38] as well as the Multi-Layer Perceptron [43] have been used [44], partially increasing the accuracy of the model. Other models such as Random Forest [45] or Extreme Gradient Boosting [46] have been also used [47], as well as a combination and comparison of many [48,49]. The fine-tuning that these models allow have not been extensively analyzed, like in adjacent areas that also care for classification, such as thermography and thermal anomaly detection [50–52], building design and structure detection [53,54] and indoor thermal comfort [55,56].

The problem with these models, and something rooted in the area when applying ML techniques, is that the chosen models have a low explainability [57], which combined with the relatively low accuracy they provide does not help assess which variables affect and in which way the perception of thermal comfort. This is more accentuated when using Deep Learning techniques, which even if providing accurate results for thermal comfort classification [58–61] they do not explain their inner behavior. To soften this issue, some researchers have been using the Shapley values [62] to provide a layer of explainability for their models, such as in the case of [63] and [27]. Notice, however, that due to the generality of the Shapley values, specific information from the chosen model is not exploited, as in the case of black box models such as Random Forest, Extreme Gradient Boosting or the Multi-Layer Perceptron, models that have provided results more accurate than those of the OTC indices when predicting the OTC classification [44,48,49].

1.3. Research gap

A notable limitation of UTCI is its limited inherent interpretability, as its generalized formulation obscures the influence and relative importance of individual variables [28]. Although the index is derived from only four environmental parameters—air temperature, mean radiant temperature, wind speed, and relative humidity—its computation follows a highly complex methodological process [23]. This process involves dimensionality reduction using Principal Component Analysis (PCA) and a fifth–order polynomial regression applied to the first principal component, yielding an expression comprising 109 terms. Such complexity restricts the direct interpretation of how each input variable contributes to the final thermal comfort classification [27]. While post–hoc approaches, including sensitivity analysis and explainable artificial intelligence methods such as SHAP, can be employed to explore variable effects, they require additional analytical layers beyond the index formulation. In contrast, the framework proposed in this study offers built–in interpretability, enabling the direct examination of how individual environmental variables influence comfort predictions and supporting a more transparent understanding of outdoor thermal perception.

1.4. Objectives & main contributions

In this paper, our objective is to present an alternative model for Outdoor Thermal Comfort (OTC) classification using the same variables as the UTCI index, in the binary case, *i.e.*, predicting either "*comfort*" or "*discomfort*", as well as their probability. UTCI was chosen over other options because of its extended use, despite the mentioned limitations. To achieve this, we used Machine Learning (ML) explainable models and techniques that exploited the RUROS dataset [64], therefore generating a classifier that provided both accurate and interpretable results. With this, our model can not only provide a prediction of the expected comfort, but also an explanation for its classification, while obtaining increased accuracy capabilities for binary outdoor thermal comfort classification. Notice that our approach does not share the thermophysiological bases in which the UTCI is founded, using instead a data-driven approach by means of the RUROS dataset. It is noteworthy that binary classification is considerably different from the equivalent temperature that the UTCI provides or the UTCI categories of classification built over it, but it was chosen due to the use of some ML techniques, which only allow binary classification, and particularly Naïve Bayes, the best-behaving model.

The RUROS dataset was one of the presented results of the homonymous project [17], and it provides a throughout set of measurements for key variables, both climatic, personal and morphological. The RUROS project itself proposed models for evaluation of the microclimate of open spaces and the resulting thermal visual and audible comfort conditions for the people using these spaces; a methodology for developing comfort maps for the area, and design guidelines for the development of open spaces. The open-source data was originally published in 2004 on the website of the Centre for Renewable Energy Sources. This database includes all the raw data from extensive outdoor comfort surveys conducted across seven European cities (Athens (GR), Thessaloniki (GR), Milan (I), Fribourg (CH), Kassel (D), Cambridge (UK) and Sheffield (UK), between July 2001 and March 2002.

To develop our classifier over it, we considered 7 models, 6 preprocessors, and the fine-tuning of the resulting model combinations. With this approach, we present a way of crossing state-of-the-art Machine Learning (ML) tools with an appropriate dataset for an explainable ML model with improved prediction capabilities. Our final model, the Naïve Bayes with downsampling, provided a balanced accuracy of 0.611 for binary OTC prediction, while at the same time showing its suitability for real-world use based on comfort probabilities. This model obtained better results in most metrics over the UTCI, PET and PMV indices, as well as results comparable to those of the Random Forest and Extreme Gradient Boosting, and outperforming the other explainable ML options.

1.5. Paper structure

The structure of this paper is organized as follows: Section 1 introduces the research context, presents a review of relevant literature—including the UTCI index and recent machine learning (ML) methods—identifies the research gap, and outlines the main contributions. Section 2 describes the materials and methods used in this study, beginning with an overview of the RUROS project, followed by a detailed explanation of the ML models, preprocessing techniques, and hyperparameter tuning procedures that have been used. The main results and their discussion are presented in Section 3, where we analyze the performance of the proposed model, explore its interpretability, and compare it against the UTCI-based classification. Finally, Section 4 summarizes the main achievements of the study and outlines possible paths for future work.

2. Materials & methods

2.1. The RUROS project

The RUROS (Rediscovering the Urban Realm and Open Spaces) project was a large-scale European project from 2001 to 2004, aiming to develop tools and guidelines for the analysis of open spaces in the urban environment [65], combining the physical environment (*i.e.* microclimate, thermal, visual and audible comfort, urban morphology, etc.) with user requirements and satisfaction [66].

By integrating these tools, RUROS aimed to inform the design and transformation of urban spaces into more sustainable and resilient environments. One of the key outputs of this project was the RUROS dataset [17], sharing the data harvested from the extensive field surveys conducted across different European cities during the project, to support future studies of OTC, e.g., [67,68].

The RUROS dataset is a comprehensive collection of thermal comfort

data, containing 9,271 observations across 58 variables, for 7 cities, representing different participants, as shown in Table 1. The surveys were carried out between 2001 and 2004 in seven European cities—Athens, Cambridge, Fribourg, Kassel, Milan, Sheffield, and Thessaloniki. Its multi–city scope enables the examination of thermal comfort patterns across a wide range of climatic contexts, thereby supporting the robustness and generalizability of subsequent model evaluations. It is important to mention that, out of the 9,271 samples, we executed a filter, keeping 6,079, 65.57% of the total. To do so, rows with one or more missing values were removed. This filter was done to ensure data stability and comparability across multiple experiments.

2.1.1. Available variables

This dataset offers a comprehensive basis for analyzing how environmental and personal factors shape thermal perception in outdoor urban environments. It includes a wide range of key **demographic** variables [69] (e.g., age, sex, nationality, education, occupation, local residency), **behavioral** factors (e.g., clothing insulation, metabolic rate, activity level, food and drink consumption), and **environmental** parameters (e.g., air temperature, wind speed, humidity, MRT), all of which are crucial for thermal comfort assessment [23].

In addition, it captures **subjective perceptions**, such as comfort, thermal sensation, glare, sound, and visual conditions, along with **social and contextual information** like reasons for being at the site, previous location, and frequency of use. This diversity of variables allows for detailed analysis of thermal comfort across different urban settings and population groups. In our case, we will focus on environmental variables *Tair*, *Tglobe*, *Wind_sp, R*h; as well as the *City* and the *Heat* variables, as they will be the only ones required for the development of our models and their comparison with the UTCI classification. Before choosing the final variables, a Random Forest model was trained and fine-tuned with only these variables and compared to another trained over all available variables in the dataset. The results showed that the increase in accuracy was less than 0.05, justifying the selection of the variables.

Specifically, the RF model trained with only the selected variables (*Tair, Tglobe, Wind_sp, Rh*) achieved a balanced accuracy of 0.66, while the RF model trained with all 58 available variables —after filtering uninformative ones such as *Date* or *City*— reached 0.69. Given the marginal improvement of 0.03, we concluded that the selected subset sufficiently captures the key determinants of outdoor thermal comfort while reducing model complexity and maintaining interpretability.

2.1.2. Target variable and classification approach

All analyses conducted in this study utilize the filtered RUROS dataset encompassing the seven cities over the 2001–2004 period, thereby ensuring consistency in both model training and evaluation. From the OTC point of view, the most relevant variable is the *Heat* variable, which represents the Actual Sensation Vote (ASV), ranging from −2, "*very cold*", to 2, "*very hot*", given by the participants. For our purposes, as we intend to use several ML models that can only process binary classification, such as the Logistic Regression (LR), the Naïve Bayes (NB), the Linear Discriminant Analysis (LDA) and the LASSO models, the *Heat* variable had to be transformed into a binary one, with the loss in information this carries. How to choose this transformation was key, as it would affect our methodology overall.

It should be highlighted that perceived OTC traditionally has been codified, in terms of UTCI, as a multiclass variable with 11 categories. The reduction to a binary variable —"comfortable" versus "uncomfortable"—, therefore, is an important simplification, as we lose the gradient of possibilities to identify different levels of discomfort. The reduction in complexity was motivated by both methodological and interpretive factors: binary classifications enable more stable model training, improve clarity in explanation, and enhance generalization. This is especially beneficial when employing interpretable machine learning techniques, which often suffer in terms of precision and transparency with increased class complexity.

To do so, we propose the *Thermal Comfort* (TC) variable as our objective variable, with a value of 1 when the answer given by the participant represented "*Thermal Comfort*" and 0 otherwise. To implement this codification in the RUROS dataset 3 main options appear for this binary recodification:

1. Defining TC as 1 when ASV is 0, "*Thermal neutrality*", and 0 otherwise.
2. Defining TC as 1 when ASV is between −1, "*cold*", and + 1, "*warm*", and 0 otherwise.
3. Directly using the *Comfort* variable, which was already binary, the answer to the question "*perception of comfort*".

To assess which one of these possibilities was the more reliable one, we analyzed the distributions of the derived TC variables. We compared these three codifications among them, *i.e.*, options 1 and 2 (TC_1, TC_2) and *Comfort* variable by itself. To execute this comparison, we deployed Fisher's Exact Test [70] and McNemar's Test [71], which regardless of the TC definition, consistently showed statistically significant differences for each pairwise comparison, with p-values below 10^{-16}. Therefore, the conclusion given by these tests was that the distributions of all three variables were different in a statistically significant manner, and not even the *Comfort* variable, that could have been used to double-check the other TC variables, could be used to do so, at least not using the proposed definition.

In particular, the differences between *Comfort* and the TC variables as presented could be explained by the difference between thermal and general comfort, highlighting the subjectivity of thermal comfort. Therefore, and not having statistically significant results to use one variable or the other, we decided to use the "strict" definition for the TC, previously presented as TC_1, because it provided a codification of OTC, as "thermal neutrality", *i.e.*, a value of zero for ASV, which clearly represented a situation of thermal comfort, compared to ranges that could be considered as discomfort, for TC_2, or not necessarily related to thermal comfort, as for the *Comfort* variable.

This binary approach is stricter than the classical one, in which OTC is classified into an ordered set of categories [11,23] for different ranges of comfort. However, it allows us to use certain models which would be impossible to deploy, and that will be shown in the following sections that are appropriate for this task. However, it is worth noting that while the UTCI index uses an equivalent temperature and from it a classification using ranges, in our case we will only use the binary classification for the reasons presented before, therefore addressing a subtask of OTC classification.

Regarding the comparison with the UTCI index, that will appear in Section 3.3, it is important to mention that, just as with the *Heat* variable, a binary recodification must be done for comparison. Therefore, using the UTCI classification, values labeled as "*no thermal stress*" are translated to "*comfort*", while all other values are considered *discomfort*. For completeness, we also include a broader UTCI comfort definition, the "extended UTCI" where UTCI classifications values span across "slight cold stress", "no thermal stress", and "moderate heat stress" are grouped under *comfort*. Effectively, this recodification means that our

Table 1
Number of entries and percentage of comfort (comfort answers / total answers) by city in the RUROS project.

City	Number of entries	Percentage	Percentage of comfort
Athens	1203	19.79	43.22
Cambridge	525	8.64	18.09
Fribourg	1313	21.60	46.91
Kassel	441	7.25	71.20
Milan	777	12.78	64.61
Sheffield	308	5.07	33.44
Thessaloniki	1512	24.87	34.26

UTCI range of comfort is, in UTCI terms, between 9 °C and 26 °C, and in the extended UTCI case, between 0 °C and 32 °C.

2.2. Machine learning

To develop our model, we crossed traditional elements of machine learning within the tidymodels framework [72], in the R language, after the proper processing of the RUROS using the Tidyverse framework [73], to provide a robust workflow for testing our configurations. Our workflow for model selection included bootstrap resampling, preprocessing techniques for data imbalance, model selection and hyperparameter fine-tuning through specific metrics. As a result, we obtained our proposed model, which we will study in Section 3.

