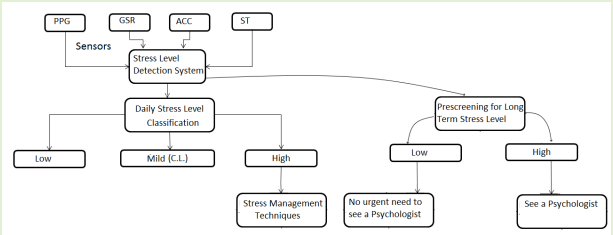
Real-Life Stress Level Monitoring Using Smart Bands in the Light of Contextual Information

Yekta Said Can, Niaz Chalabianloo, Deniz Ekiz, Javier Fernández-Álvarez, Claudia Repetto, Giuseppe Riva, Heather Iles-Smith, and Cem Ersoy

***Abstract*—An automatic stress detection system that uses unobtrusive smart bands will contribute to human health and well-being by alleviating the effects of high stress levels. However, there are a number of challenges for detecting stress in unrestricted daily life which results in lower performances of such systems when compared to semi-restricted and laboratory environment studies. The addition of contextual information such as physical activity level, activity type and weather to the physiological signals can improve the classification accuracies of these systems. We developed an automatic stress detection system that employs smart bands for physiological data collection. In this study, we monitored the stress levels of 16 participants of an EU project training every day throughout the eight days long event by using our system. We collected 1440 hours of physiological data and 2780 self-report questions from the participants who are from diverse countries. The project midterm presentations (see Figure 3) in front of a jury at the end of the event were the source of significant real stress. Different types of contextual information, along with the physiological data, were recorded to determine the perceived stress levels of individuals. We further analyze the physiological signals in this event to infer long term perceived stress levels which we obtained from baseline PSS-14 questionnaires. Session-based, daily and long-term perceived stress levels could be identified by using the proposed system successfully.**



***Index Terms*—Commercial smartwatch, mental stress, psychophysiological, emotion regulation, heart rate variability, electrodermal activity.**

I. Introduction

WEARABLE devices help measure and reduce stress, leading to significant improvements in human health and well-being. Personal health monitoring is among the most prominent ones in these fields. Researchers obtained the ability to track physical activities, well-being, daily routines with these devices. By using this information, we can improve the life quality of individuals with insightful suggestions and interventions. Stress is one of the most severe work-related health problems in Europe [1]; it is the second commonly seen issue after musculoskeletal diseases which could be caused by stress in some situations [2].

Researchers have started to use commercial smart wearable devices for detecting the stress of individuals from physiological signals. Heart rate variability and electrodermal activity are the most commonly used physiological signals in the literature [3]. Stress detection studies started in laboratory environments. The research then directed towards controlled real-life environments. Office, automobile, and classrooms are selected because they can be controlled with cameras and sensors and movements are restricted. Major research in office environments can be listed as [4] and [5]. Campus environments compose of both controlled and uncontrolled areas. They are the bridge between controlled and unrestricted daily life studies. The most prominent studies in those environments are [6], [7] and [8]. Another work in those semi-constrained settings was presented by Can *et al.* [9]. They developed a three-class stress detection framework using commercial smartwatches and wristbands. The proposed system was tested in a semi-constrained real-life setting which was an algorithmic summer camp. They achieved approximately 98%

Manuscript received October 16, 2019; revised March 27, 2020; accepted March 28, 2020. Date of publication March 31, 2020; date of current version July 6, 2020. This work was supported in part by the European Union's Horizon 2020 Research and Innovation Programme through the Marie Skłodowska-Curie under Grant 722022 and in part by the Turkish Directorate of Strategy and Budget through the TAM Project under Grant DPT2007K120610. The associate editor coordinating the review of this article and approving it for publication was Prof. Aime Lay-Ekuakille. *(Corresponding author: Yekta Said Can.)*

Yekta Said Can, Niaz Chalabianloo, Deniz Ekiz, and Cem Ersoy are with the Computer Engineering Department, Boğaziçi University, 34342 Istanbul, Turkey (e-mail: yekta.can@boun.edu.tr).

Javier Fernández-Álvarez, Claudia Repetto, and Giuseppe Riva are with the General Psychology and Communication Psychology Department, Catholic University of Milan, 20123 Milan, Italy.

Heather Iles-Smith is with the Leeds Teaching Hospitals NHS Trust/University of Leeds, Leeds LS2 9JT, U.K.

Digital Object Identifier 10.1109/JSEN.2020.2984644

classification accuracy in the task of differentiating high stress, medium stress and free day (relaxed sessions in the middle day of the event).

Recently, there are a few studies in unconstrained real-life. Gjoreski *et al.* [10] created a stress recognition system for the laboratory environment with an activity recognition model. The physiological signals were recorded during the experiment. First, they asked subjects to stay relaxed. Then they gave a mental arithmetic task to the subjects. The calibration of the subjective assessment is made with the scores gathered from the mental arithmetic task phase. They trained their model with these data. The model makes a decision every 20 minutes. Electrodermal activity (EDA), blood volume pressure (BVP), heart rate (HR), temperature, and inter-beat interval (IBI) data were obtained. The best performing model achieved a 70% recall and 95% precision. Gimpel *et al.* [11] proposed a stress recognition method by collecting smartphone data. The researchers declared that the most significant distinction from the previous studies is that their method is not based on the user declaration or supplementary wearables.

The performance of unconstrained real-life studies where the models are self-report based is lower than those in controlled environments. The reasons for this can be counted as the subjectivity of self-reports, signal distortions caused by unlimited movements and unknown context of the users [3]. In order to increase the classification performance of the proposed systems, researchers added contextual information to physiological signals. The activity type is the most commonly used information. Gjoreski *et al.* [12] proposed a method for continuous detection of stressful events using data provided from a commercial wrist device in both laboratory and real-life. They used the Empatica [13] wrist device. In the laboratory setting, a mental arithmetic task is given to the participant in order to induce stress. Features from BVP, HRV, ST (Skin Temperature), EDA and inter-beat (RR) intervals signals were computed. They achieved 83% classification accuracy on the two class problem. They achieved 72% classification accuracy with the 3-class problem. An activity recognition model that discriminates sitting, walking, running and cycling was used for everyday life settings. The activity recognition model is used to differentiate high physical activity from a mental stress elevated case. They achieved 92% classification accuracy with the model that includes the activity recognition. The training data gathered from 5 participants for 11 days.

Kostopoulos *et al.* [14], used a smart-phone based data collection application to assess stress levels. They also collected context information and surveys from the smart-phone. Mishra *et al.* [15] developed an automatic stress detection application by using heart activity signals obtained from a Moto360 smartwatch. They collected data from 23 subjects in the duration of 3 days. To improve their performance, they added activity type to physiological features. They further investigated the time of the day and day of the week and stress level relationship and they could not find any meaningful correlation. However, they demonstrated that contextual information is crucial, especially in unconstrained real-life studies.

