LSTM-based safety-oriented prediction of big toe skin temperature in extreme cold conditions for mountaineering footwear evaluation

Scientific Reports

https://doi.org/10.1038/s41598-026-52990-x

Article in Press

LSTM-based safety-oriented prediction of big toe skin temperature in extreme cold conditions for mountaineering footwear evaluation

Received: 12 January 2026

Accepted: 8 May 2026

Published online: 22 May 2026

Cite this article as: Bianca E., Marcato A., Ferri A. *et al.* LSTM-based safety-oriented prediction of big toe skin temperature in extreme cold conditions for mountaineering footwear evaluation. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-52990-x

Eleonora Bianca, Agnese Marcato, Ada Ferri & Gianluca Boccardo

LSTM-Based Safety-Oriented Prediction of Big Toe Skin Temperature in Extreme Cold Conditions for Mountaineering Footwear Evaluation

Eleonora Bianca[1,*], **Agnese Marcato**[1], **Ada Ferri**[1], **and Gianluca Boccardo**[1]

[1]Department of Applied Science and Technology, Polytechnic of Turin, Turin, Italy
[*]eleonora.bianca@polito.it

ABSTRACT

This study proposes a safety-oriented framework for the assessment of mountaineering footwear insulation in extreme cold environments. A novel metric, the Duration of Safe Exposure (DSE), is introduced and defined as the time required for big toe skin temperature to reach the conservative critical threshold of 15°C under a given combination of footwear insulation, ambient temperature, and physical activity.
Four structured experimental campaigns were conducted in a controlled climatic chamber to investigate the interaction between footwear thermal resistance, ambient temperature, and physical activity on peripheral thermoregulation. Big toe temperature was selected as a benchmark indicator due to its pronounced vasoconstrictive response during cold exposure. In total, the dataset comprised 96 time series collected across 10 experimental protocols, each corresponding to subject-specific tests under defined conditions. An LSTM-based neural network was trained to predict the temporal evolution of big toe temperature from mean skin temperature and the three operational parameters.
The evaluation framework was explicitly structured to reflect the safety-oriented objective of the study and was fully consistent with the custom loss function adopted during training, which jointly penalises trajectory deviation (RMSE) and errors in the prediction of the threshold-crossing event. Beyond regression-based performance indicators, the model was evaluated through a binary classification of DSE (Duration of Safe Exposure) occurrence, defined as the network's ability to correctly predict whether and when the big toe temperature reaches the critical 15°C threshold. Based on true and false positive/negative outcomes, classification metrics including Accuracy, Precision, Recall, and F1-score were computed to quantify the reliability of threshold detection within a predefined $\pm$10-minute tolerance window.
To further address safety implications, a dedicated safety-specific indicator, the Unsafe Error Rate (UER), was introduced to explicitly identify non-conservative DSE overestimations, i.e., cases in which the model predicts a longer safe exposure time than physiologically observed. Results indicate that unsafe predictions were rare, with most estimates being either within tolerance or conservative.
Finally, the trained network was integrated with the JOS-3 thermoregulation model to enable exploration of physiologically consistent virtual scenarios. This hybrid framework provides a quantitative, physiologically grounded approach that may support the investigation of safety-related aspects of mountaineering footwear under controlled conditions. However, its applicability remains limited to the specific experimental setup and population considered, since the experimental data were collected exclusively from young male participants, and cannot be generalised to other populations without further validation.

Introduction

Mountaineering boots are commonly classified as sports goods and, as such, are not subject to specific performance standards addressing protection against extreme cold exposure. However, these products are routinely used in high-altitude and polar environments, where prolonged exposure to low temperatures may lead to severe cold injuries, including frostbite and, in extreme cases, limb loss or life-threatening conditions[1]. Despite this high-risk context, the evaluation of their thermal performance is often based on empirical experience and manufacturer expertise rather than on physiologically validated safety criteria.
From a technical perspective, mountaineering footwear represents a highly complex, multi-layered system. A single boot may include more than twenty materials distributed across different structural regions, each required for specific mechanical, waterproofing, or insulating functions. This structural heterogeneity already complicates purely material-based thermal assessment. The complexity increases further when considering that the boot does not operate in isolation, but as part of a coupled human–product system in which thermoregulation, vascular mechanisms, metabolic heat production, and environmental conditions dynamically interact. Standard thermal manikin measurements, although valuable for evaluating thermal resistance and heat transfer properties of footwear or clothing, are based on physical heat flux and surface temperature control and do not fully capture the active physiological responses of human thermoregulation observed during real cold exposure and activity,

such as vasoconstriction, shivering, and metabolic heat production, leading to notable differences compared to human trials under similar conditions[2].

For this reason, evaluating thermal performance solely through dry material insulation values may be insufficient in extreme environments[3,4]. A safety-oriented assessment must incorporate the physiological response of the user. Previous studies[5] have identified 15 °C skin temperature as a conservative threshold below which the risk of cold-related discomfort and tissue damage significantly increases. Defining a reliable early-warning indicator capable of predicting the time required to reach this critical threshold is therefore essential for improving protective design.

Recent experimental investigations[6] demonstrated that, among foot regions, the big toe exhibits the most pronounced vasoconstrictive response and consistently records the lowest skin temperatures during cold exposure. Consequently, big toe skin temperature was selected as a benchmark indicator for defining the critical threshold in mountaineering conditions.

Human thermoregulation in cold environments can be interpreted as the dynamic interplay of three primary factors: the thermal insulation provided by clothing, ambient environmental conditions, and the level of physical activity performed by the subject[7].

To support a safety-oriented design approach, the concept of *Duration of Safe Exposure* (DSE) was introduced. DSE is defined as the maximum time an individual can remain under a specific combination of insulation, ambient conditions, and activity level before the big toe skin temperature reaches the conservative critical threshold of 15 °C.

To characterise this relationship, a series of controlled experimental campaigns were conducted on young male participants in a climatic chamber. During these campaigns, physiological parameters were recorded continuously, with particular focus on mean skin temperature and big toe temperature, while systematically varying the three governing factors: footwear insulation (defined as thermal resistance R_{cT} in $m^2 \cdot K \cdot W^{-1}$), ambient temperature (T_a in °C), and activity level (PA in met where 1 met equals to 58.2 $W \cdot m^{-2}$[8]) among the protocols.

The objective of these campaigns was to generate a structured dataset suitable for training a predictive algorithm capable of estimating the temporal evolution of big toe temperature as a function of these interacting variables. By modelling this time-dependent response, the proposed framework enables the estimation of DSE before the product deployment, thereby providing a quantitative tool to inform safety-oriented footwear design.

Predicting the Duration of Safe Exposure (DSE) requires modelling a highly non-linear and time-dependent physiological response governed by interacting environmental and behavioural variables. Several approaches have been explored in the field of time-series forecasting to model temporal dynamics, including widely adopted statistical methods such as ARIMA and SARIMAX[9,10]. These models have demonstrated robust performance in structured and quasi-stationary contexts, particularly in environmental and climate-related applications[10,11]. However, classical statistical forecasting techniques rely on assumptions of linearity and stationarity that are often not satisfied in human thermoregulatory processes. Peripheral temperature dynamics during cold exposure are characterised by non-linear feedback mechanisms, intermittent vasoconstrictive responses, and activity-dependent metabolic heat production, leading to regime shifts and complex temporal dependencies that are difficult to capture with traditional linear models.

To address these limitations, data-driven approaches based on deep learning have gained increasing relevance for modelling complex temporal physiological signals. In particular, Recurrent Neural Networks (RNNs)[12,13], and especially Long Short-Term Memory (LSTM)[14] architectures, are well suited for capturing long-term temporal dependencies in multivariate time series[15–17]. LSTMs mitigate the vanishing gradient problem typical of standard RNNs and can effectively learn non-linear relationships between multiple inputs and evolving physiological outputs.

Importantly, compared to highly parametrised attention-based architectures, LSTM models remain particularly appropriate in structured time-series settings where dataset size is moderate and temporal dynamics dominate the predictive task. Prior comparative studies[18,19] have shown that in small or task-specific datasets, simpler recurrent architectures can achieve performance comparable to or superior to more complex Transformer-based models, while reducing the risk of overfitting and excessive computational demand. Building on their demonstrated capability in biomedical and physiological signal modelling[13,20], an LSTM architecture was implemented to model the temporal evolution of big toe skin temperature using the same governing factors of the experimental campaigns as predictive inputs (R_{cT}, T_a, and PA) together with mean skin temperature as a physiological state indicator. In this way, the network learns the dynamic mapping between externally controlled conditions and peripheral thermal response, enabling the estimation of the Duration of Safe Exposure (DSE) under different operational scenarios.

Once trained and validated against experimental data, the LSTM-based framework can be extended beyond the directly tested conditions. To enable predictions under virtual scenarios not experimentally investigated, the JOS-3 thermoregulation model was employed to generate physiologically consistent input variables for the network[21]. JOS-3 is a validated multi-segment human thermoregulation model capable of simulating mean skin temperature and related physiological responses under varying environmental and activity conditions.

By coupling the physics-based JOS-3 model with the data-driven LSTM architecture, the proposed framework enables exploration of a broader range of environmental and operational scenarios. A similar hybrid modelling strategy, combining

mechanistic thermoregulation modelling with data-driven approaches, has previously been proposed in related physiological prediction contexts. Nevertheless, predictions outside the experimentally validated domain should be interpreted within the limitations inherent to the underlying thermophysiological assumptions[22].
By introducing a physiologically grounded predictive indicator into the design process, this work represents a step toward a more data-driven and safety-oriented evaluation paradigm, while acknowledging the need for further validation before practical application.

