Article

Application of Statistical Learning Algorithms in Thermal Stress Assessment in Comparison with the Expert Judgment Inherent to the Universal Thermal Climate Index (UTCI)

Peter Bröde [1,*], Dusan Fiala [2] and Bernhard Kampmann [3]

[1] Leibniz Research Centre for Working Environment and Human Factors at TU Dortmund (IfADo), Ardeystrasse 67, 44139 Dortmund, Germany

[2] ErgonSim–Human Thermal Modelling, Robert-Bosch-Str. 20, 72469 Messstetten, Germany; dfiala@ergonsim.de

[3] Department of Occupational Health Science, School of Mechanical Engineering and Safety Engineering, University of Wuppertal, 42119 Wuppertal, Germany; kampmann@uni-wuppertal.de

* Correspondence: broede@ifado.de; Tel.: +49-231-1084-225

Abstract: This study concerns the application of statistical learning (SL) in thermal stress assessment compared to the results accomplished by an international expert group when developing the Universal Thermal Climate Index (UTCI). The performance of diverse SL algorithms in predicting UTCI equivalent temperatures and in thermal stress assessment was assessed by root mean squared errors (RMSE) and Cohen's kappa. A total of 48 predictors formed by 12 variables at four consecutive 30 min intervals were obtained as the output of an advanced human thermoregulation model, calculated for 105,642 conditions from extreme cold to extreme heat. Random forests and k-nearest neighbors closely predicted UTCI equivalent temperatures with an RMSE about 3 °C. However, clustering applied after dimension reduction (principal component analysis and t-distributed stochastic neighbor embedding) was inadequate for thermal stress assessment, showing low to fair agreement with the UTCI stress categories (Cohen's kappa < 0.4). The findings of this study will inform the purposeful application of SL in thermal stress assessment, where they will support the biometeorological expert.

Keywords: bio-meteorological index; heat stress; cold stress; high-dimensional data; artificial intelligence; machine learning; UTCI

1. Introduction

Statistical or machine learning (SL) is central to artificial intelligence (AI) applications [1–3] with potential relevance to environmental risk assessment, especially in settings with high-dimensional input such as thermal stress indices [4]. There, they may assist or even attempt to replace the bio-meteorological expert judgment, as indicated by the increasing number of recent applications in diverse fields of biometeorology concerning, e.g., indoor and outdoor thermal comfort, the impact assessment of climate change-related heat stress, urban planning, the adaptation of buildings and human behavior to changing climatic conditions, or establishing a link between human physiology and thermal sensation [5–11]. The rising number of SL applications triggers a demand for quantitatively assessing the skills of statistical learning in comparison to results involving expert judgment. Our study aimed to contribute to such an assessment by utilizing as a testbed the development process of the Universal Thermal Climate Index (UTCI), a complex assessment procedure for the physiological strain related to the outdoor thermal environment [12], which soon after its release had been widely adopted and applied in core fields of biometeorology covering weather services, public health research and epidemiology, precautionary planning of heat health warning systems and urban environments, including applications to tourism and health resorts, and climate impact research [13–24].

The development of UTCI was accomplished by an international inter-disciplinary endeavor [25,26] involving the judgment of more than 40 experts from 23 countries [12]. Figure 1 visualizes the concept and major stages of the UTCI development process [25].
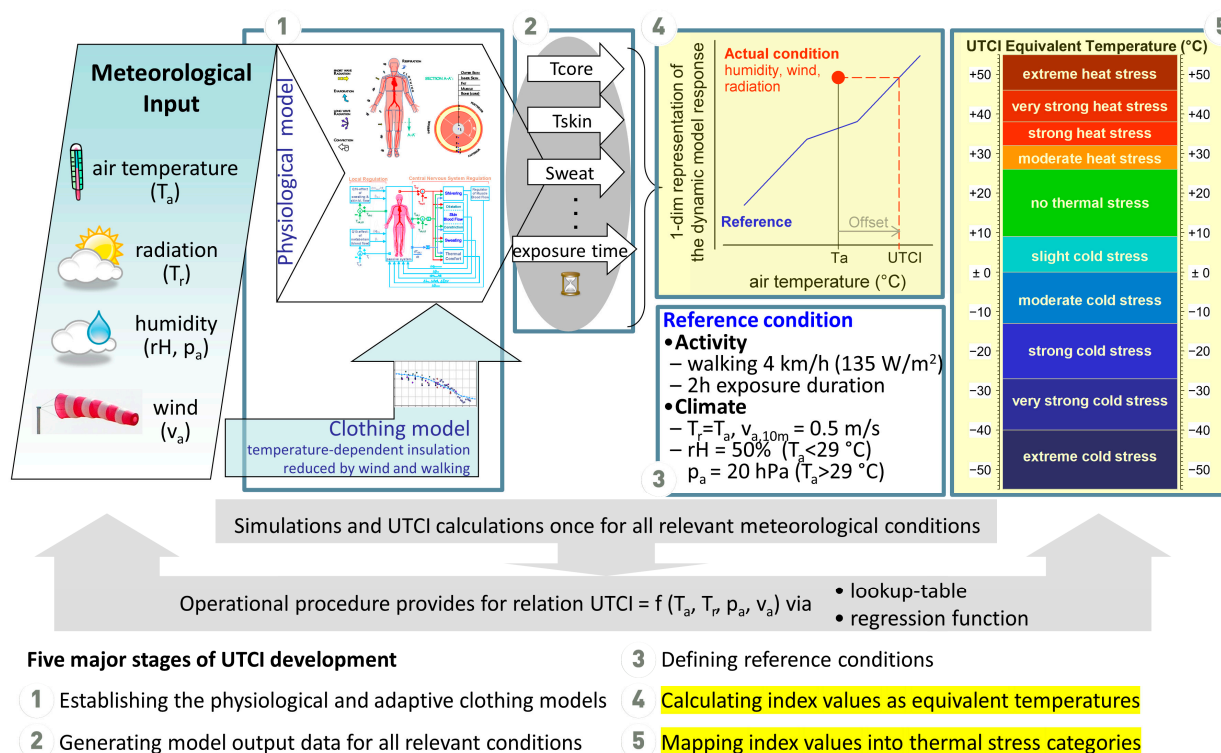
Figure 1. Stages of the development process with the elements of the operational procedure for calculating UTCI equivalent temperatures with thermal stress categories, modified from [25], where the highlighted stages 4 and 5 are of major concern in this study.

UTCI was conceptualized as an equivalent temperature (in °C) defined as the air temperature of the reference condition with the same dynamic physiological response as the actual condition. The physiological response to thermal stress was derived at stage 1 from the output of an advanced human model of thermoregulation [27], which had been coupled with an adaptive clothing model with clothing insulation changing depending on air temperature [28]. Extensive simulation runs were performed at stage 2, which also included the validation of predicted physiological responses against experimental laboratory and field data [29–32]. After the experts reached consensus about the definition of reference conditions in stage 3, stage 4 derived UTCI on an equivalent temperature scale. This involved a multivariate approach [25,33] including a dimension reduction step of the multidimensional model output to a one-dimensional strain indicator by principal component analysis. Then, UTCI values for non-reference conditions concerning wind speed, humidity, and solar and thermal radiation were identified by searching the reference condition with the same strain indicator value (Figure 1). At stage 5, for assessment purposes, a scale classifying the UTCI values into ten categories of thermal stress was added to the operational procedure [25]. The assessment scale was established by consensus and expert judgment involving the comparison of the model output values concerning the thermal state of the human body, including thermal sensation and effectors of thermoregulation, to established ergonomic limit criteria regarding cold and heat stress [25,34].

Thus, UTCI integrates human expert judgment with the high-dimensional model output from numerous simulations performed for a huge grid of relevant conditions representing the reference and non-reference climates ranging from the extreme cold to extreme heat. Therefore, UTCI and its underlying data represent a suitable test case for benchmarking the skills of SL algorithms against procedures involving expert knowledge concerning

the comprehensive assessment of human responses to thermal stress in biometeorological applications.

