Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection

Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger
Robert Bosch GmbH
Renningen, Germany
firstname.lastname@de.bosch.com

Kristof Van Laerhoven
University Siegen
Siegen, Germany
kvl@eti.uni-siegen.de

ABSTRACT

Affect recognition aims to detect a person's affective state based on observables, with the goal to e.g. improve human-computer interaction. Long-term stress is known to have severe implications on wellbeing, which call for continuous and automated stress monitoring systems. However, the affective computing community lacks commonly used standard datasets for wearable stress detection which a) provide multimodal high-quality data, and b) include multiple affective states. Therefore, we introduce WESAD, a new publicly available dataset for wearable stress and affect detection. This multimodal dataset features physiological and motion data, recorded from both a wrist- and a chest-worn device, of 15 subjects during a lab study. The following sensor modalities are included: blood volume pulse, electrocardiogram, electrodermal activity, electromyogram, respiration, body temperature, and three-axis acceleration. Moreover, the dataset bridges the gap between previous lab studies on stress and emotions, by containing three different affective states (neutral, stress, amusement). In addition, self-reports of the subjects, which were obtained using several established questionnaires, are contained in the dataset. Furthermore, a benchmark is created on the dataset, using well-known features and standard machine learning methods. Considering the three-class classification problem (*baseline vs. stress vs. amusement*), we achieved classification accuracies of up to 80 %. In the binary case (*stress vs. non-stress*), accuracies of up to 93 % were reached. Finally, we provide a detailed analysis and comparison of the two device locations (*chest vs. wrist*) as well as the different sensor modalities.

KEYWORDS

Affective computing, Emotion recognition, Stress detection, Multimodal dataset, Sensor fusion, Benchmark, User study

ACM Reference Format:
Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger and Kristof Van Laerhoven. 2018. Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. In *2018 International Conference on Multimodal Interaction (ICMI '18), October 16–20, 2018, Boulder, CO, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3242969.3242985

ICMI '18, October 16–20, 2018, Boulder, CO, USA

ACM ISBN 978-1-4503-5692-3/18/10…$15.00
https://doi.org/10.1145/3242969.3242985

1 INTRODUCTION

Affective computing is an emerging field, inspired by the vision to improve human-computer interaction by building empathic machines. Empathic machines detect the affective state of a human user, adapt their 'behaviour' accordingly, and might even exhibit own emotional traits. From a health care point of view, stress, defined as 'nonspecific response of the body to any demand upon it' [25], is a particularly interesting affective state. This is due to the harmful effects of long-term stress, which can range from headaches and troubled sleeping to an increased risk of cardiovascular diseases [4, 16, 22]. According to the British Health and Safety Executive (HSE), stress accounted for 37% of all work-related ill health cases in 2015/16 [1]. These severe side effects of stress call for automated detection methods.

In order to build a reliable stress detection system, it is important to understand that stress is primarily a physiological response to a stimulus, triggered by the sympathetic nervous system (SNS). During this response a mixture of hormones like cortisol or adrenaline are released, leading to an increased breathing/heart rate and muscle tension. These physiological changes prepare the organism for a physical reaction ('fight-or-flight'). As shown by *Kreibig et al.* [13] the physiological responses to certain emotional stimuli are also to some extent specific. A psychological model well suited for capturing affective states is Russel's circumplex model [23]. According to this model, affective states can be mapped into a 2D space, using for example the axes valence and arousal. The valence dimension indicates how negative/positive an affective state is perceived. On the arousal axis, which is known to be impacted by stress [9], the state is rated in terms of excitement.

In recent years, the specificity of the physiological responses to stress and emotional stimuli was utilised to train machine learning models to predict the affective state of a subject. Using deep neural networks, trained on audio and/or visual data, a high performance in emotion classification is achieved [17, 29]. However, these models are quite demanding in terms of computational resources and are only partially applicable on embedded devices. Classifying stress from audio samples was also successfully done, e.g., by Lu [14]. However, recording audio and/or video data continuously is in terms of privacy quite intrusive, and concerning technical feasibility difficult. Hence, these modalities are only available in specific occasions. Wearable electronic devices, in contrast, are only minimally intrusive. Devices like smart phones/watches are already popular among users. Contemporary wearables can be used to track steps and monitor other physical activities. To keep up with the current trend to quantify vital functions, a desirable next step is to infer affective states based on multimodal wearable sensor data.

Plarre et al. [20] and *Hovsepian et al.* [8] trained stress detection systems on peripheral physiological data utilising electrocardiogram (ECG), respiration (RESP) and 3-axis accelerometer (ACC) data, recorded by a chest-worn device. *Gjoreski et al.* [5] used the data of a wrist-worn device recording blood volume pulse (BVP), electrodermal activity (EDA), skin temperature (TEMP), and ACC to train a stress detection model. In order to detect emotions in response to music, *Kim et al.* [10] used ECG, RESP, EDA, and electromyogram (EMG) data. Distinguishing stress and emotions is not a trivial task, since both have a strong impact on the autonomous nervous system. However, in current affective computing research, stress and emotion detection from wearables are commonly tackled as two disjoint topics. Our work addresses this shortcoming. This is important since, for a holistic affective user model, assessing both stress and emotions is required.

As outlined above, multimodal setups have been used for stress or emotion detection tasks. However, in contrast to many other research fields, there is a lack of commonly used, standardised benchmarking datasets for stress and affect detection. Hence, it is difficult to compare results obtained by different researchers. Our work intends to address this shortcoming as well.

The main contributions of this paper are threefold:

(1) A new multimodal, publicly available dataset[1] is presented. The data has been recorded using two different devices (one chest-based and one wrist-based), each including high-resolution physiological (BVP, ECG, EDA, EMG, RESP, and TEMP) and motion (ACC) modalities.
(2) The dataset bridges the gap between previous lab studies on stress and emotions, by containing three different affective states (neutral, stress, amusement). In addition, the dataset features self-reported values on the perceived affective state of the subjects, which were obtained using several established questionnaires. These self-reports can be used to train personalised classifiers.
(3) A benchmark is created using a large amount of well-known features (extracted from physiological and motion signals) and common machine learning methods (Decision Tree (DT), Random Forest (RF), AdaBoost (AB), Linear Discriminant Analysis (LDA) and k-nearest neighbour (kNN)).