2.2.1. Approach summary

We selected our final workflow, the combination of the model and the preprocessor, by means of the following procedure, using two different phases: one for selecting the best behaving model and a second one for giving the obtained metrics and the final training of the most suitable model.

In the initial phase, we followed the steps outlined in Fig. 1 to cross-validate the workflow and the chosen model for different combinations of cities. This involved using the candidate models and preprocessors for a final cross-validation using the entire dataset, which was split into train and test sets. It is worth noting that we considered six different combinations of train/test splits, manually leaving cities not included in the training dataset for testing purposes. The selection of the split ratio was based on a 70:30 ratio between the train and test sets, as suggested in previous studies [74–76]. Other works in literature have also employed similar ratios [12]. Our goal was to select the most suitable and general model, taking into account the variability of the dataset. We crossed each of the six models with each of the six possible preprocessors, resulting in a total of 36 workflows for initial testing. Each workflow, particularly each model, was fine-tuned to obtain the best possible configuration of the models' hyperparameters using a grid of size 10. We obtained key metrics such as balanced accuracy, recall, and specificity using bootstrap resamples and ordered the workflows based on their highest balanced accuracy.

In the second phase, that would give us our final proposal, our dataset was divided into a training set and a test set, using a ratio of 70:30 over the whole dataset. Then, the models and preprocessors that gave us the best results in the previous steps would be the ones chosen for the final model.

2.2.2. Models

Focusing on the proposal of possible interpretable alternatives to the UTCI index, we chose the models shown in Table 2. Even if other models such as Random Forest or XGB have provided better results in the area [12,48], the possibilities they provide towards explainability are reduced, and even with the aid of Shapley values interpretability is still narrow [63], compared to that of models which are explainable by

Table 2
Chosen interpretable models.

Name	Codification	Name	Codification
Logistic Regression	LR	Decision Tree	DT
Lasso Logistic Regression	Lasso	Linear Discriminant Analysis	LDA
Naïve Bayes	NB	Logistic Generalized Additive Model	GAM

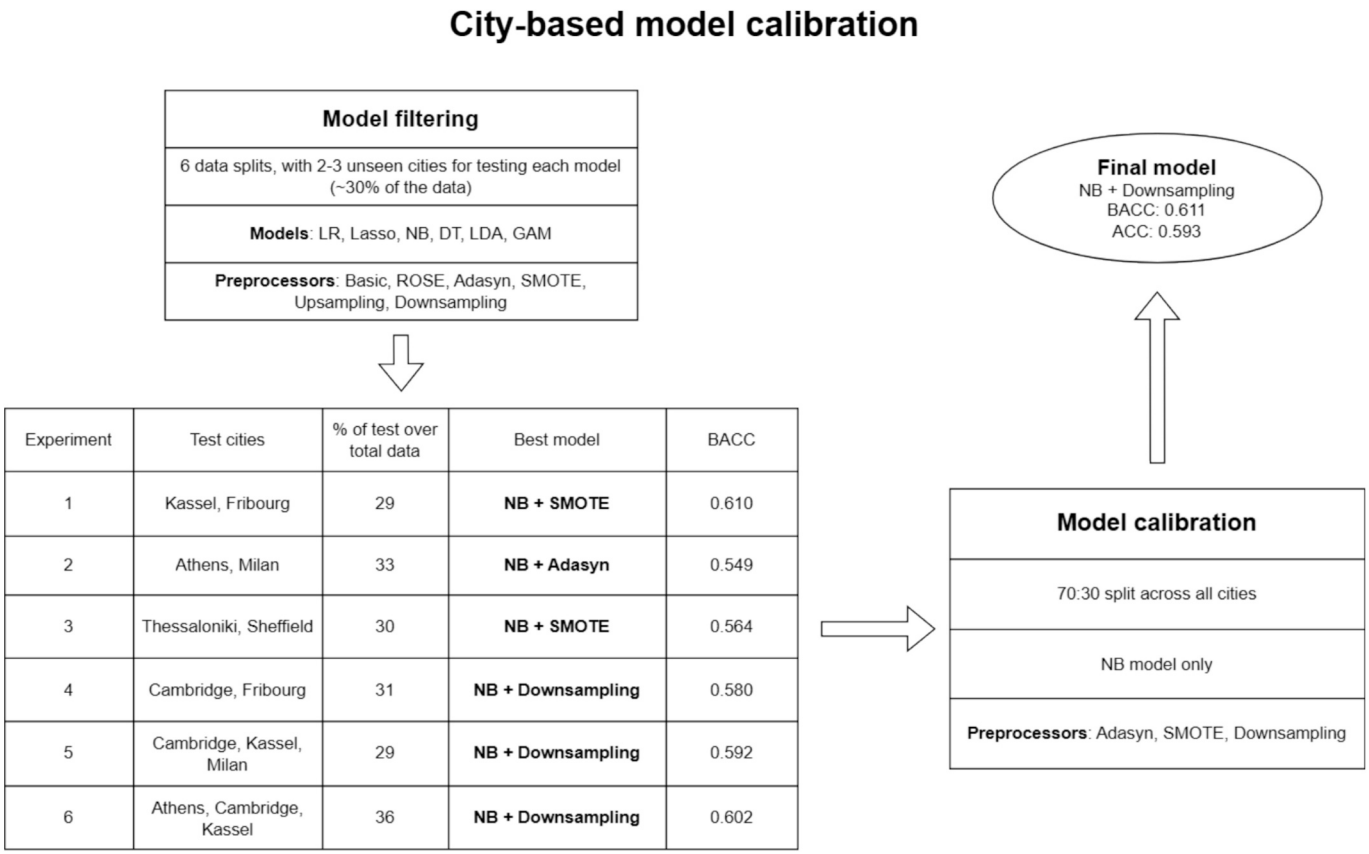

Fig. 1. Model development and calibration workflow. Key steps of the proposed approach are shown, including city–based model evaluation, selection of optimal model–preprocessing combinations, and final calibration of the Naïve-Bayes model with downsampling.

Table 3
Data imbalance in the RUROS with stratified cross-validation.

Subset	Comfortable?	Recount	Proportion
Train	No	2558	0.56
Train	Yes	2001	0.44
Test	No	853	0.56
Test	Yes	667	0.44

Table 4
Preprocessing settings.

Preprocessor	Description
Basic	Zero-variance, near-zero-variance filters and normalization
ROSE	Basic + ROSE resampling algorithm
Adasyn	Basic + Adasyn resampling algorithm
SMOTE	Basic + SMOTE resampling algorithm
Upsampling	Basic + Upsampling
Downsampling	Basic + Downsampling

Table 5
Tuned hyperparameters for the different models.

Model	Tuned parameters
Logistic Regression	None (basic logistic regression using glm)
Decision Tree	cost_complexity: Complexity parameter for pruning tree_depth: Max depth
Lasso Logistic Regression	penalty: Regularization strength mixture = 1: Lasso penalty only
Naïve Bayes	None explicitly tuned (based on data likelihood and priors, engine = klaR)
Linear Discriminant Analysis	None (uses MASS::lda, no hyperparameters to tune)
Generalized Additive Model	Automatically estimated smoothing parameters (mgcv handles penalization internally)

themselves. For completeness, some results regarding the use of these models over our dataset are included.

Therefore, we prioritized models that offer inherently interpretable behaviors, such as coefficients (as in Logistic Regression, Ridge and Lasso Regression, or Generalized Additive Models), decision rules (as in Decision Trees or RuleFit), and additive factors (as in Linear Discriminant Analysis). These models allow for a more transparent and direct understanding of the relationships between variables and outcomes, something that would be impossible using other options such as RF or XGB [57]. For a complete description of these methods one can consult [77,78].

After the process presented in the previous section, the model that provided the best results was the Naïve Bayes model, with a bootstrap-estimated balanced accuracy of 0.610 and a balanced accuracy of 0.611 over the test set.

2.2.3. Preprocessing

For each model and workflow, a corresponding preprocessing recipe (in the terms used by the Tidymodels framework) was selected. In total, 6 preprocessing options were considered, with 5 of them derived from the basic one, including as a variation the balancing methods. The basic preprocessing steps included zero-variance and near-zero-variance filters, as well as the normalization of all variables.

Due to the data imbalance, shown in Table 3, it was key to test whether some upsampling or downsampling method could provide an increase in accuracy, so we added the resampling steps associated with the Random Over-Sampling Examples (ROSE) [79], the Adaptive Synthetic Algorithm (ADASYN) [80], the Synthetic Minority Oversampling Technique (SMOTE) [81] algorithms and the basic upsampling and downsampling methods [72], in all cases using an over ratio of 0.75. A summary of the preprocessing steps can be found in Table 4Table, which were implemented using the Themis package [82].

2.2.4. Metrics

To evaluate model performance, various metrics were computed, such as balanced accuracy (BACC), accuracy, precision, recall, and specificity. These complementary metrics collectively offer a thorough insight into model performance regarding different types of errors. For a complete description for each metric, one can go to [83]. Balanced accuracy (BACC) was used over the traditional accuracy (ACC) because of data imbalance. While accuracy takes into account the overall correct predictions, it can be misleading if one class is dominant. BACC, defined as the average of recall and specificity, assigns equal importance to both classes, yielding a more reliable performance measure under imbalanced conditions, as shown in Table 1.

2.2.5. Fine-tuning

For all possible workflows it was key to do a fine-tuning step so that the best combination of hyperparameters was chosen. For each model, the list of available hyperparameters can be found in Table 5. During this step, using the bootstrap metrics for each workflow the best hyperparameters were chosen, using a grid of size 20 in a suitable range, as in [72]. We considered other metrics to guide our workflows, but the data imbalance showed us that using any other metric but the balanced accuracy provided extremely biased workflows, that would either answer always “Comfort” or “Discomfort”.

2.2.6. Workflows

Up to this point, the proposed workflows served as an integrated framework to evaluate different combinations of models and preprocessors under the same conditions, for different sets of test cities. Through cross-validation and fine-tuning, each workflow was assessed using balanced accuracy. Following this approach, we obtained the results shown in Table 6, from which our resulting workflow was composed of the Naïve Bayes model with the downsampling preprocessor. In that table we can see that, for the chosen metric, the balanced accuracy, almost all workflows, after fine-tuning, showed similar results. Therefore, it was key to analyze other metrics, to provide a reasoned

Table 6
Key metrics for the selection of the best models and preprocessors during the first phase.

Test cities	Train:Test ratio	Chosen model	BACC	UTCI BACC
Kassel, Fribourg	71:29	NB + SMOTE	0.610	0.640
Athens, Milan	67:33	NB + Adasyn	0.549	0.577
Thessaloniki, Sheffield	70:30	NB + SMOTE	0.564	0.549
Cambridge, Fribourg	69:31	NB + Downsampling	0.580	0.660
Cambridge, Kassel, Milan	71:29	NB + Downsampling	0.592	0.582
Athens, Cambridge, Kassel	64:36	NB + Downsampling	0.604	0.606

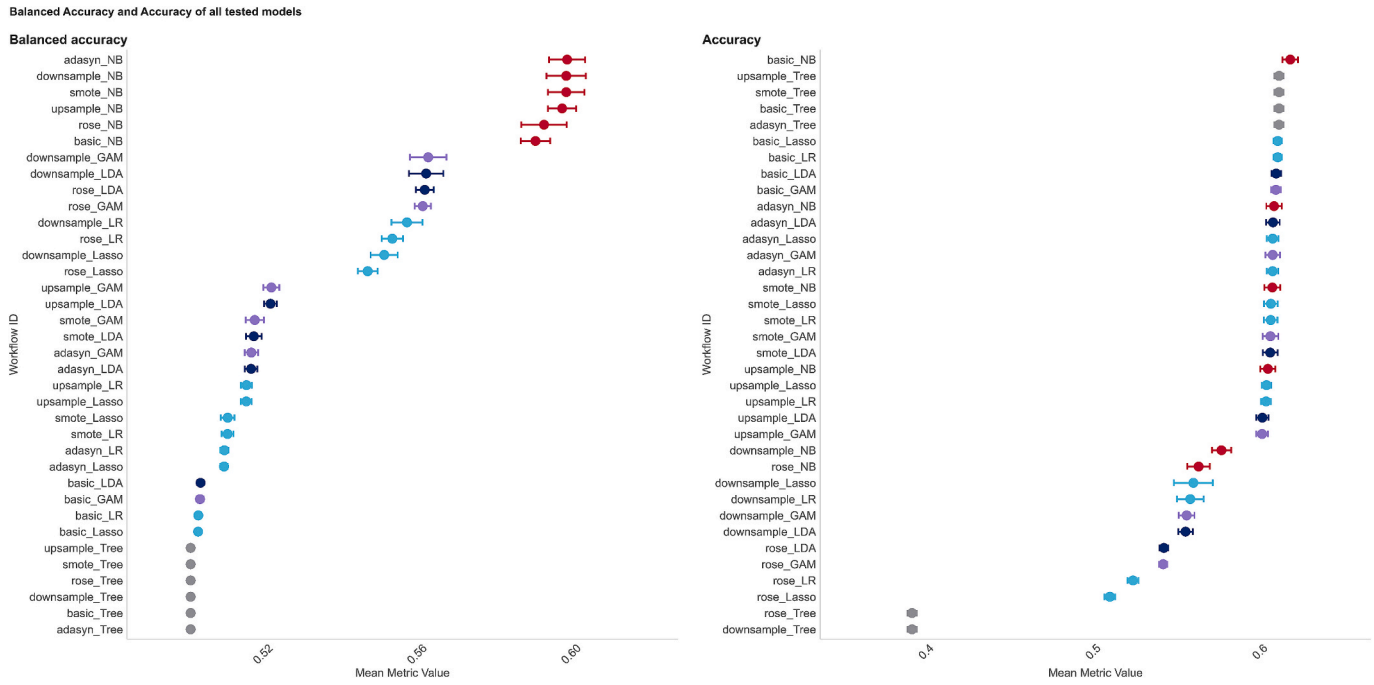

Fig. 2. (a) Balanced accuracy and (b) accuracy performance, using the bootstrap estimates for the different workflows, ordered by performance on the X-axis, leaving the data from Fribourg and Kassel for testing.

