Stress Detection and Management in Daily Life Using Wearable Sensors

by

Yekta Said CAN

B.S., Computer Engineering, Boğaziçi University, 2012

M.S., Computer Engineering, Boğaziçi University, 2014

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Graduate Program in Computer Engineering

Boğaziçi University

2020

ACKNOWLEDGEMENTS

I would like to thank my thesis supervisor Professor Cem Ersoy. His encouragement and rigorous, continuous guidance made the research easier. I have been really lucky to know him and work with him.

My deepest gratitude goes to my family, my wife Nihal CAN, my parents Muhiddin, Müşerref CAN and my little sister Asiye Sevde CAN for their unflagging love and support throughout my life. I would like to thank my daughter Ceyda CAN for bringing joy to our family.

I would like to thank our stress detection research group Deniz and Niaz. Especially in data collection periods, we worked very hard together and decrease the workload of each other.

I also would like to express my deep and sincere gratitude to all my friends Mert, Sancar, Hakan, Akif. I want to thank these nice people for the pleasant time I had with them.

I would like to thank Affectech Project which is European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 722022 and the Turkish Directorate of Strategy and Budget under the TAM Project number DPT2007K120610 for making my research easier.

ABSTRACT

Stress Detection and Management in Daily Life Using Wearable Sensors

Stress has become an integral part of our modern society. Researchers investigated ways to cope with it to alleviate its negative effects on human health, society and economy. At this point, widespread usage of smartphones, smartwatches and smart wrist-bands raised the question of whether we can detect and alleviate stress with them. Although research has traditionally been conducted in laboratory settings, a set of new studies have recently begun to be conducted in ecological environments with unobtrusive wearable devices. In this thesis, we developed a stress detection system for daily life. Unobtrusive wearable devices were used for physiological data collection. For that purpose, we used heart rate variability (HRV) and electrodermal activity (EDA) signals. Modality specific artifact detection and removal algorithms, feature extraction and advanced machine learning methods were proposed. We tested our system in a laboratory environment, restricted, semi-restricted and unrestricted real-life environments by collecting data in each environment. We proposed different techniques to improve the state of the art in real life environments. We worked on prominent environment-specific research questions. We further examined different stress alleviation methods including those which can be applied indoors. We also discussed promising techniques, alleviation methods and research challenges for daily life stress management.

ÖZET

Günlük Hayatta Giyilebilir Cihazlar Kullanılarak Stress Seviyesinin Tespiti ve Düşürülmesi

Stres artık modern toplumumuzun ayrılmaz bir parçası oldu. Araştırmacılar insan sağlığı, toplum ve ekonomi üzerindeki olumsuz etkilerini hafifletmek için stresle başa çıkmanın yollarını araştırdılar. Bu noktada akıllı telefonların, akıllı saatlerin ve akıllı bilekliklerin yaygın şekilde kullanılması, bu cihazlarla stresin algılanabileceği ve hafifletebilineceği araştırma konusunu gündeme getirdi. Her ne kadar geleneksel olarak laboratuvar ortamlarında araştırmalar yapılmış olsa da, son zamanlarda göze çarpmayan giyilebilir cihazlar içeren ekolojik ortamlarda bir dizi yeni çalışma başlatılmaya başlandı. Bu tezde günlük yaşam için bir stres algılama sistemi geliştirdik. Fizyolojik veri toplama için rahatsızlık vermeyen giyilebilir cihazlar kullanıldı. Bu amaçla kalp ve deri iletkenliği aktivitesi fizyolojik sinyalleri kullanıldı. Değişik kiplere özgü hatalı kayıt algılama ve düzeltme algoritmaları, öznitelik çıkarımı ve gelişmiş makine öğrenme yöntemleri önerildi. Bütün ortamlarda veri toplayarak sistemimizi laboratuvar ortamında, sınırlı, yarı sınırlı ve sınırsız gerçek yaşam ortamlarında test ettik. Her birinde seçilen ortama özgü araştırma sorularını inceledik. Gerçek hayattaki sistemlerin başarımlarını geliştirmek için yeni yöntemler önerdik. İç mekanlarda uygulanabilecek farklı stres iyileştirme yöntemleri önerdik ve inceledik. Ayrıca günlük yaşam stres yönetimi çalışmaları için umut vaat eden teknikler, hafifletme yöntemleri ve araştırma zorluklarını tartıştık.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

H_i	Happiness score
C_i	Cheerfulness score
c_t	memory of block t
h_t	output of block t
b_n	bias vector n
A_i	Anger score
S_i	Sadness score
F_i	Frustration score
σ	Sigmoid activation function

LIST OF ACRONYMS/ABBREVIATIONS

ABRF	AdaBoost with Random Forest Classifier
ANS	Autonomic Nervous System
AC	Affective Computing
ANN	Artificial Neural Networks
BVP	Blood Volume Pulse
BP	Blood Pressure
CBT	Cognitive Behavioural Therapy
CL	Cognitive Load
CNN	Convolutional Neural Network
DPSL	Daily Perceived Stress Level
DSRGT	Daily-life model trained with Self-Report Ground Truth
ECG	ElectroCardiogram
EDA	Electro-Dermal Activity
EEG	Electroencephalogram
EMA	Ecological Momentary Assessment
EMG	Electromyogram
ERP	Event-Related Potentials
ERVE	Emotional Responses in Virtual Environments
FD	Functional Descriptor
fMRI	functional Magnetic Resonance Imaging
GMM	Gaussian Mixture Model
GSR	Galvanic Skin Response
GWBI	General Wellbeing Index
HCI	Human Computer Interaction
HHT	Hilbert Huang Transform
HMM	Hidden Markov Models
HF	High Frequency
HR	Heart Rate

HRV	Heart Rate Variability
IBI	Interbeat Intervals
IAPS	International Affective Picture System
ICA	Independent Component Analysis
kNN	k Nearest Neighbors
LLD	Low-level descriptors
LDA	Linear Discriminant Analysis
LTPSL	Long-Term Perceived Stress Level
LSTM	Long Short Term Memory
LF	Low Frequency
LR	Logistic Regression
LSRGT	Laboratory model trained with Self-Report Ground Truth
LKGCT	Laboratory model trained with Known Context Ground Truth
MFCC	Mel Frequency Cepstral Coefficients
ML	Machine Learning
MV	Morphological Variability
MLP	Multilayer Perceptron
MIST	Montreal Imaging Stress Task
NASA-TLX	NASA Task Load Index
NCC	Nearest Class Center
PANAS	Self Assessment Manikin and Positive and Negative Affect Schedule
OLBI	Oldenburg Burnout Inventory
pLF	Prevalent Low Frequency
pHF	Prevalent High Frequency
PET	Positron Emission Tomography
PNS	Parasympathetic Nervous System
PCA	Principal Component Analysis
PD	Pupil Diameter
PPG	Photoplethysmography
PSL	Perceived Stress Level
PSQ	Perceived Stress Questionnaire

PSQI	Pittsburgh Sleep Quality Index
PSS	Perceived Stress Scale
QDA	Quadratic Discriminant Analysis
RNN	Recurrent Neural Network
RR	R peak to R peak Interval
SOM	Self Organizing Maps
SCWT	Stroop Color-Word Test
SSRS	Stress Self-Rating Scale
SCL	Skin Conductance Level
SCR	Skin Conductance Response
ST	Skin Temperature
SNS	Sympathetic Nervous System
SSST	Sing-a-Song Stress Test
SVM	Support Vector Machines
SFFS	Sequential Forward Floating Selection
SVI	Sympathovagal Balance Index
TI	Thermal Imaging
TMST	Trier Mental Stress Test
TSST	Trier Social Stress Test
VLF	Very Low Frequency
VR	Virtual Reality
WHO-5	WHO Five Well-being Index

1. INTRODUCTION

Stress is a widely used term, but it is not easy to agree on the definition because it is a subjective and hard to define phenomenon [8]. However, if scientists cannot define stress, how could they quantify it? Merriam Webster dictionary defines stress as a physical, chemical, or emotional factor that causes bodily or mental tension and may be a factor in disease causation. Stress can be non-formally defined as the body's way of reacting to any demanding or hazardous situation [9].

The stress reaction is initiated by the brain as a response to sensory inputs from the eye, nose, or ear. When the body senses a threat, it could be real or imagined; defensive mechanisms of the body initiates a rapid, automatic process called as the "fight-or-flight" reaction or the stress response to protect itself. The brain immediately sends a distress signal to the hypothalamus. Hypothalamus is analogous to the command center of the brain [10]. Hypothalamus controls involuntary body functions via the autonomic nervous system (ANS). ANS constitutes two elements, which are the sympathetic nervous system (SNS) and the parasympathetic nervous system (PNS). SNS can be considered analogous to a gas pedal in a car [10]. In fight or flight situations, after getting the distress signal, the hypothalamus activates SNS. SNS releases stress hormones such as epinephrine and cortisol, which arouse the body to act in emergencies. Heart rate increases, muscles tighten, blood pressure increases, inhalation frequency rises when the SNS is activated. Volumes of the airways in the lungs increase [10]. Oxygen in the brain increases and this causes senses to become sharper [9]. Blood sugar is also increased which makes an individual more energetic. The abovementioned changes improve the strength and stamina, shortens the reaction time, and improve the focus. If this mechanism works as desired, it helps an individual to stay focused and alert. When the danger passes, PNS, which is analogous to the brake in the car, diminishes the stress reaction.

The abovementioned stress reaction is the physiological response of an individual to a stressful event and can be named as physiological stress. Another type of stress occurs due to mental appraisal and interpretation of the stressful situations by an individual and can be called as perceived stress [11]. Perceived stress can be measured

with periodical self-reports collected from the individuals [11]. In other words, it is the perception of the stress from an individual's point of view. Theoretically, researchers expect that these two stress levels have a coherency between them. However, the perception could be subjective, it can change from person to person. Liapis et al. [12] give examples regarding the subjectivity of the self-reports. They demonstrated that although physiological data is similar for both genders in their laboratory experiments, women tend to express the stress level more than men on self-reported data. Another problem of perceived stress is that since the self-reports are collected at certain periods, they can miss the stress episodes in some cases. Stress has a negative effect on memory and people have a tendency to forget stressful events [13], especially if the survey collection period is long. These two problems may result in discrepancies between the measured physiological stress levels and perceived stress levels. The collection of perceived stress data is also more difficult than physiological data collection. Although for the collection of the physiological data, an unobtrusive wearable device can be sufficient for the perceived stress, participants should fill some surveys periodically which could create a burden on the individuals.

Stress helps the individual to survive in dangerous cases. Up to a certain level, stress can also be helpful in demanding situations such as a presentation at work, exam in school. As abovementioned, stress is a mechanism of the body that helps and saves us in critical cases. However, after it exceeds a certain level, stress is not beneficial anymore, on the contrary, it starts damaging the health, emotional state, productivity and life quality of an individual. The reason is that our nervous system cannot differentiate between emotional and physical dangers [9]. In emotional demanding situations such as an argument with a friend or a deadline, the nervous system reacts in the same way as a daily life-or-death situation. If this activation becomes frequent and an individual stresses out more, the body would be stressed most of the time, and this can result in serious health problems.

Daily life stress has become a significant issue for the modern society. Therefore, recognizing stress is one of the most commonly investigated research issues in the affective computing literature (see [14], [15], [16], [17] and [18]). Offices, among other places, contribute to the high stress most [19]. The mismatch between job demands

and abilities, time pressure and high workloads are ample reasons for the office stress. Family-related issues, illnesses and chronic injuries and emotional problems can be listed as off-the-workplace stress causes. Stress is the second most severe work-related health issue in Europe [20], after musculoskeletal illnesses which can be caused by stress in some cases [21]. The American Institute of Stress reveals that the US spends 300 billion USD per year for diseases caused by stress [22]. In 2013, work-related stress cost 25 billion euros to the EU businesses [23]. A recent public survey [24] revealed that 51% of European workers are exposed to stress at workplaces. It is anticipated that 50 - 60% of all lost working days in the European business sector are caused by work-related stress and psychosocial risks [19].

There are two types of stress as far as the duration is concerned: acute and chronic [25]. Acute stress is more common and the majority of people have experienced this kind of stress. American Psychological Association stated that demands and pressure from the recent past and near future cause acute stress [26]. The potential triggers for acute stress can be listed as athletic challenges, test-taking, or anxiety when meeting new people. The causes of chronic stress can be counted as long-standing pressures and demands as a result of socioeconomic conditions, difficulties in interpersonal relationships, or an unsatisfying career [26]. The consequences of chronic stress can be destructive if left unmanaged [27]. Since the short-term symptoms of acute stress are more observable, researchers have investigated acute stress more than chronic stress. Furthermore, since the duration of acute stress is shorter, subjects are less affected when they are induced in the experiments when compared to chronic stress.

As abovementioned, stress has noticeable effects on human health. In acute stress, possible symptoms can be listed as emotional distress, muscular ache and tension, digestive tract issues, and overarousal [28]. Overarousal related problems are more prominent which can lead to heart attacks, arrhythmias, and possible sudden death in those with preexisting heart conditions [29]. Minor effects on physiological conditions may include headaches, back pain, heartburn, stomach ache, elevated blood pressure, and rapid heartbeat [28]. Chronic stress has similar effects with acute stress. However, it causes more damage to physiological conditions. It is a significant risk factor for hypertension and coronary disease [29], [30], irritable bowel syndrome,

gastroesophageal reflux disease [31], generalized anxiety disorder, and depression [32]. These health issues also have a prominent effect on the economy such as absenteeism, staff turnover [33] and tardiness. These problems result in a decrease in productivity. Stress can also cause the "Presenteeism" problem, which can be defined as employees are present at their workplace, but they do not work at full capacity [19]. The yearly cost of absenteeism and presenteeism has been estimated at 272 billion Euros and the cost of productivity loss has been 242 billion Euros every year [23] in the whole of Europe. The long-term consequences of stress triggered a need for avoiding it when the symptoms first emerge. To prevent more damages, stress should be detected in the early stages. This has led to increasing interest in affective computing (AC) which makes use of technology to recognize the affective state of a person. Rosalind Picard of the Massachusetts Institute of Technology (MIT) published the first book on affective computing [34]. It has since become a prominent branch under the human-computer interaction area [35], [36]. It makes use of physiological and physical manifestations of an individual to detect the current emotion. Prominent elementary emotional states that can be inferred in affective computing are joy, anger, surprise, disgust, sadness, and fear [25]. Stress has been recently added to the abovementioned emotions that can be determined and smartphones, smartwatches and smart wristbands have become an integral part of our lives and have reached a widespread usage. This raised the question of whether we can detect and prevent stress with smartphones and wearable sensors. The damages of stress on human health have been known by the researchers and a significant amount of efforts have been made recently to develop an automatic stress measuring system by making use of these smart devices and advanced AC algorithms. Possible application areas of automatic stress detection systems can be listed as driver stress monitoring in automobile environments, passenger stress detection and alleviation in the commercial flights in airplanes especially for passengers with flying phobia, stress assessment of workers to augment efficiency in the workplaces (factories and offices), stress handling evaluation for applicants in job interviews and supporting psychologists in online therapy sessions by continuously monitoring stress level of patients.

In the early stress level recognition studies, data from the different physiologi-

cal modalities were collected in various experiments performed in laboratory environments. Most recent works achieved over 90% accuracy for detecting two levels of stress. However, researchers noticed that stress levels felt in the laboratory environments are different from the stress in the real-life situations [25]. Moreover, the ultimate aim of these studies is to help individuals manage their stress levels in their daily life routines. These systems should detect the stress and help them alleviate stress if it is higher than normal levels. The laboratory experiments gave researchers clues about how to build such systems. However, there are more issues to deal with when the research takes a step outside the laboratory e.g., unrestricted movement of subjects, unknown context, reliability issue of self-reports, battery life.

After laboratory environments, the research effort was directed towards detecting stress levels in restricted environments such as offices, automobiles and classrooms. Traffic jams in crowded cities, offices and workplaces and exams and courses in campus environments are among the primary causes of increased stress levels. These environments could be categorized into semi-restricted environments because the movements are limited and they can be controlled and monitored with sensors and cameras. There are a small number of studies in unlimited daily life stress level detection with smartphones and smart wearables. Due to the above-mentioned reasons, their accuracies are lower than those of restricted environments. Another research issue is to differentiate between stress and cognitive load to increase the resolution of stress detection systems and get more realistic results. However, most of the daily life stress studies only distinguish between stressed and relaxed states, which are not representative of different stress levels experienced in real-life. A high cognitive load state also frequently occurs, especially in work environments, and it should be discriminated from stress. Stress is different from cognitive load because it occurs when the individuals feel that they could not cope with the cognitive task load. However, if they can handle the cognitive load, stress may not occur. After detecting high levels of enduring stress, these systems should also decrease the stress of individuals to acceptable levels. When we examine the literature, there are a very few number of studies dealing with managing stress in the daily life.

1.1. Thesis Contributions

Our thesis addresses the following prominent research issues:

- The development of modality specific artifact detection and removal algorithms to cope with unrestricted real-life wrist movements.
- The implementation of an unobtrusive multi-level daily stress detection system.
- The effect of applying decision level smoothing and decision-making mechanisms on system performance.
- The performance evaluation of person specific, clustered according to the baseline stress levels (hybrid) and person-independent models.
- The effect of different ground truth surveys (NASA-TLX and a more suitable, less time consuming, version of it for everyday stress detection) on classification accuracies,
- Measuring the daily and session-based perceived stress levels by using only physiological signals and combining the contextual information (weather, physical activity level and activity type) with them
- Application of a guided mindfulness technique and measuring its success with smartwatch based physiological signals for reducing stress levels.
- The comparison of mobile mindfulness, traditional mindfulness and yoga methods in the context of stress management in a relieving stress
- Developing a prescreening tool for long-term perceived stress levels
- Application of James Gross's emotion regulation model [7] in the context of stress management and measuring the physiological component with smart bands.
- Application of LSTM (Long Short Term Memory) to enhance the performance of daily life perceived stress models
- Developing stress detection models in the laboratory to improve the performance of perceived stress level detection systems in the wild.

1.2. Thesis Outline

In this thesis, we briefly mention the stress induction methods, physiological signals that are used to detect stress, widely used devices for data collection and features that are used commonly for each signal, real-life data collection challenges, machine learning algorithms and stress alleviation methods are provided in Chapter 2. In Chapter 3, studies conducted in controlled laboratory environments, restricted and semi-restricted real-life environments: office, campus, car and non-restricted daily life conditions are discussed. We proposed an unobtrusive platform-independent daily life stress detection system that uses smartbands and smartwatches in Chapter 4. We tested our system in different types of environments and discussed a variety of techniques to improve the daily-life stress detection performance and alleviate high daily life stress levels in Chapters 5, 6 and 7. We then conclude with the insights and future works for the stress detection studies in Chapter 8.

2. BACKGROUND

2.1. Stress Signals

We can measure and observe stress symptoms in numerous ways. The Sympathetic Nervous System (SNS) ignites the stress reaction [37], resulting in psychological, physiological and behavioral symptoms [38]. The psychological way of measuring stress can be self-report questionnaires or being interviewed by a psychologist [19]. Therefore, automatic stress detection topics do not include this type of measurement.

The second way to detect stress is by evaluating physiological signals (see Figure 2.1 for different types of them). They include information related to the intensity and quality of the affect and experience of the subject [39]. Signals of interest include hormone levels, ElectroCardiogram (ECG), Electroencephalogram (EEG), Electro-Dermal Activity (EDA), Blood Pressure (BP), Skin Temperature (ST), Electromyogram (EMG), Respiration, Blood Volume Pulse (BVP), Pupil Diameter (PD), Eye Gaze and Blinking, Thermal Imaging (TI) and functional Magnetic Resonance Imaging (fMRI). The other two methods are investigating behavioral data and context information which are not investigated thoroughly in the literature [40].

Another way of recognizing stress is from the behavioral changes. Stress affects in behaviors of individuals. Without invasive methods and a need for extra pieces of equipment, we can measure the behavioral changes. Behavioral responses comprise of keystroke and mouse dynamics, posture, facial expressions, speech, mobile phone usage, walking pattern and text linguistics. The last property we can recognize stress from is the contextual information which consists of calendar events and location. In this section, we will discuss the most important signals for stress detection, widely used types of equipment for each signal and the most distinctive extracted features.

2.1.1. Physiological Signals

2.1.1.1. Heart Activity. One of the most prominent signals for discriminating stress comes from the heart activity because ANS influences the heart rate directly. The

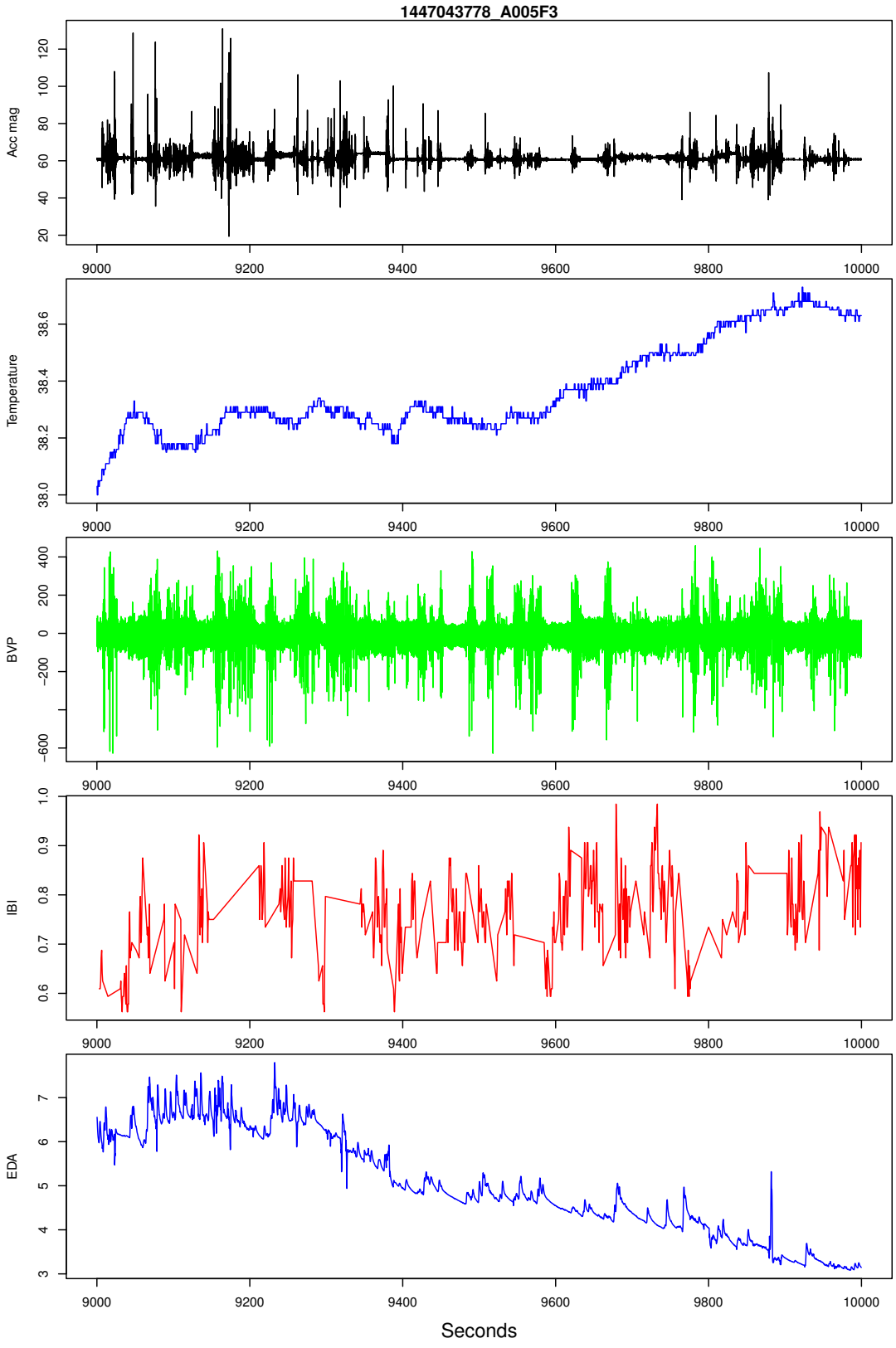

Figure 2.1: We collected different types of physiological signals from Empatica E4 (From Top to Bottom: Accelerometer, Skin Temperature, Blood Volume Pressure, InterBeat Intervals and Electrodermal Activity Data)

electrocardiogram (ECG) is employed to measure the electrical activity produced by the heart via electrodes placed on the body typically to the left arm, right arm and the left leg. A typical heartbeat has four fundamental elements which are baseline, P wave, QRS complex and T wave [41] (see Figure 2.2). The most distinctive R peak is employed for feature extraction. Heart activity can be modeled with heart rate (HR), RR intervals (IBI) and heart rate variability (HRV). Heart rate variability is the oscillation of the time between consecutive beats. IBI interval can be defined as the time between two consecutive R peaks. All of these can be inferred from R peaks. The wearable devices that are used commonly for ECG measurement are Biopacs MP150, MP35 and Shimmer Sensing 3 [41]. Features can be divided into three classes: time domain, frequency domain and non-linear features. Time domain features are the mean of RR intervals, the standard deviation of RR intervals, the root mean square of RR intervals, the percentage of the number of successive RR intervals varying more than 50 ms. Frequency domain features can be counted as Low-Frequency component, High-Frequency Component, LF/HF ratio. The most common non-linear features are entropy, complexity, Poincare Plots, recurrence and fluctuation slopes [19].

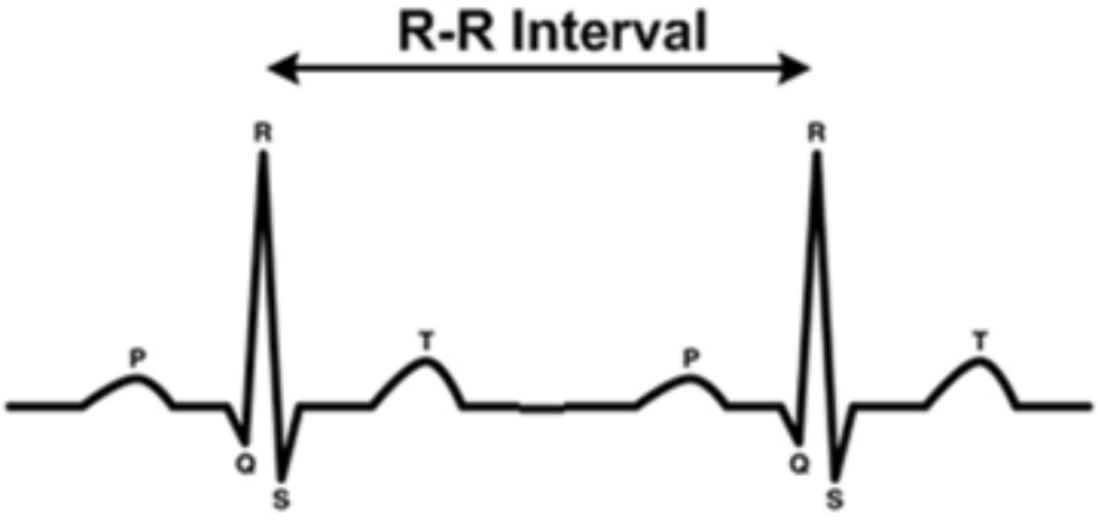

Figure 2.2: The structure of the ECG signal [1]. All P, Q, R, S points are shown.

2.1.1.2. Brain Activity. Brain activities are also affected by emotional changes and stress. The electroencephalogram (EEG) is used to measure the brain activity by placing a series of electrodes onto the scalp of the subject. BioSemi ActiveTwo, ABM B Alert X10, Emotiv EPOC+ are widely used gold standard types of equipment for measuring EEG. EEG signal is composed of four frequency bands: Alpha (8-13 Hz), Beta (13-30 Hz), Delta (0.1-4 Hz) and Theta (4-8 Hz) [19]. Alpha activities are a sign of a calm and balanced state of mind and a decrease in stressful states. Beta activity correlates with emotional and cognitive processes and increases with stress. Mean amplitudes of the EEG signal, mean amplitudes of Event-Related Potentials (ERP), alpha, beta, theta frequency bands, mean power ratios and fractal dimension features are used to detect stress.

2.1.1.3. Muscle Activity. The stress-related neural activity also affects muscles. Muscle action potentials are used to detect stress [41]. Electromyogram (EMG) measures muscle action potentials by placing electrodes on selected muscles. Facial and Trapezius muscles are regions of interest for measuring muscle activity. The mean, median, standard deviation, RMS, peak loads and gaps per minute are the features that are commonly used [19].