Methods

Experimental Design

The experimental protocol with human subjects was reviewed and approved by the Ethics Committee of the Polytechnic of Turin (Authorisation ID: 87212/2025). All participants provided written informed consent prior to participating in the study. In addition, personal data were anonymised and treated in accordance with the applicable data protection regulations.

Sampling Methods and Enrollment

Participants were eligible for participation if they met the following inclusion criteria: male gender, age between 18 and 55 years, good physical fitness confirmed by a valid medical sports certificate, and regular exposure to cold environments. All subjects had to be within a normal weight range (i.e., neither underweight nor overweight with a BMI between 18.5 and 24.9). Exclusion criteria included a history of cardiovascular, pulmonary, or musculoskeletal disease, and previous frostbite.
It should be noted that only male participants were included in the study. This choice represents a limitation, as sex-related physiological differences in thermoregulation may affect the generalisability of the results.

Protocols

Four experimental campaigns were conducted under controlled cold conditions, varying ambient temperature (–10 to –30°C), activity level, and footwear insulation.
All protocols were developed in collaboration with experienced mountaineers to accurately replicate real-life activity patterns typical of alpine expeditions.

Acquired Data

All tests were conducted in a controlled climate chamber (–10 to –30°C, 0.4–0.5 ms^{-1} air velocity). The experimental setup did not include forced high-speed ventilation or direct solar radiation. This choice was motivated by the need to ensure participant safety under severe cold exposure. Accordingly, the in vivo component was designed to investigate controlled thermophysiological responses rather than to replicate extreme meteorological scenarios.
Mean skin and local toe temperatures were continuously recorded using distributed iButton sensors and thermistors, in compliance with ISO 9886:2008. Heart rate was monitored with a chest-strap to verify activity levels and physiological effort. Further details on the instrumentation, calibration accuracy, sensor placement, and data acquisition protocols are provided in the Supplementary Material and in our previous work[6].

Data Preparation for Deep Learning

The temperature data collected during the experimental campaigns were organised into temporal arrays, each 70 minutes long, corresponding to 420 time steps (with a sampling frequency of one data point every 10 seconds). Each array represents a single test performed by a subject under a specific combination of experimental conditions. In total, 10 protocols were conducted, each with a different number of subjects (depending on participant availability for exposure to extremely cold conditions), resulting in a dataset of 96 independent time series.
It is important to note that, in this context, each test represents an independent experimental observation, while the temporal samples within each sequence are not statistically independent. Therefore, the effective sample size is determined by the number of subjects/tests rather than the number of time-series points.
This characteristic is typical of human physiological datasets, where data collection is inherently constrained by experimental complexity, cost, and ethical considerations. Similar studies in the field of physiological signal analysis and biomedical deep learning report datasets composed of a limited number of subjects, often in the order of a few tens, despite the large number of temporal samples obtained through segmentation or continuous acquisition[23–25].
Within this framework, deep learning models are applied to extract temporal patterns from structured sequences, while validation strategies are designed to ensure generalisability across subjects rather than across individual time points. Each array was labelled according to three parameters: thermal insulation of the boot model (R_{cT}), ambient temperature (T_a), and level of physical activity, as detailed in the Supporting Information.
`MinMaxScaler` normalisation was applied.

The neural network was structured to receive the following input (X) and generate the corresponding output (Y):

X: a three-dimensional matrix of shape [batch size, 420, 4], where the four features correspond to mean skin temperature, thermal insulation (R_{cT} in $Km^{-2}W^{-1}$), ambient temperature (T_a in °C), and physical activity level (PA in met).

Y: a matrix of shape [batch size, 420, 1], representing the predicted big toe temperature over time.

Deep Learning Model

Model Construction

Among RNN variants, Long Short-Term Memory (LSTM) networks were chosen for their proven ability to capture long-term temporal dependencies and non-linear thermal dynamics[26,27].
The model parameters were optimised through an extensive hyperparameter tuning process, aiming to maximise accuracy and generalisability across experimental conditions. Hyperparameter tuning was performed using Keras Tuner's `Hyperband` algorithm.
This method, introduced by Jamieson and Talwalkar[28,29], differs from traditional optimisation techniques such as Bayesian optimisation in that it exploits the iterative nature of training algorithms to adaptively allocate computational resources and stop poorly performing configurations early. In contrast to Bayesian methods, which often allocate fixed budgets and treat the model as a black box, Hyperband dynamically adapts resource allocation, enabling much faster convergence. This approach is particularly beneficial when model training is computationally intensive, as it speeds up the identification of promising hyperparameters without requiring assumptions on convergence rates, making it highly scalable and practical for real-world applications.

Final Model Architecture

After hyperparameter optimisation, the final LSTM model comprised two recurrent layers followed by a compact feedforward block.
In earlier versions of the architecture, the model underestimated the initial big toe temperature before gradually correcting itself, resulting in an observable bell-shaped bias at the beginning of the predicted sequence. To improve training stability and prevent systematic underestimation of the initial big toe temperature (observed in early architectural versions as a bell-shaped bias at sequence onset), an informed initialisation strategy was introduced during model training. Specifically, the hidden and cell states of the first LSTM layer were initialised by mapping the first measured big toe temperature of each training sequence through two dense layers (384 units, tanh activation)[30,31]. Providing the network with an informed initial state derived from the first measurement improved temporal alignment at sequence onset, reducing early prediction error and enhancing the fit to the measured temperature profile. This initialisation strategy was employed during training to stabilise convergence and improve early-sequence prediction accuracy. Once trained, the model no longer requires an external measurement of big toe temperature for inference. In virtual or experimentally untested scenarios, the LSTM generates temperature trajectories solely from the input variables (R_{cT}, T_a, and PA), as the learned internal dynamics encode appropriate initial temporal behaviour without requiring real-time physiological measurements.
The first LSTM layer included 384 units and returned full sequences to enable temporal processing by the subsequent layer. The second LSTM layer comprised 256 units, also returning sequences. A 5% dropout was applied after the LSTM block to prevent overfitting by randomly deactivating neurons during training.
The recurrent block was followed by a dense layer with 80 units and `tanh` activation, which refined the temporal features before the final prediction stage. The output layer consisted of a single neuron with linear activation, providing the estimated big toe temperature at each time step.
Figure 1 presents a schematic representation of the final model architecture used for big toe temperature prediction, including the custom initialisation strategy for hidden and cell states.

Model Validation and Evaluation

To ensure subject-independent evaluation, a deterministic protocol-stratified holdout strategy was adopted. For each experimental condition (R_{cT}, Ta, PA), one subject (the first available in the corresponding protocol folder) was selected for validation, and all trials belonging to these subjects were excluded from the training set. This procedure guarantees the absence of subject overlap between training and validation sets.
Such a subject-independent split is particularly appropriate for physiological datasets, where inter-individual variability can substantially influence model performance.
In addition to the conventional loss metrics, a custom loss function was developed to assess the ability of the model to predict the time at which the big toe temperature falls below the critical threshold of 15°C[5,6]. The mean square standard error (MSE) penalises the differences between the predicted and measured temperature values over the entire time sequence, but does

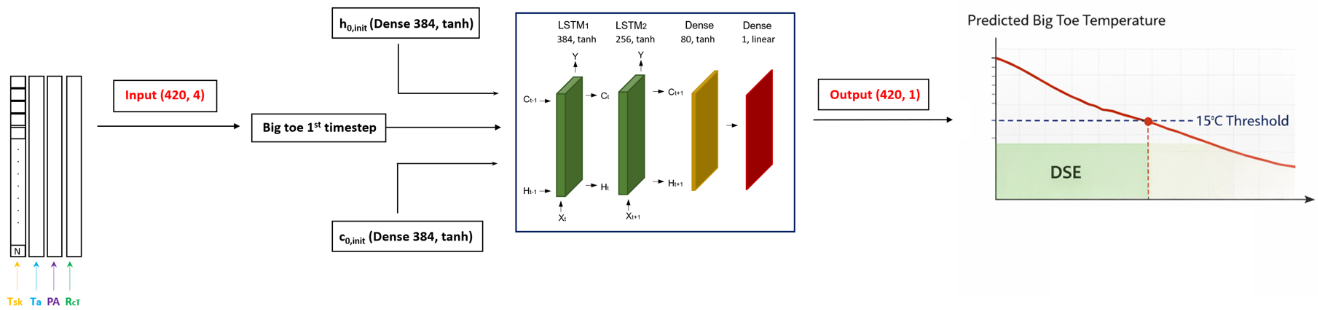

Figure 1. Compact schematic of the final LSTM model for big toe temperature prediction, including hidden and cell state initialization from the first measured value.

not explicitly capture the accuracy of the predicted time at which the critical threshold is exceeded. As this point in time is particularly relevant for thermal comfort and risk assessment, an additional component was included in the loss function. The first step was to define a differentiable function, referred to here as a *Differentiable Threshold-Crossing* (DTC) function, which estimates the index of the first time step at which the temperature sequence falls below the defined threshold. With this operator, the problem of recognising a discrete crossing event is transformed into a continuous and differentiable operation, which allows its integration into the gradient-based training of the neural network. The DTC function provides the predicted crossing time (expressed in minutes after normalisation of the time axis), both for the ground truth and for the model output.
A threshold-related loss component was then defined as the squared difference between the predicted and actual threshold crossing time.
The final custom loss function combined the conventional MSE term, which ensures accurate prediction of the entire temperature curve, with this additional threshold component weighted by an adjustable coefficient α.
For each temperature sequence $y = (y_1, \dots, y_T)$, the first time step at which the temperature falls below the threshold value τ is estimated. This is achieved by defining a masked variable:

$$m_t = \begin{cases} \mathrm{t}, & \text{if } y_t \leq \tau \\ \mathrm{T}, & \text{otherwise} \end{cases} \quad \text{for } \mathrm{t} = 1, \dots, \mathrm{T} \tag{1}$$

