AutoTherm: A Dataset and Benchmark for Thermal Comfort Estimation Indoors and in Vehicles

MARK COLLEY*, Institute of Media Informatics, Ulm University, Germany

SEBASTIAN HARTWIG*, Institute of Media Informatics, Ulm University, Germany

ALBIN ZEQIRI, Institute of Media Informatics, Ulm University, Germany

TIMO ROPINSKI, Institute of Media Informatics, Ulm University, Germany

ENRICO RUKZIO, Institute of Media Informatics, Ulm University, Germany
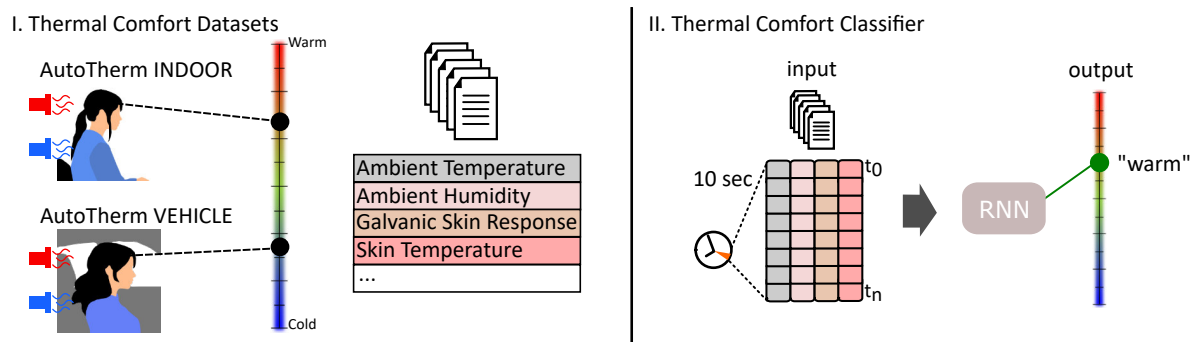
Fig. 1. We propose the AUTOTHERM DATASETS. (I.) We collected human-annotated sensor measurements during two thermal state studies (INDOOR and VEHICLE), enabling the training of neural thermal state classifiers (II.) to identify human thermal state changes from sequential sensor measurements.

Thermal comfort inside buildings is a well-studied field where human judgment for thermal comfort is collected and may be used for automatic thermal comfort estimation. However, indoor scenarios are rather static in terms of thermal state changes and, thus, cannot be applied to dynamic conditions, e.g., inside a vehicle. In this work, we present our findings of a gap between building and in-vehicle scenarios regarding thermal comfort estimation. We provide evidence by comparing deep neural classifiers for thermal comfort estimation for indoor and in-vehicle conditions. Further, we introduce a temporal dataset for indoor predictions incorporating 31 input signals and self-labeled user ratings by 18 subjects in a self-built climatic chamber. For in-vehicle scenarios, we acquired a second dataset featuring human judgments from 20 subjects in a BMW 3 Series. Our experimental results indicate superior performance for estimations from time series data over single vector input. Leveraging modern machine learning architectures enables us to recognize human thermal comfort states and estimate future states automatically. We provide details on training a recurrent network-based classifier and perform an initial performance

benchmark of the proposed dataset. Ultimately, we compare our collected dataset to publicly available thermal comfort datasets.

CCS Concepts: • **Computing methodologies** → *Machine learning*; • **Human-centered computing** → **Ubiquitous and mobile computing**.

Additional Key Words and Phrases: Machine Learning; vehicles; state recognition; dataset.

1 INTRODUCTION

The inclusion of sensors such as cameras, radars, thermometers, or lidars in today's manually driven vehicles and, most likely, in future (automated) vehicles [39] allows for novel insights into users' states and intentions. The same can be said of modern smart home systems, which automate indoor climate conditions and optimize energy efficiency. To improve user experience, there is a need for accurate and reliable recognition, interpretation, and understanding of current user states [68], as this ensures the execution of adjustments based on user preferences [85] and alleviates the user of burdensome tasks such as adjusting the temperature. Additionally, manually adjusting the temperature can lead to an overshoot [72], requiring additional interactions. Using in-vehicle sensors and machine-learning methods, current vehicles already recognize some driver states, such as level of drowsiness [58] or fatigue [96]. This is mainly a safety measure to avert potentially dangerous driving behavior. Yet, seeing as driving-related tasks will increasingly become irrelevant the higher the level of automation, recognition of other states such as emotional state [8] and intention [66] become increasingly relevant. There are various methods to determine the passenger's state, including machine-learning-based methods. Depending on whether the state recognition task is formulated as a supervised or unsupervised learning problem, labeled data is required [78]. Learning from human judgments to replicate human behavior or to mirror human preferences is a common practice [32, 33, 52, 98, 99], also indoors [42]. However, there are only a few labeled and publicly available datasets for the automotive state recognition use case, such as *drive&act* by Martin et al. [60] or the *VEmotion dataset* by Bethge et al. [8]. While there are existing datasets for indoor thermal comfort, they cannot be applied to the field of in-vehicle environments.

Already today, one relevant aspect for passengers inside vehicles is the perceived level of thermal comfort [41, 67, 77]. In building ergonomics, thermal comfort, and its influencing factors have been part of numerous research studies that resulted in different models for thermal comfort estimation, such as the commonly known predicted mean vote (PMV) index by Fanger et al. [24]. However, while thermal comfort is dependent on the exposure history [92], these works focus on a single data point while we also incorporate temporal data. Additionally, there are significant differences between indoor settings and vehicles and the required data. First, indoors, and especially offices, occupants do not have direct control access to the temperature. Second, this temperature changes only slowly. Therefore, the available datasets do not include temporal data but only provide singular data points. Additionally, the data points are intended to assess current comfort without the explicit goal of adapting the temperature. The automotive use case, on the other hand, is defined by possibly fast-changing temperatures (e.g., due to opened windows also by other users), limited variety and action performed by the users, and the current expectation of having the personally optimal temperature due to relatively easy access. Additionally, in the context of automotive state recognition systems, thermal comfort state estimation (i.e., *warm, comfortable, cold, ...*) has thus far not been explored. Consequently, no machine-learning model or dataset exists for in-vehicle thermal comfort state estimation. The impact of inputs from in-vehicle sensory units on prediction performance is also unknown. The automotive domain is particularly characterized by providing an environment known by

manufacturers (regarding size, capabilities, limitations, and heating possibilities), providing (already today) a plethora of sensor data, and restricting the user's movements and actions. Therefore, a specialized dataset is required.

Contribution Statement: In this work, we created a (1) **temporal**, labeled multi-modal dataset featuring **31** input signals (age, gender, weight, height, body fat, body core temperature, activity level, time since last meal, tiredness, clothing level, radiation temperature, emotion, RGB frame, ten body pose key points, heart rate, wrist temperature, galvanic skin response, ambient temperature, relative humidity) relevant for the task of thermal comfort state estimation, which has not yet been addressed in the context of automotive state recognition. (2) The implemented and employed logging can be used as a data-gathering blueprint for future state recognition research. (3) Thirdly, we present a machine-learning-based approach for both indoor and in-vehicle thermal comfort state recognition that takes advantage of different feature combinations to explore predictive performance and the impact of individual input modalities. (4) Fourthly, we report a feature combination study with different network architectures for thermal comfort state estimation and forecasting. (5) Finally, we evaluate our trained classifiers on existing thermal comfort datasets and report superior performance for models trained on our AUTOTHERM DATASET.

Within the remainder of this paper, we will first discuss the work related to our approach in Section 2. Second, we're providing details on our study setup, where we collect human judgments for thermal comfort in Section 3, before describing the data we collected for indoor scenarios, in Section 4 and for in-vehicle scenarios in Section 5. We then compare our dataset to existing thermal comfort datasets, in Section 6 followed by Section 7 where we conduct and present results of several thermal comfort estimation experiments. Finally, we discuss our findings and address limitations of our approach in Section 8, share our dataset and code in Section 9 and conclude in Section 10.

2 RELATED WORK

This work builds on previous work on indoor thermal comfort, factors influencing thermal comfort, and state recognition in general.

2.1 Thermal Comfort

Various factors influence one's perceived comfort level in an indoor environment, such as visual, acoustic, and environmental conditions [28]. Thermal comfort describes the level of satisfaction with one's surroundings based on thermal influences [48] and has been extensively researched in the field of building ergonomics. The PMV index [24] referenced in the ISO 7730:2006-05 and American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) 55-2020 standards [1, 48] estimates the perceived level of thermal comfort for a large group of people on a thermal sensation scale with the seven items *Cold, Cool, Slightly Cool, Comfortable, Slightly Warm, Warm, and Hot* and is based on empirical thermal comfort studies, from which an equation for thermal comfort calculation based on six main influencing factors was derived: Metabolic Rate, Clothing Insulation, Mean Radiation Temperature, Ambient Temperature, Relative Humidity, and Air Velocity.

2.2 Influences on Thermal Comfort

Research on building ergonomics indicates that environmental [71] or physiological (human thermoregulation [29]) factors and their interplay should be considered. Thermal sensation is mostly felt due to thermoreceptors on one's skin and muscles [89]. Accordingly, skin temperature has been used as an indicator of thermal perception changes (e.g., [76, 83]). When used in conjunction with body core temperature, Frank et al. [27] found that skin and body core temperature contribute similarly to thermal comfort. Sim et al. [83] demonstrate the estimation of thermal comfort based on measuring different sites around the wrist and fingertips, and Ramanathan

[76] proposed an approach for the estimation of the mean skin temperature, computed by averaging the skin temperature of different locations across the body. Another work proposed an estimation of thermal comfort from multiple physiological input streams, such as heart rate, skin temperature, or electrodermal activity [104]. Additionally, it was established that there is a range in which occupants feel thermally comfortable (thermal comfort zone) [14, 15], rather than a single temperature. This zone is influenced by the dynamics of temperature change and the direction of the change [16]. Apart from estimation based on physiological input data, it was reported that there is an influence of gender [12, 47], age [20, 31], and emotion [95] on thermal comfort perception. For instance, female occupants seem more sensitive to thermal changes that deviate from their optimal state and thus feel too cold or too hot more frequently [46]. In the elderly, deterioration of skin receptors is assumed to cause reduced thermal perception ability, especially in the limbs [31]. As for emotion, Wang and Liu [95] concluded that negative emotions have an unfavorable effect on thermal comfort. However, overall, emotions only affect thermal comfort perception during light activities such as sitting or standing.

2.3 Thermal Comfort Label Scale

Fanger's PMV index [24] is calculated in the interval [-3, 3]. Accordingly, the seven different thermal comfort states *Cold, Cool, Slightly Cool, Comfortable, Slightly Warm, Warm, and Hot* represent ranges in the defined interval rather than integers. For instance, according to the ASHRAE standard [48], the state *Comfortable* is established in the interval [-0.5, 0.5]. This more granular approach allows for reducing the initial seven-point scale to a three-point scale, where the new reduced states can be denoted as *Too Cold, Comfortable, Too Warm*. Due to its simplicity and standardized theoretical basis, the seven-point thermal sensation scale was adopted as the label set.

2.4 State Recognition Systems

Machine-learning-based state recognition was used for cognitive load detection [9, 103], driver stress detection [64], situation awareness prediction [49], affective computing [65, 107], vehicle assistance systems [58, 63, 87], and even pain recognition [94]. They mainly differ in terms of chosen input spaces, modalities, and employed learning methods (supervised or unsupervised learning). Even though systems trained on data from a single input space or single modality can perform satisfyingly, a closer resemblance to human perception can be achieved by incorporating additional modalities [5]. Therefore, multi-modal input data for the training of state recognition systems is beneficial.

Early works in multi-modal in-vehicle state recognition incorporated various vehicle parameters (pedal position, steering), environmental information (local and global vehicle position), and driving performance attributes (speed) as input signals [7, 35] to improve intention recognition in safety systems. Likewise, human action recognition experienced advances by incorporating multi-modal input signals together with feature fusion and co-learning methods [86]. Zhang et al. [105] demonstrated in-vehicle action recognition with their proposed interwoven CNN approach and a self-recorded dataset. Additionally, within the field of affective computing, the recognition of emotional states was explored. This was done as multi-modal interfaces, not just in the automotive context, could benefit from the ability to recognize and interpret one's emotions. With VEmotion, Bethge et al. [8] proposed a novel way of estimating the emotional state of drivers in real-time using driving context information such as weather, traffic, road, and car trajectory data. They demonstrated that states such as emotions can be predicted by using mainly contextual information, which is more readily available in vehicles. Generally, incorporation of insights from areas, such as emotion recognition [65, 107], cognitive load estimation [9] or next interaction method prediction [97] can contribute to creating a more holistic understanding of users' needs in automotive state recognition systems.

Nevertheless, only a few publicly available datasets can be used for further state recognition research. Therefore, developing new approaches for recognizing certain states almost always entails the acquisition of a new dataset, thereby significantly slowing down development speed while increasing task complexity. Additionally, methods for in-vehicle thermal comfort estimation from multi-modal data have not been explored, although thermal sensation was, in other settings, reported to be one of the primary influencing factors in overall comfort perception [28].

2.5 Thermal Comfort Estimation

Automated recognition of thermal comfort levels has been a research focus in energy and building ergonomics. For example, energy efficiency in office buildings or other occupant spaces could be improved through adaptation to occupants' current needs [93]. Furthermore, adaptive models for personalized thermal comfort were also correlated with an increase in occupant productivity [10] as environmental factors are better tailored to occupants' needs. Previous data-driven approaches for thermal comfort estimation showed differences in the used datasets (self-recorded or publicly available). With a self-recorded dataset featuring data from 12 participants, Zhang et al. [106] developed a building context machine-learning model that estimates thermal sensation with an accuracy of 95.4% by identifying frowning facial expressions in recorded RGB data and relating the occurrences to the currently perceived thermal sensation. Mao et al. [59]concluded that thermal comfort could be predicted using heart rate and left-arm wrist skin temperature measures by using a self-recorded dataset to train different machine-learning models. As part of the advances in thermal comfort research, gathering sufficiently large datasets for data-driven thermal comfort estimation has become a research focus, resulting in the acquisition of the ASHRAE RP-884 [19] and, more recently, the ASHRAE II [25] datasets. Both datasets accumulate environmental, personal attributes, and thermal indices, one of which is the PMV index. Moreover, both ASHRAE datasets are publicly available. Using the ASHRAE RP-884 [19], Scales [79], and US Office Buildings dataset [51] (all public), Somu et al. [84] built a machine-learning model that employs transfer learning strategies and achieved a prediction accuracy of 55%. The Scales Project dataset is a cross-national dataset (30 countries) that explores the occupants' understanding of common thermal sensation scales, such as the previously described seven-point scale. It includes thermal comfort labels based on different rating scales for thermal conditions. Additionally, personal, indoor, and outdoor environmental factors were gathered using a questionnaire. The US Office Buildings dataset is aimed at office spaces in the US and was recorded to explore human-building interactions driven by factors such as comfort and behavioral changes over time [51]. It includes data on personal attributes, indoor/outdoor variables, and labels gathered with various thermal comfort rating scales.

Francis et al. [26] presented OccuTherm, a system to estimate thermal comfort using the body shape. They conducted a sensing study in which biometrics, physical measurements (height, shoulder circumference), and subjective comfort responses were recorded. They find that an adapted personalized comfort model can improve model performance to 60% accuracy.

Quintana et al. [74] collected a longitudinal dataset of 17 participants over four weeks across 17 indoor and outdoor spaces. The dataset includes physical characteristics, background information, and personality surveys, which were assessed once. During the four-week trial, thermal preference, clothing level, metabolic rate, perceived air velocity, and location were assessed. The dataset contribution contains 1.400 unique responses across 17 indoor and outdoor spaces.

Quintana et al. [75] also introduced a cohort comfort model to reduce the necessity for personalized data. While not achieving the accuracy of personalized models, this approach opens novel ways to estimate thermal comfort without additional data.

An overview of publicly available datasets is given in Table 3. Given the low number of public datasets (six in total), most thermal comfort research is still performed on self-recorded datasets that vary in terms of included

measures and employed sensory devices. Additionally, no thermal comfort dataset for the automotive use case currently exists.

3 METHOD - THERMAL COMFORT STUDY

This work investigates thermal comfort estimation from temporal data in two scenarios, indoors and in vehicles. To assess thermal comfort for such scenarios, we conducted two studies enabling us to collect human judgments for varying thermal conditions in each scenario. For the indoor scenario, we set up a climate chamber, enabling us to control thermal conditions carefully. However, the study inside a vehicle is conducted with less control, utilizing a BMW 3 Series to capture the thermal properties of real in-vehicle conditions. In this section, we provide details on how we collected human feedback during the study using our developed application. For each subject, we gathered demographics using a self-report integrated into our application.

Self-reports and labels were gathered using a GUI that incorporates various dialogues. A keyboard was used for interaction because it requires very little space and does not require elaborate hand movements.

During the trials, a feeling of boredom quickly manifested due to the repetitive task of simply labeling one's thermal comfort level. Consequently, for the indoor study, the waiting dialogue between labeling prompts was extended to display a slide show of fractal images changing every 2 seconds. A dialog asking users to provide a rating on the 7-point thermal comfort scale was displayed with an interval of 20 seconds. While we could not provide an acclimatization period between the heating and cooling period, we let participants first fill out a demographic questionnaire and explained the scenario. This took approximately 20 min, providing sufficient time for initial acclimatization of the climate chamber. Nonetheless, this remains a limitation of our indoor dataset.

The experimental procedure followed the guidelines of the ethics committee of our university and adhered to regulations regarding the handling of sensitive and private data, anonymization, compensation, and risk aversion. Compliant with our university's local regulations, no additional formal ethics approval was required.

4 THERMAL COMFORT INDOOR DATASET

For the exploration of temporal data in conjunction with thermal comfort estimation and to provide a scientific comparison of human thermal comfort between indoor scenarios and in-vehicle scenarios, within this work, we collect human judgments for both scenarios. In this section, we provide a detailed description of our INDOOR dataset acquisition. The INDOOR dataset represents scenarios where participants are sitting, which constitutes almost 70% of a typical work day in the office and 60% at home, thereby showing the relevance and appropriateness of the scenario [17].