2 RELATED WORK

In recent years, a number of studies have been conducted with the aim to elicit and detect stress based on physiological parameters. For this purpose, stressors like public speaking, mental arithmetic or physical stressors (e.g. cold pressor) were employed [5, 8, 20]. However, these approaches focus on detecting and classifying stressful vs. non-stressful states and do not take any other affective states into account. Classical machine learning algorithms like the RF were employed to the stress classification task, achieving a 72% accuracy on a three class (no, low, high stress) problem [5]. *Kim et al.* [10] used four songs to elicit different target emotions, which were then classified using LDA, achieving a subject-independent correct classification ratio of 70%. However, the topic of combining stress and emotion detection systems has only received little attention. *Zenonos et al.* [32] presented a mood recognition system capable of distinguishing eight different moods with a subject-independent accuracy of 62.14%. However, as the system was trained and tested on only four subjects, the generalisation properties are questionable.

Although there is intensive research in the field of affective computing from wearable devices, there is only very little publicly available data. *Healey et al.* [6] published a dataset on driver stress. This dataset features ECG (496 Hz), EDA (31 Hz), RESP (31 Hz), and EMG (15.5 Hz) data. Moreover, *Picard et al.* [19] published a dataset containing physiological data recorded from one person, who is subject to eight different emotional stimuli over 20 days. More recently, *Koestra et al.* [12] published DEAP, a database for emotion analysis using physiological signals. The dataset contains electroencephalogram (EEG) (512 Hz), facial videos and peripheral physiological signals (recorded at 512 Hz, then down-sampled to 256 Hz). The data was recorded while the subjects watched 40 one-minute excerpts from music videos. The final 40 clips were chosen from a larger pool of videos, by asking volunteers to rate the clips in terms of their valence and arousal value and then choosing the ones that had the strongest rating with the smallest variance.

The way humans perceive and react to an affective stimulus is very subject dependent. Hence, personalisation is an important issue. In order to train personalised models, subjective ratings of the different affective stimuli are required. These ratings are commonly generated by self-assessment of the subjects. For instance, manikins can be used to generate personalised valence, arousal, dominance, and liking labels [12]. In the study of *Plarre et al.* [20] subjects reported their stress levels by answering five questions (Cheerful?, Happy?, Angry/Frustrated?, Nervous/Stressed?, Sad?) on a four point scale (NO, no, yes, YES). Other studies employed more complex questionnaires such as the PANAS [18] and STAI [5]. In field studies, smart phone apps offer ideal platforms for self-reports, e.g., on mood [32].

In this paper we present a novel dataset for stress and affect detection. The subjects ($n = 15$) were exposed to different affective stimuli (stress and amusement). In addition, a baseline and two meditation periods (introduced to de-excite the participants after a stimulus) were recorded. The dataset contains high resolution physiological (ECG, EDA, EMG, RESP, and TEMP) and motion (ACC) data sampled at 700 Hz from a chest-worn device, and lower resolution data from a wrist-worn device. Finally, the data of each subject is linked to several self-reports, which represent the subjective experience during an affective stimulus. The dataset is well-suited to benchmark (personalised) stress and affect detection algorithms, a first evaluation is presented in this paper.

3 DATA COLLECTION

This section provides details on the subjects, employed sensors, sensor placement, the study protocol, and the self-reports. The study was approved by the workers council and the data security officer of our research center.

3.1 Participants

Due to the defined study protocol, we specifically targeted graduate students at our research facility. Exclusion criteria, stated in the study invitation, were pregnancy, heavy smoking, mental disorders,

[1] The dataset introduced in this paper is made publicly available, and can be downloaded from: https://ubicomp.eti.uni-siegen.de/home/datasets/icmi18/.

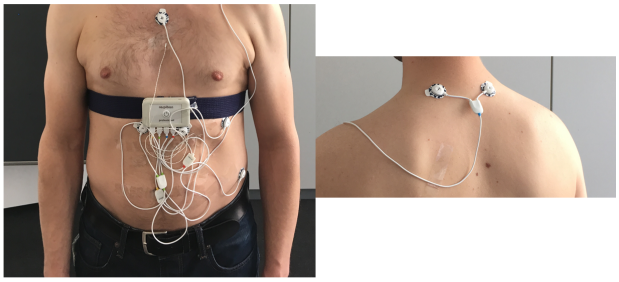

Figure 1: Placement of the *RespiBAN* and the ECG, EDA, EMG, TEMP sensors.

chronic and cardiovascular diseases. In total, 17 subjects participated in our study. Due to sensor malfunction, the data of two participants had to be discarded. The remaining 15 subjects had a mean age of 27.5 ± 2.4 years. Twelve subjects were male and the other three subjects were female.

3.2 Sensor Setup and Placement

For the data collection, we used both a chest- and a wrist-worn device: a *RespiBAN Professional*[2] and an *Empatica E4*[3], respectively. The *RespiBAN* itself is equipped with sensors to measure ACC and RESP, and can function as a hub for up to four additional modalities. Using the four analog ports, ECG, EDA, EMG, and TEMP were recorded. All signals were sampled at 700 Hz. The *RespiBan* was placed around the subject's chest (see Figure 1). The RESP is recorded via a respiration inductive plethysmograph sensor. The ECG data was recorded via a standard three point ECG. In order to allow the subject to move as freely as possible, the EDA signal was recorded on the rectus abdominis (the abdomen has a high density of sweat glands [28], hence suitable for EDA measurement) and the TEMP sensor was placed on the sternum. The EMG data was recorded on the upper trapezius muscle on both sides of the spine. In order to avoid wireless packet loss, the recorded data was stored locally and transferred to a computer for further processing after the experiment. All subjects wore the *Empatica E4* on their non-dominant hand. The *E4* records BVP (64 Hz), EDA (4 Hz), TEMP (4 Hz), and ACC (32 Hz).

3.3 Study Protocol

The goal of the study was to elicit three different affective states (neutral, stress, amusement) in the participants. In addition, the subjects were asked to follow a guided meditation in order to de-excite them after the stress and amusement conditions. The different parts of the study protocol are detailed below:

Preparation: The participants had to avoid caffeine and tobacco in the hour before the experiment was to begin. Further, the subjects were asked to do no strenuous exercise on the day of the study. Prior to the study the participants read and signed a consent form. Upon arrival at the study location, the participants were equipped with the sensors and a short sensor test was conducted. Then the *RespiBAN* and *E4* were synchronised manually via a double tap gesture.

Baseline condition: After the subjects had been equipped with the sensors, a 20 minute baseline was recorded. During the baseline the subjects were sitting/standing at a table and neutral reading material (magazines) was provided. The baseline condition aimed at inducing a neutral affective state.

