UNIVERSITY OF CALIFORNIA,
IRVINE

Objective Stress Monitoring based on Wearable Sensors in Everyday Settings

THESIS

submitted in partial satisfaction of the requirements
for the degree of

MASTER OF SCIENCE

in Computer Science

by

Hee Jeong Han

Thesis Committee:
Distinguished Professor Nikil D. Dutt, Chair
Distinguished Professor Ramesh Chandra Jain
Associate Professor Amir Rahmani

2019

DEDICATION

Thank you to

my academic advisers

who guided me in this process

and

the committee

who kept me on track.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express the deepest appreciation to my advisors Professor Nikil Dutt and Professor Amir Rahmani at University of California, Irvine. They continually and convincingly conveyed a spirit of adventure in regard to research. Without their guidance and persistent help this thesis would not have been possible.

I would also like to thank my committee member, Professor Ramesh Jain who encourage me to keep researching at University of California, Irvine. I am gratefully indebted to his very valuable comments on this thesis.

Finally, I must express my profound gratitude to my parents and to my sister for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis.

ABSTRACT OF THE THESIS

Objective Stress Monitoring based on Wearable Sensors in Everyday Settings

By

Hee Jeong Han

Master of Science in Computer Science

University of California, Irvine, 2019

Distinguished Professor Nikil D. Dutt, Chair

Monitoring stress levels has become an important part of healthcare systems for physical and mental illnesses. However, current stress monitoring systems have failed to gather personal data in an everyday context. The current state of sensor technology allows us to develop systems measuring the physiological signals, which reflect stress by wearable devices. Therefore, we propose a stress monitoring system that provides an objective daily healthcare based on personal physiological signals: electrocardiogram (ECG), photoplethysmogram (PPG), and galvanic skin response (GSR). We use the wearable devices, Shimmer3 ECG, Shimmer3 GSR+ and Empatica E4 Wristband, to monitor stress ubiquitously. We perform controlled stress experiments on 17 participants and the system successfully detects 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 artifacts affect stress assessment.

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]. People nowadays exhibit more stress due to increased mental workload from stressful environments such as work [2]. In the 2016 Stress Pulse survey, 60 percent of employees have high levels of stress and 32 percent of employees keep 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 behaviors such as smoking, irregular sleep, and poor eating habits [6].

There are two kinds of stress: short-term/acute stress and long-term/chronic stress [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. When people do not relieve long-term stress, it can cause detrimental results. Since controlling stress levels is significant in life, informing people about stress in daily life has become important [8]. Therefore, monitoring stress has become a key component of modern healthcare since it can be used for early diagnosis to prevent both physical and mental diseases.

There is no accurate method for checking stress levels because stress can be affected by many factors. That is why the most accurate method of evaluating stress levels is currently

under development. One method to measure personal stress is conducting an interview with a psychologist based on a questionnaire [9]. However, this method has a reliability problem because participants might not answer correctly or might not be able to recall memories relating to why they are stressed. In order to increase reliability, objective measurements are needed. The current healthcare systems for managing stress have been confined within clinical usage. One method consists of using leukocytes to detect hormone changes when people are stressed [10]. Although it can precisely monitor stress, it is not widely used as people cannot check their hormone levels easily.

However, stress is directly related to the autonomic nervous system (ANS), which can be measured from physiological signals [11]. There are numerous physiological methods to measure stress: heart activity, blood activity, and skin response. The current state of sensor technology allows us to develop systems measuring physiological signals reflecting stress level [2]. Monitoring personal physiological signals by wireless 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 ECG to using optical sensors such as PPG and GSR [11]. Most stress detection systems collect physiological data in controlled settings. With various stress tests (e.g., memory game, presentation), those systems invoke stress and classify stress levels. However, existing systems assess stress in controlled settings. To provide a proper stress monitoring system, everyday assessment is necessary instead of short-term laboratory based assessment.

In this paper, we propose a new stress monitoring system which provides daily healthcare in everyday settings. The system collects various physiological signals (ECG, PPG, and GSR) obtained by wearable sensors and detects stress based on those personal signals during daily activities. The proposed system is supported by all existing wearable devices. With noninvasive sensors, physiological signals are collected in controlled and everyday settings. The system analyzes physiological signals from controlled tests to define stress and build a model. In order for the system to provide daily stress assessments, the system classifies physiological signals from daily life to define stress with the model. We design the system to improve the overall quality of life for people by enabling the monitoring of stress in everyday settings.

The remainder of this paper is organized as follows. Section 2 presents prior studies on stress monitoring. Section 3 provides background information on physiological parameters related to stress. Section 4 describes our methodology for 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. 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 [12] 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 [13], 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 [14]. Furthermore, it is hard to design questionnaires and interviews clearly [15]. As a result, respondents may understand questions differently. Even though psychological experts may interpret and analyze personal answers well, questionnaires and interviews are hard modalities for capturing emotional responses or mental imbalances.

To overcome these challenges, researchers have proposed laboratory-based objective stress assessment methods through analyzing stress-related hormone levels. Using

psychological stress tests for inducing stress such as the Trier Social Stress Test (TSST), changes in hormones are used to investigate stress responses [16]. 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 [17]. Beside salivary cortisol, leukocyte is used for assessing stress [10]. White blood cells from blood respond to 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 adjusts physiological activities. The parasympathetic nervous system (PNS) and the sympathetic nervous system (SNS) are two parts of the ANS [18]. 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 [19]. Due to the development of sensor technology, many studies use heart rate, health behaviors, and other vital signals to detect individual stress. Table 1 shows a comparison between ours and related works. It summarizes deployed sensors, test setting, test period, and test activities.