Gathering ratings for thermal comfort states can be done using the thermal sensation scale referenced in the ASHRAE standard [48], which is defined in the same range as the PMV output scale proposed by Fanger et al. [24]. The thermal sensation scale comprises seven-string encodings *Cold, Cool, Slightly Cool, Comfortable, Slightly Warm, Warm, Hot* but can be represented numerically as $[-3, -2, -1, 0, +1, +2, +3]$. Additionally, a significant attribute of the PMV index is its scope, as it was designed to predict thermal comfort in steady-state environments [1, 24], yet, temperatures inside buildings can vary depending on the context, seasonal climate, geographical climate, architectural properties, heating/cooling system, etc. To simulate the temperature of buildings, these factors need to be considered for data acquisition. In particular, the rate of change is important as thermal comfort zones in cooling and heating phases were found to differ, especially with smaller temperature step changes [16]. A slow rate of temperature change in previous work was defined as 0.5 ℃/min, while a fast change rate was defined as 1.0 ℃/min [16]. Temperature ranges used in thermal comfort experiments included ranges such as [18°,35℃] [83] or even larger ranges with a minimum of 15℃ and a maximum of 40℃ [16]. Consequently, the temperature

ranges and step sizes should be selected so that the conditions for all possible thermal comfort states are met at least once within the heating and cooling phases.
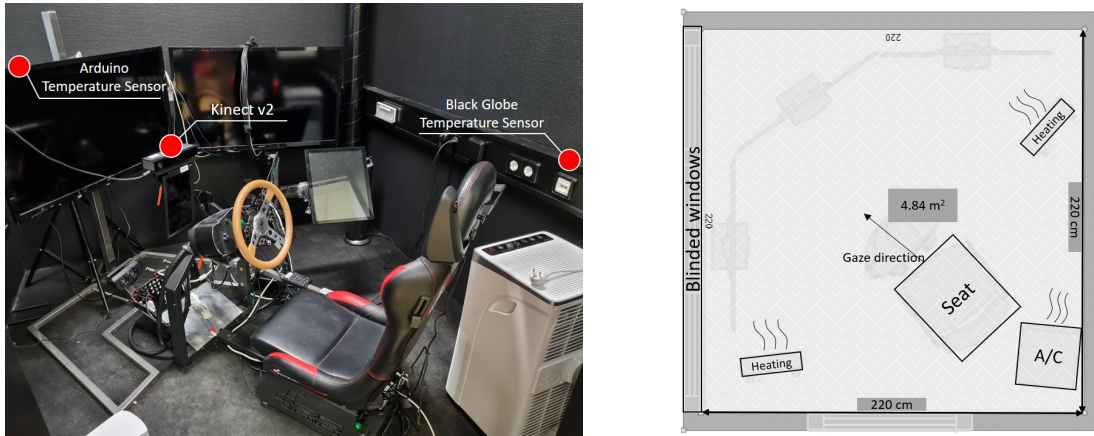
(a) The room as used for the thermal comfort recordings. The employed A/C unit can be seen on the bottom right. Temperature sensors were placed above the A/C unit and on top of the monitor on the left-hand side.

(b) The schematic room plan used for the thermal comfort recordings. A 3d version with stylized devices can be found via this link.

Fig. 2. Overview of the climate chamber.

Sensory setup. Ciuha et al. [16] used a climatic chamber and spanned 25℃ (range=[15℃, 40℃]) over 150 min. Due to the unavailability of such a professional chamber, we built a low-fidelity climatic chamber using mobile A/C devices. The room had a size of 220cm × 220cm with a height of 280cm (see Figure 2b). This results in a volume of 13550l. While previous scarcely report the room volume (e.g., Battistel et al. [6] do not report this), this is in the range of available commercial climate chambers (e.g., see Clitec). The room is painted completely black, from the carpet to the walls and ceiling to the blinds for the 3 small windows. We used the Monzana MZKA1000 Smart A/C (see Figure 2a) with a power output of 9000 BTUs as it provides cooling and heating modes and includes a smart home cloud that allows for developer access via APIs. This device enables temperature changes in the range of [16℃, 32℃]. As the A/C unit's heating capability is less powerful than its cooling mode, two additional smart home heating units (Nedis P22-2054875) were added. The smart heating components allow for temperature changes in the range of [15℃, 35℃] and include two heating modes (1.200W/2.000W). Adding multiple A/C units for heating also ensures that the area is heated at multiple locations, which allows for a more uniform temperature change. Other than the mentioned smart A/C and heating units, no other components were used to manipulate the ambient temperature (i.e., no seat heating). A series of internal tests showed that the ambient temperature range achieved most reliably within 60 min was [18.4℃, 32.0℃]. The time interval of 60 min was selected to ensure that a mean temperature change rate of 0.45℃/min is achieved for both heating and cooling. This change rate is favorable compared to a high rate of 1℃/min as large thermal changes over short periods are perceived more intensely and, therefore, shift the range at which a person feels comfortable at [16].

The PCE-WB 20SD thermometer was selected due to its logging rate of 1Hz and the ability to record ambient temperature, relative humidity, and radiation temperature with the integrated black globe components [50]. Black globe temperature is measured with the thermometer being inside a black globe. This means that this indicates how hot it feels in direct sunlight. A cheaper and easily integrable solution for temperature and humidity data is

Table 1. List of derived modalities from different input spaces for multi-modal thermal comfort estimation.

Input Space	Modalities
Personal Context	Age [20, 31], Gender [12, 47], Clothing [1, 48], Tiredness [30], Weight & Height [53], Body Fat [38], Metabolic Rate & Activity Level [1, 48, 90]
External Context	Relative Humidity [1, 44], Ambient Temperature [16, 48], Radiation Temperature [1, 4, 48], Air Velocity [1, 48]
Physiology	Heart Rate [57, 59], Wrist Skin and Body Core Temperature [12, 27, 102], Galvanic Skin Response (GSR) [57]
Visual Attributes	3D Body Pose [100, 101], RGB View [106]
Emotion	Emotional States (after Ekman [22]) [95] with neutral emotion

Arduino sensory units [2]. However, these kits do not provide the same level of accuracy as specialized measuring tools do. Nevertheless, ambient temperature and humidity data from an Arduino sensory board were included in the data logging application to be able to compare prediction performance with different frequencies and levels of accuracy for temperature and humidity streams. As for physiological signals, we used the Empatica E4 [23]. Modalities like emotion, body pose, and visual features can be captured using appropriate machine-learning models and RGB frame processing. For emotion estimation, *Serengil and Ozpinar* presented a deep learning model that includes multiple face detector backends and allows for analysis of recorded RGB frames [81, 82]. For 2D pose key point estimation, we employed OpenPose [11]. Seeing as both models can be executed in real-time, they were integrated for emotion and body pose estimation. RGB frames themselves can easily be captured using a webcam. However, for the data collection process, Microsoft's Kinect v2 for Windows [62] was selected as it provides RGB frames of size 1920×1080 and depth frames of size 512×424 [43]. As the selected body pose model estimates 2D skeleton key points, the depth frame provided by the Kinect sensor is necessary to measure the respective depth values. This way, 2D key points are extended to 3D key points (see Table 1).

4.1 Participants

We gathered data from *N*=21 participants. The gender ratio was *12F/9M* with a mean age of *M*=24.64 (*SD*=3.03, *range=[20, 33]* in years). Participants weighed *M*=69.97 (*SD*=15.02, *range=[53.00, 106.90]* in kg) at an average height of *M*=174.50 (*SD*=10.18, *range=[155.00, 198.00]* in cm) and a body fat percentage of *M*=22.00% (*SD*=5.00%, *range=[14.00%, 34.00%]*). Body temperature measured at the forehead with an infrared skin thermometer was mostly similar for all participants (*M*=36.38, *SD*=0.31, *range=[35.8, 37.2]* in ℃). Clothing levels varied only slightly, as most participants wore a short-sleeve T-shirt and trousers (mean insulation based on ASHRAE standard 55 clothing insulation tables [48] *M*=0.60, *SD*=0.05, *range=[0.45, 0.69]* in clo). Only 14% (3Y/18N) of participants reported having performed physical activities in the hour before the recording session, while the time since participants' last meals was *M*=4.23 hours (*SD*=4.92, *range=[0, 20]*) before the recording. Based on the tiredness ratings in the initial personal context assessment, participants reported having been moderately tired at the start of their recording (*M*=4.12, *SD*=1.73, *range=[2, 8]*).

4.2 Raw Dataset

The initial raw dataset included recordings from all 21 participants. Due to incorrect use of the numeric keyboard during labeling, the data of one participant was removed. Also, we removed another 2 participants due to incomplete radiation temperature readings. Therefore, the final and filtered raw dataset includes 18 participants and a separate image archive with the RGB frames. We used 16 participants for training and 2 participants to evaluate our models (see Section 7). The dataset includes corresponding body pose key-point coordinates for each

Table 2. Comparison of minimal and maximal values, means and standard deviations of the numeric features from the INDOOR and VEHICLE dataset. *ibp* stands for the (averaged) image brightness per pixel as an approximation for solar radiation.

		INDOOR dataset		VEHICLE dataset	
Participants		18		20	
Duration of recordings	minutes	60		30	
		[min, max]	*mean±std*	[min, max]	*mean±std*
Radiation Temperature	°C	[16.9, 33.6]	25.53±3.92	-	
Heart Rate	bpm	[40.0, 191.99]	82.94±13.46	[37.28, 191.99]	77.70±24.74
Wrist Skin Temperature	°C	[27.91, 36.95]	33.82±1.75	[25.55, 36.43]	30.89±2.54
Galvanic Skin Response	μS (microsiemens)	[0.0, 16.9]	1.42±2.81	[0.01, 9.19]	0.57±1.21
Ambient Temperature PCE-WB 20SD	°C	[17.1, 33.7]	25.31±3.72	-	
Ambient Temperature Arduino	°C	[17.6, 37.0]	26.82±4.52	[10.0, 35.40]	25.13±4.23
Relative Humidity	%	[12.0, 55.0]	31.35±8.86	[10.0, 67.0]	29.54±9.20
Solar Radiation	ibp	[0.16, 0.45]	0.28±0.04	[0.0, 0.88]	0.46±0.06
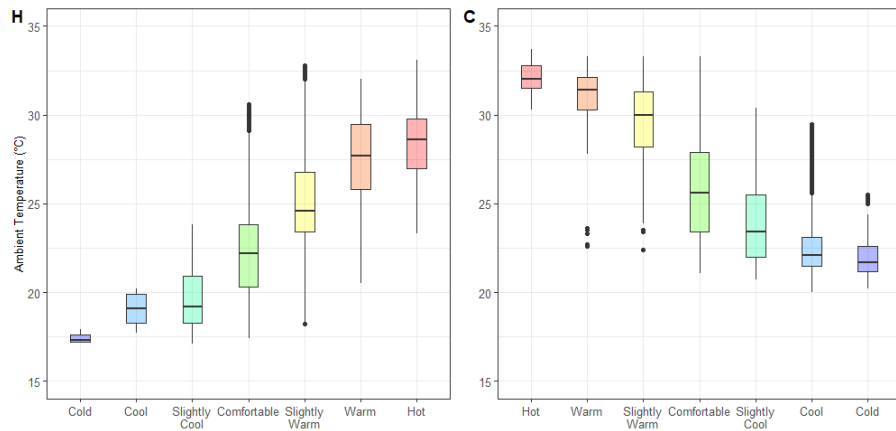
Fig. 3. Left: Labels given during heating. Right: Labels for the cooling phase. In line with previous research (e.g., [16]), thermal comfort is established at a lower ambient temperature during heating than during cooling.

recording. Due to movement out of the depth camera's field of view during the recordings, not all key points could be estimated reliably at all times, which led to empty vectors in the dataset.

The final dataset includes a total of 1.856.290 data points with extrapolated labels, each with 34 feature columns (including timestamp and label columns) from which 2927 data points were actually labeled by the user. Table 2 shows for the INDOOR dataset a descriptive evaluation for the numeric features not gathered using the self-report GUI. *Radiation and ambient temperature* are fairly similar and could replace each other during classification. Concerning the values measured with the Empatica E4 wristband, the most stable measurements were achieved for the wrist skin temperature. This is indicated by the low standard deviation (*SD=1.75*) and minimum and maximum values (*min=27.81, max=36.95*) close to the mean of 33.82°C. Contrarily, the raw measurements for the remaining physiological signals, heart rate, and GSR were far less stable. The highest outlier rate out of the numeric continuous features was found in the GSR measurements (13.7% outliers). Therefore, a data pre-processing scheme that includes outlier removal methods is required when using the dataset for classifier training.

The PMV calculation also requires the air velocity caused by body movement and not the air velocity measured by a given air velocity sensor. Thus, an approximation was used. Air velocity was not measured but deduced from the capabilities of the A/C unit as listed in the manual. The manual states that a max air volume flow of 360 m/h is possible (page 37). Therefore, we set the air velocity to static 0.1 m/s.

The raw dataset further includes two *emotion* feature columns, as participants were able to report their current emotions (after Ekman [22] with the additional *neutral* emotion; exclusive choice: "Which item describes your current emotion best?" with the options: Anger, Fear, Sadness, Disgust, Happiness, Neutral, Surprise, and Contempt; see Appendix H) using the logging GUI while also having their emotions estimated using the RGB capture and the DeepFace model. Emotions were estimated at the same frequency as emotion self-report dialogues were displayed. Most participants rated their emotions as *neutral*. This resulted in different rating distributions across emotion feature columns. Self-reported emotions were distributed as follows: *Anger 0.83%, Contempt 0.00%, Disgust 3.14%, Fear 0.00%, Happiness 8.35%, Neutral 86.38%, Sadness 0.73% and Surprise 0.49%.* The model-based emotion predictions also tended towards neutrality but less strongly: *Anger 10.99%, Disgust 0.08%, Fear 13.48%, Happiness 8.98%, Neutral 41.14%, Sadness 20.87%, and Surprise 4.46%.* This indicates a mismatch between felt emotions and detected facial expressions by DeepFace. Therefore, we refrained from using the data from the DeepFace model. We also did not use the self-report emotion data due to the very low variance. The dataset also includes two different *ambient temperature* features. As described at the beginning of this section, an external thermometer was used during the recordings to measure radiation temperature changes. However, when comparing the ambient temperature measures of the external thermometer and the Arduino sensory kit measures, it could be observed that the Arduino kit tends to react more intensely when ambient temperature changes occur. Moreover, the temperature sensor used in the sensory kit is labeled to have an accuracy of 2°C [3], whereas the external thermometer (PCE-WB 20SD) is labeled with an accuracy of 0.8°C [40]. For this reason, further analysis and visualizations of ambient temperatures are based on the values measured with the more accurate PCE thermometer. The label distribution throughout the dataset suggests that the minimum temperature during the trials was insufficient for inducing cold and cool thermal sensations. The answers given during participants' debriefing further support this assumption, as it was mentioned that *"the temperature, in the beginning, felt somewhat cool, but you get used to it quickly"* [P16] and *"it didn't get very cold, but it did get quite hot."* [P7]. The collected labels are distributed as follows: *Cold 4.49%, Cool 9.53%, Slightly Cool 22.27%, Comfortable 22.71% (this would represent the accuracy of a null model predicting only "comfortable"), Slightly Warm 13.30%, Warm 15.83%, Hot 11.88%.* A difference in thermal comfort ratings could be found between the heating and cooling phases, as illustrated in Figure 3. This is in line with previous findings that reported a thermal comfort zone shift based on previous exposure to different thermal environments [16]. Moreover, a thermal comfort zone can be seen between the labels *Slightly Cool* and *Slightly Warm*, as the temperature ranges for slightly cool, comfortable, and slightly warm states are the largest among all reported states. From a classification perspective, this indicates that classifications of states in the thermal comfort zone may be more difficult to predict based on ambient temperature alone, while colder states may be more difficult to predict due to the imbalance in frequency of occurrence in the dataset.

While some of these features might not encode much information due to inherent limitations (e.g., 3D pose limited by the available space), we emphasize that additional data might facilitate novel approaches. Therefore, we strived to collect a rich and large-scale dataset.

5 THERMAL COMFORT VEHICLE DATASET

To collect data for a real vehicle scenario, we used a BMW 3 F31 for conducting a second study and data acquisition (see Figure 4). We refer to this dataset as **VEHICLE dataset**. During the study, we utilized the application described in Section 3. To manipulate temperature, we used the built-in A/C and adjusted the temperature manually. The
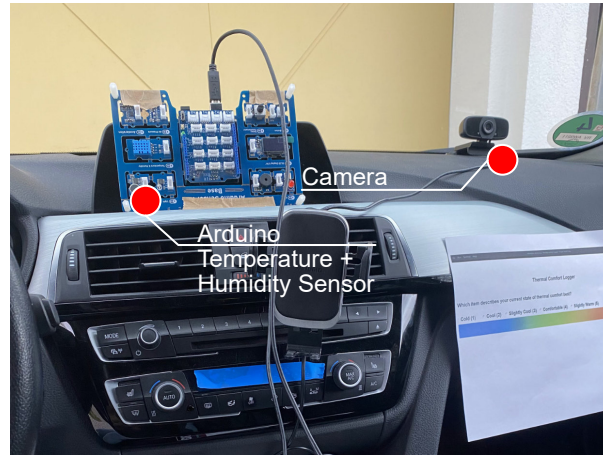
Fig. 4. Sensory setup inside the BMW 3 F31. The Arduino board was mounted on top of the middle A/C vents. The displays were covered to hide temperature information from participants. In-vehicle setup did not include a Kinect camera.
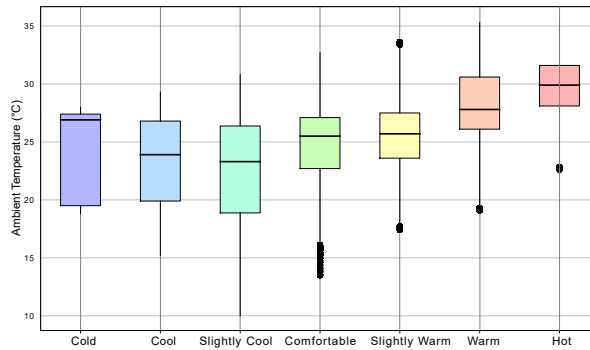
Fig. 5. Labels given during in-vehicle study.

temperature range was narrower compared to the INDOOR study due to the lower power of the built-in A/C. Participants provided data annotations in a time frame of 30 min. 14 participants provided data while the vehicle was parked, 6 while the experimenters drove the vehicle in a small town and highways in the region of Bad Waldsee, Germany. The participants remained in the passenger seat. The possible inputs were shown on a printout (see Figure 4). During the in-vehicle study, users were notified by a beep sound to provide a rating of their current thermal comfort state. We determined the input spaces and modalities shown in Table 1 such that the PMV variables (ambient temperature, relative humidity, metabolic rate, clothing insulation, radiation temperature, air velocity) [24] along with modalities for the identified physiological, visual, emotional, and personal input spaces are included. Additionally, we approximate solar radiation by computing the average brightness of an image frame extracted from the RGB video that was captured during the study.