Focusing on the data-driven stages 4 and 5 of the UTCI development (Figure 1), the aim of this study was to compare the performance of SL algorithms to the outcome of the process involving the knowledge and judgment of the UTCI expert group [12]. Thus, it is expected that the findings from this study will inform the purposeful usage of SL algorithms in supporting the biometeorological expert for thermal risk assessment.

2. Materials and Methods

Our approach was to apply selected SL algorithms deemed representative for recent applications [5–11,35–41] to the multi-dimensional data simulated over a comprehensive grid of relevant climatic conditions by the UTCI-Fiala model [27] at stage 2 of the UTCI development process (Figure 1).

2.1. UTCI Data

The operational procedure [25] was based on the dynamic physiological response characterized by the 48-dimensional model output formed by 12 variables at 4 consecutive 30 min intervals (Table 1). This output was generated for 1051 reference conditions, which were characterized by low wind speed (0.5 m/s measured 10 m above ground level), relative humidity of 50%, but water vapor pressure capped at 20 hPa for air temperatures above 29 °C, and shadowed conditions with mean radiant temperature equaling air temperature, with the latter covering the temperature range from −110 °C to +100 °C with a 0.2 K increment. Notably, for reference conditions, UTCI values are known and equal to the air temperature by definition (Figure 1). Another set of 104,591 non-reference conditions with varying levels of air temperature (−50 °C to +50 °C with 1 K increment), wind speed (0.5 to 30.3 m/s 10 m above ground), relative humidity (5% to 100%, but with maximum water vapor pressure of 50 hPa), and solar and thermal radiation (mean radiant temperature from 30 K below to 70 K above air temperature with 5 K increment) had to be valued in UTCI equivalent temperatures (°C) and classified in 10 stress categories ranging from extreme cold to extreme heat, as indicated by stages 4 and 5 in Figure 1.

For independent test purposes, these analyses were repeated with an external set of 1000 non-reference conditions, which had been simulated previously [25,42] and served as non-reference test data in this study.

Table 1. Physiological output variables obtained from the UTCI-Fiala model [27].
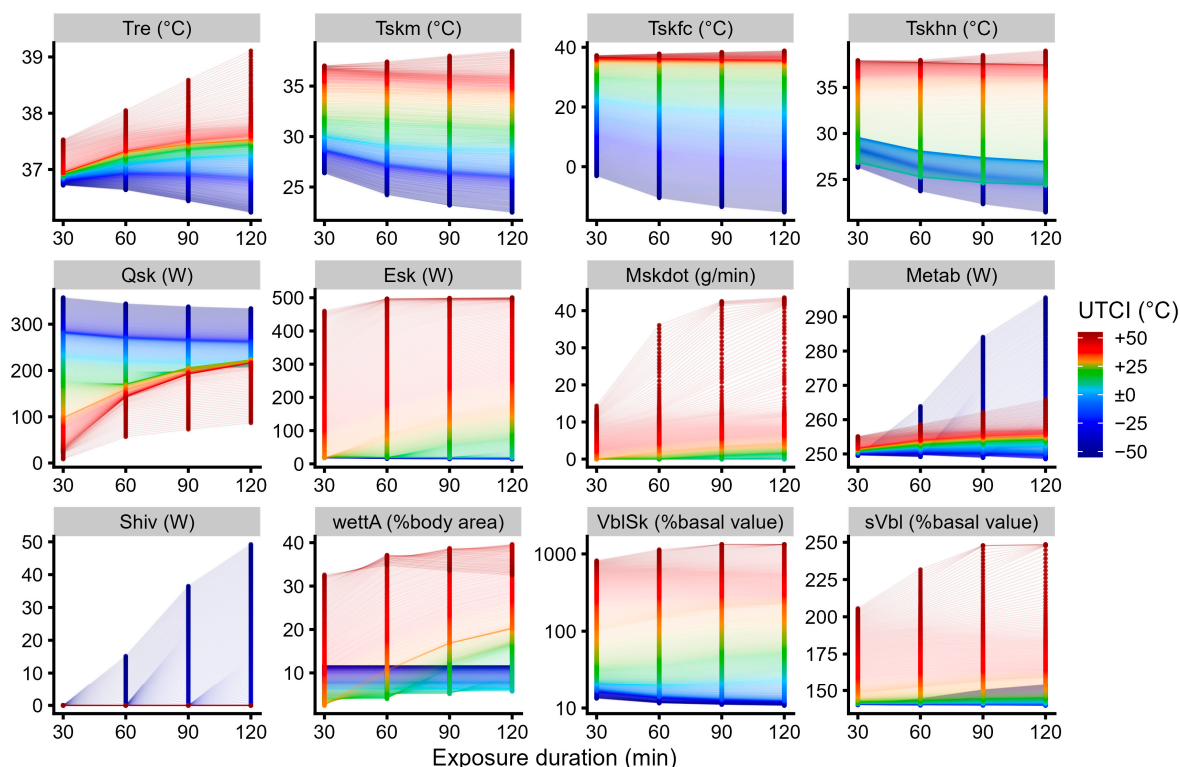
Physiological Variable [1]	Abbreviation [2]	Unit
rectal temperature	Tre	°C
mean skin temperature	Tskm	°C
facial skin temperature	Tskfc	°C
hand skin temperature	Tskhn	°C
total net heat loss	Qsk	W
evaporative (latent) heat loss	Esk	W
sweat rate	Mskdot	g/min
metabolic heat production	Metab	W
heat generated by shivering	Shiv	W
skin wetness	wettA	% of body area
skin blood flow	VblSk	% of basal value
cardiac output	sVbl	% of basal value

[1] In addition, scores of the dynamic thermal sensation (DTS) on the 7-unit ASHRAE [43] scale (−3: cold; −2: cool; −1: slightly cool; 0: neutral; +1: slightly warm; +2: warm; +3: hot) were estimated from the physiological output [44]; [2] Each output variable was calculated for 30, 60, 90, and 120 min of simulated exposure duration.

2.2. Data Analysis

To not compromise the study goal of comparing SL algorithms to expert judgment, data preprocessing using external input was limited. After inspection (Figures 2 and A1), we

log-transformed the percentage skin blood flow data (VblSK). Data cleaning was obsolete as the deterministic UTCI-Fiala model [27] did not generate any missing values.

Figure 2. Set of twelve physiological output variables (abbreviations in Table 1) at four points in time from the simulated reference conditions plotted representing the distribution of the training data with UTCI ranging from −55 °C to +55 °C. Note the log-scale applied to the skin blood flow data (VblSk) shown in the third column of the bottom row.

Splitting the reference conditions into sets of 840 training and 211 test data and using the external 1000 non-reference conditions as additional non-reference test data, we compared the results of the UTCI expert group to diverse SL algorithms in predicting UTCI equivalent temperature values from the 48 predictors, using the root-mean squared error (RMSE) as a performance metric with error defined as the deviation of UTCI from the equivalent temperatures predicted by the SL algorithms. The supervised learning techniques comprised linear regression (MLR: multiple linear regression; LASSO: least absolute shrinkage and selection operator), tree-based and ensemble methods (CART: classification and regression trees; RF: random forests; XGBoost: extreme gradient boosting), as well as support vector machines (SVN), and the non-parametric k-nearest neighbors (KNN) [1–3,45–47]. More specifically, because UTCI equivalent temperature equals air temperature for reference conditions, we could label UTCI equivalent temperatures for the reference training data and did utilize the above-mentioned regression algorithms for predicting equivalent temperature by the 48 predictors (12 variables at 4 points in time) listed in Table 1.

For comparison to the UTCI assessment scale, UTCI values were then categorized by hierarchical and k-means clustering [1] after applying principal component analysis (PCA) and t-distributed stochastic neighbor embedding t-SNE [48], respectively, to the high-dimensional UTCI model output for dimensionality reduction. In accordance with the approach of the UTCI expert group [25], we searched for ten categories using the UTCI reference data set limited to the 926 values with air temperatures between −110 °C and +75 °C, covering the range of UTCI values obtained for non-reference conditions [25]. This enabled ranking the resulting clusters by the intra-cluster mean air temperatures, which

were equal to UTCI for the reference conditions according to the equivalent temperature definition (Figure 1).