Amusement condition: During the amusement condition, the subjects watched a set of eleven funny video clips. Each clip was followed by a short neutral sequence of five seconds. Eight of the short clips were chosen from the corpus presented by *Samson et al.* [24]. The remaining three videos were chosen by the authors. In total, the amusement condition had a length of 392 seconds.

Stress condition: The subjects were exposed to the well-studied Trier Social Stress Test (TSST) [11], which consists of a public speaking and a mental arithmetic task. These tasks are known to elicit stress reliably [20], as they are social evaluative and inflict a high mental load on the subjects. In our version of the TSST, the study participants first had to deliver a five minute speech on their personal traits in front of a three-person panel, focusing on strengths and weaknesses. The subjects were told that the three panel members were human resources specialists from our research facility. In order to boost their career options, the subjects were told to try to leave the best possible impression. The study participants had three minutes to prepare their speech but they were not allowed to use their notes during the presentation. After the speech, the panel asked the subjects to count from 2023 to zero, doing steps of 17. Moreover, whenever the subjects made a mistake, they had to start over. For both tasks, the subjects were given five minutes by the panel and hence the TSST had a total length of about ten minutes. After the TSST the study participants were given a ten-minute rest period.

Meditation: The amusement and stress conditions, which both aimed at exciting the subjects, were followed by a guided meditation. The aim of this meditation was to 'de-excite' the subjects and bring them back to a close to neutral affective state. The meditation was based on a controlled breathing exercise, instructed via an audio track. Subjects followed the instructions with closed eyes, while sitting in a comfortable position. The meditation had a duration of seven minutes.

Recovery: At the end of the protocol, the sensors were again synchronised via a double tap gesture. Then, the sensors were removed and the subjects were informed that the panel members were just 'normal' researchers.

In total, the study had a duration of about two hours. Figure 2 summarises the protocol (without the preparation and the recovery period). As detailed above, our lab protocol features two major stimuli: an amusement condition and a stressful condition. These two conditions were interchanged (see Figure 2) between different subjects in order to avoid effects of order. In addition to these stimuli, a baseline and two meditation periods were recorded. In order to induce variance in the subjects' posture, the baseline, amusement and stress conditions were conducted either standing or sitting. For each condition, approximately half of the subjects were standing and the other half were sitting. During the meditation, however, all subjects were seated.

[2]http://www.biosignalsplux.com/en/respiban-professional
[3]http://www.empatica.com/research/e4/

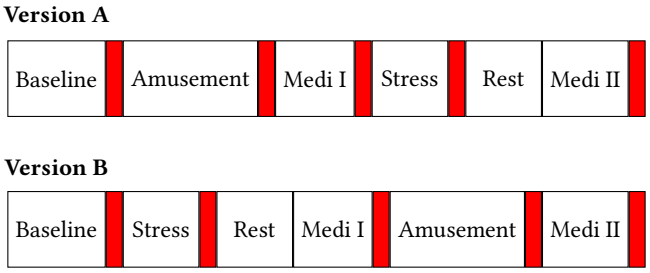

Figure 2: The two different versions of study protocol. The red/dark boxes refer to filling in self-reports.

3.4 Obtaining Ground Truth

In order to validate the study protocol, we collected five self-reports of each participant (timing indicated by red/dark boxes in Figure 2). Each of the self-reports contained several questionnaires. Firstly, participants filled in a Positive and Negative Affect Schedule (PANAS), which consists of 20 items (ten positive and ten negative items) each rated on a five point Likert scale. PANAS reliably assesses positive (PA) and negative affect (NA), which are two largely independent dimensions [30]. PA reaches from 'sad and lethargic' (low value) to 'concentrated and energetic' (high value). NA ranges from 'calmness' (low value) to 'subjective distress' (high value). Furthermore, we added the items *Stressed?*, *Frustrated?*, *Happy?*, and *Sad?*, which were scored by the subjects using the same scale as in PANAS. These items can be used to generate the same labels as used by *Plarre et al.* [20]. Secondly, similar to *Gjoreski et al.* [5], we used six items from the State-Trait Anxiety Inventory (STAI) to gain insight into the anxiety level of the participants. The items were chosen according to their factor loads [2], and scored on a four point Likert scale. Thirdly, we used Self-Assessment Manikins (SAM) to generate labels in the valence-arousal space [12]. Finally, after the TSST, nine items from the Short Stress State Questionnaire (SSSQ) [7] were added to the questionnaires in order to identify which type of stress (worry, engagement, or distress) was most prevalent in the subjects. The values from these questionnaires can be seen as subjective reports on how the participants felt during a condition and may be used to train personalised models. However, for the first evaluation presented in this paper, we used the study protocol as ground truth.

4 METHODS

The analysis and evaluation of our dataset follows the classical data processing chain, consisting of the following steps: preprocessing, segmentation, feature extraction, and classification. Details on these different steps are presented below (the first three steps are explained together since they depend on the specific sensor modality).

4.1 Feature Extraction

Segmentation of the (preprocessed) sensor signals was done using a sliding window, with a window shift of 0.25 seconds. The ACC-features were computed with a window size of five seconds, as similar window lengths are broadly applied for acceleration-based context recognition (e.g. *Reiss et al.* [21]). All features (except for statistical- and frequency-domain EMG-features, see below) based on physiological signals were computed with a window size of 60 seconds. This window size was chosen following *Kreibig et al.* [13]. In Table 1, the features extracted from the different modalities are displayed.

On the raw ACC signal different statistical features, e.g. the mean $\mu_{acc,i}$ and standard deviation $\sigma_{acc,i}$ were computed. These features were computed both for each axis separately ($i \in \{x, y, z\}$) and as absolute magnitudes, summed over all axes (3 D). In addition, the peak frequency was computed for each axis separately $f^{peak}_{acc,i}$.

On the raw ECG/BVP signal the heart beats were found based on peak detection algorithms. Using the peaks, the heart rate (HR) and corresponding statistical features (mean, standard deviation) were computed. Moreover, from the location of the heart beats the heart rate variablility (HRV) was derived, which is an important starting point for additional features. For instance, the energy in different frequency bands (f^{x}_{HRV}) was computed. The frequency bands (x) used, were the ultra low (ULF: 0.01-0.04 Hz), low (LF: 0.04-0.15 Hz), high (HF: 0.15-0.4 Hz) and ultra high (UHF: 0.4-1.0 Hz) band. In [15] the HR and HRV are described in detail.