5.1 Participants

We gathered data from *N*=20 participants, two of whom had participated in the INDOOR study. The gender ratio was *7F/13M* with a mean age of *M*=33.41 (*SD*=12.67, *range=[25, 66]* in years). Participants weighed *M*=73.52 (*SD*=23.47, *range=[53.00, 119.0]* in kg) at an average height of *M*=179.89 (*SD*=10.85, *range=[161.00, 198.00]* in cm). Body temperature measured at the forehead with an infrared skin thermometer was mostly similar for all participants (*M*=36.14, *SD*=0.15, *range=[35.8, 36.4]* in ℃). Clothing levels varied only slightly, as most participants wore a long-sleeve shirt and trousers (mean insulation based on ASHRAE standard 55 clothing insulation tables [48] *M*=0.67, *SD*=0.06, *range=[0.57, 0.81]*). None of the participants reported having performed physical activities in the hour before the recording session, while the time since participants' last meals was *M*=3.75 hours (*SD*=3.83, *range=[0, 14]*) before the recording. Based on the tiredness ratings in the initial personal context assessment, participants reported having been moderately tired at the start of their recording (*M*=2.75, *SD*=1.13, *range=[1, 5]*).

5.2 Raw Dataset

To investigate the thermal comfort of humans under real conditions, we collected a data set using a real car (i.e., a BMW 3). The study was carried out from the end of March until the end of April 2023, with varying outdoor temperatures ranging from 6 to 14℃. During the study, sensory data of the Empatica E4 for wrist skin temperature, GSR, and heart rate were recorded. Additionally, we recorded ambient relative humidity and ambient temperature using an Arduino board. We did not measure air velocity in the vehicle due to the necessity to report the velocity in relation to the participants' movement. We report the mean, standard deviation, and minimal and maximal value for the corresponding sensory readings in Table 2. For the study, we randomized the following conditions, which were each conducted one or multiple times in random order: switching off the A/C, switching to maximum heating (28 ℃), switching to maximum cooling (16 ℃), opening the window for at least 1 min. Additionally, we randomized driving the car and standstill of the car per participant for the study. The collected data set consists of 1.069.374 labels, which we extrapolate from 1597 collected labels from the user.

6 COMPARISON TO PUBLICLY AVAILABLE DATASETS

We compare our INDOOR and VEHICLE dataset to related available datasets. Table 3 shows that our AUTOTHERM INDOOR dataset has over 17 times the number of entries compared to the next largest dataset from prior work. All datasets include measures from the personal and environmental input space. The environmental variables for the INDOOR dataset, ASHRAE RP-884, ASHRAE II, and US Office Buildings datasets include the environmental measures necessary for PMV calculation. The Scales Project dataset includes weather station and indoor environment data; however, no specialized tools (e.g., black globe thermometers for radiation temperature measures or anemometers for air velocity) were used. Thus, PMV estimations cannot be computed using the Scales Project dataset. The UCLIC-Bentley Comfort (UBComfort) dataset is the closest dataset to the AUTOTHERM dataset but contains fewer data points with fewer features. However, it does include additional features such as mental relaxation level and sitting discomfort. The listed publicly available datasets include ratings that were gathered with different rating scales (*TS=thermal sensation*, *TP=thermal preference*, and *TA=thermal acceptability*). The AUTOTHERM dataset differs from the publicly available datasets by including physiology, emotion, visual signals, and participants' RGB recordings. Additionally, the AUTOTHERM dataset includes **temporal data** compared to the singular data of the other datasets. In Appendix C, Figure 8 shows an example image (blurred) for the INDOOR and VEHICLE dataset.

7 VALIDATION

We conducted several experiments on the AUTOTHERM INDOOR dataset acquired (see Section 4). First, we performed feature importance ranking using an impurity-based method. Based on the found feature importance, we

Table 3. Comparison to publicly available datasets with our AUTOTHERM DATASET. 7P=seven-point thermal sensation (*Cold, Cool, Slightly Cool, Comfortable, Slightly Warm, Warm, Hot*), 3P=three-point thermal preference (*Want Warmer, No Change, Want Cooler*), and 2P=two-point thermal comfort (*Acceptable, Unacceptable*). Measurements coded as P=Personal Context, EX=External Factors, PH=Physiological Factors, E=Emotions, and V=Visual Attributes. The ASHRAE datasets also include different environmental indices (e.g., PMV index).

Dataset	Participants	Entries	Temporal Data	Included Measurements					Rating Scales		
				P	EX	PH	E	V	7P	3P	2P
ASHRAE RP-884 [19]	25.288	25.288	✗	✔	✔	✗	✗	✗	✔	✔	✔
ASHRAE II [70]	109.033	109.033	✗	✔	✔	✗	✗	✗	✔	✔	✔
Scales Project [79]	8.225	8.225	✗	✔	✔	✗	✗	✗	✔	✔	✔
US Office Buildings [51]	24	2.503	✗	✔	✔	✗	✗	✗	✔	✔	✔
OccuTherm [26]	77	2.067	✔	✔	✔	✔	✗	✗	✗	✗	✗
LPTC [74]	17	1.403	✔	✔	✔	✔	✔	✗	✗	✔	✗
UBComfort [68]	28	587	✔	✔	✔	✔	✗	✔	✔	deducible	deducible
AUTOTHERM INDOOR	18	1.856.290(2.927*)	✔	✔	✔	✔	✔	✔	✔	deducible	deducible
AUTOTHERM VEHICLE	20	1.069.374(1.597*)	✔	✔	✔	✔	✔	✔	✔	deducible	deducible

*Number of original labels provided by the participants.

conducted a feature combination study in Section 7.7 to find the best combination of features. In Appendix E, we investigate feature importance of physical, psychological, and physiological features by performing feature permutation based on RF and LSTM model estimations. Having these insights, we considered an input vector of four features for all experiments and implemented three types of classification models. As our baseline, we use an RF classifier (Section 7.1). We also measure the performance of the PMV index on our dataset in Section 7.2. In Section 7.3, after the evaluation of the best classifier, we continue with a comparison of our dataset to the ASHRAE II [70] dataset. We compare these against deep learning models based on recurrent networks (RNN) (Section 7.4), and a combination of RNN and convolutional neural networks (CNN), in Section 7.5. Additionally, we investigate the forecasting of time series data to predict thermal comfort for a future state in Section 7.6. Next, we provide a short description of the used metrics to quantify estimation accuracy.

Thermal Comfort Metric. To measure a classifier's performance, we use three metrics of accuracy with varying scales of precision: κ_n, where n is the number of classes used to compute accuracy. First, the seven-point thermal sensation scale (*Cold, Cool, Slightly Cool, Comfortable, Slightly Warm, Warm, Hot*), which is denoted as κ_7. Second, the three-point thermal preference scale (*Cold, Comfortable, Warm*), denoted as κ_3. For computing κ_3, we reduce seven classes by mapping the classes (*Cold, Cool, Slightly Cool*) to *Cold*, (*Slightly Warm, Warm, Hot*) to *Warm*, and *Comfortable* remains. And third, the two-point thermal comfort scale (*Comfortable, Uncomfortable*), which we denote as κ_2. Note that for computing κ_2, we reduce seven classes to binary classes by mapping the classes (*Slightly Cool, Comfortable, Slightly Warm*) to *Comfortable* and the remaining classes to *Uncomfortable*.

For all experiments, our training dataset consists of data from 16 participants, while our test data consists of the two remaining participants. For the LSTM approach (the **best-performing approach**, see Section 7.4), we additionally conducted a 20-fold cross-validation (see Appendix A), showing robust performance.

AUTOTHERM represents the only approach to have acquired both in a field and an in-lab study and to focus on buildings and vehicles. Other work only focused on field studies with a focus on buildings.
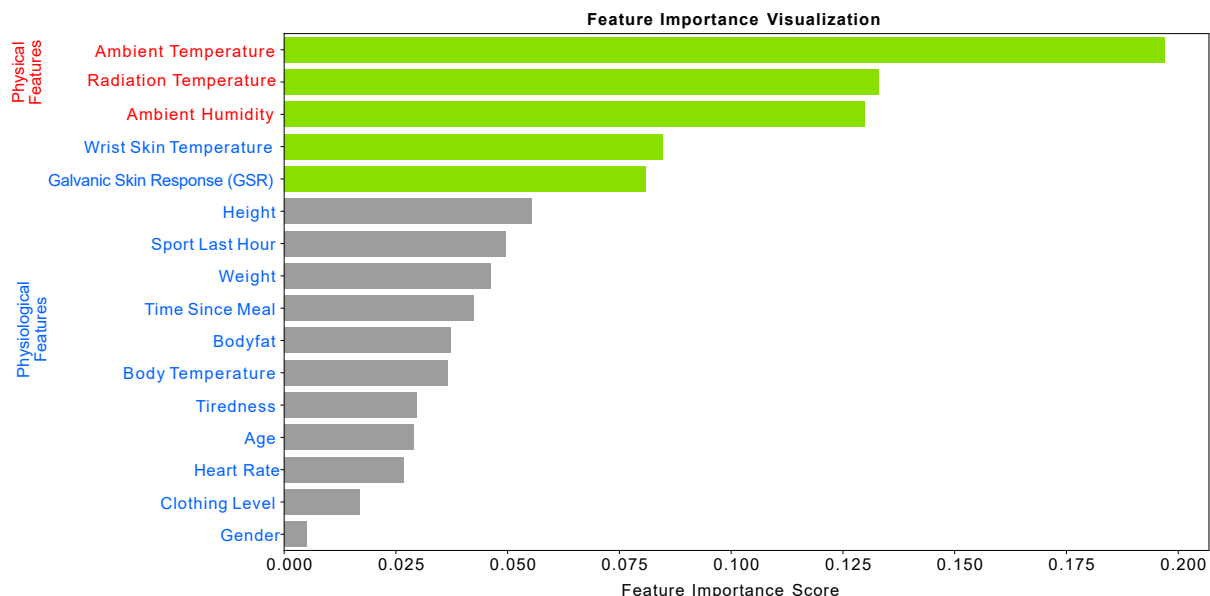
Fig. 6. Feature importance ranking based on mean accuracy impurity decrease of an RF classifier. According to this metric, the five most important features were ambient temperature, relative humidity, radiation temperature, skin temperature, and galvanic skin response, marked in green.

7.1 Random Forest Classifier

In a first step, we leverage an RF classifier to perform an impurity-based importance ranking of all available features, see Figure 6. For this task, we use the RF implementation from the sklearn library [80], which provides functions for dataset sampling, pre-processing, and training pipeline definition. While RF models provided by sklearn allow for tuning of different parameters such as the number of estimators, maximum tree depth, and maximum number of features to consider per node, the standard configuration is used at first and then later optimized using a grid search approach. For our experiments, we used 400 estimators and a maximum tree depth of 8 adopting the Gini impurity cost function [45]. We used every 100th data point to downsample the dataset. **Results.** In Table 4, we report the classification accuracy of $\kappa_7 = 47.1\%$, $\kappa_3 = 73.8\%$, and $\kappa_2 = 67.9\%$, for the RF classifier, indicating a mediocre performance.

7.2 PMV and Scale Reduction Performance

To compare the performance of our proposed models to the PMV index, we pre-computed the PMV values for all participants in the evaluation split of our dataset using the clothing level, radiation temperature, ambient temperature, and relative humidity features. For computation of PMV measures, the PyThermalComfort package [88] was used, which enables PMV computation based on the definitions in ISO7730 [1] and ASHRAE standards [48]. **Results.** A comparison with the participants' subjective thermal sensation ratings showed that the PMV index fails to accurately predict thermal comfort. The achieved prediction accuracies are $\kappa_7 = 35.9\%$, $\kappa_3 = 63.7\%$, and $\kappa_2 = 65.2\%$, suggesting that the PMV index values were mostly off by one class.

Table 4. Evaluation results of a performance comparison between different classifiers on the INDOOR dataset. In the last row, we report the performance of the PMV index computed for the corresponding PMV values.

Classifier	κ_7	κ_3	κ_2
LSTM	59.3%	83.5%	71.9%
Forecast 10 sec	56.6%	80.4%	74.4%
Forecast 5 min	55.5%	77.7%	74.6%
Forecast 10 min	50.9%	77.1%	73.3%
CNN-LSTM	48.5%	69.4%	76.0%
Random Forest	47.1%	73.8%	67.9%
PMV	35.9%	63.7%	65.2%

7.3 ASHRAE Thermal Comfort Field Measurements

In this experiment, we focus on a comparison of our INDOOR dataset to existing thermal comfort datasets. After a close inspection of publicly available datasets, the ASHRAE II [70] dataset features comparable aspects to our collected dataset, as it provides measurements for *Radiation Temperature*, *Ambient Temperature*, and *Relative Humidity*, which are also measured in our dataset. While our dataset provides sequences over time, ASHRAE II recorded single data points, which makes it unfeasible to compare against our recurrent-based methods. Instead, we optimize an RF classifier, described in Section 7.1, using the ASHRAE II dataset. After filtering out data points with incomplete measurements, a total of 31.500 data points remain. We use 80% of the ASHRAE II data for training, resulting in 25.204 data points and the remaining 6.301 data points for evaluation. Further, we use the trained classifier from Section 7.1, which is trained on our dataset, and evaluate it on ASHRAE II. Finally, we use the classifier trained on ASHRAE II and evaluate it on our dataset. We report evaluation results in Table 5.

Results. The evaluation split of our dataset seems to be more difficult than ASHRAE II. This is indicated by a worse performance of the classifier trained and evaluated on our dataset than the classifier trained and evaluated on ASHRAE II. Also, the performance of the classifier trained on ASHRAE II and evaluated on our dataset performs worse than the classifier trained on our data and evaluated on ASHRAE II, which further supports our observation. This comparison shows a big gap between both thermal comfort datasets while underlining the missing feature information when only three sensor measurements are used. Further, this experiment shows a drastic performance drop when only single thermal state values are used for state recognition. Looking at the 2-point performance measure, the classifier trained on our dataset generalizes to the ASHRAE II dataset.

Table 5. Comparison results of an RF classifier that was optimized on our INDOOR dataset (AutoTherm) from the climatic chamber and evaluated on the ASHRAE II dataset and vice versa. We report evaluation performance for three metrics: κ_7, κ_3, κ_2. Each classifier was trained using *Radiation Temperature*, *Ambient Temperature*, and *Relative Humidity* as input vectors.

Train	Evaluate	κ_7	κ_3	κ_2
AutoTherm	AutoTherm	40.6%	76.1%	69.1%
ASHRAE II	ASHRAE II	45.8%	45.8%	82.7%
AutoTherm	ASHRAE II	16.2%	22.3%	82.8%
ASHRAE II	AutoTherm	2.7%	39.7%	37.5%

7.4 Classification Using Time-Dependent Information

We investigated representing our time-dependent data as a time series of feature vectors and leveraging a recurrent neural network to process sequences of features. Thus, we implement our model in a standard encoder-decoder structure. Our feature encoder network consists of two long short-therm memory layers (LSTM) [37], followed by a decoder network, which outputs a probability distribution of classes. We use the deep learning framework PyTorch for implementing the LSTM layers, along with PyTorch-Lightning for module and training cycle management. We formulate the classification task as a regression problem by using ordinal labels [13]. For further details, we refer the reader to Appendix C. We use the mean-squared error (MSE) as a loss function to regress ordinal labels. We use the following hyperparameters: learning rate=0.00001, learning rate decay=0.99, batch size=16, and dropout=0.5. For the two LSTM layers, we use a hidden state size of 64 and a sequence length of 30. Also, we use every 10th data point, corresponding to a sequence length of 10 seconds. In our hyperparameter search (see Appendix D), we found the selection of these two hyperparameters to be the best tradeoff between the length of the input sequence and the oversampling of the same values.

Results. The training of our network resulted in an accuracy of $\kappa_7 = 59.3\%$, $\kappa_3 = 83.5\%$, and $\kappa_2 = 71.9\%$, outperforming the RF classifier, see Table 4, which is an expected result regarding the increased number of parameters of our recurrent classifier and the ability to extract neural representations of sequence data.

7.5 Vision-based Thermal Comfort Estimation

In this experiment, we evaluated estimating thermal comfort with **additional** visual data. Hence, an additional model was implemented to also incorporate RGB and body pose key points. While training an LSTM with normalized RGB tensors is possible, it introduces high redundancy and complexity as individual pixel values are processed for a sequence of images. In past works, architectures for tasks such as action recognition have been proposed, which include a feature extraction step before the LSTM component [91]. For image feature extraction, architectures that employ CNNs are often used [69]. To obtain image features from each participant's RGB frames, ResNet [34] was selected due to its ability to filter deep features in images while also introducing mechanisms to avoid the vanishing gradient problem. The PyTorch framework provides pre-trained ResNet versions [73] that can be incorporated directly into existing models. We train our network with a batch size of 4 and alter the skip rate from 10 to 3 to achieve sufficient training speed for the image-based RNN model. Image feature vector extracted by the ResNet with size 512 and the ten body pose key points (given by OpenPose [11]) are concatenated to the feature vector. We stick to the best-performing feature combination, described in Section 7.4.

Results. The classifier reached accuracies of $\kappa_3 = 48.5\%$, $\kappa_3 = 69.4\%$, $\kappa_2 = 76.0\%$, falling behind our classifier trained without image features. A quantitative comparison between all models is reported in Table 4.

7.6 Forecasting using Recurrent Classifier

In our experiments, we apply recurrent neural networks (RNNs) to sequential data to predict the state of a future point in time for a given sequence of the current time step. As RNNs are also successfully used for time series forecasting [18, 21], we investigate different forecasting ranges to explore the accuracy of predictions using various forecasting windows. We adopt the classifier presented in Section 7.4 and train it using labels from time steps later in the future. The forecasting gaps used in our experiments range from 10 seconds, 5 min, and 10 min into the future. We train our classifier using the identical training protocol as it is described in Section 7.4.

Results. In Table 4, performance results for our forecasting experiments are reported in row two until row four. The prediction performance between the 10-second window and the 5-min window differed only slightly: $\kappa_7 = 56.6\%$, $\kappa_3 = 80.4\%$, $\kappa_2 = 74.4\%$ for the 10-second window, $\kappa_7 = 55.5\%$, $\kappa_3 = 77.7\%$, $\kappa_2 = 74.6\%$ for the 5-min window. And the 10-min shows lower accuracies, $\kappa_7 = 50.9\%$, $\kappa_3 = 77.1\%$, $\kappa_2 = 73.3\%$, than both previous forecasting

performances. Experimental runs with forecasting windows exceeding 10 min showed further decreases in prediction performance.