The crossing time is then the minimum value of the masked sequence:

$$\mathrm{t}_{\mathrm{cross}} = \min_{t=1,\dots,T} m_t \tag{2}$$

To match the assumed unit of time (minutes), the index is converted using a scaling factor κ:

$$\hat{\mathrm{t}} = \frac{\mathrm{t}_{\mathrm{cross}}}{\kappa} \tag{3}$$

with $\kappa = 6$, as each time step corresponds to 10 seconds. If the measured or predicted temperature curve does not fall below the threshold value of 15 °C, the crossing time is set to the maximum observation period of 70 minutes.
The loss term associated with crossing the threshold is defined as follows

$$\mathscr{L}_{\mathrm{threshold}} = \left(\hat{\mathrm{t}}_{\mathrm{pred}} - \hat{\mathrm{t}}_{\mathrm{true}}\right)^2 \tag{4}$$

Finally, the general custom-defined loss function combines the mean squared error (MSE) over the entire sequence with the threshold-related component, weighted by a factor α:

$$\mathscr{L} = \underbrace{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{i,t}^{\mathrm{true}} - y_{i,t}^{\mathrm{pred}}\right)^2}_{\text{MSE of predicted curve}} + \underbrace{\alpha\frac{1}{N}\sum_{i=1}^{N}\left(t_i^{\mathrm{true}} - t_i^{\mathrm{pred}}\right)^2}_{\text{time-to-threshold component}} \tag{5}$$

where N is the number of sequences in each batch and T the number of time steps per sequence; $y_{i,t}^{\text{true}}$ and $y_{i,t}^{\text{pred}}$ denote the true and predicted temperature at time step t for sequence i; t_i^{true} and t_i^{pred} are the true and predicted times (in minutes) when the temperature first falls below the threshold, and α (set to 0.6) is a weighting factor that controls the relative importance of the two components of the total loss. The value was determined through empirical optimisation by monitoring validation performance across different weight configurations and selecting the value that maximised the trade-off between sequence-level accuracy and threshold-crossing prediction.

This formulation ensures that the model learns to reproduce the overall temperature dynamics while explicitly aligning the prediction of the physiologically critical threshold-crossing time.

Once the model was built, trained, and tested against the experimental data, the performance was classified using accuracy, precision, and recall metrics. These metrics were applied to assess the model's ability to correctly classify whether or not the temperature crossed the 15 °C threshold in a physiologically plausible time frame.

The *Accuracy* measures the overall proportion of correct classifications, both positive and negative. It is defined as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{6}$$

Where TP denotes true positives, TN true negatives, FP false positives, and FN false negatives. In the context of our time-series prediction problem, these terms were defined based on whether the model correctly predicted the threshold-crossing event within a tolerance of ±10 minutes. This window was chosen to account for inter-subject variability observed in the experimental trials. Table 1 summarizes the specific criteria used for classification:

Classification	Definition
TP	The predicted curve crosses the threshold within ±10 minutes of the experimental crossing time.
TN	Neither the predicted nor the experimental curves cross the threshold.
FP	The predicted curve crosses the threshold (earlier than 10 minutes before test end), but the experimental curve does not.
FN	The predicted curve does not cross the threshold, but the experimental one does.

Table 1. Definition of classification outcomes for model performance evaluation.

To ensure that model evaluation was not biased by class imbalance, the *Precision* was also considered, and it was defined as the proportion of true positive predictions among all predicted positives:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

Finally, the *Recall* (also known as sensitivity) was computed to measure the proportion of actual positive cases correctly identified by the model:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

To complement accuracy-based evaluation, the F1-score was used as a summary metric combining precision and recall into a single indicator of event-detection performance. The F1-score is defined as the harmonic mean of precision and recall:

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

While the mathematical definitions of Accuracy, Precision, Recall, and F1-score follow standard classification theory, their interpretation in the present study requires careful contextualisation. The underlying task is not a discrete binary classification problem, but a continuous time-series regression problem in which a categorical outcome is derived from a temporal event (threshold crossing).

In this framework, *Accuracy* represents the proportion of cases in which the predicted threshold-crossing behaviour (or absence) agrees with the experimental observation under the imposed $\pm$10-minute tolerance window. However, accuracy is strongly influenced by the strictness of the temporal tolerance and by the distribution of event and non-event cases. In a setting where temporal precision is constrained by physiological variability, accuracy should therefore be interpreted as a measure of agreement under conservative temporal constraints rather than as a direct indicator of predictive adequacy in a traditional binary sense.

Precision and *Recall* provide complementary information. Precision reflects the reliability of positive predictions, i.e. how often a predicted threshold-crossing event corresponds to an actual crossing within the tolerance window. In practical terms, high precision reduces the likelihood of false alarms. Recall (sensitivity), on the other hand, quantifies the model's ability to detect true threshold-crossing events. In the context of thermal safety, recall is directly related to the capacity to correctly identify potentially dangerous cooling scenarios.

Importantly, the consequences of misclassification are asymmetric. Overestimation of the Duration of Safe Exposure (DSE) may lead to increased physiological risk, whereas underestimation results in conservative decisions without direct harm. For this reason, symmetric classification metrics alone do not fully capture the safety relevance of prediction errors. The F1-score, as the harmonic mean of precision and recall, provides a balanced summary of event-detection performance but does not account for the directional asymmetry of risk.

Therefore, beyond symmetric accuracy metrics, model performance was evaluated using a safety-oriented indicator, the *Unsafe Error Rate (UER)*, defined as the proportion of cases in which the predicted Duration of Safe Exposure exceeded the experimentally observed value by more than the $\pm$10-minute tolerance threshold.

The Unsafe Error Rate (UER) was defined as:

$$\mathrm{UER} = \frac{\left|\left\{ i \mid DSE_{\mathrm{pred}}^{(i)} - DSE_{\mathrm{obs}}^{(i)} > \Delta \right\}\right|}{N} \tag{10}$$

Where, in the considered case, Δ equals 10 minutes.

Since overestimation of DSE may expose users to increased risk of cold injury, UER provides a safety-oriented evaluation metric that complements symmetric accuracy measures.

Overall, the proposed architecture integrates both continuous temperature prediction and discrete event detection, providing a comprehensive framework for assessing thermal comfort and cold protection performance.

Mean Skin Temperature from JOS-3 Model

While the LSTM model was trained and validated on experimental data, its ultimate goal is to operate within a predictive framework capable of simulating untested conditions. The broader objective of this work is to develop a tool that can identify physiology-based parameters, such as the *Duration of Safe Exposure* (DSE), to assess the thermal performance of cold-protective footwear even during the prototyping phase, without requiring physical prototypes. This challenge can be effectively addressed through the use of human thermoregulation models.

In this study, the JOS-3 thermoregulation model was selected for its advanced capability to simulate detailed human thermal responses. JOS-3 is an evolution of the previously established JOS-2 model, itself based on the classical Stolwijk model[22]. It represents the human body through 85 discrete nodes, including a central blood node, 17 arterial and 17 venous nodes, 12 superficial venous nodes, 17 nuclear nodes, 2 muscle nodes, 2 fat nodes, and 17 skin nodes. These nodes correspond to 17 anatomical segments: head, neck, chest, back, pelvis, left (L) shoulder, L arm, L hand, right (R) shoulder, R arm, R hand, L thigh, L leg, L foot, R thigh, R leg, and R foot.

JOS-3 employs a finite difference method to simulate the human thermoregulatory response, predicting both global (e.g., mean skin temperature) and local body temperatures. Although the model includes the foot, it treats it as a uniform segment, without distinguishing between distal regions such as the toes areas that are particularly susceptible to cold injuries like frostbite.

In this work, JOS-3 simulations were configured to reproduce the same boundary conditions as the experimental trials, including subject characteristics (body mass, height) and clothing composition (detailed information is reported in the supplementary material). This enabled the evaluation of thermoregulatory behaviour under controlled variations of ambient temperature, activity level, and insulation. The JOS-3 outputs were then validated using selected datasets from the human trials, confirming the model's ability to reproduce realistic mean skin temperature trends under the tested scenarios.

Once the LSTM model had been successfully trained on real data (i.e., time series of mean skin and big toe temperatures), JOS-3 could be employed to generate synthetic skin temperature profiles for untested or transient scenarios. By feeding these synthetic time series into the trained LSTM, it becomes possible to estimate big toe temperature under a variety of virtual conditions, including new combinations of climate, activity, and footwear materials.

While wind and radiative heat transfer were not directly reproduced in the climatic chamber experiments, the JOS-3 thermoregulation model allows simulation of a broader range of environmental boundary conditions, including variations in air velocity and radiative loads.

This hybrid simulation-and-prediction workflow allows for the evaluation of thermal performance at the early design stage, providing a material- and physiology-based framework for footwear development before physical prototyping. Figure 2 illustrates the overall stepwise approach.