We developed a stress level detection scheme using physiological signals. We first filtered the physiological signals to clean artifacts caused by environmental factors and the movement of individuals by developing modality-specific algorithms. We then extracted features from these signals and applied different machine learning algorithms to classify stress levels. To test our system, we collected physiological data of 16 international PhD students during the eight days of training. 1440 hours of physiological signals from Empatica E4 smart bands and 2780 questions from the self-reported questionnaires (three Nasa-TLX questionnaires and one daily questionnaire each day from the participants) are collected. To the best of our knowledge, this work is the first automatic stress detection application which takes advantage of different kinds of contextual information along with physiological signals in the real-life environment using commercial smart bands. The major research contributions of this study are the following:

- Measuring the daily and session-based perceived stress levels by using only physiological signals and combining the contextual information (weather, physical activity level and activity type) with them
- Developing a prescreening tool for long-term perceived stress levels

The structure of the remaining of the paper is as follows. In Section 2, the related work in daily life stress detection is presented. Our method for stress level detection and management and the data collection procedure are described in Section 3. Experimental results and discussion are presented in Section 4. We present the conclusion and future works in Section 5.

II. Proposed System

A. Unobtrusive Stress Detection System With Smart Bands

The stress detection mechanism implemented in this work offers the flexibility of informing the users about their stress levels in their everyday activities with no extra interruptions or constraints created by the system. The only obligation is the use of a wearable smart wristband. These instruments were worn on the participants' non-dominant hand. The smart band provides BVP, ST, EDA and 3D acceleration data. The artifacts are identified and processed. The features are obtained from the sensory signals and fed to the stress predictor machine learning algorithms. Pre-trained machine learning models are needed to use this system. Model training was performed by running the machine learning algorithms on the feature vectors with generated class labels. Figure 1 depicts the block diagram of the proposed scheme.

1) EDA Preprocessing Artifact Detection and Removal Methods: The SC (Skin Conductance) signal is highly susceptible to be contaminated by intense physical activity and changes in temperature. Consequently, impacted segments must be filtered out from the original signal. We used an EDA toolkit [16] to identify the artifacts in the SC signal, this toolkit is 95 percent accurate in the artifact detection. Technicians have labeled the artifacts manually while developing this tool and by using these labels, a machine learning model has been trained. Furthermore, as well as SC signals, 3D acceleration, and ST signals are also used to detect artifacts.

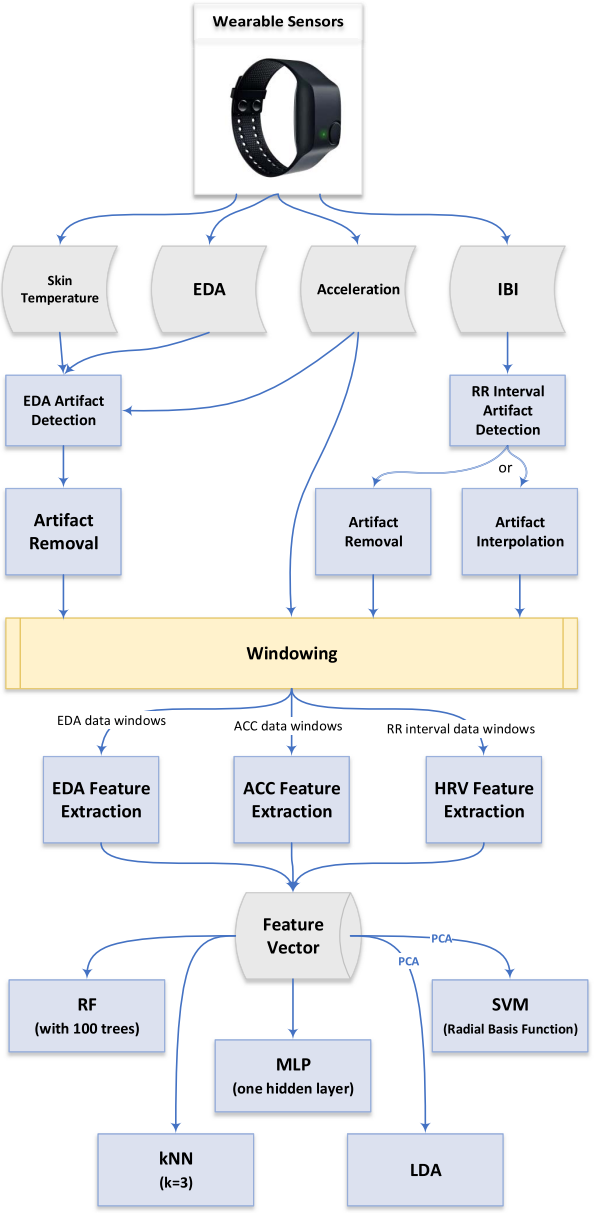


Fig. 1. Block diagram of the stress level detection system with Empatica E4 wristband.

Segments identified by this toolkit as artifacts were removed from the signals. To further enhance the capabilities of the artifact detection toolkit, we added the batch processing and segmentation features to this tool using a custom software built-in Python 2.7.

2) EDA Feature Extraction Methods: Features are obtained from the EDA signals which have undergone the artifact removal phase. Features are extracted from both components of the EDA signal which are phasic and tonic. To decompose the signal into these components, the cvxEDA tool [17] is employed. This tool utilizes convex optimization to predict the activity of the Autonomous Nervous System (ANS) using Bayesian statistics. The extracted EDA features are the Mean Tonic (Mean of the tonic component), SD Tonic (Standard deviation of the tonic component), Perc20 (20th percentile of the phasic component), Perc80 Tonic (80th percentile of the tonic component), Quartdev Tonic (Quartile deviation 75 percentile - 25 percentile of the tonic component), Strong Peaks (The number of strong peaks per 100 seconds) and Peaks Phasic (The number of peaks per 100 seconds).

a) Tonic component features: In the EDA signal, the tonic component reflects the long-term slow variations. This component is also referred to as the skin conductance level (SCL). It could be considered as the indicator of general psychophysiological activation [18].

b) Phasic component features: The phasic component in the SC signal reflects quicker variations (related to events). In response to a stimulus, skin Conductance Response (SCR) is the peaks of the phasic SC component [18]. The peak-related characteristics are obtained from the decomposed phasic component from the EDA signals.