2.1.1.4. Electrodermal Activity (EDA). EDA, also known as Galvanic Skin Response (GSR), is the change of electrical properties of the skin. Under emotional arousal and stress, body sweats and skin conductance increases. EDA can be computed by applying a small current and measure the resistance of the skin between two placed electrodes. The EDA signal is composed of two components. The first one is the Skin Conductance Level (SCL). It represents slowly changing part over the long term (Tonic). The second part is the Skin Conductance Responses (SCR) which represents the faster and event-related part of the EDA (Phasic). To measure a slowly moving baseline and extract statistics related features such as mean, standard deviation and percentiles researchers are using the Tonic (SCL) part since this part is not contaminated with peaks which will affect the baseline calculations. For arousal event detection, the Phasic (SCR) part is employed. EDA is one of the best discriminative signals along with the heart rate signal.

ProComp Infiniti, Biopac MP150, Shimmer 3 GSR+ and Empatica E4 wristband are the instruments that are used widely to measure the EDA [41]. The mean amplitude, standard deviation, minimum and maximum values, RMS, the delay between applied stimuli and response, number of peaks, peak height, rising time, recovery time, the position of maximum and minimum features are tried in the literature to determine the stress of a user [19].

2.1.1.5. Blood Volume Pulse. When HRV changes with stress stimuli, blood volume and blood pressure also change. Blood volume pulse (BVP) is the change in the blood volume for each interbeat interval. Photoplethysmography (PPG) is a low-cost optical technique to measure BVP. It uses the absorption of light by blood. After the light is emitted from a light source, different amounts of blood in the volume will absorb a different amount of light. In this manner, blood volume can be measured. UFI model 1020, Empatica E3 and E4 wristbands, Angel Sensor, Biopac Bionomadix PPGED-R are the pieces of equipment that are commonly employed to measure BVP by using PPG. Although BVP features can be used directly, in general, they are used to extract the heart rate variability or IBI features.

2.1.1.6. Skin Temperature. Skin temperature can change due to various factors including stress. The research demonstrated that arousal could cause 0.1 or 0.2 Celsius temperature change [42]. The main reason for local temperature changes is that the blood flow is controlled by SNS. By controlling other factors, the effect of stress on skin temperature can be measured.The mean, minimum, maximum and standard deviation of skin temperature features were used in the literature to determine stress.

2.1.2. Behavioral Data

2.1.2.1. Speech. Stress causes changes in the human voice generation mechanism. Pitch, speaking rate, energy and spectral characteristics are affected by stressful events. Speech is preferred by many researchers because it is noninvasive and data collection is easy in the controlled quiet environments. Pitch (mean, standard deviation, range),

higher frequency bands ratio, speaking rate, voice intensity, smoothed energy, voiced-unvoiced speech ratio, Mel Frequency Cepstral Coefficients (MFCC) are the features that are used widely to detect stress levels [19].

2.1.2.2. Facial Expressions. Stress and emotional states have a correlation with facial expression. It has been demonstrated that facial expressions reflect emotions more than self-reports [43]. Facial EMG and image recognition from cameras were used to detect facial expressions in stressful situations. The mean smile intensity, eyebrow activity and mouth activity are the main facial features for detecting stress [19].

2.1.2.3. Keystroke and Mouse Dynamics. Another important behavioral data is keystroke and mouse usage dynamics. Every individual has different writing and mouse usage speed and style. Especially the speeds of writing and mouse usage depend much on individuals [44]. These two features define the keystroke and mouse dynamics of a user. When an individual is stressed, the muscles contract more than usual and it affects keystroke and mouse dynamics of an individual. Stress can be detected by relying on this change. Keystroke and mouse dynamics are noninvasive techniques and do not require additional equipment other than the mouse and the keyboard. The most important and measurable features that distinguish stress from keystroke dynamics are dwell and flight times, pause rate, the frequency of using specific keys such as backspace and space-bar, duration of digraphs and trigraphs and key pressure [19]. Important mouse dynamics features can be listed as the acceleration, average speed of the movement, frequency of the movement, stillness, frequency of the clicks.

2.1.2.4. Body Gestures and Movements. Individuals demonstrate behavior and gesture changes when feeling stressed. These changes include but are not limited to jaw clenching, arm movements, self-touching, finger rubbing. Posture change is another sign of stress. Stress in the sitting position is investigated by the change in the center of pressure [45]. Subjects revealed more posture changes in stressful situations.

2.1.2.5. Mobile Phone Usage. Stress also affects how the individual uses smartphones. Since collecting information is easy and non-invasive, smartphone usage behavior changes are investigated in the literature. Call logs, SMS logs, app usage, types of apps, battery usage, the screen on-off frequency, internet browsing and Bluetooth proximity were used to detect stress [19].

2.1.3. Questionnaires and Surveys

Psychological stress evaluations can be collected by asking subjects to fill questionnaires and by interviewing. There are two ways to collect data from subjects in daily life (or long laboratory experiments) which are instant reporting and day reporting (or cumulative reports). People have a tendency to forget emotional peaks in 24 hours [46]. For this reason, asking questions to subjects at the end of the day can cause incorrect measurements. The alternative is to ask the subject to report stressful events instantly. The problem in this manner is that people can forget to report events. In the literature, a combination of these methods is used to increase the accuracy of reports. Perceived Stress Scale (PSS), Stress Self-Rating Scale (SSRS), NASA-TLX, Self Assessment Manikin and Positive and Negative Affect Schedule (PANAS) are the questionnaires and interviews that are commonly used in the laboratory and daily life stress experiments.

2.2. Stress Induction Tests

In order to investigate stress in controlled laboratory environments, subjects are asked to perform certain tasks in order to induce stress. In this section, the most prominent tests applied in the laboratory conditions will be briefly described.

2.2.1. Trier Social Stress Test (TSST)

The most common stress induction test is the TSST. The procedure of TSST is as follows. The total period of the test is fifteen minutes [47]. It is divided into three phases. During the test, saliva and blood are collected and the heart rate data is

Figure 2.3: SCWT Test Example [2]

recorded. The test takes place in a room with a video camera, microphone and three judges. Judges are prepared and trained to control their expressions as neutral. In the first phase, the subject is asked to prepare a presentation that has five-minute duration. The subject has a paper and pen to prepare the presentation. After the first phase, the paper is suddenly pulled back from the participant. In most versions of TSST, the subject is told that the presentation is for a job interview. In the second phase, the user is asked to present the prepared material. Judges listen to the presentation without any intervention. If the participant finishes the presentation before five minutes, they ask him or her to continue. The last phase is a mental arithmetic task in front of the judges. The subject is asked to count backwards from 1022 in steps of 13 [47]. If the participant makes a mistake, he/she is asked to begin from the start. After the last five-minute mental arithmetic task, the recovery period starts. The collection of samples and signals continues until the end of the recovery period. At the end of the test, the user is briefed about the aim of the test.

2.2.2. Stroop Color-Word Inference Test

The Stroop Color-Word Test (SCWT; Stroop, 1935) is one of the oldest and most widely applied stress induction test [48]. By changing the number of subtasks, type and number of stimuli, times for the task or scoring procedures, different variations of the test were created [49]. In the standard version of the test, a subject is asked to read the name of the color words. This subtask is called ”word reading”. The second subtask is naming the color of ink. The last subtask is to name the color of the word

that is come up at the screen. An example of the SCWT test is shown in Figure 2.3.

2.2.3. Montreal Imaging Stress Task

The Montreal Imaging Stress Task (MIST) is a variation of Trier Test [50]. It includes computerized mental arithmetic challenges and social evaluative threat elements with the help of a computer program. It is applied by an investigator. In order to be able to differentiate the effects of stress and mental arithmetic, it has three phases (rest, control and experimental). In the original test version, the subject is observed via functional magnetic resonance imaging (fMRI) or positron emission tomography (PET) in these phases. In the resting phase, the subject looks at the computer display. In this phase, subjects are not asked to perform any tasks. In the control phase, a series of mental arithmetic tasks are shown to the participant and participants answer the tasks by using the provided computer program. In the last phase which is the experimental phase, the difficulty of the tasks is increased and the time limit of the tasks is decreased. The aim is to increase the task difficulty over the participant's mental abilities. To increase the stress of the participant, the average performance of participants and estimated completion time is displayed to the subject. After the end of each task, the performance evaluation is displayed to increase the stress of the participant further.

2.2.4. Cold Pressor Test

The cold pressor test is another cardiovascular test applied by putting the participant's hand into an ice water container for a minute [51]. It can also be used to induce stress. Blood pressure and heart rate are monitored. The procedure of the test is as follows: The subject of the test is asked to put his/her hand in the cold pressor as long as they can. They are told to express if they feel any pain. If the pain becomes unbearable, they can remove their hands. This helps researchers to measure the threshold (first feeling pain) and tolerance (total time minus threshold). This version of the cold pressor task is the most widely applied one.

2.2.5. Sing-a-Song Stress Test (SSST)

Brouwer et al. [52] developed a new method for inducing stress which is named SSST. The subjects are shown neutral messages at the screen for one minute. After this neutral phase, a final message that is asking the subjects to sing aloud is shown. The participants must stand still during singing. They remarked that SSST is a quick, easy, controlled and potent way to induce mental stress [52]. They measured the ECG and EDA signals during both the neutral and singing phases and they demonstrated that two phases are considerably different and found a correlation between the stress level and singing.

2.2.6. International Affective Picture System (IAPS) Test

The IAPS is an image database created for providing a standardized set of images for emotion and attention research [53]. It has been commonly applied in psychological studies [54]. The IAPS was created by the National Institute of Mental Health Center for Emotion and Attention at the University of Florida [55]. It consists of pictures varying from simple household objects to extreme pictures which cause arousal on individuals (mutilated corpses, erotic and violent scenes). Stress induction by showing a series of pictures from the IAPS is another test that is widely applied in the literature.

2.3. Data Collection Challenges

In the laboratory environments, data collection procedures are less prone to error when compared to daily life. In the daily life, errors from incorrect placement or detached equipment can occur. Corrupted physiological signals due to body movements are very common [56]. Multi-sensor, multi-device measurements, invasive devices and storage and processing of the massive amount of data can be counted as the additional challenges in the daily life environments.

2.3.1. Movement and Improper Placement Problems

High-quality data must be accurate, complete, relevant, timely, sufficiently detailed, appropriately represented and must retain sufficient contextual information to support the decision making [57]. Wearable physiological measurement devices mentioned above satisfy most of the listed conditions and provide high-quality data. The sampling frequency should be sufficient to represent the signal well. Proper placement of the sensors should be done to avoid ambiguities and obey standards for the correct measurement of physiological data. Signals are affected by the device noise, random noise as well as by loose device-skin connections and body movements [19]. Noise in the signal can be filtered by Kalman filters, Butterworth low-pass filters, Median filters, Wiener filters, Wavelet Decomposition, etc. [19]. Artifact removal can be done via least mean squares, regression analysis, independent component (ICA) and principal component analysis (PCA) [19].

2.3.2. Data Fusion from Variety of Sensors

In most of the experiments, data is retrieved from different sensors and devices. The integration of the data has challenges. Determination of the point of data integration, i.e., before the final decision or during processing, is the first challenge. Synchronization of the data must be achieved via timestamps for the calculations to be accurate. The last challenge is the missing data from one source or more sources. Some precautions must be taken, such as using probabilistic data fusion algorithms, imputation or removal of the missing time intervals.

2.3.3. Big Data Problem

Continuous data collection accumulates a massive amount of data to be processed and stored. Onboard signal processing algorithms could be a solution to the big-data problem. Nevertheless, this increases the power consumption, decreases the battery life of the devices, reduces storage requirements, but increases the algorithm complexity [19]. Researchers need to cope with this trade-off according to their experimental

conditions.

2.3.4. Selection of Unobtrusive Devices

Instruments to measure stress should be unobtrusive and non-invasive to collect data accurately. Invasive devices can create additional stress on subjects when they are worn. Recent technologies have provided us non-invasive and completely transparent devices to the user for monitoring stress. Smart wearable systems and smartphones are commonly used to collect significant amounts of data unobtrusively and sometimes without the user being aware of it.

2.3.5. Battery Life

Charging is also a significant issue in real-life data collection. When the life of the battery becomes three or four hours (as in the case of Samsung Galaxy Gear S1, S2, S3 when all of the sensors are active), devices must be charged several times for collecting one day-long continuous data. Besides being inconvenient, the need for recharging may lead to data gaps between the times when devices are collected for recharging and handed out to the participants with the full battery again. The battery issue directs researchers towards new challenges such as designing batteries with longer lives on wearable devices, increasing the life of the battery by reducing the power consumption (i.e., disabling some sensors, decreasing the duty cycle of sensors, cutting down the brightness levels).

2.3.6. Ground Truth Collection

One of the most important differences between the research that is conducted in the laboratory and in daily life is the need for collecting the conditions of the participants. In the laboratory experiments, the condition of a subject is always known because the experiment steps are designed with the timestamps beforehand. However, in the daily life data collection, the stress condition of a subject is unknown a priori and the only way to learn this ground truth is to ask the participants. For this purpose,

some known surveys (Perceived Stress Scale (PSS), Stress Self-Rating Scale (SSRS), NASA-TLX, STAI, Self Assessment Manikin and Positive and Negative Affect Schedule (PANAS) questionnaires) are employed during a day with the determined time periods. The surveys must be distributed to all of the participants and collected from all of the participants for each session of the daily life experiment periodically. Another alternative might be using a mobile survey app developed for this purpose. The ground truth collection from the participants imposes some new challenges on researchers. The scores are subjective and change from person to person. They may not represent the true stress status of a person due to the difference between perceived and physiological stress as mentioned in Chapter 1.

2.4. Stress Alleviation Methods: Mobile Apps and Techniques

Another important question arising after detecting stress is how to relax an individual to the normal state i.e., emotion regulation. Some ancient techniques can be employed to alleviate stress, such as yoga, meditation. However, these techniques either require outdoor environments and they could not be applied without interrupting daily activities or in office environments. We investigated mobile apps and techniques along with traditional techniques such as yoga and mindfulness to alleviate stress and decrease their impact without leaving the office environments. The effect of the indoor techniques is not thoroughly investigated, but a limited number of studies have started such as [58]

2.4.1. Yoga

Yoga is an ancient Eastern practice that developed more than 2000 years ago. Although its original creator and source are uncertain, the earliest written word 'Yoga Sutra' describes the philosophy of yoga focussing on growing spirituality, regulating emotions and thoughts. Initially, the focus was on awareness of breathing and breathing exercises 'pranayama' to calm the mind and body, ultimately to reach a higher state of consciousness.

As yoga evolved, physical movement in the form of postures was included and integrated with yogic breathing 'prana' and elements of relaxation. The underlying purpose to create physical flexibility, reduce pain and unpleasant stimuli and reduce negative thoughts and emotions to calm the mind and body, thereby improving wellbeing. In the healthcare literature, the benefits are reported to be far-reaching both for mental and physical health conditions such as anxiety, depression, cardiovascular disease, cancer and respiratory symptoms. It is also reported to reduce muscular-skeletal problems and physical symptoms through increasing the awareness of the physical body.

Yoga has become a global phenomenon and is widely practiced in many different forms. Generally, all types of yoga include some elements of relaxation. Additionally, some forms include mainly pranayama and others are more physical in nature. One such practice is vinyasa flow which involves using the inhale and exhale of the breathing pattern to move through a variety of yoga postures; this leads to the movement becoming meditative. The practice often includes pranayama followed by standing postures linked together with a movement called vinyasa, (similar to a sun salutation) which helps to keep the body moving and increases fitness, flexibility and helps maintain linkage with the breath. The practice also often includes a range of seated postures, an inversion (such as headstand or shoulderstand) and final relaxation 'savasana'.

2.4.2. Mindfulness

Mindfulness involves being more present at the moment by acknowledging the here and now, often referred to as 'being present' rather than focussing on the past or future [59]. Being present may include being aware of our surroundings and the environment, or of what we are eating and drinking and physical sensations such as the sun or wind on our skin.

Acknowledging the thoughts and body are also aspects of mindfulness. Each day humans experience thousands of thoughts the majority being of no consequence. In some instances, these thoughts are repetitive and negative in nature which can lead to

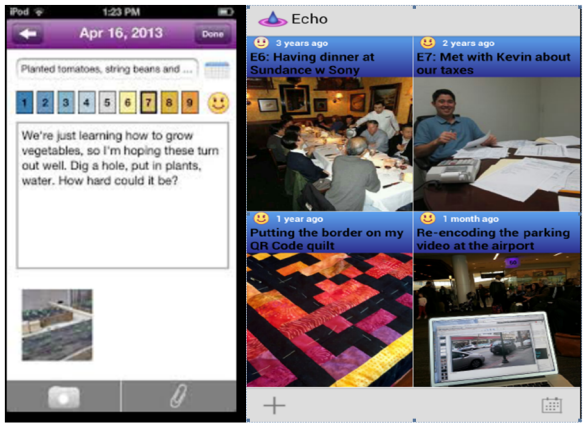

Figure 2.4: Echo App for Reflections: Screenshots [3]

increased stress and the related unpleasant physical symptoms such as feeling anxious, nausea and tension headaches. Being mindful includes an awareness of our thinking and whether we are caught up with our thoughts rather than being aware of the moment. Additionally, on a daily basis, awareness of the physical body may be minimal; being mindful includes increasing this awareness through becoming more connected with the sensations in the body. This might include experiencing the legs moving when walking, or feeling the ground under the feet or the natural way of the body whilst standing.

Mindfulness has been shown to be of benefit to physical and mental health. It is currently recommended by the National Institute for Clinical Excellence [60] as adjunctive therapy to Cognitive Behavioural Therapy (CBT) for the prevention of relapse depression.

However, it may be challenging for some individuals to do this with a multitude of distractions around them, and therefore they may choose to identify a particular time and place when and where they can sit in a comfortable position to start to become aware of their breathing and bodily sensations.

2.4.3. Echo and Emotical Apps

Isaacs et al. [61] developed two mobile apps to cope with stress. The first one is based on the user's reflection on daily events. Users are asked to write reflections to

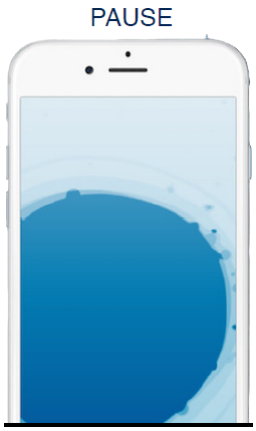

Figure 2.5: You move your finger with the circle. If you move "mindfully", the music continues and the circle grows. Otherwise, feedback stops [4]

daily events. After a while, both negative and positive reflections are showed to the user again and the user is requested to write another reflection on this event after some time. The reflection of negative and positive past both have their advantages. Negative past reflection reduced medical visits, improved immune response, better grades, reemployment, reduced absenteeism, increased subjective well-being and increased working memory [61]. Positive past reflection increases the enjoyment of life, subjective happiness and it has a positive effect (often invoked as a response to loneliness) [61]. It causes escapism which is "A vacation from the present". They evaluated these and developed a reflection mobile app "Echo". The screenshots of the app are available in Figure 2.4.

The second application from this research group is the Emotical app which aims to understand and change the mood of a person. Hollis et al. [62] listed the challenges of emotion regulation as people do not have much insight into their emotions, poor at predicting future emotions, poor at regulating mood by actions. They defined the problem and make complex inferences about what affects mood and then propose remedial actions. After the application detects high stress, it suggests some activities based on the user's information about what makes her/him happy in the beginning. It also clusters people and suggests the joyful activities of people in the same cluster. This feature provides discoveries for users to learn what makes them happy.

2.4.4. Pause and Sway iPhone Application – Tai-Chi on Screen

Cheng et al. [58] indicated that humans have 65000 thoughts on an average day. Eighty percent of these are not positive and 95% of these thoughts are from the previous day. Their starting idea was paying attention at will increases the level of happiness [58]. They developed a mobile app inspired by an ancient Chinese martial art Tai Chi. They focused on the movement based medication. Pause app does not require any special equipment besides a mobile phone and it could be used in busy, noisy daily situations [58]. Pause provides mindful touch on smartphones and helps individuals relax by giving calm feedbacks. They stated that interactive meditation works better than traditional guided meditation in noisy daily life situations, meanwhile, in quiet environments, Pause has similar results with traditional techniques. A screenshot can be seen in Figure 2.5:

2.4.5. HeartMath: Increase your 'Coherence'

HeartMath is a stress alleviation mobile application. Developers tried to increase inner balance which helps individuals to prevent, manage and reverse the harmful effects of stress. They give users some goals every day to achieve. Users breathe deeply and think of some positive memories. They are required to plug a HeartMath sensor to their phones and start the relaxation session. They will breathe along with the pacer. While breathing, they focus on a moment when they felt joy, appreciation or care by closing their eyes. By real-time coaching, it shows them how coherent they are. Coherence is a formula that HeartMath created. If a user's heart signal is more like a sinusoidal, they are more coherent. They can increase their coherence and decrease their stress levels by the following five minutes of exercises. They can track their progress to keep the motivation up. Screenshots are given in Figure 2.6.

2.4.6. Other Relaxation Techniques Used in the Literature

There are other apps in the literature. The Spire device measures breaths per minute, evaluates and if an individual is stressed, offers some respiration exercises

Figure 2.6: Heartmath Screenshots : Real-Time Feedback [5]

[63]. Similarly, Tinke app evaluates a novel metric called the Zen index [63]. If a user is stressed, the Zen index will be higher. Tinke then suggests deep breathing exercises. Another app is WellBe. WellBe advises some breathing and meditation exercises that utilize sitting in silence and some relaxing voices [63]. Akmandor et al. [64] used classical music, warm stone and good news to relax a stressed individual. Chen et al. [65] first record the respiratory pattern of participants. To relax the stressed participant, they suggest a yoga respiratory pattern which is most similar to the stressed user's respiratory pattern.

2.4.7. Insights from Stress Alleviation Apps and techniques

An ideal stress alleviation method should be applicable indoor, scientifically proven and not require any extra instrument. Pause app is the closest one to the ideal stress alleviation technique definition. HeartMath is also a significant alleviation method. The disadvantage of HeartMath is the need for an extra equipment. Echo app is more suitable for dealing with chronic stress. Yoga is another effective stress reducing technique, but it is hard to apply in daily life routines and in workplaces. Classical music, good news, warm stones may affect some people and not affect others.

2.5. Machine Learning Classification Algorithms

After features are extracted from the data, machine learning (ML) algorithms are used to classify stress levels of individuals. In this section, the most widely used traditional ML algorithms used in our classification work are presented.

2.5.1. Traditional Machine Learning Methods

2.5.1.1. Support Vector Machines (SVM). SVM creates decision planes that define decision boundaries. A decision plane can be defined as the plane that divides objects belonging to different classes. In some classification tasks, complex decision structures are needed to separate these objects into their classes correctly. Support Vector Machines are designed to cope with this kind of tasks. SVM rearranges objects using kernels which are a set of mathematical functions [66]. The objects are mapped or transformed so that they can be easily separated by less complex planes.

2.5.1.2. k Nearest Neighbors (kNN). This method relies on memory for saving instances with known outputs. Labels of these instances are known. When a new test instance is to be decided, the output of the closest known objects is examined. The majority of votes rule is applied. The output that has emerged on the majority among these neighbors is assigned to the test instance. The distance formula (Euclidean, Mahalanobis, etc.) and the number of closest objects that are being evaluated (k number) are important parameters for the kNN algorithm.

2.5.1.3. Decision Tree / Random Forest. Decision trees are ML tools which are used for regression or classification of both continuous and discrete variables [66]. The structure of this ML model inspired its name (See Figure 2.7). The decision tree mechanism is as follows: For each iteration, local regions are created recursively. It is a supervised and hierarchical model [67]. The decision tree comprises of decision nodes and leaves. Each decision divides the data. Low entropy divisions are created in this manner. Appropriate sized tree generation requires expert knowledge [66]. A Random

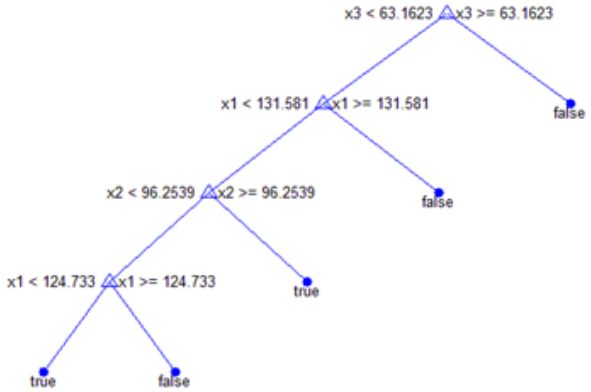

Figure 2.7: An example decision tree

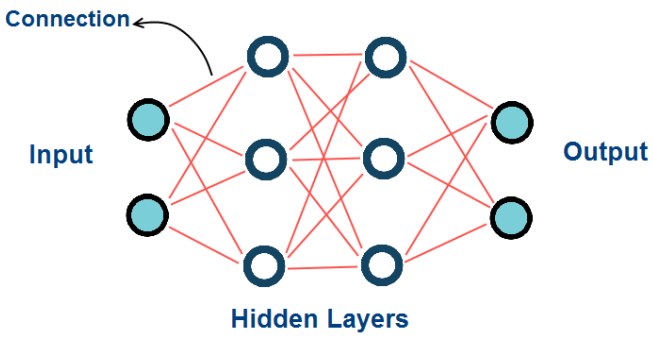

Figure 2.8: ANN Model Structure

Forest is a variation of A Decision Tree that uses multiple trees instead of a single one.

2.5.2. Artificial Neural Networks (ANN)

Nielsen defines a neural network as "A computing system made up of a number of simple, highly interconnected processing elements, which process information by its dynamic state response to external inputs" in [68]. ANNs try to mimic biological neuron structures in the brain. They have an input layer, output layer and hidden layers (See Figure 2.8). Each layer has connections and weights of these connections. Iterations of neural networks are called epochs. For each new input pattern, the weights are updated by evaluating feedbacks [66].

2.5.2.1. Multilayer Perceptron. The Multilayer Perceptron (MLP) is a feedforward artificial neural network [69]. It has a minimum of three layers which are the input layer, hidden layer(s) and the output layer. The hidden layer (s) uses the activation functions to capture nonlinear data relations. Therefore, MLPs can discriminate between classes that are nonlinearly separated [69]. We selected them as a representative of a shallow neural network to compare with the used sequential deep learning method. Unipolar Sigmoid Function was used as an activation function in hidden volumes of MLP. We used a plain MLP with only one hidden layer to compare a shallow neural network with the deep sequential one.

2.5.2.2. Long Short Term Memory (LSTM). Long Short-Term Memory (LSTM) Network is a particular type of Recurrent Neural Network (RNN). It was defined by Hochreiter and Schmidhuber in 1997 [70] to alleviate the long-term dependency problem of RNN [71]. Because of the vanishing/exploding gradient problem as a result of back-propagation through time in a standard RNN, a long sequential data could be hard to learn [71]. In LSTM, instead of an RNN cell, a gated LSTM cell is used to cope with this problem. This makes LSTM suitable for sequential data classification or regression problems. A basic structure of an LSTM cell is demonstrated in Figure 2.9. The decision of whether the information should be remembered or not is controlled by gates. The previous data is saved via LSTM cells.

There are three types of gates LSTM uses to decide whether to add or remove the information in the cell state(C_t). Forget gate decides to throw away information by using a sigmoid layer. The input gate is the second gate which selects the values to be updated. It uses a sigmoid layer when making a decision and tanh layer for creating an updated value vector (see Figure 2.9). The cell state is updated with the reevaluated output of the input gate. The decision of which parts of the cell state is selected as the final output is calculated on the updated cell state and by using a sigmoid layer, as seen in Figure 2.9.

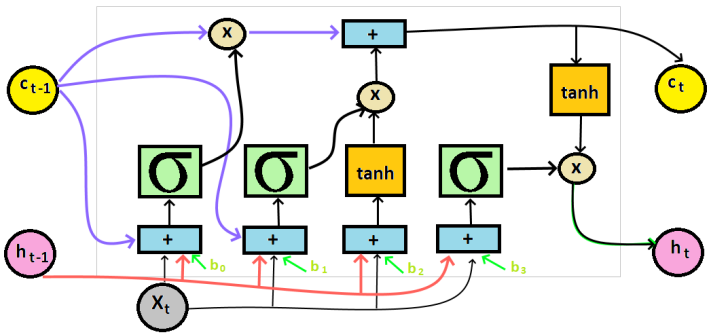

Figure 2.9: The block diagram of a LSTM cell. σ is sigmoid activation function which maps numbers to a range between 0 and 1. Tanh is a hyperbolic tangent activation function which maps numbers to a range between -1 and 1. C_t is the memory of block t, H_t is the output of block t, input data is the x, past output (block t-1) is $H_t - 1$, past memory is $C_t - 1$ and bias vectors are represented as b_0, b_1, b_2, b_3 symbols.