In [25], 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-

Table 1. Related studies and its setting of the experiment.

Related Works	Deployed Sensor	Test Setting	Period	Test Activities
Ours	**PPG, ECG, GSR**	**Lab-based**	**50mins**	**Memory Game, Mosquito Sound, IAPS Plank, Ice Test, TSST, SCWT**
		non-Lab-based	**2days**	**-**
[20]	EMG, ECH, GSR, RSP	Lab-based	20mins	Music related to emotion
[19]	ECG	Lab-based	9mins	Presentation
[21]	EEG, HR	Lab-based	not provided	Mensa Test
[22]	ECG, GSR, accelerometer	Lab-based	30mins	SCWT
[23]	EEG, GSR, PPG	Lab-based	6mins	SCWT
[24]	HR, GSR, Body Temperature	Lab-based	30mins	Tower of Hanoi 6 discs
[25]	ECG, GSR, respiration rate, blood pressure, blood oximeter	Lab-based	94mins	Memory Game, Fly Sound, IAPS, Ice Test
[26]	bioradar	Lab-based	2mins	Mathematical problem
[27]	EEG	Lab-based	10mins	Music
[28]	EEG, ECG	Lab-based	40mins	Threatening message
[29]	GSR, PPG	Lab-based	not-provided	Presentation
[2]	GSR	non-Lab-based	8hours x 4weeks	-

based experiment, which takes 94 minutes. Their machine learning based approach is trained and validated in a controlled setting where the participants are asked to carry out certain 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 artifacts).

In [22], an activity-aware mental stress detection system is proposed which also considers physical activity. The system gathers ECG, GSR and accelerometer data for 30 minutes across three activities: sitting, standing, and walking. Its experimental procedure also consists of laboratory-based stress tasks such as the Stroop Color 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 utilizes 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 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 use a combination of sensors, GSR, PPG and ECG to improve accuracy for stress identification. Furthermore, we collect daily context data from users and use it when determining stress.

Another major 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.

3. BACKGROUND

Stress can be measured by several physiological indicators such as heart activity, blood activity and skin response. We use three physiological parameters, ECG, PPG, and GSR, and extract various features from those signals to classify stress. In this section, we describe the concept of physiological parameters and features.

3.1. ECG and PPG

The ECG is a measure of the electrical activity of the heart during each cardiac cycle [30]. The ECG uses electrodes to measure electrical signals produced by depolarization and repolarization of the heart [31]. 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 [22]. Heart rate (HR) and heart rate variability (HRV) are calculated from the R-R interval.

The PPG is a measure of the electrical activity of the blood during each cardiac cycle [11]. The PPG uses an optical pulse to measure a change in blood volume and blood pressure. Since the changes of HR and HRV can be observed 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 [32]. HRV variables change in response to stress. A decrease in HRV variables has been found to be associated with stress [33]. We focus on time-domain, frequency-domain, and non-linear HRV variables.

3.1.1. 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 [34]. In the time-domain analysis, Table. 2 shows several time-domain parameters that we focus on.

Table 2. 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

3.1.2. Frequency Domain HRV

Frequency-domain variables estimate the distribution of absolute or relative power into certain frequency bands [34]. Table 3 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. In addition, the LF and HF components are respectively associated with the sympathetic nervous system (SNS) and parasympathetic nervous system (PNS) activities in the nervous system [35][36]. 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 low parasympathetic activity, which is characterized by a decrease in the HF and an increase in the LF [33].

Table 3. 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

3.1.3. Nonlinear Domain HRV

Non-linear indices quantify the unpredictability of a time series [37]. Non-linear indices have a correlation with time-domain indices and frequency-domain indices. Table 4 shows several non-linear parameters that we focus on. A Poincar plot is a graph plotting every RR interval against the prior interval [34]. Poincar plot analysis shows patterns within a sequence of values from successive R-R intervals. It does not affect changes of the R-R intervals rapidly [38]. SD1 describes the width of the ellipse and has a correlation with HF. SD2 describes the length of the ellipse and has a correlation with LF. SD1/SD2 which is related to LF/HF, is used to measure stress in sympathetic activity.

Table 4. Non-linear domain HRV.

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

3.2 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 of the skin to conduct electricity to detect increased stress [11]. 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.

4. METHODOLOGY

The stress monitoring system provides an assessment of stress levels to the user. We collect physiological signals: ECG, PPG, GSR. We have two kinds of settings: controlled setting and everyday setting. To find a correlation between stress and physiological signals, we do offline laboratory-based stress tests when collecting personal signals from wearable devices. With the collected physiological signals, we process raw signals to extract features. We make a model using 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 project overview in a controlled setting.

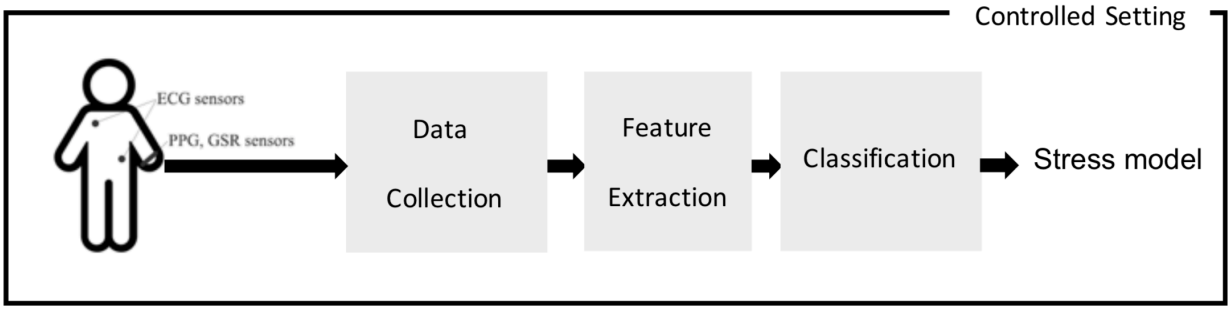

Figure 1. Project overview in controlled setting.