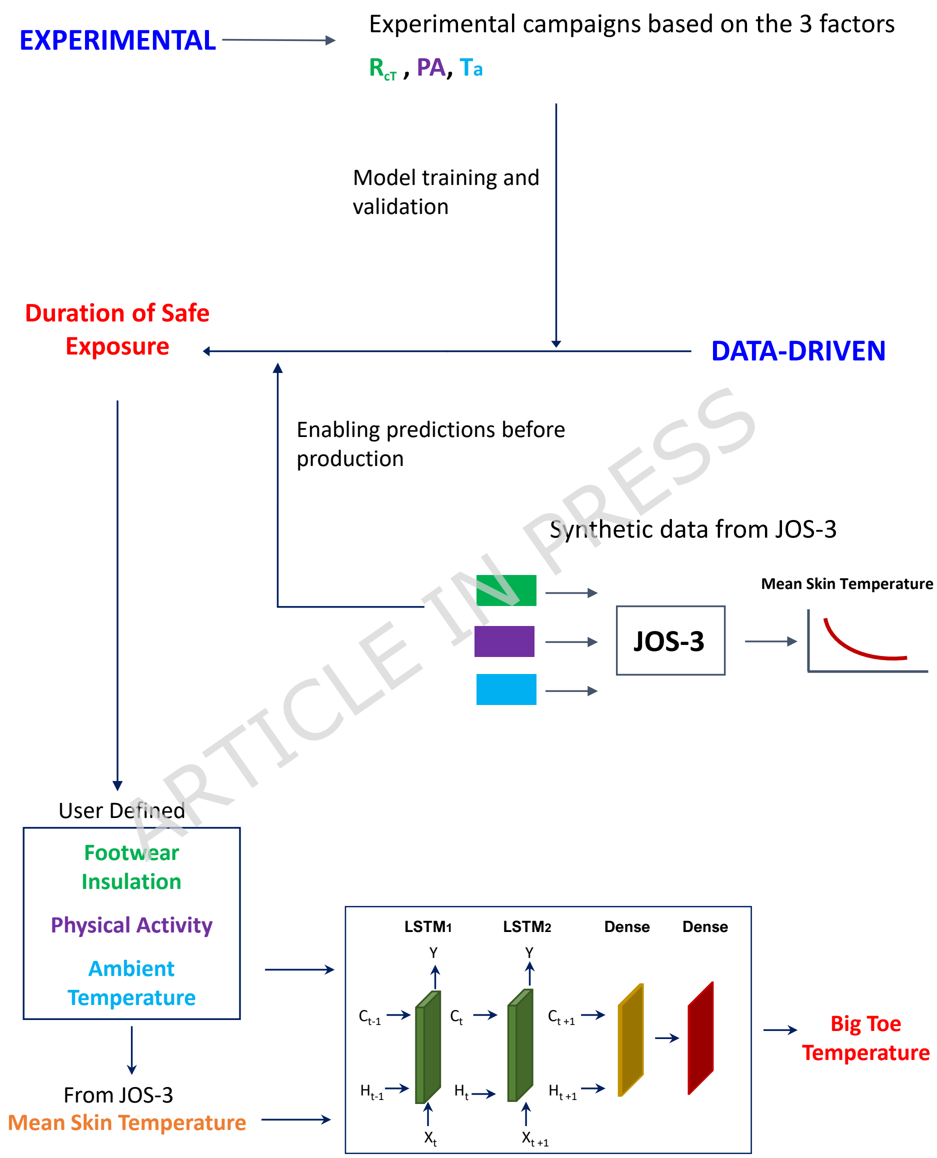

Figure 2. Workflow of the project from experimental to synthetic data.

Results & Discussion

LSTM predictions

The performance of the trained LSTM model was evaluated under the Leave-One-Subject-Out validation framework described in the Methods section. Results are presented in terms of both continuous curve-level agreement and event-based DSE estimation.
The model generally reproduced the temporal evolution of big toe temperature with high fidelity. Table 2 summarises the predicted threshold-crossing times across all tested conditions, along with the corresponding classification outcomes.

Table 2. Observed and predicted Duration of Safe Exposure (DSE) across all validation cases. $\Delta t = DSE_{\text{pred}} - DSE_{\text{obs}}$. A value of 70 min indicates that the 15°C threshold was not reached within the observation window. The asterisk (*) indicates cases where both curves crossed the threshold, but the predicted crossing time differed from the observed one by more than ±10 min; these were assigned to the closest event-based class.

Condition	DSE_{obs} (min)	DSE_{pred} (min)	Δt (min)	Class.
(A, 0, –10)	41.43	55.47	14.04	FN*
(A, 0, –17)	39.76	47.78	8.02	TP
(A, 1, –10)	69.83	70.00	0.17	TN
(A, 1, –17)	69.83	53.29	-16.54	FP*
(C, 2, –15)	59.81	63.32	4.84	TP
(B, 2, –20)	70.00	70.00	0.00	TN
(E, 1, –13)	70	54.13	-13.03	FP
(D, 1, –13)	30.91	57.30	26.39	FN*
(E, 0, –13)	49.12	49.95	0.83	TP
(D, 0, –13)	59.47	49.95	-9.52	TP

True positives (TP) were correctly identified across a variety of ambient temperatures and activity levels, confirming the model's ability to track threshold-crossing events within the observation window. The LSTM performed consistently across multiple combinations of footwear insulation (R_{cT}), ambient temperature (T_a), activity level, including non-monotonic cooling trends characterised by transient dynamics.
Overall, the network produced 4 true positives (TP), 2 true negatives (TN), 2 false positives (FP), and 2 false negatives (FN), corresponding to an accuracy of 60%, a precision of 67%, a recall of 67%, and an F1-score of 0.67. It is worth noting that reported accuracy values in small, structured datasets must be interpreted relative to task complexity and evaluation criteria. As observed in prior small-corpus comparisons, even state-of-the-art architectures may exhibit moderate absolute accuracy when class boundaries are strict or data variability is high[18]. In the present work, the tolerance window imposes a constraint comparable to the lower bound of experimentally observed inter-subject variability, making the evaluation criterion intentionally conservative.
These metrics must be interpreted in the context of the evaluation framework. Unlike standard binary classification problems, the present task involves continuous time-series regression, and categorical outcomes are derived by imposing a strict ±10-minute tolerance window on the predicted threshold-crossing time.
Under such a conservative temporal constraint, even moderate shifts in predicted crossing time, within the range of experimentally observed inter-subject variability, may be counted as misclassifications. Consequently, the reported accuracy does not reflect random prediction behaviour, but rather agreement under deliberately restrictive timing requirements. Precision indicates that two-thirds of predicted threshold-crossing events correspond to actual crossings within tolerance, while recall reflects the model's ability to identify more than half of true cooling events under this strict criterion. The F1-score summarises this balance between detection capability and false-alarm control.
Taken together with the safety-oriented analysis based on the Unsafe Error Rate (UER), these results suggest that the moderate symmetric metrics primarily reflect the stringency of the evaluation window and physiological variability, rather than a fundamental limitation of the predictive framework.
However, from a safety perspective, prediction errors are not symmetric in their practical meaning. Overestimation of the DSE may expose users to increased risk of cold injury, whereas underestimation results in conservative decisions without direct physiological harm. For this reason, model performance was additionally evaluated using the Unsafe Error Rate (UER).
Observing Table 2, 3 out of 10 cases met this criterion, corresponding to a UER of 20%. In the remaining cases, predictions were either within the tolerance window or conservative. This safety-oriented interpretation provides a more meaningful assessment of the model's practical reliability than symmetric accuracy metrics alone.

To provide a more comprehensive assessment of model performance, the classification results were complemented with continuous time-based error metrics. When both the experimental and predicted temperature curves crossed the threshold (i.e., valid crossing events), the mean absolute error of the crossing time was 10.78 min, indicating good temporal alignment between predicted and observed events. In censored cases, where the threshold was reached only in the experimental data and the predicted crossing time was set to the maximum observation duration of 70 min, the mean censored error was 24.64 min. These results indicate that, even when the threshold event was missed, the predicted temperature trajectories remained physiologically plausible and did not exhibit erratic behaviour.
In addition, a time-to-event tolerance analysis was performed to evaluate the sensitivity of the classification metrics to the temporal acceptance window used for defining correct threshold-crossing prediction. Figure 3 illustrates the variation of accuracy, precision, and recall as a function of the temporal tolerance τ applied to the predicted Duration of Safe Exposure (DSE). A prediction is classified as correct when the absolute difference between predicted and experimental crossing times satisfies $|DSE_{\text{pred}} - DSE_{\text{obs}}| \leq \tau$.
The choice of ± 10 minutes as the primary evaluation criterion was not arbitrary. Experimental data revealed that, within the same protocol condition, inter-subject variability in observed DSE ranged between approximately 10 and 18 minutes. To ensure a conservative and safety-oriented evaluation, the lower bound of this physiological variability (10 minutes) was selected as the tolerance threshold. This implies that the classification window is intentionally stricter than the full experimentally observed dispersion across participants.