The EDA is controlled by the sympathetic nervous system (SNS), and hence it is particularly sensitive to high arousal states. First, a 5 Hz lowpass filter was applied to the raw EDA signal, similar to related work [26, 27]. Then, statistical features were computed (e.g. mean, standard deviation, dynamic range, etc.). Furthermore, the raw EDA signal consists of a tonic (referred to as skin conductance level (SCL)) and a phasic (skin conductance response (SCR)) component. The SCL represents a slowly varying baseline conductivity, while the SCR is a short term response to a stimulus. In order to separate these two components, the method proposed by *Choi et al.* [3] was applied. After separating the SCL and SCR, additional features, e.g. number of peaks in the SCR ($\#_{SCR}$), were computed. Details about the EDA-related features can be found in *Choi et al.* [3] and *Healey et al.* [6].

Two different processing chains were applied to the raw EMG signal. In the first chain, the DC component was removed by applying a highpass filter. Then, the filtered signal was cut into 5-second windows, and statistical and frequency-domain features (e.g. peak frequency) were computed. In addition, the spectral energy ($PSD(f_{EMG})$) was computed in seven evenly spaced frequency bands from 0 to 350 Hz. Following the second processing chain, a lowpass filter (50 Hz) was applied to the raw EMG signal. Next, the processed signal was segmented into 60-second windows. On these windows different peak features, e.g. number $\#^{peaks}_{EMG}$ and mean amplitude μ^{Amp}_{EMG}, were computed. For a more detailed description of EMG-based features, we refer the reader to *Wijsman et al.* [31].

Before computing features on the RESP signal, a bandpass filter (cut off frequencies: 0.1 and 0.35 Hz) was applied. Next, a peak detector was used to identify minima and maxima. Following *Plarre et al.* [20] the mean and standard deviation of the inhalation/exhalation (μ_I, σ_I, μ_E, and σ_E) were computed. In addition, the ratio between inhalation and exhalation (I/E), stretch $range_{RESP}$, inspiration volume vol_{insp}, respiration rate $rate_{RESP}$, and respiration duration were derived $\sum_{RESP}$ [20].

Table 1: List of extracted features. Abbreviations: # = number of, $\sum$ = sum of, STD = standard deviation.

	Feature	Description
ACC	$\mu_{ACC,i}$, $\sigma_{ACC,i}$ $i \in \{x,y,z,3D\}$	Mean, STD for each axis separately and summed over all axes
	$\| \int_{ACC,i} \|$ $i \in \{x,y,z,3D\}$	Absolute integral for each/all axes
	$f^{peak}_{ACC,j}$ $j \in \{x,y,z\}$	Peak frequency for each axis i
ECG and BVP	μ_{HR}, σ_{HR}	Mean, STD of the HR
	μ_{HRV}, σ_{HRV}	Mean, STD of the HRV
	$NN50$, $pNN50$	# and percentage of HRV intervals differing more than 50 ms
	$TINN$	Triangular interpolation index
	rms_{HRV}	Root mean square of the HRV
	f^{x}_{HRV} $x \in \{ULF,LF,HF,UHF\}$	Energy in ultra low, low, high, and ultra high frequency component of the HRV
	$f^{LF/HF}_{HRV}$	Ratio of LF and HF component
	$\sum^{f}_{x}$ $x \in \{ULF,LF,HF,UHF\}$	$\sum$ the freq. components in ULF-HF
	rel^{f}_{x}	Relative power of freq. component
	LF_{norm}, HF_{norm}	Normalised LF and HF component
EDA	μ_{EDA}, σ_{EDA}	Mean, STD of the EDA signal
	min_{EDA}, max_{EDA}	Min and max value
	∂_{EDA}, $range_{EDA}$	Slope and dynamic range
	μ_{SCL}, σ_{SCL}, σ_{SCR}	Mean, STD of the SCR/SCL
	$corr(SCL,t)$	Correlation btw SCL and time
	$\#_{SCR}$	# identified SCR segments
	$\sum^{Amp}_{SCR}$, $\sum^{t}_{SCR}$	$\sum$ SCR startle magnitudes and response durations
	$\int_{scr}$	Area under the identified SCRs
EMG	μ_{EMG}, σ_{EMG}	Mean, STD of EMG signal
	$range_{EMG}$	Dynamic range
	$\| \int_{EMG} \|$	Absolute integral
	$\tilde{\pi}_{EMG}$	Median of the EMG signal
	P^{10}_{EMG}, P^{90}_{EMG}	10th and 90th percentile
	μ^{f}_{EMG}, $\tilde{f}_{EMG}$,	Mean, median and
	f^{peak}_{EMG}	Peak frequency
	$PSD(f_{EMG})$	Energy in seven bands
	$\#^{peaks}_{EMG}$	# peaks
	μ^{Amp}_{EMG}, σ^{Amp}_{EMG}	Mean, STD of peak amplitudes
	$\sum^{Amp}_{EMG}$, $\bar{\sum}^{Amp}_{EMG}$	$\sum$ and normalised $\sum$ of peak amplitudes
RESP	μ_x, σ_x $x \in \{I,E\}$	Mean, STD of inhalation (I) and exhalation (E) duration
	I/E	Inhalation/exhalation ratio
	$range_{RESP}$, vol_{insp}	Stretch, Volume
	$rate_{RESP}$	Breath rate
	$\sum_{RESP}$	Respiration duration
TEMP	μ_{TEMP}, σ_{TEMP}	Mean, STD of the TEMP
	min_{TEMP}, max_{TEMP}	Min, max TEMP
	$range_{TEMP}$	Dynamic range
	∂_{TEMP}	Slope

On the raw TEMP signal common statistical features (mean, standard deviation, min, max, etc.) were computed. In addition, the slope of the signal ∂_{TEMP} is used as a feature.

4.2 Classification Algorithms

The extracted features, detailed above, serve as input for the classification step. Five machine learning algorithms were applied and compared within our benchmark: Decision Tree (DT), Random Forest (RF), AdaBoost (AB), Linear Discriminant Analysis (LDA), and k-Nearest Neighbour (kNN). As the entire data processing chain was implemented in Python, we used the scikit-learn implementation of the aforementioned classifiers. For the AB ensemble learner, decision tree was used as base estimator. For each of the decision-tree-based classification algorithms (DT, RF, AB), information gain was used to measure the quality of splitting decision nodes, and the minimum number of samples required to split a node was set to 20. The number of base estimators was set to 100 for both of the ensemble learners (RF and AB). Moreover, a LDA and a kNN (with k=9) classifier were used for classification.