3) Heart Activity Preprocessing Artifact Detection and Removal Methods: Improperly worn wristbands and ceaseless movements of the subjects also contaminate the heart activity signals obtained from smart wristbands. We created an artifact handling tool in MATLAB with the batch processing capacity to solve this intricacy. MATLAB built-in tools and Marcus Vollmer's HRV toolbox [19] are used for processing the heart activity signal. First, with 50 percent overlapping, the data is split into 2 minutes long segments. After the segmentation stage, the artifact detection percentage rule (also used in Kubios [20]) is utilized. In this rule, each data point is compared with the local average around it. The data point is marked as an artifact if the difference is higher than a predetermined threshold percentage. The common threshold value widely used in the literature is defined as 20% difference [5]. In our physiological signal analysis toolkit, we removed the inter-beat intervals marked as artifacts.

a) Time domain features: HRV's time-domain features from the RR interval time series are calculated. The most distinctive features are selected from the literature [10], [20], [21] and [22] which are the Mean RR (Mean value of the RR intervals), STD RR (standard deviation of the inter-beat interval), RMSSD (Root mean square of successive difference of the RR intervals), pNN50 (Percentage of the number of successive RR intervals varying more than 50ms from the previous interval), HRV triangular index (Total number of RR intervals divided by the height of the histogram of all RR intervals measured on a scale with bins of 1/128 s), TINN (Triangular interpolation of RR interval histogram) and SDSD (Related standard deviation of successive RR interval differences).

b) Frequency domain features: Since some of the heart peaks are missing due to the artifacts, we first interpolated the RR intervals to 4Hz before extracting the frequency domain features. Then the Fast Fourier Transform (FFT) was used. As a second option, we applied the Lomb-Scargle periodogram, which is implemented to convert for non-equidistant sampled signals to the frequency domain. We extracted features by using both methods. Extracted frequency domain HRV features are LF (Power in the low-frequency band 0.04-0.15 Hz), HF (Power in the high-frequency band 0.15-0.4 Hz), LF/HF (Ratio of LF-to-HF), pLF (Prevalent low-frequency oscillation of heart rate), pHF (Prevalent

high-frequency oscillation of heart rate), VLF (Power in the very low-frequency band 0.00-0.04 Hz).

c) Wavelet domain features: We also extracted wavelet-based features. Among these features, four of them are Autoregressive model (AR) coefficients of order four [23]. Sixteen features are Shannon entropy (SE) values for the maximal overlap discrete wavelet packet transform (MODPWT) at level 4 [24]. Two features are from Multifractal wavelet leader estimates of the second cumulant of the scaling exponents and the range of Holder exponents, or singularity spectrum [25]. Lastly, multiscale wavelet variance estimates are extracted [26]. An unbiased estimate of the wavelet variance is employed. It needs that only levels with at least one wavelet coefficient unaffected by border conditions are employed in the variance estimations. For each window, four features are extracted with the Daubechies wavelet. In total, we have 26 wavelet-based features: 4 AR features, 16 Shannon entropy values, two fractal estimates, and four wavelet variance estimates.

4) Accelerometer Feature Extraction Methods: The sensor data of the accelerometer is used for two distinct reasons. First, we obtained the features shown in Figure 1 from this sensor. This sensor was also used, as stated above, to cleanse the EDA signal. Extracted accelerometer features are Mean X (Mean acceleration over x axis), Mean Y (Mean acceleration over y axis), Mean Z (Mean acceleration over z axis), Mean ACC MAG (Mean acceleration over acceleration magnitude axis) and Energy (FFT energy over mean acceleration magnitude). We further obtained step count and stillness features from the accelerometer data. Stillness, which is a value from 0 to 1 indicates the ratio of activeness throughout a session.

5) Skin Temperature: The skin temperature data is employed for artifact detection in the EDA signal. After our data has been split into segments, various modalities should be merged into one feature vector. Particularly, the heart activity signal which begins with a delay to calculate heartbeat per minute in the beginning. All signals must be synchronized. For each segment, we included start and end timestamps and each modality is merged using a custom python script.

6) Machine Learning Classifier Algorithms: To discriminate stress from the cognitive load, the Weka machine learning toolkit [27] is employed. There are several pre-processing features in the Weka toolkit which can be applied to the data before beginning the classification process. Based on the number of instances in each class, our dataset is not balanced. By removing samples from the majority class, we resolved this problem. Hence, the biasing of the classifiers towards the class with more instances was prevented. In this research, we used five distinct classification algorithms to detect distinct levels of stress: MultiLayer Perceptron (MLP), Random Forest (RF) (with 100 trees), K- nearest neighbors (kNN) (n=1-4), Linear discriminant analysis (LDA), Principal Component Analysis (PCA) and support vector machine (SVM) with a radial basis function. 10-fold stratified cross-validation was used. The machine learning algorithms' hyperparameters are fine-tuned with grid search. Finally, the best performing models are listed.

7) Dimensionality Reduction: We have 41 (26 from wavelet-based, 15 from time and frequency domains) features extracted from the HRV signal. We implemented a correlation-based feature selection (CBFS) on the combined HRV features. CBFS is accessible via the Weka machine learning package [27]. The CBFS technique excludes the features that are less correlated with that of the output class. We chose the fifteen most significant features for combined HRV features (see Figure 2). This approach is implemented for all classifiers. We further performed PCA-based dimensionality reduction to generate models based on LDA and SVM, where the covered variance is chosen as 0.95. Since we have seven features from the EDA signal, feature selection and dimensionality reduction methods are not implemented for EDA features.

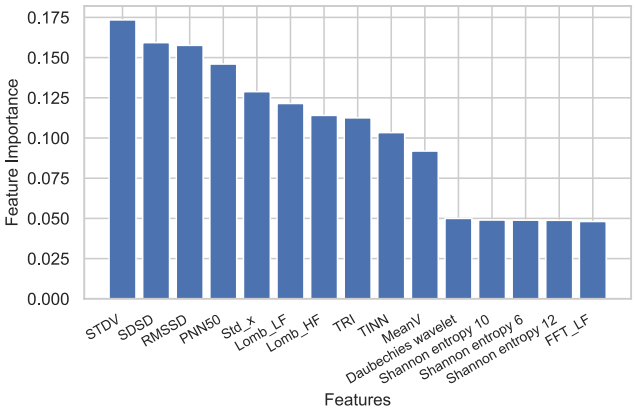


Fig. 2. The 15 best features selected for the HRV signal.

B. Description of the Data Collection Procedure

Analysis of the data obtained during the eight-day AffecTech training event in Istanbul-Turkey was carried out to assess our proposed stress level monitoring system in real-life environments. AffecTech is a Marie Skłodowska-Curie Innovative Training Network program funded by Horizon 2020 framework funded by the European Commission. The AffecTech initiative is an international cooperative study network aimed at developing and improving private and person-specific and eventually low-cost yet efficient wearable health systems to assist people suffering from affective disorders such as depression, anxiety, and bipolar disorders. We collected 16 participants' physiological data in this research. One participant had to leave the session because of a fault on one of the Empatica E4 devices. However, fifteen subjects successfully finished all stages of the data collection session. All of the participants were PhD students with distinct areas of research and expertise within the associated disciplines. Participants come from various countries with different nationalities (two from Iran, two from Spain, two from Italy, one from Argentina, one from Pakistan, one from Switzerland, one from Belarus, one from France, one from England, one from Barbados, one from Turkey and one from Bulgaria). We emphasize participants' distinct nationalities because different cultures can have different stress responses and to the best of our knowledge, our research is the first to evaluate this sort of diversified data with wearable devices. This event was held for eight consecutive days. Before the research, the 14 item version