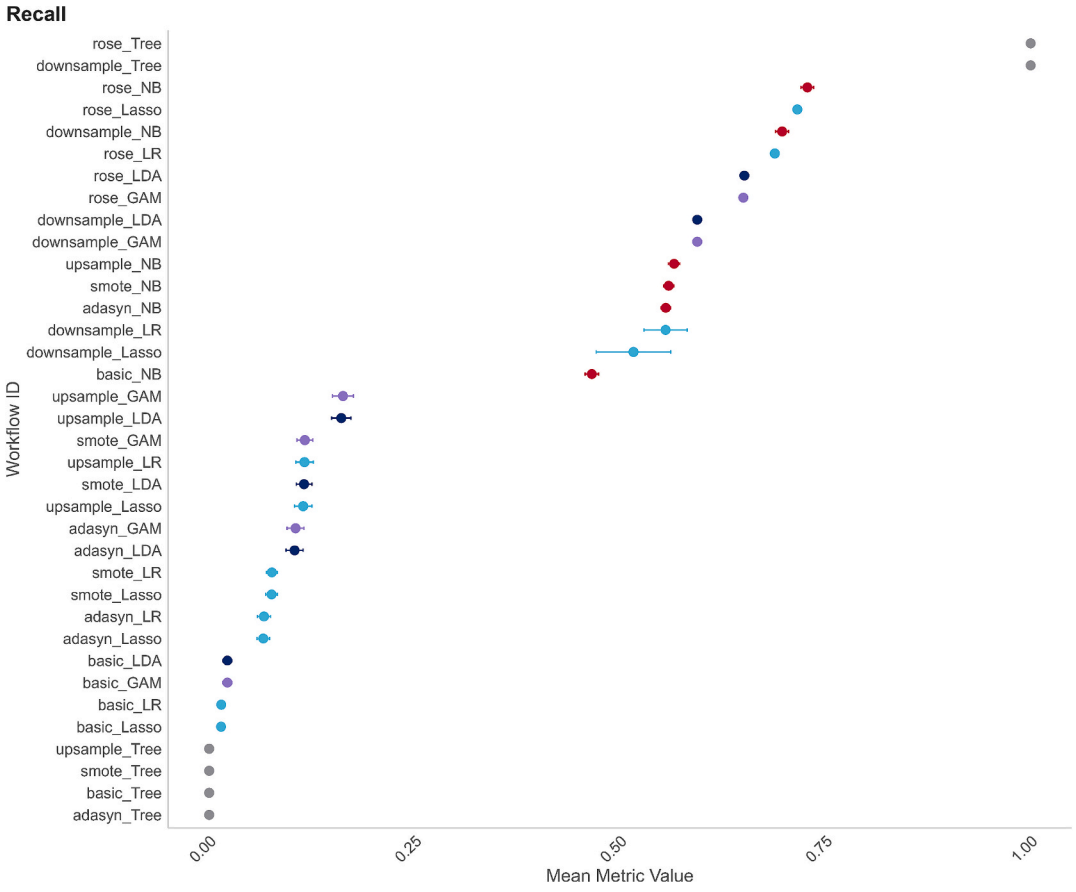

Fig. 3. Recall estimates for the different workflows, ordered by performance on the X-axis, leaving the data from Fribourg and Kassel for testing.

selection of the final workflow.

As a way of obtaining this additional information and justifying the selection of our metric, we plotted in Figs. 2 and 3, for the particular train/test split in which Fribourg and Kassel were left out for testing, the key metrics obtained during the training of the workflows, for each possible combination of model and preprocessor, therefore showing that in all cases most models provided similar results, with slight changes in balanced accuracy and accuracy, as expected.

3. Results

In the following section, the main results of our work are presented, as the output of the workflow described in the previous section. We obtained, as already mentioned, the Naïve Bayes with the downsampling technique as the final model, model that henceforth will be called NBD for simplicity.

3.1. Proposed model

In all cases, it was seen that the model providing the best results, using the Bootstrap estimator over 5 samples with 25% of the training set for validation, was the Naïve Bayes, with a mean balanced accuracy (BACC) across samples of 0.58, compared to that of the UTCI classification of 0.60. The preprocessor, however, varied across samples: Adasyn was chosen once, SMOTE twice and downsampling thrice. Therefore, we chose the Naïve Bayes algorithm as well as these 3 preprocessors as the options for our final model. The key results of this phase can be seen in Table 7, as well as a comparison with the BACC obtained by the UTCI index, giving a classification of "*comfort*" when the UTCI value is between 9 and 26 °C, *i.e.*, the range for the stress category "*no thermal stress*", and "*discomfort*" otherwise. It is noteworthy that this UTCI classification yields better results in 4 out of 6 experiments, something that can be expected knowing that it is a global index, while our models have only access to a reduced amount of data from certain cities and are testing on completely different ones. For completeness, we will also consider the extended UTCI classification, in which the categories labelled by UTCI as "*moderate heat stress*" (26–32 °C), "*no thermal stress*" (9–26 °C) and "*slight cold stress*" (0–9 °C) are all merged into "*comfort*". The results given by this alternative can be seen in Table 7, but as it shows worse results than those of the "strict" UTCI, it is not further analyzed.

To fully understand our selection of the objective metric, the Balanced Accuracy (BACC) and the results shown in Fig. 2a, it can be useful to consider Fig. 2b, for the Accuracy of bootstrap estimates, and Fig. 3, for the Recall estimates. When looking at the accuracy estimates, one can identify that even if the Naive Bayes model in its basic setting is performing better than other models, multiple classification tree models with different preprocessors are clearly close in performance. This, however, doesn't correspond at all with the results shown in Fig. 2a, in which the tree models obtain an extremely low balanced accuracy.

The underlying problem these models are showing resides on the imbalance of the original data. Even when using algorithms such as the ones previously described in Section 2.2.2, some models like the classification trees cannot take that imbalance into account. This problem was shared by other models, LDA, GAM, and LR, when using the basic preprocessor, that didn't include any sort of modification to the data to address the data imbalance. The explanation for this problem, in the tree model case, can be clearly seen in Fig. 3, in which the recall, the true positive rate, is shown for the same experiments. In this figure, the tree models obtain either 0 or 1 in recall. This is something that in practice can only be obtained by models that predict the same in all cases, in our case, either "Comfort" or "Discomfort".

To confirm this theory, other metrics such as precision, recall, and specificity were considered. In all cases, just as shown before for the tree models, even with the preprocessors addressing the balance issue, we obtained biased models, as shown by extreme values of 0 and 1 in one of the mentioned metrics. This allows us to justify our decision to use the BACC metric instead of the usual accuracy, as it could be biased towards models not considering the data imbalance. For completeness, in the Appendix the plots for the precision and the specificity can be seen, obtained under the same conditions as the previous plots.

As a result of the previous phase, the Naïve Bayes algorithm was tested with the preprocessors chosen in the previous step, using the same Bootstrap validation split for obtaining the best combination of tunable parameters in the model. Knowing the best combination of tunable parameters and preprocessors, all options provided similar results, showing the stability of the Naïve Bayes technique, but in the end downsampling showed the best results in the training set, being selected then as the final model for our proposal. It is noteworthy that, even if similar results were obtained with other preprocessors, downsampling is way faster and easier, avoiding the overfitting that other resampling methods such as the SMOTE or the ROSE algorithms could carry.

3.2. NBD results

As shown in Table 7, the performance indicators for the NBD model suggest that it has an appropriate capability in correctly classifying thermal comfort. The balanced accuracy of 0.611 shows a slight improvement over the results in Table 6, which ranged from 0.549 up to 0.610, manifesting the advance the model experienced when using more general data. With this result, the model shows its ability to differentiate between comfort and discomfort, while considering class imbalances at the same time. Its overall accuracy of 0.593 indicates that the model gets the prediction (either comfort or discomfort) right about 59% of the time. The precision of 0.526 implies that a little more than half of the comfort predictions are accurate, hinting at a propensity for false positives. Nevertheless, a recall rate of 0.756 signifies that the model excels in detecting actual cases of comfort, successfully identifying around 76% of them. Conversely, a specificity of 0.467 indicates a limited ability to recognize discomfort, revealing that the model often fails to detect negative (uncomfortable) conditions. In other words, the model is more effective at identifying comfort than at maintaining a balanced detection of both comfort and discomfort.

These results, while showing moderate improvement, reveal an important limitation: the overall accuracy of any method tested—ours or the UTCI-based baselines—rarely surpasses 60%. This raises a legitimate concern: despite considerable modeling effort and complex tuning, the gain in predictive performance appears modest. However, it is precisely in this context that a fair comparison with the UTCI becomes essential. The similarity in accuracy between the machine learning models and the UTCI-based classifications suggests that both approaches are effectively leveraging the available input variables to their fullest extent. Rather than reflecting model weakness, the limited ceiling in accuracy likely speaks to an intrinsic challenge in the task itself. Thermal comfort perception is influenced not only by measurable environmental and physiological factors but also by highly subjective, personal, and situational elements that are not easily captured and modelled. Therefore, while the accuracy may seem modest in absolute terms, it represents a reasonable upper bound given the nature of the problem, and it validates the effectiveness of both modeling strategies in interpreting the available data.

Table 7
Main classification metrics for NBD, UTCI and extended UTCI.

Metric	NBD	UTCI	Extended UTCI
Balanced accuracy	0.611	0.585	0.563
Accuracy	0.593	0.563	0.524
Precision	0.526	0.501	0.477
Recall	0.756	0.766	0.888
Specificity	0.467	0.404	0.239

Compared to the results reported by Rodríguez-Gallego et al. (2024), our model shows comparable overall accuracy (0.593 vs. 0.59 with Random Forest and MLP) but significantly higher sensitivity (0.756 vs. 0.34 and 0.20, respectively), indicating a stronger ability to detect comfort cases. However, it's important to note that these previous models were not designed specifically for binary classification of thermal comfort, which limits the direct comparability. This issue is even more pronounced in other recent works, such as those by Tian & Lin [44] and Avci [48], where multi-class approaches are adopted to reflect finer-grained thermal sensation scales. As such, while performance metrics may appear similar on the surface, differences in classification strategy and target definitions make strict comparisons challenging.

3.2.1. Non-explainable models

For completeness, other non-explainable models were used to provide further comparison. In particular, the Random Forest (RF) and the Extreme Gradient Boosting (XGB) were tested in the same manner, using only downsampling and with the same workflow settings (cross-validation, fine-tuning) as the previously presented NBD. Notice that this was not the key objective of this article, but it is important to include these results to see the baselines provided by black-box models. The results for these models can be found in Table 8, where the results for NBD are included again for comparison.

The table previously discussed shows that our NBD model achieves a balanced accuracy (BACC) of 0.611, which is comparable to the more complex Random Forest with SMOTE (0.650) and XGBoost with ADASYN (0.662). In terms of overall accuracy, the NBD model scores 0.593, again slightly lower than RF (0.655) and XGB (0.663). Despite this modest disparity, there's a significant trade-off involved: ensemble methods such as RF and XGB function as black-box models, restricting interpretability and complicating the comprehension of variable influence. Conversely, the probabilistic framework of NBD offers transparent decision boundaries, enabling a clear insight into how each feature contributes to comfort classification. This clarity is particularly beneficial in environmental and public health contexts, where explainability is crucial.