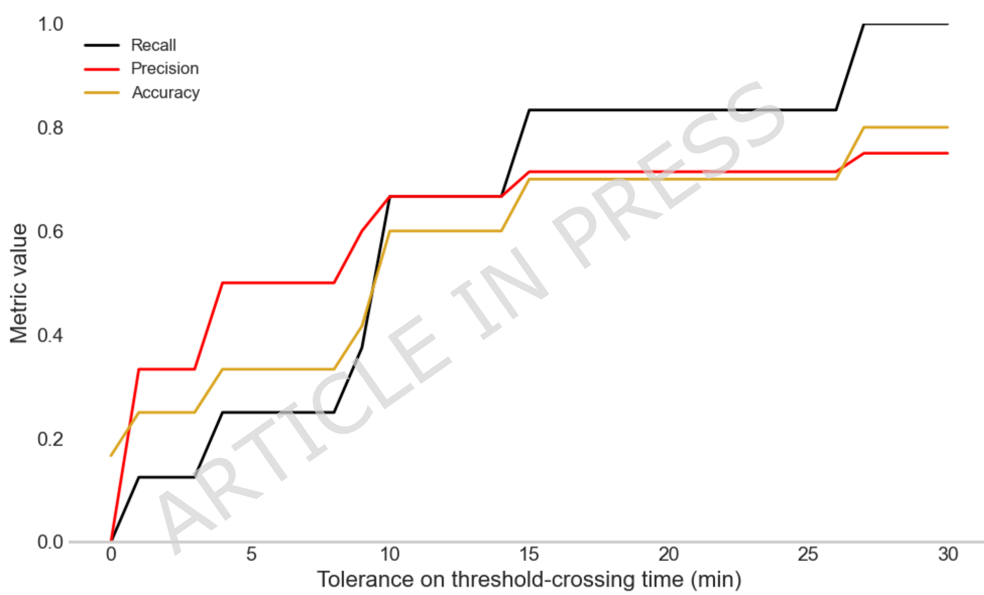

Figure 3. Time-to-event tolerance curve for the threshold-crossing prediction. Accuracy, precision, and recall are shown as a function of the temporal tolerance τ applied to the predicted crossing time.

As shown in Figure 3, relaxing the tolerance leads to a rapid increase in accuracy, precision, and recall. This behaviour indicates that many apparent misclassifications under the ± 10-minute criterion arise from moderate temporal shifts within the natural range of physiological variability rather than from fundamental prediction failure. Therefore, the moderate classification metrics observed under the strict ± 10-minute threshold should be interpreted as reflecting the conservativeness of the evaluation framework rather than an intrinsic limitation of the model's predictive capability.
In addition to accuracy, the F1-score was calculated as a complementary performance metric. By jointly considering precision and recall, the F1-score offers a more informative assessment of threshold-crossing detection than accuracy alone, especially in the presence of class imbalance and a limited number of test cases. The F1-score, with a value of 0.67, confirms that the network achieves a reasonable balance between reliable event detection and sensitivity to true cold-risk conditions.
Representative examples of accurate predictions are shown in Figure 4. The agreement between measured and predicted curves highlights the network's ability to reproduce both the overall trajectory and the critical threshold event.

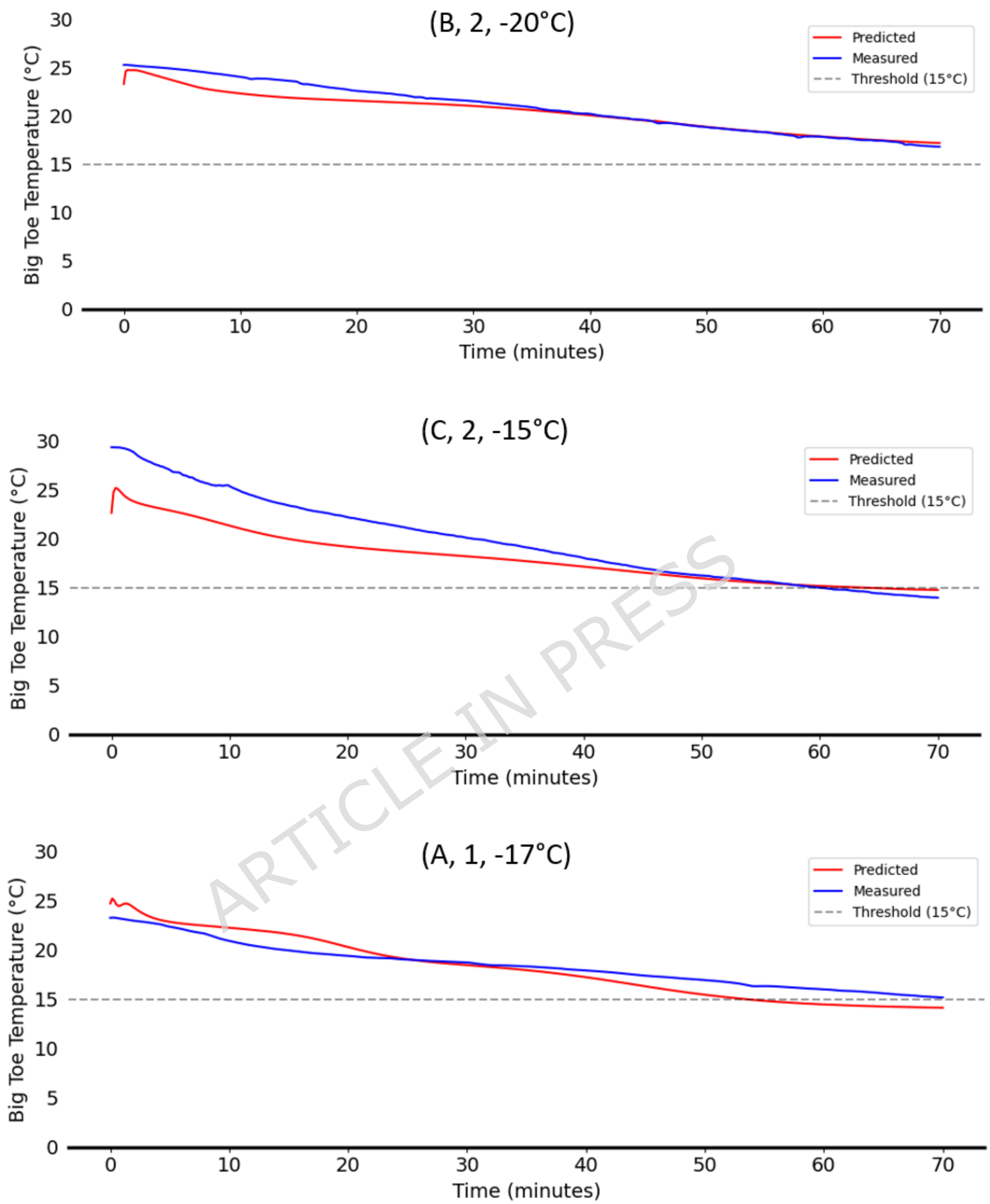

Figure 4. Representative examples of accurate LSTM predictions. The network successfully captures both the shape of the temperature curve and the timing of the threshold-crossing event.

As observed in the initial time steps, a short transient increase in predicted temperature may appear before stabilisation, corresponding to the dynamic correction of the model's initial state.

Nevertheless, certain limitations remain. In some cases, such as conditions (D, 1, –13) (Figure 6), the network failed to accurately estimate the threshold-crossing time, resulting in false negatives (FN) or significant deviations between predicted and observed timings.

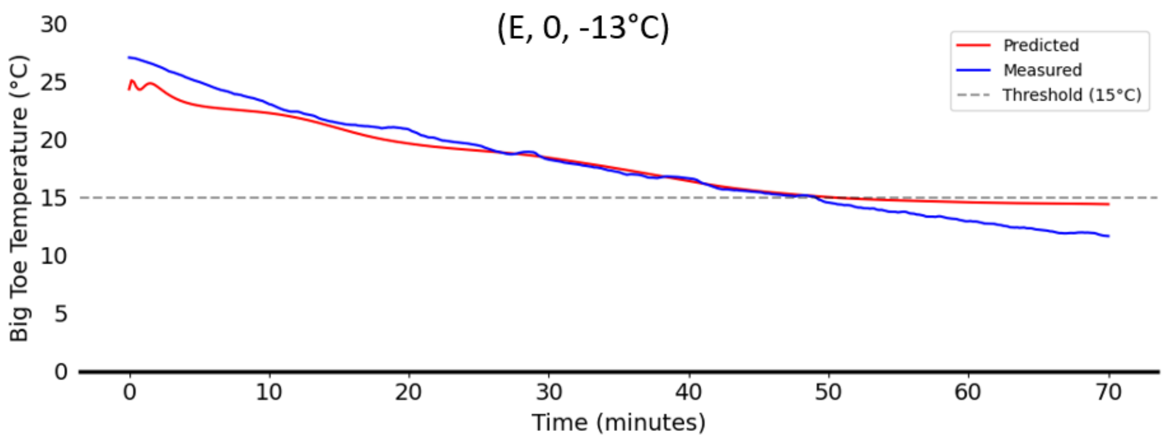

Figure 5. Representative example of accurate LSTM predictions. The network successfully captures both the shape of the temperature curve and the timing of the threshold-crossing event.

