Journal of Medical Engineering & Technology

ISSN: 0309-1902 (Print) 1464-522X (Online) Journal homepage: https://www.tandfonline.com/loi/ijmt20

Objective stress monitoring based on wearable sensors in everyday settings

Hee Jeong Han, Sina Labbaf, Jessica L. Borelli, Nikil Dutt & Amir M. Rahmani

To cite this article: Hee Jeong Han, Sina Labbaf, Jessica L. Borelli, Nikil Dutt & Amir M. Rahmani (2020): Objective stress monitoring based on wearable sensors in everyday settings, Journal of Medical Engineering & Technology, DOI: 10.1080/03091902.2020.1759707

To link to this article: https://doi.org/10.1080/03091902.2020.1759707

JOURNAL OF MEDICAL ENGINEERING & TECHNOLOGY
https://doi.org/10.1080/03091902.2020.1759707

Taylor & Francis
Taylor & Francis Group

INNOVATION

Objective stress monitoring based on wearable sensors in everyday settings

Hee Jeong Han[a], Sina Labbaf[a], Jessica L. Borelli[c], Nikil Dutt[a] and Amir M. Rahmani[a,b]

[a]Department of Computer Science, University of California, Irvine, CA, USA; [b]School of Nursing, University of California, Irvine, CA, USA; [c]School of Social Ecology, University of California, Irvine, CA, USA

ABSTRACT

Monitoring people's stress levels has become an essential part of behavioural studies for physical and mental illnesses conducted within the biopsychosocial framework. There have been several stress assessment studies in laboratory-based controlled settings. However, the results of these studies do not always translate effectively to an everyday context. The current state of wearable sensor technology allows us to develop systems measuring the physiological signals reflecting stress 24/7 while capturing the context. In this paper, we present a stress monitoring system that provides objective daily stress measurements in everyday settings based on three physiological signals: electrocardiogram (ECG), photoplethysmogram (PPG), and galvanic skin response (GSR) using Shimmer3 ECG, Shimmer3 GSR+, and Empatica E4 wearable sensors. We perform controlled stress assessment experiments on 17 participants in which we successfully detect stress with a 94.55% accuracy for 10-fold cross-validation and an 85.71% accuracy for subject-wise cross-validation. In everyday settings, the system assesses stress with an 81.82% accuracy. We also examine whether motion artefacts affect stress assessment and filter the low-confidence readings to minimise false alarms.

ARTICLE HISTORY
Received 25 July 2019
Accepted 20 April 2020

KEYWORDS
Stress monitoring; physiological signals; biosignal processing; wearable sensors; internet-of-things

1. Introduction

Stress is defined by Hans Selye as the body's response to one or more stimuli that have disrupted its mental or physical equilibrium [1]. In contrast to the environments in response to which our stress reaction system evolved, people nowadays exhibit more stress due to the increased mental workload from stressful environments such as work [2]. In the 2016 Stress Pulse survey, 60 percent of employees have reported high levels of stress, and 32 percent of employees report constant but manageable stress levels [3]. Prolonged periods of stress are associated with wear and tear on the system [4], resulting in higher rates of disease, including psychological illnesses. For instance, it has been shown that stress correlates with heart disease, asthma, obesity, and diabetes [5], and also, can lead to maladaptive health behaviours such as smoking, irregular sleep, and poor eating habits [6].

Emotion consists of multiple different components – the experiential or subjective component, which can often be thought of in terms of how the emotional experience feels to the person experiencing the change from neutral or euthymic states; the behavioural component, or the action tendency, which is the emotion's expression in the external world, and can take the form of actions such as running away in the case of fear or gritting one's teeth in the case of anger; and finally, the physiological component of emotion includes the internal changes at the level of the autonomic nervous system (sympathetic nervous system, parasympathetic nervous system) or stress hormones (detectable in cortisol, salivary alpha-amylase) that occur in the individual's biological system as a reaction to the emotion. Behavioural scientists argue that these different indices of emotion provide non-overlapping sources of information regarding emotion; consistent with this argument, it is not unusual for these different aspects of emotion to be weakly correlated or even unrelated to one another.

Stress, one type of emotion, can be categorised into two classes: short-term/acute or long-term/chronic [7]. Short-term stress is caused by pressures and demands in the recent past or near future. For example, test anxiety can cause short-term stress. However, long-term stress occurs when there are long-standing pressures and demands. An unsatisfying interpersonal relationship or career can cause long-term stress. Long-term stress can cause detrimental

CONTACT Hee Jeong Han heejh@uci.edu Department of Computer Science, University of California, Irvine, CA 92697, USA

impacts, which necessities effective stress monitoring and management [8].

Accurately assessing stress is challenging as stress can be affected by different factors.One method of measuring the experiential or subjective aspect of the emotion of stress is through the use of self-report assessments using questionnaires [9]. While this method has certain advantages, such as the ability to capture the phenomenology of the individual's experiences, it lacks reliability and feasibility to be used in everyday settings. Taking a questionnaire every day can cause survey fatigue [10], which can make the respondents overwhelmed by the periodic surveys. Also, it cannot accurately reflect the stress level of the participants when they are in certain situations such as during sleep, and often results in a high volume of missing data. In order to enhance reliability, the use of objective stress assessments has been recently proposed. Many of these methods are only feasible within clinical usage. One method consists of using leukocytes to detect hormone changes related to stress [11]. Although this method has proven to be precise, it is not a feasible method for continuous monitoring due to cost and feasibility issues.

As described above, another approach to assessing the emotion of stress is to monitor the activation of the autonomic nervous system (ANS), which provides an index of the physiological aspect of the emotional response. Stress activates the ANS, and this activation can be detected through monitoring the changes in physiological signals [12]. Several physiological signs are correlating with stress such as heart rate, heart rate variability, respiration rate, blood pressure, galvanic skin response. The current state of sensor technology allows us to develop systems measuring physiological signals reflecting stress levels [2]. Monitoring physiological signals using wearable sensors enables the continuous tracking of personal stress status. Stress monitoring systems are currently moving from using traditional physiological sensors such as electroencephalography (EEG) and electrocardiogram (ECG) towards using low-cost and comfortable optical sensors such as photoplethysmogram (PPG) and galvanic skin response (GSR) which are often used in wristbands or smart rings [12]. Most of the existing stress assessment systems collect physiological data in controlled settings [2,13,14]. In these studies, first stress is invoked using various stress tests (e.g. memory game, presentation) and then a predictive model is used to classify the stress level; however, these methods have not been tested in everyday settings.

In this paper, we propose using a stress monitoring system evaluated in both controlled as well as everyday settings. The system collects various physiological signals (ECG, PPG, and GSR) collected by wearable sensors and building machine-learning based stress assessment models based on these signals. These physiological signals are collected using non-invasive sensors in controlled and everyday settings. Our models are tested in everyday settings to examine how motion artefacts affect stress assessment to be able to filter the low-confidence readings to minimise false alarms.

The remainder of this paper is organised as follows. Section 2 provides background information on physiological parameters related to stress. Section 3 presents prior studies on stress monitoring. Section 4 describes our methodology for the experimental protocol, data collection, processing of physiological signals, feature selection, and classification. Section 5 presents the experimental results. Section 6 concludes the paper and states future work.