NBD notably achieves the highest recall rate (0.756) among all models, exceeding both RF (0.607) and XGB (0.651), which demonstrates its superior capability to identify discomfort scenarios—a critical feature for early-warning or thermal risk applications. Its precision (0.526) and specificity (0.467), though lower compared to RF (0.607, 0.693) and XGB (0.609, 0.673), indicate a strategic preference for detection over reduction in false negatives, an appropriate choice for preventive decision-making. In conclusion, while ensemble methods boast slightly improved overall metrics, NBD offers a balanced and interpretable alternative, ensuring solid predictive power, excelling in pinpointing critical discomfort instances, and promoting clear, evidence-based evaluations of Outdoor Thermal Comfort (OTC).

4. Discussion

In this section, we evaluate the performance of our Naïve Bayes with Downsampling (NBD) model against the Universal Thermal Climate Index (UTCI). Both approaches are compared under a consistent binary classification framework and standard performance metrics, using confusion matrices and key indicators such as balanced accuracy, recall, specificity, and precision. Our goal is to determine whether we achieve measurable improvements with a data–driven, interpretable model relative to the threshold–based UTCI, while maintaining practical relevance for the classification process.

Table 8
Classification metrics for RF and XGB.

Metric	NBD	RF + SMOTE	XGB + Adasyn
Balanced accuracy	0.611	0.650	0.662
Accuracy	0.593	0.655	0.663
Precision	0.526	0.607	0.609
Recall	0.756	0.607	0.651
Specificity	0.467	0.693	0.673

4.1. UTCI comparison

As seen in Table 9, in comparison to the UTCI classification, the NBD model demonstrates a slightly more effective ability for discerning comfort versus discomfort, as evidenced by its increased balanced accuracy of 0.611 compared to 0.585. UTCI exhibits a small advantage over NBD in recall (0.766 compared to 0.756), signifying a slightly superior capability in recognizing comfort scenarios. However, NBD compensates for this with a significantly greater specificity (0.467 against 0.404), implying a more reliable identification of discomfort conditions. NBD demonstrates a higher precision at 0.526 compared to 0.501 for UTCI, indicating fewer incorrect comfort predictions. Regarding overall accuracy, NBD also outperforms UTCI, scoring 0.593 against 0.563. This contributes to a slight but overall improvement in key classification metrics and abilities. Notice that the extended UTCI was not included in these sections as its metrics were worse than that of our binary classifier version of the UTCI.

When contrasting NBD with the extended UTCI classification, the differences are more marked, due to the limitations the latter presents. Although the lenient UTCI boasts the highest recall (0.888), indicating it identifies nearly every comfort case, this comes with a trade-off of significantly low specificity (0.239) and balanced accuracy (0.563). This suggests a pronounced tendency towards predicting comfort, which may lead to an increase in false positives. Conversely, NBD provides a more balanced and accurate model, resulting in a superior trade-off between identifying comfort and distinguishing discomfort.

4.1.1. Statistical validation of NBD

To assess if our method, the Naïve Bayes with Downsampling (NBD) model's observed enhancements over the UTCI baseline are statistically significant, we conducted formal significance testing on performance metrics derived from repeated cross-validation and bootstrap resampling.

Specifically, we employed the Wilcoxon signed-rank test [84] to compare paired accuracy and balanced accuracy scores across folds, as this non-parametric test does not require normally distributed differences. Furthermore, we calculated 95% bootstrap confidence intervals for the mean differences to measure the uncertainty in performance improvements. A p-value under $1.25 \cdot 10^{-5}$ was considered a sign of statistically significant improvement, enabling us to determine that there was a statistically significant difference between the capabilities of these two models.

Using a bootstrap estimate [85] over 2000 resamples built over the resulting predictions for both the UTCI and our NBD, we obtained a 95% confidence interval that showed an improvement between 2.43% and 6.51%. Therefore, for binary OTC classification, it can be concluded that the NBD, trained using balanced accuracy, provides a consistent and substantial benefit over UTCI rather than a random fluctuation.

4.2. Comparison with other indices

In Table 10 we provide a comprehensive comparison that includes our model's outcomes alongside those from the NBD, as well as respective results for two indices: PET and PMV [11,18]. For transforming PET and PMV into binary classifications, thresholds rooted in established comfort standards were employed. Consistent with the ISO 7730 guideline [19], PMV readings from −0.5 to 0.5 were deemed “Comfort,” with values beyond this range classified as “Discomfort.” The PET range

Table 9
NBD vs. UTCI. Confusion matrices for train (left) and test (right) subsets.

NBD	Real comfort	Real discomfort	NBD	Real comfort	Real discomfort
Predicted comfort	1466	1254	Predicted comfort	504	455
Predicted discomfort	535	1304	Predicted discomfort	163	398
UTCI	Real comfort	Real discomfort	UTCI	Real comfort	Real discomfort
Predicted comfort	511	508	Predicted comfort	592	649
Predicted discomfort	156	345	Predicted discomfort	75	204

Table 10
Main classification metrics for NBD, PET and PMV.

Metric	NBD	PET	PMV
Balanced accuracy	0.611	0.518	0.499
Accuracy	0.593	0.544	0.538
Precision	0.526	0.450	0.420
Recall	0.756	0.352	0.240
Specificity	0.467	0.684	0.757

was selected to be from 18 to 26 °C. This range was chosen following previous research: Matzarakis et al. [86] identified 18–23 °C as "comfortable" for Central European climates, Matzarakis et al. [87] extended this range up to 26 °C, and Blazejczyk et al. [88] similarly assumed PET comfort intervals between 18 °C and 26 °C across various European locations. These adjustments ensured uniform binary classification among all indices prior to performance evaluation. The *ladybug-comfort* Python library facilitated the calculation of these indices for our dataset [89].

Based on the analysis, the NBD model showed superior predictive prowess in most metrics. It scored the highest in balanced accuracy (0.611), precision (0.526), recall (0.756), and overall accuracy (0.593), surpassing the results of PET and PMV, which had lower scores in these categories. Although the NBD model's specificity (0.467) was less than PMV (0.757) and PET (0.684), its substantially higher recall underscores its enhanced capability to detect "Comfort" instances within the dataset. These findings suggest that the NBD model offers more dependable classification of thermal comfort conditions compared to PET and PMV, which are based on conventional comfort thresholds. Notably, both PMV and PET were more inclined to predict "Discomfort," as indicated by their high specificity and low recall. This problem, while present in our model, is more pronounced in these two indices.

4.3. Interpretability

One of our objectives was to provide an interpretable tool, using the same variables as the UTCI index, but showing further explanations about its inner workings. As the chosen model has been the Naïve Bayes method, this is achieved as it provides explainability through its simple probabilistic structure, allowing us to understand how each variable or feature contributes to the prediction by the values of the conditional probabilities. Knowing that this method uses a posterior estimate of the probability of belonging to a certain class (comfort/discomfort) proportional to the fixed prior probability and the likelihood of the observation, we can analyze both the prior estimate as well as the likelihoods of each feature given the class and identifying which variables most influence the classification into comfort or discomfort. This helps us understand the patterns that dictate the moel's decisions.

To address the substantial correlation between air temperature and globe temperature ($r = 0.94$), which somewhat compromises the model's independence assumption, an extra iteration was executed that omitted globe temperature, called NBD3 to remember that we are only using three variables. This alternative analysis maintains the conditional independence structure, ensuring the results remain interpretable while allowing for comparisons with UTCI-based models, which utilize all four variables.

Table 11
Main classification metrics for NBD and NBD3.

Metric	NBD	NBD3
Balanced accuracy	0.611	0.617
Accuracy	0.593	0.603
Precision	0.526	0.534
Recall	0.756	0.738
Specificity	0.467	0.497

As we can see in Table 11, training the model without the globe temperature slightly improved our results. In particular, balanced accuracy increased from 0.611 to 0.617, overall accuracy rose from 0.593 to 0.603, and precision improved from 0.526 to 0.534, indicating a modest gain in correctly identifying both classes. While recall decreased slightly from 0.756 to 0.738, the increase in specificity from 0.467 to 0.497 suggests that NBD3 is better at detecting uncomfortable conditions, leading to a more balanced classification performance overall. These changes highlight that removing the globe temperature did not hinder the model and may even enhance its robustness in distinguishing between comfort and discomfort.

In Fig. 4 we can see the conditional distribution of comfort across the different features considered in our model, representing the likelihood of a certain value being observed depending on the comfort classification. It can be seen that it is in the middle ranges when the different features provide the highest effect on the classification of comfort, with the case of globe temperature, air temperature and relative humidity showing and increased probability of comfort for the mean values. In particular, both relative humidity and air temperature have a noteworthy effect on the probabilities when showing small values for the first and high for the second, accounting for the effect of dry conditions and extreme heat.

As a further visualization towards the understanding of the model, Fig. 5 illustrates how the NBD model predicts comfort probabilities across air temperatures (*tair*). The model performs well at temperature extremes and around the neutral comfort zone (18–20 °C), where classification is clearer. However, it struggles in transitional ranges (e.g., 12–15 °C and 20–25 °C), where comfort perception is more variable. This highlights both the strengths of Naive Bayes in modeling clear patterns and its limitations in more ambiguous regions, while still offering interpretable outputs useful for analysis.

Finally, to evaluate the model's confidence in its predictions, we can plot the distribution of predicted probabilities for the "Comfort" and the "Discomfort" classes. The density curves shown in Fig. 6 illustrate how often the model assigns probabilities along the range from 0 to 1. Knowing that for comfort our likelihood will be of around 0.6 and 0.7 most often, our model will be more often confident about predicting comfort over discomfort, which will return with a likelihood of 0.4 usually. In particular, the overlap between the curves shows the

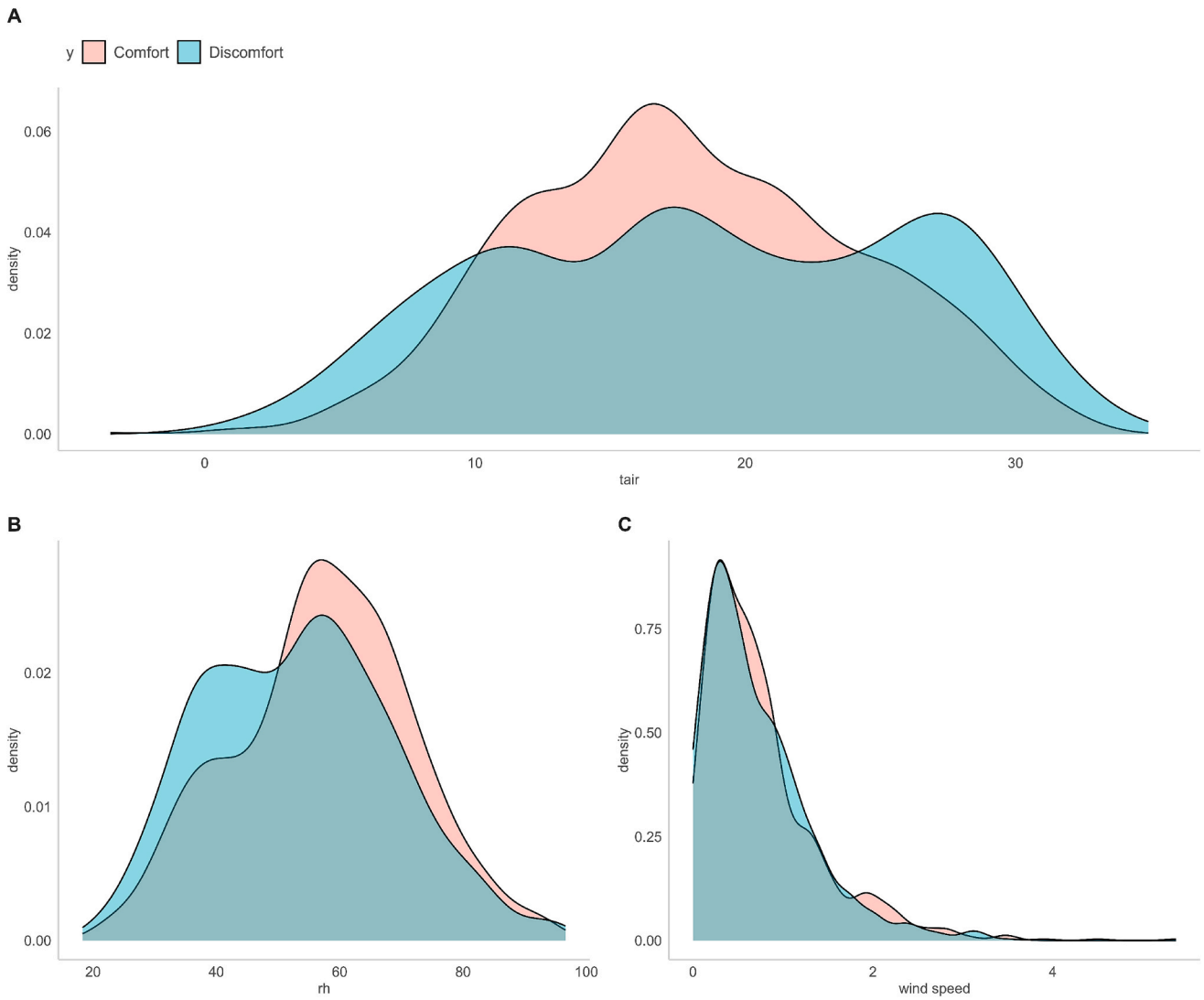

Fig. 4. Conditional distribution of Comfort across the 3 microclimatic variables used in NBD3: a) Air temperature; b) Relative humidity and c) Wind speed.

uncertainty this model suffers, which will be further analyzed in section 3.7.