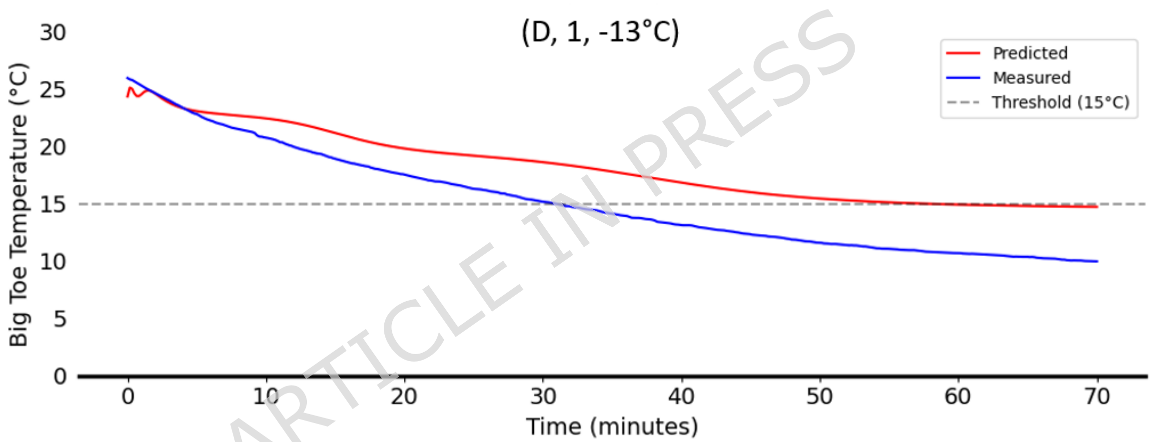

Figure 6. Example of inaccurate prediction performance. The threshold crossing was not identified, although the overall temperature trajectory was correctly reproduced.

For condition (C, 2, –15), the predicted curve did not fully overlap with the measured data but still correctly identified the absence of a threshold event, corresponding to a true negative outcome. Conversely, for condition (D, 1, –13), both the shape of the curve and the threshold timing were misestimated, resulting in a false negative classification. This suggests that, while the network captured the general cooling trend, it was less precise in locating the threshold event.

Two main factors account for the degradation of performance observed in these cases. First, reduced accuracy occurs under conditions that were under-represented in the training dataset.

Because subjects were deterministically selected for validation in each protocol, the resulting validation set included individuals with heterogeneous physiological responses. In several cases, the network accurately reproduced both the overall temperature trend and the threshold-crossing time, suggesting that prediction errors were primarily associated with inter-individual variability rather than systematic model bias.

Importantly, no subject overlap occurred between training and validation sets, ensuring that the observed variability reflects genuine subject-specific responses rather than information leakage.

From a practical perspective, the joint interpretation of accuracy, precision, and recall provides a clearer understanding of model reliability. Accuracy quantifies overall consistency across conditions; precision reflects confidence in predicted threshold events (avoiding false alarms); and recall measures the ability to detect all true events (minimising missed cold exposures). Although the absolute values of these metrics remain moderate, their combined interpretation highlights a meaningful trade-off between sensitivity and safety, ensuring that the model remains conservative in predicting cold-risk scenarios.

Temperature prediction based on JOS-3

Once the LSTM model has been trained, its main purpose was to predict the DSE, defined as the time at which the temperature of the big toe reaches the critical threshold of 15°C. For this purpose, the network relies on the time series of mean skin temperature provided by the JOS-3 thermoregulation model.

The JOS-3 model generates the skin temperature evolution based on several key inputs. These include environmental conditions such as air temperature, mean radiant temperature, relative humidity, and air velocity, which can be adjusted to represent different scenarios of cold exposure. The model also requires information about physical activity and posture, which influence metabolic heat production and heat distribution in the body. Another essential input is the clothing insulation of the different body segments. In our study, this is defined on the basis of the standardised clothing used in the experimental campaigns. Details of the clothing ensemble are shown in the supplementary materials. However, the insulation provided by the footwear, as well as the environmental conditions and physical activity, can be adjusted by the user depending on the scenario under investigation.

Based on these inputs, JOS-3 calculates the temporal evolution of the mean skin temperature, which serves as input for the LSTM network. The network then predicts the corresponding temperature curve of the big toe and estimates the time at which the critical threshold is reached, which is a quantitative measure of the cold risk. Before extending the workflow to completely untested virtual scenarios, the big toe temperature predictions derived from synthetic mean skin temperature series generated by the JOS-3 model were first compared with corresponding experimental cases within the range of the investigated protocols. This comparison aimed to verify the internal consistency of the hybrid framework when replacing experimentally measured mean skin temperature with physiologically simulated inputs.

An example of this comparison is shown in Figure 7. The temperature profile predicted from the JOS-3 synthetic input closely follows the experimental trend within the tested domain. In the example shown, the experimental Duration of Safe Exposure (DSE) was 46.3 minutes, while the corresponding predicted value was 45.11 minutes.

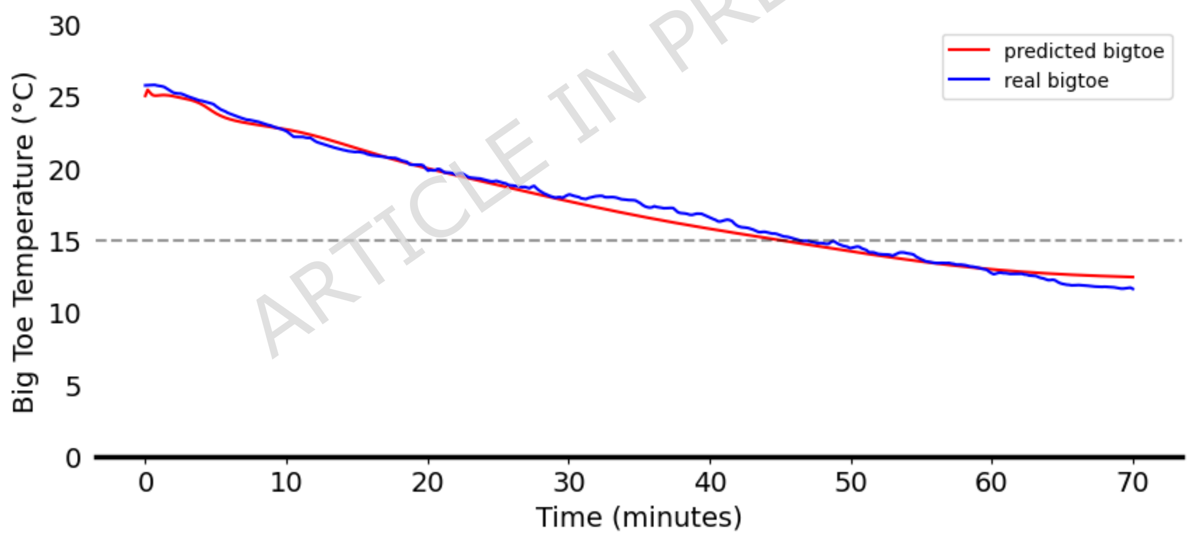

Figure 7. Comparison between experimental and LSTM-predicted big toe temperature curves using synthetic mean skin temperature data generated by JOS-3.

This agreement indicates that the coupled JOS-3–LSTM framework produces physiologically coherent estimates under conditions comparable to those experimentally investigated.

The four cases shown in Figures 8–11 illustrate the ability of the LSTM model to consistently capture both thermophysiological trends and the influence of operational conditions on threshold-crossing behaviour.

In Figure 8, where the subject is at rest with high boot insulation under severe ambient temperature, the predicted time to threshold (57 min) aligns with expected physical behaviour. In Figure 9, insulation and activity remain unchanged, but the ambient temperature is increased to –10°C; in this case, the predicted trajectory does not reach the critical threshold, demonstrating the model's sensitivity to environmental variation.

Figure 10 presents a scenario with high activity and reduced insulation (R_{cT}=1.1) at –10°C, resulting in a longer predicted time to threshold (64 min), indicating that metabolic heat production can partially compensate for lower insulation. Finally,

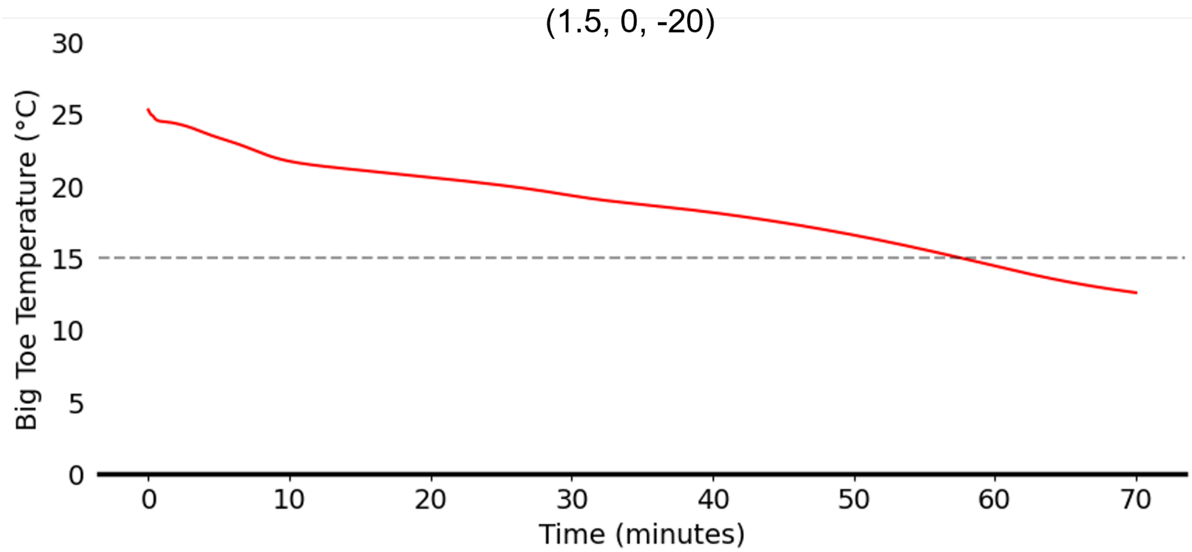

Figure 8. LSTM prediction under stillness with high boot insulation and severe ambient temperature. The predicted time to reach the 15°C threshold is 57 minutes.