We also collect physiological signals in a everyday setting through wearable devices to find daily stress levels. With everyday data, we do feature extraction and prediction through the model from the controlled setting to get personal stress levels. The proposed system overcomes the problem of current stress monitoring methods, which cannot analyze personal stress obtained in everyday activities. Figure 2 shows the project overview in the everyday setting.

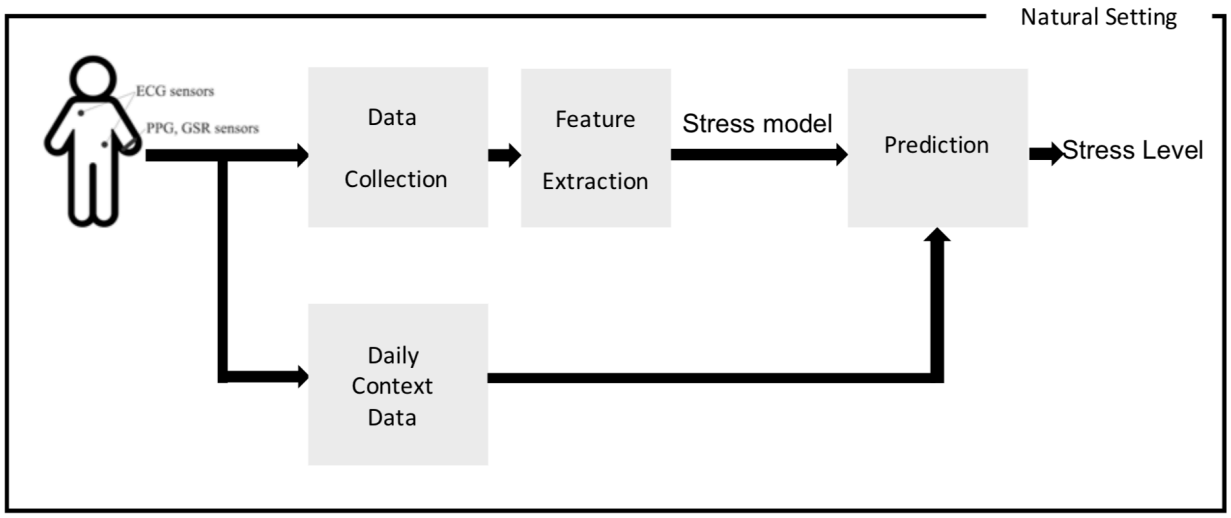

Figure 2. Project overview in the everyday setting.

4.1 Device

We use Shimmer3 ECG, Shimmer3 GSR+, and Empatica E4 Wristband devices. We collected the ECG data from Shimmer3 ECG, and the PPG, and the GSR data from Shimmer3 GSR+. Empatica E4 Wristband is used for gathering the PPG, and the GSR data.

4.2 Data collection in controlled setting

In a controlled setting, we do laboratory-based stress tests. Laboratory-based stress tests consist of several stress tasks for inducing short-term stress. During stress tests, the expected result is 1 for stress tasks, and 0 otherwise. We implement two kinds of laboratory-based tests. Figure 3 is our experimental procedure. The ECG has a sampling rate of 512 Hz, however, the PPG and the GSR have a sampling rate of 128 Hz.

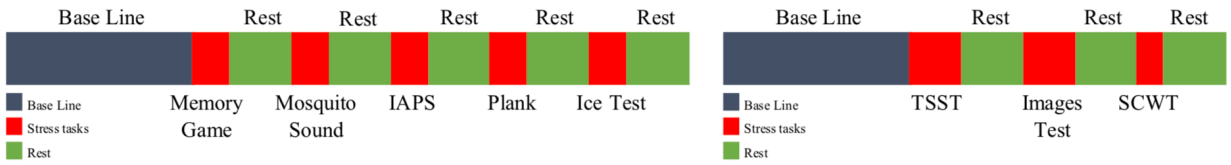

Figure 3. Test procedure collecting physiological signals related to stress.

The first laboratory-based test takes approximately 50 minutes for each participant. It consists of five stress tasks: Memory game, Mosquito sound, Images Test, Plank, and Ice Test [25]. Each stress task lasts for 2 minutes, followed by a 6 minute rest period between each stress task. Baseline is the first stage of the test to collect basic personal physiological signals. We asked participants to meditate while listening to classical music. Rest is a rest period for the participants after each stress task where they would listen to 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 would listen to a mosquito sound with a black screen to prevent distraction [40]. Images Test is a period for the participants to see selected pictures from the International Affective Picture System (IAPS) [41]. IAPS provides affect ratings of pictures. Plank is a period for the participants to do 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 [25].

The second laboratory-based test also takes approximately 50 minutes for each participant. It consists of three stress tasks: TSST, Images Test, and SCWT. TSST is a laboratory

procedure used to reliably induce stress in human research participants [16]. TSST consists of two stress tasks: speech and math. In the speech portion, there is a preparation step and a presentation step for a given topic, which are each 5 minute period. After the presentation, we ask participants to count backwards from 1022 subtracting 13 for 5 minutes. SCWT which is based on the Stroop Effect, provides colored word lists [42]. Participants read those word lists while following the instructions.