4.4. Real-world applications

The performance shown by our NBD model allows it to be used as a real and alternative tool for OTC prediction, offering an alternative to the UTCI. By offering probabilistic predictions rather than fixed-threshold classifications, NBD enables urban planners to translate likelihood estimates into spatial heat-risk maps that highlight where and when thermal discomfort is most probable. This approach, based on probability, offers more flexibility, helping adaptive planning strategies by pointing out key areas based only on meteorological data. Therefore, this model can be directly fed into risk-based decision frameworks, enabling planners to allocate resources according to projected discomfort probabilities rather than static index boundaries.

Converting outdoor thermal comfort into a binary outcome ("comfortable" vs. "uncomfortable") simplifies the original multiclass scale for OTC. This reduction trades some granularity (e.g., distinctions between "slightly warm" and "very hot") but allows the models to work with probabilities and to offer more robustness against variability. Apart from the methodological benefits, this binary classification is also driven by practical reasons. In the realms of public policy and urban planning, decision-makers frequently need a straightforward and actionable benchmark to determine if outdoor spaces are comfortable for the majority of users. This simplification allows authorities to pinpoint key areas for intervention, evaluate adherence to thermal comfort standards, and direct urban design strategies aimed at enhancing well-being in public areas. Consequently, adopting a binary framework not only enhances the interpretability of models but also aligns research findings with the requirements of policy execution and urban management.

The current NBD configuration offers a balanced compromise—providing interpretable, probabilistic, and readily deployable predictions that strengthen the bridge between thermal comfort modeling and urban-scale planning practice.

4.5. OTC challenges

To understand the limitations of OTC classification, as well as the results obtained by all the models presented in our work, it can be key to

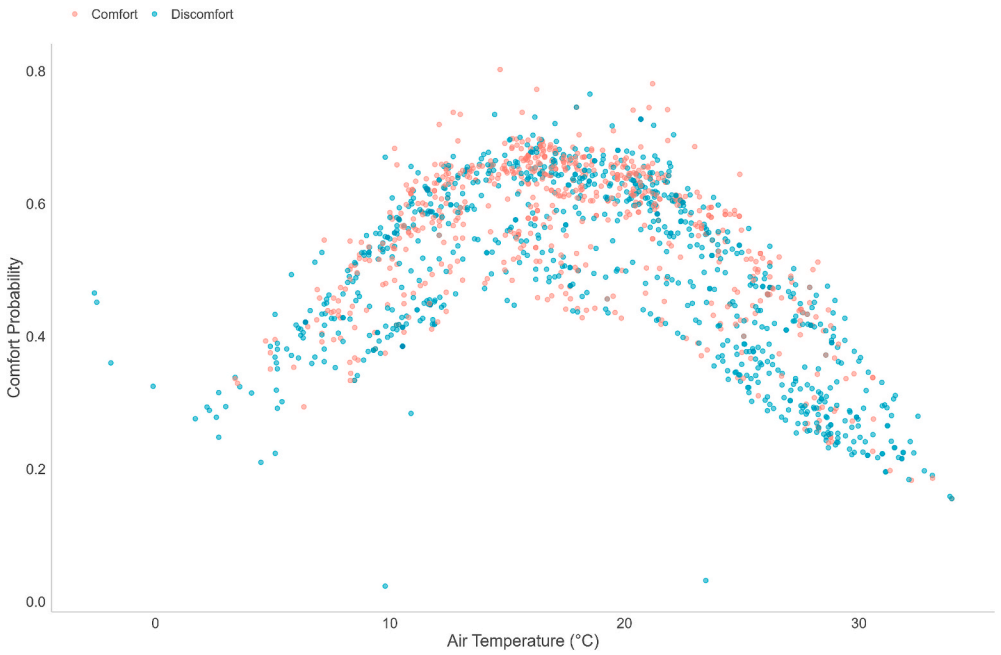

Fig. 5. Comfort probability prediction obtained with NBD3 depending on air temperature (°C), for each sample in the test set.

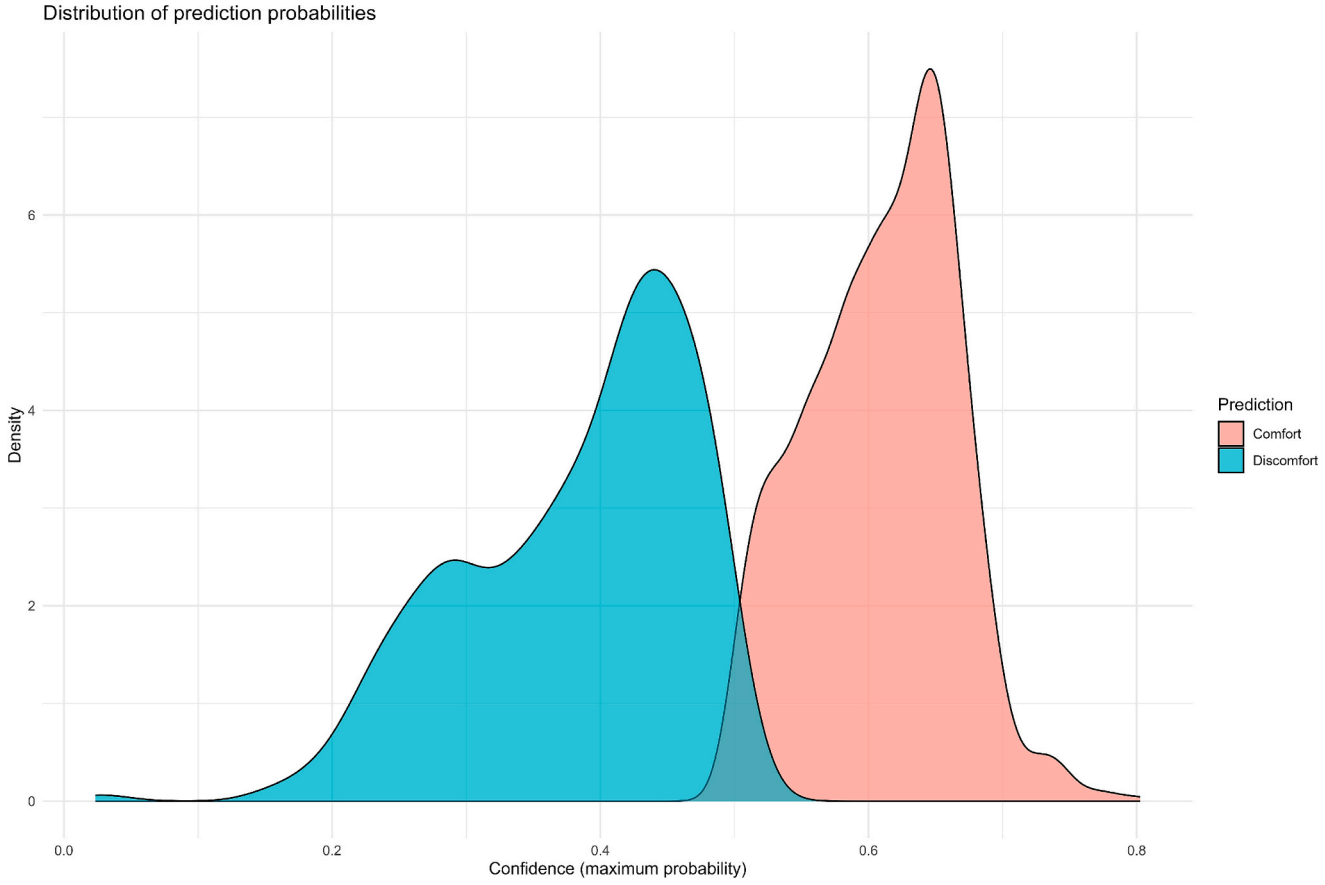

Fig. 6. Conditional distributions for comfort and discomfort for the NBD3 model.

see the variation of the comfort perception over small changes in key variables, an issue that can greatly limit the capabilities of the usual models.

Analyzing the distribution of comfort proportions across narrowly defined temperature and relative humidity categories, as illustrated in Fig. 7, reveals noticeable variability even within groups of small size. The comfort proportion within each category can differ significantly, even among those with comparable environmental conditions. In

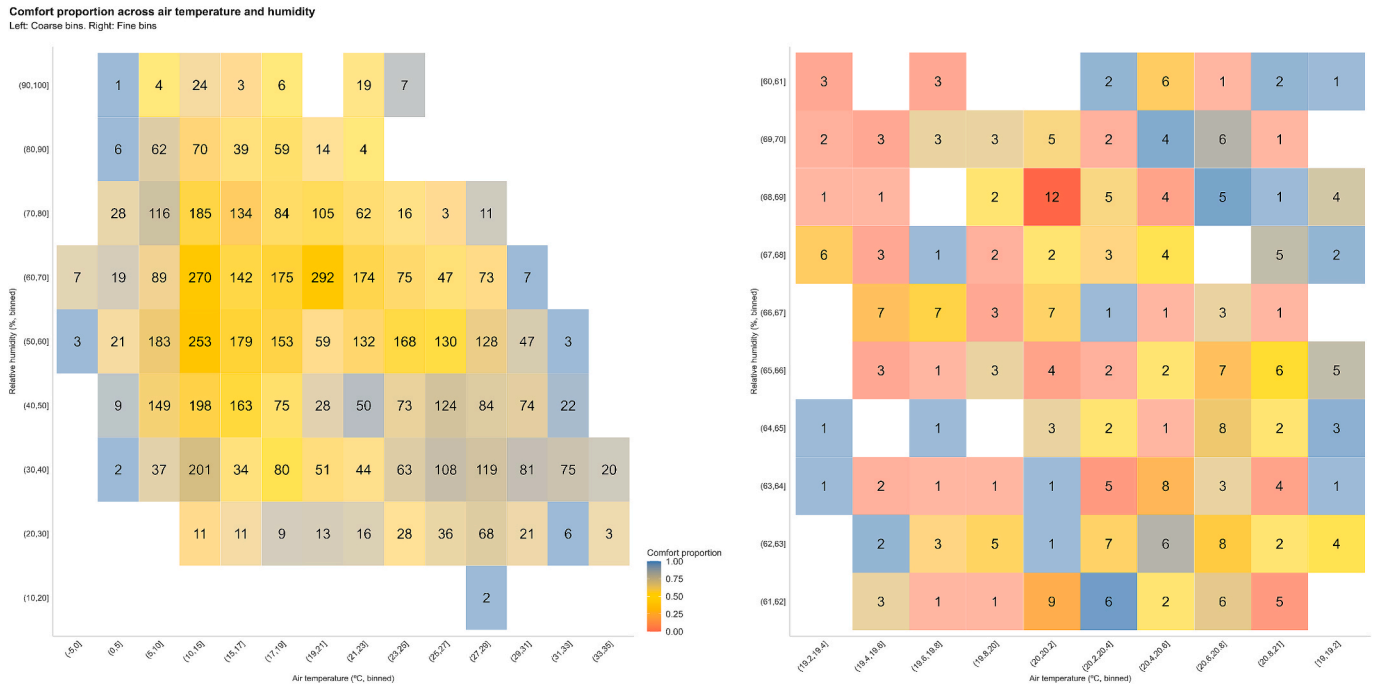

Fig. 7. Comfort proportion across (a) binned and (b) finer-binned air temperature (°C) and relative humidity (%). Full dataset.

particular, for air temperatures between 15 and 21 °C and relative humidity of 50 to 60%, we can see that the proportion of people in comfort is closer to 0.5 than it is to extreme values (0 or 1). This can greatly affect the models built over this data, as even for relatively small bins the variability in perception is still present.

In this case, again, we find areas in which the proportion of comfort is extremely close to 0.5, while at the same time showing a great number of values, especially for air temperatures between 17 and 21 °C and 50 to 70%. The same situation can be seen for the crossing of air temperature and wind speed in Fig. 8.