3. LITERATURE REVIEW

In this section, we provide a literature review of the stress detection studies. Firstly, the controlled laboratory environment where the research has first started is investigated. This chapter is divided into subsections by taking the used physiological signal into account: heart activity, electrodermal activity, brain activity, speech data, camera-based studies and multi-modal measurement research. The works that use stress alleviation methods after detecting stress are presented at the end of the chapter.

We classify outside the laboratory environments as restricted, semi-restricted and unrestricted. Automobile and office are restricted environments since the movement is limited and environments can be controlled with sensors and cameras. University campuses are semi-restricted environments. Some parts of campus environments are more constrained (i.e. classrooms, offices) and some parts are less constrained. This section is examined as a transition between restricted and unrestricted environments. Lastly, we present the literature in unrestricted daily life. We provide an insight subsection for each environment which includes the most successful machine learning (ML) algorithms, discriminative physiological signals and features. We also highlight the outstanding studies for each environment.

3.1. Controlled Laboratory Environment: Early Works

Stress detection research in the laboratory environments are the preliminary works that give researchers ideas about which sensors, features and machine learning algorithms to use in daily life studies. In most of the laboratory studies, ground truth collection methods are not required because the cognitive load and relaxation activity of the participants are known a priori. The prominent studies took place in controlled laboratory environments are provided in Table 3.1. These experiments can be counted as the preliminary works for the stress detection research in the daily life. When we examine Table 3.1, we can deduce that the EDA and HRV combination yields the best results in the laboratory environments. Four out of five works, with the

Table 3.1: Stress Detection Experiments in Controlled Laboratory Environments

Article	Stress Signal	Stress Test	Method	# of Classes	Accuracy %
[72](2012)	HRV	Stress in the traffic	Minimum Distance Classifier	3 (Low, Medium, High Stress)	90
[73] (2018)	HRV	SCWT	SVM	2 (Rest (R), Stress(S))	70
[74](2010)	EDA - PPG	Hyperventilation and Talk Prep.	kNN - FDA	2 (S,R)	95
[75](2010)	EDA	MIST	LDA - SVM-RBF - NCC	2 (Stress, Cognitive Load)	82.8
[76](2011) (Extension of [74])	EDA - PPG	Hyperventilation and Talk Prep.	Fuzzy Logic	2 (S, R)	99
[77](2015)	EDA - PPG	TSST	SVM	2 (S,R)	80
[78](2013)	Speech	TSST	SVM	2 (S,R)	72
[79](2011)	ECG- EMG - EDA	Arithmetic, Puzzle, Memory Tasks	Bayes, kNN, Fisher Least Square	2 (S, R)	80
[80](2016)	PPG - EDA - Respiration - Thermal Cam	Lie Detection	DecisionTree	2 (S, R)	73
[81](2016)	EEG	Arithmetic Task	SVM	4 (Neutral, Medium, Low, High Stress)	89
[82](2015)	Body Movements	Arithmetic Task	SVM	2 (S, R)	77
[83](2016) (Extension of [82])	Body Movements - EMG - EDA - Respiration	Arithmetic Task	SVM	2 (Stress, Relax)	85
[84] (2017)	Facial Cues	Social Exposure and Stressful Media (IAPS)	kNN - SVM - Naive Bayes	3 (Neutral, Relax, Stressed)	91.68
[85](2011)	Pupil Diameter	SCWT	FuzzySVM	2 (S, R)	70
[86](2014)	Pupil Diameter	IAPS	DecisionTree	2 (Stress, Relaxed)	90
[87](2017)	EDA - PPG - Speech - Accelerometer	TSST	Adaboost	2 (Stress, Relax)	94
[88](2015)	EDA - Accelerometer - Bluetooth	-	Logistic Regression	2 (Stressed, Unstressed)	91
[89] (2017)	HRV - EDA	Puzzles	F-state Machine	3 (Low, High stress, High alert)	0.984(f-measure)
[90](2012)	Temperature - Heat Flux - EDA- Respiration - Accelerometer	Arithmetic Task, Cold Pressor and Noisy Sounds	Naive Bayes	2 (Stress, Relaxed)	82
[64] (2017)	ECG - GSR - respiration - Blood Pressure - Blood Oximeter	Ice test and IAPS	SVM - kNN	2 (Stressed, Relax)	95.8
[91](2014)	Added unsupervised activity clustering to [90]	Arithmetic Task, Cold Pressor and Noisy Sounds	Naive Bayes - Logistic Regression	2	65
[92], [93] (2017) (Extension of [92])	Respiration	SCWT - Mental Computation	CNN	2 (Stressed, Relax)	84.59
[94](2014)	EEG - ECG - EMG - EOG	Mental and Memory Task	ANN	3 (Relaxed, Mental, Fatigue Stress)	80
[95](2015)	Facial Blood Flow	SCWT	Multiple Regression	2 (S, R)	88.6
[96](2014)	Facial Blood Flow	TSST	Binary Classifier	2 (S, R)	88
[12](2015)	EDA	Fail Scenarios	LDA	2 (S, R)	98.88
[97](2016)	Human Gaze - Mouse Click	Arithmetic Task	Random Forest	2 (S, R)	66
[98](2017)	EDA - PPG	Ice Bucket, SCWT, Mental Arithmetic Task, Singing	SVM, Logistic Regression, Random Forest	2 (S, R)	0.79(f-measure)

accuracies higher than 95%, use this combination as the physiological signals. LDA, SVM, kNN and Fuzzy Logic classifiers have achieved the highest accuracies with these signals (SVM was applied in [77] which is a more recent paper). In [81], researchers achieved 89% accuracy in 4-class stress classification by using EEG signals. However, for daily life stress detection, EEG measuring devices are obtrusive for individuals and they would not be comfortable wearing them in their daily life routines. Almost all of the studies in Table 4, employed a 2-class classification. To get more realistic results, stress detection resolution must be increased (i.e., the number of stress levels should be increased). The laboratory experiments further demonstrated the importance of using multimodality for increasing the detection accuracies as in the case of EDA - HRV combination.

3.2. Restricted Environments

3.2.1. Office Environment

One of the first stress detection works in the office environments was conducted by Hernandez et al. [99]. They investigated stress levels of the call center employees. They used skin conductance signals from an Affectiva QTM sensor for this purpose. They also evaluated the self-reports and call logs to detect stressful and non-stressful calls. For one week, nine employees answered 1500 calls. They altered the loss function of SVM to react better to changing priors when classifying calls. They achieved 78% classification accuracy for individuals and 73% for the general model.

Arnrich et al. [45] investigated the relationship between the posture changes

Table 3.2: Stress Detection Experiments in Office Environments

Article	Stress Signal	Method	# of Classes	Accuracy %
[99](2011)	EDA	SVM	2 (Stressed, Relaxed)	78
[45](2010)	Posture Changes	SOM - XY-fused Kohonen Network	2 (Stress, Cognitive Load)	73
[100](2016)	Accelerometer	Naive Bayes - Decision-Tree	3 (Low, medium, high stress)	60
[101](2013)	ECG	LDA	3 (Low, Medium, High load)	85

and stress levels. The Montreal Imaging Stress Task was applied to induce stress.

They investigated stress by comparing with a control condition where subjects have to perform the same task without time pressure and social stress. They calculated the Center of Pressure (CoP) from 1024 pressure sensor elements. To discriminate between cognitive load and stress, Self Organizing Maps (SOM) and XY-fused Kohonen Network were used. They observed that the variance in the sidewards is increasing in the stress condition. They achieved 73% classification accuracy when discriminating between stress and cognitive load.

Garcia-Ceja et al. [100] detected stress using only the accelerometer sensor of a smartphone. The advantages of the accelerometers are low privacy concerns, low power consumption, widespread usage in electronic devices such as smartphones, smart bands, fitness trackers. The Oldenburg burnout inventory (OLBI) questionnaire was used to collect reported stress data from a subject three times a day. They extracted 34 features in both the time and frequency domains. Naive Bayes and decision tree classifiers are used. They measured their classification accuracies in three different ways. The first one is user-specific results which have the highest accuracies, but they need more labeled data. The second model is the general model which has the lowest accuracy as expected. The novelty in this paper is the development of the third model which is the similar user model. Two users are selected similar by using some of their data and k-means clustering. Then, the model was applied to these users. The accuracy results (60%) are between the user-specific and the general model values. As future work, instead of collecting data all the time, only phone handling times can be used for data collection.

Kocielnik et al. [102] developed a stress detection scheme for a work environment. A DTI-2 sensor wristband was used to collect the EDA data. They further collected calendar information and the Self-Assessment Manikin Questionnaire user feedback. They preprocessed the data by removing the signal from the first 15 minutes and last 10 seconds, moments when the band lost contact with the skin and smoothing the signal. They did not provide classification accuracy, but they stated that the data is meaningful and provides new information.

Cinaz et al. [101] investigated workload levels in an office scenario. They used subjective ratings from NASA-TLX. They also had an objective reference from collected

salivary cortisol samples. Zephyr Bioharness was employed to gather ECG data. They had a calibration phase where the subject's responses were measured. They defined three tasks with the low, medium and high workload for the calibration. After the calibration phase, subjects are asked to perform their office works for one hour. Time and frequency domain features were used. They measured the mental load of office work and classified as low, medium or high. The LDA was the best performing classifier. They correctly classified mental workloads of 6 out of 7 subjects when subjective ratings were taken into consideration.

3.2.1.1. Insights from the Experiments in Office Environments. Office environments worked as a bridge in the transition from stress detection in the laboratory environments to the real-life stress detection research. The movement of individuals is limited in the offices and the environment can be controlled with cameras and sensors. Having said that, offices and other workplaces are among places that contribute to stress levels of individuals. When the research in office environments is investigated (Table 3.6), we can see that works that employed EDA and HRV have the highest accuracies. SVM and LDA were the best-performing classifiers.

3.2.2. Automobile Environment

Stress in automobile environments has been another interesting topic for researchers (See Table 3.7). Lee et al. [103] developed a stress detection system for drivers by using a wearable glove that has a photoplethysmogram (PPG) and inertial motion unit (IMU) orientation sensor which has an accelerometer, gyroscope and magnetometer on it. They gathered data from twenty-eight subjects and used the Euro Truck driving simulator with a throttle pedal, brake pedal and steering wheel. A driving behavior survey (DBS) was filled by subjects. They rated anxious driving behavior using Likert type scale 1-7. They measured cardiac rhythms (PRV) using PPG. For both sensors, time and frequency domain features were extracted and an SVM classifier was applied. The best result for 2-class classification was 95% using SFS-SVM with an RBF kernel function (Table 3.7).

Lee et al. [104] proposed a driver stress detection system using only an inertial motion unit. Their experimental setup was the same as [103]. As the stress level references, they used GSR, self-report surveys and facial expressions. Eight subjects participated in the experiments. They extracted 46 features and reduced to 22 with an insignificant loss in the classification accuracy. By using SVM classifiers, they achieved approximately 94% accuracy for discriminating low stress from high stress. As future work, they plan to investigate which conditions cause an increase in the driver's stress.

Chen et al. [105] detected driver stress by evaluating ECG, EDA and respiration data. Time, frequency and wavelet domain features were extracted. PCA and Sparse Bayesian Learning (SBL) machine learning algorithms were applied. They achieved 99% accuracy on a three-class classification.

Munla et al. [106] proposed a driver stress detection system. They used the

Table 3.3: Stress Detection Experiments in Automobile Environments

Article	Stress Signal	Method	# of Classes	Accuracy %
[103](2017)	PPG - IMU	SVM-RBF	2 (Stressed, Relax)	95
[104](2016)	IMU	SVM	2 (Stressed, Relax)	94
[106](2015)	ECG	SVM-RBF - kNN - RBF	2 (Stressed, Relax)	83
[107](2015)	EDA - EMG - ECG	kNN - SVM	3 (Low, medium, High stress)	98
[108](2015)	ECG	DecisionTree	3 (Low, medium, High stress)	88
[105] (2017)	ECG - EDA - Respiration	SVM-ELM	3	99
[109](2013)	HRV	kNN	2 (Stressed, Relax)	97

Automobile Driver database (DRIVEDB) [110], including ECG signals collected from 16 different individuals driving around Boston, Massachusetts. From the data, they extracted time domain, frequency domain, nonlinear and time-frequency domain (such as wavelet and STFT) features. As classifiers, they used SVM-RBF, kNN and RBF. They achieved 83% classification accuracy when discriminating between stress and no stress conditions. They plan to extend their work into distinguishing more stress levels.

Ghaderi et al. [107] presented a stress detection scheme for automobile drivers. They used DRIVEDB from the PHYSIONET database created by Jennifer Healey and Rosalind Picard [111]. The data were recorded from drivers in Boston. They used GSR, EMG, ECG signals. They selected reported the best features from the literature for these signals. For 100 second intervals, they classified the driver stress as low, medium and high. By using kNN and SVM, they achieved 98% classification accuracy over

this database. They stated that respiration is the most discriminative signal to detect stress.

Another research based on the DRIVEDB database was carried out by Keshan et al. [108]. They only used ECG data from the database. They stated that their difference was to develop personalized stress analysis consisting of three levels of stress: low, medium and high. ECG has become less invasive with wearable patches and demonstrates a high correlation with stress. They used Naive Bayes, Logistic Regression, Multilayer Perceptron, SVM, J48, kNN and random forest classifiers. They used extracted features alone and in groups. The decision tree gave the best results in most cases. The single feature that discriminates stress solely best was the difference in the average number of beats. They achieved 88% classification accuracy with three classes of stress.

DriveDB database was used in [109] to measure driver stress levels. Wang et al. extracted features by employing trend-based and parameter-based methods. In trend-based methods, long-term statistical features were calculated. Parameter-based methods extracted features from 5-minutes HRV analysis. The kernel-based class separability (KBCS) technique was employed to select features. LDA and PCA were used to reduce feature dimensions. LDA, PCA and KBCS were used in different combinations. They showed that the best accuracy was achieved when all three methods were used with trend-based features. Long-term features are more distinctive than parameter based short-term features. They reached 97% classification accuracy by using a kNN classifier.

Stress levels of bus drivers were investigated in [13]. Rodrigues et al. developed a bus driver stress detection scheme. They used the VitalJacket biodevice to measure ECG. As a reference, they used pressing the button for stressful events and at the end of the day, a questionnaire filled by bus drivers. They gave the extra information and visualization using Google Maps and GPS data to help the driver remember the stressful event. They collected data from 36 drivers. Another significant result is that they found stressful locations in Porto, Portugal which have tight roads, low visibility crossroads, specific intersections. This information can be helpful to municipalities to alleviate problems in these areas. The LF component in HRV was the most discrimi-

native feature. They also demonstrated that experience and stress levels are inversely proportional by using the background data from drivers.

3.2.2.1. Insights from the Experiments in Automobile Environments. In the automobile environment, movements are restricted and sensors like cameras can be integrated into the environment. In that sense, this environment is similar to the offices and workplaces. People tend to be stressed in automobile environments, especially in crowded cities with traffic jams. In the literature, most works have used the driveDB database [111]. Fifty minutes ECG, EMG, EDA, respiratory sensor data were collected from 24 drivers in Boston in this database. EDA and ECG signal combination has achieved the best two performances in the automobile environment. For the driveDB database, SVM and kNN classifiers achieved the best accuracies (up to 99% [105]) in a 3-class classification. Another significant research in this environment is the bus driver stress detection in Porto [13]. They stated that stress has a negative impact on memory and this affects the accuracy of the ground truth questionnaires. They added photos of stressful events for drivers to help them remember the important times at the end of the day. They further demonstrated that the LF component in the HRV is the most distinctive feature.

3.3. Semi-Restricted Environments

3.3.1. University Campus Environments: Student Stress

Detecting stress in university campus environments would help students to increase their academic success and quality of their lives. Wang et al. [116] conducted a very wide range of research on college students. They developed a studentLife application that collects activity, sleep, conversation data and EMA (Ecological Momentary Assessment) questionnaires filled by participants. Their aim was to evaluate mental well-being, stress, loneliness and their correlation with academic performance. They made the dataset publicly available. They showed that sociability, location changes, activity of students decreases when a midterm approaches. They found a correlation

Table 3.4: Stress Detection Experiments in University Campus Environments

Article	Stress Signal	Method	# of Classes	Accuracy %
[112](2016)	ECG (Ultra Short Term Features)	SVM - DecisionTree - NaiveBayes - MultilayerPerceptron	2 (Stressed, Relax)	80
[113](2015)	Accelerometer - Speech - GPS - Wifi - Proximity - Call Statistics - Light Sensor	SVM - J48 - Bagging - RandomForest - Ordinal classifier	3 (Not stressed, Slightly stressed, Stressed)	60
[114](2014)	Call-SMS Statistics - Bluetooth - Weather Conditions - Personality Traits	Random Forest	2 (Stressed, Relax)	72
[115](2012)	GPS - Social Interaction(Bluetooth Connections) - Call-SMS Statistics	-	2 (Stressed, Relax)	53

between GPA and conversation and the indoor mobility of students. Although the research is not about stress detection, the data collected from students can be used to recognize stress.

Castaldo et al. [112] investigated student stress during an oral exam by using ECG. The difference from other HRV using articles is that they detected stress using ultra short-term HRV (3 minutes). They recorded controlled resting base condition from the same student after a vacation. 18 features were extracted from the time and frequency domains. Non-linear features were also used. 12 out of 18 features correlate with stress. Naive Bayes, Decision Tree, SVM and Multilayer Perceptron classifiers were applied to the data. They achieved the best accuracy with the C4.5 tree classifier which is approximately 80%. The strong sides can be listed as the measurement of daily life stress of students during an oral exam and ultra short-term HRV usage.

In earlier research, Gjoreski et al. [113] developed a student stress detection scheme by using the data from smartphones which was collected during the study in [116]. They used the activity (from the accelerometer), audio classification (silence, voice, noise), GPS, WiFi, conversation (whether there is one nearby), time and duration of calls, light sensor data and questionnaire data from subjects. 47 features were extracted from these data. They applied three different approaches. The first is the general model for all subjects with Leave One Student Out testing, they obtained 43% classification accuracy when discriminating not stressed, slightly stressed and stressed cases. They divided students who react similarly into clusters and calculated inside

cluster accuracy, but results are the same with the general model. The last classification was dividing data of each student into two parts. The first part of the data were used as a training and the rest of the student data were used as the test data. They called this method learning with a calibration phase and achieved 60% classification accuracy. They stated that low classification accuracy caused by subjectiveness of the perceived stress and difficulty of determining stress using only smartphones. For future works, they plan to add voice analysis from smartphones and social media analysis.

Bogomolov et al. [117], [114] proposed a scheme that automatically measures the stress of graduate students. They collected data from 117 subjects for approximately a year. The selected features were the user's mobile phone activity (call and SMS logs, Bluetooth proximity hits), weather conditions, personality traits. They divided weather conditions into six: mean temperature, pressure, total precipitation, humidity, visibility and wind speed, namely. They also classified personality traits into the Big five: extraversion, neuroticism, agreeableness, conscientiousness and openness to experience. They decided personality by applying the Big five questionnaire. The stress reference was determined from a self-perceived stress level questionnaire filled at the end of a day. As a classifier, a number of classification algorithms were applied, but random forest outperformed other algorithms. They reduced 500 features into 32 features by applying the Pearson correlation analysis. They achieved 72% classification accuracy. The weaknesses of this research can be counted as a biased sample (all grad students) and interaction with people who are not part of the study was not taken into account. Their final aim is to integrate the non-obtrusive system into the clinical support or office self-monitoring applications.

Bauer et al. [115] presented a mobile phone stress detection scheme. They defined a two week exam period as a stressful time and after an exam a two week period as a stress-free time on participating students who were given an Android mobile phone with the required logging software. They divided behavioral patterns into three classes: location behavior, social interaction behavior, call and SMS behavior. They defined Regions of Interests (ROI) and tracked it with GPS and Wi-Fi. The k-means algorithm was employed to determine these regions. They measured social interaction with the number of different Bluetooth device connections. They measured the difference in

Figure 3.1: Wearable Devices for collecting data in Daily Life (Left Top: Fitbit, Right Top: Sony Smartband, Middle: Samsung Galaxy Gear S3, Bottom: Empatica E4)

behavior between the exam period and the free period. They showed that participants changed their behavior 53% on average. For future works, they plan to add new behavioral features (outgoing SMS, phone call duration, etc.), detect behavior change in real time and add a mobile questionnaire.

3.3.1.1. Insights from the Campus Environments. The campus environment is the most similar environment to the daily life experiments among the restricted environments. Movements are not limited and campus environments are less controllable with sensors when compared with offices and automobile environments. Therefore, the accuracy of the classification schemes is relatively lower (See Table 3.8). The ECG signal and the decision tree classifier achieved the highest accuracy in 2-class classification. Most works only used features extracted from the smartphones. Wearables with more sensors (such as EDA, temperature, accelerometer) can be integrated into these systems to improve the performance of these systems.

The ultimate aim of the studies is to detect the stress levels of individuals in their daily lives (See Table 3.9). Data is collected with unobtrusive wearable devices (Figure 3.5). The number of participants and data collection intervals are provided in Table 3.10. Ciman et al. [118] developed a stress detection scheme by analyzing the

smartphone usage patterns. They divided the experiment into two parts. The first part took place in the controlled lab environment. They developed an Android application with search and write tasks. In these tasks, users tap, scroll, swipe and text input gestures were recorded. They provided a stressor task to induce stress on the subject. For reference, 5-point Likert scale stress states were reported using the Experience Sampling Method. There were 13 subjects for this part. They achieved approximately 80% stress detection classification accuracy using SVM, NN, kNN and decision tree classifiers. For the second part which they called "in the wild", they collected used application types, physical activity of the user, light values of screen and events related to a mobile phone screen. The subject is using the smartphone in his/her own daily life in this part. They achieved approximately 70% classification accuracy. The weakness of the second part in the used application type evaluation is the issue of causality. They reported that they do not know whether the used application causes the stress or stress causes the used application type. For instance, a researcher might notice that when a subject is stressed, the used application type is generally social media. However, they could not conclude whether if the social media app usage caused stress. Because another possibility might be that the increased stress level caused social media app usage. They emphasized the causality issue in that case. As future works, they plan to build a remote assessment system for data collection and add wearable devices to the system.

Gjoreski et al. [119] designed a stress detection scheme in both daily life and laboratory. They used an Empatica wrist device. The mental arithmetic task was used to induce stress. From BVP, HR, ST, GSR and RR sensors, 63 features were extracted. In the laboratory environment, they discriminate stress from no stress with 83% accuracy. They also introduced a three-class classifier with no stress, low stress and high stress with 72% classification accuracy. However, for the daily life, they used activity recognition which discriminates sitting, walking, running and cycling. The activities are enumerated according to their intensity (i.e., lying corresponds to 1 to running corresponds to 5). For the stress detection intervals, the average intensity is calculated in this manner. The reason behind using an activity recognizer and intensity calculation is to distinguish an intense physical activity from a stressful situation by giving

context information to the daily life stress detector as an input. They divide a day into one-hour episodes. The subjects record stressful events by pushing a button. They achieved 76% and 92% stress detection classification accuracies for no-context and with context information (using the activity recognizer) cases respectively. They used 11 days of daily life data from 5 users. As future works, they plan to discriminate more stress levels also. Personalization depending on age, gender, fitness condition is also another planned future work. They expanded their work in [120] by creating a stress detection model for the laboratory environment (2-min interval), activity recognizer and context information. In the laboratory, they recorded a one-day relax baseline and mental arithmetic task as the stress phase. They mapped the subjective scores to stress labels in this phase. Daily life system uses these models and decides for every 20 minutes. HR, BVP, IBI, EDA and temperature data were obtained. They applied all ML algorithms in the MATLAB Weka toolkit. They achieved 70% recall rate with 95% precision.

3.4. Nonrestricted Daily Life

Gimpel et al. [128] presented a stress detection scheme by using the smartphone data. They stated that their most significant innovation from similar works is that their system is not based on user inputs or additional devices. They published an Android app and analyzed the data from users. They made use of 36 hardware and software sensors. Their research was in progress; thus, they did not provide results. However, they found out that high smartphone usage, average battery temperature, the maximum number of running applications and the frequency of switching the display on are related to high stress. They also pointed out that in daily life, we do not know whether high smartphone usage causes stress or vice versa. Furthermore, they reported that the perceived stress is not necessarily the same as the actual stress.

Another daily life stress detection scheme using smartphones was presented in [121]. Sysoev et al. used audio, gyroscope, accelerometer, ambient light sensor data, screen mode changing frequency, self-assessment and activity type. They did not

Table 3.5: Stress Detection Experiments in Unrestricted Daily Life

Article	Stress Signal	Method	# of Classes	Accuracy %
[118](2016)	Mobile Application Usage Pattern - Physical Activity - Light Sensor - Screen Events	SVM, ANN, kNN	2 (Stressed, Relax)	70
[119](2016)	BVP - SkinTemperature - EDA - RR - HeartRate (Without Context Info)	RandomForest	2 (Stressed, Relax)	76 (With Context Information 92)
[121](2015)	Speech - Gyroscope - Accelerometer - Light Sensor - Screen Events - Activity Type	RandomForest - Simple Logic - DecisionTree	2 (Stressed, Relax)	77
[122](2014)	ECG - SkinTemperature - Respiration - Accelerometer- EDA	SVM - kNN - ANN - RandomForest	2 (Stressed, Relax)	73
[123](2013)	Call-SMS Statistics - GPS - Screen On/Off - Accelerometer - EDA	SVM - SVM-RBF - kNN	2 (Stressed, Relax)	75
[124](2013)	HRV - Speech - Physical Activity - GPS - Accelerometer	Logistic Regression	3-class(Low, moderate, high stress)	61
[125] (2015)	ECG + Respiratory + Accelerometer	SVM	2 class(Stressed, Relax)	72
[11] (2011)	ECG + Respiratory	J48, J48 + Adaboost, SVM, (HMM (Hidden Markov Model) for daily life)	2 class(Stressed, Relax)	0.71(correlation with self-reports)
[126](2012)	Speech	GMM	2 (Stressed, Relax)	80.5
[120] (2017)	HRV - EDA - Temperature	Weka Toolkit	2 class(Stressed, Relax)	70 (precision with 95% recall)
[127] (2018)	Usage Data for different application categories	HMM with MPM	2 (Stressed, Relax)	68

use any wearable sensors. Self-assessment stress level based seven scale questionnaire NASA-TLX was used as a reference. They combined activity recognition with their stress detection system to increase its performance. They achieved approximately 77% classification accuracy which corresponds to 3.8% increase when compared to the stress detection scheme not using the activity recognizer. They created a separate model for the standing activity among others which results in 1.5% accuracy increase.

Maier et al. [129] presented their work in progress. They used HRV with activity and contextual data to increase the accuracy of the stress detection scheme which uses ECG. They monitored the stress level of a user and when the predetermined threshold was exceeded. Users can quit from the stressful event or some relaxation methods are provided to them when they are stressed. Their app has an automatic adaptation for individuals. The app collects data from the user about emotion and energy. Physical activity from the accelerometer, the location from the GPS, change of location and

time of the day information were added to HRV to increase the accuracy. They applied a particular neural network classifier (BINN). However, they did not provide its accuracy. They plan to adapt the work for mentally ill people in addition to healthy people. They also plan to add the Perceived Stress Questionnaire (PSQ) as a reference.

Muaremi et al. [122] presented a system for detecting stress from the sleep patterns. They collected data from 10 Hajj pilgrims. Both physical and physiological data measurements are obtained. ECG/HRV, respiration, body temperature, GSR, upper body posture sensors, accelerometers on body and arms were the data sources. Chest strap and Empatica wristband were used to collect data. After features were extracted, SVM, kNN, NN, RF(Random Forest) and logistic regression classifiers were applied to the data. SVM obtained the best classification accuracy. ECG/HRV, upper body activity and sleep duration were the most distinctive modalities. 73% maximum classification accuracy was achieved. For future works, in addition to the existing features, sleep stages, the number of wake-ups and the percentage of time that the subject is lying but not sleeping could be taken into consideration.

Sano et al. [123] developed a stress recognition scheme using a wrist sensor, a mobile phone and surveys. Mobile phone usage was modeled from a call, SMS, location change (mobility) and screen on/off features. Accelerometer and EDA data were collected from a wrist sensor. They obtained information about sleep, caffeine and alcohol usage, stress, mood and tiredness from the surveys. Sequential Forward Floating Selection (SFFS) was employed to find the most discriminative features. SVM with linear and RBF kernels and kNN were applied for classification. PCA was also employed with classifiers. Acceleration during the second sleep phase and between 6-9 pm, few or short SMS and a small number of screen display on/off behavior between 6-9 pm or 9-12 am were found to be correlated with high stress. By employing screen on/off, mobility, call, acceleration and EDA data, they achieved 75% classification accuracy in daily life with two classes.