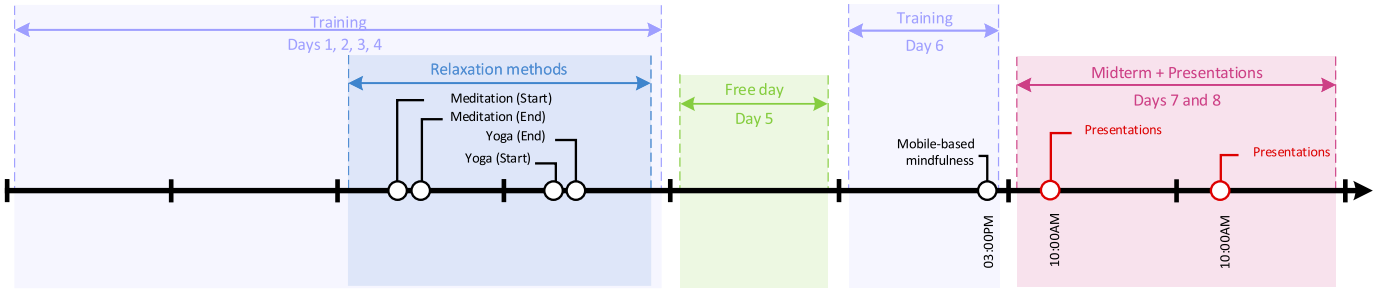


Fig. 3. Time-line depicting eight days of the training event. Presentations, relaxations and lectures are highlighted.

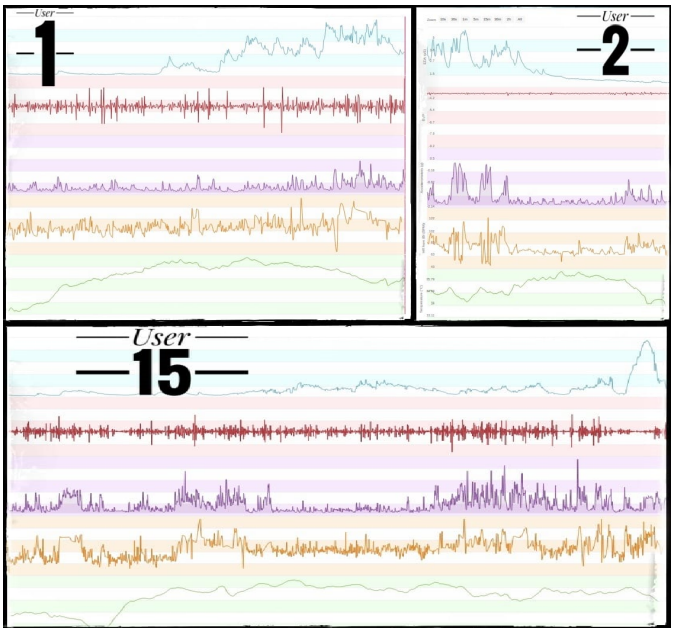


Fig. 4. Sample data collected from three different users in the presentation event. The turquoise signal in the top is EDA, the purple one in the middle is acceleration, the orange one is average HR and the bottom signal is ST. The second signal from the top (shown in red) is the raw blood volume pressure data.

of the Perceived Stress Scale (PSS) [28] was obtained from each participant. This questionnaire is adopted as the baseline. Session-based self-reports comprised of six questions from the Nasa Task Load Index (NASA-TLX) [29], daily perceived stress questionnaire (a mixture of questions from [30], [31] and [32]), and the physiological signal data from Empatica E4 (see Figure 4), all were collected during the event. The gender split is six females and nine males. The participants' average age is 28. A total of 2780 self-report responses were collected (from three session-based and one daily questionnaire per day). The training week focused on clearly specified training tasks and pursuits. An innovative set of design and implementation workshops and training programs were planned and implemented in multiple week-long series of informative workshops and presentations to ensure that the fellows have developed the required target skills, knowledge, and values. Participants obtained practical experience in installing and using wearables and, subsequently, analyzing their sensor data. Participants had to present their prior works to two evaluators from the European Union at the end of the training week, where they received feedback on their progress.

To study the impacts of the emotion regulation on stress, sessions of yoga, guided mindfulness and mobile-based mindfulness were conducted. The event's timeline is shown in Figure 3.

C. Ethics

The procedure used in this research is endorsed with the approval number 2018/16 by Boğaziçi University's Institutional Review Board for Research with Human Subjects. Each participant received a consent form prior to the data acquisition explaining the experimental procedure and its benefits and implications for both the society and the subject. The procedure was also vocally described. All data are also stored anonymously.

D. Data Description

The psychophysiological signal data was obtained using the Empatica E4 smart band while subjects were awake throughout the eight days of the AffecTech training. The data included IBI, EDA, ACC (Accelerometer), and ST. 27.39% of the data is collected from free times (free day and after training until subjects slept 17:00-22:00), 43.83% of the data comes from lectures in training, 11.41% is the presentation session and relax sessions consist of 17.35% of the data. The data is randomly undersampled (most commonly used technique [33]) in order to overcome the class imbalance problem.

E. Measuring Perceived Stress Levels Using Different Types of Self-Reports

The perceived stress of individuals was measured throughout the event with two types of self-reports.

1) Session-Based Self-Report for Perceived Stress Measurement: The first collected self-report is the Frustration item of the raw NASA-TLX [34]. We asked the following question to the participants for each session:

> *How irritated, stressed, and annoyed versus content, relaxed, and complacent did you feel during the task?*

We measured the session based perceived stress levels by using the first question. In order to validate that the participants experienced different perceived stress levels in different contexts (lecture, relaxation, presentation), we used the

TABLE I
T-TEST RESULTS FOR SESSION TUPLE COMPARISON OF PERCEIVED STRESS LEVELS USING SELF-REPORTS

Session Tuple	t-test statistic	p-value
Yoga - Presentation	-4.0027	$p<0.005$
Guided Mindfulness - Presentation	-5.4905	$p<0.005$
Mobile Mindfulness - Presentation	-4.2677	$p<0.005$

Frustration item (see Section 4.5) from the NASA-TLX [34]. The distribution of answers is demonstrated in Figure 6. We aim to demonstrate that the induced perceived stress levels (obtained from self-report answers) differ in relaxation, lecture and presentation sessions (high stress). Thus, we used the t-test (in R programming language) to the perceived stress self-report responses of relaxation types (mobile mindfulness, yoga and traditional mindfulness) versus presentation pairs. The paired t-test is applied to assess the separability of relaxation (low/no stress context) and presentation (high stress context). The degree of freedom is 15. We used the variance test for each session tuple, and equal variance could not be identified in any of the tuples. So, the variance is selected as unequal. 99.5% confidence intervals were employed. The t-test results (P-values and test statistics) are presented in Table I. For all tuples, the null hypothesis stating that the perceived stress of the relaxation method is not less than the presentation session is rejected. We can infer that the perceived stress levels of participants are reduced during relaxation sessions and increased during presentation sessions. Different stress levels are induced in each session of the training event.