2. Background

Stress can be measured by monitoring several physiological indicators such as heart activity, blood activity, and skin response. In this section, we describe the concept of physiological parameters and features.

2.1. ECG and PPG

ECG is a measure of the electrical activity of the heart during each cardiac cycle [15]. The ECG uses electrodes to measure electrical signals produced by depolarisation and repolarization of the heart [16]. A typical heart rate consists of a P wave, a QRS complex, and a T wave. The R-R interval is the time interval between adjacent R peaks in the ECG [14]. Heart rate (HR) and heart rate variability (HRV) are calculated from the R-R interval.

PPG is a measure of the electrical activity of the blood during each cardiac cycle [12]. PPG uses an optical pulse to measure a change in blood volume and blood pressure. Since the changes in HR and HRV can be observed from the changes in blood volume pulse measured from the skin, we intend to extract HR and HRV.

In a stressful situation, HR increases. HRV is the fluctuation in the time intervals between adjacent heartbeats [17]. HRV variables change in response to stress. A decrease in HRV variables has been found to be

associated with stress [18]. We focus on time-domain, frequency-domain, and non-linear HRV variables.

2.1.1. Time domain HRV

Time-domain variables of HRV show the amount of variability in measurements of the interbeat interval (IBI), which is the time period between successive heartbeats [19]. In the time-domain analysis, Table 1 shows several time-domain parameters that we focus on.

2.1.2. Frequency domain HRV

Frequency-domain variables estimate the distribution of absolute or relative power into certain frequency bands [19]. Table 2 shows several frequency-domain parameters that we focus on. The LF band is related to short-term blood pressure variation. The HF band is related to breathing rate. Also, the LF and HF components are respectively associated with the sympathetic nervous system (SNS) and parasympathetic nervous system (PNS) activities in the nervous system [20,21]. The analysis involved in assessing frequency-domain HRV analysis lies in the energy ratio of LF to HF content. The most frequently reported stress factor associated with variation in HRV variables was a low parasympathetic activity, which is characterised by a decrease in the HF and an increase in the LF [18].

2.1.3. Nonlinear domain HRV

Non-linear indices quantify the unpredictability of a time series [22]. Non-linear indices correlate with time-domain indices and frequency-domain indices. Table 3 shows several non-linear parameters that we focus on. A Poincaré plot is a graph plotting every R–R interval against the prior interval [19]. Poincaré plot analysis shows patterns within a sequence of values from successive R-R intervals. It does not affect changes in the R-R intervals rapidly [23]. SD1 describes the width of the ellipse and correlates with HF. SD2 describes the length of the ellipse and correlates with LF. SD1/SD2, which is related to LF/HF, is used to measure stress in sympathetic activity.

Table 1. Time domain HRV.

Parameter	Description
SDRR	the standard deviation of R-R intervals
SDSD	the standard deviation of successive differences between adjacent R-R intervals
RMSSD	the root mean square of successive differences between adjacent R-R intervals
pNN20	the percentage of adjacent intervals that differ from each other by more than 20ms
pNN50	the percentage of adjacent intervals that differ from each other by more than 50ms

Table 2. Frequency domain HRV.

Parameter	Description
LF	the low-frequency band (0.04–0.15 Hz)
HF	the high-frequency band (0.15–0.4Hz)
LF/HF	the ratio of LF to HF

2.2. GSR

GSR is a measure of skin conductance during activity changes. Skin conductance based on sweat gland activity that activates in response to high stress is indicative in the skin to conduct electricity to detect increased stress [12]. Since the sweat gland reacts to the SNS, an increase in sweating causes an increase of skin conductance in a stressful situation. Therefore, it can be used as an indicator of stress. We focus on a parameter of GSR, skin conductance.

3. Related works

Assessing stress has been widely studied in psychology. The most popular subjective methods used for this purpose are questionnaires and interviews. Holmes and Rahe [24] established the Social Readjustment Rating Scale, which became the quantitative standard. Since then, several questionnaire or interview based methods have been proposed for measuring stress through self-assessment. For instance, in [25], the Stress Assessment Questionnaire (SAQ) is proposed as an on-line self-reporting assessment tool. The SAQ contains 16 areas, where each area has 8 items for understanding symptoms of stress in different contexts such as relationships, parenting, and work. Each participant self-assesses and reports her/his stress level on a scale from 1 to 5.

Even though questionnaires and interviews are practical and allow researchers to gather subjective information from a large number of participants, these methods suffer from a number of disadvantages. First, it is possible that respondents may not answer truthfully or may ignore some questions. Some respondents answer based on what they may think is socially acceptable or desirable [26]. Furthermore, it is hard to design questionnaires and interviews clearly [27]. As a

Table 3. Non-linear domain HRV.

Parameter	Description
SD1	the Poincaré plot standard deviation perpendicular to the line of identity
SD2	the Poincaré plot standard deviation along the line of identity
SD1/SD2	the ratio of SD1 to SD2

result, respondents may understand questions differently. Even though psychological experts may interpret and analyse personal answers thoroughly, questionnaires and interviews are hard modalities for capturing emotional responses or mental imbalances.

To overcome these challenges, researchers have proposed the use of laboratory-based objective stress assessment methods – for instance, by analysing stress-related physiological reactivity in response to standardised laboratory stressors. The use of these standardised stressors enables researchers to test participants' physiological stress reactivity under controlled parameters. For instance, using psychological stress tests for inducing stress such as the Trier Social Stress Test (TSST), changes in hormones are used to investigate stress responses [28]. In these methods, salivary cortisol was found to correlate with stress, and since then, it has been used as a stress factor in clinical usage [29]. Besides salivary cortisol, leukocytes are used for assessing stress [11]. White blood cells from the blood response can also be used to assess stress hormones. Even though these methods can provide a valid stress assessment, they are not feasible to be used in real-time continuous remote stress monitoring in everyday settings due to their issues in terms of cost, delay and need for physical samples.

To provide a real-time stress assessment, researchers focus on the ANS, which has the advantage of providing a moment-to-moment window into people's physiological arousal. The parasympathetic nervous system (PNS) and the sympathetic nervous system (SNS) are two parts of the ANS [30]. The PNS is responsible for moving the body during rest. On the other hand, the SNS is responsible for "fight or flight" responses to protect the body. Under stress, the SNS forces the body's systems to action [8]. Due to the development of sensor technology, many studies use heart rate, health behaviours, and other vital signals to detect individual stress. Table 4 summarises related work and positions our work with respect to deployed sensors, test setting, test period, test activities, and stress classes.

In [13], an automatic stress detection and an alleviation system is proposed based on five physiological signs: ECG, GSR, respiration rate, blood pressure, and peripheral capillary oxygen saturation (SpO2). The data is collected from 32 participants through a laboratory-based experiment, which takes 94 min. Their machine learning based approach is trained and validated in a controlled setting where the participants are asked to carry out specific stress tests (e.g. fly sound or ice tests). However, this approach suffers from two limitations in terms of its use in everyday settings: 1) it is not feasible to deploy some of the sensors used in this study (e.g. blood pressure, SpO2) in continuous monitoring and 2) the approach does not consider disturbances and challenges existing in daily life (e.g. motion artefacts).