Metrics of agreement between the UTCI stress categories and the classification found by clustering were derived from the confusion matrix by calculating the overall accuracy, defined as the proportion of the UTCI stress categories that were classified correctly, and by Cohen's kappa, defined as $(p_o - p_e)/(1 - p_e)$, with p_o denoting the observed proportion of agreement and p_e the hypothetical proportion of agreement due to chance [49]. The calculations were performed with the statistical software R version 4.3.3 [50] using the packages caret [51], xgboost [47], yardstick [52], tidyverse [53], and cowplot [54].

3. Results

Figure 2 illustrates the distribution of the training data, i.e., the 48-dimensional set of twelve physiological output variables at four points in time predicted by the UTCI-Fiala model [27] for the UTCI reference conditions with air temperatures ranging from −55 °C to +55 °C. Similarly, Figure A1 depicts the corresponding data for the combined reference and non-reference conditions depending on air temperature.

While several parameters, like rectal, mean skin, and facial temperatures, as well as skin blood flow, showed a time-dependent pattern of values increasing from cold to heat stress, other parameters showed non-monotonous relationships, e.g., metabolic heat production increasing both in the cold due to shivering, and in the heat due to the so-called Q_{10}-effect, i.e., the rise in metabolic rate associated with increasing body core temperature. The UTCI-Fiala model had implemented an increase in metabolic rate of about 7% per 1 K rise in core temperature in accordance with the results of climate chamber experiments [27,55]. These non-linear and non-monotonous patterns of stress–strain relationships with respect to the air temperature-dependent physiological responses, as well as shivering only occurring in the cold, posed a particular challenge to the learning algorithms.

3.1. UTCI Equivalent Temperature Calculation

UTCI and the equivalent temperatures predicted by the diverse algorithms were highly correlated, with the squared correlation (R^2) exceeding 0.95 for any dataset and with any method, as shown in Appendix A by Figure A2.

However, performance differences became obvious in Figure 3A, showing the typical prediction error RMSE obtained from the separate algorithms for the three sets of training and test data, respectively. While multiple linear regression and LASSO worked well for UTCI reference conditions, RMSE increased to more than 7 °C for the non-reference data, approximately corresponding to one heat stress category deviation (Figure 1). RMSE values of similar magnitude for the non-reference test data occurred with CART and SVM, whereas gradient boosting lowered the prediction error. The best performance was observed with RF and KNN, both showing an excellent fit for the reference data with RMSE < 0.2 °C and a good agreement with RMSE of about 3 °C for non-reference UTCI, where RMSE for KNN was below the prediction error obtained for calculating UTCI from meteorological input by the polynomial regression function provided by the operational procedure (Figure 1) [25]. Among the tree-based algorithms, CART performance was inferior to RF and gradient boosting (XGBoost) in all cases, in accordance with earlier observations in different application areas [46]. The results for RMSE were confirmed by the corresponding analyses utilizing the squared correlation coefficient (R^2) and the mean absolute error (MAE) as alternative performance metrics, as shown in Figure A2.
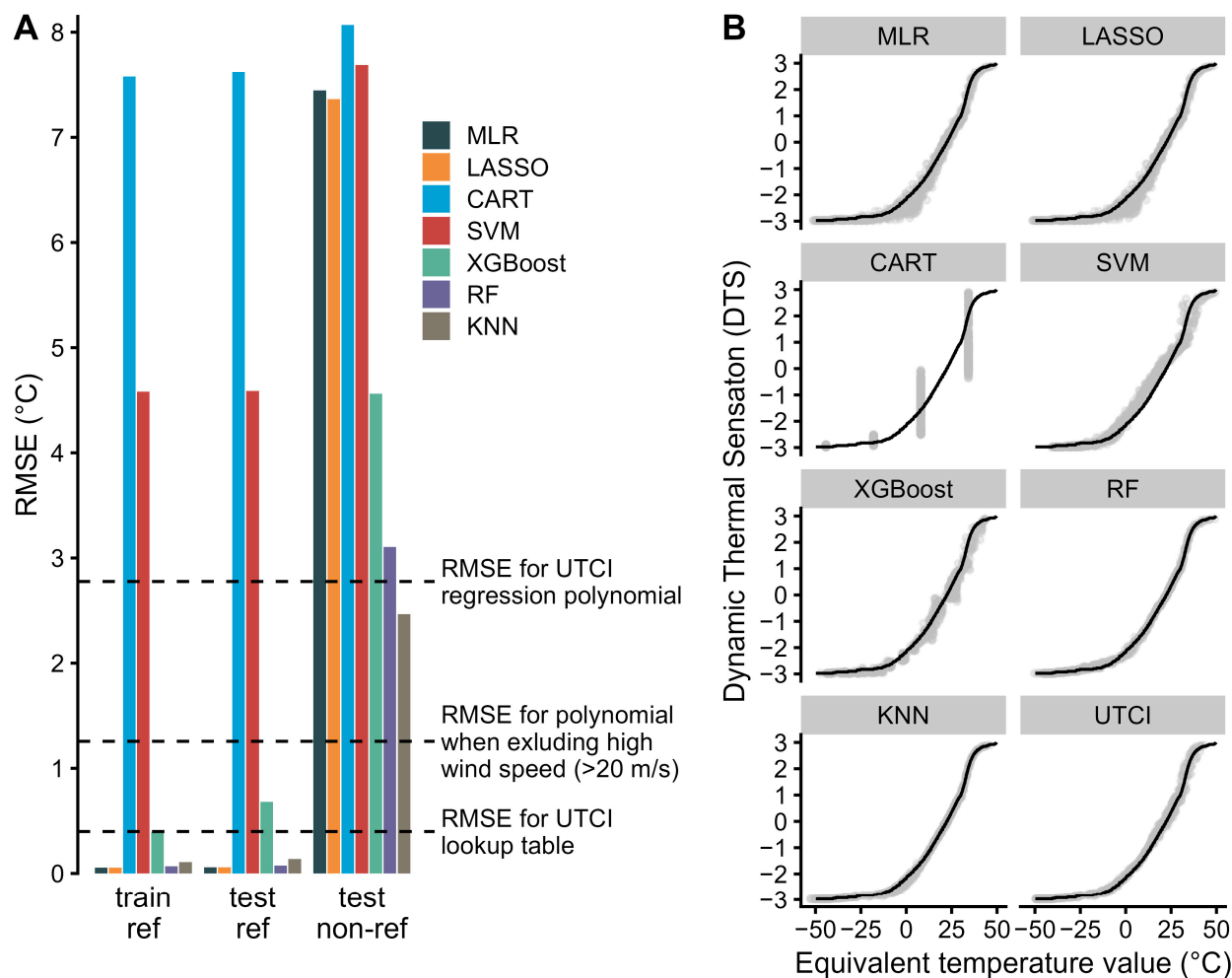
Figure 3. (**A**) Root-mean squared error (RMSE) for predicting UTCI equivalent temperatures from 48 physiological variables by different supervised SL algorithms (MLR: multiple linear regression; LASSO: least absolute shrinkage and selection operator; CART: classification and regression trees; SVM: support vector machines; XGBoost: extreme gradient boosting; RF: random forests; KNN: k-nearest neighbors) for the training reference (train ref) data and two test datasets for reference (ref) and non-reference (non-ref) UTCI conditions. For comparison, horizontal reference lines indicate the RMSE for calculating UTCI from meteorological input using the polynomial regression function or the lookup table provided by the operational procedure [25]. (**B**) Dynamic thermal sensation votes from the non-reference test data depending on the equivalent temperature predicted by the SL algorithms from (**A**) in comparison to the original UTCI (**lower right panel**), with lines indicating the values for the reference conditions.

As an essential requirement for thermal stress indices, identical index values should indicate equivalent thermal strain [12]. Thus, Figure 3B evaluates this criterion for the equivalent temperatures determined by the diverse SL algorithms in comparison to the original UTCI concerning the dynamic thermal sensation (DTS) from the non-reference test data averaged over the four simulated points in time (30, 60, 90, and 120 min). The resulting pattern for the best-performing algorithms, RF and KNN, was similar to the original UTCI, showing a small variation around the reference values, whereas the other algorithms, especially CART, showed larger and partly systematic deviations (Figure 3B).