7.7 Feature Combination Study

To assess the combination of input features to our networks, we conducted a feature combination study based on the top-k ranked features by the feature importance method (see Figure 6). We selected $k = 5$, covering the most important features while also yielding a manageable number of combinations. For the experiment, we selected n features out of k, where $n \in [3, 4, 5]$. In total, this results in 16 combinations of input features. For each training run, we use the LSTM model with identical hyperparameters, as described in Section 7.4.

Results. In Table 6, we report performance results of all training runs. These results indicate that the best prediction results can be achieved using the following input features: *Galvanic Skin Response, Relative Humidity, Ambient Temperature, and Wrist Skin Temperature*, as demonstrated in Section 7.4. Using the five most important features reduces κ_7 of the classifier to 55.9%, κ_3 to 79.8%, and κ_2 to 72.1%. For detailed analysis results of the feature combination study, incorporating up to 8 features, as well as an investigation of SHAP values [56], we refer the reader to Appendix F.

Table 6. In this experiment, we report the accuracy of our classification model, ablating different combinations of input features. We highlight the feature combination performing best. During the feature combination study, we investigate the top 5 ranked features by our feature importance analysis, Figure 6: Radiation temperature (RT), wrist skin temperature (WS), galvanic skin response (GL), ambient temperature (AT), relative humidity (HU). In the last row, we include baseline results of a null model outputting the most frequent label only.

GL	AT	HU	RT	WS	κ_7	κ_3	κ_2
	Features					Classification	
X	X	X		X	**59.3%**	**83.5%**	71.9%
	X	X	X	X	59.1%	78.4%	73.5%
	X		X	X	57.7%	79.1%	73.8%
	X	X		X	57.1%	79.6%	72.7%
X	X		X	X	56.2%	79.2%	72.8%
X	X	X	X	X	55.9%	79.8%	72.1%
X	X			X	55.8%	79.4%	72.8%
X	X		X		55.4%	79.2%	72.3%
X	X	X	X		55.0%	78.0%	72.7%
	X	X	X		52.5%	78.6%	69.6%
X	X	X			50.5%	73.9%	73.1%
X			X	X	47.7%	70.8%	**74.6%**
		X	X	X	45.6%	70.5%	70.8%
X		X	X		42.0%	64.9%	69.0%
X		X	X	X	41.5%	67.5%	70.5%
X		X		X	41.1%	67.3%	68.1%
Null Model					24.8%	24.8%	63.2%

7.8 Experiments On In-Vehicle and Combined Dataset

Having collected a real scenario dataset, as described in Section 5, enables us to investigate thermal comfort estimation under real conditions and examine the gap to indoor conditions. To do so, we train two additional classifiers using the same LSTM model architecture and the same hyperparameters as described in Section 7.7 with the best feature combination. We use data from 16 participants during training and the remaining data for

Table 7. Comparing experimental results of five models trained on different datasets. The upper table shows performance results for the in-vehicle test dataset. The models in the bottom rows are tested using the in-lab test dataset. In this experiment, all models were trained using five input features: galvanic skin response, heart rate, relative humidity, ambient temperature, and skin temperature. Note that the evaluation reported in Table 6 used a different combination of features (radiation temperature used instead of heart rate, as radiation temperature was not available in-vehicle). Additionally, we include baseline results of a null model outputting the most frequent label only.

Train	Evaluate	κ_7	κ_3	κ_2
INDOOR	VEHICLE	39.2%	55.0%	67.5%
VEHICLE	VEHICLE	45.2%	**63.8%**	67.7%
Combined	VEHICLE	**48.1%**	61.4%	**69.3%**
Null Model	VEHICLE	29.6%	42.4%	64.0%
INDOOR	INDOOR	52.0%	76.4%	71.3%
Combined	INDOOR	**56.9%**	**79.7%**	**72.6%**
Null Model	INDOOR	24.8%	24.8%	63.2%

evaluation. In exception, during the in-vehicle study, five sensory measurements are recorded: galvanic skin response, heart rate, relative humidity, ambient temperature, and skin temperature, which are used as input to our classifier. In this experiment, we compared and combined the scenarios in-vehicle and buildings. Therefore, we evaluate all classifiers on our INDOOR dataset originating from the in-vehicle study, as described in Section 5, and on our INDOOR dataset, see Section 4.

Results. First, we trained one classifier solely on the VEHICLE dataset, leading to a performance of $\kappa_7 = 45.2\%$, $\kappa_3 = 63.8\%$, and $\kappa_2 = 67.7\%$. When combining both datasets (INDOOR and VEHICLE dataset), we achieved an accuracy of $\kappa_7 = 48.1\%$, $\kappa_3 = 61.5\%$, and $\kappa_2 = 69.3\%$ (see third row in Table 7). Additionally, we evaluate the classifier, which was trained on the INDOOR dataset, using the VEHICLE dataset to compare the difference between building and vehicle scenario, and we report the results in the first row of Table 7 (i.e., $\kappa_7 = 39.2\%$, $\kappa_3 = 55.0\%$, and $\kappa_2 = 67.7\%$). Looking at the evaluation results of the two models, one trained and tested on the INDOOR dataset, and the other trained and tested on the VEHICLE dataset, there is a performance difference. We attribute the lower accuracy for the VEHICLE dataset compared to the INDOOR dataset (second vs. fourth row in Table 7) to the less controlled environment and fewer data entries. Further, looking at the classifier that was trained on INDOOR dataset, its performance for the vehicle scenario lags behind the performance of the classifier that was trained using the VEHICLE dataset. This shows a gap between building and in-vehicle data. Ultimately, when combining both data sets (third and last row in Table 7), it shows that both classifiers benefit from additional data and the test performance for INDOOR dataset and VEHICLE dataset increases. We draw the conclusion from these experiments that there is a gap between both datasets and that it can be narrowed by increasing training data, yet a gap between both scenarios remains.

8 DISCUSSION

In the following, we discuss the results of the dataset acquisition, model training, and feature importance analysis, as well as similarities and differences to previous approaches in the field of state recognition.

8.1 Temporal Data for Thermal Comfort Estimation

For our RF classifier, an evaluation accuracy of 47.1% was reached on the seven-point thermal sensation scale. RF models were previously used to predict thermal preference (three-point scale) and thermal state (two-point scale), where accuracies between 76% (three-point) [54] - 94% (two-point) [12] were achieved using physiological signals. The performance of the RF model employed in this work was similar, with a prediction accuracy of

73.8% for the three-point scale and 67.9% for the two-point scale. However, the LSTM-based classifier achieved a prediction accuracy > 59% on the seven-point scale (see Table 4). We attribute the difference to the inherent mechanism, as the RF models were not processed in sequences but on an individual line-by-line basis. Therefore, changes over time could not be incorporated into the model's estimation. This is supported by the improved accuracy when employing data-downsampling. Previous RF models for thermal comfort estimation [12, 54] did not employ data-downsampling, as data was already labeled more sparsely in comparison to the thermal comfort dataset that was recorded at 30Hz.

8.2 Visual Features for Thermal Comfort Estimation

Comparing LSTM and CNN-LSTM, the LSTM model was optimized using ambient temperature, humidity, skin temperature, and galvanic skin response as input features, which were identified as the four most relevant features for thermal sensation prediction, while CNN-LSTM model features were selected based on real-world applicability. The LSTM architecture achieved a higher evaluation accuracy score than the CNN-LSTM. However, radiation temperature and even physiological features such as heart rate or GSR may not be readily available in future buildings. Furthermore, the meaningfulness of the raw extracted visual features via the ResNet block could be increased by, for instance, applying facial expression or motion detection [106], before concatenation with the remaining features, as this allows the model to predict thermal comfort on more filtered feature representations. Consequently, the LSTM model would not be guaranteed to outperform the CNN-LSTM architecture in a real-world scenario. The achieved accuracy on the dataset (59.3%) should thus be seen as an initial dataset benchmark.

8.3 Data Acquisition for State Recognition Datasets

The data acquisition in this work differed from previous approaches. Firstly, a controlled (low fidelity) climate chamber was employed over 60 min. Additionally, instead of questionnaire-based data collection, data were recorded using a data logging application that enabled direct labeling by the participants. The logging differed from previous approaches (e.g., [54, 55, 59]) as the synchronization of the various sensory inputs, as well as the labeling interface, were centralized in a single application that allows for dense sampling *without post-labeling*. Therefore, we assume that the labeling accuracy and log timing were greatly increased, as participants could report their current state by performing a minimal number of interactions with a numeric keypad. Due to the personal thermal comfort zone [14, 15], which is person-dependent, a calibration of the model during usage seems appropriate. This could be done, for example, by directly asking users (e.g., "Are you currently feeling cold, comfortable, or hot?") or by treating the manual adjustment of the A/C as input ("want warmer", "want cooler", with the amplitude of change being an indicator of the strength of this desire). However, Quintana et al. [75], with their cohort comfort model approach, showed a potential avenue to reduce required personal data.

8.4 State Recognition For Interaction Design

In line with work by Stampf et al. [85], our work helps determine the current user's state in an AV. Our dataset has relevance both for manual and especially for automated driving, which is one of the raised points by Stampf et al. [85]. With this dataset, we contribute one aspect to enable developers to create a "digital twin" that could enable novel interaction design. We envision numerous interaction possibilities here, as the uncertainty of the prediction and severity of the deducted action have to be considered. Additionally, there are multiple ways to use these states, for example, by directly adjusting values or asking the user whether they want a value adjusted. These interactions and effects will become highly relevant with ever-better recognition.

8.5 Limitations and Future Work

A low-fidelity climatic chamber was built using three (two heaters and one cooler) commercial smart-home A/C units. As the minimum temperature during the recordings was 17.1°C, cold sensations could not successfully be induced for all participants, which led to an unbalanced label distribution in the dataset. Additionally, radiation types alter comfort modeling, therefore, the heater types most likely had an impact on the reported thermal comfort [36]. Moreover, physiological signals measured using the Empatica E4 wristband were found to have a high outlier rate (up to 13.7%) for heart rate and GSR measurements. Additionally, while methods for boredom aversion during the recordings were implemented, we assume that the overall level of boredom in participants remained high due to the length of the recordings. This potential boredom, in turn, may have led to unwanted influences on emotion ratings. Furthermore, the measurements for body fat and body temperature were taken using commercial-grade measuring tools, likely affecting the accuracy of the recorded values. The recorded dataset also contains a bias due to the repeated heating pattern in all data recordings. Future work should address the dataset bias by adding further data recorded using different heating patterns and a more diverse participant population to provide more balanced data for state recognition. Regarding thermal comfort classification, future architectures should focus on processing visual and environmental features that can be collected reliably in future real-world scenarios. The comparability to other datasets is also limited because the ASHRAE II [25] dataset contains international data from singular observations, and our dataset is limited to one country Germany. Also, the variance in clothing for the participants was low, given the short timeframe we used to gather data (3 weeks in total). Also, while there is an influence of age on thermal comfort perception [20, 31], our dataset is limited to data from mostly younger participants (M=24.64, SD=3.03). To overcome such limitations observed in the current study on indoor thermal comfort and vehicle experiments, several improvements are proposed: First, diversifying the participant pool is crucial. Expanding the participant count will enhance the statistical significance of the findings, while including individuals from various demographics will provide insights into diverse thermal comfort perceptions. Secondly, the experimental conditions should be broadened. This can be achieved by implementing different rates of thermal changes, for example, by abruptly or very subtly increasing or decreasing the temperature to further study how this influences individual thermal comfort state. At the same time, future studies should focus on exposing participants to an equally distributed temperature spectrum to study a wide variety of thermal comfort scenarios and conduct longitudinal studies to observe the effects of transient thermal phenomena over time. Thirdly, conducting comparative studies across different environments and internationally could illuminate specific thermal comfort challenges and strategies, highlighting the influence of various external factors on thermal perceptions. Fourthly, to mitigate subjective differences in the raters' individual temperature range in which they feel comfortable [61], relative thermal comfort changes should be investigated. To do so, the minimal and maximal temperature values representing the minimum and maximum of the comfort scale need to be retrieved per participant and used to normalize temperature measurements. Lastly, addressing behavioral and psychological factors through targeted studies would illuminate how individual actions and mental states impact thermal comfort, further enriching our understanding of this multifaceted topic. Implementing these suggestions could significantly refine the study's approach, leading to a more comprehensive understanding of thermal comfort dynamics in both indoor and vehicular environments.

9 DATASET AND CODE AVAILABILITY

The code is available at https://github.com/az16/thermal-comfort-classification. The data is available online at HuggingFace (https://huggingface.co/datasets/kopetri/AutoTherm).

10 CONCLUSION

This work explored machine-learning-based thermal comfort recognition as a relevant target state space for future AV interaction concepts. First, relevant influencing factors on thermal comfort and the affiliated sensory devices were reviewed and filtered based on their level of obtrusiveness and their prevalence in previous thermal comfort studies used in the field of building ergonomics (e.g., [24, 27, 59]). Then, a thermal comfort experiment and a data logging application were designed to record a dataset. This dataset was collected in two scenarios: buildings and in-vehicle. This dataset includes a total of 31 different features (age, gender, weight, height, body fat, body temperature, clothing, sport, meal timing, tiredness, radiation temperature, ambient temperature, wrist skin temperature, GSR, heart rate, relative humidity, RGB frames, ten pose key points, metabolic rate, air velocity, and two emotion features). A thermal comfort study with $N = 18$ participants in a self-built climatic chamber and $N = 20$ participants in a BMW 3 Series was conducted. The datasets were then analyzed regarding feature importance based on correlation coefficients, model-estimated data impurity scores, and model-estimated permutation scores. The four most relevant features for thermal comfort estimation with the recorded data were ambient temperature, galvanic skin response, skin temperature, and relative humidity. Three different classifiers (RF, LSTM, and CNN-LSTM) were implemented and trained using the recorded dataset to assess the feasibility of thermal comfort recognition. A thermal comfort estimation accuracy of 59.3% in the climatic chamber and 45.2% in the vehicle was achieved using the implemented LSTM model and the dataset labels given on a seven-point scale (combined: 48.1%). The resulting classifier configuration was used for further experiments employing state forecasting and different label ranges (three-point and two-point scales). The reduction of labeling scales increased prediction performance in all implemented classifiers. The recorded dataset and the data logging application will be made available publicly to encourage further data recording projects for state recognition by providing a template for user-labeled data acquisition. This work provides the first dataset and reference implementation for state-of-the-art thermal comfort estimation. It helps developers and designers to gain a better understanding of their users, as well as helps introduce novel automated responses to detected states.

ACKNOWLEDGMENTS

REFERENCES

[1] ISO/TC 159/SC 5. 2005. *DIN EN ISO 7730:2006-05, Ergonomie der thermischen Umgebung - Analytische Bestimmung und Interpretation der thermischen Behaglichkeit durch Berechnung des PMV- und des PPD-Indexes und Kriterien der lokalen thermischen Behaglichkeit (ISO_7730:2005); Deutsche Fassung EN_ISO_7730:2005*. Technical Report. Beuth Verlag GmbH. https://doi.org/10.31030/9720035
[2] Arduino. 2022. Arduino Sensor Kit. https://sensorkit.arduino.cc/, last accessed on 29 JUNE 2022.
[3] Arduino. 2022. How to use DHT11 with Arduino. https://create.arduino.cc/projecthub/pibots555/how-to-connect-dht11-sensor-with-arduino-uno-f4d239, last accessed on 29 JULY 2022.
[4] Ibrahim Atmaca, Omer Kaynakli, and Abdulvahap Yigit. 2007. Effects of Radiant Temperature on Thermal Comfort. *Building and Environment* 42, 9 (2007), 3210–3220. https://doi.org/10.1016/j.buildenv.2006.08.009
[5] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2017. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP (05 2017). https://doi.org/10.1109/TPAMI.2018.2798607
[6] Laura Battistel, Andrea Vilardi, Massimiliano Zampini, and Riccardo Parin. 2023. An investigation on humans' sensitivity to environmental temperature. *Scientific Reports* 13, 1 (2023), 21353.
[7] Holger Berndt, Jorg Emmert, and Klaus Dietmayer. 2008. Continuous Driver Intention Recognition with Hidden Markov Models. In *2008 11th International IEEE Conference on Intelligent Transportation Systems*. IEEE, New York, NY, USA, 1189–1194. https://doi.org/10.1109/ITSC.2008.4732630
[8] David Bethge, Thomas Kosch, Tobias Grosse-Puppendahl, Lewis L. Chuang, Mohamed Kari, Alexander Jagaciak, and Albrecht Schmidt. 2021. VEmotion: Using Driving Context for Indirect Emotion Prediction in Real-Time. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 638–651.

https://doi.org/10.1145/3472749.3474775

[9] Pradipta Biswas and Gowdham Prabhakar. 2018. Detecting drivers' cognitive load from saccadic intrusion. *Transportation Research Part F-traffic Psychology and Behaviour* 54 (2018), 63–78.

[10] Ana Maria Bueno, Antonio Augusto de Paula Xavier, and Evandro Eduardo Broday. 2021. Evaluating the Connection between Thermal Comfort and Productivity in Buildings: A Systematic Literature Review. *Buildings* 11, 6 (June 2021), 244. https://doi.org/10.3390/buildings11060244

[11] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2021. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 1 (2021), 172–186. https://doi.org/10.1109/TPAMI.2019.2929257

[12] Tanaya Chaudhuri, Deqing Zhai, Yeng Chai Soh, Hua Li, and Lihua Xie. 2018. Random Forest Based Thermal Comfort Prediction from Gender-Specific Physiological Parameters Using Wearable Sensing Technology. *Energy and Buildings* 166 (2018), 391–406. https://doi.org/10.1016/j.enbuild.2018.02.035

[13] Jianlin Cheng, Zheng Wang, and Gianluca Pollastri. 2008. A neural network approach to ordinal regression. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. IEEE, IEEE, New York, NY, USA, 1279–1284.