In [14], an activity-aware mental stress detection system is proposed, which also considers the physical activity. The system gathers ECG, GSR, and accelerometer data for 30 min across three activities: sitting, standing, and walking. Its experimental procedure also consists of laboratory-based stress tasks such as the Stroop Colour and Word Test (SCWT) and mental arithmetic. This study detects mental stress affected by physical activities. However, even though it provides a relationship between stress and physical activities, it does not consider the daily context when determining stress, i.e. the system is not deployed in everyday settings.

In [2], a GSR-based pattern recognition system is proposed for stress assessment. Even though this work utilises non-laboratory data to find stress levels, it only uses a GSR sensor which is not as sensitive compared to other mechanisms. The data is collected from five

Table 4. Related studies and its setting of the experiment.

Related Works	Deployed Sensor	Test Setting	Period	Test Activities	Stress Classes
Ours	PPG, ECG, GSR	Lab-based	50mins	Memory Game, Mosquito Sound, IAPS Plank, Ice Test, TSST, SCWT	2 levels (Binary)
		non-Lab-based	2days	–	2 levels
[31]	EMG, ECH, GSR, RSP	Lab-based	20mins	Music related to emotion	4 levels
[8]	ECG	Lab-based	9mins	Presentation	2 levels
[32]	EEG, HR	Lab-based	not provided	Mensa Test	2 levels
[14]	ECG, GSR, accelerometer	Lab-based	30mins	SCWT	2 levels
[33]	EEG, GSR, PPG	Lab-based	6mins	SCWT	3 levels
[34]	HR, GSR, Body Temperature	Lab-based	30mins	Tower of Hanoi 6 discs	2 levels
[13]	ECG, GSR, respiration rate, blood pressure, blood oximeter	Lab-based	94mins	Memory Game, Fly Sound, IAPS, Ice Test	2 levels
[35]	bioradar	Lab-based	2mins	Mathematical problem	2 levels
[36]	EEG	Lab-based	10mins	Music	2 levels
[37]	EEG, ECG	Lab-based	40mins	Threatening message	4 levels
[38]	GSR, PPG	Lab-based	not-provided	Presentation	3 levels
[2]	GSR	non-Lab-based	8hours x 4weeks	–	2 levels

persons during working hours for four weeks. However, the paper concludes that GSR data is not sufficient for determining levels of stress with high accuracy. It also states that contextual data is needed when detecting stress in daily activities. Even though the users are suggested to record their feelings, the paper does not use this information. In our approach, we employ a combination of sensors, GSR, PPG, and ECG to improve accuracy for stress identification. Furthermore, we collect daily context data from users and use them when determining stress.

Another significant difference is that these previous studies have been confined to laboratory environments, making it impractical to build a stress monitoring system for daily life usage. In contrast, our study collects various physiological data (ECG, PPG, and GSR) in both laboratory-based tests and everyday data collection and thus, defines stress levels in everyday life.

4. Methodology

Our stress monitoring system provides an assessment of stress levels using three main physiological signs: ECG, PPG, GSR. Our research has been conducted in two different settings: controlled setting and everyday setting. To find a correlation between stress and physiological signals, we perform offline laboratory-based stress tests to collect bio-signals from wearable devices. We then process the raw signals to extract features, build predictive models using these features, and find the relationship between each feature and stress. We assume that stress is labelled in binary: whether each participant is stressed or not. Figure 1 shows the process overview in a controlled setting. Figure 1(a) shows the training process to build a predictive stress model. Figure 1(b) shows the inference process to find the relationship between each feature and stress by the stress model.

We also collect physiological signals in an everyday setting through wearable devices to find daily stress levels. With everyday data, we perform feature extraction and prediction using the models trained in the controlled setting to get personal stress levels. Figure 2 shows the process overview in the everyday setting.

4.1. Wearable sensors

We use *Shimmer3 ECG*, *Shimmer3 GSR+*, and *Empatica E4* wearable sensors. We collected ECG signals from *Shimmer3 ECG* and PPG and GSR signals from *Shimmer3 GSR+*. *Empatica E4* Wristband is also used for gathering PPG and GSR signals. Table 5 summarises wearable sensors, wearing type of each sensor, signals collected from each sensor, and sample rate of each signal.

4.2. Data collection in the controlled setting

In the controlled setting, we conduct laboratory-based stress tests. Laboratory-based stress tests consist of several tasks for inducing short-term stress. A total of 17 participants (13 male, 4 female), ranging between

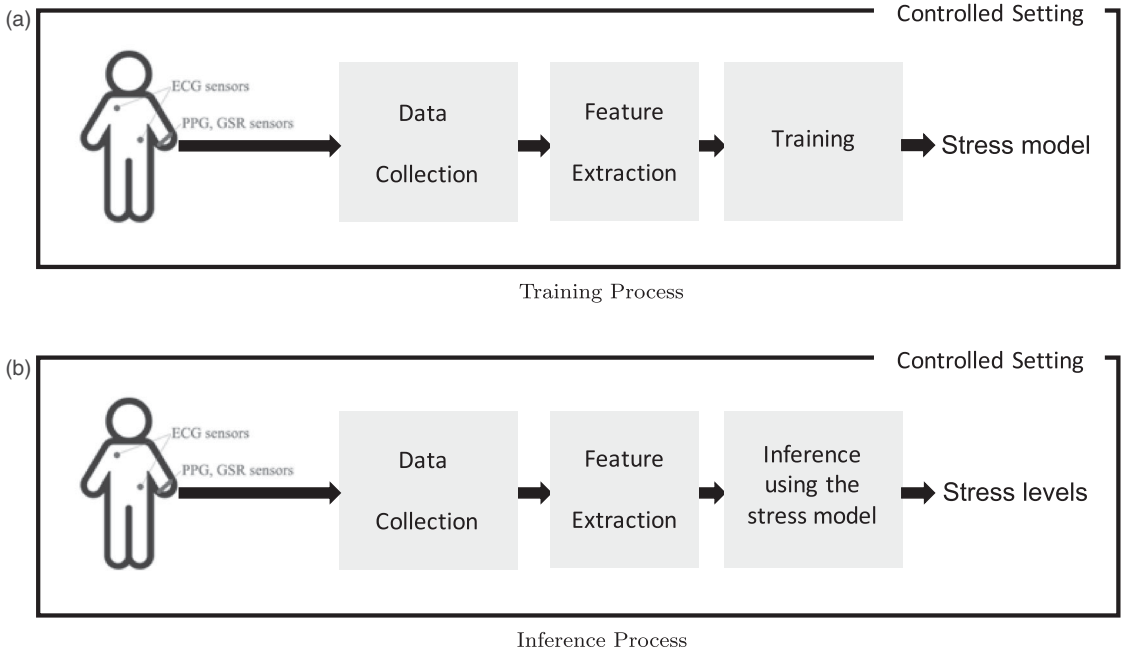

Figure 1. Process overview in the controlled setting. (a) Training Process; (b) Inference Process

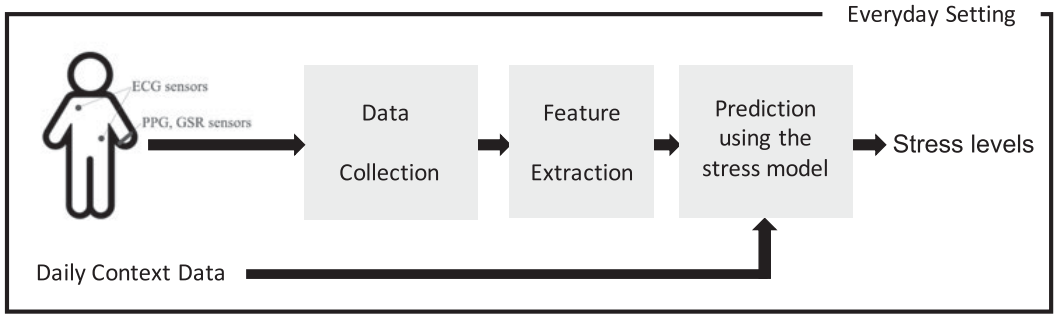

Figure 2. Process overview in the everyday setting.