2) Daily Self-Report Questionnaire for Perceived Stress Measurement: The daily self-report questionnaire was comprised of six questions. Two questions were measuring rumination [32], the other two questions were measuring worry [30] and the last two questions were measuring reappraisal [31]. We selected rumination, worry and reappraisal questions from mentioned prominent questionnaires (Brief State Rumination Inventory [32], State-Reappraisal Inventory [31] and Penn State Worry Questionnaire [30]) because they are the main components for causing stress and they are linked to depression and anxiety [35]. We measured the daily perceived stress levels of the participants by using the collected daily questionnaire which is demonstrated below. The response measure was the Likert scale with answers from 1 to 6.

Worry Question 1 My worries are overwhelming me
Worry Question 2 I know I should not worry about things, but I just cannot help it
Rumination Question 1 Right now, it is hard for me to shut off negative thoughts
Rumination Question 2 Right now, I am thinking: "why can't I handle things better?"
Reappraisal Question 1 I'm trying to think that things could be much worse
Reappraisal Question 2 I'm trying to think of positive aspects of the events

Fig. 5. The data collection device is shown. The electrodes in the bottom right image are for EDA sensors. PPG and ST sensors are presented in the bottom middle image.

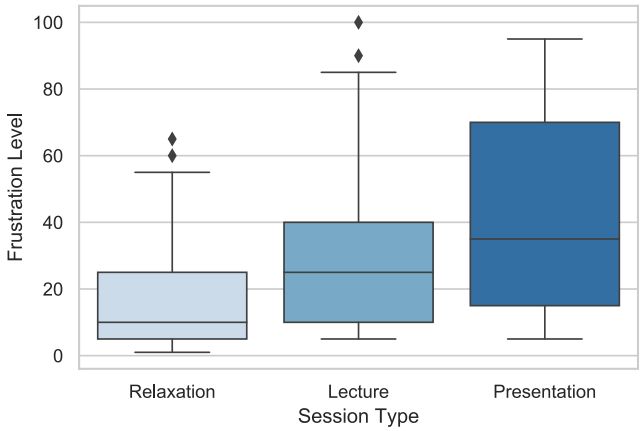


Fig. 6. The distribution of self reports for different session types.

III. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we examined the effect of context in measuring session-based (3 hours) perceived stress levels and in predicting daily stress levels separately. We further developed and tested a long-term perceived stress level pre-screening tool by evaluating physiological signals.

A. Measuring the Session-Based Perceived Stress Levels

To enhance the performance of our system, we made use of the contextual information. Weather information and known context (activity type i.e., lecture, presentation, relaxation activities) are added to the physiological features to improve the performance of the session-based perceived stress level detection system. The performance is evaluated without context, in the light of weather related context and with activity type information.

1) Measuring Perceived Stress Levels Without Context: We collected the session-based self-reports during the event in every session. We further asked the participants to fill these self-reports in the evening in their free time while wearing Empatica E4 wristbands (see Figure 5). 2-class and 3-class session-based perceived stress scores are presented

TABLE II
PREDICTING THE 3-CLASS PERCEIVED STRESS LEVEL (PSL) FROM THE HRV (TIME AND FREQUENCY DOMAIN FEATURES) AND THE EDA DATA COLLECTED FROM THE EVENT. THE PSL IS CALCULATED FROM THE FRUSTRATION SCALE. WEATHER-RELATED FEATURES FOR THESE SESSIONS ARE ADDED

Algorithm	Accuracy			
	HRV	HRV + weather	EDA	EDA + weather
MLP	47.74	54.05	46.28	55.42
RF	51.35	59.46	36.57	56.17
kNN	53.15	**60.36**	43.42	57.71
LDA	51.35	59.46	45.14	57.14
SVM	51.35	**60.36**	45.71	52.00

TABLE III
PREDICTING THE 2- CLASS PERCEIVED STRESS LEVEL (PSL) FROM THE HRV (TIME AND FREQUENCY DOMAIN FEATURES) AND EDA DATA COLLECTED FROM THE EVENT. THE PSL IS CALCULATED FROM THE FRUSTRATION SCALE. WEATHER-RELATED FEATURES FOR THESE SESSIONS ARE ADDED

Algorithm	Accuracy			
	HRV	HRV + weather	EDA	EDA + weather
MLP	69.62	**81.01**	50.00	72.29
RF	55.69	74.68	57.43	68.92
kNN	59.49	75.94	54.73	68.24
LDA	53.16	69.62	50.68	71.62
SVM	64.55	75.94	49.32	56.08

in Tables III and II. The results obtained by adding wavelet based features to time and frequency domains are presented at Tables VI and VII. By using EDA and HRV signals, our system achieved a maximum 69.62% classification accuracy on 2-class and 53.15% 3-class classification. Another important finding is that the HRV signal achieves higher detection accuracies with all algorithms. Especially when we added wavelet based features to time and frequency domain features and applied a feature selection, maximum accuracy increases to 73.40% for 2-class and 65.53% for 3-class classification. Since we collected self-reports during the training event (10:00-17:00) and free time in the evening (17:00-22:00), our system suffered from the problems with perceived stress measurement in unconstrained real-life mentioned in [9]. The negative effect of stress on memory and subjectivity of self-reports are among the most prominent problems with self-reports and they contributed to the decrease in the performance of our system.

2) Session-Based Perceived Stress Measurement With Weather-Related Context: The association of the effects of the changes in weather conditions with stress and mood has been demonstrated in the literature [36], [37]. We further investigated the effect of weather-related context into our perceived stress detection system. Air Temperature at 2 meters high above the surface (degrees Celsius), atmospheric pressures at the weather station and sea level (millimeters of mercury), changes in atmospheric pressure over the last three hours, relative humidity at the height of 2 meters above the earth's surface, wind speed, total cloud cover and amount of precipitation data was extracted for each session and added to the physiological data. Weather information is gathered from the Windguru wind and weather forecasting website [38]. The weather-related context information increases performance of our system drastically as it can be seen in Tables II and III (accuracies also increase in the wavelet added results see Tables VI and VII). Air pressure, humidity, precipitation and cloud ratio are selected among the top features when used with EDA and HRV signals. These results demonstrated that weather has an impact on the perceived stress of individuals and provides additional information for a stress detection system.