A total of 17 participants (13 male, 4 female), with ages between 20 and 27 years, participated in the laboratory-based experiment. All of them are undergraduate or graduate students.

4.3 Data collection in everyday setting

In a everyday setting, we collect physiological signals in daily life. We ask participants to wear a smart wrist band, Empatica E4 Wristband, and a Shimmer ECG device. We also collect daily context data with a stress label. Participants label whether they are stressed out or not every 30 minutes. The ECG has a sampling rate of 512 Hz, however, the PPG has a sampling rate of 64 Hz and the GSR has a sampling rate of 4 Hz. 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. We implement preprocessing and feature extraction. Each raw data has two steps for preprocessing due to

noise: filtering and smoothing. Figure 4 shows the procedure of preprocessing feature extraction.

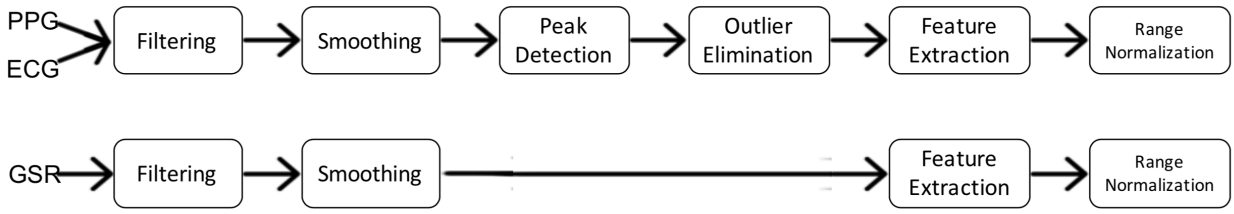

Figure 4. The procedure of preprocessing feature extraction.

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 moving average filter to smooth the data. From the filtered ECG data, we extract the mean of HR. We also extract the mean 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 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 function that is used for the ECG.

In order to extract skin conductance, the GSR signal needs to be preprocessed with the median filter and moving average filter. The median filter is used for removing noise, and

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 will increase noise [44]. Computing some loosely correlated features may also not be useful because of their complexity. For instance, frequency domain and independent features have non-linear computational complexity and calculating them has computational overhead. Especially in local implementations in IoT systems, these overheads are considerable. Thus, we decide to select features which are more related to stress.

In order to find the best subset of features, we use a greedy stepwise method [45]. In this method, it starts from an empty set. It adds features which 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 5-nearest-neighbor classifier, correlation-based feature selection method, and information gain. By using the method, we decide to use the features shown in Table 5.

Table 5. 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 Classification

The bias of physiological data can vary by using personal data sets or general data sets [46]. Personal data sets contain data collected from the same person and general data sets contain data from other subjects. In order to test the efficiency of our classifier, we need to test it in both cases.

We use several classifiers, which are K-nearest-neighbor (kNN) with k ∈ {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 [47]. SVM finds hyper-planes to divide data-points into different classes [48]. We used the Weka implementation of LIBSVM [49]. Naive Bayes classifiers predict the result based on the probabilities of each feature's probabilistic knowledge [50]. Naive Bayes classifiers act differently based on the distribution of data-points [51].

5. EXPERIMENTAL RESULTS

In this section, we present our experimental results in the controlled and everyday settings. First, we validate our developed stress models using three different classification algorithms (i.e., kNN, SVM, and Naive Bayes). We test whether the classifiers generalize 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 analyze techniques to mitigate them. We use everyday self-report stress label as ground truth. We also collect context data (e.g., running, walking, eating, etc.) to evaluate the effect of noise such as motion artifacts on the decisions in everyday settings. To examine how a combination of features affect 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 [52] 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 minutes of data for the controlled experiment as training data. Out of these, 234 minutes are during stressful tasks and 133 are during the baseline (i.e., labeled 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 minutes data are provided with self-reported stress labels. Labels are associated with the stress reported for each 30 minutes.

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 set of experiments: 1) with all features, and 2) with selected features (presented in Section 4.5). In addition, we analyze 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 generalize across data-points with 10-fold cross-validation [53]. Figure 5 shows the accuracy of three different classifier, 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.

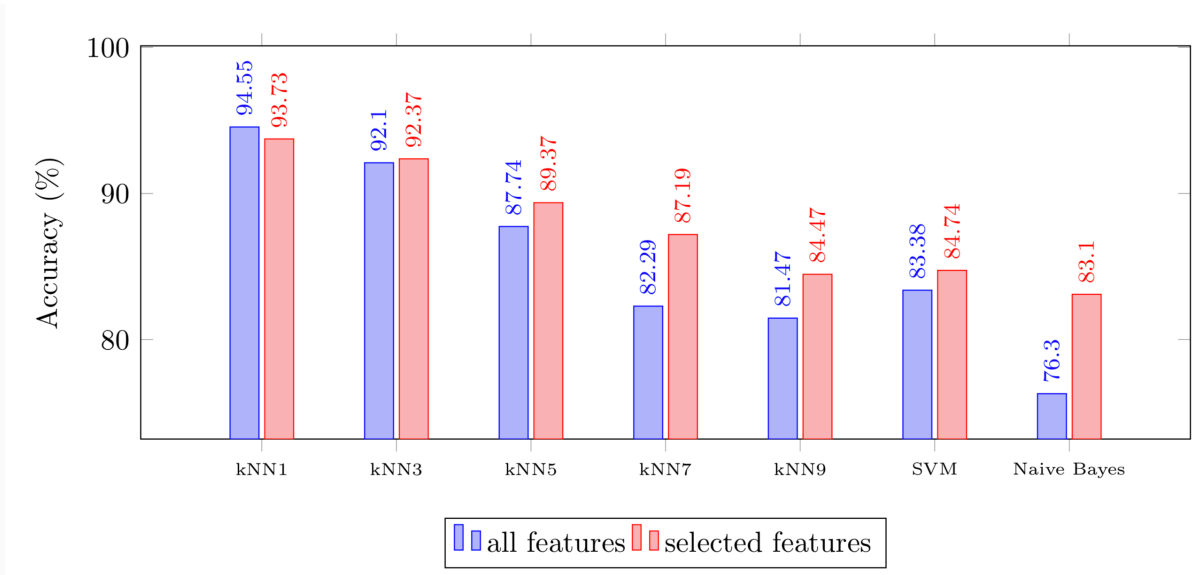

Figure 5. 10-fold leave data points out cross-validation accuracy of the different classifiers using the different number of features.