4.3 Evaluation Metric

We used accuracy and F_1-score as evaluation metrics. Accuracy represents the number of correctly classified instances out of all samples. The F_1-score is defined as the harmonic mean of precision, indicating the reliability of the results in a certain class, and recall, representing a measure of completeness. To obtain the final F_1-score, precision and recall were computed for each class separately and then averaged. Applying the F_1-score is recommended for unbalanced classification tasks, which is the case when using WESAD (since the various conditions were carried out at different lengths during the study protocol). All models were evaluated using the leave-one-subject-out (LOSO) cross-validation (CV) procedure. Hence, the results indicate how a model would generalise and perform on data of a previously unseen subject.

5 RESULTS AND DISCUSSION

This section provides first an analysis of the collected self-reports. Second, detailed results on the evaluation of the recorded sensor data and processing chain are given, including a discussion on the importance of the different sensor modalities and extracted features. For the data analysis and evaluation presented here, we only consider the data recorded during the baseline, stress (TSST), and amusement parts of the study protocol (see Figure 2.)

5.1 Evaluation of the Self-reports

In this work, the analysis of the self-reported measures (see subsection 3.4) has been used to verify that the design of the experimental conditions was suitable to manipulate the subjects' affective state as desired. Table 2 shows the results (mean and standard deviation) of the three measures and subscales, respectively.

Comparing the self-reports after the amusement and baseline condition reveals that the amusement condition had the desired effect: the subjects report slightly higher scores on valence and arousal (dimensional approach, DIM) and less anxiety (STAI). However, the effect of the condition is rather small. In contrast, the impact of the stress condition is pronounced, across all questionnaires. The analysis of the SSSQ scores indicates that the subjects felt more engaged and worried than distressed during the TSST task (Engagement: 11.7 ± 2.3, Distress: 6.0 ± 2.9, Worry: 10.6 ± 2.3).

Table 2: Evaluation of the questionnaires.

	PANAS		STAI	DIM	
	positive	negative		valence	arousal
Baseline	25.5±6.0	12.3±2.0	10.8±1.9	6.7±0.9	2.5±0.9
Stress	31.3±4.7	22.0±6.4	18.5±2.0	4.5±1.6	6.8±1.8
Amusement	25.8±5.1	11.4±2.1	9.3±2.0	7.5± 0.6	3.0±1.6

The high 'Engagement' score might result from the subjects' high motivation to perform well in the given task. The high 'Worry' score suggests that the subjects were determined to give a good impression on the panel. In our opinion, these scores demonstrate that most subjects believed our cover story of the TSST.

After the stress condition, the PANAS showed increased scores with respect to positive (PA) and negative affect (NA). The high PA score indicates that subjects felt energised and concentrated during the TSST, which coincides with the high engagement values reported in the SSSQ. The elevated NA score indicates an increased level of subjective distress. The DIM scores support these observations, indicating an increase in arousal and a decrease in valence. Moreover, the STAI shows elevated values after the TSST, as expected for subjects in a stressful condition. The statistical difference between the baseline and stress conditions were confirmed with the Wilcoxon signed-rank test. Overall, the experimental protocol (especially with respect to the stress condition) is considered suitable to induce the desired affective states.

5.2 Evaluation of Sensor Modalities and Extracted Features

Based on the affective states of the study protocol (baseline, stress, and amusement condition), we distinguish two classification tasks. First, a three-class problem was defined: *baseline vs. stress vs. amusement*. Results on this classification task are presented in Table 3. Second, a binary classification task was defined by combining the states *baseline* and *amusement* to a *non-stress* class, posing the *stress vs. non-stress* classification problem. Results of this classification task are presented in Table 4. For both classification tasks, 16 different modality combinations are evaluated:

- each of the four modalities of the wrist-based device separately (ACC, BVP, EDA, and TEMP)
- each of the six modalities of the chest-based device separately (ACC, ECG, EDA, EMG, RESP, and TEMP)
- all modalities of one device (wrist or chest)
- all physiological modalities of one device (same as last entry, but without ACC)
- all modalities from both devices (wrist and chest) together
- all physiological modalities from both devices together (same as last entry, but without ACC)

Finally, the evaluation was performed using each of the five machine learning algorithms, specified previously. Each setup (defined by the classification task, applied classifier, and included sensor modalities) was run five times, to report mean and standard deviation of the evaluation metrics (F_1-score and accuracy). Since LDA and kNN are deterministic classifiers, only the mean values are reported.

The data considered in this paper (belonging to the three affective states of interest) amount to approximately 36 minutes per subject. With 15 subjects and using a sliding window of 0.25 seconds, approximately 133000 windows were generated. Out of these windows, 53 % belong to the *baseline* class, 30 % represent the *stress* class, and 17 % originate from the *amusement* condition. In the last two rows of Table 3 the baseline F_1-score/accuracy of a random and a sophisticated guesser on the three-class problem are displayed. The random guesser is defined to choose one of the three possible classes at random, thus reaching an accuracy of 33 % and a F_1-score of 32 %. In contrast, the sophisticated guesser would always choose the majority class. Hence, a sophisticated guesser would reach an accuracy of 53%. However, its' F_1-score would only be 32 %. In the two last rows of Table 4, the same type of random and sophisticated guesser are presented for the binary classification task.

Comparing the performance of the employed algorithms, on the three-class task (Table 3) and binary classification task (Table 4), it becomes apparent that the ensemble-based methods (RF, AB) and the LDA reached similar classification scores. Depending on the input modalities, these classifiers reach scores up to 80 % for the three-class problem and up to 93 % for the binary task, respectively. Concluding from Table 3 and Table 4, the kNN had the overall worst performance, reaching accuracies of at most 60 % on the three-class problem, and 78 % in the binary task.

Using only motion-based features (wrist and/or chest ACC) leads to considerably lower classification scores compared to results obtained using physiological features. This suggests that the physiology-based features provide a deeper insight into the affective states of the subjects than the motion patterns. Moreover, we can rule out the possibility that our classifiers only learned to distinguish between motion patterns characteristic for the conditions of the protocol.