[14] Ursa Ciuha and Igor B. Mekjavic. 2016. Regional Thermal Comfort Zone in Males and Females. *Physiology & Behavior* 161 (2016), 123–129. https://doi.org/10.1016/j.physbeh.2016.04.008

[15] Urša Ciuha and Igor B. Mekjavic. 2017. Thermal Comfort Zone of the Hands, Feet and Head in Males and Females. *Physiology & Behavior* 179 (Oct. 2017), 427–433. https://doi.org/10.1016/j.physbeh.2017.07.020

[16] Ursa Ciuha, Kunihito Tobita, Adam McDonnell, and Igor Mekjavic. 2019. The Effect of Thermal Transience on the Perception of Thermal Comfort. *Physiology & Behavior* 210 (July 2019), 112623. https://doi.org/10.1016/j.physbeh.2019.112623

[17] Stacy A Clemes, Sophie E O'connell, and Charlotte L Edwardson. 2014. Office workers' objectively measured sedentary behavior and physical activity during and outside working hours. *Journal of occupational and environmental medicine* 56, 3 (2014), 298–303.

[18] Jerome Connor and Les Atlas. 1991. Recurrent neural networks and time series prediction. In *IJCNN-91-Seattle international joint conference on neural networks*, Vol. 1. IEEE, IEEE, New York, NY, USA, 301–306.

[19] Richard de Dear and G. S. Brager. 1998. Developing an Adaptive Model of Thermal Comfort and Preference. *ASHRAE Transactions* 104, 1 (1998), 145–167.

[20] S. Del Ferraro, S. Iavicoli, S. Russo, and V. Molinaro. 2015. A Field Study on Thermal Comfort in an Italian Hospital Considering Differences in Gender and Age. *Applied Ergonomics* 50 (2015), 177–184. https://doi.org/10.1016/j.apergo.2015.03.014

[21] Luca Di Persio and Oleksandr Honchar. 2017. Recurrent neural networks approach to the financial forecast of Google assets. *International journal of Mathematics and Computers in simulation* 11 (2017), 7–13.

[22] Paul Ekman. 1992. An Argument for Basic Emotions. *Cognition and Emotion* 6, 3-4 (1992), 169–200.

[23] Empatica. 2022. E4 Wristband | Real-time Physiological Signals | Wearable PPG, EDA, Temperature, Motion Sensors. https://www.empatica.com/research/e4, last accessed on 13-MARCH-2022.

[24] Poul O Fanger et al. 1970. *Thermal comfort. Analysis and applications in environmental engineering.* Danish Technical Press., Kopenhagen, Denmark.

[25] Veronika Földváry Ličina, Toby Cheung, Hui Zhang, Richard de Dear, Thomas Parkinson, Edward Arens, Chungyoon Chun, Stefano Schiavon, Maohui Luo, Gail Brager, Peixian Li, Soazig Kaam, Michael A. Adebamowo, Mary Myla Andamon, Francesco Babich, Chiheb Bouden, Hana Bukovianska, Christhina Candido, Bin Cao, Salvatore Carlucci, David K. W. Cheong, Joon-Ho Choi, Malcolm Cook, Paul Cropper, Max Deuble, Shahin Heidari, Madhavi Indraganti, Quan Jin, Hyojin Kim, Jungsoo Kim, Kyle Konis, Manoj K. Singh, Alison Kwok, Roberto Lamberts, Dennis Loveday, Jared Langevin, Sanyogita Manu, Cornelia Moosmann, Fergus Nicol, Ryozo Ooka, Nigel A. Oseland, Lorenzo Pagliano, Dušan Petráš, Rajan Rawal, Ramona Romero, Hom Bahadur Rijal, Chandra Sekhar, Marcel Schweiker, Federico Tartarini, Shin-ichi Tanabe, Kwok Wai Tham, Despoina Teli, Jorn Toftum, Linda Toledo, Kazuyo Tsuzuki, Renata De Vecchi, Andreas Wagner, Zhaojun Wang, Holger Wallbaum, Lynda Webb, Liu Yang, Yingxin Zhu, Yongchao Zhai, Yufeng Zhang, and Xiang Zhou. 2018. Development of the ASHRAE Global Thermal Comfort Database II. *Building and Environment* 142 (Sept. 2018), 502–512. https://doi.org/10.1016/j.buildenv.2018.06.022

[26] Jonathan Francis, Matias Quintana, Nadine von Frankenberg, Sirajum Munir, and Mario Bergés. 2019. OccuTherm: Occupant Thermal Comfort Inference Using Body Shape Information. In *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation* (New York, NY, USA) *(BuildSys '19)*. Association for Computing Machinery, New York, NY, USA, 81–90. https://doi.org/10.1145/3360322.3360858

[27] Steven M. Frank, Srinivasa N. Raja, Christian F. Bulcao, and David S. Goldstein. 1999. Relative contribution of core and cutaneous temperatures to thermal comfort and autonomic responses in humans. *Journal of Applied Physiology* 86, 5 (1999), 1588–1593. https://doi.org/10.1152/jappl.1999.86.5.1588 arXiv:https://doi.org/10.1152/jappl.1999.86.5.1588 PMID: 10233122.

[28] Monika Frontczak and Pawel Wargocki. 2011. Literature Survey on How Different Factors Influence Human Comfort in Indoor Environments. *Building and Environment* 46, 4 (2011), 922–937. https://doi.org/10.1016/j.buildenv.2010.10.021

[29] A. Pharo Gagge and Richard R. Gonzalez. 2011. *Mechanisms of Heat Exchange: Biophysics and Physiology.* John Wiley and Sons, Ltd, Hoboken, New Jersey, USA, 45–84. https://doi.org/10.1002/cphy.cp040104 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/cphy.cp040104

[30] B Griefahn and C Künemund. 2001. The Effects of Gender, Age, and Fatigue on Susceptibility to Draft Discomfort. *Journal of Thermal Biology* 26, 4 (Sept. 2001), 395–400. https://doi.org/10.1016/S0306-4565(01)00050-X

[31] Slava Guergova and André Dufour. 2011. Thermal Sensitivity in the Elderly: A Review. *Ageing Research Reviews* 10, 1 (2011), 80–92. https://doi.org/10.1016/j.arr.2010.04.009

[32] S. Hartwig, M. Schelling, C. v. Onzenoodt, P.-P. Vázquez, P. Hermosilla, and T. Ropinski. 2022. Learning Human Viewpoint Preferences from Sparsely Annotated Models. *Computer Graphics Forum* 41, 6 (2022), 453–466. https://doi.org/10.1111/cgf.14613 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14613

[33] Sebastian Hartwig, Christian van Onzenoodt, Pedro Hermosilla, and Timo Ropinski. 2023. ClusterNet: A Perception-Based Clustering Model for Scattered Data. *ArXiv* abs/2304.14185 (2023). https://api.semanticscholar.org/CorpusID:258352838

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, New York, NY, USA, 770–778.

[35] Lei He, Chang-fu Zong, and Chang Wang. 2012. Driving Intention Recognition and Behaviour Prediction Based on a Double-Layer Hidden Markov Model. *J. Zhejiang Univ. - Sci. C* 13, 3 (March 2012), 208–217. https://doi.org/10.1631/jzus.C11a0195

[36] Thomas Hirn, Alexander Kirmas, Damian Backes, and Lutz Eckstein. 2021. Impact of different radiation types on thermal comfort modelling. *Comfort Congress 2021* (2021), 6 pages.

[37] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[38] Charlie Huizenga, Zhang Hui, and Edward Arens. 2001. A Model of Human Physiology and Comfort for Assessing Complex Thermal Environments. *Building and Environment* 36, 6 (July 2001), 691–699. https://doi.org/10.1016/S0360-1323(00)00061-5

[39] Rasheed Hussain and Sherali Zeadally. 2019. Autonomous Cars: Research Results, Issues, and Future Challenges. *IEEE Communications Surveys and Tutorials* 21, 2 (2019), 1275–1313. https://doi.org/10.1109/COMST.2018.2869360

[40] PCE Instruments. 2022. Datenlogger PCE-WB 20SD. https://www.pce-instruments.com/deutsch/messtechnik/messgeraete-fuer-alle-parameter/datenlogger-pce-instruments-datenlogger-pce-wb-20sd-det_377956.htm?_list=qr.art&_listpos=7, last accessed on 29 JULY 2022.

[41] Pascal Jansen, Mark Colley, and Enrico Rukzio. 2022. A Design Space for Human Sensor and Actuator Focused In-Vehicle Interaction Based on a Systematic Literature Review. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 56 (jul 2022), 51 pages. https://doi.org/10.1145/3534617

[42] Prageeth Jayathissa, Matias Quintana, Mahmoud Abdelrahman, and Clayton Miller. 2020. Humans-as-a-sensor for buildings—intensive longitudinal indoor comfort models. *Buildings* 10, 10 (2020), 174.

[43] Jichao Jiao, Libin Yuan, Weihua Tang, Zhongliang Deng, and Qi Wu. 2017. A Post-Rectification Approach of Depth Images of Kinect v2 for 3D Reconstruction of Indoor Scenes. *ISPRS International Journal of Geo-Information* 6 (11 2017), 349. https://doi.org/10.3390/ijgi6110349

[44] Shenglan Jing, Baizhan Li, Meilan Tan, and Hong Liu. 2013. Impact of Relative Humidity on Thermal Comfort in a Warm Environment. *Indoor and Built Environment* 22, 4 (Aug. 2013), 598–607. https://doi.org/10.1177/1420326X12447614

[45] Lou Jost. 2006. Entropy and diversity. *Oikos* 113, 2 (2006), 363–375.

[46] S. Karjalainen. 2007. Gender Differences in Thermal Comfort and Use of Thermostats in Everyday Thermal Environments. *Building and Environment* 42, 4 (2007), 1594–1603. https://doi.org/10.1016/j.buildenv.2006.01.009

[47] S. Karjalainen. 2012. Thermal Comfort and Gender: A Literature Review. *Indoor Air* 22, 2 (2012), 96–109. https://doi.org/10.1111/j.1600-0668.2011.00747.x

[48] Brad Kelechava. 2021. *ANSI/ASHRAE 55-2020: Thermal Environmental Conditions for Human Occupancy.* Technical Report. American Society of Heating, Refrigerating and Air-Conditioning.

[49] Gwangbin Kim, Jieun Lee, Dohyeon Yeo, Eunsol An, and SeungJun Kim. 2023. Physiological Indices to Predict Driver Situation Awareness in VR. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing* (<conf-loc>, <city>Cancun, Quintana Roo</city>, <country>Mexico</country>, </conf-loc>) *(UbiComp/ISWC '23 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 40–45. https://doi.org/10.1145/3594739.3610687

[50] Im Langel. 2022. Datenblatt für Hygrometer PCE-WB 20SD. , 2 pages. https://www.pce-instruments.com/deutsch/api/getartfile?_fnr=891730&_dsp=inline, last accessed on 13-MARCH-2022.

[51] Jared Langevin. 2019. Longitudinal Dataset of Human-Building Interactions in U.S. Offices. *Scientific Data* 6, 1 (Nov. 2019), 288. https://doi.org/10.1038/s41597-019-0273-5

[52] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. 2023. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192* (2023).

[53] Aleksandra Lipczynska, Asit Mishra, and Stefano Schiavon. 2020. Experimental Evaluation of the Effect of Body Mass on Thermal Comfort Perception. 11th Windsor Conference on Thermal Comfort, Windsor, UK, 403–415.

[54] Shichao Liu, Stefano Schiavon, Hari Prasanna Das, Ming Jin, and Costas J. Spanos. 2019. Personal Thermal Comfort Models with Wearable Sensors. *Building and Environment* 162 (Sept. 2019), 106281. https://doi.org/10.1016/j.buildenv.2019.106281

[55] Siliang Lu, Weilong Wang, Shihan Wang, and Erica Cochran Hameen. 2019. Thermal Comfort-Based Personalized Models with Non-Intrusive Sensing Technique in Office Buildings. *Applied Sciences* 9, 9 (Jan. 2019), 1768. https://doi.org/10.3390/app9091768

[56] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[57] Silvia Angela Mansi, Giovanni Barone, Cesare Forzano, Ilaria Pigliautile, Maria Ferrara, Anna Laura Pisello, and Marco Arnesano. 2021. Measuring Human Physiological Indices for Thermal Comfort Assessment through Wearable Devices: A Review. *Measurement* 183 (Oct. 2021), 109872. https://doi.org/10.1016/j.measurement.2021.109872

[58] Dietrich Manstetten, Frank Beruscha, Hans-Joachim Bieg, Fanny Kobiela, Andreas Korthauer, Wolfgang Krautter, and Claus Marberger. 2020. The Evolution of Driver Monitoring Systems: A Shortened Story on Past, Current and Future Approaches How Cars Acquire Knowledge About the Driver's State. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, Oldenburg Germany, 1–6. https://doi.org/10.1145/3406324.3425896

[59] Haomin Mao, Shuhei Tsuchida, Yongbeom Kim, Rintaro Kanada, Takayuki Hori, Tsutomu Terada, and Masahiko Tsukamoto. 2021. A Thermal Comfort Estimation Method by Wearable Sensors. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*. ACM, Virtual Event Republic of Korea, 603–610. https://doi.org/10.1145/3412841.3441941

[60] Manuel Martin, Alina Roitberg, Monica Haurilet, Matthias Horne, Simon Reiß, Michael Voit, and Rainer Stiefelhagen. 2019. Drive&Act: A Multi-Modal Dataset for Fine-Grained Driver Behavior Recognition in Autonomous Vehicles. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Seoul, Korea (South), 2801–2810. https://doi.org/10.1109/ICCV.2019.00289

[61] Larissa Arakawa Martins, Veronica Soebarto, and Terence Williamson. 2022. A systematic review of personal thermal comfort models. *Building and Environment* 207 (2022), 108502.

[62] Microsoft. 2022. Kinect for Windows. https://docs.microsoft.com/en-us/windows/apps/design/devices/kinect-for-windows, last accessed on 29 JUNE 2022.

[63] Brendan Morris, Anup Doshi, and Mohan Trivedi. 2011. Lane Change Intent Prediction for Driver Assistance: On-Road Design and Evaluation. In *2011 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, New York, NY, USA, 895–901. https://doi.org/10.1109/IVS.2011.5940538

[64] Salman Muhammad, Hyunkyu Jang, Youngtae Noh, Seungwan Jin, Dayoung Jeong, Hoyoung Choi, Kyungsik Han, and Hyangmi Kim. 2023. A Contactless and Non-Intrusive System for Driver's Stress Detection. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing*. Association for Computing Machinery, New York, NY, USA, 58–62.

[65] Surabhi S. Nath, Vishaal Udandarao, and Jainendra Shukla. 2021. It's LeVAsa Not LevioSA! Latent Encodings for Valence-Arousal Structure Alignment. In *8th ACM IKDD CODS and 26th COMAD* (Bangalore, India) *(CODS COMAD 2021)*. Association for Computing Machinery, New York, NY, USA, 238–242. https://doi.org/10.1145/3430984.3431037

[66] M. Nilsson, Serge Thill, and Tom Ziemke. 2015. Action and intention recognition in human interaction with autonomous vehicles.

[67] D. J. Oborne. 1978. Passenger Comfort — an Overview. *Applied Ergonomics* 9, 3 (Sept. 1978), 131–136. https://doi.org/10.1016/0003-6870(78)90002-9

[68] Temitayo Olugbade, Youngjun Cho, Zak Morgan, Mohamed Abd El Ghani, and Nadia Bianchi-Berthouze. 2021. Toward Intelligent Car Comfort Sensing: New Dataset and Analysis of Annotated Physiological Metrics. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 1–8. https://doi.org/10.1109/ACII52823.2021.9597393

[69] Keiron O'Shea and Ryan Nash. 2015. An Introduction to Convolutional Neural Networks. arXiv:1511.08458 [cs]

[70] Thomas Parkinson, Federico Tartarini, Veronika Földváry Ličina, Toby Cheung, Hui Zhang, Richard de Dear, Peixian Li, Edward Arens, Chungyoon Chun, Stefano Schiavon, et al. 2022. ASHRAE global database of thermal comfort field measurements. *Methods* 2022 (2022), 07–15.

[71] Ken Parsons. 2002. *Human Thermal Environments: The Effect of Hot, Moderate And.* CRC Press, London, 44–49.

[72] Alexandra Galina Petre, James Brusey, and Ross Wilkins. 2019. An Examination of Comfort and Sensation for Manual and Automatic Controls of the Vehicle HVAC System. In *Automotive Technical Papers*. SAE International. https://doi.org/10.4271/2019-01-5005

[73] PyTorch. 2022. PyTorch ResNet. https://pytorch.org/vision/stable/models/resnet.html, last accessed on 29 JUNE 2022.

[74] Matias Quintana, Mahmoud Abdelrahman, Mario Frei, Federico Tartarini, and Clayton Miller. 2021. Longitudinal Personal Thermal Comfort Preference Data in the Wild. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems* (Coimbra, Portugal) *(SenSys '21)*. Association for Computing Machinery, New York, NY, USA, 556–559. https://doi.org/10.1145/3485730.3493693

[75] Matias Quintana, Stefano Schiavon, Federico Tartarini, Joyce Kim, and Clayton Miller. 2023. Cohort comfort models — Using occupant's similarity to predict personal thermal preference with less data. *Building and Environment* 227 (2023), 109685. https://doi.org/10.1016/j.buildenv.2022.109685

[76] N. L. Ramanathan. 1964. A New Weighting System for Mean Surface Temperature of the Human Body. *Journal of Applied Physiology* 19, 3 (May 1964), 531–533. https://doi.org/10.1152/jappl.1964.19.3.531

[77] Christian Rommelfanger, Christoph van Treeck, and Louis Fischer. 2020. Evaluation of Thermal Comfort in the Development of Air Conditioning. *ATZ worldwide* 122, 7 (2020), 42–47.

[78] R. Sathya and Annamma Abraham. 2013. Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification. *IJARAI* 2, 2 (2013), 34–38. https://doi.org/10.14569/IJARAI.2013.020206

[79] Marcel Schweiker, Amar Abdul-Zahra, Maíra André, Farah Atrash, Hanan Al-Khatri, Rea Alprianti, Hayder Alsaad, Rucha Amin, Eleni Ampatzi, Alpha Yacob Arsano, Azadeh Montazami, Elie Azar, Bahareh Bannazadeh, Amina Batagarawa, Susanne Becker, Carolina Buonocore, Bin Cao, Joon-Ho Choi, Chungyoon Chun, and Mahsa Zomorodian. 2019. The Scales Project, a Cross-National Dataset on the Interpretation of Thermal Perception Scales. *Scientific Data* 6 (Nov. 2019), 6 pages. https://doi.org/10.1038/s41597-019-0272-6

[80] scikit. 2022. scikit-learn Machine Learning in Python. https://scikit-learn.org/stable/, last accessed on 15 MAY 2022.