The overlap shown in these crossings of variables suggests that environmental factors alone cannot accurately predict comfort, emphasizing the inherent uncertainty in anticipating human thermal perception. Consequently, the variability limits our model's performance, and the concentration of observations in specific categories influences prediction of confidence. While the model can capture overall trends, the overlap and scarcity in certain categories place a ceiling on the potential accuracy and balanced performance, leading to some unavoidable misclassifications. This can partially explain the limitation of all the considered models including UTCI, NBD, NBD3 and non-explainable models, with the apparent optimal value for balanced accuracy of 0.6 approximately. However, even when using additional variables from the dataset, our experiments showed an increase in accuracy below 0.05, for both the NBD and the NBD3 models.

For a complete analysis of the problem, we also executed a gradient analysis for the crossing between air temperatures and relative humidities. For this, we define it as the mean absolute difference between the proportion of comfort between one cell and its neighbors, understanding cell in the sense given by Fig. 7.

In Fig. 9a we represent, using the same wide bins as in Fig. 7a, the obtained gradients in each cell. The proportion of comfort changes in a rather smooth manner, with most gradient values close to 0. This behavior changes only for extreme cases, the edge cases like the reduced temperatures below 0 °C or with a relative humidity over 80%. Comparing this plot with Fig. 7aFigure, we can see that these edge cases are a minority, with a change in comfort proportion highly related to the number of samples in that bin. This distribution of values and gradients can be confirmed by looking at a finer scale.

In Fig. 9b we provide the gradient plot for the same bins used in Fig. 7a, for representing a finer grid over the crossing of 60–70% of relative humidity and 19–21 °C for air temperature. Confirming our previous hypothesis observed for the edge cases in the wide grid, when we reduce the scale and the number of samples considered, the variability greatly increases. We can observe that in many cases the mean gradient abruptly changes from values close to 1 to others extremely close to 0. This behavior is hidden on the broader scale but using the finer grid we can assess that even for a small scale, the gradient shows significant changes in cases extremely close to each other. For any model trained on this data, these abrupt changes limit the potential for an accurate prediction using these variables.

Knowing that the data from the RUROS dataset was taken under different circumstances (e.g. seasons, locations, cultures), can partially explain this variability in a range in which additional variables can have a key effect on the OTC. The ranges that were studied in this section represent an area in which both OTC perception and UTCI show extreme variability too. As shown, using only the selected variables can explain up to a certain point. It is noteworthy, however, that we run the same experiments with more variables, and the increase in accuracy, less than 0.05, was not worth the loss of explicability. To do so, we tested both the same workflow and a RF model, with the same variables as well as with the whole dataset, including variables such as the clothing thermal insulation, age and the metabolic activity, while at the same time avoiding redundant variables. Therefore, to account for the subjective behavior behind this variability, obtaining an even bigger dataset would be key.

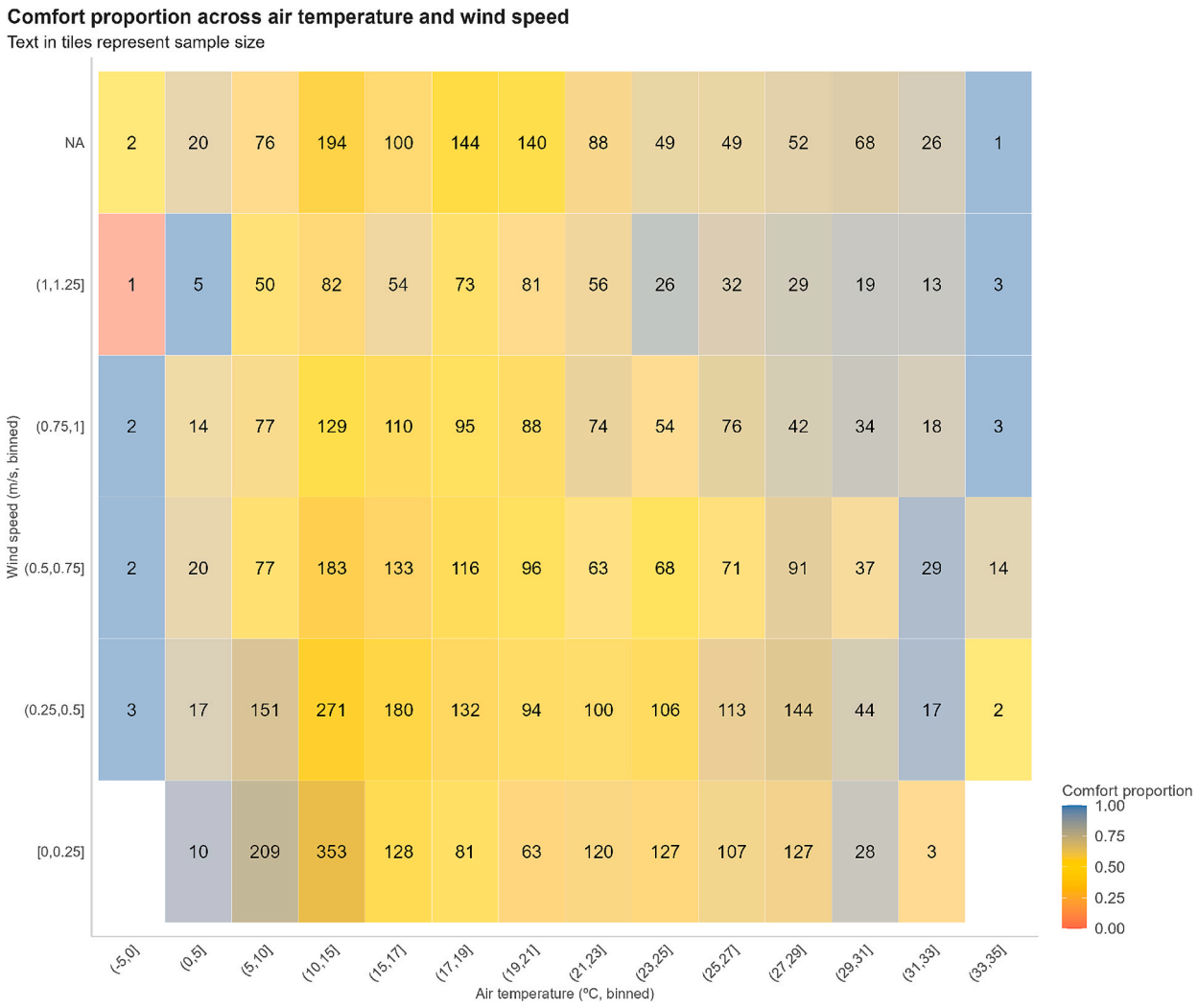

Fig. 8. Comfort proportion across binned air temperature (°C) and binned wind speed (m/s). Full dataset.

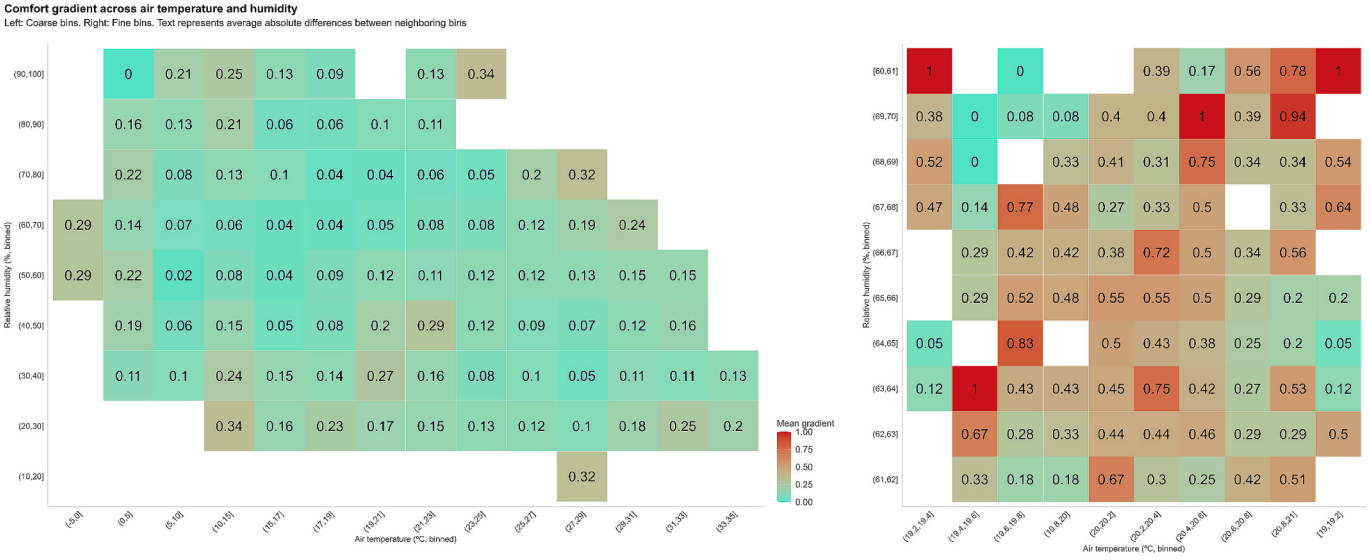

Fig. 9. Gradient of the comfort proportion across binned and finer-binned air temperature (°C) and relative humidity (%). Full dataset.

5. Conclusion

In this study, the RUROS dataset was used to evaluate and compare different workflows against the classification given by the UTCI index, obtaining as a result a Naïve Bayes downsampling (NBD) model that provided competitive results when predicting the OTC. The goal was to provide a more interpretable alternative to the UTCI index, giving insight into its classification for Outdoor Thermal Comfort (OTC).

5.1. Achievements

After evaluating and fine-tuning 6 different interpretable models and their combinations with 7 possible preprocessors in different configurations of training and test splits for the RUROS according to geographical location, it was shown that in all cases the Naïve Bayes model outstood the rest of the models, for 3 different preprocessors. Then, we compared the NB model and the 3 chosen preprocessors using a bootstrap estimate over the training set, now using 75% of the data from all cities. The combination that provided the best results was the NB with downsampling, which was then used as our final model to analyze over the test set.

Our NBD model demonstrates improved results compared to the traditional UTCI index when both are reduced to a binary classification problem of comfort versus discomfort, both in the strict case —comfort equals thermal neutrality— as well as in the loose case —comfort is both neutrality and slight discomfort. Notably, the model achieves a balanced accuracy above 0.611 and a high overall accuracy of 0.593, indicating good sensitivity to identifying comfort and discomfort cases, compared to the results given by the UTCI index for the same task, with a balanced accuracy of 0.585 and an accuracy of 0.563. These results confirm that probabilistic models can capture meaningful patterns in the data when paired with appropriate preprocessing, such as downsampling.

One noteworthy advantage of our proposal is that it provides both an accurate prediction as well as an explanation for it, using the posterior probability estimates for the different comfort classifications according to the features that have been used. This can be used to understand how a certain set of variables affect the resulting thermal perception, and in particular the probability of comfort against that of discomfort, something that can be of great help when addressing the problem of comfort optimization.

5.2. Future work & limitations

It is important to notice that, even if our models show slightly better results than the UTCI index for Outdoor Thermal Comfort classification, this comparison is extremely demanding for the index, as we have reduced both models for binary classification, for the task of predicting thermal comfort or discomfort. As a result, it would be useful to further compare our proposal with the UTCI index in the overall classification, *i.e.* in which comfort has the 11 categories included in the UTCI index. Another possibility would be to use the Mean Radiant Temperature (MRT) instead of the globe temperature, which was chosen as a primitive and straightforward value, instead of the MRT, which is often obtained through an estimate from the other climatic variables, following the UNE-EN ISO 7726:2002 standard [90].

These options could limit our possibilities, as some models cannot be used for multinomial classification, such as in the case of logistic regression, but it would provide a more accurate representation of the OTC. While the use of Thermal Sensation Vote (TSV) or Thermal Comfort Vote (TCV) targets can be informative, they add greater categorical complexity and demand more data, which is not always compatible with models like the Naïve Bayes or the Logistic Regression methods used in this work. Another approach would be to investigate employing black-box models, including Random Forest and XGBoost, together with explainable AI methods such as SHAP, to further improve predictive performance while preserving interpretability. Notice, however, that it has an increased complexity, and even with the binary cases ML models like the ones used in this work have had trouble when predicting comfort perception.

Future work should also consider the characteristics of the data used for model development. The RUROS dataset employed in this study, while a valuable and widely cited reference, was collected in European urban settings during 2001–2002. Its historical and regional nature may limit the transferability of results to contemporary or non-European contexts. Over the past two decades, thermal expectations, clothing patterns, and urban morphologies have evolved, potentially influencing thermal comfort perceptions. Therefore, future research should validate the proposed models using more recent datasets from varied climates and cultural settings, expanding beyond the European chosen dataset, to strengthen model generalizability.