Kostopoulos et al. [130] proposed a stress detection application for the Android operating system called StayActive. They employed sleeping patterns, social interaction and physical activity to detect stress. They used a combination of offline mathematical model and an online machine learning algorithm to detect stress and sug-

Table 3.6: Campus Environment and Daily Life Experiment Details

Article	Devices	# of Participants	Experiment Duration
[127](2018)	Smart Phone	28	4 days
[130](2017)	Smartphone	5	4 weeks
[112] (2016)	3-lead ECG	42	2 days
[118](2016)	Smartphone	25	4 weeks
[119](2016)	Smart Wrist (Empatica E4)	5	55 days
[129](2015)	Smart Wrist(HRV data only)	35	4 weeks
[125](2015)	ECG + Respiratory + Accelerometer (Wearable Sensors)	23 (3 good EMA reports)	1 week
[116](2014)	Smartphone	48	10 weeks
[117](2014)	Smartphone	111	7 month
[123](2013)	Smartphone + Smart Wrist	18	5 days
[124](2013)	Smartphone + Smart Wrist(HRV)	35	4 weeks
[115](2012)	Smartphone	7	4 weeks
[11](2012)	ECG + Respiratory (Wearable Sensors)	17	2 days

gested some relaxation activities when stress is recognized. They collected user stress feedback by using the modified Circumplex Model of Affect by James A. Russel [131]. They added a relaxed-stressed question to this questionnaire. They used the sleeping hours as a feature for the sleep patterns. A punishment was given to lack of sleep and excessive sleep. They measured the social interaction by using the number of touches to screen, number of calls and SMSs. To compute feature weights, they asked the user how important they consider these features. These were the preliminary results. They plan to extend the number of people and add some physiological signals to the study as future works.

Muaremi et al. [124] presented a stress detection scheme for iOS devices. They gathered data from both smartphones and wearable chest belts. HRV data were collected from the chest belts and audio, physical activity, communication data were gathered from smartphones. They divided the day into four quarters and users were asked to fill out the questionnaire and record their voice in each quarter. This questionnaire was a modified version of the Positive and Negative Affect Schedule (PANAS) questionnaire. They removed some questions because the subjects complained about the length of the questionnaire. Before going to sleep, a subject provides stress level information for the whole day. They anonymized the speech and used the openSMILE library for the audio feature extraction. Physical activity data were gathered using GPS and accelerometer. Social interaction features were call events, number of calls,

duration and ratio of incoming and outgoing calls. For the HRV data, time, frequency domain and nonlinear features were extracted. They modeled and scored daily and long-term stress for each individual. The stress model is dynamically changing in time. They removed highly correlated features and obtained 13 smartphone and 10 HRV features. After the feature selection, they noticed that accelerometer, calendar, battery features were not selected because subjects are not using the smartphones as a calendar, they used their personal computers for that purpose. Another reason was that they put smartphones on the table and activity data becomes useless. They further stated that the HRV data is more distinctive than the smartphone features. By employing logistic regression models, they achieved 61% classification accuracy for three-class (low, medium and high stress) classification with all features. As future works, they plan to add comparative questionnaires with past time, apply SVM and random forest classifiers, develop a scheme that has sleep stage awareness and measure HRV during a day in addition to the sleep times.

In [125], researchers developed a stress detection scheme for continuous monitoring. They used a chest band as a respiratory sensor, a 2-lead ECG and a 3-axis accelerometer for data collection. They first designed a laboratory experiment with a baseline, public speech, arithmetic test and ice cold test sessions which are used to train a model with 24 (training)-26 (test) participants. They further collected field data from 23 participants in the wild. Participants are required to fill 15 questionnaires. They used the adapted version of the PSS survey which includes five questions. The SVM classifier was applied. In Laboratory testing, they achieved 89% recall. On the other hand, they obtained 72% accuracy in the wild. As future works, they planned to increase the accuracy of the system by applying new data processing methods, develop a stress management scheme, add some context information such as GPS, light exposure and social interaction.

Lu et al. [126] detected stress from human voice gathered from smartphones. They separately evaluated indoor and outdoor environment cases. They created three experimental setups: a job interview, marketing jobs and a neutral task. They used GSR data as a reference. For each task, they divide the audio into voice and non-voice parts. They implemented different algorithms for speaker segmentation for indoor and

outdoor environments. For the outdoor environment, they employed two smartphones for speaker segmentation, one in the shoulder another in the waist. Used features were pitch (F0), spectral energy distribution, speaking rate, nonlinear TEO transformation results and MFCC. The classification was performed based on GMM and likelihood functions. They employed three models: a universal model, personalized model and individual-adapted model (by using MAP). They stated that pitch features and speaking rates are the most distinctive features for stress detection. General models achieved 69% (outdoor and indoor average) classification accuracy whereas personalized models achieved 80.5%. Supervised adapted models achieved 10% higher than the general, 2% lower accuracy than the personalized model.

Reimer et al. [132] described their pilot mobile system on stress recognition. They chose only HRV signal stating that EDA is not wearable which is not a true statement. They also added the context information like GPS, physical activity, time of a day and week. Their main novelty in this app is the automatic user calibration. The algorithm sets thresholds to discriminate three stress levels at the initial learning phase. 35 healthy subjects participated in the experiment for four weeks. The PSQ questionnaire was collected at the end of each week. They did not present the accuracy of the system.

Labeled data is an essential problem for daily life stress detection algorithms. In most of the studies, collecting labeled data is a bottleneck and available labeled data has the subjectivity problem. Vildjounaite et al. [127] proposed an unsupervised stress detection scheme. They used mobile phone usage data and applied HMM with MPM for classification. 28 people participated in the experiment with an average of 4 days. They achieved maximum 68% accuracy for semi-personal data.

In [11], the authors collected data from 21 subjects in the laboratory environment and gathered self-reports. They developed two stress recognition models. The first one is the physiological classifier which uses ECG and respiratory signals for detecting the stress of an individual. They divide the experiment into the baseline, public speaking, mental arithmetic and cold pressor sessions. By applying J48, J48+Adaboost and SVM classifiers, they obtained 90.17% accuracy in a two-class (stressed or non-stressed) classification task. They further developed a perceived stress model to estimate the stress perception. This model takes the outputs of a physiological classifier as an input and

predicts the accumulation and decay of the stress level of a participant by employing HMM. They calibrate the parameters of the model for each participant. This model has achieved 0.72 median correlation with self-reports in the laboratory environment. In a two-day field study with 17 participants, the perceived stress model has 0.71 correlation with the self-reports. As future works, the researchers have planned to add more modalities, develop new methods to clean the signals during a physical activity and increase the number of the perceived levels of the stress.

3.4.1. Insights from Daily Life Experiments

The ultimate aim of researchers is to detect stress levels and help individuals to cope with high stress in their daily lives. However, there are some issues to deal with when taking a step outside the laboratory (see Section 2.3). The stress detection accuracies of daily life schemes are lower than in restricted environments and laboratory environments. Smartphone usage statistics and wearable sensor stress detection schemes have accuracies between 70% and 80%. The combination of smartphone usage data and physiological signals (such as PPG and EDA) from an unobtrusive wearable sensor data would increase the stress detection accuracy. One of the most significant issues in the daily life data collection is the ground truth collection and the reliability of the questionnaire data (see Section 2.3). Researchers developed an unsupervised stress detection for daily life by employing HMM in [11] and [127]. Unsupervised schemes may eliminate the need to collect the ground truth questionnaires. Gjoreski et al. [120] used activity recognition to have knowledge about context and improve their stress detection performance. Other context data such as information on whether the subject is inside/outside of a building by employing light sensors and GPS, calendar load, weekend/weekday knowledge, social interaction knowledge from Bluetooth sensors can be added to increase the performance of these schemes. The duration of experiments and the number of participants are limited in the literature (see Table 3.10). To get a better understanding of stress behaviors and discriminative features, these values should be improved. Lastly, after detecting stress, some stress alleviation methods specific for individuals should be suggested to recover from the negative effects of stress. We men-

tioned the scientifically proven stress alleviation methods in Section 2.4. We selected the methods that can be applied without interrupting daily routines and can also be used in indoor environments.

4. UNOBTRUSIVE STRESS LEVEL DETECTION SYSTEM

In this thesis, an unobtrusive stress detection system that uses heart activity, skin conductance, accelerometer and skin temperature signals for recognizing multiple stress levels is described (see Figure 4.1 for the high level system description). In order to eliminate the artifacts caused by unlimited movements in real-life scenarios, preprocessing tools specific for each modality were developed and used. The most distinctive features used in the literature for each signal were also selected and extracted. After the feature extraction phase, the best performing classifier algorithms were applied to the feature set. Our system works with the data collected from both Samsung Gear S smartwatches and Empatica E4 smart bands even though they have different software platforms and sensor types. Each modality-specific preprocessing and feature extraction tool is described in detail.

4.1. EDA Preprocessing Artifact Detection and Removal Methods

Intense physical activity and temperature changes contaminate the EDA signal. Therefore, affected segments should be filtered out from the original signal. In order to detect the artifacts in the EDA signal, we used the EDA Explorer software [133] which is developed specifically for this signal. While developing this tool, experts labeled the artifacts manually. They trained an SVM model by using the labeled data. Accelerometer and skin temperature signals are also used for artifact detection. EDA Explorer achieved 95% accuracy for detecting artifacts. We removed the parts that this tool detects as artifacts from our signals. We further added batch processing and segmentation to this tool as new properties (see Figure 4.2).

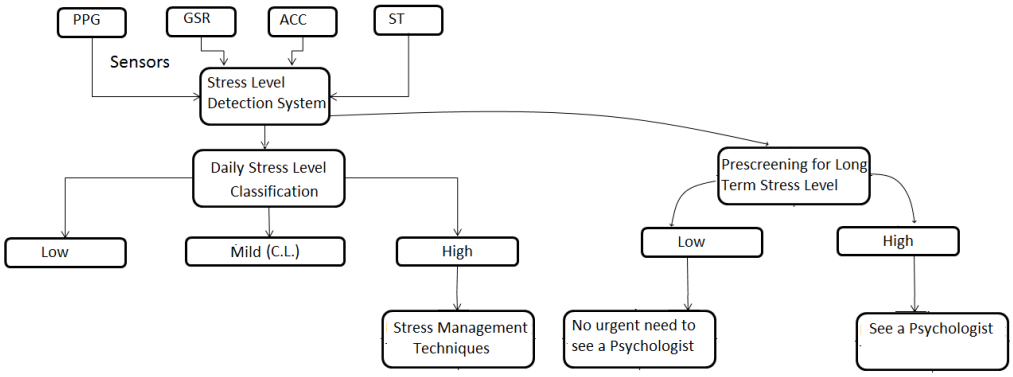

Figure 4.1: The high level description of the proposed system.

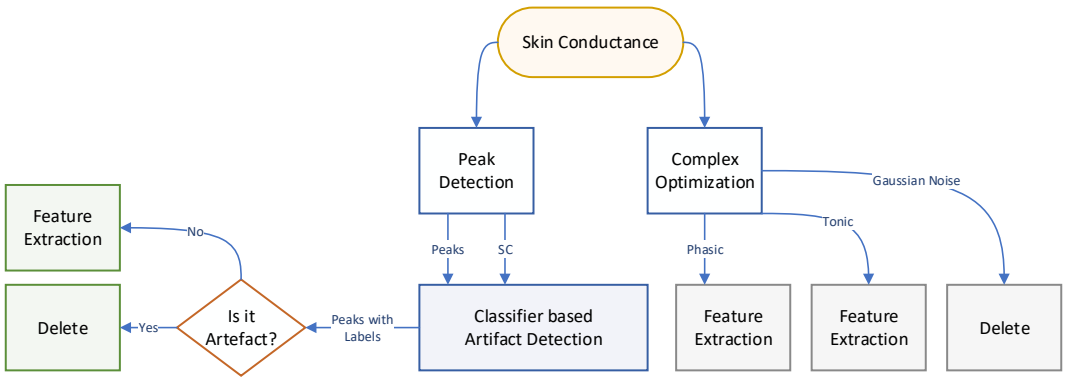

Figure 4.2: The detailed EDA processing module for artifact removal and feature extraction. We selected the non-artifact peaks and features from tonic and phasic components are extracted from the clean signal.

4.2. EDA Feature Extraction Methods

After the artifact removal phase, features are extracted from the EDA signal which has two components phasic and tonic, namely. We used the cvxEDA tool [134] for decomposing the signal into these components. This tool uses convex optimization to estimate the Autonomic Nervous System (ANS) activity which is based on Bayesian statistics. After this tool is applied, our algorithm extracted features of both components.

Table 4.1: HRV, EDA and Acceleration features and their definitions.

Feature	Description
Heart Rate Variability Features	
Mean RR	Mean value of the RR intervals
STD RR	Standard deviation of the inter-beat interval
RMSSD	Root mean square of successive difference of the RR intervals
pNN50	Percentage of the number of successive RR intervals varying more than 50ms from the one
HRV triangular index	Total number of RR intervals divided by the height of the histogram of all RR intervals measured on a scale with bins of 1/128 s
TINN	Triangular interpolation of RR interval histogram
LF	Power in low-frequency band (0.04-0.15 Hz)
HF	Power in high-frequency band (0.15-0.4 Hz)
LF/HF	Ratio of LF-to-HF
pLF	Prevalent low-frequency oscillation of heart rate
pHF	Prevalent high-frequency oscillation of heart rate
VLF	Power in very low-frequency band (0.00-0.04 Hz)
SDSD	Related standard deviation of successive RR interval differences
Acceleration Features	
Mean X	Mean acceleration over x axis
Mean Y	Mean acceleration over y axis
Mean Z	Mean acceleration over z axis
Mean ACC MAG	Mean acceleration over acceleration magnitude axis
Energy	FFT energy over mean acceleration magnitude
Electrodermal Activity Features	
Mean Tonic	Mean of the phasic component
SD Tonic	Standard deviation of phasic component
Perc20	20th percentile of the phasic component
Perc80 Tonic	80th percentile of the phasic component
Quartdev Tonic	Quartile deviation (75 percentile - 25 percentile) of the phasic component
Strong Peaks Phasic	The number of strong peak per 100 seconds
Peaks Phasic	The number of peaks per 100 sec.

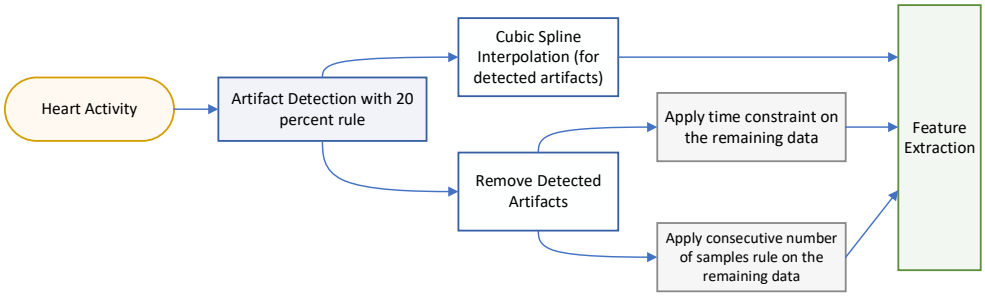

Figure 4.3: The detailed HRV processing module for artifact removal and feature extraction. After the removal of artifacts, we used the first option of the system which is the interpolation of the removed points since it has higher performance than appying more constraints on both time and samples [6]. Lastly, HRV features are extracted.

4.2.1. Tonic Component Features

The tonic component in the EDA signal represents the long term slow changes. This component can also be called as the skin conductance level (SCL). It could be regarded as the indicator of general psychophysiological activation [42] and can depend highly on individuals [42]. The values generally can rise up to 15 ms and above 20 ms values are regarded as highly unlikely [42]. The tonic component is used since long term changes should not be overestimated with event-related fast changes. For this purpose, the phasic part is subtracted from the EDA signal. After the decomposition of EDA signal, we extracted mean, standard deviation, 20 percentile, 80 percentile and quartile deviation (75 percentile - 25 percentile) which are the most distinctive features in the literature [19], [120] from the tonic component (see Table 4.1).

4.2.2. Phasic Component Features

The phasic component represents faster (event-related) differences in the EDA signal. Peaks of EDA that happens as a reaction to a stimulus is also called Skin Conductance Response (SCR) [42]. It happens with a delay after the stimulus [135]. After we decompose the phasic component from the EDA signal peak related features e.g., peak per 100 seconds, a strong peak per 100 seconds are calculated. The peaks with more than 1 micro siemens response are identified as strong peaks.

4.3. Heart Activity Preprocessing Artifact Detection and Removal Methods

Unlimited movement of subjects and improperly worn devices also contaminates the HRV signal collected from smartwatches and smart bands. In order to address this issue, we developed an artifact handling tool in MATLAB that has the batch processing capability. First, the data is divided into 50% overlapping segments [101] as recommended in the related literature. The artifact detection percentage rule (also employed in Kubios [136]) is applied after the segmentation phase. In this rule, each data point is compared with the local average around it. If the difference is more than a predetermined threshold percentage, the data point is labeled as an artifact. The threshold is defined as 20% difference which is commonly selected in the literature [101]. In our tool, we further developed two options after an artifact is detected: interpolation or further filtering. We described these methods in detail in Section 4.3.1 and 4.3.2 (see Figure 4.3).

4.3.1. Artifact Detection Percentage Threshold - Removal

The first option is removing the artifact data point from the signal. However, after the data is removed, this creates holes in segments which makes it difficult to evaluate them as a whole. In order to overcome this issue, we implemented two filters: the minimum consecutive time and the minimum consecutive number of samples. The minimum consecutive time constraint dictates a minimum non-interrupted (with deleted artifact holes) time series with a determined length on a segment to be evaluated and to extract features. Similarly, the minimum consecutive number of samples filter dictates a determined number of consecutive samples. By applying these filters, we ensure that segments that have too many artifacts distributed among the clear data are not evaluated and affect the performance of our system.

4.3.2. Artifact Detection Percentage Threshold - Interpolation

After detecting the artifacts, another option would be to replace them with interpolation. The choice of the interpolation technique is a critical decision. The interpolation function should be similar to the heart signal. To this end, we selected shape preserving cubic spline interpolation and applied the built-in MATLAB function. The tool has a further batch processing feature. Parameters such as the length of local mean, the percentage of artifact detection rule, the minimum consecutive time and data sample constraints are parameters that could be changed in our tool.

4.4. Heart Activity Feature Extraction Methods

In order to extract features from the HRV signal, MATLAB built-in tools and Marcus Vollmer HRV toolbox [137] are used. The features could be examined in time and frequency domain categories.

4.4.1. Time Domain Features

We searched the literature and selected the most distinctive features in the time domain. Mean value of the heart rate (Mean HR), standard deviation of the inter-beat interval (STD RR), mean value of the inter-beat (RR) intervals (Mean RR), root mean square of successive difference of the RR intervals (RMSSD), the percentage of the number of successive RR intervals varying more than 50 ms from the previous interval (pNN50), the total number of RR intervals divided by the height of the histogram of all RR intervals measured on a scale with bins of 1/128 s (HRV triangular index), and triangular interpolation of RR interval histogram (TINN) are selected and extracted from the HRV signal.

4.4.2. Frequency Domain Features

Features from the frequency domain are also extracted. However, since the heart peaks are not equidistant from each other, FFT could not be applied directly. We first preprocess the signal for equal distant samples (resample) and then applied FFT. As an alternative, we applied Lomb-Scargle periodogram [138] which is developed for this type of signal to convert to the frequency domain. We extracted features from both methods. Low frequency power (LF), high frequency power (HF), very low frequency power (VLF), prevalent low frequency (pLF), prevalent high frequency (pHF), the ratio of LF to HF (LF/HF) (Preprocessing +FFT), LF, HF, LF/HF (Lomb-Scargle) features are selected and extracted from frequency domain representation of the HRV signal.

4.5. Accelerometer Feature Extraction Methods

The accelerometer sensor data is used for two different purposes. Firstly, we extracted features from this sensor. As mentioned above, this sensor was also employed to clean the EDA data along with the skin temperature sensor as a second use. The mean value of 4-axis and the frequency domain energy of magnitude were the extracted features.

4.6. Data Fusion

After we divided our data into segments (between 60-480 seconds) different modalities should be aligned in time. Especially, the HRV signal starts with a delay (to calculate the heartbeat per minute at the start) and all signals should be synchronized. We included start and end timestamps for each segment and each modality is merged with a script if their time intervals are overlapping.

4.7. Feature Selection

4.7.1. Feature Selection

We applied correlation-based feature (CBF) [139] selection using the Weka [140] tool. The importance of the features is shown in Figure 4.4. Since our goal is to develop a stress detection model that works in daily life settings, we conducted experiments with 1 to 20 best features for the unrestricted daily life data. We achieved the best results with ten features for HRV+EDA, five features for HRV, and five features for EDA. We applied the classifiers to the selected features.

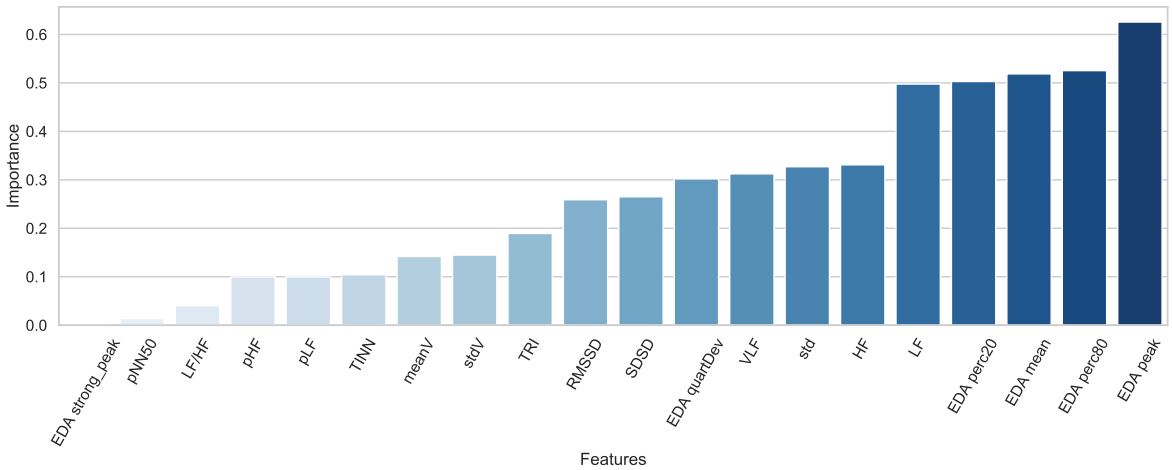

Figure 4.4: Features listed in order of importance based on correlation-based feature selection for the DDSR model. EDA-peaks is the feature that has the highest importance, whereas EDA-strong-peak has the lowest.

4.7.2. Principal Component Analysis

Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms variables into uncorrelated principal components. It is widely used in data analysis, previous studies reported that when it is used with machine learning algorithms, their classification performance enhanced. We applied PCA implementation of the Weka package. This implementation allows us to select the covered variance. We selected covered variance as 80% in order to avoid over-fitting.

4.7.2.1. Preparation of the Data for ML Algorithms. Although we have evenly distributed data in terms of known context labels in some datasets, we have class imbalance problem in others. We overcome this problem by randomly undersampling the extra samples of the majority class, which is the most commonly used procedure for imbalanced datasets [141]. We further applied normalization on features to prevent overfitting. Lastly, we convert the numeric labels to the nominal type as an input to the Weka toolkit classification algorithms.

4.8. Machine Learning Classifier Algorithms

The Weka toolkit [142] and the LSTM architecture in Keras [143] are used for discrimination of stress from the cognitive load. The Weka toolkit is the most commonly used and comprehensive machine learning platform in the literature. It has several preprocessing features before classification. Numeric to nominal transformation among them was used to convert the stress level column into a nominal class attribute. We used the LSTM architecture to increase the performance of our daily life perceived stress detection besides traditional ML algorithms.

In this study, we employed six different machine learning classification algorithms to recognize different stress levels:

a) Principal component analysis (PCA) and support vector machine (SVM) with a linear kernel
b) Random Forest (with 100 trees)
c) K-nearest neighbours (n=1)
d) PCA and Linear discriminant analysis (LDA)
e) MLP (Shallow Neural Network)
f) LSTM

The first five classifiers are chosen because they are the most prominent traditional

ML algorithms. MLP (Shallow Neural Network), Random Forest, kNN (One of the Simplest), SVM and LDA are the most commonly used and successful classification algorithms for stress detection [120], [144]. In order to evaluate the performance of the classifiers, we applied 10-fold cross-validation by partitioning the original sample into a training set to train the model, and a test set to evaluate it, then we changed the test and training sets until each partition is used for the test set. We further created separate training (80%) and test sets (20%) and evaluated the results to compare with 10-fold cross-validation results. Our system is capable of detecting stressed, cognitively loaded and baseline states by using the mentioned algorithms.

We further applied LSTM to improve and compare the performance of our system with traditional machine learning algorithms in Chapter 7.

4.8.1. Parameter Tuning

For each classifier, there are some parameters to adjust such as k value of kNN, regularization parameter and kernel type of SVM, maximum depth, minimum samples in a leaf, maximum number of trees parameters in RF, number of layers, batch size, dropout, number of neurons, optimizer type and learning rate parameters of deep neural network algorithms (MLP and LSTM). In the validation set, we fine-tuned these parameters and selected the best ones. The accuracies are obtained in separate test sets.

5. STRESS DETECTION IN A LABORATORY ENVIRONMENT

We conducted two experiments in the laboratory environments. In the first one, we measured the level of cognitive load on subjects. The cognitive load level was induced by making the subjects solve 4D puzzles in different difficulty levels in VR glasses. In the second experiment, we induced stress on participants by applying TSST test [145]. After that, we applied a mindfulness technique with breathing exercises. The PSS-5 (for ambulatory environments) and known context were used as ground truth labels. In these experiments, subjects wore an Empatica device and physiological signals are recorded. We used our method described in Chapter 4 to detect stress levels.

5.1. Measuring Cognitive Load and Insight

5.1.1. Experiment Design

This study [146] is originally designed to identify the emergence of learning trajectories that underpin the process inherent in problem-solving. Potential moments of insight or Aha! moments are particularly interesting. An exploratory design where a single stream of participants performs problem-solving tasks within a VR environment is used. A system which requires a subject to gain mastery over the original concept of 4D space, in particular, mastery over manipulation of a 4D construct the hypercube was implemented.

5.1.1.1. Participants. Participants were recruited from (blinded for anonymity) student and staff populations. In total, 24 participants completed the study (17 male, 7 female) with ages ranging from 18 to 45. There were no inclusion criteria with respect to the domain of expertise. Participants were compensated for their time with a $50 voucher. The study was approved by the (blinded for anonymity) ethics committee.

5.1.1.2. Apparatus. A system which involves a learner interactively manipulating rotations of a 4D cube (hypercube) and attempting to gain mastery over it is implemented. Rather than using the originally proposed visual and interactive mediums, the system is implemented for use with modern immersive technology. The implementation runs a fully immersive VR environment with an HTC head-mounted display (HMD) for visualization, and two HTC Vive controllers are used together to manipulate the rotations of the hypercube.

The task is introduced aspect to the system by presenting the user with two hypercubes, one which they manipulate, and one which is static and pre-rotated. The task for the participant is to manipulate their hypercube to match (with some built-in error tolerance) the rotation of the second static hypercube.

5.1.1.3. Procedure Overview. Participants put on the HMD and are given the two controllers. For the first three minutes of the experience, participants are exposed to a single hypercube which they can manipulate to get used to the environment, the controllers, and movements within the space. After three minutes, the participants were presented with a second hypercube and their goal was to try to match the hypercubes in terms of rotations. There were 30 puzzle cubes to match in total, and subjects were able to switch through the list of hypercubes by using a panel with forward and backward arrows on it. If subjects completed all 30 of the hypercube puzzles, the VR segment of the study would end, otherwise they were asked to remain for the full hour within the puzzle system to complete as many puzzles as they could. Upon completion of the study, participants were thanked and remunerated. For more details about the experiment, please see the article [146].

5.1.1.4. Physiological Measurements. E4 wristband sensor was used which is produced by Empatica (www.empatica.com) to measure physiological signals. The E4 sensor measures: electrodermal activity (EDA), blood volume pulse (BVP), heart-rate (HR), peripheral skin temperature, motion through an accelerometer, and it contains an internal real-time clock. We were particularly interested in the EDA, HR, temperature,

and accelerometer data. EDA and HRV signals are used as the primary measure of emotional response and represent the sensed reality dimension of data in the ERVE (Emotional Responses in Virtual Environments) methodology. Accelerometer and temperature data were used to wrangle our measured raw data, in particular to detect and clean artifacts. More detail about the system is given in Chapter 4 on how these measures were analyzed and used.