[81] Sefik Ilkin Serengil and Alper Ozpinar. 2020. LightFace: A Hybrid Deep Face Recognition Framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE, IEEE, New York, NY, USA, 23–27. https://doi.org/10.1109/ASYU50717.2020.9259802

[82] Sefik Ilkin Serengil and Alper Ozpinar. 2021. HyperExtended LightFace: A Facial Attribute Analysis Framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*. IEEE, IEEE, New York, NY, USA, 1–4. https://doi.org/10.1109/ICEET53442.2021.9659697

[83] Soo Young Sim, Myung Jun Koh, Kwang Min Joo, Seungwoo Noh, Sangyun Park, Youn Ho Kim, and Kwang Suk Park. 2016. Estimation of Thermal Sensation Based on Wrist Skin Temperatures. *Sensors* 16, 4 (April 2016), 420. https://doi.org/10.3390/s16040420

[84] Nivethitha Somu, Anirudh Sriram, Anupama Kowli, and Krithi Ramamritham. 2021. A Hybrid Deep Transfer Learning Strategy for Thermal Comfort Prediction in Buildings. *Building and Environment* 204 (Oct. 2021), 108133. https://doi.org/10.1016/j.buildenv.2021.108133

[85] Annika Stampf, Mark Colley, and Enrico Rukzio. 2022. Towards Implicit Interaction in Highly Automated Vehicles - A Systematic Literature Review. *Proc. ACM Hum.-Comput. Interact.* 6, MHCI, Article 191 (sep 2022), 21 pages. https://doi.org/10.1145/3546726

[86] Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. 2022. Human Action Recognition From Various Data Modalities: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022), 1–20. https://doi.org/10.1109/TPAMI.2022.3183112

[87] Wang Tao, Yuzhi Chen, Xingchen Yan, Wenyong Li, and Dong Shi. 2020. Assessment of Drivers' Comprehensive Driving Capability Under Man-Computer Cooperative Driving Conditions. *IEEE Access* PP (Aug. 2020), 1–1. https://doi.org/10.1109/ACCESS.2020.3016834

[88] Federico Tartarini and Stefano Schiavon. 2020. Pythermalcomfort: A Python Package for Thermal Comfort Research. *SoftwareX* 12 (July 2020), 100578. https://doi.org/10.1016/j.softx.2020.100578

[89] Ronald Terjung (Ed.). 2011. *Comprehensive Physiology* (first ed.). Wiley, Hoboken, New Jersey, USA, 3–5. https://doi.org/10.1002/cphy

[90] Ahmet Uğursal and Charles H. Culp. 2013. The Effect of Temperature, Metabolic Rate and Dynamic Localized Airflow on Thermal Comfort. *Applied Energy* 111 (Nov. 2013), 64–73. https://doi.org/10.1016/j.apenergy.2013.04.014

[91] Amin Ullah, Jamil Ahmad, Khan Muhammad, Muhammad Sajjad, and Sung Wook Baik. 2018. Action Recognition in Video Sequences Using Deep Bi-Directional LSTM With CNN Features. *IEEE Access* 6 (2018), 1155–1166. https://doi.org/10.1109/ACCESS.2017.2778011

[92] K.B. Velt and H.A.M. Daanen. 2017. Thermal sensation and thermal comfort in changing environments. *Journal of Building Engineering* 10 (2017), 42–46. https://doi.org/10.1016/j.jobe.2017.02.004

[93] Michal Veselý and Wim Zeiler. 2014. Personalized Conditioning and Its Impact on Thermal Comfort and Energy Performance – A Review. *Renewable and Sustainable Energy Reviews* 34 (June 2014), 401–408. https://doi.org/10.1016/j.rser.2014.03.024

[94] Steffen Walter, Sascha Gruss, Markus Kächele, Friedhelm Schwenker, philipp Werner, Ayoub Al-Hamadi, Adriano Andrade, Gustavo Moreira, and Harald Traue. 2015. Data Fusion for Automated Pain Recognition. In *Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare*. ICST, Istanbul, Turkey, 261–264. https://doi.org/10.4108/icst.pervasivehealth.2015.259166

[95] Haiying Wang and Lu Liu. 2020. Experimental Investigation about Effect of Emotion State on People's Thermal Comfort. *Energy and Buildings* 211 (2020), 109789. https://doi.org/10.1016/j.enbuild.2020.109789

[96] Qiong Wang, Jingyu Yang, Mingwu Ren, and Yujie Zheng. 2006. Driver Fatigue Detection: A Survey. In *2006 6th World Congress on Intelligent Control and Automation*, Vol. 2. IEEE, New York, NY, USA, 8587–8591. https://doi.org/10.1109/WCICA.2006.1713656

[97] Jannik Wolf, Marco Wiedner, Mohamed Kari, and David Bethge. 2021. HMInference: Inferring Multimodal HMI Interactions in Automotive Screens. In *13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, Leeds United Kingdom, 230–236. https://doi.org/10.1145/3409118.3475145

[98] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023. Human Preference Score: Better Aligning Text-to-Image Models with Human Preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, New York, NY, USA, 2096–2105.

[99] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. *arXiv preprint arXiv:2304.05977* (2023).

[100] Bin Yang, Xiaogang Cheng, Dengxin Dai, Thomas Olofsson, Haibo Li, and Alan Meier. 2019. Real-Time and Contactless Measurements of Thermal Discomfort Based on Human Poses for Energy Efficient Control of Buildings. *Building and Environment* 162 (Sept. 2019),

106284. https://doi.org/10.1016/j.buildenv.2019.106284

[101] Bin Yang, Xiaojing Li, Yingzhen Hou, Alan Meier, Xiaogang Cheng, Joon-Ho Choi, Faming Wang, Huan Wang, Andreas Wagner, Da Yan, Angui Li, Thomas Olofsson, and Haibo Li. 2020. Non-Invasive (Non-Contact) Measurements of Human Thermal Physiology Signals and Thermal Comfort/Discomfort Poses -A Review. *Energy and Buildings* 224 (2020), 110261. https://doi.org/10.1016/j.enbuild.2020.110261

[102] Ye Yao, Zhiwei Lian, Weiwei Liu, and Qi Shen. 2008. Experimental Study on Physiological Responses and Thermal Comfort under Various Ambient Temperatures. *Physiology & Behavior* 93, 1 (2008), 310–321. https://doi.org/10.1016/j.physbeh.2007.09.012

[103] Bo Yin, Natalie Ruiz, Fang Chen, and M. Asif Khawaja. 2007. Automatic Cognitive Load Detection from Speech Features. In *Proceedings of the 2007 Conference of the Computer-Human Interaction Special Interest Group (CHISIG) of Australia on Computer-Human Interaction: Design: Activities, Artifacts and Environments - OZCHI '07*. ACM Press, Adelaide, Australia, 249. https://doi.org/10.1145/1324892.1324946

[104] Hiroki Yoshikawa, Akira Uchiyama, Yuki Nishikawa, and Teruo Higashino. 2019. Combining a Thermal Camera and a Wristband Sensor for Thermal Comfort Estimation. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. ACM, London United Kingdom, 238–241. https://doi.org/10.1145/3341162.3343813

[105] Chaoyun Zhang, Rui Li, Woojin Kim, Daesub Yoon, and Paul Patras. 2020. Driver Behavior Recognition via Interwoven Deep Convolutional Neural Nets With Multi-Stream Inputs. *IEEE Access* 8 (2020), 191138–191151. https://doi.org/10.1109/ACCESS.2020.3032344

[106] Qinglian Zhang, Kai Tang, Xiaolei Liu, Qianqian Wang, and Kazushige Ouchi. 2021. A Frown-Based Thermal Comfort Detection Method of Facial Emotion Recognition. In *2021 The 4th International Conference on Machine Learning and Machine Intelligence*. ACM, Hangzhou China, 1–7. https://doi.org/10.1145/3490725.3490726

[107] Wei-Long Zheng, Wei Liu, Yifei Lu, Bao-Liang Lu, and Andrzej Cichocki. 2019. EmotionMeter: A Multimodal Framework for Recognizing Human Emotions. *IEEE Transactions on Cybernetics* 49, 3 (March 2019), 1110–1122. https://doi.org/10.1109/TCYB.2018.2797176

A CROSS-VALIDATION OF THE RECURRENT NETWORK CLASSIFIER

In our experiment from Section 7.4, we used data from 16 participants during the training of our models. Then, we evaluate performance on the data from the remaining 2 participants. Now, we investigate the ability of our network to generalize to unseen data. Therefore, we adopt 9 fold cross-validation, repeating the optimization of our classifier. To do so, we create 9 folds with two participants each for evaluation. The remaining participants are used for training. Then, we train 9 classifiers, using the same hyperparameters overall training runs. In Table 8, we report individual performance measures, as well as the mean and standard deviation for all runs. The mean accuracy of our classifier is $\kappa_7 = 54.8\% \pm 1.7\%$, $\kappa_3 = 78.2\% \pm 1.7\%$ and κ_2 $72.3\% \pm 0.7\%$.

Table 8. Cross-validation results over 9 folds with 16 participants in the training split and remaining 2 participants for testing. In the last two rows, we report the mean and std for all metrics.

Run	Classification		
	7-point Acc	3-point Acc	2-point Acc
0	54.4%	79.4%	70.9%
1	52.3%	76.0%	72.8%
2	57.0%	79.8%	72.8%
3	54.3%	77.7%	72.1%
4	57.8%	81.1%	72.9%
5	55.6%	79.0%	71.9%
6	54.7%	77.9%	72.5%
7	53.8%	77.9%	71.4%
8	52.9%	75.2%	73.2%
Mean	54.8%	78.2%	72.3%
Std	1.7%	1.7%	0.7%
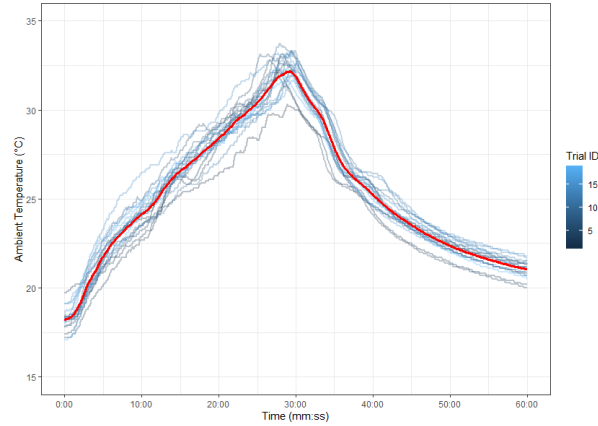
Fig. 7. Ambient temperature conditions for the full recording duration during the INDOOR study, grouped by trial ID. The mean heating profile is included (red).

B AMBIENT TEMPERATURE

Figure 7 illustrates the ambient temperature profile that each participant was exposed to. It can be observed that the defined temperature ranges were met in each trial during the recording sessions.

C DATA PREPARATION AND AUGMENTATION

Firstly, outliers were defined as outside of the mean with three times the standard deviation. This was also done per label group for the ambient temperature. That means that for a given label, data points with an ambient temperature of $+/-3 * std$ of the mean were excluded. This was done to filter falsely attributed labels. Then, we used one-hot encoding for categorical features and overall data down-sampling for the RF classifier.

For the LSTM and CNN-LSTM models, the same outlier filtering, one-hot-encoding, and down-sampling steps were employed. However, as sequences are expected as input to the LSTM models, a sliding window approach was implemented, where, for every index in the created data frame, a sequence of length n is created, such that, for instance, a window size of $w = 30$ would result in the following data frame scheme.

Simple data augmentation for continuous features was also implemented. Gaussian noise sampled from a Gaussian distribution with the parameters $\mu = 0.00$ and $\delta = 0.30$ was added to the continuous variables through element-wise addition. For image data, a more extensive data augmentation scheme was applied, as RGB frames were recorded at a size of 1920x1080, which is too large for efficient training. Thus, first, a central 1000x1000 crop was extracted and resized to 224x224. The resized image is then randomly rotated up to 5°, randomly flipped horizontally, and then normalized so that all three RGB channels are given as values between 0 and 1

Finally, as the thermal sensation scale has an ordinal scaling level, providing simple integer targets as prevalent in the raw dataset ignores the rank information. Therefore, the labels were changed according to a scheme proposed by Cheng et al. [13]. Ordinal labels are transformed into binary vectors of size k (here 7), where k is the number of ranks given on the ordinal scale. The binary vectors are instantiated as zero-vectors and then filled with ones from left to right, depending on the rank of the ordinal label. The resulting label transformations for the thermal sensation scale (seven-point scale) were:

- $-3 \longrightarrow [1, 0, 0, 0, 0, 0, 0]$
- $-2 \longrightarrow [1, 1, 0, 0, 0, 0, 0]$
- $-1 \longrightarrow [1, 1, 1, 0, 0, 0, 0]$

- $0 \longrightarrow [1, 1, 1, 1, 0, 0, 0]$
- $1 \longrightarrow [1, 1, 1, 1, 1, 0, 0]$
- $2 \longrightarrow [1, 1, 1, 1, 1, 1, 0]$
- $3 \longrightarrow [1, 1, 1, 1, 1, 1, 1]$

The resulting binary vectors were used as labels for the LSTM and CNN-LSTM models. Split sizes were predefined, so data from 16 participants were used for training, and the remaining data files for validation/testing.

In-Lab study   In-Vehicle study
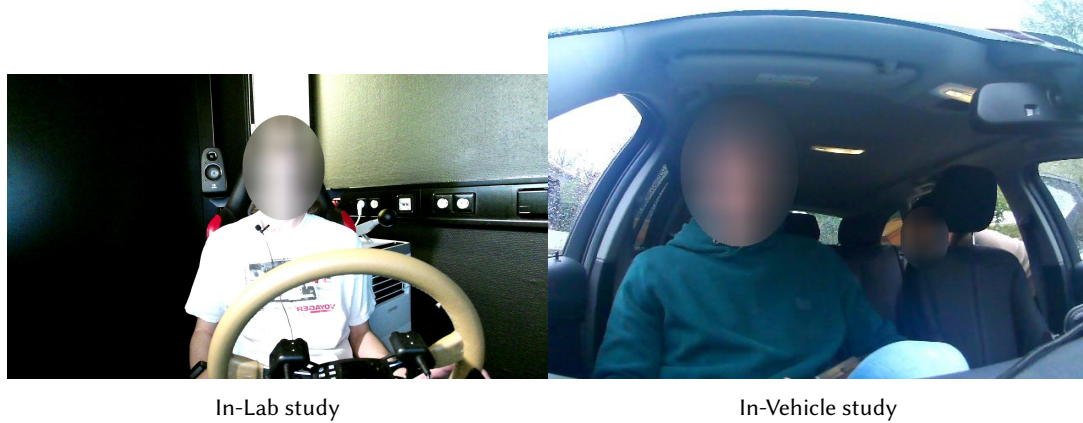
Fig. 8. Left: Image taken from the setup during the in-lab study. Right: Image taken from inside the BMW 3 during the in-vehicle study.

D LSTM HYPERPARAMETER SEARCH

For the LSTM model (without visual features), initial trial runs on the data showed severe over-fitting symptoms (i.e., training loss steadily declined while validation loss increased continuously). The model was then modified such that the hidden state size of the LSTM cells and the input sequence length could be included in the hyperparameter search. The optimization parameters for the initial LSTM model, therefore, included batch size, learning rate, sequence length, down-sampling rate, drop-out, number of LSTM layers, and the hidden state size. A training script using the parameter ranges shown in Table 9 was employed, and each configuration was run initially run for 20 epochs. The mean-squared error (MSE) was used as a loss function. The highest validation accuracy using the fore-named parameters was achieved at 0.62 with the configuration: learning rate=0.00001, hidden states=64, sequence length=300 (i.e., 10 seconds), batch size=16, layers=2, and drop-out=0.5. The same feature combination was used for all the runs. After analyzing the loss graphs for the training and validation data, it became notable that improvement of the model decreases quickly after the learning rate decreases far below 0.00001. Moreover, over-fitting was less severe with smaller hidden state sizes ([32-64]). One reason for the quick drop in model improvement was assumed to be the learning rate decay, which was thus far implemented as a reduction by $x^{0.99999}$. To explore the learning rate decay's influence on the model improvement, it was also added to the optimization parameters.

Keeping the previously optimized parameters the same while varying only learning rate decay and down-sampling rate, additional training runs were started and trained for up to a maximum of 100 epochs or until over-fitting was detected. The number of training epochs was increased as a response to the low learning rate and decreased learning rate decay. Optimization of the remaining parameters resulted in a validation accuracy

Parameter	Range
batch size	{4, 8, 16}
learning rate	{0.001, 0.0001, 0.00005, 0.00001}
learning rate decay	{0.99999, 0.999999, 0.9999999}
hidden states	{10, 16, 32, 64, 96, 128, 256}
lstm layers	{1, 2, 3}
sequence length	{10, 50, 100, 150, 300, 600}
drop-out	{0.2, 0.5}
features	{clothing, radiation temperature, ambient temperature, ambient humidity, GSR, heart rate, skin temperature}

Table 9. The parameter space used for grid search optimization of the LSTM and CNN-LSTM models.

of 0.71 with the configuration learning rate=0.00001, hidden states=64, layers=2, down-sampling=10, sequence length=30 (i.e., 10 seconds), batch size=16, and drop-out=0.5, and learning rate decay=0.9999999.