To summarize, in this article we have proposed data-driven models whose results are comparable to those of the UTCI index. We trained and tested the different options over the RUROS dataset, using the same variables and providing explanations about the nature of the predictions. With this, comparison with further models and indices will be more accessible to researchers focusing on the classification of Outdoor Thermal Comfort.

CRediT authorship contribution statement

José-Antonio Rodríguez-Gallego: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation. **Eduardo Diz-Mellado:** Writing – review & editing, Methodology, Investigation, Formal analysis. **Marialena Nikolopoulou:** Writing – review & editing, Resources, Formal analysis, Data curation. **Tomás Chacón-Rebollo:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Carlos Rivera-Gómez:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization. **Carmen Galán-Marín:** 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: Carmen Galan-Marin reports equipment, drugs, or supplies was provided by Spain Ministry of Science and Innovation. Carmen Galan-Marin reports financial support and article publishing charges were provided by Spain Ministry of Science and Innovation. Eduardo Diz-Mellado reports financial support was provided by Spain Ministry of Science, Innovation and Universities.

Acknowledgements

This work has been supported by the project PID2021-124539OB-I00 funded by MCIN/AEI/ 10.13039/501100011033 and by “ERDF A way of making Europe”, project TED2021-129347B-C21 funded by MCIN/AEI/ 10.13039/501100011033, postdoctoral grant to Diz-Mellado JDC2023-050478-I funded by MCIN/AEI/ 10.13039/501100011033 and by the “European Union NextGenerationEU/PRTR”.

Appendix

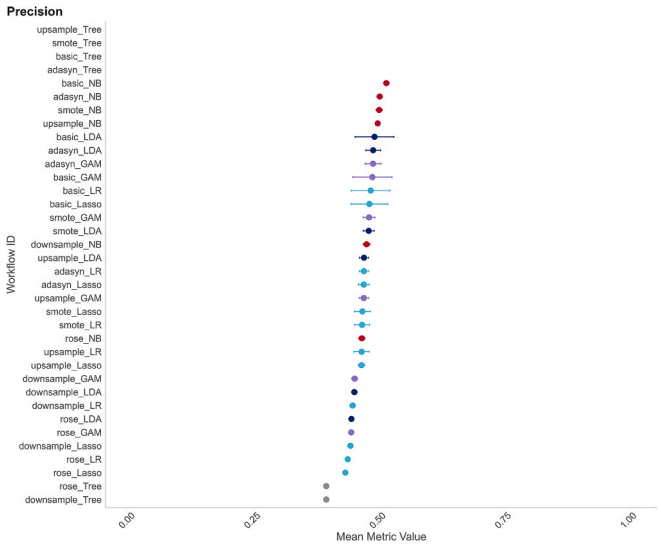

Fig. 10. Precision estimates for the different workflows, ordered by performance on the X-axis, leaving the data from Fribourg and Kassel for testing.

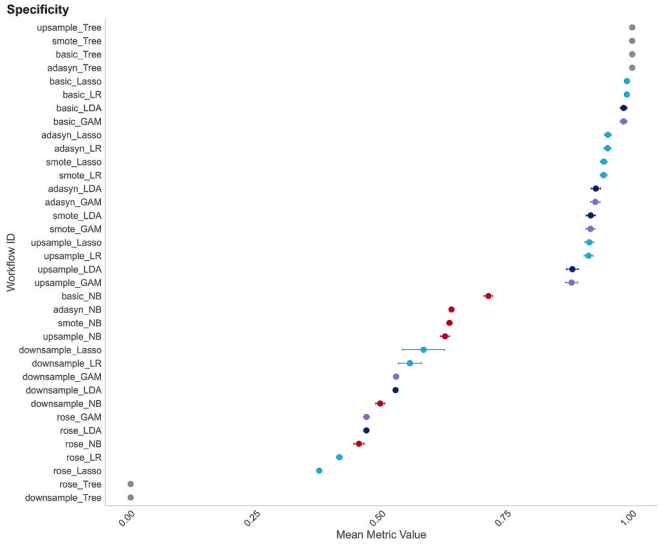

Fig. 11. Specificity estimates for the different workflows, ordered by performance on the X-axis, leaving the data from Fribourg and Kassel for testing.

Data availability

Data will be made available on request.

References

[1] D. Ormandy, V. Ezratty, Health and thermal comfort: from WHO guidance to housing strategies, Energy Policy 49 (2012) 116–121, https://doi.org/10.1016/j.enpol.2011.09.003.
[2] H. Wang, L. Liu, Experimental investigation about effect of emotion state on people's thermal comfort, Energ. Buildings 211 (2020) 109789, https://doi.org/10.1016/j.enbuild.2020.109789.
[3] S. Haddad, W. Zhang, R. Paolini, K. Gao, M. Altheeb, A. Al Mogirah, et al., Quantifying the energy impact of heat mitigation technologies at the urban scale, Nat. Cities 1 (2024) 62–72, https://doi.org/10.1038/s44284-023-00005-5.
[4] S. Kawakubo, M. Sugiuchi, S. Arata, Office thermal environment that maximizes workers' thermal comfort and productivity, Build. Environ. 110092 (2023), https://doi.org/10.1016/j.buildenv.2023.110092.
[5] D. Lai, Z. Lian, W. Liu, C. Guo, W. Liu, K.L. Chen, A comprehensive review of thermal comfort studies in urban open spaces, Sci. Total Environ. 742 (2020) 140092, https://doi.org/10.1016/j.scitotenv.2020.140092.
[6] S.W. Kim, R.D. Brown, Urban heat island (UHI) intensity and magnitude estimations: a systematic literature review, Sci. Total Environ. 779 (2021) 146389, https://doi.org/10.1016/j.scitotenv.2021.146389.
[7] S.E. Cleland, W. Steinhardt, L.M. Neas, J.J. West, A.G. Rappold, Urban heat island impacts on heat-related cardiovascular morbidity: a time series analysis of older adults in US metropolitan areas, Environ. Int. 178 (2023) 108005, https://doi.org/10.1016/j.envint.2023.108005.
[8] Y. Wu, B. Wen, A. Gasparrini, B. Armstrong, F. Sera, E. Lavigne, et al., Temperature frequency and mortality: Assessing adaptation to local temperature, Environ. Int. 187 (2024) 108691, https://doi.org/10.1016/j.envint.2024.108691.
[9] F. Binarti, M.D. Koerniawan, S. Triyadi, S.S. Utami, A. Matzarakis, A review of outdoor thermal comfort indices and neutral ranges for hot-humid regions, Urban Clim. 31 (2020) 100531, https://doi.org/10.1016/j.uclim.2019.100531.
[10] G.R. McGregor, universal thermal comfort index (UTCI), Int. J. Biometeorol. 56 (2012) 419, https://doi.org/10.1007/s00484-012-0546-6.
[11] P. Höppe, The physiological equivalent temperature–a universal index for the biometeorological assessment of the thermal environment, Int. J. Biometeorol. 43 (1999) 71–75, https://doi.org/10.1007/s004840050118.

[12] Z.Q. Fard, Z. Sadat Zomorodian, S. Sadat Korsavi, Application of machine learning in thermal comfort studies: a review of methods, performance and challenges, Energ. Buildings 256 (111771) (2022), https://doi.org/10.1016/j.enbuild.2021.111771.

[13] T. Mamani, R.F. Herrera, F. Muñoz-La Rivera, E. Atencio, Variables that affect thermal comfort and its measuring instruments: a systematic review, Sustainability 14 (3) (2022) 1773, https://doi.org/10.3390/su14031773.

[14] M. Nikolopoulou, Outdoor thermal comfort, Front. Biosci. 3 (2011) 1552–1568, https://doi.org/10.2741/245.

[15] K. Parsons, Human thermal environments: the effects of hot, moderate, and cold environments on human health, comfort and performance, CRC Press, 2007.

[16] Ashrae, ANSI/ASHRAE Standard 55-2020: thermal environmental conditions for human occupancy, American Society of Heating, Refrigerating and Air-Conditioning Engineers, 2021.

[17] M. Nikolopoulou, S. Lykoudis, Thermal comfort in outdoor urban spaces: analysis across different European countries, Build. Environ. 41 (2006) 1455–1470, https://doi.org/10.1016/j.buildenv.2005.05.031.

[18] P.O. Fanger, Assessment of man's thermal comfort in practice, Occup. Environ. Med. 30 (1973) 313–324, https://doi.org/10.1136/oem.30.4.313.

[19] International Organization for Standardization (ISO). (1994). 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 (ISO 7730:1994). ISO.

[20] Y.-C. Chen, A. Matzarakis, Modified physiologically equivalent temperature—basics and applications for western European climate, Theor. Appl. Climatol. 132 (2018) 1275–1289, https://doi.org/10.1007/s00704-017-2158-x.

[21] A.P. Gagge, J.A. Stolwijk, Y. Nishi, An effective temperature scale based on a simple model of human physiological regulatiry response, Memoirs of the Faculty of Engineering, Hokkaido University 13 (Suppl) (1972) 21–36.

[22] W. Ji, Y. Zhu, H. Du, B. Cao, Z. Lian, Y. Geng, C. Yang. Interpretation of standard effective temperature (SET) and explorations on its modification and development. Building and Environment, 210 (2022), 108714, https://ui.adsabs.harvard.edu/link_gateway/2022BuEnv.21008714J/doi:10.1016/j.buildenv.2021.108714.

[23] P. Bröde, D. Fiala, K. Blazejczyk, I. Holmér, G. Jendritzky, B. Kampmann, G. Havenith, Deriving the operational procedure for the universal thermal climate Index (UTCI), Int. J. Biometeorol. 56 (2012) 481–494, https://doi.org/10.1007/s00484-011-0454-1.

[24] F. Schaudienst, F.U. Vogdt, Fanger's model of thermal comfort: a model suitable just for men? Energy Procedia 132 (2017) 129–134, https://doi.org/10.1016/j.egypro.2017.09.658.

[25] A. Tseliou, I.X. Tsiros, S. Lykoudis, M. Nikolopoulou, An evaluation of three biometeorological indices for human thermal comfort in urban outdoor areas under real climatic conditions, Build. Environ. 45 (2010) 1346–1352, https://doi.org/10.1016/j.buildenv.2009.11.009.

[26] A. Matzarakis, F. Rutz, H. Mayer, Modelling radiation fluxes in simple and complex environments: basics of the RayMan model, Int. J. Biometeorol. 54 (2) (2010) 131–139.

[27] C. Hu, H. Zeng, Decoding spatial patterns of urban thermal comfort: explainable machine learning reveals drivers of thermal perception, Environ. Impact Assess. Rev. 114 (2025) 107895, https://doi.org/10.1016/j.eiar.2025.107895.

[28] K. Pantavou, G. Theoharatos, M. Santamouris, D. Asimakopoulos, Outdoor thermal sensation of pedestrians in a Mediterranean climate and a comparison with UTCI, Build. Environ. 66 (2013) 82–95, https://doi.org/10.1016/j.buildenv.2013.02.014.

[29] E. Diz-Mellado, V.P. López-Cabeza, C. Rivera-Gómez, C. Galán-Marín, J. Rojas-Fernández, M. Nikolopoulou, Extending the adaptive thermal comfort models for courtyards, Build. Environ. 203 (2021) 108094, https://doi.org/10.1016/j.buildenv.2021.108094.

[30] E. Diz-Mellado, M. Nikolopoulou, V.P. López-Cabeza, C. Rivera-Gómez, C. Galán-Marín, Cross-evaluation of thermal comfort in semi-outdoor spaces according to geometry in Southern Spain, Urban Clim. 49 (2023) 101491, https://doi.org/10.1016/j.uclim.2023.101491.

[31] P.R. Höppe, Heat balance modelling, Experientia 49 (1993) 741–746, https://doi.org/10.1007/bf01923542.

[32] R.J. Cureau, I. Pigliautile, I. Kousis, X. Huang, E. Bou-Zeid, A.L. Pisello, On the performance of human thermal stress models in the outdoors against observations, Energ. Buildings 115837 (2025), https://doi.org/10.1016/j.enbuild.2025.115837.

[33] Y.-C. Chen, W.-N. Chen, C.-C.-K. Chou, A. Matzarakis, Concepts and new implements for modified physiologically equivalent temperature, Atmos. 11 (2020) 694, https://doi.org/10.3390/atmos11070694.

[34] D. Fiala, G. Havenith, P. Bröde, B. Kampmann, G. Jendritzky, UTCI-Fiala multi-node model of human heat transfer and temperature regulation, Int. J. Biometeorol. 56 (2012) 429–441, https://doi.org/10.1007/s00484-011-0424-7.

[35] D. Fiala, et al., Dynamic simulation of human heat transfer and thermal comfort, De Montfort University Leicester, UK, 1998.