Table 6 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.

Table 6. 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**
[20]	EMG, ECH, GSR, RSP	92.0
[22]	ECG, GSR, accelerometer	92.4
[23]	EEG, GSR, PPG	81.8
[24]	HR, GSR, Body Temperature	84.5
[25]	ECG, GSR, respiration rate, blood pressure, blood oximeter	89.3
[26]	bioradar	94.4
[27]	EEG	80.1
[28]	EEG, ECG	79.6

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 generalize 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 corresponds too closely to a particular set of data in a high bias. Overfitted classifiers perform worse on validation [54]. Since SVM can avoid overfitting appropriate, it shows the best accuracy rather than other classifiers.

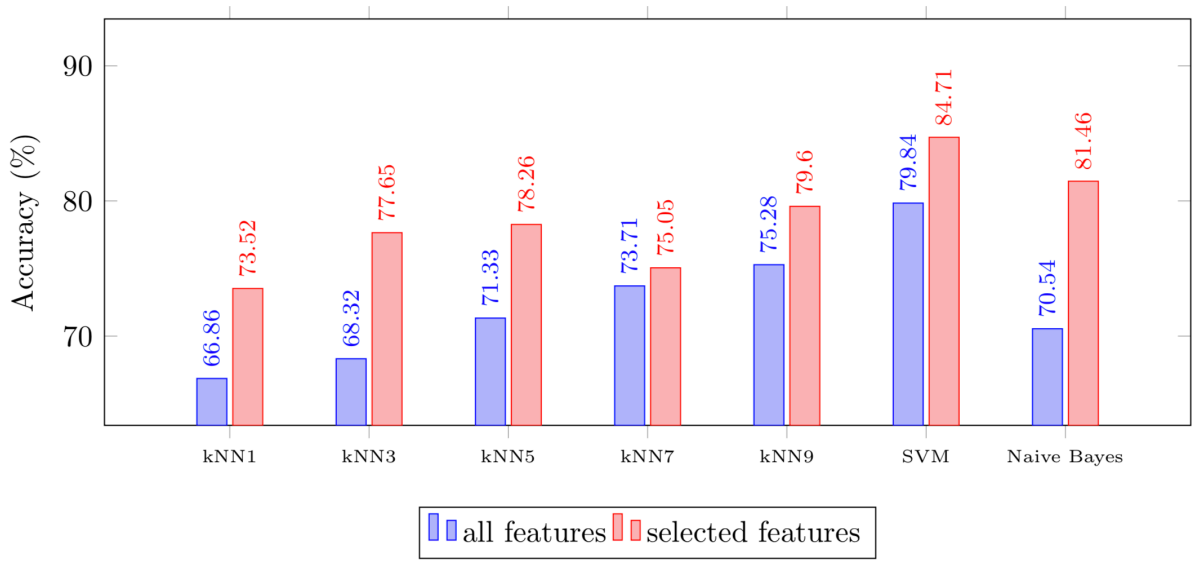

Figure 6. 10-fold leave subjects out cross-validation accuracy of the different classifiers using the different number of features.

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 (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 which has used subjectwise cross-validation in this context.

Table 7. 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**	**84.3**
[22]	ECG, GSR, accelerometer	80.9

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 minutes. 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 [55]. We use two third majority to consider a prediction reliable.

5.2.1. Cross-validation accuracy without activity recognition

We evaluate 340 minutes everyday data, which have self-assessed stress labels. It includes various kinds of activities such as sitting, walking and eating. Figure 7 shows 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 from all features. This is because loosely correlated features are removed.