3.2. UTCI Assessment Scale and Thermal Stress Categories

As illustrated by Figure 4A, the UTCI-Fiala model output projected to the first two dimensions by PCA, which explained about 90% of total variance, exhibited a one-

dimensional structure along the UTCI categories from extreme cold to extreme heat, which followed an increasing course concerning the first principal component (dim1) in accordance with the original analysis [25]. Similarly, a one-dimensional structure also emerged with t-SNE; however, the resulting curve followed a non-monotonous pattern in each of the two dimensions (dim1 and dim2).
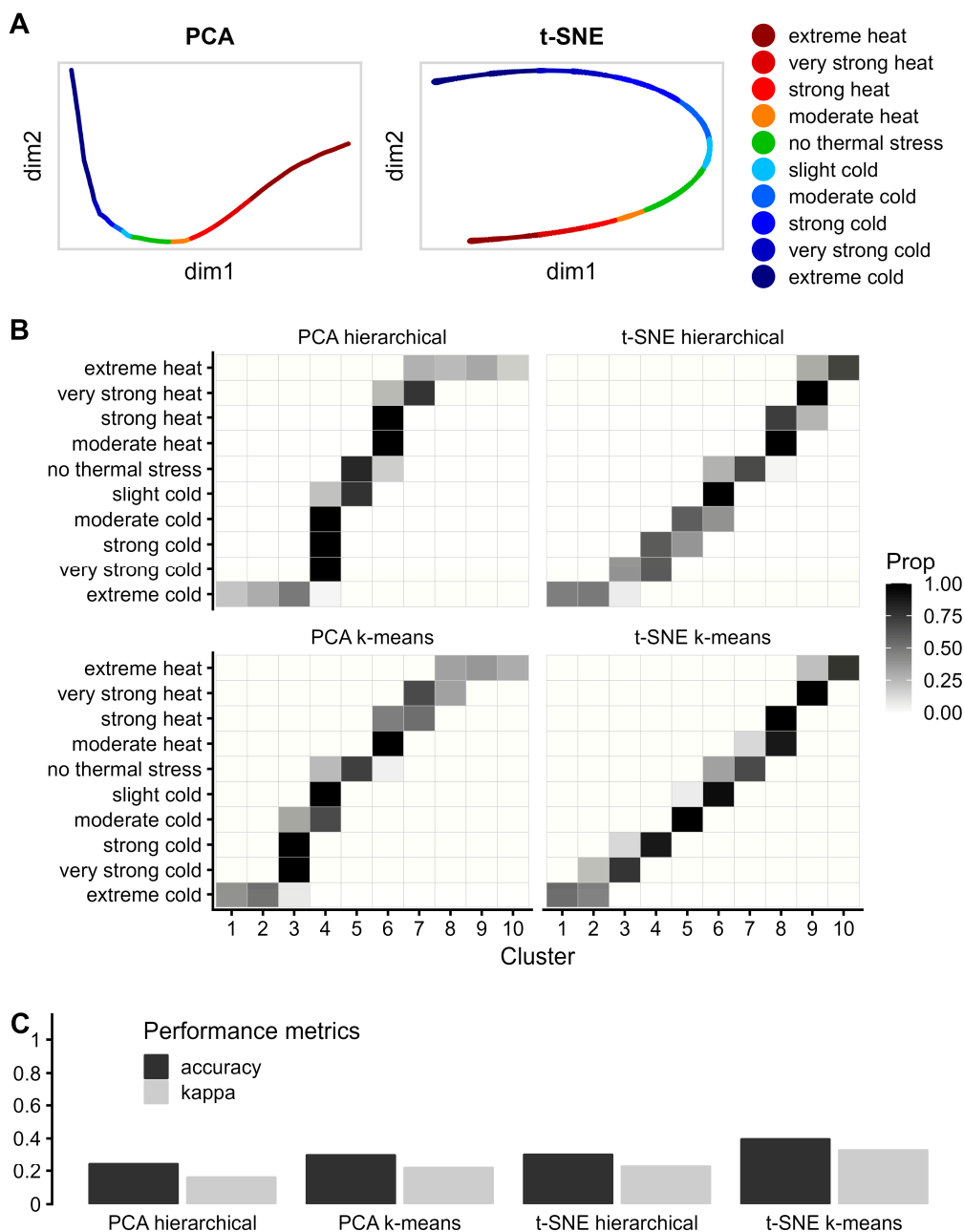
Figure 4. (**A**) Two-dimensional projections from principal component analysis (PCA), and t-distributed stochastic neighbor embedding (t-SNE), respectively, of the reference data (48 physiological variables plus 4 calculated dynamic thermal sensations) colored by the ten UTCI categories. (**B**) Confusion matrices showing the proportion (Prop) of conditions in the original ten UTCI stress categories (*y*-axis) assigned to the corresponding number of groups by hierarchical (upper panels) and k-means-clustering (**lower panels**) applied to the projections from PCA and t-SNE, respectively. (**C**) Metrics (accuracy and Cohen's kappa) on the agreement of the UTCI stress categories with the classifications by the different clustering algorithms applied to PCA and t-SNE projections.

Figure 4B shows the confusion matrices comparing the groups formed by hierarchical or k-means clustering following dimensionality reduction by PCA and t-SNE, respectively, to the categories of the UTCI assessment scale. Although correct classifications followed a trend from extreme cold to heat, discrepancies concerning the intermediate UTCI categories were obvious.

Consequently, the quantification of agreement in Figure 4C resulted in low to fair levels of agreement, according to Cohen's kappa, varying between 0.1 and 0.4 [49]. The different methods, as defined by the combinations of the dimension reduction (PCA vs. t-SNE) with the clustering algorithms (hierarchical vs. k-means), performed similarly with k-means clustering following t-SNE yielding the highest scores for both accuracy (0.395) and the kappa coefficient (0.33), respectively.

4. Discussion

Our study provided a quantitative assessment of the skills of SL algorithms, focusing on the data-driven stages 4 and 5 of the UTCI development process (Figure 1). Notably, both stages did involve consensus and expert judgment, e.g., for selecting the relevant subsets of two exposure durations (30 min and 120 min) and seven physiological output variables (Tre, Tskm, Tskfc, Mskdot, Shiv, wettA, and VblSk, with abbreviations from Table 1) in equivalent temperature calculation (stage 4). Similarly, the assessment scale from stage 5 was based on consensus about limit criteria derived from external expert knowledge from thermophysiology and ergonomics [25]. While Section 4.1 compares the skills of SL to the UTCI approach concerning these two stages of the development process, it should be noted that expert knowledge and judgment were inherently needed at all stages shown in Figure 1, especially with respect to the definition of the reference conditions in stage 3 as a crucial requirement for equivalent temperature calculation. Therefore, Section 4.2 relates our findings to corresponding results based on ensemble modeling as established for biological aging research [56,57], which is presented here in a novel context of thermal index development and could enable the calculation of equivalent temperatures without requiring the explicit definition of reference conditions, thus waiving stage 3 requiring expert judgment from the development process (Figure 1).

4.1. UTCI Approach Compared to SL Algorithms

Our results concerning the equivalent temperature index UTCI indicate that recent machine learning algorithms such as random forests or k-nearest neighbors were capable of producing equivalent temperatures deviating from UTCI with an RMSE of about 3 °C for the general, non-reference use case. For comparison, this deviation was similar to, or in the case of the k-nearest neighbor regression algorithm, even below the RMSE associated with calculating UTCI from the meteorological input variables using the polynomial regression function [25]. The non-parametric k-nearest neighbor regression algorithm might be better suited to capture non-monotonous stress–strain relationships like metabolic rate increasing under both cold and heat stress conditions (Figures 2 and A1), which could explain the superior performance of k-nearest neighbor regression compared to the other SL algorithms. The deviations found for the two best-performing algorithms might be considered acceptable given that uncertainties in the measurement or estimation of the mean radiant temperature in application scenarios could lead to absolute deviations above 6 K concerning the UTCI calculation by the polynomial regression function, which might even increase to more than 8 K when taking the other meteorological input variables (air temperature, humidity, wind speed) into account [58]. In addition, the values of dynamic thermal sensation (DTS) showed only limited variability at equivalent temperatures produced by random forest and k-nearest neighbors (Figure 3B), which was of similar magnitude as observed for the original UTCI. As identical equivalent temperature values should indicate identical thermal strain [12], this supports the potential usefulness of random forest and k-nearest neighbor regression for thermal index development.