Participants are required to wear the E4 bracelet for the duration of the study. The bracelet was firmly fitted to the wrist/forearm of the participant. They can be worn on the ankle/leg if needed, but the wrist was appropriate for the study. Upon conclusion of each participant's session, the device is plugged into a PC containing the E4 software and the data is automatically uploaded.

5.1.1.5. Hypercube Difficulty. The observational measure was participant successes when they solved a hypercube puzzle. When this happened the system recorded which hypercube was solved (and therefore the hypercube difficulty) and the time-stamp of the event. In this case, the computer acts as an automatic observer of the environment. Hypercubes were categorized into easy, medium, or hard difficulty bins of which there were approximately equal numbers of each in the list of 30 hypercube puzzles.
Rotational complexity refers to the combinations of 4D rotational planes (*xw, yw, zw, xy, xz, yz*) that are rotated, and the extent of the rotation. The rotational complexity in 4D space in terms of rotational planes is not intuitive, i.e. we expect that rotation in only one plane must be easier than rotations in all six. In fact, certain rotational planes complement each other. For instance, if the '*xw*' plane only is rotated 75 degrees, it will be a much harder ghostcube to solve than if the '*xw*' and '*xy*' planes are both rotated 75 degrees. Based on these differences, three categories of difficulty ratings were assigned to the hypercubes resulting in nine easy difficulties, 11 medium difficulties, and 10 hard difficulties. Rotational complexity was also rated by an expert user of the system whose ratings were similar to the difficulties established by the rotational analysis.

5.1.2. Analysis

We explore with the following analysis what insights can be gained from the implementation of ERVE. We describe the required steps for first processing the data, and then how we analyze it in the context of the other dimensions of data. We will firstly briefly report on participants' performance and reporting of Aha! moments during the experiment.

Participant solutions from the ghostcube task are expected to be a strong indicator for learning achievements. Out of 30 hypercubes presented to participants, they were able to achieve an average 13.54 with a standard deviation of 7.57. In terms of difficulty, participants were able to achieve in total: 168 easy hypercubes, 138 medium hypercubes, and 39 difficult hypercubes.

13 out of 24 of the participants reported Aha! moments during the experiment where the mean number of reported moments is 1.79 with a standard deviation of 3.24. The total number of recorded Aha! moments is 43. There were two observable outliers where one reported a total of 13 Aha! moments, and the other reported 11. We used our system presented in Chapter 4, to detect and remove artifacts and extract features. With the EDA and Heart Activity data preprocessed, we conduct the classification with respect to observed data (ghostcube solutions) and reported data (Aha! moments).

5.1.2.1. Classification of Cognitive Load. We first analyze the relationship between EDA data and solution events. We estimate cognitive load trends based on the solution difficulties established earlier. During easier solutions, participants would be experiencing less cognitive difficulty, and harder solutions require more cognitive effort. The difficulty levels of the ghostcubes (easy, medium, and hard) solved by participants in the experiment were used as labels for machine learning algorithms (1, 2, and 3, respectively). In order to classify three cognitive load levels, we have used the Weka toolkit [142]. These classes were imbalanced due to the nature of the data where "hard" labels represent the minority class. We employed a re-sampling method from the Weka

toolkit to balance the data (i.e., added samples of the minority class) to prevent classifiers from biasing towards the majority class. To ensure an exhaustive analysis, we tested our EDA features (extracted earlier) using multiple classifiers. We have applied five different classifiers on the cognitive load data:

a. PCA and SVM with linear basis function
b. MultiLayer Perceptron (7-5-3) (MLP)
c. K-nearest neighbours (k=1)
d. J48 Decision Tree
e. Random Forest (RF, 100 trees)

These classifiers are selected due to their common application for physiological signal processing in the literature. All classifiers in the Weka toolkit were run with the algorithms' default values. For each classifier, results are validated with 10-fold cross validation. Results have been provided for 3-class classification of low, medium and high cognitive load levels (see Figure 5.1). The resulting average classification accuracy across the five different classifiers was 48.36% when discriminating using three difficulty levels of cognitive load. The most successful classifier for EDA data was the Random Forest approach which achieved 50.83% accuracy with 7.62% variance.

We applied the same process to the HRV data collected from the Empatica E4 device using the same cognitive labels. We were able to discriminate the three cognitive load levels with a resulting average classification accuracy of 82.79%. For the HRV data, the most successful classifier is the Random Forest approach which achieved 91.75% with variance 4.87% (see Figure 5.1 (HRV-CL)).

5.1.2.2. Classification of Aha! Moments. Aha! experiences (moments of insight) are times the participants felt that they made a conceptual breakthrough in determining a solution. These events occur rarely. In the duration of our experiment, less than two of these moments occur in sixty minutes on average. This means that the detection of the events is not trivial. EDA is excellent for determining the arousal level, since it can

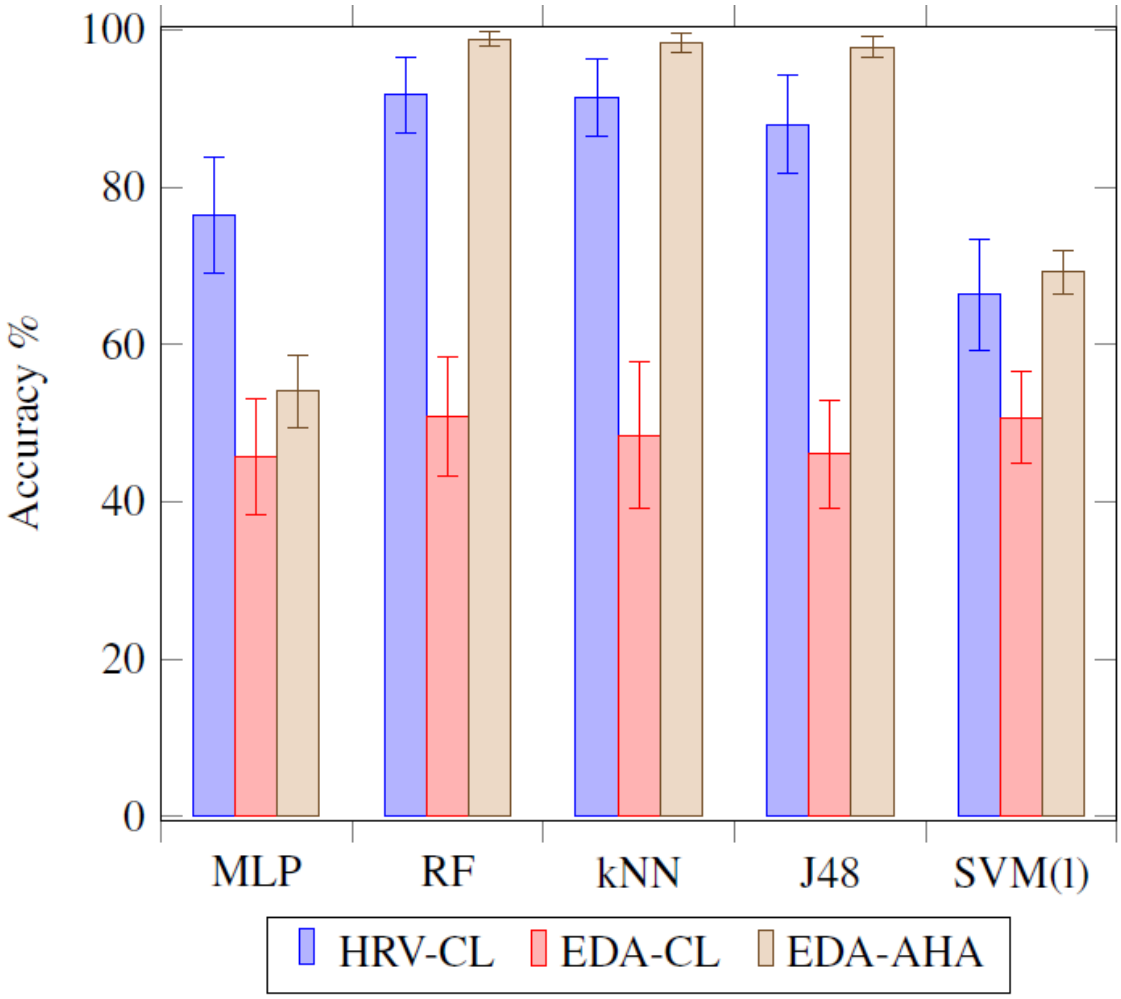

Figure 5.1: Bar plot showing all five classifiers, indicating that Cognitive Load classification using HRV features (HRV-CL) and Aha moment (insight) classification using EDA features (EDA-AHA) yield the highest prediction accuracy. Cognitive Load classified using EDA features (EDA-CL) yields low prediction rates.

assess changes in the SNS (sympathetic nervous system) [147]. Since these moments represent a form of arousal, we used the EDA signals which can detect instant arousal of individuals. We have employed the same EDA tools as specified above in Chapter 4 for feature extraction of the EDA data.

We were able to classify against multiple classes for the cognitive load classification due to our difficulty categorization. Aha! moments cannot be categorized in such a way, so a binary classification is required. We divided each 60 minute session into 60 segments, each one minute long. A one or a zero was assigned to a segment depending on whether a participant reported an Aha! moment within that minute or not. These values are then given to the classifier.

By applying the Weka toolkit with the same machine learning algorithms, we achieved an average accuracy of 83.65%. The most successful classifier was again the Random Forest approach achieving an accuracy of 98.81% with a variance of 0.9 (see Figure 5.1 (EDA-AHA)).

5.1.3. Interpretation

The average classification accuracy for the cognitive load with the EDA was 48.36% with the lowest classifier yielding an accuracy of 45.8%, and the highest classifier yielding 50.83%. The primary reason the EDA classification has returned low accuracy is due to the short consecutive segments of cognitive load data. Puzzle completion times were often quite close to each other (in the order of 10 seconds). The more delayed nature of the EDA makes it more difficult to isolate significant emotional responses against such frequent cognitive events. For instance, a user's EDA will rise as they spend 3 minutes on a level-3 puzzle and then upon completion, they are presented with a level-1 puzzle which they complete in 10 seconds. The EDA takes longer to stabilize making a correct decision difficult for the classifier.

The HRV data does not have the issue of delay as it is one of the first responses to the sympathetic nervous system. This explains the consistently higher classification

accuracy of cognitive load against using the HRV data of 82.78%. By looking at the HRV data alone, we are able to predict the difficulty of the environmental subject matter at the approximate rate of the achieved accuracy. For example, we could tell when a subject is struggling with a particular problem, or when they are finding particular problems easy.

We chose the EDA signal for Aha! moment classification. It is excellent for determining the arousal level, because it can determine changes in the SNS [147]. Since AHA moments represent a form of arousal, this is the perfect fit. We are looking for whether there is an Aha! moment in each minute interval, the delay problem for the EDA signal is also alleviated. One minute interval is enough for the recovery of the EDA signal to normal levels. The Aha! moment classification was higher again with an average accuracy of 83.66%, meaning that we have identified emotional responses in the EDA data which indicate the subject has had, is having, or is about to have an Aha! moment. Classifier performances are aligned with recent investigations of classification algorithms for wearable sensor data [6].

5.2. Inducing and Measuring Stress in Laboratory Environments

5.2.1. Experiment Design

We further collected controlled room data from a separate experiment participants. We conducted a psychological experiment, TSST, proven to be an academically correct way of inducing stress after the baseline condition we created at the beginning of the experiment. This experiment has been conducted on 14 different participants who are university students aged between 20 and 25.

The experiment which takes approximately 1-hour can be described in steps as:

- Set Up
- Pre-Stress Measurements (Baseline)
- The TSST (Inducing Stress)

- Post-stress Recovery Measurements (Recovery)

The communication language between interviewers and the participant was Turkish except in Section 5.2.1.3.vi and 5.2.1.3.vii. Please note that the mother tongue of all participants is Turkish. In addition to that, they know English as a foreign language. This circumstance affects the stress induction in Section 5.2.1.3.vi and 5.2.1.3.vii.

5.2.1.1. Set Up.

(i) Preparation of experiment areas: Camera should be set. Empatica E4 should be ready.
(ii) Interviewers should keep eye contact with the participant. Their gestures and mimics should be neutral.
(iii) The participant is informed with the procedure and then signs the consent form.
(iv) The participant wears the smartband (Empatica E4).
(v) The participant is asked to turn off his/her phone in order to eliminate distraction.

5.2.1.2. Pre-stress Measurements.

(i) The participant fills out the PSS-14 form.
(ii) The participant is told to stay in the waiting area and get rest for 10 minutes. There can be neutral contents, such as magazines, presented to the participant for this period.
(iii) The participant fills out the PSS-5 (ambulatory PSS) form. This questionnaire was first created by Cohen et al. [148] and used for measuring perceived stress in ambulatory settings in [11] (see Figure 5.2).

5.2.1.3. The TSST.

(i) The participant is directed to the interview area.
(ii) TSST speech preparation period: this is read to the participant: "This is the

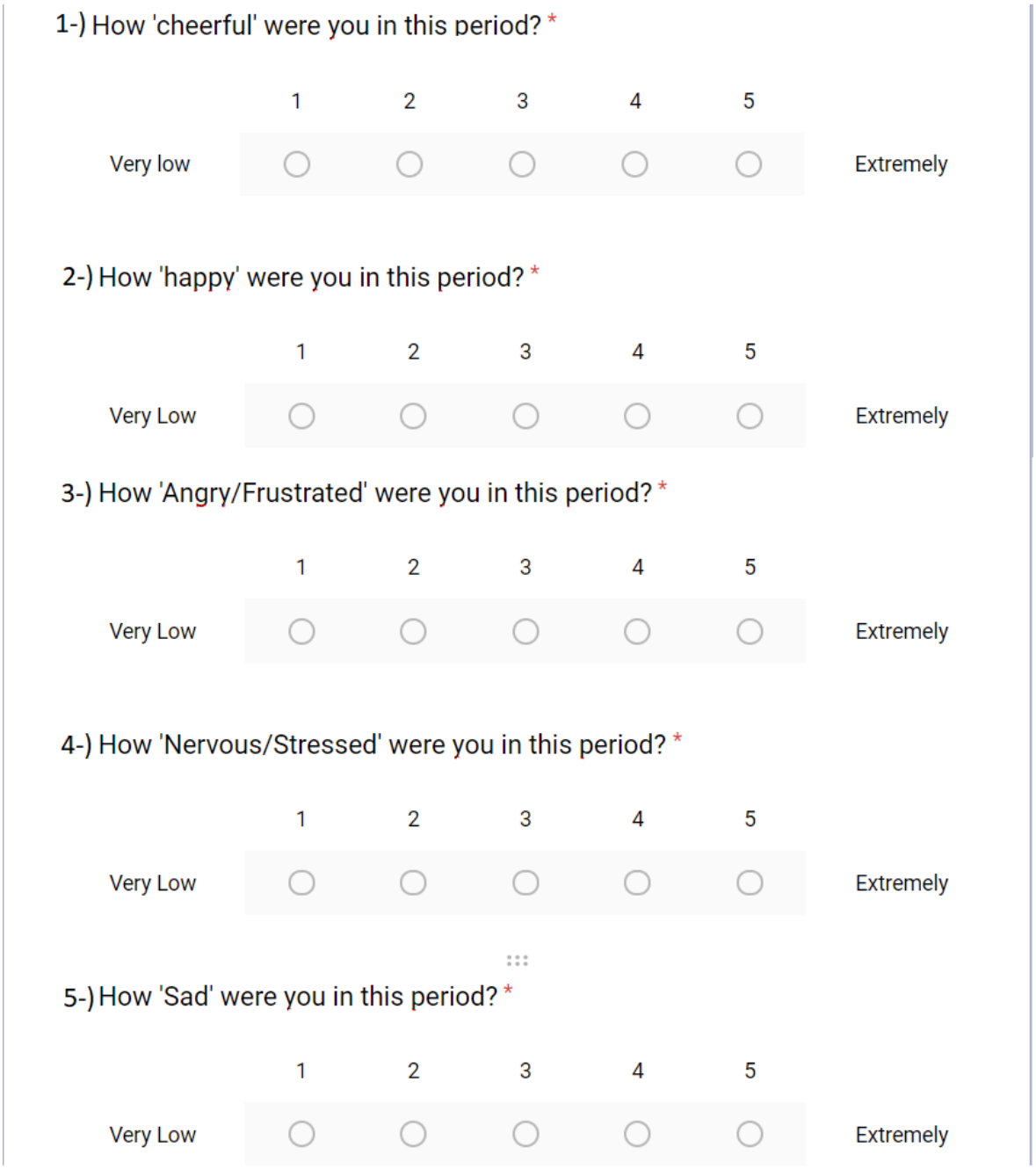

Figure 5.2: The PSS-5 questionnaire used in the experiment.

speech preparation portion of the task; you are to mentally prepare a five-minute speech describing why you study [name of the degree that the participant studies/studied] and why you would be a good candidate for your ideal job. Your speech will be videotaped and reviewed by the psychologists that we conduct the research with. You have five minutes to prepare and your time begins now."

(iii) The participant prepares his/her speech. There should be a digital timer in the room set to five minutes. Interviewers should leave the room.

(iv) This is read to the participant at the end of speech preparation period: "This is the speech portion of the task. You are to deliver a speech describing why you study [name of the degree that the participant studies/studied] and why you would be a good candidate for your ideal job. You should speak for the entire five-minute time period. Your time begins now.". Interviewers should start the recording of the camera.

(v) TSST speech performance period: If the participant stops during this period, interviewers allow him/her to stay silent for around 20 seconds and then prompt: "You still have time remaining."

(vi) At the end of 2 minutes, the interviewers say: "We want you to pursue the rest of your speech in English".

(vii) At the end of 2.5 minutes, if the participant does not attempt to reply to both questions, interviewers prompt the participant to answer the other question.

(viii) At the end of the speech performance period, the communication between interviewers and the participant resumes in Turkish. Interviewers reset the timer to 5 minutes and read the following to the participant: "During the final five-minute math portion of this task you will be asked to sequentially subtract the number 13 from 1,022. You will verbally report your answers aloud, and be asked to start over from 1,022 if a mistake is made. Your time begins now." If the participant makes any mistake, the interviewer says the following: "That is incorrect, please start over from 1,022." (Figure 5.5)

(ix) Participant fills out the PSS-5 form.

5.2.1.4. Post-stress Recovery Measurements.

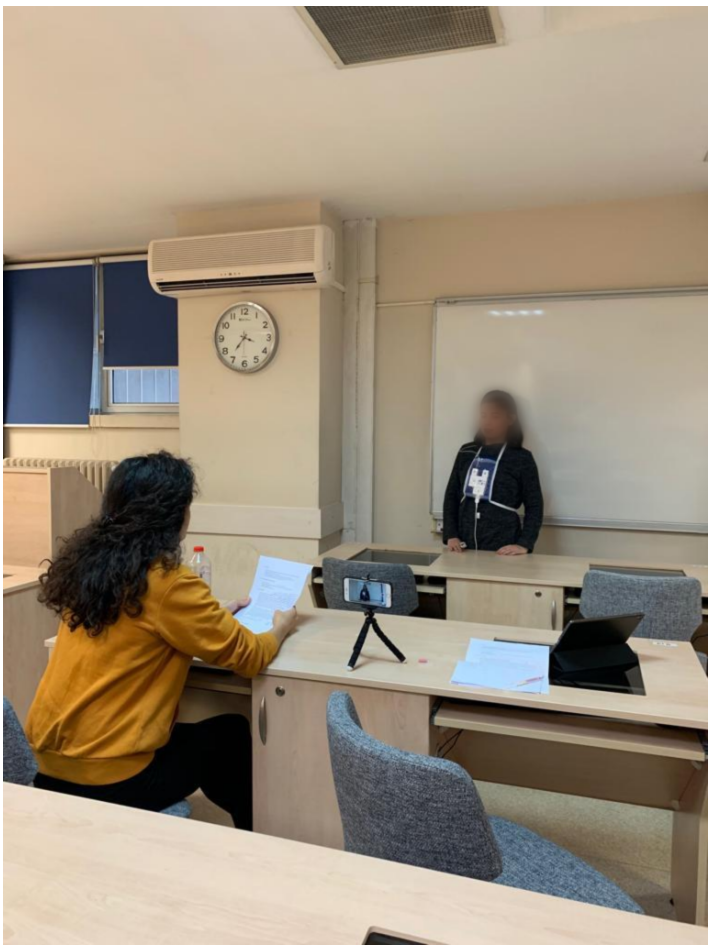

Figure 5.3: An example scene from the TSST phase in our experiment. The participant is presenting at this moment in front of the neutral experimenter.

(i) Participants are directed to the couch as a relaxing place.

(ii) The participant wears an Apple Watch given to him/her at this stage, following the breathing exercise built in the Apple Watch for 1 minute and then following a mindfulness video, for 4 remaining minutes, on a comfortable couch (Figure: 5.4), sitting or lying as the participant prefers.

(iii) Interviewers should leave the room after giving the Apple Watch.

(iv) At the end of five minutes long recovery period, interviewers return the room and participant fills out the PSS-5 form.

5.2.2. Results

We used our system described in Chapter 4 to detect and remove artifacts and extract features. In this laboratory experiment, our aim is to differentiate between the stress and baseline states with high accuracy. In this section, we investigated the performance of our stress detection scheme in two different manners. The first one is using the known context as the ground truth. We enumerated the different states as 1: baseline, 2: TSST (stress). We further provide these labels as classes to the machine learning algorithm. The second way is to use the perceived stress levels collected from self-reports as the ground truth. In order to measure the perceived stress levels, we collected PSS-5 which is appropriate for ambulatory environments. For these five questions, positive emotions (happy and cheerful) are evaluated inversely. In other words, if a participant states 6: extremely happy from 1-6, it is evaluated as 1 because happiness and cheerfulness are inversely proportional to stress levels. On the other hand, anger, sadness and frustration are evaluated proportionally when calculating the score (see Equation 5.1).

$$\text{PercStress} = (7\text{-H}_i) + (7\text{-}C_i) + A_i + S_i + F_i \quad (5.1)$$

where H_i: Happiness score, C_i: Cheerfulness score, A_i: Anger score, S_i: Sadness score and F_i: Frustration score. Individual scores on the PSS can range from 0 to 30 with higher scores indicating higher perceived stress. Scores ranging from 0-15 would be considered low stress and scores ranging from 15-30 would be considered high perceived stress. Furthermore, we have class imbalance problem in the PSS scores. We overcome this problem by deleting the extra samples of the majority class to avoid overfitting. The performance of our system is presented in Table 5.1:

Table 5.1: Stress Detection Accuracies with Different ML Algorithms - 2 class classification. On the left side, stress recognition results which are only using HRV and EDA signals are presented. On the right side, context information with accelerometer data is also added.

Algorithm	**w.r.t. Perceived Stress**			**w.r.t. Known Context**		
	HR	**EDA**	**HRV + EDA**	**HR**	**EDA**	**HRV + EDA**
Random Forest	83.3	86.7	84.9	57.8	66.7	80.39
kNN	**94.4**	86.4	89.7	68.6	73.8	77.1
SVM	88.9	77.3	**92.3**	74.3	73.8	77.1
LDA	**94.4**	72.7	89.7	68.6	76.2	80
Multilayer Perception	83.3	77.3	87.2	62.9	76.2	82.9

5.2.3. Interpretation

We applied feature selection together with machine learning models to avoid overfitting. We successfully differentate stress with baseline states as seen in Table 5.1 (with a maximum of 94.4 %). Perceived stress level detection accuracies are always higher than physiological stress level detection in the known context because participants may experience different stress levels than the expected level of the context. Some participants may experience lower stress in the TSST while preparing presentation, presenting in a foreign language or counting tasks. This proves that using the ground truth as known context labels or perceived stress labels have a significant influence on the performance of the system. Combination of the two pyhsiological signals always achieves higher accuracy than the modality with minimum accuracy alone. However, it is not the maximum accuracy at all conditions. We could state that it has observable effect on some conditions. Overall, we successfully differentiated between baseline and stress states in the laboratory environment.

Figure 5.4: The physiological signals of the baseline state is recorded in this couch. Subjects are asked to read magazines about car, design, sports during 10 minutes.

6. STRESS DETECTION IN A SEMI-RESTRICTED REAL-LIFE ENVIRONMENT

In real-life, stress detection performance is always lower than laboratory environments due to reasons mentioned in Chapter 2. After detecting the stress in the laboratory environment successfully, we took a step into the real life. In this chapter, we made experiments on semi-restricted environments and we proposed some methods to increase the performance of real-life stress detection systems. Participants can move freely in these environments however they are part of a real-life event that has a pre-arranged program. In the first part, we examined the effect of the personal stress-level clustering and decision-level smoothing on the data we collected from public school teachers in the ILKYAR event. We further examined the optimum stress detection resolution, effect of different self-reports and different number of classes. In the second part, we examined the effect of adding contextual information to our real life stress detection systems.

6.1. Personal stress-level clustering and decision-level smoothing to enhance the performance of ambulatory stress detection with smartwatches

Stress detection studies have started in the laboratory environment and then the direction of this research shifted towards real-life ambulatory environments. However, most of these studies only distinguish between stressed and relaxed states, which is not representative of the vast array of possible nuances that define the continuum from relaxation to the highest stress levels. Furthermore, cognitive load state also frequently occurs, especially in work environments and it should be added as a different class from stress to increase the stress detection resolution for more realistic results. An important research issue is that since everyone has specific stress responses, person-independent models generally have low accuracies for finding stress levels. If the data of individuals are used for creating models, it might be insufficient or requires long-term data col-

lection. A hybrid approach for clustering people by using their stress responses might solve the low accuracy problem without requiring long term data collection from individuals. Furthermore, after the machine learning algorithm assigns labels to frames, a decision level smoothing technique could increase the performance of stress detection systems. As an example, for 1-minute frames, it is not logical for a person to have a stressed frame followed by a relaxed state and one stressed frame after that. The middle frame, in this case, could be regarded as an error and smoothed.

After detecting the stress level of individuals, researchers should recover from the stressed state to the baseline state. Stress management mechanisms should be applied to achieve this recovery. Traditional yoga and meditation are ancient methods for stress alleviation. However, they could not be employed in office environments during daily work routines. As a consequence, mobile apps or indoor techniques should be applied to manage stress in the daily life without interrupting daily routines. The benefit of these indoor techniques are not examined comprehensively, but a limited number of studies have commenced [58]. A close examination of the literature permits to identify that most of the stress recognition studies do not offer relaxation methods after detecting stress which is needed to recover the subjects to their baseline states.

In this section, we improved our multi-level stress detection system which uses unobtrusive smart wearable devices. To the best of our knowledge, this study [149] is the first one to propose personal stress level clustering and decision-level smoothing to enhance person-independent stress detection models by using smartwatches. We first preprocessed the signals to clear the artifacts caused by unconstrained real-life motions. We further extracted discriminative features from the selected physiological signals. Lastly, machine learning algorithms are applied to classify different stress levels. To test our algorithms, collected physiological data in the ILKYAR summer school seminars took place at Bogazici University. We induced stress on primary school teachers following a cognitive load session in this event and then apply a stress management technique to recover them to their baseline states. Baseline signals and self-reports from these teachers are also collected at the beginning of this event. Our decision level smoothing and stress level clustering methods are applied after the classification algorithms are applied. Guided mindfulness was further used to alleviate the stress levels.

Our research addresses four original research issues in this chapter:

- The effect of applying decision level smoothing and decision-making mechanisms on system performance.
- The performance evaluation of person specific, clustered according to the baseline stress levels (hybrid) and person-independent models.
- The effect of different ground truth surveys (NASA-TLX and a more suitable, less time consuming version of it for everyday stress detection) on classification accuracies,
- Application of a guided mindfulness technique and measuring its success with smartwatch based physiological signals for reducing stress levels.

6.1.1. Experiment Design

6.1.1.1. Event Description: ILKYAR Summer School Seminars for Teachers. Every year, teachers from public schools of different cities in Turkey are gathered and participate in seminars which are given by university lecturers. Thirty-two teachers participated in the ILKYAR summer school. A seminar session took approximately three hours. We collected data during this session with different wrist-worn wearable devices which are the combination of Samsung Gear S2, Samsung Gear S and Empatica E4.

We first described the general outline of the study and delivered informed consent forms to the teachers. The experiment started after volunteering participants signed these forms. All participants wore the wrist-worn wearables and turned on the devices for data collection. At the beginning of the event, Pittsburgh Sleep Quality Index (PSQI), General Wellbeing Index (GWBI), WHO-five Well-being Index (WHO-5) Well Being and Perceived Stress Scale (PSS) baseline questionnaires were collected. We explained these questionnaires to the participants and baseline signals are recorded afterward.