Table 5. The list of wearable devices and its specifications.

Device	Wearing Type	Sensors	Sampling Rate
Shimmer3 ECG	Chest strap	ECG	512 Hz
Shimmer3 GSR+	Wristband	PPG	128 Hz
		GSR	128 Hz
Empatica E4	Wristband	PPG	64 Hz
		GSR	4 Hz

20 and 27 years of age, participated in the laboratory-based controlled experiments. During stress tests, the expected result is "1" (i.e. stress) for stress tasks, and "0" (i.e. no stress) otherwise. We implement two types of laboratory-based tests. Figure 3 shows our two experimental procedures.

The first laboratory-based test takes approximately 50 min for each participant. It consists of five stress tasks: Memory game, Mosquito sound, Images Test, Plank, and Ice Test [13]. Each stress task lasts for 2 min, followed by a 6 min rest period between each stress task. *Baseline* is the first stage of the test to collect essential personal physiological signals. We ask participants to meditate while listening to relaxing classical music. *Rest* is a rest period for the participants after each stress task where they also listen to relaxing classical music. This period is needed to reduce stress, which is incurred from the previous stress task. *Memory Game* is a card game in which all of the cards are laid face down, and two cards are flipped face up over each turn. The goal of the game is to turn pairs of matching cards [39]. *Mosquito sound* is a period where the participants listen to a mosquito sound with a black screen [40]. *Images Test* is a period for the participants to look at selected pictures from the International Affective Picture System (IAPS) [41]. IAPS provides affect ratings of pictures. *Plank* is a period for the participants to perform a plank for two minutes while putting their palms up to prevent sensor distortion. *Ice Test* is a period for the participants to put their right hand inside an ice cup [13].

The second laboratory-based test also takes approximately 50 min for each participant. It consists of three stress tasks: The Trier social stress test (TSST), Images Test, and the Stroop Colour and Word Test (SCWT). *TSST* is a laboratory procedure used to reliably induce stress in human research participants [28]. TSST consists of two stress tasks: speech and maths. In the speech portion, there is a preparation step and a presentation step for a given topic, which are each 5 min period. After the presentation, we ask participants to count backward from 1022 subtracting 13 for 5 min. *SCWT*, which is based on the Stroop Effect, provides coloured word lists [42]. Participants read those word lists while following the instructions.

4.3. Data collection in the everyday setting

In an everyday setting, we collect physiological signals in daily life. We ask participants to wear Empatica E4 Wristband, and a Shimmer ECG chest strap. We also collect daily context data labelled with self-reported stress levels. Participants report whether they feel stressed or not every 30 min. We have 3 subjects (1 female) in the non-laboratory-based experiment.

4.4. Preprocessing and feature extraction

Before extracting features, the data needs to be preprocessed. Two steps of preprocessing are performed on the signals to remove noise: filtering and smoothing. Figure 4 shows the procedure of preprocessing and feature extraction.

In order to extract HR and HRV, ECG and PPG signals need to be preprocessed using proper digital signal processing techniques [43]. For the ECG data, we use band-pass filters to remove noise and use the moving average filter to smooth the data. From the filtered ECG data, we extract the mean value of HR. We also extract the mean value of time-domain and non-linear HRV variables. However, for frequency-domain HRV variables, we use Fast Fourier Transform and Power Spectral Density analysis to study how power is distributed as a function of frequency, which allows an autonomic balance to be quantified at any given

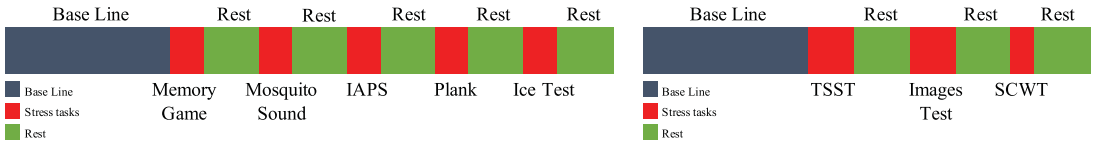

Figure 3. Test procedure collecting physiological signals related to stress.

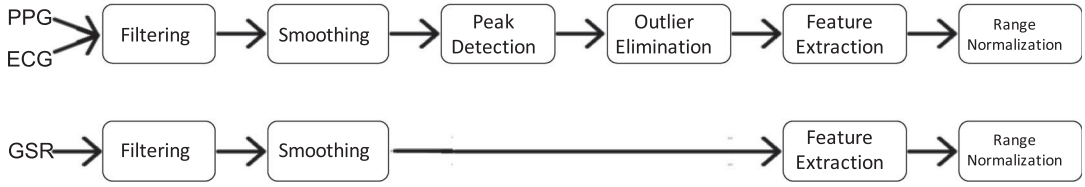

Figure 4. The procedure of preprocessing and feature extraction.

time. After preprocessing, we extract the mean of frequency-domain HRV variables.

For the PPG data, we use band-pass and moving average filters to remove noise and smooth the data. From the filtered PPG data, we then extract the mean values of HR. We also extract the mean of time-domain, frequency-domain, and non-linear HRV variables using the same feature extraction functions used for the ECG.

In order to extract skin conductance, the GSR signal needs to be processed with median and moving average filters. A median filter is used for removing noise, and a moving average filter is used for smooth the data. After preprocessing, we extract the mean values of skin conductance. We also extract the gradient of skin conductance to calculate the variation.

4.5. Feature selection

Using all features is not necessarily helpful as they may not help in increasing accuracy. If a feature is not related to stress, having it among related features may increase noise [44]. Computing some loosely correlated features may also not be useful because of the computational complexity. For instance, frequency domain and independent features have non-linear computational complexity. Especially in local implementations in Internet-of-Things [45,46] based systems, these overheads are considerable. Thus, we decide to select features that are more correlated to stress.

In order to find the best subset of features, we adopt a greedy stepwise method [47]. This method starts from an empty set. It adds features that increase accuracy and removes features that decrease it. We continue doing these two steps until we reach a set of features in which adding no new feature or removing any selected feature can increase accuracy. To evaluate the accuracy of each subset, we use a 5-nearest-neighbor classifier, correlation-based feature selection method, and information gain. Based on this method, the features are shown in Table 6 in bold are selected.