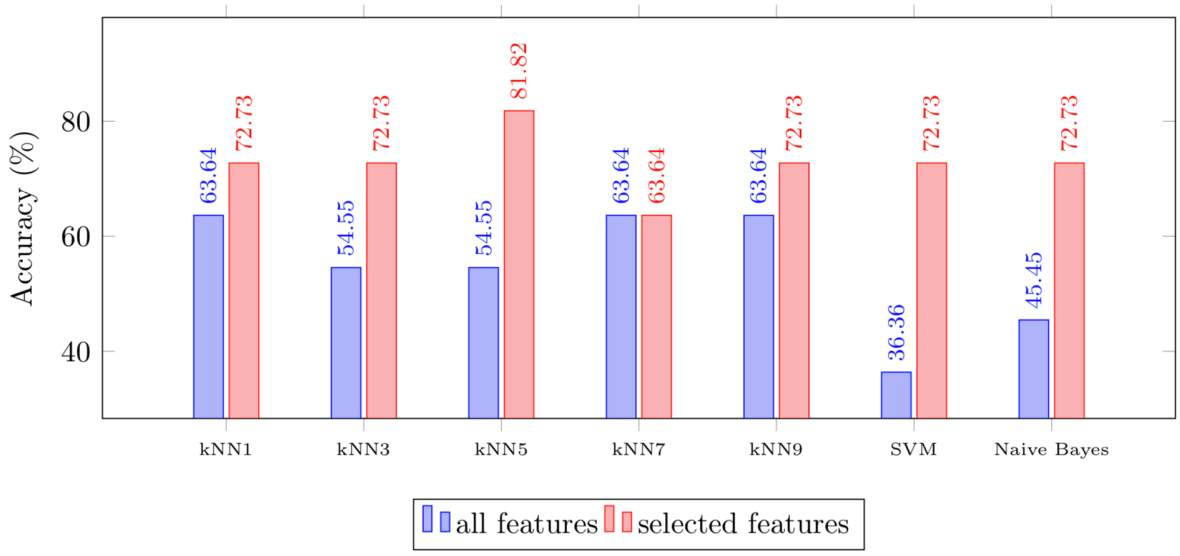

Figure 7. Everyday stress assessment accuracy of the different classifiers using the different number of features.

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 easy 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.

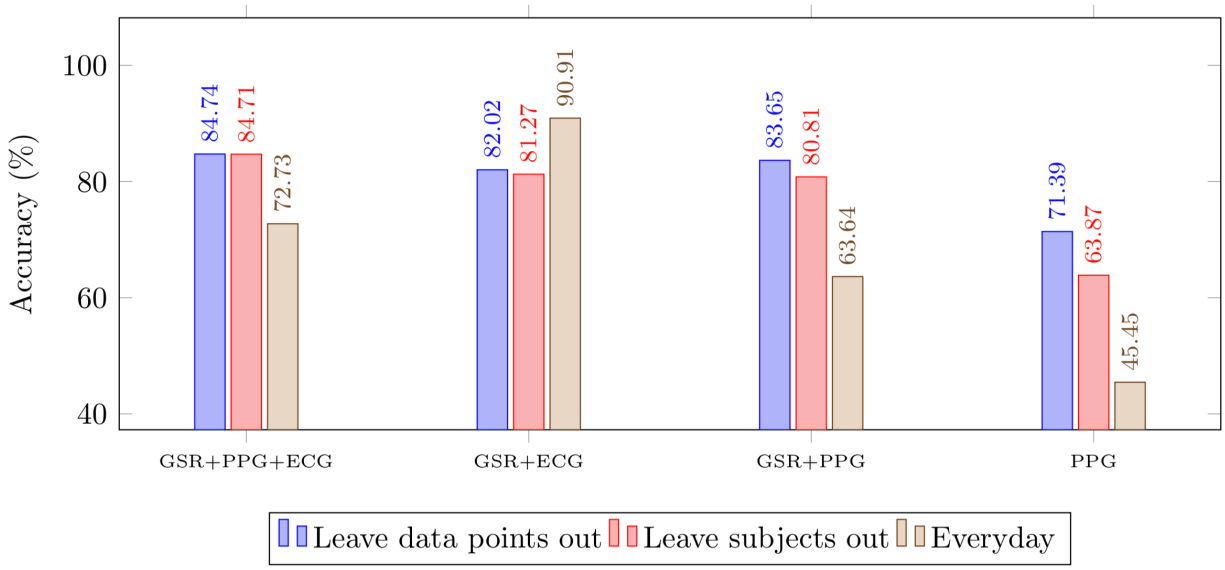

Figure 8. Comparison of feature combination with feature selection.

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 [56]. Even though PPG is rising to use in a personal healthcare system, PPG is sensitive to a movement [57]. 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 affect PPG signal increasing noise.

5.2.2. Cross-validation accuracy with activity recognition

Since noise is increased in PPG signal in everyday settings, we examine how much motion artifacts affect PPG signal. We extract activities based on personal daily context report. We contain data-points during low intensive activities such as sitting, but exclude data-points during high intensive activity such as walking. Figure 9 shows 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 artifacts, 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 for decrease false positive rate.

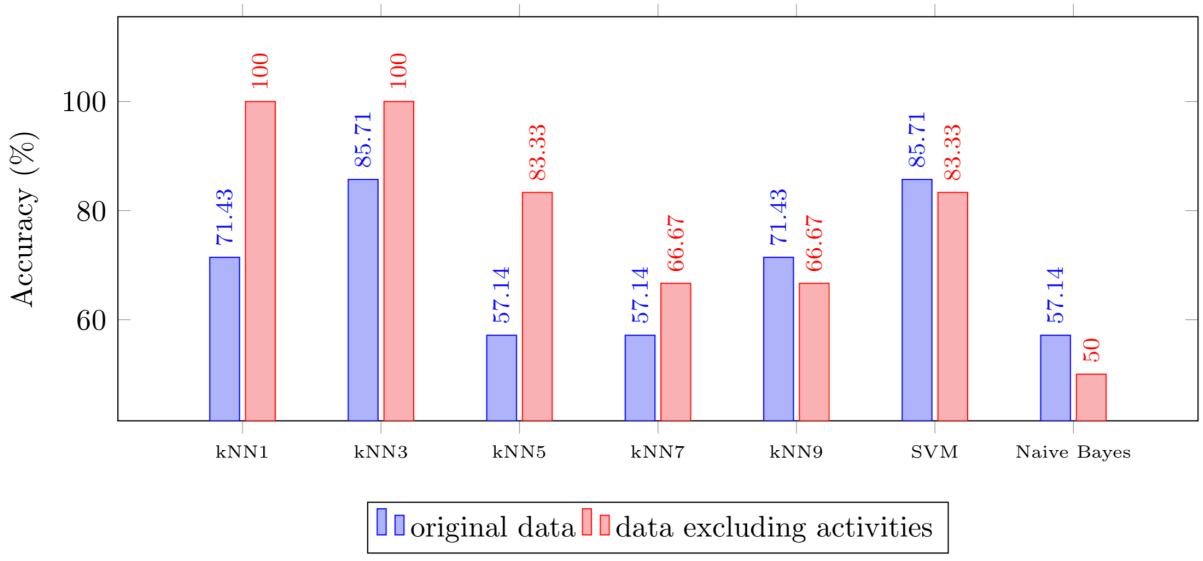

Figure 9. Comparison of everyday stress assessment accuracy between original data and data excluding activities with feature selection.