In summary, our results suggest that selected SL algorithms may be helpful tools to derive summarizing metrics for complex environmental assessment tasks characterized by high-dimensional data describing stress and strain. Notably, these useful features concerned a narrow, but important data-analytical task within the UTCI development (stage 4 in Figure 1), which itself had been based on multivariate statistical methods [25], bearing some similarities with the SL algorithms applied in this study. From the opposite perspective, the close similarities of UTCI with equivalent temperatures calculated by modern machine learning algorithms suggest that the UTCI approach still appears appropriate considering recent data analytic developments.

The dimensionality-reduction results by PCA and t-SNE in Figure 4A both suggested a one-dimensional structure of thermos-physiological strain, concurring with the original analyses presented as stage 4 of Figure 1 [25]. In addition, the clustering algorithms were able to identify a trend from extreme heat to extreme cold. However, they did not reliably discriminate between the intermediate categories, with only marginal differences between the diverse algorithms. This demonstrates the discrepancy between clustering algorithms searching for patterns in the data and forming groups of data of comparable size on the one hand [59] and the definition of UTCI stress categories based on thermo-physiological knowledge and ergonomics reasoning on the other hand [25,34].

4.2. Alternative Approach by Ensemble Modeling

Ensemble modeling techniques such as random forests and gradient boosting, which had shown good performance in this study (Figure 3), aim at enhancing the predictive capabilities of simple models by averaging single predictions of numerous models and have been successfully applied to the assessment of human biological features changing with aging by a single quantity termed 'biological age' [56,57]. Biological age is defined as the chronological age of a population reference associated with identical levels of single or composite aging biomarkers, where the population reference is usually taken as a means of a comprehensive sample obtained from large-scale clinical or epidemiological studies [57]. Thus, the concept of biological age bears close similarities with equivalent temperature-based thermal indices [12], with age taking the role of temperature [60].

Therefore, as additional analysis, we applied the widely used Klemera-Doubal method (KDM) [56], which predicts the outcome of interest, here air temperature replacing age, as a weighted average of the predictions obtained by inverting the individual regression lines, as shown in Figure A1 for the predicted 48 physiological responses from the combined reference and non-reference datasets. The weights are chosen by the KDM algorithm proportional to the squared correlation coefficients. We calculated KDM equivalent temperatures using the R package BioAge [57], and compared them in Figure 5 to the UTCI values determined for the external test data [42]. Both types of equivalent temperatures were highly correlated with UTCI values above KDM equivalent temperatures in the heat and below KDM in the cold (Figure 5).

It is important to note that the KDM algorithms allow for the determination of equivalent temperatures without defining reference conditions, thus omitting stage 3 from Figure 1. In KDM, the reference conditions are unknown, but implicitly referring to those conditions representing the 'population average' in the underlying data, with air temperature equaling the predicted equivalent temperature after the complex ensemble modeling algorithm involving inverting and weighting individual regression functions. Concerning the combined reference and non-reference data shown in Figure A1, this population average will be characterized by higher values of wind speed and mean radiant temperature compared to the UTCI reference conditions. This explains, at least partly, the pattern observed in Figure 5, because the non-reference test data will be further away from the UTCI reference than from the implicit population average in KDM in terms of wind speed and mean radiant temperature. This will increase the offset, i.e., the shift of equivalent to air temperature as shown in Figure 1 for UTCI compared to KDM, and lead to higher UTCI values in the heat and lower values in the cold.
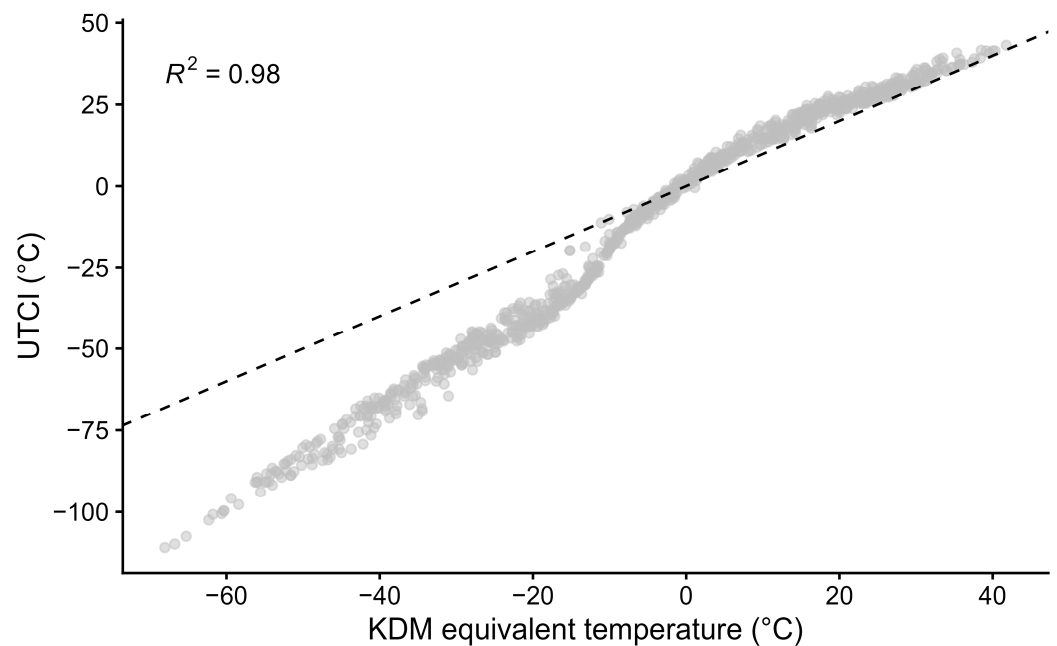
Figure 5. UTCI values and KDM equivalent temperatures for the external test data comprising 1000 non-reference conditions with a squared correlation coefficient (R^2) and a dashed line of identity (y = x).

As the KDM procedure removes the necessity of finding consensus about the definition of the reference conditions, it could simplify the development process of an equivalent temperature index. On the other hand, it will make the resulting equivalent temperature less interpretable due to the unspecified reference conditions. Consequently, the utilization of the KDM algorithm for the development of an equivalent temperature index will require the most careful design of the underlying database, which will implicitly also define the (unknown) reference to which all conditions will be compared.

Hence, we believe that the UTCI approach with a specified reference will promote results that are more interpretable in terms of thermal stress and physiological strain, which forms an essential requirement for assessment procedures concerning the human responses to the thermal environment [12,61].

Of course, the KDM algorithm could also be used to derive predictive equations for the equivalent temperatures from the reference data only and then apply those to the non-reference data, in the same manner as it was accomplished for the other SL algorithms in Section 3.1. Implementing this analysis, KDM showed inferior performance, with RMSE exceeding 23 °C for the reference and 42 °C for the non-reference conditions, far above the deviations reported for the other SL algorithms in Figure 3A. According to this result, the application of KDM for the development of equivalent temperature indices will be limited to the special case when the definition of reference conditions will not be possible or warranted.

4.3. Limitations and Outlook

As a limitation concerning generalizability, the database underlying the UTCI development was generated by a deterministic model of human thermoregulation. Thus, it lacks random variation induced by, e.g., inter-individual or day-to-day variability of the human physiological responses, which might have complicated the analyses of this study. However, the best-performing SL algorithms for equivalent temperature prediction, i.e., concerning stage 4 in Figure 1, like random forests and k-nearest neighbors, had already been suggested and successfully applied in other fields pertaining to climatic change and thermal stress [35–41].