Table 6. Extracted features from sensors, selected features in bold.

Sensor	Features
ECG	HR, SDRR, SDSD, RMSSD, **pNN20**, pNN50, LF, HF, LF/HF, SD1, SD2, **SD1/SD2**
PPG	HR, **SDRR**, SDSD, RMSSD, pNN20, pNN50, **LF**, HF, **LF/HF**, **SD1**, SD2, SD1/SD2
GSR	**Skin conductance**

4.6. Machine learning based classification

The bias of physiological data can vary by using personal data sets or general data sets [48]. Personal data sets contain data collected from the same person (within), and general data sets contain data from other subjects (between). In order to test the efficiency of our classifier, we test it in both cases.

We use several machine learning based classification algorithms such as K-nearest-neighbor (kNN) with $k \in \{1, 3, 5, 7, 9\}$, support vector machine (SVM), and Naive Bayes classifier. kNN is a method that uses k nearest data-points and does a majority vote to predict the result [49]. SVM finds hyper-planes to divide data-points into different classes [50]. We used the Weka implementation of LIBSVM [51]. Naive Bayes classifiers predict the result based on the probabilities of each feature's probabilistic knowledge [52]. Naive Bayes classifiers act differently based on the distribution of data-points [53].

5. Experimental results

In this section, we present our experimental results in the controlled and everyday settings. First, we validate

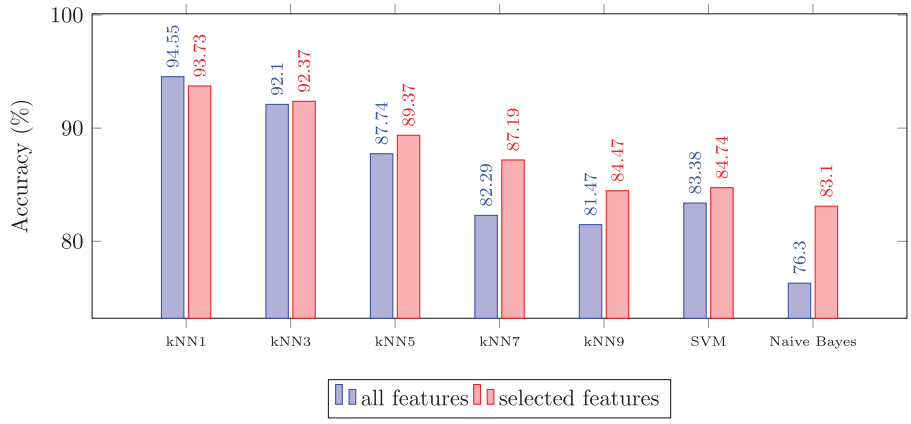

Figure 5. 10-fold leave data-points out cross-validation accuracy of the different classifiers using the different number of features.

our developed stress models using three different classification algorithms (i.e. kNN, SVM, and Naive Bayes). We test whether the classifiers generalise across data-points as well as across subjects. We then apply the classifier on everyday data to predict stress, observe and study the contextual factors affecting the results, and analyse techniques to mitigate them. We use everyday self-report stress label as ground truth (i.e. reference point). We also collect context data (e.g. running, walking, eating, etc.) to evaluate the effect of noise such as motion artefacts on the decisions in everyday settings. To examine how a combination of features affects stress detection accuracy, we create four groups of bio-signals: GSR + PPG + ECG, GSR + PPG, GSR + ECG, and only PPG. The rationale to study the PPG only case is the fact that this is the most dominant, cost-effective, and convenient method used in wearables such as smart bands, watches, and rings, making it the most feasible monitoring method for everyday settings.

We use the Weka software package [54] for classification and prediction. We collect stress data from 17 participants in our controlled setting. Our participants are college students between the age of 20 to 27. The 25 features mentioned in Section 4.5 are extracted from the multi-modal signals for each subject during the tests. Out of these features, 12 are extracted from ECG, 12 from PPG, and 1 feature from GSR. We collect features from each signal for a window size of a minute resulting in 367 min of data for the controlled experiment as training data. Out of these, 234 min are during stressful tasks, and 133 are during the baseline (i.e. labelled as no-stress).

In the everyday setting, we collect stress data from one participant excluded in the controlled setting. We extract the same 25 features for every minute. 340 min of data are provided with self-reported stress labels. Labels are associated with the stress reported for every 30 min.

5.1. Stress assessment in a controlled setting

To objectively assess the stress in a controlled setting, we build a stress model using different classifiers (kNN, SVM, and Naive Bayes). We conducted two different sets of experiments: i) with all features, and ii) with selected features (presented in Section 4.5). In addition, we analyse the data from two different perspectives: data-points vs. subjects. In the data-points view, we treat the data points similarly regardless of the participant they were collected from whereas in the subjectwise analysis, we group each individual's data.

5.1.1. Leave data-points out cross-validation accuracy

We evaluate the accuracy when the classifiers generalise across data-points with 10-fold cross-validation [55]. Figure 5 shows the accuracy of three different classifiers, kNN, SVM, and Naive Bayes. The best accuracy when all features are used belongs to kNN1, which is equal to 94.55%. Similarly, kNN1 performs best when selected features are used with the accuracy of 93.73%. As the test data is chosen randomly from all participants, we expect to find data-points from the same subject in both testing and training sets. This makes kNN1 a better classifier as it eliminates the effect of other subjects in the result more than the others.

Table 7 shows a comparison between our work and the related work in terms of the deployed sensors and the obtained accuracy. As can be seen from the table, our obtained accuracy (94.55%) is the highest

compared to the related work. Note that all these works also report their accuracy in control settings.

5.1.2. Leave subjects out cross-validation accuracy

Since the population is rather small (only 17 subjects) it can result in a high bias on individual subjects. To isolate the effect of such bias in the accuracy of the classifiers, we also evaluate the accuracy when the classifiers generalise across subjects with 10-fold cross-validation. Figure 6 shows leave subjects out cross-validation accuracy among the kNN, SVM, and Naive Bayesian classifiers. As can be seen from the figure, the best accuracy for the all features case belongs to the SVM, which is equal to 79.84%. Similarly, the best accuracy for the selected features also belongs to SVM, which is 84.71%. Classifiers correspond too closely to a particular set of data in a high bias. Overfitted classifiers perform worse on validation [56]. Since SVM can avoid overfitting appropriately, it shows the best accuracy rather than other classifiers.

Table 8 shows a comparison between our work and the related work in terms of the deployed sensors and the obtained accuracy. As can be seen from the table, our obtained accuracy (84.71%) is the highest compared to the related work. Note that to the best of our knowledge, there is only one work in the literature that has used subjectwise cross-validation in this context.

Table 7. Comparison of leave data-points out cross-validation accuracy between related studies and ours in general model.