E   FEATURE PERMUTATION IMPORTANCE
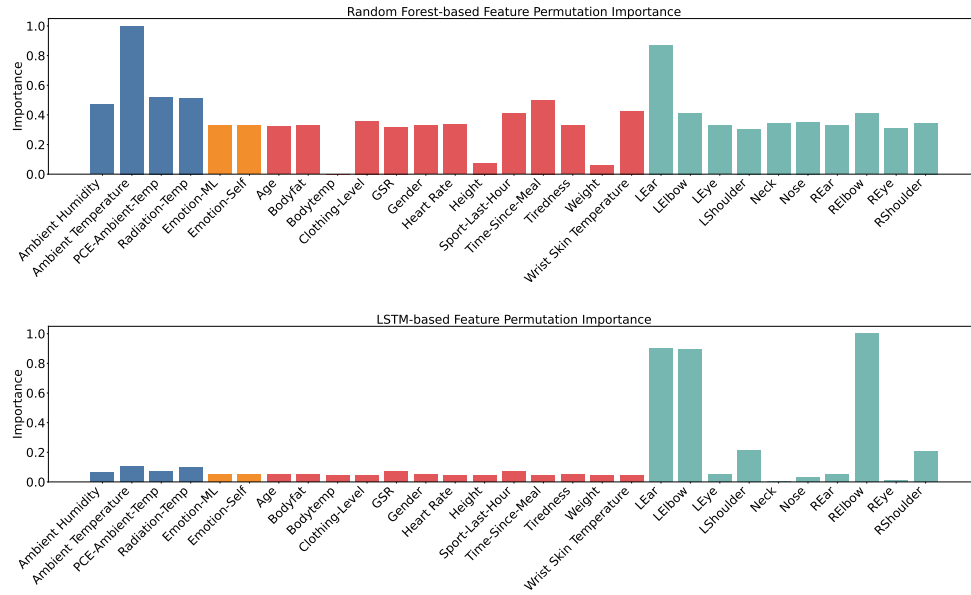
Fig. 9. Feature permutation importance is computed for both models, (top) Random Classifier and (bottom) LSTM-base model. We group by physical, psychological, physiological and OpenPose [11] features. While the results for the random forest indicate more importance to the physical and some physiological features, for the LSTM-based, a similar conclusion can be drawn. However, the differences are only marginal. However, features derived from RGB Frames using OpenPose [11] seem to matter only in some cases (left ear for random forest and left ear and elbows for the LSTM model). We attribute this inconsistency to occluded body parts in the recordings due to sitting posture.

F EXTENDED FEATURE COMBINATION STUDY

In this experiment, we investigate the importance of different features. We train 255 classifiers using all possible feature combinations using the following 8 features: Body temperature (BT), PCE ambient temperature (PCE), heart rate (HR), galvanic skin response (GL), ambient temperature from Arduino sensor (AT), relative humidity from Arduino sensor (HU), radiation temperature (RT), wrist skin temperature (WS). We report performance results for classification and regression measures in Table 10, Table 11, Table 12, Table 13, Table 14 and Table 15. Additionally, we compute the absolute SHAP values [56], averaged for the INDOOR dataset, see Figure 10.

Table 10. Top 50 runs of feature combination study.

BT	PCE	HR	GL	AT	HU	RT	WS	7-point Acc	3-point Acc	2-point Acc
			X	X	X		X	59.3%	83.5%	71.9%
			X		X	X	X	59.1%	78.4%	73.5%
			X			X	X	57.7%	79.1%	73.8%
	X		X		X		X	57.5%	78.4%	74.4%
			X		X		X	57.1%	79.6%	72.7%
	X	X	X	X			X	57.2%	80.3%	72.4%
	X	X	X	X	X		X	56.5%	80.3%	72.0%
	X		X				X	56.5%	79.4%	73.3%
	X	X	X			X	X	56.4%	81.5%	72.4%
			X		X	X	X	56.2%	79.2%	72.8%
	X		X	X			X	56.1%	79.2%	72.9%
	X	X	X	X			X	56.0%	79.5%	73.2%
			X	X	X	X	X	55.9%	79.8%	72.1%
			X	X			X	55.8%	79.4%	72.8%
	X		X	X		X		55.5%	78.8%	72.7%
		X	X	X			X	55.5%	78.1%	73.3%
			X				X	55.5%	79.1%	70.9%
			X	X		X		55.4%	79.2%	72.3%
	X	X	X	X			X	55.3%	77.9%	73.4%
	X	X	X				X	55.2%	78.3%	72.6%
	X	X	X	X			X	54.8%	78.5%	72.4%
			X		X			54.5%	78.8%	72.1%
		X	X	X		X	X	54.4%	77.5%	72.9%
	X		X		X	X	X	54.4%	78.9%	71.0%
		X	X				X	54.3%	77.6%	72.0%
	X		X					54.3%	78.8%	69.9%
	X		X	X				54.1%	78.5%	71.6%
	X	X	X	X	X			54.1%	78.9%	71.7%
	X		X	X	X			53.9%	77.5%	71.7%
		X	X	X			X	53.8%	76.9%	72.3%
		X	X		X	X	X	53.8%	77.7%	71.2%
		X	X	X	X	X	X	53.6%	77.0%	72.6%
	X	X	X				X	53.3%	76.9%	72.5%
	X		X	X				53.3%	76.9%	71.7%
		X	X	X			X	52.9%	77.3%	70.7%
X	X		X	X	X		X	52.8%	76.4%	72.2%
	X	X	X			X		52.6%	75.6%	72.3%
	X		X			X		52.6%	76.6%	73.5%
			X	X				52.4%	75.4%	71.3%
	X	X	X	X		X		52.3%	76.8%	70.9%
	X		X	X				52.1%	74.4%	73.5%
	X		X	X	X			52.1%	80.5%	68.3%
		X	X			X	X	52.1%	76.5%	70.6%
			X					52.1%	74.7%	72.6%
		X	X	X	X		X	52.0%	76.4%	71.3%
	X	X	X	X		X	X	51.9%	75.9%	72.4%
	X		X		X		X	51.7%	76.3%	71.3%
		X	X	X	X	X	X	51.3%	77.2%	68.9%
		X	X			X		51.3%	74.1%	72.6%
			X	X	X			51.3%	78.2%	68.3%

Table 11. Top 51 - 100 runs of feature combination study.

| Features | | | | | | | | Classification | | |
BT	PCE	HR	GL	AT	HU	RT	WS	7-point Acc	3-point Acc	2-point Acc
		X		X				51.3%	74.0%	72.7%
X	X	X			X	X	X	51.2%	78.1%	69.5%
X		X	X					51.2%	74.6%	72.3%
		X	X	X				51.2%	75.2%	71.6%
	X	X	X		X	X	X	51.0%	74.3%	71.4%
X		X			X	X	X	50.9%	77.2%	69.9%
X	X	X	X	X			X	50.8%	76.8%	70.6%
			X	X	X			50.8%	75.3%	72.1%
	X	X		X		X	X	50.7%	76.4%	69.8%
X	X	X	X	X	X		X	50.7%	77.4%	69.9%
		X	X	X	X			50.6%	73.7%	73.0%
		X	X			X	X	50.6%	73.9%	72.3%
	X	X	X				X	50.6%	75.2%	70.9%
	X	X		X				50.5%	76.2%	71.8%
X	X	X	X		X	X	X	50.4%	74.7%	71.5%
	X	X	X	X				50.3%	72.8%	73.1%
X	X	X	X		X		X	50.3%	75.0%	71.3%
X		X	X		X	X	X	50.2%	75.0%	71.1%
		X	X	X	X			55.0%	78.0%	72.7%
X	X	X		X	X		X	50.0%	75.6%	70.8%
	X	X	X			X	X	49.8%	72.8%	73.2%
	X	X		X	X	X	X	49.8%	73.9%	70.7%
	X	X			X	X	X	49.8%	73.2%	71.7%
	X	X	X	X	X	X	X	49.8%	74.9%	70.2%
	X	X		X	X	X		49.6%	72.2%	73.1%
X		X				X	X	49.5%	72.1%	71.7%
	X	X			X		X	49.4%	73.5%	70.8%
	X	X					X	49.3%	72.8%	73.7%
	X		X			X	X	49.0%	71.6%	73.6%
	X	X	X					48.7%	71.5%	72.9%
	X		X	X	X			48.7%	78.5%	69.5%
			X			X	X	48.5%	70.8%	70.2%
	X				X	X	X	48.5%	71.5%	72.5%
X			X		X	X	X	48.2%	70.5%	73.8%
X	X	X	X	X	X	X	X	48.2%	73.4%	69.2%
	X	X			X	X	X	48.2%	70.8%	72.2%
X	X	X			X		X	48.2%	71.7%	71.4%
	X		X				X	48.1%	72.0%	73.3%
	X		X		X	X	X	48.0%	73.7%	70.8%
		X	X		X	X	X	48.0%	71.7%	72.4%
X	X	X		X			X	48.0%	75.1%	68.6%
X	X				X		X	47.9%	73.8%	69.0%
X					X		X	47.9%	74.7%	68.5%
	X						X	47.7%	70.6%	73.1%
X		X	X	X	X	X	X	47.7%	73.3%	68.9%
X	X	X		X		X	X	47.7%	71.9%	71.2%
		X	X	X	X			47.6%	74.7%	72.2%
X	X			X	X		X	47.6%	72.6%	66.9%
	X				X	X	X	47.5%	71.6%	71.7%
X	X			X	X		X	47.4%	72.5%	70.7%

Table 12. Top 101 - 150 runs of feature combination study.

BT	PCE	HR	GL	AT	HU	RT	WS	7-point Acc	3-point Acc	2-point Acc
X		X		X		X	X	47.3%	73.0%	70.4%
X			X	X			X	47.1%	71.2%	69.4%
		X	X	X	X	X		47.1%	78.1%	68.7%
X	X	X				X	X	47.1%	72.1%	68.9%
X			X	X		X	X	46.7%	71.0%	69.9%
X			X			X	X	46.6%	69.2%	73.0%
X		X			X	X	X	46.6%	70.1%	69.9%
X	X					X	X	46.5%	67.4%	70.1%
X		X		X	X		X	46.4%	73.9%	68.1%
X	X		X			X	X	46.3%	67.3%	71.0%
						X	X	46.3%	68.8%	73.8%
X	X	X	X				X	46.3%	69.4%	71.1%
X	X		X				X	45.9%	67.3%	70.0%
	X				X			45.7%	69.5%	74.9%
X		X		X		X		45.7%	69.0%	73.0%
	X	X	X	X	X	X		45.7%	77.0%	67.9%
	X	X		X	X			45.7%	68.5%	72.9%
	X					X	X	45.7%	69.0%	73.6%
X	X			X		X	X	45.4%	70.8%	69.7%
X	X	X					X	45.3%	69.7%	69.1%
X					X	X	X	45.3%	66.0%	72.1%
X	X				X	X	X	45.2%	69.4%	69.0%
X		X	X	X			X	45.2%	71.4%	68.4%
X		X	X			X	X	45.0%	69.4%	71.2%
	X	X	X		X		X	44.9%	69.8%	70.8%
X	X	X		X	X	X	X	44.7%	71.1%	68.8%
X	X	X		X				44.4%	68.3%	71.8%
X			X		X			44.3%	71.2%	68.9%
X	X	X	X			X	X	44.2%	69.5%	68.8%
X		X			X		X	44.2%	73.6%	68.1%
X	X		X		X	X	X	44.2%	69.6%	70.7%
	X		X				X	44.2%	70.1%	71.9%
	X		X		X	X		44.1%	70.3%	66.1%
X		X	X		X		X	44.1%	62.1%	69.0%
X	X	X	X	X		X	X	43.6%	68.7%	68.7%
	X				X		X	43.6%	69.8%	69.5%
X			X	X		X	X	43.5%	66.8%	69.9%
X		X	X	X	X	X		43.3%	67.8%	70.5%
X			X	X	X	X	X	42.8%	66.8%	71.0%
X	X	X		X	X			42.3%	67.2%	71.6%
			X		X	X	X	42.1%	68.5%	70.4%
X		X	X	X		X	X	42.0%	75.6%	62.9%
X					X	X	X	42.0%	70.7%	66.9%
X	X		X	X	X			41.9%	66.6%	70.3%
X				X		X	X	41.9%	65.7%	69.5%
X				X	X			41.9%	66.1%	71.5%
X	X				X			41.7%	65.8%	62.8%
					X	X		41.6%	63.2%	71.1%
X	X			X		X		41.6%	66.6%	71.5%
X		X		X			X	41.5%	70.0%	68.8%

Table 13. Top 151 - 200 runs of feature combination study.

		Features						Classification		
BT	PCE	HR	GL	AT	HU	RT	WS	7-point Acc	3-point Acc	2-point Acc
					X	X	X	41.4%	67.3%	70.1%
X			X				X	41.4%	65.6%	66.8%
	X	X	X		X	X		41.3%	69.3%	69.1%
X	X			X	X	X	X	41.3%	72.6%	62.4%
X			X		X		X	41.3%	64.9%	61.0%
X					X			41.2%	64.1%	72.2%
X		X	X	X	X		X	40.9%	65.0%	65.7%
X	X	X	X	X	X	X		40.8%	65.8%	71.2%
	X	X			X	X		40.8%	68.1%	70.5%
X	X			X	X			40.7%	67.2%	69.5%
X		X	X	X	X			40.7%	65.9%	70.8%
X		X		X	X	X		40.6%	66.1%	70.8%
X			X	X	X			40.4%	67.6%	69.2%
X		X		X	X			40.4%	68.0%	68.1%
			X		X		X	40.4%	67.6%	64.7%
X		X	X		X			40.3%	64.0%	61.2%
X	X			X			X	40.2%	61.7%	61.2%
X	X	X	X	X	X			40.1%	63.7%	71.1%
X		X	X			X		40.1%	62.9%	68.6%
X	X		X	X			X	40.1%	63.5%	68.6%
X			X	X				40.0%	65.0%	69.2%
X					X	X		39.8%	63.0%	71.9%
X	X		X	X				39.7%	64.7%	69.5%
X	X		X	X	X	X		39.7%	64.6%	68.5%
		X			X		X	39.5%	64.7%	66.0%
					X		X	39.4%	64.7%	63.4%
X	X	X		X		X		39.4%	64.6%	68.8%
X			X	X			X	39.3%	74.1%	63.2%
X		X			X	X		39.3%	66.4%	64.5%
X	X	X			X	X		39.3%	65.0%	63.7%
X	X		X	X		X		39.3%	64.3%	69.5%
X					X	X	X	39.0%	63.5%	70.4%
X		X			X			39.0%	65.0%	60.3%
X		X	X	X				39.0%	65.6%	69.7%
	X	X				X		38.7%	67.9%	72.3%
X	X		X					38.6%	63.1%	69.3%
	X				X	X		38.6%	67.9%	67.4%
			X			X		38.5%	56.4%	73.3%
X			X	X				38.5%	63.8%	69.0%
		X			X	X		38.5%	62.2%	71.8%
X			X					38.5%	62.5%	69.1%
X	X	X	X	X				38.4%	65.5%	68.9%
		X			X	X		38.2%	58.2%	72.9%
X			X	X	X		X	38.1%	63.2%	63.6%
X	X				X	X		38.1%	60.8%	64.9%
X			X	X	X	X		38.0%	62.7%	67.8%
	X	X	X		X			38.0%	60.7%	70.3%
	X	X			X			37.8%	59.6%	68.5%
	X	X			X	X		37.8%	58.9%	70.5%
X	X	X			X			37.8%	55.7%	66.5%

Table 14. Top 201 - 250 runs of feature combination study.

			Features					Classification		
BT	PCE	HR	GL	AT	HU	RT	WS	7-point Acc	3-point Acc	2-point Acc
X			X	X		X		37.7%	61.2%	69.7%
			X		X			37.6%	65.7%	67.5%
X	X						X	37.5%	58.9%	70.1%
X	X	X	X		X	X		37.4%	61.6%	64.0%
X	X		X	X	X	X	X	37.4%	61.5%	67.4%
X		X	X		X	X		37.3%	56.2%	71.0%
	X		X		X			37.3%	59.4%	68.4%
						X		37.1%	56.4%	74.6%
X			X		X	X		37.1%	56.1%	60.2%
	X							37.1%	58.7%	73.1%
X	X		X		X	X		37.0%	60.2%	70.7%
	X	X						36.7%	55.9%	73.5%
X	X	X	X		X			36.6%	62.6%	68.5%
	X		X					36.4%	58.5%	73.7%
X	X		X	X		X	X	36.2%	55.4%	58.5%
	X				X			36.0%	58.5%	68.1%
X	X			X	X	X		36.0%	61.1%	69.4%
					X			36.0%	59.8%	63.3%
X	X	X					X	35.9%	60.7%	70.3%
	X	X	X					35.6%	55.9%	72.6%
	X	X					X	35.5%	56.6%	74.4%
	X	X	X				X	35.4%	62.8%	73.3%
X		X	X				X	35.3%	56.5%	69.2%
	X	X		X			X	34.8%	62.2%	62.3%
X	X		X		X			34.6%	55.6%	61.5%
	X	X					X	34.5%	57.2%	71.5%
X	X	X	X				X	34.3%	61.2%	67.5%
X	X	X	X	X			X	34.2%	58.6%	64.2%
X	X	X	X					34.2%	55.3%	65.3%
X			X				X	33.6%	55.3%	73.1%
X							X	33.3%	55.2%	72.1%
X	X	X						33.2%	56.5%	64.4%
	X					X		32.6%	60.6%	65.3%
X	X		X				X	32.5%	54.9%	67.9%
	X	X		X				32.3%	62.9%	64.4%
X	X						X	32.3%	50.9%	63.4%
X	X		X					31.6%	54.8%	62.5%
X	X	X		X	X	X		31.3%	57.0%	56.8%
X	X							30.6%	46.9%	60.2%
X		X				X		30.1%	43.1%	66.3%
	X							26.6%	29.5%	64.8%
X		X	X				X	26.1%	41.7%	60.8%
X		X						25.7%	32.7%	63.2%
X		X	X					25.0%	40.1%	58.8%
	X	X					X	24.7%	24.7%	63.5%
	X	X						24.7%	24.7%	63.5%
X			X					24.6%	24.6%	63.4%
			X				X	24.6%	24.6%	63.4%
			X					24.6%	24.6%	63.4%
X			X				X	24.6%	24.6%	63.4%

Table 15. Top 251 - 255 runs of feature combination study.