While selected SL algorithms were competitive with procedures involving expert judgment for equivalent temperature prediction (stage 4 in Figure 1), assessment scales (stage 5 in Figure 1) automatically derived by clustering algorithms may still require expert knowledge for their refinement. Future enhancements at this stage may be possible by including further artificial intelligence tools such as expert systems or large language models working with knowledge databases [62,63].

Notably, the KDM approach, transferred in this study from biological aging research to a novel context, might facilitate the development of equivalent temperature indices by avoiding the necessity of specifying reference conditions (stage 3 in Figure 1) in future applications.

5. Conclusions

In summary, a potential supportive role for the utilized SL algorithms when analyzing high-dimensional input in thermal index development can be concluded from their adequate performance in equivalent temperature prediction. On the other hand, the low agreement of the assessment scale defined by the UTCI expert group with clustering-based thermal stress categories suggested that statistical learning algorithms will not (yet) fully replace the knowledgeable experts in biometeorological and inter-disciplinary research. Nevertheless, we believe that the findings from our study will contribute to informing biometeorological experts on the purposeful application of SL algorithms in thermal risk assessment.

Appendix A

This appendix contains supplemental Figures A1 and A2.

Figure A1 illustrates the distribution and correlation of air temperature with the twelve physiological responses at four points in time predicted by the UTCI-Fiala model [27] for the combined reference and non-reference datasets, i.e., training and test data, including the individual regression lines. The latter were used by the KDM algorithm [56,57] for ensemble modeling of the equivalent temperature as averaged predictions from the inverted regression lines weighted by the squared correlation coefficients, as discussed in Section 4.2.

Figure A2 presents the squared correlation coefficients (R^2) and the mean absolute error (MAE) as goodness-of-fit metrics complementing the root mean squared error (RMSE) presented in Figure 3A for the UTCI equivalent temperature predictions from diverse SL algorithms.
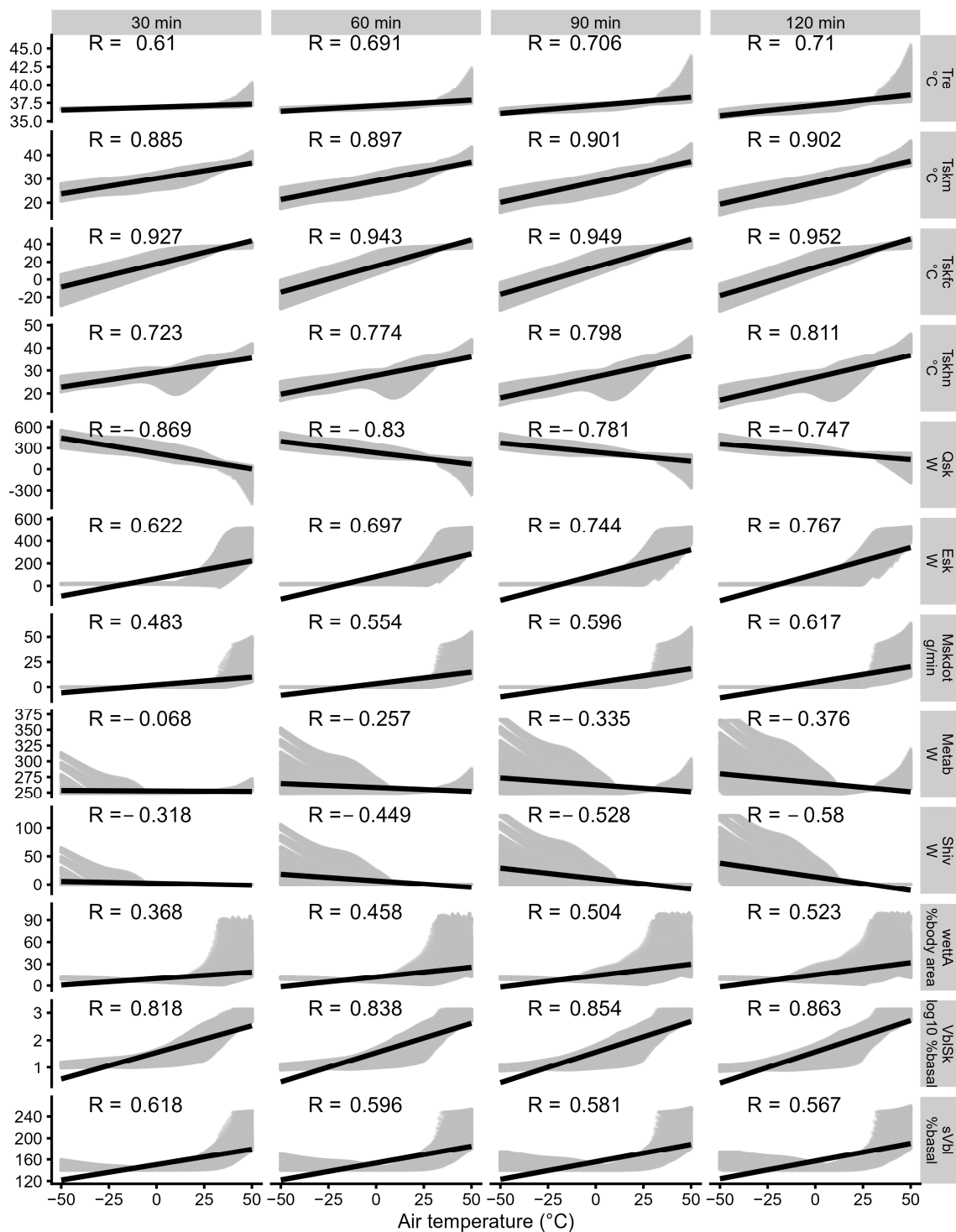
Figure A1. Set of twelve physiological output variables (abbreviations in Table 1) at four exposure times from the combined reference and non-reference conditions plotted related to air temperatures ranging from −50 °C to +50 °C with correlation coefficients (R) and linear regression lines. Note the log transformation applied to the skin blood flow data (VblSk) shown in the second row from the bottom up.
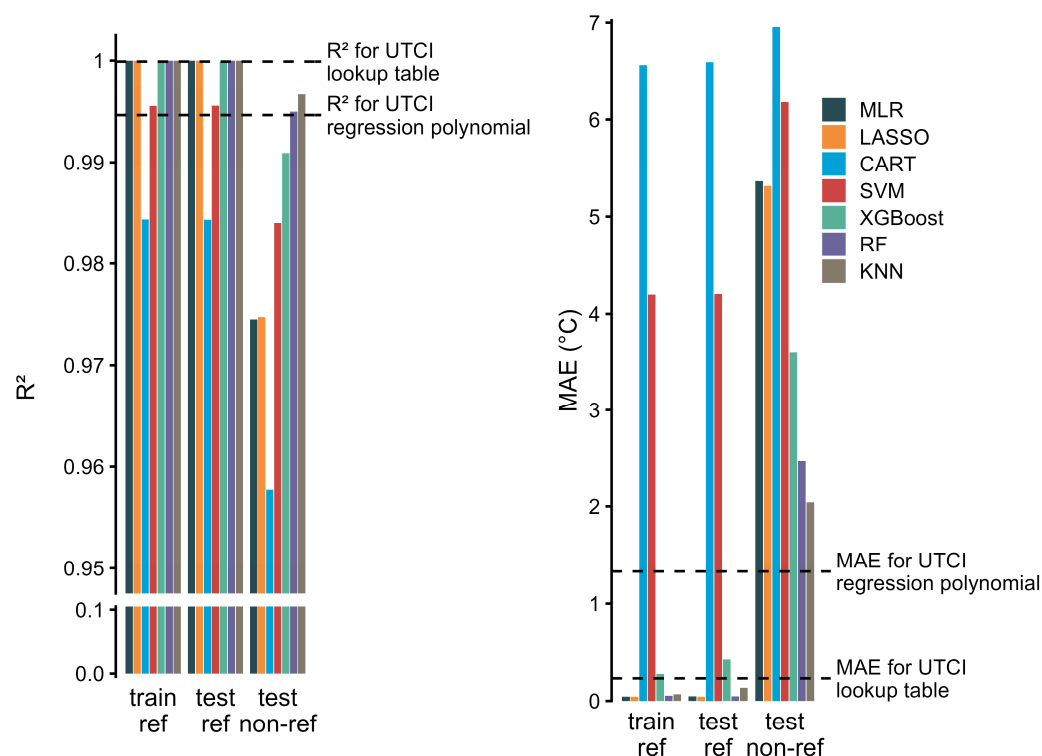
Figure A2. Squared correlation coefficients (R^2, **left panel**) and mean absolute error (MAE, **right panel**) as goodness-of-fit metrics in addition to the RMSE presented in Figure 3A for predicting UTCI equivalent temperatures from 48 physiological variables by different supervised SL algorithms (MLR: multiple linear regression; LASSO: least absolute shrinkage and selection operator; CART: classification and regression trees; SVM: support vector machines; XGBoost: extreme gradient boosting; RF: random forests; KNN: k-nearest neighbors) for the training reference (train ref) data and two test datasets for reference (ref) and non-reference (non-ref) UTCI conditions. For comparison, horizontal reference lines indicate the corresponding metrics for calculating UTCI from meteorological input using the polynomial regression function or the lookup table provided by the operational procedure [25].