Related Works	Deployed Sensors	Accuracy
Ours	**PPG, ECG, GSR**	**94.6**
[31]	EMG, ECH, GSR, RSP	92.0
[14]	ECG, GSR, accelerometer	92.4
[33]	EEG, GSR, PPG	81.8
[34]	HR, GSR, Body Temperature	84.5
[13]	ECG, GSR, respiration rate, blood pressure, blood oximeter	89.3
[35]	bioradar	94.4
[36]	EEG	80.1
[37]	EEG, ECG	79.6

5.2. Stress assessment in the everyday setting

We predict the stress level in the everyday setting through the stress model. We split everyday data into minutes, extract the features, and run them through the stress model. To get an accuracy of everyday stress prediction, we use a binary self-described stress level as ground truth. Participants report their self-assessment of stress level every 30 min. Since we have the stress model from the controlled setting, we use a majority vote to prevent an unstable prediction for data-points due to its inherent noise cancellation property [57]. We use two-third majority to consider a prediction reliable.

5.2.1. Cross-validation accuracy without activity recognition

We evaluate 340 min of daily data from participants across 2 days, which have self-assessed stress labels. Each data point also includes various kinds of activities such as sitting, walking, and eating. Figure 7 shows the cross-validation accuracy of everyday data among three different classifiers: kNN, SVM, and Naive Bayes. The best accuracy from all features belongs to kNN1, kNN7, and kNN9, which is 63.64%. The best accuracy from selected features belongs to kNN5, which is 81.82%. We observe that the result from selected features shows higher accuracy than the result

Table 8. Comparison of leave subject out cross validation accuracy between related studies and ours in general model.

Related Works	Deployed sensors	Accuracy
Ours	**PPG, ECG, GSR**	**84.7**
[14]	ECG, GSR, accelerometer	80.9

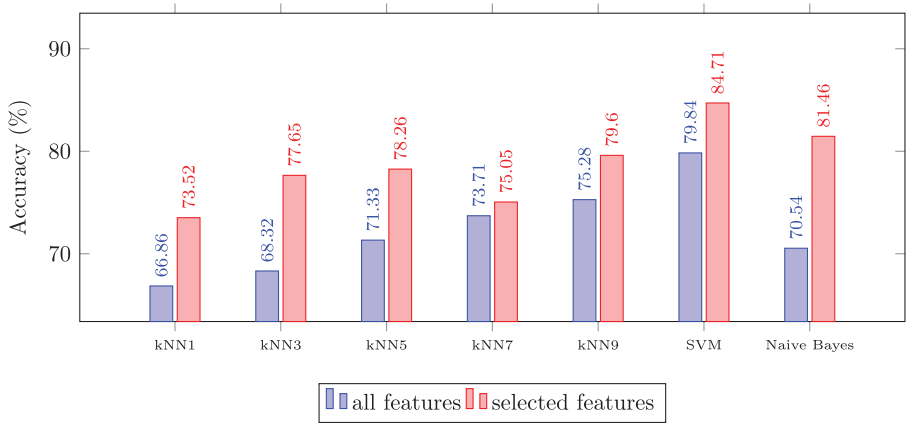

Figure 6. 10-fold leave subjects out cross-validation accuracy of the different classifiers using the different number of features.

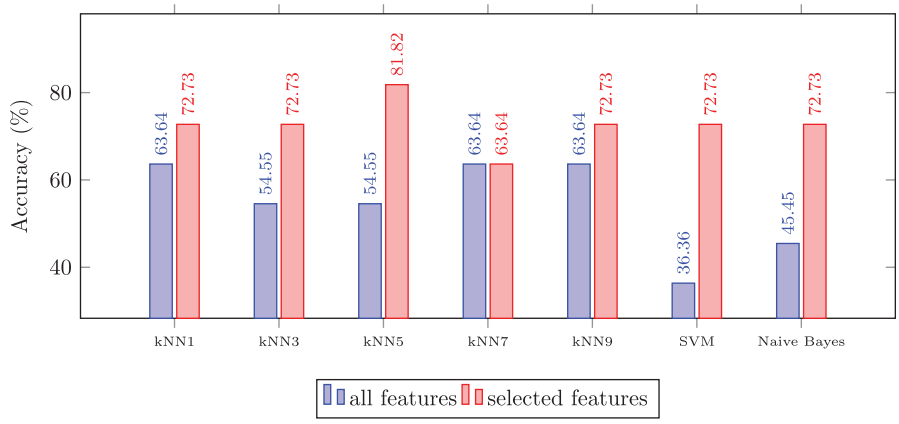

Figure 7. Everyday stress assessment accuracy of the different classifiers using the different number of features.

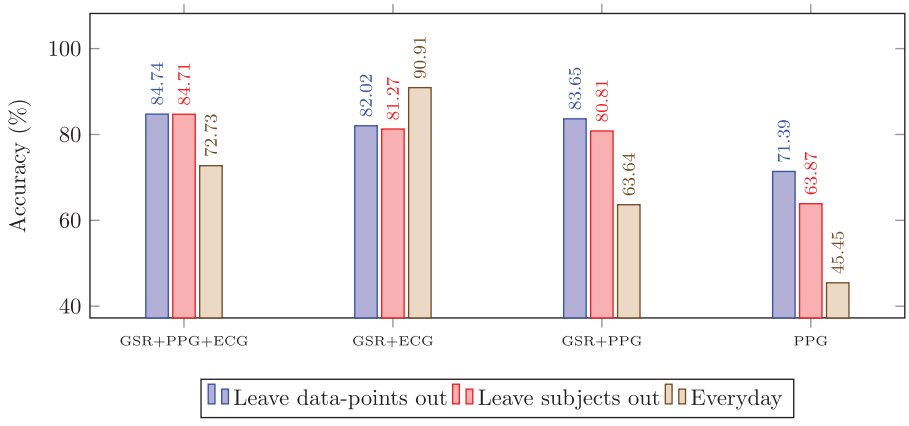

Figure 8. Comparison of feature combination with feature selection.

from all features. This is because loosely correlated features are removed.

We make subsets of features to examine how a combination of physiological signals affect stress assessment accuracy: GSR + PPG + ECG, GSR + PPG, GSR + ECG, and only PPG. The first group, GSR + PPG + ECG, is all signals collected in this study. The second group, GSR + PPG, is chosen because sensors of GSR and PPG are comfortable to wear on the body. The third group, GSR + ECG, is chosen in order to compare with the second group. The last group consists of only PPG, which is the most available physiological sensor. Figure 8 shows a comparison of each group's stress accuracy. We present the results in the controlled settings (leave data-points out cross-validation and leave subjects out cross-validation) and the everyday setting with selected features mentioned in Section 4.5. All groups show a similar result in the controlled settings. However, in the everyday setting, GSR + ECG group shows the best accuracy, which is 90.91%. It is the highest accuracy in the everyday settings. Compared to GSR + ECG, GSR + PPG group shows a worse result, and a result is getting worse when we consider only PPG to assess stress.

ECG signal is required to detect heart activities reliably, but the acquisition of ECG is not easy in daily life. PPG signal, which is more convenient when collecting data, is an alternative method to track heart activities [58]. Although PPG is increasingly being used in a personal healthcare setting, PPG is sensitive to motion artefacts [59]. In this paper, since participants are sitting during stress tests, the accuracy among combinations of signals shows similarity in the controlled settings. However, participants have no choice but to move in everyday settings, which increases the PPG signal noise.