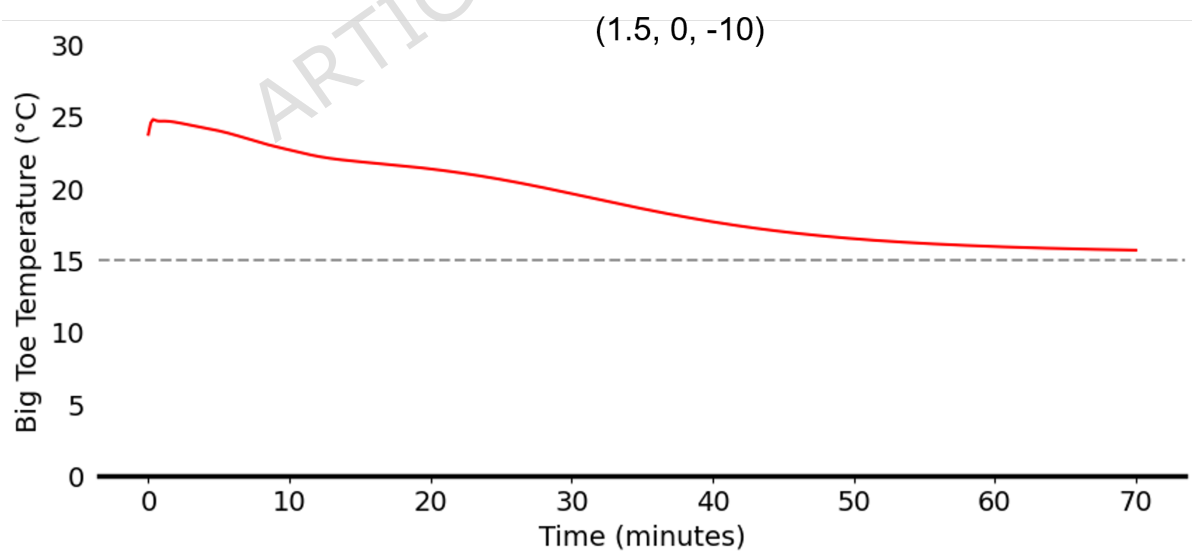

Figure 9. LSTM prediction under stillness with the same insulation as in Figure 8, but at an ambient temperature of –10°C. The predicted curve does not reach the safety threshold within the observation window.

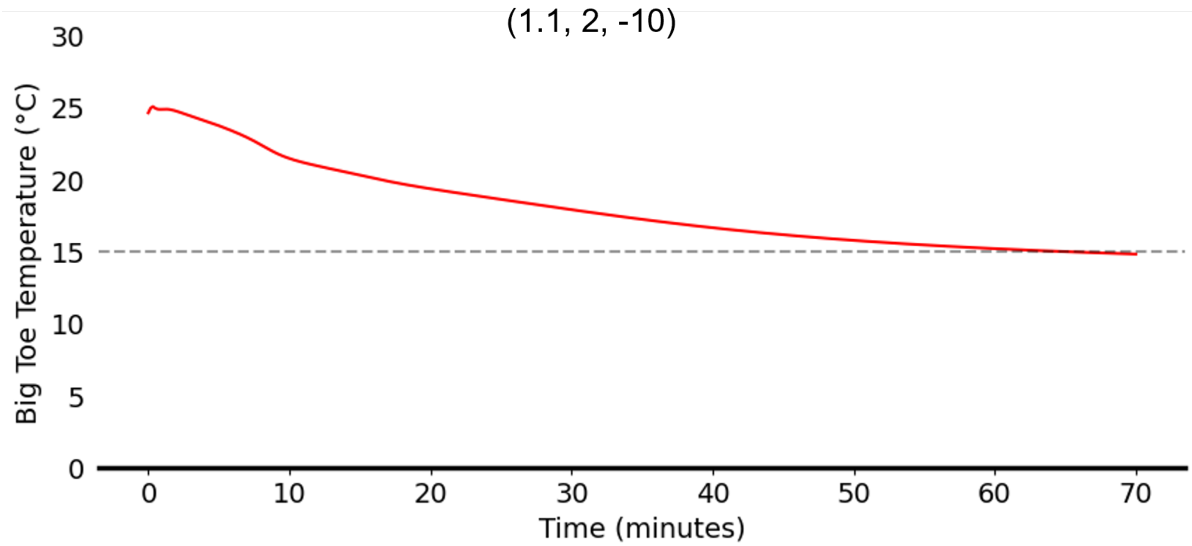

Figure 10. LSTM prediction under high physical activity with reduced boot insulation ($R_{cT} = 1.1$) at –10°C. The predicted time to threshold is 64 minutes.

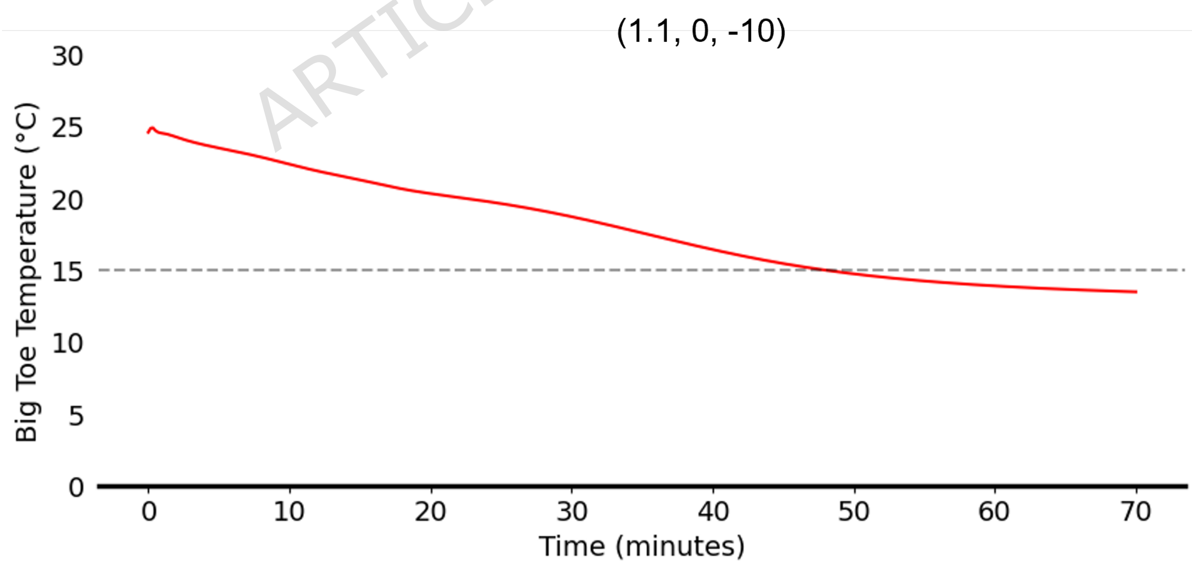

Figure 11. LSTM prediction under stillness with low boot insulation at –10°C. The predicted time to threshold is 47 minutes, highlighting the influence of insulation on toe cooling.

Figure 11 shows stillness under low insulation at –10°C, with a reduced time to threshold (47 min), confirming the dominant role of insulation in peripheral cooling dynamics.
Overall, these examples demonstrate that the model responds coherently to variations in insulation, ambient temperature, and activity level, supporting its suitability for safety-oriented assessment of peripheral cooling risk.