Following the baseline session, the lecture about the research in the Computer Engineering Department was briefly given. The length of the lecture session was about 45 minutes. We gave a break after the lecture which was about 10 minutes. When they returned, we told them there is going to be an important exam and we measure

their performance. The exam consisted of arithmetic tasks inspired by Trier Mental Stress Test. The exam lasted for 20 minutes. The last part of the experiment was the recovery session. They listened to some relaxation music, they are told to take deep breathes and think of their positive memories. In this way, we tried to reduce their stress levels with the method inspired by the HeartMath app [5]. After all sessions (lecture, exam and recovery), raw NASA-TLX questionnaires are collected from the participants about the sessions. The procedure of the experiment is shown with the chronological order in Figure 6.1:

6.1.1.2. Participants and Apparatus. All participants are public school teachers within the age range of 25 to 40. There were ten female and 22 male participants. For the wearable devices used in the experiment, we searched the market with certain criteria. The devices should be non-obtrusive, provide the raw physiological data (with official SDK), give the Inter-Beat-Interval (IBI) from HRV. IBI is used for heart rate variability (HRV) measurement. HRV is an important indicator of stress. An optional feature would be providing EDA data. EDA is another important physiological signal for especially arousal detection. Empatica E4 satisfied all our expectations. We also investigated the Samsung branded Gear S1, S2, and S3 smartwatches. However, with the S3, Samsung stopped providing IBI intervals with its original SDK. Previous Samsung smartwatches (S1 and S2) give access to IBI raw data. Microsoft Band 2 also has the two signals. However, raw data access is not provided with the original SDK. It is removed from the official website. Apple smartwatches also do not provide raw data. We further evaluated the Polar chest band. However, it could be considered as an obtrusive device for real-life settings. Therefore, Empatica E4, Samsung S1 and Samsung S2 off-the-shelf devices were selected for the experiment.

When the battery lives of the devices were compared, Empatica E4 outperformed Samsung S smartwatches. It collects data for approximately two days. On the other hand, Samsung smartwatches can collect data for approximately 4 hours when all sensors are active. However, it is important to note that Empatica E4 is developed for research purposes and it is more expensive than Samsung commercial smartwatches. Samsung devices provide data via the Bluetooth connection. Conversely, Empatica E4

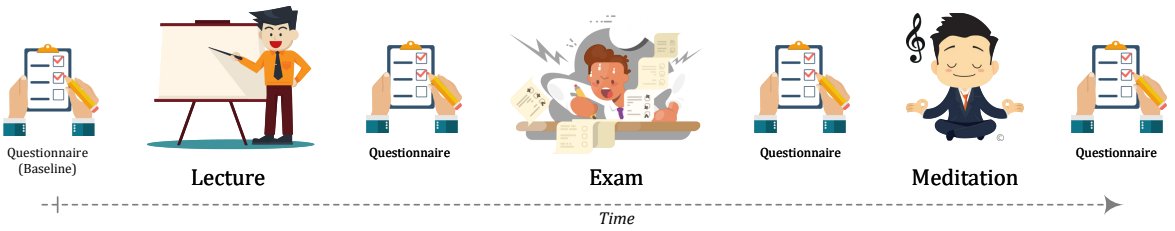

Figure 6.1: The timeline of the event is demonstrated. After each session, self reports were collected. We further took the baseline questionnaires at the beginning of the event.

has cloud support and data could be downloaded from the cloud server. Data acquired by both devices are available in CSV format.

As far as sensors are concerned, Empatica E4 devices have four sensors for measuring the acceleration, photoplethysmography, electrodermal activity and the skin temperature while Samsung Gear watches lack EDA and skin temperature sensors but instead they are equipped with Gyro and Barometer sensors. In this study, we used PPG, EDA and ACC sensors.

6.1.1.3. Ethics. The procedure of the methodology used in this study is approved by the Institutional Review Board for Research with Human Subjects of Bogazici University with the approval number 2018/16. Each subject signed a consent form explains the procedure of the experiment and its aims and implications to both the society and the subject before the data recording begins. We further described the procedure verbally to the participants. The recorded data are stored anonymously.

6.1.2. Experimental Results and Discussion

In this section, we examined two different issues. We conducted an experiment that consists of lecture, exam and recovery sessions. We first investigated the success of our experiment in creating three different psychological states by using the collected raw NASA-TLX self-reports. T-test was applied to measure the separability of these sessions from each other. We further applied the same procedure after modifying raw

Table 6.1: T-test results for Session Tuple Comparisons of Perceived Workload using RAW-TLX.

Session Tuple	**Test statistic**	**p-value**
Recovery - Exam	-13.073	$1.869e^{-14}$
Recovery - Lecture	-3.5886	0.0005645
Lecture - Exam	-9.723	$3.121e^{-11}$

NASA-TLX to demonstrate the separability when the modified version is used. By doing that, we showed that our experiment creates three different psychological states and raw NASA-TLX and modified version can show the difference between these states.

6.1.2.1. Clustering of Workload and Context. In this study, we examined the workload in different contexts for our experiment. As mentioned above, NASA-TLX is a tool which can be used to evaluate the perceived workload. By examining self-reports collected from participants after lecture, exam and recovery sessions, the success of inducing different workloads for these sessions is investigated. The t-test in the R programming language is used to measure whether these sessions are different or not in terms of the perceived workload. The paired t-test is used to evaluate the separability of each session. The degree of freedom is 31 (Since N=32 and N-1=31). We applied the variance test to each session tuple, we could not identify equal variance in any of the session tuples. Thus, we selected the variance as unequal. We used 95% confidence intervals. The t-test results are provided in Table 6.1. For all tuples, the null hypothesis stating that recovery is greater than or equal to lecture, the lecture is greater than or equal to exam and recovery is greater than or equal to exam sessions are rejected. The following p-values and test statistics are provided in Table 6.1. The perceived workload levels of participants for the exam, lecture and recovery sessions are observed to be significantly different.

From low to high perceived workloads, it can be sorted as the following: *Recovery* < *Lecture* < *Exam*. The boxplot for session tuples and all sessions are provided in Figure 6.2.

Table 6.2: Paired T-test results for Session Tuple Comparisons of Perceived Stress using Frustration score.

Session Tuple	**Test statistic**	**p-value**
Recovery - Exam	-8.3929	$1.762e^{-9}$
Recovery - Lecture	-2.6391	0.01289
Lecture - Exam	-6.7003	$1.702e^{-7}$

6.1.2.2. Clustering of Survey Data and Context after NASA-TLX modification. After we modified the NASA-TLX for measuring perceived stress levels, we further examined the self-reports in different sessions. In other words, if our stress induction, cognitive load induction and recovery sessions are successful in terms of creating different perceived stress levels, self-reports for each session should be separable. The same methods with Section 6.1.2.1 are applied for perceived stress levels. The results are provided in Table 6.2. The perceived stress levels of the t-test are sorted from low to high as *Recovery* < *Lecture* < *Exam*. Recovery session is designed as lower/mild stress and lecture also creates a cognitive load which is close to mild / low stress tasks in terms of perceived stress. The similarity between these sessions is demonstrated with the t-test. The boxplot for session tuples and all sessions are provided in Figure 6.2.

For all tuples, the null hypothesis states the above comparisons of Section 6.1.2.2 are rejected for perceived stress. The following p-values and test statistics are provided in Table 6.2. The perceived stress levels of participants are also determined to be significantly different as in Section 6.1.2.1. They can be sorted as the following, *Recovery* < *Exam* < *Lecture*. The boxplot for session tuples and all sessions are provided in Figure 6.2. As in the case of perceived workload, all of the mean value of the distributions are different, however in terms of distribution, recovery and lecture sessions are the most similar tuples. The recovery session is designed as lower/mild stress and lecture also creates a cognitive load which is close to mild / low stress tasks in terms of perceived stress. The similarity between these sessions is expected.

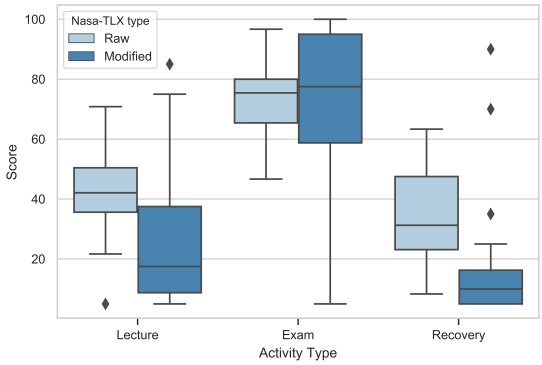

Figure 6.2: The box plots of the raw and modified Nasa Task Load Index (NASA-TLX) self-report scores during recovery, examination and lecture sessions.

6.1.2.3. Clustering of Participants with their Baseline Surveys. Our 32 participants filled the Perceived Stress Scale – 14 (PSS-14) questionnaires at the beginning of the event. We calculated the scores of each individual. This questionnaire consists of questions regarding the stress level felt during the last month before the experiment. By using these scores, we clustered the participants into low stress, medium stress and high stress classes. These classes demonstrate the condition of participants before the experiment. The top 33 percentile is assigned to high stress, low 33 percentile is assigned to low stress and the remaining 34 percentile is assigned to medium stress classes. We used these clusters to develop specific models in the next section.

6.1.2.4. Effect of Modalities to Stress Detection Performance using Known Context. In this study, we investigated the effect of two different modalities on classification accuracies. We further examined the effect of combining these two modalities. Heart rate variability (HRV) and electrodermal activity (EDA) are among the most discriminative physiological signals for stress level detection. HRV achieved higher classification accuracies than EDA with all the classifiers. From this result, we can deduce that HRV is a more discriminative physiological signal than EDA. Another important finding was the combination of these two signals achieves higher stress level classification accuracies than these signals alone in most cases (see Table 6.3).

Table 6.3: Effect of number of different modalities and combination of them on the system performance. Note that number of classes are fixed at 2 (stressed and recovery) and window size is 60 seconds.

Algorithm	**Accuracy (%)**		
	HR	**EDA**	**Combined**
MLP	76.37	62.78	79.88
RF	82.70	81.82	85.63
kNN	73.87	73.29	74.13
LDA	74.15	61.36	68.39
SVM	79.54	66.19	68.96

6.1.2.5. Perceived Stress, Perceived Workload and Physiological Stress Level Classification.

We measured physiological stress and perceived stress separately in this study. Normally, it is expected that when individuals exhibit signs of high physiological stress, they also have high perceived stress levels. However, this might not be the case in some situations [150]. Therefore, the two stress types are examined independently. The ground truth in physiological stress was selected as the stress level of known context. The perceived workload levels are measured with the answers from the whole NASA-TLX questionnaire. In the perceived stress case, we used the stress level obtained from the modified NASA-TLX self-report answers. The answer scale collected in lecture, exam and recovery sessions is divided into three levels. The low stress level is between 0 and 30 in NASA-TLX self-report (scaled from 0-100), the medium stress level is assigned if the answer is between 35-75 and the high stress level is assigned if the self-report is equal or above 80. We first examine the performance of modified the NASA-TLX questionnaire for measuring perceived stress levels. The correlation with known context stress labels increased from 0.32 to 0.54 with the modified NASA-TLX when compared with the whole NASA-TLX. This justifies the modification for better perceived stress level measurement. Since we collected answers in lecture, exam and recovery sessions, we observed that low and mild stress levels are seen in lecture and recovery sessions. Thus we combined recovery and lecture into the low/mild stress class. In the analysis of self-reports in Section 6.1.2.2, the two classes have closer behavior as

expected. The exam session is determined as the high stress class. The same procedure is also applied to the known context ground truth labels and recovery and lecture sessions are merged into one low / mild stress context. The exam session is labeled as the high stress context. The perceived workload, perceived stress and physiological stress detection accuracy results are provided in Tables 6.4 and 6.5. The perceived workload results always have the best performances with all classifiers. The workload level (physical, mental, temporal demand, performance, effort and frustration) could be easily differentiated by individuals with self-reports. The prediction accuracy of the perceived stress is weaker than the perceived workload. Individuals may not always perceive the true stress levels or express them with self-reports [150] and these issues might result in the appearance of lower system performance. The lowest classification results are obtained when measuring physiological stress levels with known context ground truth labels. Although we prove the success of our stressors from the self-reports, some participants could not be induced with the desired levels of stress or cognitive load. Our data collection did not take place in a laboratory. Instead, we recorded data in a real-life event. Thus, some participants may not be cognitively loaded in lectures or stressed in exams in a semi-controlled real-life event and this decreases the accuracy of stress detection accuracy with known context ground truth labels.

6.1.2.6. Personalized, Cluster-specific and Person-Independent Models. Since the stress reaction of individuals has a unique pattern, the ideal way to develop automatic stress detection models is to use the individual's data. However, in most cases, there is not enough personal data for developing this kind of model. Another way is to develop models from all collected data and apply this one model to all people. However, the accuracy of this model is expected to be lower than the personalized model because of the mentioned person-specific stress reactions. In this section, we offered a hybrid approach. As mentioned in Section 6.1.2.3, we clustered participants by using their baseline stress levels. By using self-report answers regarding the month before the experiment, we divided them into low stress, medium stress and high stress clusters. We then develop models for each cluster separately since we expect people in the same

Table 6.4: Physiological stress, perceived workload and stress detection accuracies, EDA signal, the number of distinguished classes is 2 (recovery - cognitive load (low / mild stress), exam (high stress)) and window size is 240 seconds. The NASA-TLX Perceived Workload and Perceived Stress scores are divided into three classes. Two classes (low and mild) in lecture and recovery sessions are combined in both known context and perceived workload and stress evaluation

Algorithm	**Accuracy with different Ground Truth Labels**		
	Physiological Stress	**Perceived Workload**	**Perceived Stress**
MLP	77.00	92.00	88.50
RF	82.00	94.50	90.00
kNN	78.50	93.00	89.50
LDA	78.00	92.00	89.50
SVM	77.00	91.00	88.00

Table 6.5: Physiological stress, perceived workload and stress detection accuracies, HRV signal, the number of distinguished classes is 2 (recovery - cognitive load (low / mild stress), exam (high stress)) and window size is 240 seconds. The NASA-TLX Perceived Workload and Perceived Stress scores are divided into three classes. Two classes (low and mild) in lecture and recovery sessions are combined in both known context and perceived workload and stress evaluation

Algorithm	**Accuracy with different Ground Truth Labels**		
	Physiological Stress	**Perceived Workload**	**Perceived Stress**
MLP	78.24	92.64	86.23
RF	84.19	94.52	89.51
kNN	79.96	92.33	85.29
LDA	78.56	94.21	88.42
SVM	78.87	94.52	88.73

Table 6.6: Effect of general, personalized and clustered models on system performance, EDA signal. Note that number of distinguished classes is 3 (relax, cognitively loaded, stressed) and window size is 120 seconds

Algorithm	**Models**		
	Personalized	**Clustered**	**General**
MLP	74.25	62.55	59.50
RF	74.45	70.57	64.46
kNN	71.50	66.69	57.44
LDA	70.18	58.91	58.88
SVM	75.04	63.64	63.64

Table 6.7: Effect of general, personalized and clustered models on system performance, HRV signal. Note that number of distinguished classes is 3 (relax, cognitively loaded, stressed) and window size is 120 seconds

Algorithm	**Models**		
	Personalized	**Clustered**	**General**
MLP	79.87	70.08	66.69
RF	81.16	74.87	71.57
kNN	76.17	67.73	59.37
LDA	78.08	67.96	66.80
SVM	74.94	70.82	66.26

cluster might have similar physiological reactions to our stimuli. As seen in Tables 6.6 and 6.7, person-independent models have the lowest stress classification accuracies, whereas personalized models obtained the best results with all classifiers. Our hybrid models have accuracies lower than personal and higher than person-independent models. When the data is not enough for personal models, our hybrid approach could be used to increase the performance of the system.

6.1.2.7. Effect of Stress Detection Interval and Resolution to Classification Accuracies.

Another important research issue we want to address is to find the optimal interval for

Table 6.8: Effect of stress resolution to stress detection accuracies. Number of distinguished classes is fixed at 2 (recovery, stressed), EDA signal

Algorithm	**Accuracy in different stress detection resolutions**			
	60s	**120s**	**240s**	**480s**
MLP	62.78	69.04	73.61	69.56
RF	81.82	75.00	75.00	52.17
kNN	73.29	73.80	72.22	56.52
LDA	61.36	59.52	66.67	65.21
SVM	66.19	69.04	76.38	73.91

Table 6.9: Effect of stress resolution to stress detection accuracies. Number of distinguished classes is fixed at 2 (recovery, stressed), HRV signal

Algorithm	**Accuracy in different stress detection resolutions**			
	60s	**120s**	**240s**	**480s**
MLP	76.37	76.56	75.22	60.24
RF	82.7	83.79	78.76	73.49
kNN	73.87	76.95	77.43	61.44
LDA	79.54	79.68	81.42	71.08
SVM	74.15	76.75	72.56	65.06

stress detection studies. In other words, since stress reaction has certain physiological characteristics, there might be an optimum time interval that the stress level could be detected more easily from biofeedbacks of individuals. We carried out experiments with 60, 120, 240 and 480 second intervals. In 9 out of 10 experiments (5 classifiers with EDA and 5 classifiers with HR), the best accuracies are found with 120 - 240 second intervals (see Tables 6.8 and 6.9). However, researchers should also take into account the employed classifier algorithm when determining the optimal interval for stress detection.

6.1.2.8. Effect of Number of Recognized Stress Levels to Classification Accuracies. The effect of the number of recognized stress level classes to the accuracies is also exam-

Table 6.10: Effect of number of stress levels to stress detection accuracies. Note that window size is fixed to 120 seconds and enumerated classes are as follows : 1 (cognitive load - lecture), 2 (relax), 3 (stressed- exam), 4 (recovery- stress management), EDA signal

Algorithm	**Classes**				
	1 vs 2	**1 vs 3**	**3 vs 4**	**2 vs 3**	**1 vs 2 vs 3**
MLP	73.94	49.52	69.04	77.63	59.50
RF	78.16	72.58	75.00	83.42	64.46
kNN	73.16	64.42	73.81	78.16	57.44
LDA	71.31	50.96	61.31	76.05	58.88
SVM	73.42	55.77	59.52	76.84	59.92

ined. As mentioned, our experiment has four different sessions: baseline, lecture, exam and recovery with guided mindfulness. It is assumed that the lecture will induce a cognitive load, the exam will induce stress on the participants. We tried to bring them back to their baseline states by applying guided mindfulness with a relaxing music. We experimented with different tuples from these four sessions. Lastly, we examined the performance of our system on three class classification.

Three classes are selected as stressed, baseline and cognitive load. Lecture vs. stress is the most difficult to distinguish session tuple with both types of signals, as seen in Tables 6.10 and 6.11. This is because of the similarity of physiological reactions of cognitive load and stress behaviors. Exam vs. recovery, lecture vs. baseline and exam vs. baseline session tuples could be differentiated with relatively higher accuracies with both modalities. Another important finding is that exam and recovery sessions can be distinguished with accuracies similar to (and sometimes higher than) lecture vs. baseline and exam vs. baseline session tuples. This shows that our recovery session successfully alleviates the stress of participants and decrease their stress level. The three class classification accuracy is similar to lecture - exam tuple but less than other tuples. The difficulty in distinguishing these two sessions is also interfering with the

Table 6.11: Effect of number of stress levels to stress detection accuracies. Note that window size is fixed to 120 seconds and enumerated classes are as follows : 1 (cognitive load - lecture), 2 (relax), 3 (stressed- exam), 4 (recovery- stress management), HRV signal

Algorithm	**Classes**				
	1 vs 2	**1 vs 3**	**3 vs 4**	**2 vs 3**	**1 vs 2 vs 3**
MLP	82.25	65.83	76.17	74.43	66.69
RF	83.53	73.35	84.76	80.33	71.57
kNN	75.26	68.03	76.95	73.37	59.37
LDA	82.93	65.83	76.75	75.79	66.26
SVM	83.15	69.43	79.68	77.91	66.8

performance of the three class classification system. However, even with these three classes, we have similar accuracies with reported systems differentiating stress from cognitive load in laboratory settings [75], [45].

6.1.2.9. Increasing Accuracies with Decision Level Smoothing. The classification errors can be corrected by examining the results from a high level perspective. We examined our decisions for 60 second intervals with this perspective. We search for cases which are not likely to occur when logically evaluated and add some rules on top of our system. Our rule was correcting changes with unusually high frequency. In other words, if a subject is found out to be stressed in one window, not stressed in the consecutive one and stressed again in the next window; we determined this case as highly unlikely and an error of our system. We applied our logic on top of our system automatically. The maximum accuracy of our system increase from approximately 82% to 92% with EDA and HRV signals (see Tables 6.12 and 6.13). The performances of all classifiers increase significantly with decision level smoothing.

Table 6.12: High Level Accuracy Calculation and Decision Level Smoothing Accuracy Results with EDA signal. Note that number of classes is fixed at 2 (stressed and recovery) and window size is 60 seconds.

Algorithm	**Decision level smoothing and high level Accuracy calc.**		
	EDA	**Decision Level Smoothing**	**HighLevelAcC**
MLP	62.78	70.17	81.25
RF	81.82	92.89	100.00
kNN	73.29	86.07	93.75
LDA	61.36	62.78	68.75
SVM	66.19	70.17	81.25

6.1.2.10. High level accuracy calculation for stress detection. We divide all experiment data into 60 second windows and test each window separately when calculating the accuracy. However, in real-life, detecting stress for particular sessions and time intervals might gain more importance. Thus, we propose a different stress level detection accuracy calculation. For all sessions, we labeled all small windows and applied majority voting afterwards for N (number of windows in a session) consecutive intervals in a sliding window fashion. To put it another way, our system labels sessions by the majority of labels of consecutive small windows. We called this method as "high-level accuracy calculation". In this way, the accuracy for 2-class stress level detection goes up to 94.44% with HRV signal and 100% with EDA signal (see Tables 6.12 and 6.13). If the aim is to identify stress levels in specific sessions, high-level accuracy calculation could be used to increase the performance.

Table 6.13: High Level Accuracy Calculation and Decision Level Smoothing Accuracy Results with HRV signal. Note that number of classes is fixed at 2 (stressed and recovery) and window size is 60 seconds.

Algorithm	**Decision level smoothing and high level Accuracy calc.**		
	HR	**Decision Level Smoothing**	**HighLevelAcC**
MLP	76.36	88.18	94.44
RF	82.70	92.12	90.74
kNN	73.87	87.32	90.74
LDA	74.15	85.30	85.30
SVM	79.54	89.62	87.03

6.1.3. Interpretation

We proposed new models and methods for improving multi-level real-life stress detection systems using unobtrusive off-the-shelf smartwatches and smart bands. We tested our algorithms in real-life settings which include baseline, cognitive load, stress and recovery sessions of 32 participants in a summer school. First, the effect of our hybrid personal stress level clustering was examined. In the person-independent model, the data of all participants are divided into training and test parts. The personalized model uses the data of each participant for developing a model. On the other hand, in our new hybrid model, we first cluster the participants into low, medium and high stress levels by examining their baseline self-reports. In this method, we develop specific models for each cluster. The personalized model has the highest and the person-independent model has the lowest accuracy. Our hybrid model has accuracies in between. It could be used in cases where there is not enough data of participants to develop personalized models. In these situations, our hybrid models will increase the accuracy of the system when compared with person-independent models.

Furthermore, the perceived stress, workload and physiological stress were investigated. We started with successfully classifying perceived workload level (3-class) using

NASA-TLX. The minimum classification accuracy is 91% and the maximum accuracy is 94.52% for 3-classes. After that, we used the modified version of NASA-TLX to measure the perceived stress levels and compared with physiological stress levels. Modifying the NASA-TLX increased the correlation with known context labels from 0.32 to 0.54. When the performance of 3-level physiological and perceived stress detection classification accuracies are compared, perceived stress levels are always detected more successfully with all classifiers. Some participants might feel a different stress level than known context labels. This might decrease the performance of the physiological stress level detection system.

We further tested a decision-level smoothing method using the fact that stress levels of participants do not oscillate instantaneously. Our maximum accuracies with using a single modality are around 80% in 2-class classification (81.82% with EDA, 82.70%). To increase the performance of the system, results were examined with a high-level perspective. We applied an additional logical rule on top of our classifier to correct some misclassifications. With decision level smoothing, the classification accuracies increased to around 90 % with both modalities (92.89% maximum). We further developed a session-based stress classifier. The majority voting among windows of every session was applied to decide the assigned class. We obtained a maximum accuracy of 94.44% with HR, 100% with EDA signals. When the stress level of a session is needed to be calculated, this method could be applied.

We improved our platform independent stress level detection system which works with off-the-shelf smartwatches and smart bands. We tested our algorithms in a real-life setting and obtained successful classification accuracies. As mentioned, personal stress level clustering and decision-level smoothing increased the performance of our system considerably. We also applied stress alleviation methods and proved their effectiveness. Our system could be easily adapted to the daily life of individuals without interrupting their routines. The study has limitations that should be mentioned. With regard to the measurement, we have initially included NASA-TLX which is a cognitive workload scale. In order to measure the perceived stress levels, as explained in the methods section, we have selected the frustration subscale which is the most representative for that purpose. Nevertheless, future studies specifically focusing on the perceived stress, could

better include specific scales such as Daily Stress Inventory [151], Daily Experiences Survey [152] or Perceived Stress Scale [148].

6.2. Stress Level Monitoring using Smart-Bands in the Light of Contextual Information and Management with Yoga and Mindfulness in Real-Life Events

In this section, we examined the effect of contextual information and stress management techniques on the data collected from 16 PhD students who are EU H2020 Affectech project partners. Our main contributions are:

- The effect of the ancient and contemporary relaxation methods in the context of stress management
- Measuring the daily perceived stress levels (DPSL) and session-based perceived stress levels (SBPSL) by using only physiological signals and combining the contextual information (weather, physical activity level and activity type) with them
- The comparison of mobile mindfulness, traditional mindfulness and yoga methods in the context of stress management for relieving stress
- Developing a prescreening tool for long-term perceived stress levels (LTPSL)
- Application of emotion regulation model in the context of stress management and measuring the physiological component with smart bands.

6.2.1. Experiment Design

6.2.1.1. Description of the Data Collection Procedure. Evaluation of our proposed stress level monitoring mechanism in real-life settings was carried out by means of the analysis of the data collected in the eight days long AffecTech training event in Istanbul-Turkey. AffecTech is a Marie Skłodowska-Curie Innovative Training Network program funded by Horizon 2020 (H2020) framework established by the European Commission. The AffecTech project is an international collaborative research network that aims to develop and improve personal and person-specific and finally low-cost yet effective wearable health technologies to help the individuals who suffer from affective disorders like de-

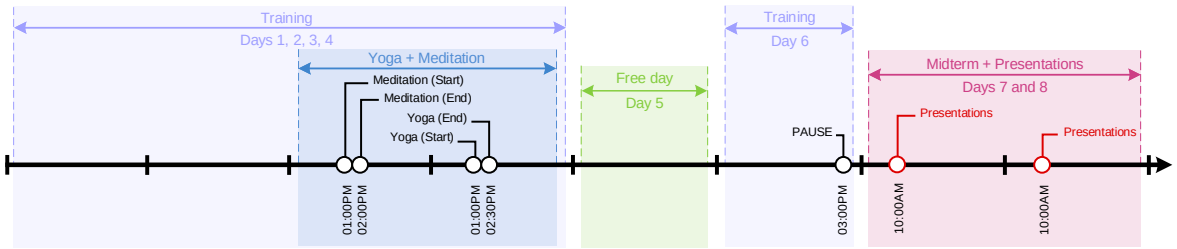

Figure 6.3: Time-line depicting eight days of the training event. Presentations, relaxations and lectures are highlighted.

pression, anxiety and bipolar disorder.

We collected physiological data of 16 participants for the study. Due to a fault on one of the Empatica E4 devices, one participant had to quit the study. 15 subjects completed all stages of the study successfully. All of the participants are Ph.D. students with different research and expertise fields within the associated disciplines. Participants are from different countries with diverse nationalities (Two from Iran, two from Spain, two from Italy, one from Argentina, one from Pakistan, one from China, one from Switzerland, one from Belarus, one from France, one from England, one from Barbados, one from Turkey and one from Bulgaria) gathered to participate in the training event. We emphasize the diverse nationality of the participants because different cultures can have different stress reactions and to the best of our knowledge, our study is the first one to analyze this type of diversified data with wearables. This event was held for eight consecutive days. Prior to the study, each participant received the 14 item version of the Perceived Stress Scale (PSS) [148]. This questionnaire is used as the baseline. The physiological signal data from Empatica E4 and session-based self-reports comprises of six questions of NASA-TLX [153] and daily perceived stress questionnaire (a combination of questions from [154], [155] and [156]) were collected during the event. The gender split is 6 females and 9 males. The average age of the participants is 28. A total of 2780 self-report questions (from 3 session-based and 1 daily questionnaire each day) were collected. The training week is concentrated on clearly defined training tasks and activities. In order to ensure that the fellows have developed the required target skills, knowledge, and values, an innovative set of design and implementation workshops and training programs are planned and implemented in multiple week-long series of informative workshops and presentations. Participants

got hands-on experience in installing, using wearables and the analysis of their sensor data. At the end of the training week, participants had to make a presentation about their previous works to two evaluators from the European Union where they received feedback about their progress.