5.2.2. Cross-validation accuracy with activity recognition

Since the PPG signal noise typically increases in everyday settings, we examine how much motion artefacts affect the PPG signal. We extract activities based on personal daily context report. We retain data-points during low intensive activities such as sitting but

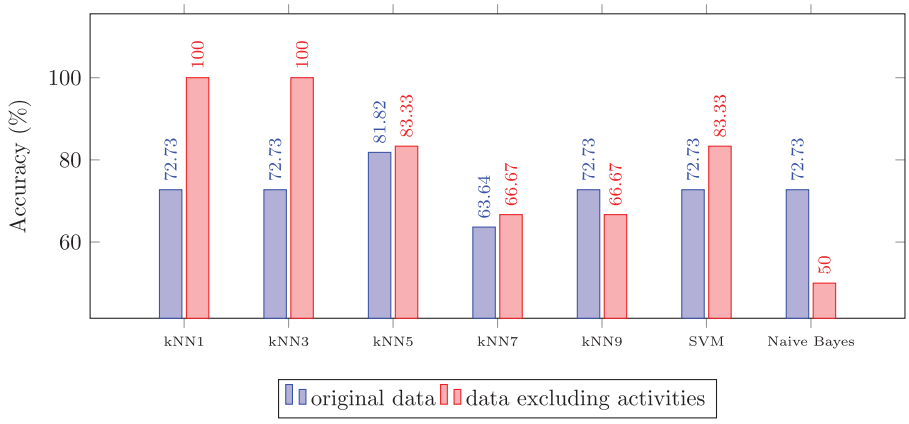

Figure 9. Comparison of everyday stress assessment accuracy between original data and data excluding activities with feature selection.

exclude data-points during high intensive activity such as walking. Figure 9 shows a comparison of everyday stress assessment between cross-validation accuracy without activity recognition and cross-validation accuracy with activity recognition. The best accuracy is 85.71% without activity recognition, on the other hand, the best accuracy is 100.00% with activity recognition. When we exclude rapid motion artefacts, the result presents better accuracy because we exclude unreliable data. In fact, high intensive activities can lead to inaccurate signal processing. Therefore, we need to exclude those activities to decrease false-positive rates.

6. Discussion

Our objective stress monitoring system performs stress assessment in both the controlled and the everyday settings. The overall system is shown to have 94.55% accuracy in the controlled setting and 85.71% accuracy in the everyday setting. We assume that stress is labelled in binary: whether each participant is stressed or not. This is indeed a limitation in our work as well as in the majority of the related works (Table 4, since, due to the nature of the stress tests, it only allows us to distinguish the stress level in binary. In the future, we intend to conduct new experiments to assess stress with finer granularity. The new stress tests will include fine-grained labels such as no-stress, low-stress, medium-stress, and high-stress.

Examining everyday data is more challenging than controlled data. Physiological data is sensitive to motion artefact, environmental noise, etc. Unlike the controlled setting where participants are monitored while sitting, in everyday settings, many activities such as walking, running, and eating involving movement can affect the signal quality as well as the body responses. These activities can cause noise in the physiological data and affect the accuracy of the features extracted from them. In Section 5, we showed that detecting reliable PPG data is vital for objective stress monitoring. We, therefore, used the automatic PPG confidence assessment technique proposed in [60] only to apply the calcification on high-confidence signals. This technique uses a convolutional neural network based model trained with reliable and unreliable samples of PPG. It reports a confidence rate every minute. In everyday data, the PPG confidence assessment method reported 244 min out of 340 min of our data as reliable, meaning that the PPG signals for the other 96 min were unreliable because of motion artefacts. This results in preventing false positives in our stress assessment. It should be noted that PPG is more prone to motion artefacts and noise compared to ECG, which makes the signal correction almost impossible in several cases.

7. Conclusions

We proposed a stress monitoring system that was tested for everyday stress assessment. We designed, implemented, and analysed the system offering not only high accuracy stress detection in the controlled setting but also reasonable predictions in the everyday setting. We performed controlled stress assessment experiments on 17 participants and everyday setting monitoring on 1 volunteer. Our results demonstrate 94.55% accuracy in the generalised model for stress detection while showing 85.71% accuracy when the classifier generalises across subjects. The accuracy of the system in the everyday setting is 81.82%. Our system is compared against related studies in terms of

the sensors used, accuracy in the generalised model, test sets, test period, and test activities.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was partially supported by the following grants: NSF Smart and Connected Communities grant CNS-1831918, NSF WiFiUS grant CNS-1702950, the Academy of Finland grants 313448 and 313449 (PREVENT project), and the Academy of Finland grants 316810 and 316811 (SLIM project).

ORCID

Jessica L. Borelli http://orcid.org/0000-0001-8471-6732

References

[1] Selye H. The stress of life; 1956.
[2] Bakker J, Pechenizkiy M, Sidorova N. What's your current stress level? detection of stress patterns from gsr sensor data. Proceedings of the Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on; IEEE; 2011. p. 573–580.
[3] 2016 compsych stresspulse survey; 2016.
[4] Sterling, P., Eyer, J. Allostasis: a new paradigm to explain arousal pathology. In: S. Fisher , J. Reason, editors. Handbook of life stress, cognition and health. New York (NY): John Wiley & Sons;1988. p. 629–649.
[5] Selye H. Stress in health and disease. Boston (MA): Butterworth-Heinemann; 2013.
[6] Glanz K, Schwartz MD. Stress, coping, and health behavior. Health behavior and health education: theory, research, and practice. Francisco (CA): Jossey-Bass. Vol. 4; 2008; p. 211–236.
[7] Miller LH, Smith AD, Rothstein L. The stress solution: an action plan to manage the stress in your life. New York (NY): Pocket; 1994.
[8] Schubert C, Lambertz M, Nelesen R, et al. Effects of stress on heart rate complexity–a comparison between short-term and chronic stress. Biol Psychol. 2009;80(3):325–332.
[9] Andreou E, Alexopoulos EC, Lionis C, et al. Perceived stress scale: reliability and validity study in greece. Int J Environ Res Public Health. 2011;8(8):3287–3298.
[10] Egleston BL, Miller SM, Meropol NJ. The impact of misclassification due to survey response fatigue on estimation and identifiability of treatment effects. Statist Med. 2011;30(30):3560–3572.
[11] Davis A, Maney D, Maerz J. The use of leukocyte profiles to measure stress in vertebrates: a review for ecologists. Funct Ecol. 2008;22(5):760–772.
[12] Greene S, Thapliyal H, Caban-Holt A. A survey of affective computing for stress detection: Evaluating technologies in stress detection for better health. IEEE Consumer Electron Mag. 2016;5(4):44–56.
[13] Akmandor AO, Jha NK. Keep the stress away with soda: stress detection and alleviation system. IEEE Trans Multi-Scale Comp Syst. 2017;3(4):269–282.
[14] Sun FT, Kuo C, Cheng HT, et al. Activity-aware mental stress detection using physiological sensors. Proceedings of the International Conference on Mobile Computing, Applications, and Services; Springer; 2010. p. 282–301.
[15] Dupre A, Vincent S, Iaizzo PA. Basic ECG theory, recordings, and interpretation. In: Handbook of cardiac anatomy, physiology, and devices. Totowa (NJ): Humana Press; 2005. p. 191–201.
[16] Klabunde R. Cardiovascular physiology concepts. Philadelphia (PA): Lippincott Williams & Wilkins; 2011.
[17] McCraty R, Shaffer F. Heart rate variability: new perspectives on physiological mechanisms, assessment of self-regulatory capacity, and health risk. Glob Adv Health Med. 2015;4(1):46–61.
[18] Kim HG, Cheon EJ, Bai DS, et al. Stress and heart rate variability: a meta-analysis and review of the literature. Psychiatry Investig. 2018;15(3):235–245.
[19] Shaffer F, Ginsberg J. An overview of heart rate variability metrics and norms. Front Public Health. 2017;5:258.
[20] Healey J, Picard RW. Detecting stress during real-world driving tasks using physiological sensors. IEEE Trans Intell Transport Syst. 2005;6(2):156–166.
[21] Acharya UR, Joseph KP, Kannathal N, et al. Heart rate variability. In: Advances in cardiac signal processing. Berlin, Heidelberg: Springer; 2007. p. 121–165.
[22] Stein PK, Reddy A. Non-linear heart rate variability and risk stratification in cardiovascular disease. Indian Pacing Electrophysiol J. 2005;5(3):210–220.
[23] Behbahani S, Dabanloo NJ, Nasrabadi AM. Ictal heart rate variability assessment with focus on secondary generalized and complex partial epileptic seizures. Adv Biores. 2013;4(1):50–58.
[24] Holmes TH, Rahe RH. The social readjustment rating scale. J Psychosom Res. 1967;11(2):213–218.
[25] Stress assessment questionnaire; 2012. https://ptc.bps.org.uk/test-review/stress-assessment-questionnaire.
[26] Patten ML. Questionnaire research: a practical guide. Abingdon-on-Thames (UK): Routledge; 2016.
[27] Phellas CN, Bloch A, Seale C. Structured methods: interviews, questionnaires and observation. Res Soc Cult. 2011;3:181–205.
[28] Kirschbaum C, Pirke KM, Hellhammer DH. The "trier social stress test"–a tool for investigating psychobiological stress responses in a laboratory setting. Neuropsychobiology. 1993;28(1–2):76–81.
[29] Hellhammer DH, Wüst S, Kudielka BM. Salivary cortisol as a biomarker in stress research. Psychoneuroendocrinology. 2009;34(2):163–171.
[30] Langley JN. The autonomic nervous system (pt. i). 1921. Brain. 1903;26(1):1–26.
[31] Wagner J, Kim J, André E. From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification.