Conclusions

This study aimed to introduce a more scientifically grounded and safety-oriented approach for evaluating the thermal performance of mountaineering boots, which are commonly classified as sports goods and are therefore not subject to specific physiological safety standards. In high-altitude and extreme cold environments, inadequate thermal protection of the feet may lead to severe cold-related injuries. For this reason, a quantitative framework capable of integrating material performance with human thermophysiological response is essential to support design decisions beyond empirical practice. For this reason, developing quantitative approaches that integrate material performance with human thermophysiological response represents an important step beyond empirical practice.
The proposed framework is centred on the concept of *Duration of Safe Exposure* (DSE), defined as the time (in minutes) during which a subject can remain under a given combination of footwear insulation (R_{cT}), ambient temperature (T_a), and physical activity level (PA) before the big toe skin temperature reaches the conservative critical threshold of 15°C. The big toe was selected as a benchmark indicator based on experimental evidence demonstrating its pronounced vasoconstrictive response and its tendency to reach the lowest temperatures among foot regions during cold exposure.
The methodological approach combined controlled experimental campaigns with a data-driven modelling strategy. The experimental component enabled systematic investigation of the interaction between footwear insulation, environmental conditions, and activity level, while accounting for inter-subject variability within the investigated population (young male participants representative of the experimental sample). The resulting dataset provided the basis for training a Long Short-Term Memory (LSTM) network to predict the temporal evolution of big toe temperature as a function of mean skin temperature, R_{cT}, T_a, and PA.
After training on experimental data, the predictive framework was extended through integration with the JOS-3 thermoregulation model. In this hybrid configuration, JOS-3 generates physiologically consistent mean skin temperature inputs corresponding to specified operational conditions, allowing the LSTM to estimate big toe temperature and DSE under controlled or simulated scenarios without requiring additional in vivo measurements.
Model performance was evaluated using both event-based classification metrics and safety-oriented indicators. Event classification (true positives, false negatives, etc.) was defined based on the model's ability to predict threshold-crossing events within a ± 10 minute tolerance window. In addition to conventional metrics (accuracy, precision, recall), performance was assessed using the Unsafe Error Rate (UER), defined as the proportion of cases in which the predicted DSE exceeded the experimentally observed value by more than the tolerance threshold. This distinction is particularly relevant in a safety-oriented context, as overestimation of DSE may expose users to increased physiological risk, whereas underestimation results in conservative decisions. In the present validation set, most predictions were either within tolerance or conservative, while the UER remained limited.
However, these results should be interpreted with caution. The limited dataset size, the controlled experimental conditions, and the inherent variability of human thermophysiological responses may affect the generalisability of the findings. In addition, the JOS-3 thermoregulation model, while providing a physiologically grounded framework, represents an approximation of human thermal regulation and may not fully capture inter-individual variability or localised responses such as those occurring in distal extremities. In addition, a limitation of this study is that the experimental dataset includes only young male participants within a restricted range of age and body composition. Therefore, the proposed framework cannot be assumed to apply to other populations, including female users, older individuals, or subjects with different physiological or circulatory characteristics.
Future developments may include expanding the dataset to incorporate broader demographic variability, refining the loss formulation to further emphasise safety-oriented constraints, and exploring alternative sequence-modelling architectures (e.g., attention-based or transformer-inspired models) to enhance robustness under highly dynamic cooling conditions.
Overall, this work should be interpreted as a preliminary and population-specific contribution toward the development of physiologically informed, data-driven approaches for the analysis of peripheral cooling in cold environments. Further validation is required before extending the proposed framework to broader populations or real-world scenarios.

Supplementary Material

Supplementary information, including details of the experimental setup, additional figures and extended results, is available in the *Supplementary Material*.

Data Availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request. The code and full training and evaluation pipeline are publicly available on GitHub (`https://github.com/comfolab/toe-temperature-dse-lstm`) and archived on Zenodo (doi: `10.5281/zenodo.18851663`).

References

1. Kuklane, K. *Footwear for cold environments Thermal properties, performance and testing.* (Arbetslivsinstitutet, 1999).
2. West, A. M. *et al.* A thermal foot manikin as a tool for footwear evaluation and development. *Proc. Inst. Mech. Eng. Part P: J. Sports Eng. Technol.* **237**, 34–46, DOI: 10.1177/1754337120952229 (2023).
3. Bianca, E., Buffo, A., Vanni, M. & Ferri, A. Optimizing thermal insulation in footwear: A numerical simulation approach with cad modeling. *Int. J. Therm. Sci.* **218**, 110110, DOI: 10.1016/j.ijthermalsci.2025.110110 (2025).
4. E., B., F., D., A., F. & G., H. Heat and mass transfer in footwear: Exploring the moisture evaporation and condensation cycle. *J. Ind. Textiles* DOI: 10.1177/15280837251325783 (2025).
5. Kuklane, K. Protection of feet in cold exposure. *Ind. Heal.* **47**, 242–253, DOI: 10.2486/indhealth.47.242 (2009).
6. Bianca, E., Dotti, F., Orrico, F. & Ferri, A. Thermoregulation of feet in cold environments: A study on alpinism. *Appl. Ergonomics* **116**, 104205, DOI: 10.1016/j.apergo.2023.104205 (2024).
7. Parsons, K. *Human thermal environments: the effects of hot, moderate, and cold environments on human health, comfort and performance* (CRC press, 2007).
8. Fanger, P. O. Assessment of man's thermal comfort in practice. *Occup. Environ. Medicine* **30**, 313–324, DOI: 10.1136/oem.30.4.313 (1973).
9. Naiqian, G., Yuxin, G. & Xuelian, S. Global temperature forecast based on arima model. In *4th International Conference on Communication and Information Systems (ICCIS)*, DOI: 10.1109/ICCIS49662.2019.00026 (2019).
10. Saa, E. D. & Ranathunga, L. Comparison between arima and deep learning models for temperature forecasting. arXiv preprint arXiv:2011.04452, DOI: 10.48550/arXiv.2011.04452 (2020).
11. Elshewey, A. M. *et al.* A novel wd-sarimax model for temperature forecasting using daily delhi climate dataset. *Sustainability* **15**, 1–15, DOI: 10.3390/su15010757 (2023).
12. Nketiah, E. A., Agyeman, M. G. & Oppong, A. K. Recurrent neural network modeling of multivariate time series and its application in temperature forecasting. *PLOS ONE* **18**, e0285713, DOI: 10.1371/journal.pone.0285713 (2023).
13. Liang, Z., Wang, X., Li, H. & Li, J. Body temperature prediction with recurrent neural network and its variants. In *2021 11th International Conference on Intelligent Control and Information Processing (ICICIP)*, 281–286, DOI: 10.1109/ICICIP53737.20 (IEEE, 2021).
14. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780, DOI: 10.1162/neco.1997.9.8.1735 (1997).
15. Acikmese, Y. & Alptekin, S. E. Prediction of stress levels with lstm and passive mobile sensors. *Procedia Comput. Sci.* **159**, 658–667, DOI: 10.1016/j.procs.2019.09.085 (2019).
16. Cortes, V. M. P., Chatterjee, A. & Khovalyg, D. Dynamic personalized human body energy expenditure: Prediction using time series forecasting lstm models. *Biomed. Signal Process. Control.* **87**, 105381, DOI: 10.1016/j.bspc.2024.105381 (2024).
17. Morishima, S., Fujimoto, Y., Yamaguchi, T. & Saito, K. Human body skin temperature prediction based on machine learning. *Artif. Life Robotics* **26**, 103–108, DOI: 10.1007/s10015-020-00617-5 (2021).
18. Ezen-Can, A. A comparison of lstm and bert for small corpus, DOI: 10.48550/arXiv.2009.05451 (2020). 2009.05451.
19. Al-Selwi, S. M. *et al.* Rnn-lstm: From applications to modeling techniques andbeyond—systematic review. *J. King Saud Univ. - Comput. Inf. Sci.* DOI: 10.1016/j.jksuci.2024.102068 (2024).
20. Liu, K., Liu, Y., Wang, C., Wang, J. & Chen, Y. A machine learning approach to predict outdoor thermal comfort using local skin temperatures. *Sustain. Cities Soc.* **59**, 102216, DOI: 10.1016/j.scs.2020.102216 (2020).
21. Han, X., Zhang, L., Xu, S. & Li, Y. Forecasting human core and skin temperatures: A long-term series approach. *Big Data Cogn. Comput.* **8**, 197, DOI: 10.3390/bdcc8120197 (2024).

22. Takahashi, Y. *et al.* Thermoregulation model jos-3 with new open source code. *Energy Build.* **231**, 110575, DOI: 10.1016/j.enbuild.2020.110575 (2021).

23. Pham, T. D. Time–frequency time–space lstm for robust classification of physiological signals. *Sci. reports* **11**, 6936 (2021).

24. Saeedinia, S. A., Jahed-Motlagh, M. R., Tafakhori, A. & Kasabov, N. K. Diagnostic biomarker discovery from brain eeg data using lstm, reservoir-snn, and neucube methods in a pilot study comparing epilepsy and migraine. *Sci. Reports* **14**, 10667 (2024).

25. Liao, Y. *et al.* Emotion recognition with multiple physiological parameters based on ensemble learning. *Sci. Reports* **15**, 19869 (2025).

26. Lipton, Z. C., Berkowitz, J. & Elkan, C. A critical review of recurrent neural networks for sequence learning, DOI: 10.48550/arXiv.1506.00019 (2015).

27. Mienye, I. D., Swart, T. G. & Obaido, G. Recurrent neural networks: A comprehensive review of architectures, variants, and applications. *Information* **15**, 517, DOI: 10.3390/info15090517 (2024).

28. Jamieson, K. & Talwalkar, A. Non-stochastic best arm identification and hyperparameter optimization. In Gretton, A. & Robert, C. C. (eds.) *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, vol. 51 of *Proceedings of Machine Learning Research*, 240–248 (PMLR, Cadiz, Spain, 2016).

29. Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A. & Talwalkar, A. Hyperband: A novel bandit-based approach to hyperparameter optimization. *J. Mach. Learn. Res.* **18**, 1–52 (2018).

30. Kemeth, F. P. *et al.* Initializing lstm internal states via manifold learning. *J. Nonlinear Sci.* **231**, DOI: 10.48550/arXiv.2104.13101 (2021).

31. Karim, F., Majumdar, S., Darabi, H. & Chen, S. Lstm fully convolutional networks for time series classification. *IEEE Access* **6**, 1662–1669, DOI: 10.1109/ACCESS.2017.2779939 (2017).

Author contributions statement

E.B., A.F., conceived the experiment(s), E.B. drafted the manuscript, conducted the experiments, and performed data analysis, A.M. revised the manuscript A.F., and G.B. designed the experiments and revised the manuscript, A.F. provided the funding. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Funding Declaration

This project took place at the end of the doctoral program, funded by PNRR Mission 4, Component 2 “Dalla Ricerca all’Impresa” – Investimento 3.3 “Introduzione di dottorati innovativi che rispondono ai fabbisogni di innovazione delle imprese e promuovono l’assunzione dei ricercatori dalle imprese”, through Ministerial Decree no. 352 of 9 April 2022.