[36] H.H. Pennes, Analysis of tissue and arterial blood temperatures in the resting human forearm, J. Appl. Physiol. 1 (2) (1948) 93–122, https://doi.org/10.1152/jappl.1948.1.2.93.

[37] J. Li, J. Niu, C.M. Mak, T. Huang, Y. Xie, Exploration of applicability of UTCI and thermally comfortable sun and wind conditions outdoors in a subtropical city of Hong Kong, Sustain. Cities Soc. 52 (2020) 101793, https://doi.org/10.1016/j.scs.2019.101793.

[38] S. Chindapol, J. Blair, P. Osmond, D. Prasad, A suitable thermal stress index for the elderly in summer tropical climates, Procedia Eng. 180 (2017) 932–943, https://doi.org/10.1016/j.proeng.2017.04.253.

[39] F. Lindberg, B. Holmer, S. Thorsson, SOLWEIG 1.0 – Modelling spatial variations of 3D radiant fluxes and mean radiant temperature in complex urban settings, Int. J. Biometeorol. 52 (7) (2008) 697–713, https://doi.org/10.1007/s00484-008-0162-7.

[40] M. Bruse, H. Fleer, Simulating surface–plant–air interactions inside urban environments with a three dimensional numerical model, Environ. Model. Software 13 (3–4) (1998) 373–384, https://doi.org/10.1016/S1364-8152(98)00042-5.

[41] B. Wang, Y.K. Yi, Developing an adapted UTCI (universal thermal climate index) for the elderly population in China's severe cold climate region, Sustain. Cities Soc. 69 (2021) 102813, https://doi.org/10.1016/j.scs.2021.102813.

[42] K. Pantavou, M. Santamouris, D. Asimakopoulos, G. Theoharatos, Empirical calibration of thermal indices in an urban outdoor Mediterranean environment, Build. Environ. 80 (2014) 283–292, https://doi.org/10.1016/j.buildenv.2014.06.001.

[43] S. Haykin, Neural networks: a comprehensive foundation, Prentice Hall PTR, 1994.

[44] X. Tian, Z. Lin, Predicting personalized thermal comfort in stratified micro-environments using turbulent jet theories and data-driven models, Build. Environ. 230 (2023) 110009, https://doi.org/10.1016/j.buildenv.2023.110009.

[45] T.K. Ho, Random decision forests. Proceedings of 3rd international conference on document analysis and recognition, 1 (1995) 278–282, https://doi.org/10.1109/ICDAR.1995.598994.

[46] T. Chen, C. Guestrin, Xgboost: a scalable tree boosting system, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785.

[47] D. Wei, L. Yang, Z. Bao, Y. Lu, H. Yang, Variations in outdoor thermal comfort in an urban park in the hot-summer and cold-winter region of China, Sustain. Cities Soc. 77 (2022) 103535, https://doi.org/10.1016/j.scs.2021.103535.

[48] A.B. Avci, Machine learning-based prediction of thermal comfort: exploring building types, climate, ventilation strategies, and seasonal variations, Build. Res. Inform. 1–18 (2025), https://doi.org/10.1080/09613218.2025.2462932.

[49] J.A. Rodríguez-Gallego, E. Diz Mellado, T. Chacón Rebollo, C. Galán-Marín, C. Rivera-Gómez, Pedestrians' urban thermal comfort: a machine learning assessment through transect walks, (2024), http://dx.doi.oirg/10.46354/i3m.2024.sesde.008.

[50] W. Cai, L. Huang, Z. Zou, Actively-exploring thermography-enabled autonomous robotic system for detecting and registering HVAC thermal leaks, Autom. Constr. 152 (2023) 104901, https://doi.org/10.1016/j.autcon.2023.104901.

[51] G. Park, M. Lee, H. Jang, C. Kim, Thermal anomaly detection in walls via CNN-based segmentation, Autom. Constr. 125 (2021) 103627, https://doi.org/10.1016/j.autcon.2021.103627.

[52] E. Vollmer, J. Ruck, R. Volk, F. Schultmann, Detecting district heating leaks in thermal imagery: comparison of anomaly detection methods, Autom. Constr. 168 (2024) 105709, https://doi.org/10.1016/j.autcon.2024.105709.

[53] H. Huang, Y. Cai, C. Zhang, Y. Lu, A. Hammad, L. Fan, Crack detection of masonry structure based on thermal and visible image fusion and semantic segmentation, Autom. Constr. 158 (2024) 105213, https://doi.org/10.1016/j.autcon.2023.105213.

[54] M.S. Lystbæk, Machine learning-driven processes in architectural building design, Autom. Constr. 178 (2025) 10637, https://doi.org/10.1016/j.autcon.2025.106379.

[55] H. Zahid, O. Elmansoury, R. Yaagoubi, Dynamic predicted mean vote: an IoT-BIM integrated approach for indoor thermal comfort optimization, Autom. Constr. 129 (2021) 103805, https://doi.org/10.1016/j.autcon.2021.103805.

[56] V.G. Zakka, M. Lee, R. Zhang, L. Huang, S. Jung, T. Hong, Non-invasive vision-based personal comfort model using thermographic images and deep learning, Autom. Constr. 168 (2024) 105811, https://doi.org/10.1016/j.autcon.2024.105811.

[57] N. Burkart, M.F. Huber, A survey on the explainability of supervised machine learning, J. Artif. Intell. Res. 70 (2021) 245–317, https://doi.org/10.48550/arXiv.2011.07876.

[58] S.Y. Chan, C.K. Chau, Development of artificial neural network models for predicting thermal comfort evaluation in urban parks in summer and winter, Build. Environ. 164 (2019) 106364, https://doi.org/10.1016/j.buildenv.2019.106364.

[59] S.S. Shahrestani, Z.S. Zomorodian, M. Karami, F. Mostafavi, A novel machine learning-based framework for mapping outdoor thermal comfort, Adv. Build. Energy Res. 17 (2023) 53–72, https://doi.org/10.1080/17512549.2022.2152865.

[60] J. Zhang, F. Zhang, Z. Gou, J. Liu, Assessment of macroclimate and microclimate effects on outdoor thermal comfort via artificial neural network models, Urban Clim. 42 (2022) 101134, https://doi.org/10.1016/j.uclim.2022.101134.

[61] G. Zhong, Convolutional neural network model to predict outdoor comfort UTCI microclimate map, Atmos. 13 (2022) 1860, https://doi.org/10.3390/atmos13111860.

[62] L.S. Shapley, et al., A value for n-person games, Princeton University Press Princeton, 1953.

[63] R. Guo, B. Yang, Y. Guo, H. Li, Z. Li, B. Zhou, F. Wang, Machine learning-based prediction of outdoor thermal comfort: Combining Bayesian optimization and the SHAP model, Build. Environ. 254 (2024) 111301, https://doi.org/10.1016/j.buildenv.2024.111301.

[64] M. Nikolopoulou, S. Lykoudis, RUROS project outdoor comfort database (Vol. 41). RUROS project outdoor comfort database, 2004 (Vol. 41). Zenodo. doi:10.5281/zenodo.14275070.

[65] E.R. CORDIS, Rediscovering the urban realm and open spaces (RUROS). Retrieved 07 2025, from https://cordis.europa.eu/project/id/EVK4-CT-2000-00032.

[66] M. Nikolopoulou, Thermal comfort models for open urban spaces. In Designing Open Spaces in the Urban Environment: a Bioclimatic Approach. Centre for

Renewable Energy Sources, EESD, FP5, (2004). Retrieved from http://www.cres.qr/kape/education/design_guidelines_en.pdf.
[67] L. Chen, N. Kántor, M. Nikolopoulou, Meta-analysis of outdoor thermal comfort surveys in different European cities using the RUROS database: the role of background climate and gender, Energ. Buildings 256 (2022) 111757, https://doi.org/10.1016/j.enbuild.2021.111757.
[68] A. Kotopouleas, M. Nikolopoulou, S. Lykoudis, From indoors to outdoors and in-transition; thermal comfort across different operation contexts. Proceedings 8th Windsor Conference: Rethinking Comfort, 2018.
[69] R.G. Sales, A.A. Rodríguez Sousa, E. Yáñez, L. Blanco Cano, D. Raffin, L. Jatar, et al., Degree of importance of demographic and socio-cultural factors in environmental perception: bases for the design of public policies in Argentina and Spain, Environ. Dev. Sustain. 26 (4) (2024) 9005–9024, https://doi.org/10.1007/s10668-023-03079-2.
[70] P. Sprent, Fisher exact test, in: M. Lovric (Ed.), International Encyclopedia of Statistical Science, Springer, Berlin, Heidelberg, 2011, https://doi.org/10.1007/978-3-642-04898-2_253.
[71] Q. McNemar, Note on the sampling error of the difference between correlated proportions or percentages, Psychometrika 12 (2) (1947) 153–157, https://doi.org/10.1007/bf02295996.
[72] M. Kuhn, H. Wickham, Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles, (2020). Retrieved from https://www.tidymodels.org.
[73] H. Wickham, M. Averick, J. Bryan, W. Chang, L.D. McGowan, R. François, H. Yutani, Welcome to the tidyverse, J. Open Sour. Softw. 4 (2019) 1686, https://doi.org/10.21105/joss.01686.
[74] M. Han, R. May, X. Zhang, X. Wang, S. Pan, Y. Da, Y. Jin, A novel reinforcement learning method for improving occupant comfort via window opening and closing, Sustain. Cities Soc. 61 (2020) 102247, https://doi.org/10.1016/j.scs.2020.102247.
[75] J. Kim, S. Schiavon, G. Brager, Personal comfort models–a new paradigm in thermal comfort for occupant-centric environmental control, Build. Environ. 132 (2018) 114–124, https://doi.org/10.1016/j.buildenv.2018.01.023.
[76] O. Taylor, P. Ezekiel, V. Emma, Smart system for thermal comfort prediction on residential buildings using data-driven model with random forest classifier, Eur. J. Electr. Eng. Comput. Sci. 5 (4) (2021) 40–45, https://doi.org/10.24018/ejece.2021.5.4.346.
[77] E.S. Olivas, J.D. Guerrero, M. Martinez-Sober, J.R. Magdalena-Benedito, L. Serrano, et al., Handbook of research on machine learning applications and trends: algorithms, methods, and techniques: algorithms, methods, and techniques, IGI Global (2009), https://doi.org/10.4018/978-1-60566-766-9.
[78] P. Radivojac, M. White, Machine learning handbook. Machine Learning Handbook, 2019.
[79] N.L. Torelli, G. Menardi, N. Torelli, et al., ROSE: a package for binary imbalanced learning, R J. 6 (1) (2014) 82–92.
[80] H. He, Y. Bai, E.A. Garcia, S. Li, ADASYN: adaptive synthetic sampling approach for imbalanced learning, in: Proceedings of the IEEE International Joint Conference on Neural Networks, 2008, https://doi.org/10.1109/IJCNN.2008.4633969.
[81] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357, https://doi.org/10.1613/jair.953.
[82] E. Hvitfeldt, themis: Extra Recipes Steps for Dealing with Unbalanced Data. Retrieved from R package version 1.0.3, (2025), https://themis.tidymodels.org, https://github.com/tidymodels/themis.
[83] A. Burkov, The hundred-page machine learning book, Andriy Burkov, Quebec City, QC, Canada, 2019, p. p. 32.
[84] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics Bull. 1 (6) (1945) 80–83.
[85] B. Efron, Bootstrap methods: another look at the jackknife, in: Breakthroughs in Statistics: Methodology and Distribution, New York, NY, Springer, New York, 1992, pp. 569–593.
[86] A. Matzarakis, H. Mayer, M.G. Iziomon, Applications of a universal thermal index: physiological equivalent temperature (PET), Int. J. Biometeorol. 43 (2) (1999) 76–84, https://doi.org/10.1007/s004840050119.
[87] A. Matzarakis, F. Rutz, H. Mayer, Modelling radiation fluxes in simple and complex environments—application of the RayMan model, Int. J. Biometeorol. 51 (4) (2007) 323–334, https://doi.org/10.1007/s00484-006-0061-8.
[88] K. Blazejczyk, Y. Epstein, G. Jendritzky, H. Staiger, B. Tinz, Comparison of UTCI to selected thermal indices, Int. J. Biometeorol. 56 (3) (2012) 515–535, https://doi.org/10.1007/s00484-011-0453-2.
[89] Ladybug Tools LLC, Ladybug Comfort (Version 0.41.6), (2023). https://github.com/ladybug-tools/ladybug-comfort.
[90] International Organization for Standardization (ISO). Ergonomics of the thermal environment - Instruments for measuring physical quantities, (2002). (UNE-EN ISO 7726:2002). UNE_EN_ISO_7726_2002.