6. DISCUSSION

Our new objective stress monitoring system provides stress assessment within 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. Although the system provides the best accuracy in the controlled setting, daily stress monitoring can improve its accuracy. Examining everyday data is more problematic than controlled data. Physiological data is sensitive to movement. In experimental tests, participants are sitting; however, in everyday settings, many activities such as walking, running, and eating involve movement. 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 important for objective stress monitoring. We suggest an engine, which provides the PPG confidence rate automatically [58]. The system compares PPG to ECG and finds a reliable period of PPG. It uses raw signals and labels to give confidence rate every minute. For labels, we set 5 beats per minute as the threshold [59]. When the difference between heart rate from PPG and heart rate from ECG is within the threshold, a label is marked reliable. Otherwise, it is marked as unreliable data. The system uses a convolutional neural network. In everyday data, the PPG confidence engine reports 244 minutes over 340 minutes, so that the PPG data for 96 minutes is unstable because of motion artifacts. Through the engine, we can prevent the false reporting of the PPG.

7. CONCLUSION AND FUTURE WORK

We propose a new stress monitoring system that supports everyday stress assessment. We designed, implemented, and analyzed the system giving not only high accuracy stress detection in the controlled setting but also a good prediction in the everyday setting. This system is shown to have 94.55% percent accuracy in the generalized model for stress detection and shown to have 85.71% percent accuracy when the classifier generalizes across subjects. In the everyday setting, this system is shown to have 81.82% percent accuracy. Our system is compared with previous related studies shown in Table. 1, which compares sensors used, accuracy in the generalized model, test sets, test period, and test activities.

In order to use the stress monitoring system, there are more features to implement. **First**: taking activity and movement into account can help detect stress even during movements. **Second**: collecting data in an everyday setting by a high-frequency device is not practical due to battery life, cost, and convenience. We need to extract the same features from less accurate devices but provide equal accuracy. **Third**: all acute stress is not always bad. In order to detect bad acute stress, we need to find the correlation between bad acute stress and people's lifestyle. Finding acute stress is the first step toward finding episodic acute stress, and chronic stress which is more harmful and more difficult to find.

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. In: Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on; IEEE; 2011. p. 573–580.

[3] 2016 compsych stresspulse survey ; 2016.

[4] Sterling P. Allostasis: a new paradigm to explain arousal pathology. Handbook of life stress, cognition and health. 1988;.

[5] Selye H. Stress in health and disease. Butterworth-Heinemann; 2013.

[6] Glanz K, Schwartz MD. Stress, coping, and health behavior. Health behavior and health education: Theory, research, and practice. 2008;4:211–236.

[7] Miller LH, Smith AD, Rothstein L. The stress solution: An action plan to manage the stress in your life. Pocket; 1994.

[8] Choi JB, Dimsdale J, Bardwell W, et al. Effects of stress on heart rate complexitya comparison between short-term and chronic stress. Biological Psychology. 2008;80(3):325– 332.

[9] Andreou E, Alexopoulos EC, Lionis C, et al. Perceived stress scale: reliability and validity study in greece. International journal of environmental research and public health. 2011; 8(8):3287–3298.

[10] Davis A, Maney D, Maerz J. The use of leukocyte profiles to measure stress in vertebrates: a review for ecologists. Functional Ecology. 2008;22(5):760–772.

[11] 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 Electronics Magazine. 2016;5(4):44–56.

[12] Holmes TH, Rahe RH. The social readjustment rating scale. Journal of psychosomatic research. 1967;.

[13] Stress assessment questionnaire [https://ptc.bps.org.uk/test-review/stress-assessment- questionnaire]; 2012.

[14] Patten ML. Questionnaire research: A practical guide. Routledge; 2016.

[15] Phellas CN, Bloch A, Seale C. Structured methods: interviews, questionnaires and obser-vation. Researching society and culture. 2011;3.

[16] 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.

[17] Hellhammer DH, Wu¨st S, Kudielka BM. Salivary cortisol as a biomarker in stress research. Psychoneuroendocrinology. 2009;34(2):163–171.

[18] Langley JN. The autonomic nervous system (pt. i).. 1921;.

[19] Schubert C, Lambertz M, Nelesen R, et al. Effects of stress on heart rate complexity a comparison between short-term and chronic stress. Biological psychology. 2009;80(3):325–332.

[20] Wagner J, Kim J, Andr´e E. From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification. In: 2005 IEEE international conference on multimedia and expo; IEEE; 2005. p. 940–943.

[21] Taelman J, Vandeput S, Spaepen A, et al. Influence of mental stress on heart rate and heart rate variability. In: 4th European conference of the international federation for medical and biological engineering; Springer; 2009. p. 1366–1369.

[22] Sun FT, Kuo C, Cheng HT, et al. Activity-aware mental stress detection using physiological sensors. In: International Conference on Mobile Computing, Applications, and Services; Springer; 2010. p. 282–301.

[23] Das D, Bhattacharjee T, Datta S, et al. Classification and quantitative estimation of cognitive stress from in-game keystroke analysis using eeg and gsr. In: Life Sciences Conference (LSC), 2017 IEEE; IEEE; 2017. p. 286–291.