TABLE IV
PREDICTING THE 3- CLASS PERCEIVED STRESS LEVEL (PSL) FROM THE HRV (TIME AND FREQUENCY DOMAIN FEATURES) AND EDA DATA COLLECTED FROM THE EVENT. THE PSL IS CALCULATED FROM THE FRUSTRATION SCALE. KNOWN CONTEXT IS ADDED TO THE FEATURES

Algorithm	Accuracy			
	HRV	HRV + known context	EDA	EDA + known context
MLP	47.74	53.69	46.28	61.90
RF	51.35	64.43	36.57	55.84
kNN	53.15	61.75	43.42	56.71
LDA	51.35	64.43	45.14	61.03
SVM	51.35	64.43	45.71	**65.36**

3) Finding Perceived Stress Levels by Adding the Known Context: The known context of individuals can be also used for improving stress detection systems as mentioned in Section 2. The EU training event has a lecture, presentation, relaxation and free (17:00-22:00) sessions. Since the context is unknown in free sessions, we did not use these sessions for this section. We enumerated the known context for these sessions as Relaxation:0, Lecture:1 and Presentation:2. We added these enumerated known context information to the physiological features to detect session-based perceived stress levels. As it can be seen in Tables IV and V, adding the known context information increases the system performance 20-25% (accuracies also increase in the wavelet added results see Tables VI and VII). These results demonstrate that the known context information is crucial for daily stress detection systems. We presented the best results in different forms after combining all features (HRV, known context, weather) and selecting the best feature set by using grid search in Tables VIII, IX, X and XI. In these tables, we presented the best results of our system with different metrics than accuracy such as Kappa statistics, f-measure, ROC Area and confusion matrices for 2-class and 3-class classifications.

B. Daily Stress Level Detection

In this section, we first detected the perceived daily stress level by using the questionnaire in Section II.D.2 as a self-report. In the second part, we differentiated the daily

TABLE V
PREDICTING THE 2- CLASS PERCEIVED STRESS LEVEL (PSL) FROM THE HRV (TIME AND FREQUENCY DOMAIN FEATURES) AND EDA DATA COLLECTED FROM THE EVENT. THE PSL IS CALCULATED FROM THE FRUSTRATION SCALE. KNOWN CONTEXT DATA FOR THESE SESSIONS ARE ADDED TO THE FEATURE VECTOR

Algorithm	Accuracy			
	HRV	HRV + known context	EDA	EDA + known context
MLP	69.62	69.33	50.00	74.03
RF	55.69	65.33	57.43	70.13
kNN	59.49	72.00	54.73	68.83
LDA	53.16	**76.00**	50.68	74.89
SVM	64.55	73.33	49.32	75.76

TABLE VI
PREDICTING THE 3- CLASS PERCEIVED STRESS LEVEL (PSL) FROM THE HRV DATA COLLECTED FROM THE EVENT. WAVELET BASED FEATURES ARE ALSO ADDED. THE PSL IS CALCULATED FROM THE FRUSTRATION SCALE. WEATHER-RELATED FEATURES AND KNOWN CONTEXT ARE ADDED TO THE FEATURE SET

Algorithm	Accuracy		
	HRV	HRV + weather	HRV + known context
MLP	53.10	54.80	67.23
RF	65.53	63.84	66.10
kNN	55.93	**65.53**	59.89
LDA	50.07	53.67	64.40
SVM	55.93	**67.23**	68.92

TABLE VII
PREDICTING THE 2- CLASS PERCEIVED STRESS LEVEL (PSL) FROM THE HRV DATA COLLECTED FROM THE EVENT. WAVELET BASED FEATURES ARE ALSO ADDED. THE PSL IS CALCULATED FROM THE FRUSTRATION SCALE. WEATHER-RELATED FEATURES AND KNOWN CONTEXT ARE ADDED TO THE FEATURE SET

Algorithm	Accuracy		
	HRV	HRV + weather	HRV + known context
MLP	71.75	78.53	79.66
RF	73.40	77.97	76.27
kNN	71.18	**76.84**	76.27
LDA	55.55	76.27	78.53
SVM	68.33	**79.66**	75.14

physiological signals with and without physical context information (step count and stillness). We chose three different days in the training event which have similar physical activity intensity for differentiation. The effect of physical activity related context is investigated.

1) Daily Perceived Stress Level Prediction by Evaluating Rumination, Reappraisal and Worry Elements: We collected self-reports every day to measure stress by evaluating

TABLE VIII
PREDICTING THE 2- CLASS PERCEIVED STRESS LEVEL (PSL) FROM THE BEST PERFORMED FEATURE SET. THE PSL IS CALCULATED FROM THE FRUSTRATION SCALE

	Accuracy	F-measure	ROC Area	Kappa Statistics
MLP	82.59	0.82	0.90	0.65
RF	75.92	0.75	0.76	0.51
kNN	72.00	0.73	0.72	0.48
LDA	69.36	0.69	0.70	0.43
SVM	71.15	0.71	0.71	0.45

TABLE IX
PREDICTING THE 3- CLASS PERCEIVED STRESS LEVEL (PSL) FROM THE BEST PERFORMED FEATURE SET. THE PSL IS CALCULATED FROM THE FRUSTRATION SCALE

	Accuracy	F-measure	ROC Area	Kappa Statistics
MLP	61.90	0.61	0.76	0.43
RF	61.03	0.60	0.74	0.42
kNN	56.71	0.58	0.68	0.38
LDA	61.03	0.60	0.72	0.42
SVM	65.36	0.61	0.76	0.44

TABLE X
CONFUSION MATRIX OF 3-CLASS PSL OF SVM CLASSIFIER

	low stress	medium stress	high stress
low stress	43	50	37
medium stress	15	94	21
high stress	4	4	122

TABLE XI
CONFUSION MATRIX OF 2-CLASS PSL OF MLP CLASSIFIER

	low stress	high stress
low stress	106	29
high stress	18	117

rumination, worry and reappraisal elements (see the questionnaire in Section II.D.2). These elements are selected because they contribute to the high-stress levels most. Physiological signals are collected from the participants and daily perceived stress levels (DPSL) of individuals collected from self-reports are used as ground truth labels to these signals. The 2-class DPSL classification accuracies are presented in Table XII. We achieved a maximum of 68.85% classification accuracy which is similar to the reported performances in the literature.

2) Daily Physiological Stress Level Detection With Physical Activity Related Contextual Information: Physical activity is known to reduce stress levels [39]. We selected three days of the training: one from the beginning (Day 2), one from the middle (D4) and one from the end (D8) and try to differentiate daily perceived stress levels by using the HRV features. We do not expect considerable changes in terms of physical activity stemming from the schedule. We further investigated the effect

TABLE XII
PREDICTING THE DAILY PERCEIVED STRESS LEVEL (DPSL) FROM THE PHYSIOLOGICAL DATA COLLECTED FROM THE EVENT. DPSL IS CALCULATED FROM THE QUESTIONNAIRE WHICH IS COLLECTED DAILY AND COMPOSES OF RUMINATION, WORRY AND REAPPRAISAL QUESTIONS

Algorithm	Accuracy	
	HRV	**EDA**
MLP	52.78	57.38
RF	50.00	**68.85**
kNN	58.33	62.31
LDA	44.54	65.58
SVM	47.22	59.62

TABLE XIII
DAILY STRESS LEVEL DIFFERENTIATION ACCURACIES BY USING THE ONLY HRV (TIME AND FREQUENCY DOMAIN FEATURES) AND WITH THE ADDITION OF PHYSICAL ACTIVITY RELATED CONTEXT DATA (STILLNESS AND STEP COUNT)

Algorithm	Accuracy		
	HRV	**HRV + Stillness**	**HRV + StepCount**
MLP	55.81	60.46	67.44
RF	69.76	72.09	**76.74**
kNN	53.49	53.49	65.12
LDA	62.79	62.79	74.42
SVM	60.47	65.11	72.10

of physical activity related contextual features to the performance of the daily perceived stress detection system. For this purpose, we extracted stillness and step count features from the accelerometer signal by using the EDAExplorer [16]. Stillness measures the daily physical activity of an individual. The range of stillness is between 0 and 1. Step count is also calculated from the accelerometer signal. When these contextual features are added to the signal, our DPSL detection accuracies are increased considerably (see Table XIII). These results show that physical activity related to the contextual features are also important for the daily perceived stress detection schemes.