In the three-class problem the accuracies using one of the wrist-based physiological modalities range from 59 % to 70 %. Using one of the physiological chest-based modalities on the same classification problem, accuracies between 54 % and 72 % are reached. In the binary classification task the accuracies using a wrist-based input modality range from 69 % to 86 % and the accuracies using one of the chest-based modalities range from 67 % to 88 %. In both classification tasks the RESP is a particularly strong chest-based modality leading to the best result of a single modality. Besides the stress-related changes in the respiration, this can be partially explained considering the fact that the study participants spoke during the TSST. Hence, the classifiers might have partially learned to distinguish between speaking (stress condition) and non-speaking episodes (baseline and amusement condition). In both classification tasks, using only the TEMP data, either chest or wrist-based, as input leads to low classification scores. Obviously, TEMP is not a well-suited modality to solely base the classification of affective states upon. Comparing the results obtained using only the wrist- or chest-based EDA data, the latter seems to hold more relevant information leading to somewhat higher accuracies in both classification tasks. In contrast, comparing the performance of classifiers solely relying on the BVP or ECG data, the former leads to slightly higher accuracies. The results reached using all physiological chest-based modalities (three-class accuracy: 80 %, binary accuracy: 93 %) are higher than the ones obtained using all physiological wrist-based

Table 3: Evaluation of the given modalities and classifiers on the three-class (*baseline vs. stress vs. amusement*) classification task. Abbreviations: DT = Decision Tree, RF = Random Forest, AB = AdaBoost DT, LDA = Linear discriminant analysis, kNN = k-nearest neighbour

	DT		RF		AB		LDA		kNN	
	F_1-score	Accuracy	F_1-score	Accuracy	F_1-score	Accuracy	F_1-score	Accuracy	F_1-score	Accuracy
Motion:										
ACC wrist	43.91 ± 1.16	53.71 ± 0.91	46.50 ± 0.26	56.40 ± 0.16	46.38 ± 0.64	**57.20 ± 0.57**	36.27	47.73	37.20	45.54
ACC chest	42.18 ± 0.4	51.14 ± 0.29	41.96 ± 0.29	53.48 ± 0.29	44.28 ± 0.75	**56.56 ± 0.70**	34.61	48.84	31.07	40.29
Wrist:										
BVP	51.15 ± 0.31	57.57 ± 0.22	53.83 ± 0.11	64.09 ± 0.12	53.29 ± 0.16	64.46 ± 0.21	54.72	**70.17**	50.97	59.44
EDA	45.48 ± 0.17	54.36 ± 0.27	45.74 ± 0.06	56.57 ± 0.05	49.06 ± 0.59	59.85 ± 0.42	42.72	**62.32**	45.20	54.98
TEMP	41.46 ± 0.24	47.42 ± 0.36	41.85 ± 0.19	48.67 ± 0.21	41.19 ± 0.24	49.39 ± 0.23	40.89	**58.96**	38.97	44.32
Wrist physio	57.13 ± 0.86	63.34 ± 1.00	66.33 ± 0.36	**76.17 ± 0.42**	64.24 ± 0.39	73.62 ± 0.55	58.18	68.85	50.85	58.54
Chest:										
ECG	51.69 ± 0.35	57.81 ± 0.36	52.24 ± 0.33	60.36 ± 0.22	52.48 ± 0.38	61.71 ± 0.40	56.03	**66.29**	47.77	54.76
EDA	43.88 ± 0.20	48.49 ± 0.29	42.40 ± 0.55	45.00 ± 0.61	48.33 ± 0.31	54.06 ± 0.45	46.83	**67.07**	37.26	40.03
EMG	34.65 ± 0.21	41.00 ± 0.19	38.10 ± 0.47	48.20 ± 0.51	37.68 ± 0.24	48.03 ± 0.24	37.72	**53.99**	35.97	42.73
RESP	59.08 ± 0.21	65.97 ± 0.20	60.69 ± 0.15	70.27 ± 0.14	61.76 ± 0.34	71.94 ± 0.30	60.09	**72.37**	45.86	60.45
TEMP	41.27 ± 0.29	47.53 ± 0.28	42.46 ± 0.24	48.40 ± 0.26	40.76 ± 0.8	47.98 ± 0.60	30.96	**55.68**	35.18	43.32
Chest physio	55.10 ± 0.92	58.62 ± 1.07	64.60 ± 0.54	71.37 ± 0.58	72.51 ± 0.17	**80.34 ± 0.43**	74.43	79.35	51.09	57.31
All wrist	43.62 ± 1.33	53.98 ± 1.79	62.86 ± 0.65	74.85 ± 0.20	64.12 ± 0.98	**75.21 ± 0.77**	63.24	70.74	37.20	45.54
All chest	53.06 ± 0.50	57.68 ± 0.40	60.80 ± 1.00	68.76 ± 1.35	64.89 ± 0.81	74.74 ± 0.94	72.49	**76.50**	38.39	46.18
All physio	55.71 ± 0.93	62.57 ± 0.80	64.23 ± 0.97	73.33 ± 0.95	71.10 ± 0.78	**79.86 ± 0.62**	72.48	78.19	52.94	59.61
All modalities	58.05 ± 1.61	63.56 ± 1.73	64.08 ± 1.68	74.97 ± 1.11	68.85 ± 0.89	**79.57 ± 0.93**	71.56	75.80	48.70	56.14

Baseline	Random Guessing		Sophisticated guessing	
	F_1-score	Accuracy	F_1-score	Accuracy
	31.66	33.33	23.13	53.12

Table 4: Evaluation of the given modalities and classifiers on the binary (*stress vs. non-stress*) classification task.