Proceedings of the 2005 IEEE international conference on multimedia and expo; IEEE; 2005. p. 940–943.
[32] Taelman J, Vandeput S, Spaepen A, et al. Influence of mental stress on heart rate and heart rate variability. Proceedings of the 4th European conference of the international federation for medical and biological engineering; Springer; 2009. p. 1366–1369.
[33] Das D, Bhattacharjee T, Datta S, et al. Classification and quantitative estimation of cognitive stress from in-game keystroke analysis using EEG and GSR. Proceedings of the Life Sciences Conference (LSC); 2017. IEEE; IEEE; 2017. p. 286–291.
[34] Ciabattoni L, Ferracuti F, Longhi S, et al. Real-time mental stress detection based on smartwatch. Proceedings of the 2017 IEEE International Conference on Consumer Electronics (ICCE); IEEE; 2017. p. 110–111.
[35] Fernández JRM, Anishchenko L. Mental stress detection using bioradar respiratory signals. Biomed Signal Process Control. 2018;43:244–249.
[36] Liao CY, Chen RC, Tai SK. Emotion stress detection using eeg signal and deep learning technologies. Proceedings of the 2018 IEEE International Conference on Applied System Invention (ICASI); IEEE; 2018. p. 90–93.
[37] Xia L, Malik AS, Subhani AR. A physiological signal-based method for early mental-stress detection. Biomed Signal Process Control. 2018;46:18–32.
[38] Koussaifi M, Habib C, Makhoul A. Real-time stress evaluation using wireless body sensor networks. Proceedings of the 2018 Wireless Days (WD); IEEE; 2018. p. 37–39.
[39] Memory game; 2007. https://www.mathworks.com/matlabcentral/fileexchange/14059-memory-a-k-a-concentration.
[40] Mosquito fly sound; 2014. https://www.youtube.com/watch?v=PYnVlOoxZWw.
[41] Lang PJ. International affective picture system (IAPS): affective ratings of pictures and instruction manual. Technical report. Gainesville (FL): University of Florida; 2005.
[42] Stroop JR. Studies of interference in serial verbal reactions. J Exp Psychol. 1935;18(6):643–662.
[43] Gupta A, Bhandari S. ECG noise reduction by different filters - a comparative analysis. Int J Res Comp Commun Technol. 2015;4(7):424–431.
[44] Deng Y, Wu Z, Chu CH, et al. Evaluating feature selection for stress identification. Proceedings of the 2012 IEEE 13th International Conference on Information Reuse & Integration (IRI); IEEE; 2012. p. 584–591.
[45] Firouzi F, Rahmani AM, Mankodiya K, et al. Internet-of-things and big data for smarter healthcare: from device to architecture, applications and analytics; 2018.
[46] Mieronkoski R, Azimi I, Rahmani AM, et al. The internet of things for basic nursing care–a scoping review. Int J Nurs Stud. 2017;69:78–90.
[47] Hall MA. Correlation-based feature selection for machine learning Hamilton (NZ): University of Waikato; 1999.
[48] Das D, Datta S, Bhattacharjee T, et al. Eliminating individual bias to improve stress detection from multimodal physiological data. Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); IEEE; 2018. p. 5753–5758.
[49] Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. Am Statistician. 1992;46(3):175–185.
[50] Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–297.
[51] Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol. 2011;2(3):1–27:27. http://www.csie.ntu.edu.tw/cjlin/libsvm.
[52] Hand DJ, Yu K. Idiot's bayes–not so stupid after all? Int Statistical Rev. 2001;69(3):385–398.
[53] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training. Proceedings of the eleventh annual conference on Computational learning theory; ACM; 1998. p. 92–100.
[54] Hall M, Frank E, Holmes G, et al. The weka data mining software: an update. SIGKDD Explor Newsl. 2009;11(1):10–18.
[55] Kohavi R, et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the International Joint Conference on Articial Intelligenc, Montreal, Canada. Vol. 14; 1995. p. 1137–1145.
[56] Hawkins DM. The problem of overfitting. J Chem Inf Comput Sci. 2004;44(1):1–12.
[57] James G. Majority vote classifiers: theory and applications [dissertation]. Stanford University; 1998.
[58] Pinheiro N, Couceiro R, Henriques J, et al. Can ppg be used for HRV analysis? Proceedings of the 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); IEEE; 2016. p. 2945–2949.
[59] Pietilä J, Mehrang S, Tolonen J, et al. Evaluation of the accuracy and reliability for photoplethysmography based heart rate and beat-to-beat detection during daily activities. Proceedings of the 6th Embec & NBC 2017. Springer; 2017. p. 145–148.
[60] Naeinia EK, Azimib I, Rahmania AM, et al. A real-time ppg quality assessment approach for healthcare internet-of-things. Procedia Comput Sci. 2019;151(C).