C. Pre-Screening Long-Term Perceived Stress Levels by Evaluating Physiological Signals

Participants were asked to complete the PSS questionnaires as baseline surveys. PSS is used to measure the monthly stress levels of individuals. We divided the scores of the questionnaires into high and low-stress classes. If the scale is above 10 over 25, it is labeled as high stress and otherwise, it is labeled as low stress. We selected 10 as the threshold because it is the average score of all participants. When the physiological features collected for all sessions are used, we are able to predict the general stress level of a person successfully. The HRV signal achieves approximately 80% accuracy, whereas the EDA signal has a maximum of 73.59% accuracy of finding the general stress level of a person in the last month (see Table XIV). These results are promising because they demonstrated that physiological signals could be used for pre-screening long-term perceived stress levels and our system can advise users to see a psychologist or adopt stress-relieving actions if it detects a high level of long-term stress by examining their physiological signals.

TABLE XIV
PREDICTING THE LONG TERM STRESS LEVEL (LSTL) FROM THE PHYSIOLOGICAL DATA COLLECTED FROM THE EVENT. LSTL IS CALCULATED FROM THE PSS QUESTIONNAIRE REGARDING THE LAST MONTH BEFORE THE EVENT

Algorithm	Accuracy	
	HRV	**EDA**
MLP	**80.09**	70.17
RF	76.85	73.58
kNN	78.24	71.59
LDA	68.98	71.59
SVM	75.93	68.75

IV. CONCLUSION

In this study, we developed an automatic stress detection system which makes use of smart bands. This system is non-obtrusive, comfortable and suitable for daily life usage. To test our system, we collected eight days of data of 16 subjects in the EU project training, where they faced a real-life stressor. The participants are coming from different countries and they have diversified cultures. The diversity is prominent because stress reactions of different cultures could be different and measuring the stress levels of those groups is more difficult than homogeneous culture groups [40]. To the best of our knowledge, this study is the first one that collects a long time physiological data from a multicultural group and detects their stress levels by adding different kinds of contextual information to the physiological signals. 1440 hours of data (12 hours in a day) and 2780 (from 3 session-based and one daily questionnaire each day) self-report questions were collected during this eight-day event from each participant for measuring the perceived stress. We collected data both in the event sessions and after the event in participants' free times for 12 hours a day which shows that our study monitors the daily life stress. EDA and HRV signals are collected to detect physiological stress. The classification performance for both 2 and 3-class session-based, daily and long-term perceived stress levels was presented. Our long-term perceived stress detection system predicts the PSS long term stress levels with approximately 80% accuracy. This result is important because it could be used as a pre-screening tool for psychologists. It could advise people to see a psychologist if high-level long term perceived stress level is detected. The results showed that the selection of a stress scale has an important effect on the performance of the system. Our results show that weather-related, physical activity-related and activity type information improved the system performance. When the weather information in addition to the physiological signals is used, our

model achieved 81% maximum classification accuracy with HRV signal and 72% with EDA signal in 2-class perceived stress level classification. Adding the known context (activity type) information also increased the performance. The 2-class classification accuracy is 79.66% with HRV signal and approximately 76% with EDA signal in 2-class perceived stress level classification. In order to increase the daily physiological stress level detection performance, we further used physical activity based contextual information: stillness and step count, namely. These context data increased our 3-day daily perceived stress level classification performance in terms of validation accuracy to 72% with stillness and 76% with step count information. These results might indicate the correlation between the activity level and the daily stress. As future work, we plan to apply a long-short term memory (LSTM) classifier [41] to the physiological data to increase the performance of our system.

REFERENCES

[1] *European Agency for Safety and Health at Work Campaign Guide Managing Stress and Psychological Risks at Work*, Eur. Agency Safety Health Work, Bilbao, Spain, 2013.

[2] T. W. Colligan and E. M. Higgins, "Workplace stress: Etiology and consequences," *J. Workplace Behav. Health*, vol. 21, no. 2, pp. 89–97, Jul. 2006.

[3] Y. S. Can, B. Arnrich, and C. Ersoy, "Stress detection in daily life scenarios using smart phones and wearable sensors: A survey," *J. Biomed. Informat.*, vol. 92, Apr. 2019, Art. no. 103139.

[4] E. Garcia-Ceja, V. Osmani, and O. Mayora, "Automatic stress detection in working environments from Smartphones' accelerometer data: A first step," *IEEE J. Biomed. Health Informat.*, vol. 20, no. 4, pp. 1053–1060, Jul. 2016.

[5] B. Cinaz, B. Arnrich, R. La Marca, and G. Tröster, "Monitoring of mental workload levels during an everyday life office-work scenario," *Pers. Ubiquitous Comput.*, vol. 17, no. 2, pp. 229–239, Feb. 2013.

[6] M. Gjoreski, H. Gjoreski, M. Lutrek, and M. Gams, "Automatic detection of perceived stress in campus students using smartphones," in *Proc. Int. Conf. Intell. Environ.*, Jul. 2015, pp. 132–135.

[7] A. Bogomolov, B. Lepri, M. Ferron, F. Pianesi, and A. S. Pentland, "Pervasive stress recognition for sustainable living," in *Proc. IEEE Int. Conf. Pervas. Comput. Commun. Workshops (PERCOM WORKSHOPS)*, Mar. 2014, pp. 345–350.

[8] G. Bauer and P. Lukowicz, "Can smartphones detect stress-related changes in the behaviour of individuals?" in *Proc. IEEE Int. Conf. Pervas. Comput. Commun. Workshops*, Mar. 2012, pp. 423–426.

[9] Y. S. Can, N. Chalabianloo, D. Ekiz, and C. Ersoy, "Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study," *Sensors*, vol. 19, no. 8, p. 1849, Apr. 2019.

[10] M. Gjoreski, M. Luštrek, M. Gams, and H. Gjoreski, "Monitoring stress with a wrist device using context," *J. Biomed. Informat.*, vol. 73, pp. 159–170, Sep. 2017.