	DT		RF		AB		LDA		kNN	
	F_1-score	Accuracy	F_1-score	Accuracy	F_1-score	Accuracy	F_1-score	Accuracy	F_1-score	Accuracy
Motion:										
ACC wrist	55.36 ± 0.47	64.08 ± 0.49	59.02 ± 0.78	69.96 ± 0.55	61.70 ± 0.80	**71.69 ± 0.45**	44.93	60.02	52.72	63.80
ACC chest	61.92 ± 0.83	71.75 ± 0.53	59.91 ± 0.25	72.87 ± 0.08	62.17 ± 0.45	**73.87 ± 0.30**	57.52	72.05	47.79	57.81
Wrist:										
BVP	78.27 ± 0.17	81.39 ± 0.15	81.35 ± 0.15	84.18 ± 0.11	81.23 ± 0.15	84.10 ± 0.13	83.08	**85.83**	78.94	82.06
EDA wrist	70.95 ± 0.37	76.21 ± 0.27	70.88 ± 0.20	76.29 ± 0.14	75.34 ± 0.57	**79.71 ± 0.43**	69.86	78.08	68.30	73.13
TEMP wrist	63.15 ± 0.18	68.22 ± 0.19	62.90 ± 0.10	67.82 ± 0.11	62.27 ± 0.25	67.11 ± 0.34	56.37	**69.24**	60.18	64.46
Wrist physio	82.37 ± 0.21	84.88 ± 0.11	86.10 ± 0.29	**88.33 ± 0.25**	85.86 ± 0.20	88.05 ± 0.18	83.77	86.46	78.93	81.96
Chest:										
ECG	77.01 ± 0.37	80.17 ± 0.29	79.64 ± 0.15	82.78 ± 0.11	80.20 ± 0.25	83.37 ± 0.20	81.31	**85.44**	75.39	79.19
EDA chest	69.88 ± 0.41	73.55 ± 0.44	73.63 ± 0.18	77.51 ± 0.23	71.97 ± 0.26	75.50 ± 0.29	74.51	**81.70**	66.64	69.73
EMG	47.06 ± 0.20	56.25 ± 0.05	49.42 ± 0.35	63.44 ± 0.18	50.84 ± 0.44	62.88 ± 0.31	52.49	**67.10**	51.84	58.74
RESP	79.92 ± 0.19	83.03 ± 0.17	84.33 ± 0.10	86.63 ± 0.08	84.64 ± 0.06	86.87 ± 0.06	85.61	**88.09**	69.17	75.67
TEMP chest	57.40 ± 0.08	64.33 ± 0.07	56.75 ± 0.25	64.75 ± 0.28	55.03 ± 0.27	63.46 ± 0.21	41.00	**69.49**	51.64	58.25
Chest physio	81.29 ± 0.22	84.18 ± 0.20	90.44 ± 0.66	92.01 ± 0.51	87.11 ± 0.57	89.76 ± 0.48	91.47	**93.12**	77.27	81.05
All wrist	78.71 ± 0.53	82.19 ± 0.44	84.11 ± 0.31	**87.12 ± 0.24**	80.11 ± 0.93	83.98 ± 0.75	84.05	86.88	52.72	63.80
All chest	78.26 ± 0.46	81.29 ± 0.38	90.04 ± 0.84	91.70 ± 0.75	89.57 ± 0.61	91.58 ± 0.46	91.07	**92.83**	64.20	69.70
All physio	83.03 ± 1.61	85.16 ± 1.50	86.02 ± 0.55	87.91 ± 0.54	87.78 ± 1.38	89.77 ± 1.17	90.93	**92.51**	79.44	83.16
All modalities	80.83 ± 1.13	83.60 ± 1.08	85.71 ± 0.63	87.74 ± 0.60	83.88 ± 0.93	87.00 ± 0.78	90.74	**92.28**	69.14	74.20

Baseline	Random Guessing		Sophisticated guessing	
	F_1-score	Accuracy	F_1-score	Accuracy
	47.96	50.00	41.15	69.94

Table 5: Confusion matrix of the best setup trained on the three-class problem.

True / Estimated	Baseline	Stress	Amusement
Baseline	64577	1408	4444
Stress	3968	34997	899
Amusement	12153	2374	7773

modalities (three-class accuracy: 76 %, binary accuracy: 88 %). When both wrist- and chest-based physiological modalities are combined, an accuracy of 79 %/92 % is reached for the three-class/binary problem, respectively. This is no improvement compared to results achieved using only the chest-based physiological modalities. This indicates that if the chest-based modalities are available, the wrist-based modalities become redundant. Nevertheless, the classification scores reached using only the physiological wrist-based modalities are promising, especially considering the minimal intrusive nature of the device used.

Overall, the best performance result (in terms of accuracy) on each of the classification task is:

- 80.34 % (three-class problem, using all chest-based physiological modalities, AB classifier)
- 93.12 % (binary case, using all chest-based physiological modalities, LDA classifier)

These results are comparable to the work of *Gjoreski et al.*[5], who reported an accuracy of 72 % on a three-class problem (no, low, and high stress) and an accuracy of 83 % in the binary case. In Table 5 the confusion matrix of the best classifier trained on the three-class classification problem is displayed. The values indicate that the classifier was able to distinguish well between the baseline and the stress class. However, distinguishing between the classes baseline and amusement was difficult. The explanation for this is twofold. First, the physiological changes elicited by amusement are subtle. Second, the self-reports indicate (see Table 2) that the subjects' affective state was less influenced by the amusement condition compared to the stress condition.

Using all physiological features and the LDA classifier, the subject-specific accuracies range from 69% to 98% and from 82% to 100%, in the three-class classification problem and the binary case, respectively. However, only weak correlations were found between the subject-specific accuracies and the self-report value (e.g. arousal/valence) differences between the various affective states. Nevertheless, the large inter-subject differences emphasise the need for personalisation methods.

In order to assess the feature importance, a decision tree was trained for both the three-class and the binary classification task, using all available sensor modalities as input. The feature importance is computed according to the Gini importance (which reflects the reduction of the Gini criterion brought by the feature under consideration). The results of this experiment are displayed in Table 6. In both cases (three-classes and binary classification) the two most important features ($\sigma_E^{RESP,}$, and μ_{HR}^{ECG}) were alike. This suggests that the classifier in the three-class problem first learned to distinguish between *stress* and *non-stress* states, before it learned to classify the *baseline* and *amusement* classes.

Table 6: Feature importance for the three-class and binary classification task considering all modalities.

Importance	Three-class	Importance	Binary Task
0.23	$\sigma_E^{RESP,\,chest}$	0.35	$\sigma_E^{RESP,\,chest}$
0.11	$\mu_{HR}^{ECG,\,chest}$	0.20	$\mu_{HR}^{ECG,\,chest}$
0.07	min_{TEMP}^{wrist}	0.09	max_{TEMP}^{wrist}
0.06	$\mu_{ACC,\,3D}^{chest}$	0.07	$range_{EDA}^{wrist}$
0.05	$range_{EDA}^{wrist}$	0.05	$\#_{SCR}^{chest}$

6 CONCLUSION

We presented WESAD, a multimodal dataset for wearable stress and affect detection. In contrast to other available datasets, WESAD features all physiological modalities commonly integrated in commercial and medical devices: blood volume pulse (BVP), electrocardiogram (ECG), electrodermal activity (EDA), electromyogram (EMG), respiration (RESP), body temperature (TEMP), and three-axis acceleration (ACC). By using these modalities, we hope that our dataset will enable and support the development of new affect recognition systems. The study protocol aimed at inducing three different affective states (neutral, stress, amusement). Self-reports on these states were collected from the study participants.

For benchmarking, we used standard physiological and motion features and common machine learning methods. On a three-class (*baseline vs. stress vs. amusement*) problem we achieved classification accuracies of up to 80 %. Considering a binary classification problem (*stress vs. non-stress*), accuracies of up to 93 % were reached. These results should be interpreted with caution due to the limitations of WESAD, regarding the number of subjects and the lack of age and gender diversity. Nevertheless, since using the LOSO evaluation scheme, our results indicate that generalisation is possible. We also performed a detailed analysis on the importance of the two device locations as well as the different sensor modalities. Our results suggest that a chest-based device leads to the overall best classification results and by adding data of a wrist-based device no further improvement is achieved. However, the results obtained using only a wrist-based device are promising, especially considering the minimal intrusive nature of such a device.