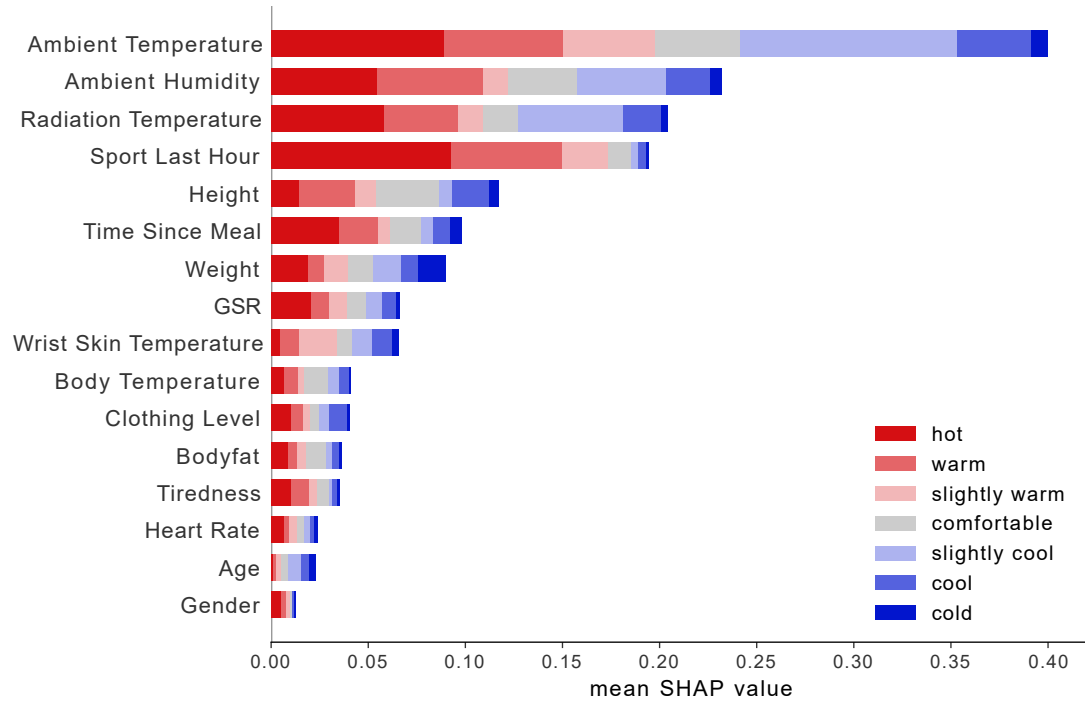|			Features					Classification		
BT	PCE	HR	GL	AT	HU	RT	WS	7-point Acc	3-point Acc	2-point Acc
							X	24.6%	27.8%	63.1%
	X						X	24.5%	24.5%	63.2%
X	X						X	24.5%	24.5%	63.2%
X							X	24.5%	24.5%	63.1%
X								24.5%	24.5%	63.1%

Fig. 10. SHAP values for sixteen features computed for the test split of the INDOOR dataset using the trained RF. For each feature and class, we report the mean absolute SHAP value.
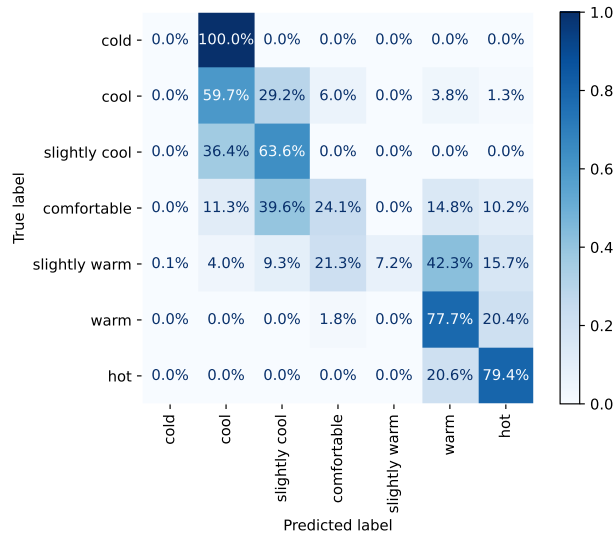
G CONFUSION MATRICES

Fig. 11.   Random Forest confusion matrix. *Cold* classes are misclassified entirely, and *Comfortable* classes are rarely classified correctly. Warmer labels are predicted with higher accuracy.
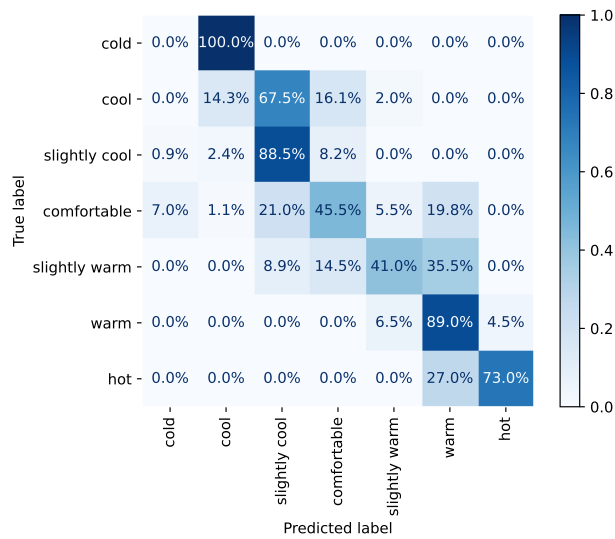

Fig. 12.   LSTM confusion matrix. *Comfortable* and *Cold* states are mostly wrongly classified. Warmer labels were classified more reliably.
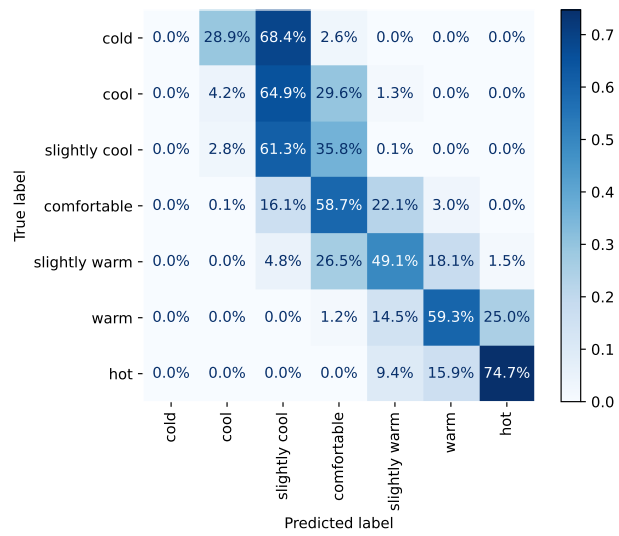
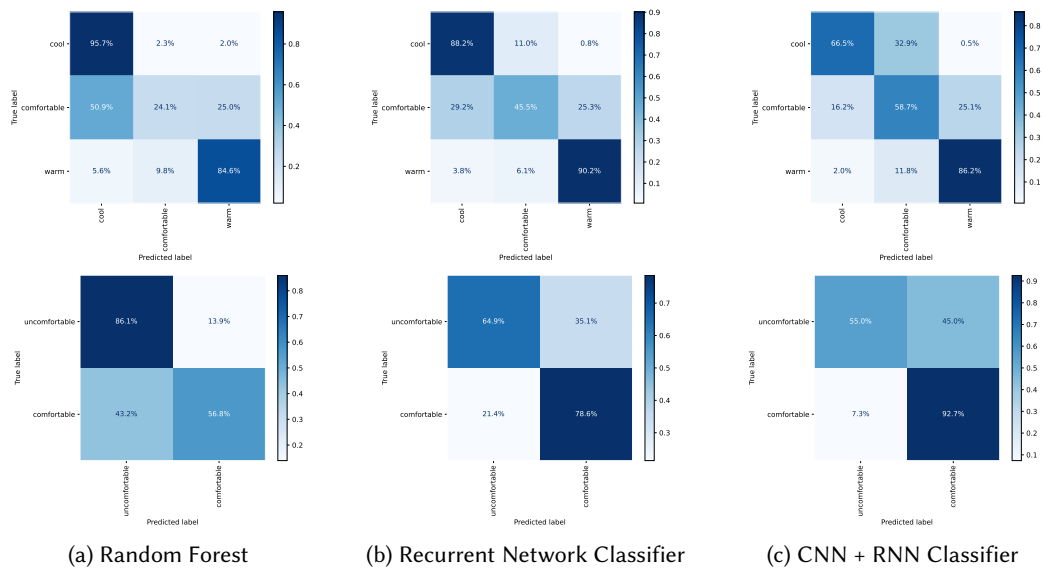Fig. 13. The CNN-LSTM confusion matrix. Most classes apart from *Hot* states are wrongly classified.

(a) Random Forest   (b) Recurrent Network Classifier   (c) CNN + RNN Classifier

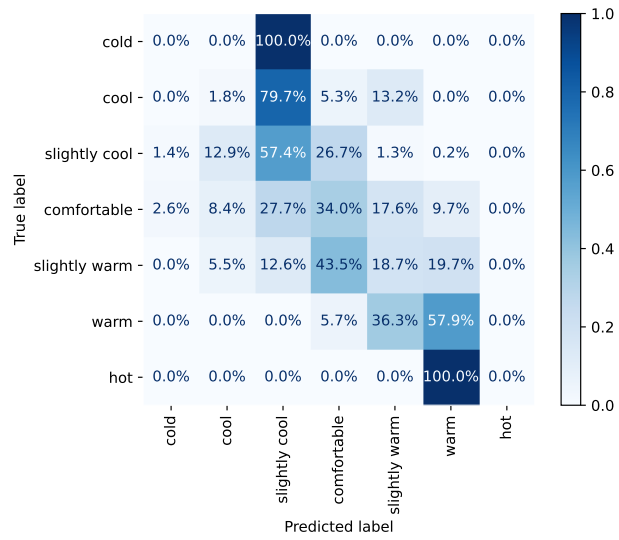Fig. 14. Confusion matrices for three-point and two-point classification.

Fig. 15. The confusion matrix resulting from PMV index calculated labels. It can be seen that all classes were mostly wrongly classified using the PMV index.

H THERMAL COMFORT STUDY INTRODUCTORY SLIDES

We provide the slides used to explain the procedure and the data logging interface to participants. Most slides included screenshots from the actual data logger GUI to ensure that explanations given during the introduction were supported by visualizations of the actual logging application.
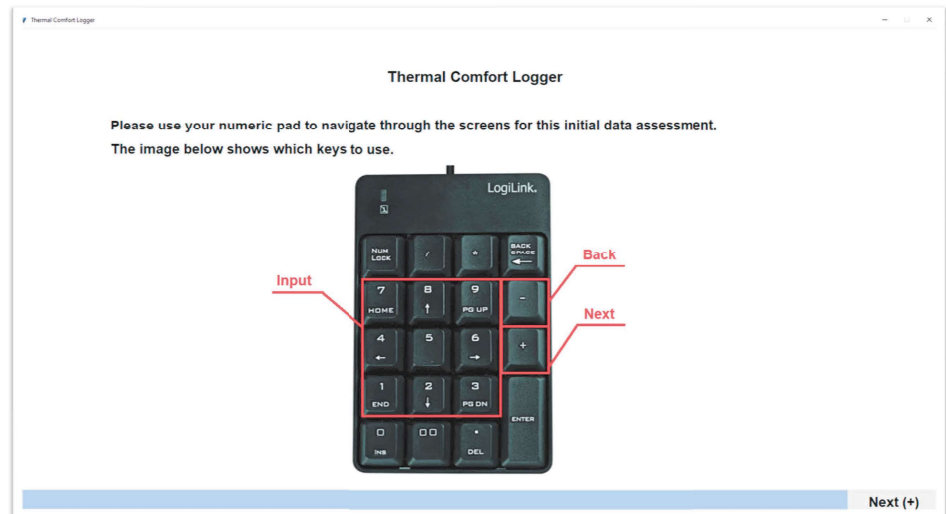
Thermal Comfort Study

Introductory Slides

Navigation

Once the logging application was started, use **only the marked keys** on the provided **numeric pad to your right.**

Follow the supervisor's instructions and feel free to ask questions if anything is unclear.

Thermal Comfort Logger

*Enter your age:

Questions with an ***** are
mandatory and cannot be skipped.

Back (-)

Next (+)

Thermal Comfort Logger

*Which gender do you identify as?

○ **Male (1)**

● **Female (2)**

○ **Non-Binary (3)**

○ **Prefer not to tell (4)**

Back (-)

Next (+)

Thermal Comfort Logger

Enter your weight in kg (e.g. 93,5):

Optional questions can be skipped by pressing **+** on the numeric pad

Back (-)   Next (+)

Thermal Comfort Logger

Enter your height in cm (e.g. 153,5):

Back (-)   Next (+)

Thermal Comfort Logger

Enter your bodyfat percentage (e.g. 15,5):

[   ]

Back (-)   Next (+)

Thermal Comfort Logger

*Enter your body temperature (e.g. 36,5):

[   ]

Back (-)   Next (+)

Thermal Comfort Logger

*Did you perform any physical activities/sports within the last hour?

◉ **No (0)**   ○ **Yes (1)**

Back (-)   Next (+)

Thermal Comfort Logger

*How much time has passed since your last meal (in full hours): | |

Back (-)   Next (+)

Thermal Comfort Logger

*Please rate how tired you're feeling at the moment.

Not at all	○	○	○	○	○	○	○	○	○	○	Very Tired
	1	2	3	4	5	6	7	8	9	10	

Thermal Comfort Logger

Enter the ID of the clothing level that describes your current clothing best.

Note: Shoes, socks, and underwear are already included in all clothing levels.

ID	Description
1	Trousers, short-sleeve shirt
2	Trousers, long-sleeve shirt
3	Trousers, long-sleeve sweater
4	Trousers, long-sleeve vest
5	Sweatpants, short-sleeve shirt
6	Sweatpants, long-sleeve shirt
7	Sweatpants, long sleeve sweater

*Enter your ID here: [ ]

Thermal Comfort Logger

*Which item describes your current emotion best?

- ○ Anger (1)
- ○ Fear (2)
- ○ Sadness (3)
- ○ Disgust (4)
- ○ Happiness (5)
- ○ Neutral (6)
- ○ Surprise (7)
- ○ Contempt (8)

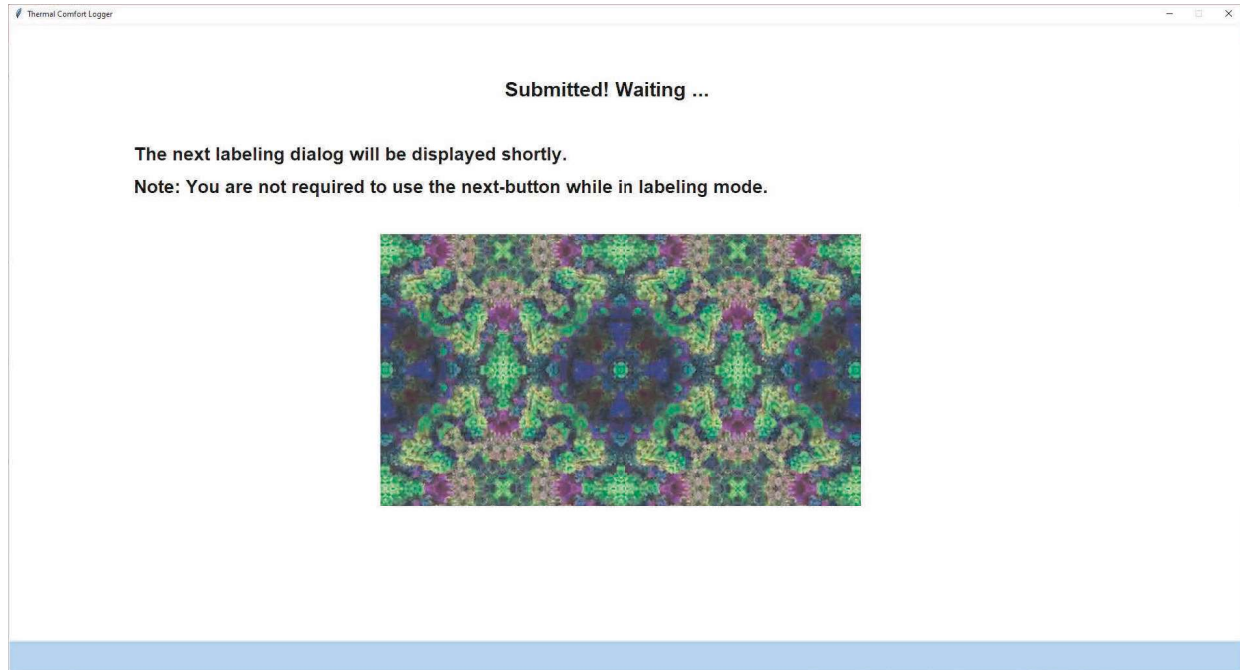| Back (-)		Next (+)

Thermal Comfort Logger

*Which item describes your current state of thermal comfort best?

○ Cold (1)　　○ Cool (2)　　○ Slightly Cool (3)　　○ Comfortable (4)　　○ Slightly Warm (5)　　○ Warm (6)　　○ Hot (7)

Once you've arrived at this page, do not press submit until the supervisor tells you to.
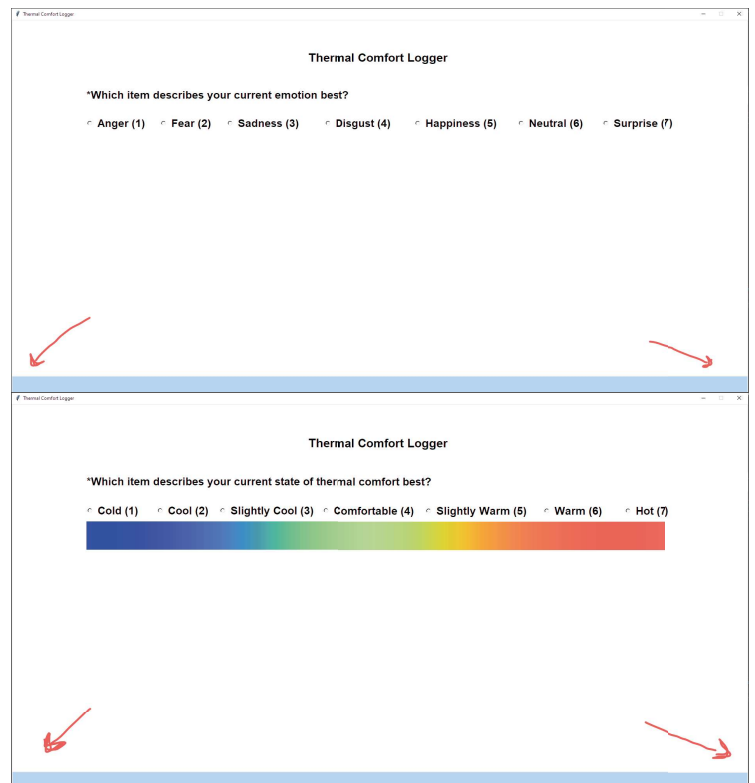
Back (-)		Submit (+)

Labeling

While in labeling mode you'll notice that there are no next/back buttons. **This is intended.**

In this mode, only use the numeric input keys to label, the page will change itself.

Other Important Hints

- If you feel unwell during the recording session, let the supervisor know.
- The interface can be displayed in german.
- **Do not use your phone or any other devices.**
- **Do not hold keyboard keys or press repeatedly in case of slow downs.**
- **If your numpad stops working you might've pressed the numlock button. Press it again to activate the numpad.**