References

1. Berk, R.A. *Statistical Learning from a Regression Perspective*; Springer International Publishing: Cham, Switzerland, 2020. [CrossRef]
2. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: New York, NY, USA, 2009. [CrossRef]
3. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning: With Applications in R*; Springer: New York, NY, USA, 2013. [CrossRef]
4. Matzarakis, A. Curiosities about Thermal Indices Estimation and Application. *Atmosphere* **2021**, *12*, 721. [CrossRef]
5. Ali, A.; Jayaraman, R.; Azar, E.; Maalouf, M. A comparative analysis of machine learning and statistical methods for evaluating building performance: A systematic review and future benchmarking framework. *Build. Environ.* **2024**, *252*, 111268. [CrossRef]
6. Aparicio-Ruiz, P.; Barbadilla-Martín, E.; Guadix, J.; Muñuzuri, J. Predicting the clothing insulation through machine learning algorithms: A comparative analysis and a practical approach. *Build. Simul.* **2024**, *17*, 839–855. [CrossRef]
7. Hamed, M.M.; Al-Hasani, A.A.J.; Nashwan, M.S.; Sa'adi, Z.; Shahid, S. Assessing the growing threat of heat stress in the North Africa and Arabian Peninsula region connected to climate change. *J. Clean. Prod.* **2024**, *447*, 141639. [CrossRef]
8. Guo, R.; Yang, B.; Guo, Y.; Li, H.; Li, Z.; Zhou, B.; Hong, B.; Wang, F. Machine learning-based prediction of outdoor thermal comfort: Combining Bayesian optimization and the SHAP model. *Build. Environ.* **2024**, *254*, 111301. [CrossRef]
9. Wang, J.; Li, Q.; Zhu, G.; Kong, W.; Peng, H.; Wei, M. Recognition and prediction of elderly thermal sensation based on outdoor facial skin temperature. *Build. Environ.* **2024**, *253*, 111326. [CrossRef]
10. Wang, M.; Gou, Z. Gaussian Mixture Model based classification for analyzing longitudinal outdoor thermal environment data to evaluate comfort conditions in urban open spaces. *Urban Clim.* **2024**, *53*, 101792. [CrossRef]
11. Choi, Y.; Seo, S.; Lee, J.; Kim, T.W.; Koo, C. A machine learning-based forecasting model for personal maximum allowable exposure time under extremely hot environments. *Sustain. Cities Soc.* **2024**, *101*, 105140. [CrossRef]
12. Jendritzky, G.; de Dear, R.; Havenith, G. UTCI—Why another thermal index? *Int. J. Biometeorol.* **2012**, *56*, 421–428. [CrossRef]

13. Geletič, J.; Lehnert, M.; Krč, P.; Resler, J.; Krayenhoff, E.S. High-Resolution Modelling of Thermal Exposure during a Hot Spell: A Case Study Using PALM-4U in Prague, Czech Republic. *Atmosphere* **2021**, *12*, 175. [CrossRef]

14. Dimitriadou, L.; Nastos, P.; Zerefos, C. Defining Heatwaves with Respect to Human Biometeorology. The Case of Attica Region, Greece. *Atmosphere* **2021**, *12*, 1100. [CrossRef]

15. Zeng, D.; Wu, J.; Mu, Y.; Deng, M.; Wei, Y.; Sun, W. Spatial-Temporal Pattern Changes of UTCI in the China–Pakistan Economic Corridor in Recent 40 Years. *Atmosphere* **2020**, *11*, 858. [CrossRef]

16. Basarin, B.; Lukić, T.; Matzarakis, A. Review of Biometeorology of Heatwaves and Warm Extremes in Europe. *Atmosphere* **2020**, *11*, 1276. [CrossRef]

17. Staiger, H.; Laschewski, G.; Matzarakis, A. Selection of Appropriate Thermal Indices for Applications in Human Biometeorological Studies. *Atmosphere* **2019**, *10*, 18. [CrossRef]

18. Di Napoli, C.; Pappenberger, F.; Cloke, H.L. Verification of Heat Stress Thresholds for a Health-Based Heat-Wave Definition. *J. Appl. Meteorol. Climatol.* **2019**, *58*, 1177–1194. [CrossRef]

19. Potchter, O.; Cohen, P.; Lin, T.-P.; Matzarakis, A. A systematic review advocating a framework and benchmarks for assessing outdoor human thermal perception. *Sci. Total Environ.* **2022**, *833*, 155128. [CrossRef]

20. Urban, A.; Di Napoli, C.; Cloke, H.L.; Kyselý, J.; Pappenberger, F.; Sera, F.; Schneider, R.; Vicedo-Cabrera, A.M.; Acquaotta, F.; Ragettli, M.S.; et al. Evaluation of the ERA5 reanalysis-based Universal Thermal Climate Index on mortality data in Europe. *Environ. Res.* **2021**, *198*, 111227. [CrossRef]

21. Di Napoli, C.; Barnard, C.; Prudhomme, C.; Cloke, H.L.; Pappenberger, F. ERA5-HEAT: A global gridded historical dataset of human thermal comfort indices from climate reanalysis. *Geosci. Data J.* **2020**, *8*, 2–10. [CrossRef]

22. Di Napoli, C.; Pappenberger, F.; Cloke, H.L. Assessing heat-related health risk in Europe via the Universal Thermal Climate Index (UTCI). *Int. J. Biometeorol.* **2018**, *62*, 1155–1165. [CrossRef]

23. Ioannou, L.G.; Tsoutsoubi, L.; Mantzios, K.; Vliora, M.; Nintou, E.; Piil, J.F.; Notley, S.R.; Dinas, P.C.; Gourzoulidis, G.A.; Havenith, G.; et al. Indicators to assess physiological heat strain—Part 3: Multi-country field evaluation and consensus recommendations. *Temperature* **2022**, *9*, 274–291. [CrossRef]

24. Brimicombe, C.; Di Napoli, C.; Cornforth, R.; Pappenberger, F.; Petty, C.; Cloke, H.L. Borderless Heat Hazards With Bordered Impacts. *Earth's Future* **2021**, *9*, e2021EF002064. [CrossRef]

25. Bröde, P.; Fiala, D.; Blazejczyk, K.; Holmér, I.; Jendritzky, G.; Kampmann, B.; Tinz, B.; Havenith, G. Deriving the operational procedure for the Universal Thermal Climate Index (UTCI). *Int. J. Biometeorol.* **2012**, *56*, 481–494. [CrossRef] [PubMed]

26. McGregor, G.R. Special issue: Universal Thermal Climate Index (UTCI). *Int. J. Biometeorol.* **2012**, *56*, 419. [CrossRef] [PubMed]

27. Fiala, D.; Havenith, G.; Bröde, P.; Kampmann, B.; Jendritzky, G. UTCI-Fiala multi-node model of human heat transfer and temperature regulation. *Int. J. Biometeorol.* **2012**, *56*, 429–441. [CrossRef] [PubMed]