[11] H. Gimpel, C. Regal, and M. Schmidt, "myStress: Unobtrusive smartphone-based stress detection," in *Proc. Eur. Conf. Inf. Syst.*, 2015, pp. 1–13.

[12] M. Gjoreski, H. Gjoreski, M. Luštrek, and M. Gams, "Continuous stress detection using a wrist device: In laboratory and real life," in *Proc. 2016 ACM Int. Joint Conf. Pervas. Ubiquitous Computing: Adjunct (UbiComp)*. New York, NY, USA: ACM, 2016, pp. 1185–1193

[13] (2018). *Empatica*. Accessed: Dec. 2018. [Online]. Available: https://www.empatica.com/

[14] P. Kostopoulos, A. I. Kyritsis, M. Deriaz, and D. Konstantas, "Stress detection using smart phone data," in *eHealth 360°*. New York, NY, USA: Springer, 2017, pp. 340–351.

[15] V. Mishra *et al.*, "Investigating the role of context in perceived stress detection in the wild," in *Proc. ACM Int. Joint Conf. Int. Symp. Pervas. Ubiquitous Comput. Wearable Comput. (UbiComp)*, 2018, pp. 1708–1716.

[16] S. Taylor, N. Jaques, W. Chen, S. Fedor, A. Sano, and R. Picard, "Automatic identification of artifacts in electrodermal activity data," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2015, pp. 1934–1937.

[17] A. Greco, G. Valenza, A. Lanata, E. Scilingo, and L. Citi, "CvxEDA: A convex optimization approach to electrodermal activity processing," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 4, pp. 797–804, Apr. 2016.

[18] C. Kappeler-Setz, *Multimodal Emotion and Stress Recognition*. ETH Zürich, Zürich, Switzerland, 2012.

[19] M. Vollmer, "Hrvtool—An open-source MATLAB toolbox for analyzing heart rate variability," in *Proc. Comput. Cardiol. (CinC)*, Sep. 2019, pp. 1–4.

[20] M. P. Tarvainen, J. P. Niskanen, J. A. Lipponen, P. O. Ranta-Aho, and P. A. Karjalainen, "Kubios HRV—A software for advanced heart rate variability analysis," in *Proc. 4th Eur. Conf. Int. Fed. Med. Biol. Eng.*, J. Vander Sloten, P. Verdonck, M. Nyssen, and J. Haueisen, Eds. Berlin, Germany: Springer, 2009, pp. 1022–1025.

[21] A. Alberdi, A. Aztiria, and A. Basarab, "Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review," *J. Biomed. Informat.*, vol. 59, pp. 49–75, Feb. 2016.

[22] S. Greene, H. Thapliyal, and A. Caban-Holt, "A survey of affective computing for stress detection: Evaluating technologies in stress detection for better health," *IEEE Consum. Electron. Mag.*, vol. 5, no. 4, pp. 44–56, Oct. 2016.

[23] Q. Zhao and L. Zhang, "ECG feature extraction and classification using wavelet transform and support vector machines," in *Proc. Int. Conf. Neural Netw. Brain*, vol. 2, 2005, pp. 1089–1092.

[24] T. Li and M. Zhou, "ECG classification using wavelet packet entropy and random forests," *Entropy*, vol. 18, no. 8, p. 285, 2016.

[25] R. F. Leonarduzzi, G. Schlotthauer, and M. E. Torres, "Wavelet leader based multifractal analysis of heart rate variability during myocardial ischaemia," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol.*, Aug. 2010, pp. 110–113.

[26] E. A. Maharaj and A. M. Alonso, "Discriminant analysis of multivariate time series: Application to diagnosis based on ECG signals," *Comput. Statist. Data Anal.*, vol. 70, pp. 67–87, Feb. 2014.

[27] F. Eibe, M. Hall, and I. Witten, *The Weka Workbench. Online Appendix for' Data Mining: Practical Machine Learning Tools and Techniques*. San Mateo, CA, USA: Morgan Kaufmann, 2016.

[28] S. Cohen, T. Kamarck, and R. Mermelstein, "A global measure of perceived stress," *J. Health Social Behav.*, vol. 24, no. 4, p. 385, Dec. 1983.

[29] S. G. Hard and L. E. Stavenland, "Development of NASA-TLX (task load index): Results of empirical and theoretical research," *Adv. Psychol.*, vol. 52, pp. 139–183, Jan. 1988.

[30] T. J. Meyer, M. L. Miller, R. L. Metzger, and T. D. Borkovec, "Development and validation of the Penn state worry questionnaire," *Behav. Res. Therapy*, vol. 28, no. 6, pp. 487–495, 1990.

[31] T. Ganor, N. Mor, and J. D. Huppert, "Development and validation of a state-reappraisal inventory (SRI)," *Psychol. Assessment*, vol. 30, no. 12, p. 1663, 2018.

[32] I. Marchetti, N. Mor, C. Chiorri, and E. H. W. Koster, "The brief state rumination inventory (BSRI): Validation and psychometric evaluation," *Cognit. Therapy Res.*, vol. 42, no. 4, pp. 447–460, Aug. 2018.

[33] W. Zhang, R. Ramezani, and A. Naeim, "WOTBoost: Weighted oversampling technique in boosting for imbalanced learning," 2019, *arXiv:1910.07892*. [Online]. Available: http://arxiv.org/abs/1910.07892

[34] S. G. Hart, *NASA Task Load Index (TLX). Volume 1.0; Paper and Pencil Package*, NASA AMES Res. Center, Mountain View, CA, USA, 1986.

[35] E. J. Lewis, K. L. Yoon, and J. Joormann, "Emotion regulation and biological stress responding: Associations with worry, rumination, and reappraisal," *Cognition Emotion*, vol. 32, no. 7, pp. 1487–1498, Oct. 2018.

[36] E. Howarth and M. S. Hoffman, "A multidimensional approach to the relationship between mood and weather," *Brit. J. Psychol.*, vol. 75, no. 1, pp. 15–23, Feb. 1984.

[37] J. L. Sanders and M. S. Brizzolara, "Relationships between weather and mood," *J. Gen. Psychol.*, vol. 107, no. 1, pp. 155–156, Jul. 2010.

[38] *Windguru*. Accessed: Feb. 2020. [Online]. Available: https://www.windguru.cz/

[39] (2018). *Physical Activity Reduces Stress*. Accessed: Dec. 2018. [Online]. Available: https://adaa.org/understanding-anxiety/related-illnesses/other-related-conditions/stress/physical-activity-reduces-st

[40] A. P. Smith, E. J. K. Wadsworth, C. Shaw, S. Stansfeld, K. Bhui, and K. Dhillon, "Ethnicity, work characteristics, stress and health," Health Saf. Executive, Bootle, U.K., Res. Rep. 308, 2005.

[41] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," in *Proc. 9th Int. Conf. Artif. Neural Netw. (ICANN)*, vol. 2, Sep. 1999, pp. 850–855.