[24] Ciabattoni L, Ferracuti F, Longhi S, et al. Real-time mental stress detection based on smartwatch. In: 2017 IEEE International Conference on Consumer Electronics (ICCE); IEEE; 2017. p. 110–111.

[25] Akmandor AO, Jha NK. Keep the stress away with soda: Stress detection and alleviation system. IEEE Transactions on Multi-Scale Computing Systems. 2017;3(4):269–282.

[26] Fern´andez JRM, Anishchenko L. Mental stress detection using bioradar respiratory signals. Biomedical Signal Processing and Control. 2018;43:244–249.

[27] Liao CY, Chen RC, Tai SK. Emotion stress detection using eeg signal and deep learning technologies. In: 2018 IEEE International Conference on Applied System Invention (ICASI); IEEE; 2018. p. 90–93.

[28] Xia L, Malik AS, Subhani AR. A physiological signal-based method for early mental-stress detection. Biomedical Signal Processing and Control. 2018;46:18–32.

[29] Koussaifi M, Habib C, Makhoul A. Real-time stress evaluation using wireless body sensor networks. In: 2018 Wireless Days (WD); IEEE; 2018. p. 37–39.

[30] Dupre A, Vincent S, Iaizzo PA. Basic ecg theory, recordings, and interpretation. In: Handbook of cardiac anatomy, physiology, and devices. Springer; 2005. p. 191–201.

[31] Klabunde R. Cardiovascular physiology concepts. Lippincott Williams & Wilkins; 2011.

[32] McCraty R, Shaffer F. Heart rate variability: new perspectives on physiological mechanisms, assessment of self-regulatory capacity, and health risk. Global Advances in Health and Medicine. 2015;4(1):46–61.

[33] Kim HG, Cheon EJ, Bai DS, et al. Stress and heart rate variability: A meta-analysis and review of the literature. Psychiatry investigation. 2018;15(3):235.

[34] Shaffer F, Ginsberg J. An overview of heart rate variability metrics and norms. Frontiers in public health. 2017;5:258.

[35] Healey J, Picard RW, et al. Detecting stress during real-world driving tasks using physiological sensors. IEEE Transactions on intelligent transportation systems. 2005;6(2):156–166.

[36] Acharya UR, Joseph KP, Kannathal N, et al. Heart rate variability. In: Advances in cardiac signal processing. Springer; 2007. p. 121–165.

[37] Stein PK, Reddy A. Non-linear heart rate variability and risk stratification in cardiovascular disease. Indian Pacing and Electrophysiology Journal. 2005;5(3):210.

[38] Behbahani S, Dabanloo NJ, Nasrabadi AM. Ictal heart rate variability assessment with focus on secondary generalized and complex partial epileptic seizures. Advances in Bioresearch. 2013;4(1).

[39] Memory game [https://www.mathworks.com/matlabcentral/fileexchange/14059-memory-a-k-a-concentration]; 2007.

[40] Mosquito fly sound [https://www.youtube.com/watch?v=PYnVlOoxZWw]; 2014.

[41] Lang PJ. International affective picture system (iaps): Affective ratings of pictures and instruction manual. Technical report. 2005;.

[42] Stroop JR. Studies of interference in serial verbal reactions. Journal of experimental psychology. 1935;18(6):643.

[43] Gupta A, Bhandari S. ECG Noise Reduction By Different Filters - A Comparative Analysis. International Journal of Research in Computer and Communication Technology. 2015; 4(7):424–431.

[44] Deng Y, Wu Z, Chu CH, et al. Evaluating feature selection for stress identification. In: 2012 IEEE 13th International Conference on Information Reuse & Integration (IRI); IEEE; 2012. p. 584–591.

[45] Hall MA. Correlation-based feature selection for machine learning; 1999.

[46] Das D, Datta S, Bhattacharjee T, et al. Eliminating individual bias to improve stress detection from multimodal physiological data. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); IEEE; 2018. p. 5753–5758.

[47] Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician. 1992;46(3):175–185.

[48] Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995;20(3):273–297.

[49] Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology. 2011;2:27:1–27:27. Software available at http:// www.csie.ntu.edu.tw/ cjlin/libsvm.

[50] Hand DJ, Yu K. Idiot's bayesnot so stupid after all? International statistical review. 2001; 69(3):385–398.

[51] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training. In: Proceedings of the eleventh annual conference on Computational learning theory; ACM; 1998. p.92–100.

[52] Hall M, Frank E, Holmes G, et al. The weka data mining software: an update. ACM SIGKDD explorations newsletter. 2009;11(1):10–18.

[53] Kohavi R, et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Ijcai; Vol. 14; Montreal, Canada; 1995. p. 1137–1145.

[54] Hawkins DM. The problem of overfitting. Journal of chemical information and computer sciences. 2004;44(1):1–12.

[55] James G. Majority vote classifiers: theory and applications [dissertation]. Stanford University; 1998.

[56] Pinheiro N, Couceiro R, Henriques J, et al. Can ppg be used for hrv analysis? In: 201638th Annual International Conference of the IEEE Engineering in Medicine and BiologySociety (EMBC); IEEE; 2016. p. 2945–2949.

[57] Pietil¨a 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. In: Embec & nbc 2017. Springer; 2017. p. 145–148.

[58] Naeinia EK, Azimib I, Rahmania AM, et al. A real-time ppg quality assessment approach for healthcare internet-of-things. 2018;.

[59] Kobayashi H. Effect of measurement duration on accuracy of pulse-counting. Ergonomics.2013;56(12):1940–1944