28. Havenith, G.; Fiala, D.; Blazejczyk, K.; Richards, M.; Bröde, P.; Holmér, I.; Rintamaki, H.; Ben Shabat, Y.; Jendritzky, G. The UTCI-clothing model. *Int. J. Biometeorol.* **2012**, *56*, 461–470. [CrossRef] [PubMed]

29. Psikuta, A.; Fiala, D.; Laschewski, G.; Jendritzky, G.; Richards, M.; Blazejczyk, K.; Mekjavic, I.B.; Rintamäki, H.; de Dear, R.; Havenith, G. Validation of the Fiala multi-node thermophysiological model for UTCI application. *Int. J. Biometeorol.* **2012**, *56*, 443–460. [CrossRef] [PubMed]

30. Kampmann, B.; Bröde, P.; Fiala, D. Physiological responses to temperature and humidity compared to the assessment by UTCI, WGBT and PHS. *Int. J. Biometeorol.* **2012**, *56*, 505–513. [CrossRef] [PubMed]

31. Bröde, P.; Krüger, E.L.; Rossi, F.A.; Fiala, D. Predicting urban outdoor thermal comfort by the Universal Thermal Climate Index UTCI—A case study in Southern Brazil. *Int. J. Biometeorol.* **2012**, *56*, 471–480. [CrossRef] [PubMed]

32. Bröde, P.; Kampmann, B. Temperature–Humidity-Dependent Wind Effects on Physiological Heat Strain of Moderately Exercising Individuals Reproduced by the Universal Thermal Climate Index (UTCI). *Biology* **2023**, *12*, 802. [CrossRef]

33. Bröde, P.; Fiala, D.; Kampmann, B.; Havenith, G.; Jendritzky, G. Der Klimaindex UTCI—Multivariate Analyse der Reaktion eines thermophysiologischen Simulationsmodells. In Proceedings of the 55. Kongress der Gesellschaft für Arbeitswissenschaft, Dortmund, Germany, 4–6 March 2009; pp. 705–708.

34. Bröde, P.; Blazejczyk, K.; Fiala, D.; Havenith, G.; Holmér, I.; Jendritzky, G.; Kuklane, K.; Kampmann, B. The Universal Thermal Climate Index UTCI Compared to Ergonomics Standards for Assessing the Thermal Environment. *Ind. Health* **2013**, *51*, 16–24. [CrossRef]

35. Aguilera, J.J.; Korsholm Andersen, R.; Toftum, J. Prediction of Indoor Air Temperature Using Weather Data and Simple Building Descriptors. *Int. J. Environ. Res. Public Health* **2019**, *16*, 4349. [CrossRef] [PubMed]

36. Benita, F.; Tunçer, B. Exploring the effect of urban features and immediate environment on body responses. *Urban For. Urban Green.* **2019**, *43*, 126365. [CrossRef]

37. Berrang-Ford, L.; Sietsma, A.J.; Callaghan, M.; Minx, J.C.; Scheelbeek, P.F.D.; Haddaway, N.R.; Haines, A.; Dangour, A.D. Systematic mapping of global research on climate and health: A machine learning review. *Lancet Planet. Health* **2021**, *5*, e514–e525. [CrossRef] [PubMed]

38. Kim, J.; Schiavon, S.; Brager, G. Personal comfort models—A new paradigm in thermal comfort for occupant-centric environmental control. *Build. Environ.* **2018**, *132*, 114–124. [CrossRef]

39. Liu, K.; Nie, T.; Liu, W.; Liu, Y.; Lai, D. A machine learning approach to predict outdoor thermal comfort using local skin temperatures. *Sustain. Cities Soc.* **2020**, *59*, 102216. [CrossRef]

40. Mantzios, K.; Ioannou, L.G.; Panagiotaki, Z.; Ziaka, S.; Périard, J.D.; Racinais, S.; Nybo, L.; Flouris, A.D. Effects of Weather Parameters on Endurance Running Performance: Discipline Specific Analysis of 1258 Races. *Med. Sci. Sports Exerc.* **2021**, *54*, 153–161. [CrossRef] [PubMed]

41. Shin, J.-Y.; Min, B.; Kim, K.R. High-resolution wind speed forecast system coupling numerical weather prediction and machine learning for agricultural studies—A case study from South Korea. *Int. J. Biometeorol.* **2022**, *66*, 1429–1443. [CrossRef] [PubMed]

42. Bröde, P. UTCI-Test-Data. Zenodo: 2021. Available online: https://zenodo.org/records/5503968 (accessed on 4 October 2022).

43. ASHRAE. *Standard 55—Thermal Environmental Conditions for Human Occupancy*; ASHRAE Inc.: Atlanta, GA, USA, 2004.

44. Fiala, D.; Lomas, K.J.; Stohrer, M. First principles modeling of thermal sensation responses in steady-state and transient conditions. *ASHRAE Trans.* **2003**, *109*, 179–186.

45. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]

46. Genuer, R.; Poggi, J.-M. *Random Forests with R*; Springer International Publishing: Cham, Switzerland, 2020. [CrossRef]

47. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; Association for Computing Machinery: San Francisco, CA, USA, 2016; pp. 785–794.

48. van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

49. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]

50. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2024.

51. Kuhn, M. Building Predictive Models in R Using the caret Package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]

52. Kuhn, M.; Vaughan, D.; Hvitfeldt, E. Yardstick: Tidy Characterizations of Model Performance—R Package Version 1.3.1. 2024. Available online: https://CRAN.R-project.org/package=yardstick (accessed on 19 April 2024).

53. Wickham, H.; Averick, M.; Bryan, J.; Chang, W.; McGowan, L.D.A.; François, R.; Grolemund, G.; Hayes, A.; Henry, L.; Hester, J.; et al. Welcome to the Tidyverse. *J. Open Source Softw.* **2019**, *4*, 1686. [CrossRef]

54. Wilke, C.O. Cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'—R Package Version 1.1.3. 2024. Available online: https://CRAN.R-project.org/package=cowplot (accessed on 18 April 2024).

55. Kampmann, B.; Bröde, P. Heat Acclimation Does Not Modify Q10 and Thermal Cardiac Reactivity. *Front. Physiol.* **2019**, *10*, 1524. [CrossRef] [PubMed]

56. Klemera, P.; Doubal, S. A new approach to the concept and computation of biological age. *Mech. Ageing Dev.* **2006**, *127*, 240–248. [CrossRef]

57. Kwon, D.; Belsky, D.W. A toolkit for quantification of biological age from blood chemistry and organ function test data: BioAge. *GeroScience* **2021**, *43*, 2795–2808. [CrossRef] [PubMed]

58. Weihs, P.; Staiger, H.; Tinz, B.; Batchvarova, E.; Rieder, H.; Vuilleumier, L.; Maturilli, M.; Jendritzky, G. The uncertainty of UTCI due to uncertainties in the determination of radiation fluxes derived from measured and observed meteorological data. *Int. J. Biometeorol.* **2012**, *56*, 537–555. [CrossRef] [PubMed]

59. Härdle, W.; Simar, L. *Applied Multivariate Statistical Analysis*, 2nd ed.; Springer: Berlin, Germany, 2007.

60. Bröde, P.; Claus, M.; Gajewski, P.D.; Getzmann, S.; Wascher, E.; Watzl, C. From Immunosenescence to Aging Types–Establishing Reference Intervals for Immune Age Biomarkers by Centile Estimation. *Int. J. Mol. Sci.* **2023**, *24*, 13186. [CrossRef]

61. Notley, S.R.; Mitchell, D.; Taylor, N.A.S. A century of exercise physiology: Concepts that ignited the study of human thermoregulation. Part 3: Heat and cold tolerance during exercise. *Eur. J. Appl. Physiol.* **2024**, *124*, 1–145. [CrossRef]

62. Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; et al. A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.* **2024**, *15*, 39. [CrossRef]

63. Asemi, A.; Ko, A.; Nowkarizi, M. Intelligent libraries: A review on expert systems, artificial intelligence, and robot. *Libr. Hi Tech* **2021**, *39*, 412–434. [CrossRef]