During the eight days of training and presentations, psychophysiological data were collected from 16 participants during the training event from Empatica E4 smart band while they are awake. For studying the effects of emotion regulation on stress, yoga, guided mindfulness and mobile-based mindfulness sessions were held. The timeline of the event is shown in Figure 6.3.

6.2.1.2. Session-based Self-Report for Perceived Stress Measurement. The first collected self-report is the Frustration item of the raw NASA-TLX [157]. We asked the following question to the participants for each session:

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

We measured the session based perceived stress levels by using the first question. Likert scale 10 was used scaling the answer.

6.2.1.3. Daily Self-Report Questionnaire for Perceived Stress Measurement. The daily self-report questionnaire was comprised of six questions. Two questions were measuring rumination [156], the other two questions were measuring worry [154] and the last two questions were measuring reappraisal [155]. We selected rumination, worry and reappraisal questions from mentioned prominent questionnaires (Brief State Rumination Inventory [156], State-Reappraisal Inventory [155] and Penn State Worry Questionnaire [154]) because they are the main components for causing stress and they are linked to depression and anxiety [158]. The collected daily questionnaire is demonstrated below. By using this questionnaire, we measured the daily perceived stress levels of the participants. The response measure was the Likert scale with answers from 1 to 6.

Worry Question 1 My worries are overwhelming me

Worry Question 2 I know I should not worry about things, but I just cannot help it

Rumination Question 1 Right now, it is hard for me to shut off negative thoughts

Rumination Question 2 Right now, I am thinking: "why can't I handle things better?"

Reappraisal Question 1 I'm trying to think that things could be much worse

Reappraisal Question 2 I'm trying to think of positive aspects of the events

6.2.1.4. Ethics. The procedure used in this study is approved by the Institutional Review Board for Research with Human Subjects of Boğaziçi University with the approval number 2018/16. Prior to the data acquisition, each participant received a consent form that explains the experimental procedure and its benefits and implications to both the society and the subject. The procedure was also explained vocally to the subject. The data collection procedure and all of the interventions in this research fully meet the 1964 Declaration of Helsinki [159]. All of the data are stored anonymously.

6.2.1.5. Stress Management Scheme using Yoga and Mindfulness . During the real-life events, the stress level of the participants is assumed to increase day by day because they have a presentation where they have to report their work to the EU project evaluators on the last day. To manage their stress levels, we applied yoga and mindfulness sessions on two separate days. These sessions lasted approximately 1 hour. Regarding James Gross's Emotion Regulation model [7], we modified the situation and helped them reduce the thoughts of the end of the training presentation. They wore Empatica E4 wristbands during these sessions. In addition to the physiological signals coming from Empatica E4 wristbands, participants' blood pressure values are also recorded before and after the relaxation sessions to demonstrate that we successfully manage the stress levels by modifying the situation according to the Gross's model (see Figure 6.4).

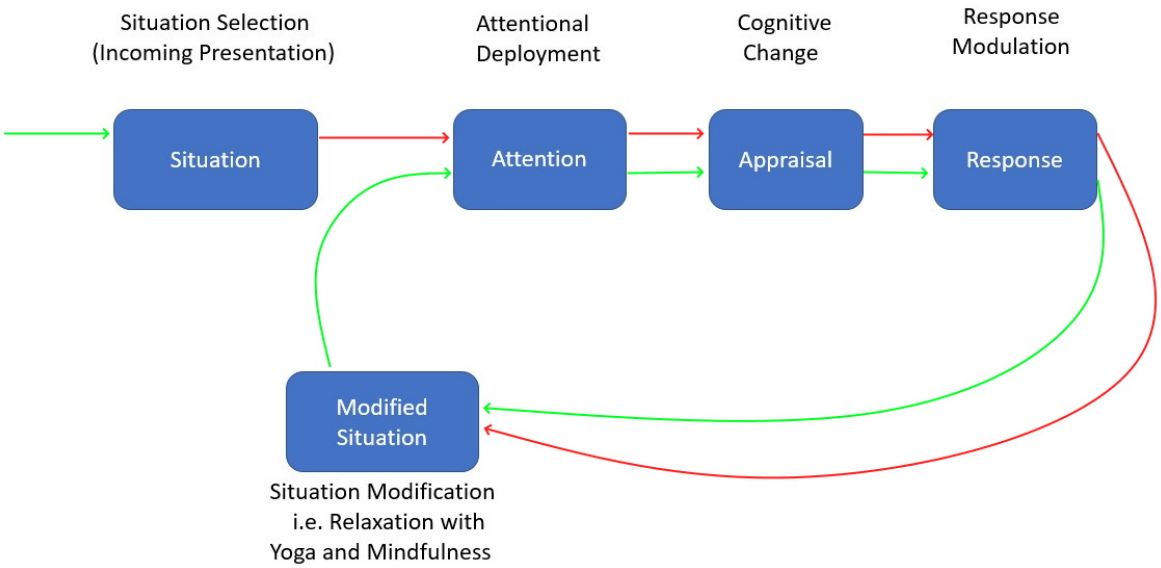

Figure 6.4: Application of James Gross's Emotion Regulation model [7] in the context of stress management.

6.2.2. Experimental Results and Discussion

The perceived stress of individuals was measured throughout the event with two types of self-reports. We further investigated the effect of varying contextual data on perceived stress level detection performance.

6.2.2.1. Measuring the Session-based Perceived Stress Levels. We collected the session-based self-reports during the event in every session. We further asked the participants to fill these self-reports in the evening in their free time while wearing Empatica E4 wristbands. 2-class and 3-class session-based perceived stress scores are presented in Tables 6.14 and 6.15. By using EDA and HRV signals, our system achieved a maximum 69.62% classification accuracy on 2-class and 53.15% 3-class classification. Another important finding is that HRV signal achieves higher detection accuracies with all algorithms. Since we collected self-reports during the training event and free time in the evening (17:00-22:00), our system suffered from the problems with perceived stress measurement in unconstrained real-life mentioned in [6]. The negative effect of stress on memory and subjectivity of self-reports are among the most prominent problems with self-reports and they contributed to the decrease in the performance of our

Table 6.14: 2-class perceived stress detection results by using the self-reports obtained from the frustration question. The scale of answers divided into two classes: low stress if the scale is less than 50, high stress otherwise.

Algorithm	**Accuracy**	
	EDA	**HRV**
MLP	50.00	55.69
RF	57.43	69.62
kNN	54.73	59.49
LDA	50.68	64.55
SVM	49.32	53.16

Table 6.15: 3-class perceived stress detection results by using the self-reports obtained from the frustration question. The scale of answers divided into three classes: low stress if scale is less than 35, medium stress if scale is between 35 and 70 and high stress is scale is higher than 70.

Algorithm	**Accuracy**	
	EDA	**HRV**
MLP	46.28	51.35
RF	36.57	47.74
kNN	43.42	53.15
LDA	45.14	51.35
SVM	45.71	51.35

system.

6.2.2.2. Pre-screening LTPSL by Evaluating Physiological Signals. Participants were asked to complete the PSS questionnaires as baseline surveys. PSS is used to measure the monthly stress levels of individuals. We divided the scores of the questionnaires into high and low-stress classes. If the scale is above 10 over 25, it is labeled as high stress and otherwise, it is labeled as low stress. We selected 10 as the threshold because it is the average score of all participants. When the physiological features collected for all sessions are used, we are able to predict the general stress level of a person successfully.

Table 6.16: Predicting the long term stress level (LSTL) from the physiological data collected from the event. LSTL is calculated from the PSS questionnaire regarding the last month before the event.

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

The HRV signal achieves approximately 80% accuracy, whereas the EDA signal has a maximum of 73.59% accuracy of finding the general stress level of a person in the last month (see Table 6.16). These results are promising because they demonstrated that physiological signals could be used for pre-screening long-term perceived stress levels and our system can advise users to see a psychologist or adopt stress relieving actions if it detects a high level of long-term stress by examining their physiological signals.

6.2.2.3. DPSL prediction by evaluating Rumination, Reappraisal and Worry Elements.

We further collected self-reports every day to measure stress by evaluating rumination, worry and reappraisal elements. These elements are selected because they contribute the high-stress levels most. Physiological signals are collected from the participants and daily perceived stress levels (DPSL) of individuals collected from self-reports are used as ground truth labels to these signals. The 2-class DPSL classification accuracies are presented in Table 6.17. We achieved a maximum of 68.85% classification accuracy which is similar to the reported performances in the literature.

6.2.2.4. Measuring Perceived Stress Levels in the Light of Contextual Information. To enhance the performance of our system, we made use of the contextual information. Activity related measures, weather information and known context (lecture, presentation, relaxation activities) are added to the physiological features to improve the

Table 6.17: Predicting 2-class daily perceived stress level (DPSL) from the physiological data collected from the event. DPSL is calculated from the questionnaire which is collected daily and composes of rumination, worry and reappraisal questions.

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

performance of the perceived stress level detection system.

6.2.2.5. Session-based perceived stress measurement with weather-related Context. The association of the effects of the changes in weather conditions with stress and mood has been demonstrated in the literature [160] [161]. We further investigated the effect of weather-related context into our perceived stress detection system. Air Temperature at 2 meters high above the surface (degrees Celsius), atmospheric pressures at weather station and sea level (millimeters of mercury), changes in atmospheric pressure over last three hours, relative humidity at a height of 2 meters above the earth's surface, wind speed, total cloud cover and amount of precipitation data were extracted for each session and added to the physiological data. Weather information is gathered from Windguru wind and weather forecasting website [162]. The weather-related context information increases performance of our system drastically as seen in Tables 6.18, 6.19. Air pressure, humidity, precipitation and cloud ratio are selected among the top features when used with EDA and HRV signals. These results demonstrated that weather has an impact on the perceived stress of individuals and provides additional information for a stress detection system.

6.2.2.6. Finding Perceived Stress Levels by Adding the Known Context. The known context of individuals can be also used for improving stress detection systems as men-

Table 6.18: Predicting the 3-class perceived stress level (PSL) from the HRV and the EDA data collected from the event. The PSL is calculated from the Frustration scale. Weather-related features for these sessions are added.

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

Table 6.19: Predicting the 2- class perceived stress level (PSL) from the HRV and EDA data collected from the event. The PSL is calculated from the Frustration scale. Weather-related features for these sessions are added.

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

Table 6.20: Predicting the 3- class perceived stress level (PSL) from the HRV and EDA data collected from the event. The PSL is calculated from the Frustration scale. Known context is added to the features.

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

tioned in Chapter 2. The EU training event has a lecture, presentation, relaxation and free (17:00-22:00) sessions. Since the context is unknown in free sessions, we did not use these sessions for this section. We enumerated the known context for these sessions as Relaxation:0, Lecture:1 and Presentation:2. We added these enumerated known context information to the physiological features to detect session-based perceived stress levels. As seen in Tables 6.20 and 6.21, adding the known context information increases the system performance 20-25%. These results demonstrate that the known context information is crucial for daily stress detection systems. We further investigated performance of the system by adding the data without context information. After 17:00, the training has finished and free time started. By adding Unknown Context:0 labels to that times, we recalculated the results as shown in Tables 6.23 and 6.22. Adding free times data and the label "Unknown Context" increased the performance of our system. This might be caused by the fact that the new data is relatively relaxed data and perceived self-reports might align with the newly added context label more than the other known context labels.

6.2.2.7. DPSL Detection with Physical Activity Related Contextual Information. Physical activity is known to reduce stress levels [163]. We selected three days of the training: one from the beginning (Day 2), one from the middle (D4) and one from the end (D8) and tried to differentiate daily perceived stress levels by using the HRV features.

Table 6.21: Predicting the 2- class perceived stress level (PSL) from the HRV and EDA data collected from the event. The PSL is calculated from the Frustration scale. Known context data for these sessions are added to the feature vector.

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

Table 6.22: Predicting the 3- class perceived stress level (PSL) from the HRV and EDA data collected from the event. The PSL is calculated from the Frustration scale. Known context is added to the features. Unknown context class is also added.

Algorithm	**Accuracy**			
	HRV	**HRV + known context**	**EDA**	**EDA + known context**
MLP	47.74	68.64	46.28	63.36
RF	51.35	62.37	36.57	65.35
kNN	53.15	65.83	43.42	62.04
LDA	51.35	65.26	45.14	64.69
SVM	51.35	67.1	45.71	67.32

Table 6.23: Predicting the 2- class perceived stress level (PSL) from the HRV and EDA data collected from the event. The PSL is calculated from the Frustration scale. Known context data for these sessions are added to the feature vector. Unknown context class is also added.

Algorithm	**Accuracy**			
	HRV	**HRV + known context**	**EDA**	**EDA + known context**
MLP	69.62	86.31	50.00	74.59
RF	55.69	88.68	57.43	74.92
kNN	59.49	89.21	54.73	72.61
LDA	53.16	77.89	50.68	76.56
SVM	64.55	82.63	49.32	77.55

We do not expect considerable changes in terms of physical activity stemming from the schedule. We further investigated the effect of physical activity related contextual features on the performance of the daily perceived stress detection system. For this purpose, we extracted stillness and step count features from the accelerometer signal by using the EDAExplorer [133]. Stillness measures the daily physical activity of an individual. The range of stillness is between 0 and 1. Step count is also calculated from the accelerometer signal. When these contextual features are added to the signal, our DPSL detection accuracies are increased considerably (see Table 6.24). These results show that physical activity related to the contextual features are also important for the daily perceived stress detection schemes.

6.2.2.8. Effectiveness of Yoga, Mindfulness and Mobile Mindfulness (Pause). We applied three different relaxation methods to manage the stress levels of individuals. In order to measure the effectiveness of each method, we examined how easily these physiological signals in the relaxation sessions can be separated from high stress presentations. If it can be separated from high stress levels with higher classification performance, it could be inferred that they are more successful at reducing stress. As seen in Tables 6.25 and 6.26, mobile mindfulness has lower success in reducing stress levels. Yoga has the highest classification performance with both HRV and EDA signals. HRV signal is

Table 6.24: Daily stress level differentiation accuracies by using the only HRV and with the addition of physical activity related context data (stillness and step count).

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

Table 6.25: The classification accuracy of the recovery sessions using stress management methods and stressful sessions using EDA.

Algorithm	**Accuracy**		
	Guided Mindfulness	**Yoga**	**Mobile Mindfulness**
MLP	65.71	78.57	75.00
RF	67.14	87.14	67.64
kNN	64.29	82.86	77.94
LDA	65.71	80.00	51.47
SVM	70.00	72.86	58.82

found to be more successful for differentiating stressful sessions from the relax sessions.

6.2.2.9. Evaluating the Performance of Yoga and Mindfulness with BP . In this section, we evaluate the effect of stress management tools such as yoga and mindfulness on blood pressure. We measured the diastolic blood pressure, systolic blood pressure and pulse from a medical-grade blood pressure monitor before and after yoga and mindfulness sessions. In order to decrease the recall bias, we measured each modality three times and took the mean value of the results. Then we applied one-sample t-test on the difference of mean values. The results are shown in Table 6.27. Guided mindfulness decreased the pulse by 5.75%, where guided yoga increased it by 8.06%. Guided yoga decreased the systolic blood pressure by 5.81% and the diastolic blood pressure by 1.93%. Guided mindfulness decreased the systolic blood pressure by 1.31% and in-

Table 6.26: The classification accuracy of the recovery sessions using stress management methods and stressful sessions using HRV.

Algorithm	**Accuracy**		
	Guided Mindfulness	**Yoga**	**Mobile Mindfulness**
MLP	90.00	97.50	93.94
RF	97.50	95.00	87.89
kNN	90.00	90.00	93.93
LDA	87.50	87.50	75.75
SVM	85.00	80.00	81.82

Table 6.27: The difference of the mean diastolic blood pressure, the mean systolic blood pressure and the mean pulse, before and after sessions of guided mindfulness and guided yoga. (* $p < 0.05$)

Activity	**Systolic**	**Diastolic**	**Pulse**
Guided Mindfulness	-1.31%	1.75%*	-5.75%*
Guided Yoga	-5.81%*	-1.93%	8.06%*

creased the diastolic blood pressure by 1.75%. In the guided mindfulness we explored a significant increase in the diastolic blood pressure ($p < 0.05$) and a significant decrease in the pulse ($p < 0.05$). In the guided yoga we explored a significant decrease in the systolic blood pressure ($p < 0.05$) and a significant increase in the pulse ($p < 0.05$). Guided yoga seems to be more effective than guided mindfulness for decreasing systolic and diastolic blood pressure. On the other hand, guided mindfulness seems to be more effective than guided yoga for decreasing the pulse due to the activity involved in yoga.

6.2.3. Interpretation

In this section, to test our system, we collected eight days of data of 16 subjects in the EU project training, where they faced a real-life stressor. The participants are coming from different countries and they have diversified cultures. The diversity is prominent because stress reactions of different cultures could be different and measuring the stress levels of those groups is more difficult than homogeneous culture groups [164]. To the best of our knowledge, this study is the first one that collects

a long time physiological data from a multicultural group and measures their stress levels as well as offering stress management techniques using wearables. 1440 hours of data (12 hours in a day) and 2780 (3 session-based and 1 daily questionnaire each day) self-report questions were collected during this eight-day event from each participant for measuring the perceived stress. We collected data both in the event sessions and after the event in participants' free times for 12 hours a day which shows that our study monitors the daily life stress. EDA and HRV signals are collected to detect physiological stress. The classification performance for both 2 and 3-class session-based, daily and long-term perceived stress levels were presented. Our long-term perceived stress detection system predicts the PSS long term stress levels with approximately 80% accuracy. This result is important because it could be used as a pre-screening tool for psychologists. It could advise people to see a psychologist if high-level long term perceived stress level is detected. The results showed that the selection of a stress scale has an important effect on the performance of the system. Our results show that weather-related, physical activity-related and activity type information improved the system performance. When the weather information in addition to the physiological signals is used, our model achieved 81% maximum classification accuracy with HRV signal and 72% with EDA signal in 2-class perceived stress level classification. Adding the known context (activity type) information also increased the performance. The 2-class classification accuracy is approximately 76% with both signals. In order to increase the daily perceived stress level detection performance, we further used physical activity based contextual information: stillness and step count, namely. These context data increased our 3-day daily perceived stress level classification performance in terms of validation accuracy to 72% with stillness and 76% with step count information. These results might indicate the correlation between the activity level and the daily stress. We also showed that a combination of different modalities increased stress detection performance and provided the most discriminative features. When the known context is used as the label, we achieved 98% accuracy for 2-class and 85% accuracy for 3-class physiological stress detection. Most of the studies in the literature, only detect stress levels of individuals. As mentioned previously, we further manage the participants' stress levels with yoga, mindfulness, and mobile mindfulness application while moni-

toring their stress levels. We investigated the success of each stress management system by the separability of physiological signals from high-stress sessions. We demonstrated that yoga and traditional mindfulness performed better than the mobile mindfulness application.

7. STRESS DETECTION IN AN UNRESTRICTED REAL-LIFE ENVIRONMENT

Lastly, we performed experiments in unrestricted real life (daily life) environments. In these environments, the movements of participants are unlimited and they are not following any prearranged program or schedule. Therefore, the accuracies are lower when compared with other mentioned environments. The subjects are wearing our smart wrist-worn devices in their daily lives that we have not got any context information. Session-based self-reports are used as the ground truth. In the first experiment, we applied LSTM to the collected physiological data. To the best of our knowledge, this is the first study that applies LSTM for daily life perceived stress detection. In the second part, to increase the performance of daily life stress detection studies, we pretrained an ML model in laboratory environments and tested on daily life data.

7.1. Daily Perceived Stress Level Detection Using LSTM Networks

We developed a multi-modal sensing platform that uses heart rate variability and accelerometer features to assess low and high daily perceived stress. We compared the performance of traditional machine learning algorithms with a shallow neural network and a deep sequential neural network. We evaluated the discriminative ability of features with Random Forest, K-nearest Neighbour, Support Vector Machine, Naive Bayes, MLP and LSTM. We obtained promising results on discrimination of low and high stress with the consumer grade off-the-shelf smartwatches in non-restricted daily life.

7.1.1. Experiment Design

7.1.1.1. Data Collection. Seventeen participants joined this study. Answers of 7854 questions from NASA-TLX questionnaire and 374 hours of physiological signals are

Table 7.1: NASA-TLX factors, rating scales and questions.

Factors	**Rating Scale**	**Questions**
Mental Demand	Low - High	How much mental and perceptual activity did you spend for this task?
Physical Demand	Low - High	How much physical activity did you spend for this task?
Temporal Demand	Low - High	How much time pressure did you feel in order to complete this task?
Performance	Good - Poor	How successful do you think you were in accomplishing the goals of the task
Effort	Low - High	How hard did you have to work to accomplish your level of performance?
Frustration	Low - High	How insecure, discouraged, irritated, stressed, and annoyed were you during this task?

collected from these participants. The ages were between 23 and 32. The gender distribution is 13 men and 4 women. We discarded data of the five participants (4 men 1 woman) from the study due to the misuse of the data collection application. The first version of our data collection application did not have a haptic feedback functionality. The subjects who were dropped earlier when this functionality did not exist and they forgot to fill the questionnaire in time. We added the haptic feedback feature on the next versions of our application and the collaboration of the users on filling the questionnaire after the physiological data collection increased. The data is collected over a month from each participant. The resulting amount of questions became 5544 and the duration of the total recording time became 264 hours. Participants were asked to wear a Samsung Gear S2 smartwatch during their everyday life without any restriction. The procedure of the data collection session in non-restricted everyday life is described in Figure 7.1. Participants recorded their physiological signals in their daily life environment with the application that we developed for Samsung Gear S2. We wanted to make the dataset as rich as possible. Evaluating the physiological data of a participant on the same day may create a bias. In order to discard this bias, participants are allowed to start our application once a day for an hour, i.e., one trial per day per participant is collected. The length of a trial is 60 minutes. At the end of the session, the participant has received a strong vibration signal from the smartwatch and received a NASA-TLX questionnaire containing 21 questions with 6 scales. The perceived stress score is determined with the NASA-TLX questionnaire (see Table

7.1). We wanted them to complete at least 20 sessions, most of them provided more than 20 sessions. Generally, participants completed the experiment in a month. The maximum total number of collected trials per person is 29 and the minimum total number of trials is 20. The trials with the answers of questionnaires, where the time-span between the daily life session and the completion of NASA-TLX are more than 30 minutes and sessions where the data quality is very low are removed. The answers of 5544 questions from NASA-TLX questionnaire and 264 hours of physiological data from 12 participants are provided to the rest of the proposed system. The histogram of the remaining NASA-TLX scores (N=264) are shown in Figure 7.2 with a normal density curve.

7.1.1.2. Smartwatch Framework. Samsung Gear S2 is equipped with the ambient light sensor, PPG sensor for heart rate monitoring, 3D inertial measurement unit with 3D accelerometer and gyroscope and pedometer. Its price is more affordable compared to research smart bands. Samsung Gear Series run on the Tizen platform. We developed a data collection application for the Tizen Platform (Wearable 2.3.2). This framework uses the javascript for the development. Our application gathers the inter-beat (RR) interval and 3D acceleration. In order to compute RR intervals, the smartwatch samples PPG with 100 Hz and calculates each successive beat. The sampling rate of the 3D accelerometer is 20 Hz. We used Samsung Gear S2's haptic feedback functionality to inform participants that they should fill the questionnaire using their smartphones.

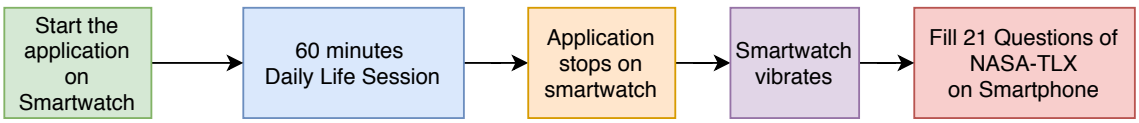

Figure 7.1: The experiment procedure in non-restricted everyday life environment.

7.1.1.3. Collection of Self-Reports. NASA-TLX [153] is used to measure the perceived stress of individuals. First, the subject has to rate each item phase with 6 items on a scale from 0 to 100 that best indicate his experience in the task. The rating consists of the following items: mental demand, physical demand, temporal demand, own performance, effort, and frustration. Next, pairwise comparison of each scale is

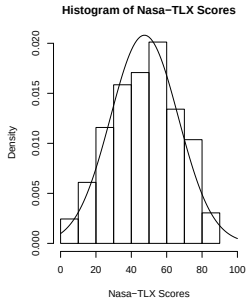

Figure 7.2: Histogram of Nasa-TLX scores.

demonstrated to the subject who is asked to indicate which of the items represents the most important contributor to the stress. Based on these ratings, the total stress was computed as a weighted average. This makes NASA-TLX suitable for measuring the perceived stress in different types of tasks. NASA-TLX can be implemented on mobile phone [165], paper [153] or computer [166] [167]. The six rating scales, questions and endpoints of the mobile implementation of the NASA-TLX is shown in Table 7.1. In this study, we used the official application of the NASA-TLX [165] for the iPhone and our implementation of NASA-TLX for Android. NASA-TLX can also be implemented on the smartwatch framework, however screen size of the current smartwatches make it difficult to fill the questionnaire for participants.

Definition of the ground-truth in daily life is not easy and it is still an open issue. We propose to record the self-reported questionnaires at the end of each hour and label each session according to their score. Some of the models require a laboratory calibration session for each individual [168] [119]. Our model does not require such an enrollment procedure. We modified the approach used in the work of Sano et al. [169]. The sessions are sorted according to the NASA-TLX scores. We divided the dataset into two parts, first 50% as high stress and last 50% as low stress.

7.1.1.4. Ethics. The procedure of the methodology used in this study is approved by the ethical committee of our university. Prior to the data acquisition, each participant received a consent form that explains the experimental procedure and its benefits and implications to both the society and the subject. The procedure was also explained

Table 7.2: The effect of different neurons and dropout parameters of LSTM networks on classification performance.

Neurons/Dropout	0.0	0.2	0.4	0.6
50 Neurons	67.50%	67.50%	65.62%	64.37%
100 Neurons	66.87%	66.87%	65.62%	65.62%
150 Neurons	65.62%	67.50%	65.62%	63.12%
200 Neurons	65.62%	**70.00%**	65.00%	64.37%
250 Neurons	66.25%	68.12%	66.87%	63.75%

Table 7.3: The effect of different classification algorithms on the performance of the system.

Algorithm	**Accuracy**
Random Forest	57.92%
PCA + SVM(RBF)	65.24%
KNN (N=3)	57.92%
Naive Bayes	62.80%
MLP	60.97%
LSTM	**70.00%**

vocally to the subject. The data collection procedure and all of the interventions in this research fully meet the 1964 Declaration of Helsinki [159]. All of the data are stored anonymously.

7.1.2. Experimental Results and Discussion

The effects of the selection of the number of neurons and the recurrent dropout rate of LSTM are shown in Table 7.2. We evaluated the number of neurons from 50 to 250 and the amount of recurrent dropout from 0.0 to 0.6. The results show that the selection of parameters has a significant impact on the classification accuracy. LSTM networks outperformed traditional machine learning algorithms. The best performed traditional machine learning model is selected as PCA + SVM (RBF) with an accuracy

of 65.24%. LSTM improved the performance by 4.76% (see Table 7.3). To the best of our knowledge, our study is the first study using LSTM for daily perceived stress detection.