Further work is required to take the self-reports into account. These self-reports could be used to create personalised models which are able to predict the affective state of a specific person. Moreover, the meditation period could be added as an additional class, posing a four-class classification problem. The dataset introduced in this paper is publicly available, and can be downloaded from https://ubicomp.eti.uni-siegen.de/home/datasets/icmi18/. We invite the research community to consider it for algorithm development and benchmarking.

ACKNOWLEDGMENT

We would like to thank Rahel Milla and Esther Bosch for many fruitful discussions. Furthermore, we thank all the study participants for their participation.

REFERENCES

[1] 2016. HSE on work related stress. http://www.hse.gov.uk/statistics/causdis/-ffstress/index.htm. (2016). Accessed: 2017-09-06.

[2] B. Barker, H. Barker, and A. Wadsworth. 1977. Factor analysis of the items of the state-trait anxiety inventory. *Journal of Clinical Psychology* 33, 2 (1977), 450–455.

[3] J. Choi, B. Ahmed, and R. Gutierrez-Osuna. 2012. Development and evaluation of an ambulatory stress monitor based on wearable sensors. *IEEE Transactions on Information Technology in Biomedicine* 16, 2 (2012).

[4] G. Chrousos and P. Gold. 1992. The concepts of stress and stress system disorders: overview of physical and behavioral homeostasis. *Jama* 267, 9 (1992), 1244–1252.

[5] M. Gjoreski, H. Gjoreski, and M. Gams. 2016. Continuous stress detection using a wrist device: In laboratory and real life. In *UbiComp '16*. 1185–1193.

[6] J. Healey and R. Picard. 2005. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems* 6, 2 (2005), 156–166.

[7] W. Helton and K. Näswall. 2015. Short stress state questionnaire. *European Journal of Psychological Assessment* (2015).

[8] K. Hovsepian, M. al'Absi, and S. Kumar. 2015. cStress: Towards a gold standard for continuous stress assessment in the mobile environment. In *UbiComp '15*. 493–504.

[9] A. Johnson and E. Anderson. 1990. Stress and arousal. (1990).

[10] J. Kim and E. André. 2008. Emotion recognition based on physiological changes in music listening. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 12 (2008), 2067–2083.

[11] C. Kirschbaum, K. Pirke, and D. Hellhammer. 1993. The Trier Social Stress Test – a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology* 28, 1-2 (1993), 76–81.

[12] S. Koelstra, C. Muhl, and I. Patras. 2012. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing* 3, 1 (2012), 18–31.

[13] S. Kreibig. 2010. Autonomic nervous system activity in emotion: A review. *Biological Psychology* 84, 3 (2010), 394–421.

[14] H. Lu, D. Frauendorfer, and T. Choudhury. 2012. StressSense: Detecting stress in unconstrained acoustic environments using smartphones. In *UbiComp'12*. 351–360.

[15] M. Malik. 1996. Task force of the European society of cardiology and the north American society of pacing and electrophysiology. Heart rate variability. Standards of measurement, physiological interpretation, and clinical use. *Eur Heart J.* 17 (1996), 354–381.

[16] B. McEwen and E. Stellar. 1993. Stress and the individual: mechanisms leading to disease. *Archives of Internal Medicine* 153, 18 (1993), 2093–2101.

[17] S. Mirsamadi, E. Barsoum, and C. Zhang. 2017. Automatic speech emotion recognition using recurrent neural networks with local attention. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2227–2231.

[18] A. Muaremi, B. Arnrich, and G. Tröster. 2013. Towards measuring stress with smartphones and wearable devices during workday and sleep. *BioNanoScience* 3, 2 (2013), 172–183.

[19] R. Picard, E. Vyzas, and J. Healey. 2001. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23, 10 (2001), 1175–1191.

[20] K. Plarre, A. Raij, and M. Scott. 2011. Continuous inference of psychological stress from sensory measurements collected in the natural environment. In *10th International Conference on Information Processing in Sensor Networks (IPSN)*. 97–108.

[21] A. Reiss and D. Stricker. 2012. Introducing a new benchmarked dataset for activity monitoring. In *16th International Symposium on Wearable Computers (ISWC)*. 108–109.

[22] R. Rosmond and P. Björntorp. 1998. Endocrine and metabolic aberrations in men with abdominal obesity in relation to anxio-depressive infirmity. *Metabolism* 47, 10 (1998), 1187–1193.

[23] J. Russell. 1979. *Affective space is bipolar.* American Psychological Association.

[24] A. Samson, S. Kreibig, and J. Gross. 2016. Eliciting positive, negative and mixed emotional states: A film library for affective scientists. *Cognition and Emotion* 30, 5 (2016), 827–856.

[25] H. Selye. 1976. Stress without distress. In *Psychopathology of Human Adaptation*. 137–146.

[26] C. Setz, B. Arnrich, J. Schumm, R. La Marca, G. Tröster, and U. Ehlert. 2010. Discriminating stress from cognitive load using a wearable EDA device. *IEEE Transactions on Information Technology in Biomedicine* 14, 2 (2010), 410–417.

[27] F.-T. Sun, C. Kuo, H.-T. Cheng, S. Buthpitiya, P. Collins, and M. Griss. 2012. Activity-aware mental stress detection using physiological sensors. In *Mobile Computing, Applications, and Services*. 211–230.

[28] N. A. Taylor and C. A. Machado-Moreira. 2013. Regional variations in transepidermal water loss, eccrine sweat gland density, sweat secretion rates and electrolyte composition in resting and exercising humans. *Extreme Physiology & Medicine* 2, 4 (2013).

[29] P. Tzirakis, G. Trigeorgis, and S. Zafeiriou. 2017. End-to-end multimodal emotion recognition using deep neural networks. *CoRR* (2017).

[30] D. Watson, L. Clark, and A. Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology* 54, 6 (1988), 1063.

[31] J. Wijsman, B. Grundlehner, and H. Hermens. 2010. Trapezius muscle EMG as predictor of mental stress. In *Wireless Health 2010 (WH '10)*. 155–163.

[32] A. Zenonos, A. Khan, and M. Sooriyabandara. 2016. HealthyOffice: Mood recognition at work using smartphones and wearable sensors. In *PerCom Workshops*.