7.2. Development of Daily Life Perceived Stress Models Based on Laboratory Tests

As mentioned previously in Chapter 2, the performance of daily life stress detection studies is lower than those in restricted real-life environments and laboratory environments. In this section, we trained machine learning models in laboratory environments and tested them in the wild [145]. In the laboratory environments, we have more controlled situations and we can record perceived stress ground truth from self-reports. When the ML model is formed in the laboratory by using these self-reports and used in the wild, accuracies between daily life and laboratory studies can be achieved. The noise coming from the self-reports of unknown context situations in daily life can be eliminated. We can examine the studies in the literature in five different groups (see Figure 7.3). The first group develops a laboratory model with known context labels and tests in the same environment. Since the stressor levels (i.e., the context participants are in) are known at any time in laboratory experiments, they could be used as the ground truth labels for machine learning (ML) models and we called this type as Laboratory-to-Laboratory Known Context (LLKC) models. The second type is using collected self-report labels in laboratory environments and tests the created model in the same environment. We call this type as Laboratory-to-Laboratory-Self-Report (LLSR) models. The third type is using self-report questionnaires collected in the wild and test the model in the same environment. We named this model as Daily-to-Daily-Self-Report (DDSR) models. Since we could not monitor the everyday life of participants and get the ground truth all the time, a known context does not exist in daily life environments. The laboratory data could be used to enhance daily life stress detection models. If the known context labels in the laboratory are used for an ML model development, we call this fourth type as Laboratory-to-Daily-Known-Context (LDKC) models. On the contrary, if the self-reports in the laboratory are

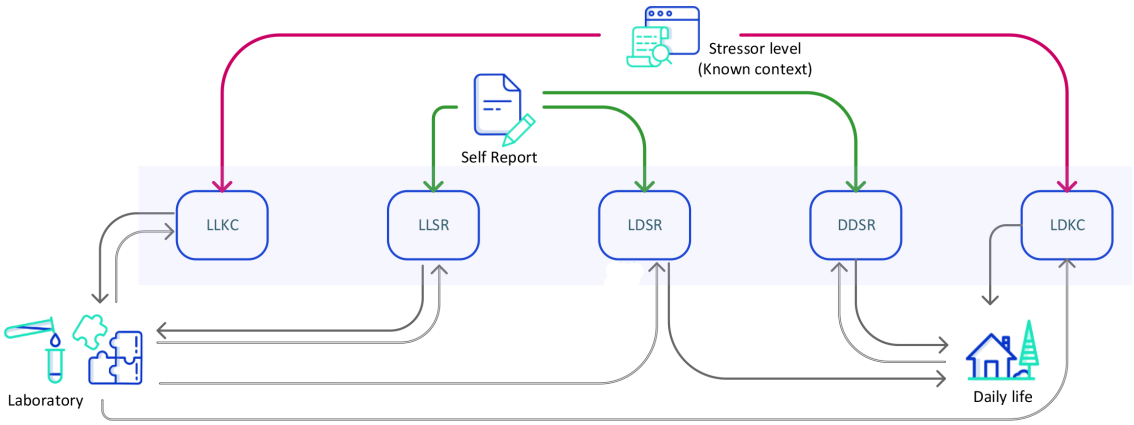

Figure 7.3: Demonstration of the models which are using different ground truth labels and training environments.

used as labels and the developed models are tested in the wild, we name the fifth type as Laboratory-to-Daily-Self-Report (LDSR). The performance of the LLKC and LLSR models was presented in Chapter 5.

7.2.1. Experiment Design

In this section, we collected daily life physiological data of 14 participants of laboratory stress level detection experiment in Chapter 5. After the controlled room experiments were finished, we gave the Empatica E4 devices to all participants. They are told to wear Empatica E4 devices for twelve hours in seven days between 9 AM and 9 PM. These days are not necessarily consecutive. For each three hour session, they filled the PSS-5 as the ground truth. In order to make sure of this, we sent them e-mails in every 3 hours to remind them to fill the PSS-5. This questionnaire is strongly correlated with PSS-14 and appropriate for ambulatory settings. We have got 1176 hours of daily life data with 294 ecological momentary assessment (EMA) questionnaires.

All participants were senior undergraduate students at Bogazici University. Since these participants already attended our laboratory stress induction tests, we used the laboratory data for training machine learning models and daily life data for testing. Empatica E4's flash memory can store up to 60 hours of data. Therefore, we transferred the recorded sessions every 3 days to the cloud storage maintained by the E4

manager application.

The procedure of the methodology used in this study is approved by the Institutional Review Board for Research with Human Subjects of Bogazici University with the approval number 2018/16. Each subject signed a consent form explains the procedure of the experiment and its aims and implications to both the society and the subject before the data recording begins. We further described the procedure verbally to the participants. The recorded data are stored anonymously.

7.2.2. Experimental Results and Discussion

In the laboratory environment stress detection experiment (see Chapter 5), we have two types of ground truth. The first one is the PSS-5 self report. We collected this questionnaire after each session. The second one is the known context. In other words, we have the knowledge of which context the participant is in at any time. We know whether the participant is in the TSST test or baseline session. On the other hand, in the wild, we have only self reports of individuals which is among the reasons why daily life stress detection performances are low and there is still a room for improvement in daily life stress detection research.

We used both laboratory ground truth labels separately while developing machine learning models for laboratory environments and testing in daily life. The model is trained with the laboratory self-report ground truth is called LSRGT where the model is trained with the laboratory known context ground truth is called LKCGT. We expect that LSRGT will have higher performance than LKCGT because we have only perceived stress labels (questionnaires) in the wild. Participants might experience different perceived stress than the known context implies. We further developed a model which is trained and tested in daily life data which we called DSRGT (Daily Self Report Ground Truth). We compared these three models in Table 7.4. As expected LSRGT has the best performance whereas DSRGT model has the lowest performance. We can state that collecting laboratory data and training a stress level detection model with that data improves the stress level detection performance in the wild.

Table 7.4: The classification accuracy of the different ground truth and training data. As physiological signals, EDA - HRV combination is used.

Algorithm	**Accuracy**		
	DSRGT	**LKCGT**	**LSRGT**
MLP	63.50	64.73	72.20
RF	61.90	68.87	72.61
kNN	65.90	70.53	72.20
Logistic Regression	70.6	62.65	73.81
SVM	71.4	71.78	73.44

8. CONCLUSION

In this thesis, we presented our platform independent stress level detection system which uses physiological signals collected by unobtrusive smartbands and smartwatches. Since our system employs unobtrusive wearable devices, it can easily be used in the daily life of individuals. It can track the stress levels in real-time and intervene if an extreme level of stress is detected. As physiological signals, we used HRV and EDA signals. We further tested some traditional and novel stress management techniques to alleviate high stress levels.

Physiological signals can be collected with medical grade devices which have high data quality. However, they are obtrusive. On the other hand, wrist-worn devices are prone to artifacts due to unrestricted hand movements and intense physical activities. For each physiological signal, modality-specific artifact detection, removal and feature extraction methods should be developed. As an example, the EDA signal is affected by temperature changes and physical activity, whereas the HRV is mainly affected by physical activities. So, an artifact detection tool should use temperature and accelerometer data for the EDA signal and only accelerometer signal for the HRV signal. We can further detect artifacts by analyzing the signal itself to see abnormal behaviors. If the heart rate changes drastically in a few seconds, it is probably affected by movements. After detecting artifacts, appropriate interpolation methods should be employed to replace the removed data. After feature extraction, traditional machine learning algorithms and a shallow neural network were applied for classification. In addition, we applied LSTM which is an appropriate recurrent neural network for sequential data. We described our physiological signal processing and machine learning modules which took mentioned constraints into account in Chapter 4.

In order to evaluate the performance of our system, we tested it on the field. We first collected data in the laboratory environment with Empatica E4 wristbands. We successfully discriminate between stress and baseline states. In a different experiment, participants solve puzzles with different difficulty levels and we successfully classified the levels by using our system. We further capture an 'Aha!' moment which is a kind of arousal. Since EDA is good at detecting arousal moments, we used only EDA and

detect these moments successfully.

We took a step outside the laboratory after successful laboratory trials. We used personal stress level clustering and decision level smoothing methods to improve the relatively low performances of real-life stress level detection systems. In order to test and evaluate our system, we collected physiological data from 32 participants of a summer school with wrist-worn unobtrusive wearable devices. This event is comprised of baseline, lecture, exam and recovery sessions. In the recovery session, a stress management method was applied to alleviate the stress of the participants. Since patterns of stress are ideographic, person-independent models have generally lower accuracies. On the contrary, person-specific models have higher accuracies, but they require a long-term data collection period. In this experiment, we developed a hybrid approach of personal level stress clustering by using baseline stress self-reports to increase the success of person-independent models without requiring a substantial amount of personal data. We further added decision level smoothing to our unobtrusive smartwatch based stress level differentiation system to increase the performance by correcting false labels assigned by the machine learning algorithm. By using our system, we were able to differentiate the 3-levels of stress successfully. We further substantially increase our system's performance by applying the mentioned methods. We demonstrated the success of the stress reduction methods by analyzing physiological signals and self-reports.

We also added contextual information to increase the performance of our system in a semi-restricted real-life environment. To evaluate the performance of our system, we monitored the stress levels of 16 participants in an 8 day long EU project training event every day. 1440 hours of physiological data and 2780 self-report questions were collected. We applied different types of stress management techniques in the event. The project presentations in front of a jury at the end of the event were the source of significant real stress. Session-based, daily and long-term perceived stress levels can be identified by using the proposed system. Contextual information such as weather, activity type and physical activity level significantly increased our performance.

Finally, we tested our system in unrestricted real-life environments. We employed LSTM and laboratory pre-trained models to improve the performance of daily life stress level detection systems. Since LSTM is suitable for sequential data, we ap-

plied it to our daily life physiological data which is collected with Samsung Galaxy Gear S2 smartwatches. We have recorded 374 hours of daily life data and 7854 answers to stress questions from 17 subjects in their real life environments over a month. The model was trained and evaluated with the real-life physiological data coming from different days. We used our multi-modal sensing platform that employs heart rate variability and accelerometer features to assess low and high stress. We compared the performance of machine learning and deep learning algorithms. We evaluated the discriminative ability of features with traditional machine learning algorithms and Long Short-Term Memory Networks. Another novel technique we proposed is developing a model in laboratory environments and testing on daily life data. We collected one week unrestricted daily life data of 14 individuals who participated in our laboratory stress detection experiment (in Chapter 5). We achieved a 5% accuracy increase with LSTM and approximately 10% accuracy increase with pre-trained laboratory models. Our techniques could be applied to improve performances of daily life stress level detection studies which have relatively lower accuracies when compared to other environments.

Stress is among the causes of many diseases in our modern society. By using the recent advances in wearable technologies, an unobtrusive stress level detection system may help to improve our lives. However, this system should also offer stress management techniques that could be applied indoors. We developed a successful stress level detection system and offered some techniques to improve its performance in real life. We also tested and demonstrated the performance of several stress alleviation methods in decreasing stress levels. As future work, we plan to develop personalized perceived stress models and to analyze their effect on system performance. Furthermore, especially commercial smartwatches have a short battery life problem. When all the sensors are active, the battery life of these devices could be shortened to 3-4 hours. As another future study, to be able to monitor users continuously, we plan to increase the battery life of the commercial grade off-the-shelf smartwatches by selecting the best duty cycle parameters for stress detection. It is believed that the contribution of this thesis will be useful for both academic and industry users and can be utilized to detect and alleviate high stress levels in real time.

REFERENCES

1. *The structure of the ECG Signal*, https://commons.wikimedia.org/wiki/File:ECG-RRinterval.svg, "Accessed at February 2020".

2. *Online Stroop Color Word Test Example*, http://opencoglab.org/stroop/, "Accessed at February 2020".

3. "Echo App", https://play.google.com/store/apps/details?id=com.childishideas.echo.android, Accessed at December 2019.

4. *Pause App*, https://itunes.apple.com/us/app/pause-relaxation-at-your-fingertip/id991764216?mt=8, Accessed at December 2018.

5. *HeartMath App*, https://itunes.apple.com/us/app/inner-balance/id569278747?mt=8, Accessed at December 2019.

6. Can, Y. S., N. Chalabianloo, D. Ekiz and C. Ersoy, "Continuous Stress Detection Using Wearable Sensors in Real Life: Algorithmic Programming Contest Case Study", *Sensors*, Vol. 19, No. 8, p. 1849, 2019.

7. Gross, J. J., "The emerging field of emotion regulation: An integrative review", *Review of general psychology*, Vol. 2, No. 3, pp. 271–299, 1998.

8. *Definition of Stress*, 2018, https://www.stress.org/what-is-stress/, accessed at December 2018.

9. *Stress: Symptoms, Causes and Effects*, 2018, https://www.helpguide.org/articles/stress/, Accessed at December 2018.

10. *Understanding the stress response*, 2018, http://www.health.harvard.edu/, ac-

cessed at December 2018.

11. Plarre, K., A. Raij, S. M. Hossain, A. A. Ali, M. Nakajima, M. Al'absi, E. Ertin, T. Kamarck, S. Kumar, M. Scott, D. Siewiorek, A. Smailagic and L. E. Wittmers, "Continuous inference of psychological stress from sensory measurements collected in the natural environment", *Proceedings of the 10th ACM/IEEE International Conference on Information Processing in Sensor Networks*, pp. 97–108, April 2011.

12. Liapis, A., C. Katsanos, D. Sotiropoulos, M. Xenos and N. Karousos, "Stress Recognition in Human-computer Interaction Using Physiological and Self-reported Data: A Study of Gender Differences", *Proceedings of the 19th Panhellenic Conference on Informatics*, PCI '15, pp. 323–328, ACM, New York, NY, USA, 2015.

13. Rodrigues, J. G. P., M. Kaiseler, A. Aguiar, J. P. S. Cunha and J. Barros, "A Mobile Sensing Approach to Stress Detection and Memory Activation for Public Bus Drivers", *IEEE Transactions on Intelligent Transportation Systems*, Vol. 16, No. 6, pp. 3294–3303, Dec 2015.

14. S., A. A., S. P., S. V., S. K. S., S. A., T. J. Akl, P. S. P. and M. Sivaprakasam, "Electrodermal Activity Based Pre-surgery Stress Detection Using a Wrist Wearable", *IEEE Journal of Biomedical and Health Informatics*, Vol. 24, No. 1, pp. 92–100, 2020.

15. Smets, E., W. De Raedt and C. Van Hoof, "Into the Wild: The Challenges of Physiological Stress Detection in Laboratory and Ambulatory Settings", *IEEE Journal of Biomedical and Health Informatics*, Vol. 23, No. 2, pp. 463–473, 2019.

16. Garcia-Ceja, E., V. Osmani and O. Mayora, "Automatic Stress Detection in Working Environments From Smartphones' Accelerometer Data: A First Step", *IEEE Journal of Biomedical and Health Informatics*, Vol. 20, No. 4, pp. 1053–1060, July 2016.

17. Hernando, A., J. Lázaro, E. Gil, A. Arza, J. M. Garzón, R. López-Antón, C. de la Cámara, P. Laguna, J. Aguiló and R. Bailón, "Inclusion of Respiratory Frequency Information in Heart Rate Variability Analysis for Stress Assessment", *IEEE Journal of Biomedical and Health Informatics*, Vol. 20, No. 4, pp. 1016–1025, 2016.

18. Wang, Z., A. Parnandi and R. Gutierrez-Osuna, "BioPad: Leveraging off-the-Shelf Video Games for Stress Self-Regulation", *IEEE Journal of Biomedical and Health Informatics*, Vol. 22, No. 1, pp. 47–55, 2018.

19. Alberdi, A., A. Aztiria and A. Basarab, "Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review", *Journal of Biomedical Informatics*, Vol. 59, pp. 49 – 75, 2016.

20. *European agency for safety and health at work campaign guide managing stress and psychological risks at work*, 2013, "https://eguides.osha.europa.eu/stress/IE-EN/story-content/external-files/", "Accessed at February 2020.".

21. Colligan, T. W. and E. M. Higgins, "Workplace stress: Etiology and consequences", *Journal of workplace behavioral health*, Vol. 21, No. 2, pp. 89–97, 2006.

22. *Stress is killing you*, 2018, http://www.stress.org/stress-is-killing-you/, accessed at December 2018.

23. B. Cosemans, R. G. D. F. K. V., Marlen Cosmar, "Calculating the cost of work-related stress and psychosocial risks", *European Agency for Safety and Health at Work*, 2014.

24. "European Agency for Safety and Health at Work, European Opinion Poll on Occupational Safety and Health", *European Agency for Safety and Health at Work*, 2013, "http://dx.doi.org/10.2802/55505".

25. Picard, R. W., "Automating the Recognition of Stress and Emotion: From Lab to Real-World Impact", *IEEE MultiMedia*, Vol. 23, No. 3, pp. 3–7, July 2016.

26. England, M. J., C. T. Liverman, A. M. Schultz and L. M. Strawbridge, "Epilepsy across the spectrum: Promoting health and understanding.: A summary of the Institute of Medicine report", *Epilepsy Behavior*, Vol. 25, No. 2, pp. 266 – 276, 2012.

27. et. al, P. R., "Incidence and mechanisms of cardiorespiratory arrests in epilepsy monitoring units (MORTEMUS): a retrospective study", *The Lancet Neurology,Elsevier*.

28. "American Psychology Association. Stress: The different kinds of stress. [Online]", `http://www.apa.org/helpcenter/stress-kinds.aspx`, "Accessed at February 2020".

29. Krantz, D. S., K. S. Whittaker and D. S. Sheps, "Psychosocial risk factors for coronary heart disease: Pathophysiologic mechanisms.", *American Psychological Association*, 2011.

30. Pickering, T. G., "Mental stress as a causal factor in the development of hypertension and cardiovascular disease", *Current hypertension reports*, Vol. 3, No. 3, pp. 249–254, 2001.

31. Mönnikes, H., J. Tebbe, M. Hildebrandt, P. Arck, E. Osmanoglou, M. Rose, B. Klapp, B. Wiedenmann and I. Heymann-Mönnikes, "Role of stress in functional gastrointestinal disorders", *Digestive Diseases*, Vol. 19, No. 3, pp. 201–211, 2001.

32. Herbert, J., "Fortnightly review: Stress, the brain, and mental illness", *BMJ*, Vol. 315, No. 7107, pp. 530–535, 1997.

33. M. Milczarek, E. G., Elke Schneider, "OSH in figures, stress at work, fact and figures", *European agency for safety and health at work*, 2009.

34. Picard, R. W., *Affective Computing*, MIT Press, Cambridge, MA, USA, 1997.

35. Picard, R. W., "Affective Computing for HCI.", *HCI (1)*, pp. 829–833, 1999.

36. Picard, R. W., "Affective computing: challenges", *International Journal of Human-Computer Studies*, Vol. 59, No. 1, pp. 55–64, 2003.

37. Kurniawan, H., A. V. Maslov and M. Pechenizkiy, "Stress detection from speech and Galvanic Skin Response signals", *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, pp. 209–214, June 2013.

38. Kaklauskas, A., E. Zavadskas, M. Seniut, G. Dzemyda, V. Stankevic, C. Simkevičius, T. Stankevic, R. Paliskiene, A. Matuliauskaite, S. Kildiene, L. Bartkiene, S. Ivanikovas and V. Gribniak, "Web-based Biometric Computer Mouse Advisory System to Analyze a User's Emotions and Work Productivity", *Engineering Applications of Artificial Intelligence*, Vol. 24, No. 6, pp. 928 – 945, 2011.

39. Zimmermann, P., S. Guttormsen, B. Danuser and P. Gomez, "Affective Computing–A Rationale for Measuring Mood With Mouse and Keyboard", *International Journal of Occupational Safety and Ergonomics*, Jan. 2003, `http://www.cse.unr.edu/ sushil/class/ps/papers/03-JOSE.pdf`.

40. Sysoev, M., A. Kos, U. Sedlar and M. Pogacnik, "Sensors Classification for Stress Analysis: Toward Automatic Stress Recognition", *2014 International Conference on Identification, Information and Knowledge in the Internet of Things*, pp. 117–121, Oct 2014.

41. Greene, S., H. Thapliyal and A. Caban-Holt, "A Survey of Affective Computing for Stress Detection: Evaluating technologies in stress detection for better health", *IEEE Consumer Electronics Magazine*, Vol. 5, No. 4, pp. 44–56, Oct 2016.

42. Kappeler-Setz, C., *Multimodal emotion and stress recognition*, Ph.D. Thesis, ETH

Zurich, 2012.

43. Lerner, J. S., R. E. Dahl, A. R. Hariri and S. E. Taylor, "Facial expressions of emotion reveal neuroendocrine and cardiovascular stress responses", *Biological psychiatry*, Vol. 61, No. 2, pp. 253–260, 2007.

44. Can, Y. S. and F. Alagoz, "A Secure Biometric Identification Technique Using Spread Spectrum Audio Watermarking", *2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS)*, pp. 345–350, April 2016.

45. Arnrich, B., C. Setz, R. L. Marca, G. Troster and U. Ehlert, "What Does Your Chair Know About Your Stress Level?", *IEEE Transactions on Information Technology in Biomedicine*, Vol. 14, No. 2, pp. 207–214, March 2010.

46. McDuff, D., A. Karlson, A. Kapoor, A. Roseway and M. Czerwinski, "AffectAura: an intelligent system for emotional memory", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 849–858, ACM, 2012.

47. Kirschbaum, C., K.-M. Pirke and D. H. Hellhammer, "The 'Trier Social Stress Test'–a tool for investigating psychobiological stress responses in a laboratory setting", *Neuropsychobiology*, Vol. 28, No. 1-2, pp. 76–81, 1993.

48. Stroop, J. R., "Studies of interference in serial verbal reactions.", *Journal of experimental psychology*, Vol. 18, No. 6, p. 643, 1935.

49. Lezak, M. D., *Neuropsychological assessment*, Oxford University Press, USA, 2004.

50. Dedovic, K., R. Renwick, N. K. Mahani, V. Engert, S. J. Lupien and J. C. Pruessner, "The Montreal Imaging Stress Task: using functional imaging to investigate

the effects of perceiving and processing psychosocial stress in the human brain", *Journal of Psychiatry and Neuroscience*, Vol. 30, No. 5, p. 319, 2005.

51. Lowery, D., R. B. Fillingim and R. A. Wright, "Sex differences and incentive effects on perceptual and cardiovascular responses to cold pressor pain", *Psychosomatic medicine*, Vol. 65, No. 2, pp. 284–291, 2003.

52. Brouwer, A.-M. and M. A. Hogervorst, "A new paradigm to induce mental stress: the Sing-a-Song Stress Test (SSST)", *Frontiers in neuroscience*, Vol. 8, p. 224, 2014.

53. Lang, P. J., "International affective picture system (IAPS): Affective ratings of pictures and instruction manual", *Technical report*, 2005.

54. Lang, P. and M. M. Bradley, "The International Affective Picture System (IAPS) in the study of emotion and attention", *Handbook of emotion elicitation and assessment*, Vol. 29, 2007.

55. "IAPS definition. [Online]", `https://en.wikipedia.org/wiki/International-Affective-Picture-System`, "Accessed at February 2020".

56. Giannakakis, G., D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis and M. Tsiknakis, "Review on psychological stress detection using biosignals", *IEEE Transactions on Affective Computing*, pp. 1–1, 2019.

57. Wyatt, J. and J. Liu, "Basic concepts in medical informatics", *Journal of Epidemiology & Community Health*, Vol. 56, No. 11, pp. 808–812, 2002.

58. Cheng, P., A. Lucero and J. Buur, "PAUSE: Exploring Mindful Touch Interaction on Smartphones", *Proceedings of the 20th International Academic Mindtrek Conference*, AcademicMindtrek '16, pp. 184–191, ACM, New York, NY, USA, 2016.

59. *Exercise: A guide to Tai Chi. [Online]*, 2019, `https://www.nhs.uk/live-well/exercise/guide-to-tai-chi`, Accessed at February 2020.

60. *Depression in adults: Recognition and management. Clinical guideline [CG90]*, 2009, "Accessed at February 2020".

61. Isaacs, E., A. Konrad, A. Walendowski, T. Lennig, V. Hollis and S. Whittaker, "Echoes from the Past: How Technology Mediated Reflection Improves Well-being", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pp. 1071–1080, ACM, New York, NY, USA, 2013.

62. Hollis, V., A. Konrad, A. Springer, M. Antoun, C. Antoun, R. Martin and S. Whittaker, "What Does All This Data Mean for My Future Mood? Actionable Analytics and Targeted Reflection for Emotional Well-Being", *Hum.-Comput. Interact.*, Vol. 32, No. 5-6, pp. 208–267, Nov. 2017.

63. Thapliyal, H., V. Khalus and C. Labrado, "Stress Detection and Management: A Survey of Wearable Smart Health Devices", *IEEE Consumer Electronics Magazine*, Vol. 6, No. 4, pp. 64–69, Oct 2017.

64. Akmandor, A. O. and N. K. Jha, "Keep the Stress Away with SoDA: Stress Detection and Alleviation System", *IEEE Transactions on Multi-Scale Computing Systems*, Vol. 3, No. 4, pp. 269–282, Oct 2017.

65. Chen, K., W. Fink, J. Roveda, R. D. Lane, J. Allen and J. Vanuk, "Wearable sensor based stress management using integrated respiratory and ECG waveforms", *2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pp. 1–6, June 2015.

66. Hill, T., P. Lewicki and P. Lewicki, *Statistics: methods and applications: a comprehensive reference for science, industry, and data mining*, StatSoft, Inc., 2006.

67. Alpaydin, E., "Introduction to machine learning. 2004", *Cambridge, Massachusettes: The MIT Press Google Scholar*.

68. Caudill, M., "Neural Networks Primer, Part I", *AI Expert*, Vol. 2, No. 12, pp. 46–52, Dec. 1987.

69. Tunca, C., *Gait Analysis and Fall Risk Asessment with Wearable Inertial Sensors*, Ph.D. Thesis, Bogazici University, 2019.

70. Hochreiter, S. and J. Schmidhuber, "Long short-term memory", *Neural computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

71. Alhagry, S., A. A. Fahmy and R. A. El-Khoribi, "Emotion Recognition based on EEG using LSTM Recurrent Neural Network", *International Journal of Advanced Computer Science and Applications*, Vol. 8, No. 10, 2017.

72. Costin, R., C. Rotariu and A. Pasarica, "Mental stress detection using heart rate variability and morphologic variability of EeG signals", *2012 International Conference and Exposition on Electrical and Power Engineering*, pp. 591–596, Oct 2012.

73. Castaldo, R., L. Montesinos, P. Melillo, S. Massaro and L. Pecchia, "To What Extent Can We Shorten HRV Analysis in Wearable Sensing? A Case Study on Mental Stress Detection.", H. Eskola, O. Väisänen, J. Viik and J. Hyttinen (Editors), *EMBEC & NBC 2017*, pp. 643–646, Springer Singapore, Singapore, 2018.

74. de Santos Sierra, A., C. S. Avila, J. G. Casanova, G. B. del Pozo and V. J. Vera, "Two Stress Detection Schemes Based on Physiological Signals for Real-Time Applications", *2010 Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pp. 364–367, Oct 2010.

75. Setz, C., B. Arnrich, J. Schumm, R. La Marca, G. Tröster and U. Ehlert, "Dis-

criminating Stress from Cognitive Load Using a Wearable EDA Device", *Trans. Info. Tech. Biomed.*, Vol. 14, No. 2, pp. 410–417, Mar. 2010.

76. de Santos Sierra, A., C. S. Avila, J. G. Casanova and G. B. del Pozo, "A Stress-Detection System Based on Physiological Signals and Fuzzy Logic", *IEEE Transactions on Industrial Electronics*, Vol. 58, No. 10, pp. 4857–4865, Oct 2011.

77. Sandulescu, V., S. Andrews, D. Ellis, N. Bellotto and O. M. Mozos, "Stress detection using wearable physiological sensors", *International Work-Conference on the Interplay Between Natural and Artificial Computation*, pp. 526–532, Springer, 2015.

78. Soury, M. and L. Devillers, "Stress Detection from Audio on Multiple Window Analysis Size in a Public Speaking Task", *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 529–533, Sept 2013.

79. Wijsman, J., B. Grundlehner, H. Liu, H. Hermens and J. Penders, "Towards mental stress detection using wearable physiological sensors", *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 1798–1801, Aug 2011.

80. Abouelenien, M., M. Burzo and R. Mihalcea, "Human Acute Stress Detection via Integration of Physiological Signals and Thermal Imaging", *Proceedings of the 9th ACM International Conference on Pervasive Technologies Related to Assistive Environments*, PETRA '16, pp. 32:1–32:8, ACM, New York, NY, USA, 2016.

81. Vanitha, V. and P. Krishnan, "Real time stress detection system based on EEG signals", pp. 271–275, 2016.

82. Aigrain, J., S. Dubuisson, M. Detyniecki and M. Chetouani, "Person-specific behavioural features for automatic stress detection", *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 03, pp. 1–6, May 2015.

83. Aigrain, J., M. Spodenkiewicz, S. Dubuisson, M. Detyniecki, D. Cohen and M. Chetouani, "Multimodal stress detection from multiple assessments", *IEEE Transactions on Affective Computing*, Vol. PP, No. 99, pp. 1–1, 2016.

84. Giannakakis, G., M. Pediaditis, D. Manousos, E. Kazantzaki, F. Chiarugi, P. G. Simos, K. Marias and M. Tsiknakis, "Stress and anxiety detection using facial cues from videos", *Biomedical Signal Processing and Control*, Vol. 31, pp. 89–101, 2017.

85. Mokhayeri, F. and M. R. Akbarzadeh-T, "Mental Stress Detection Based on Soft Computing Techniques", *2011 IEEE International Conference on Bioinformatics and Biomedicine*, pp. 430–433, Nov 2011.

86. Baltaci, S. and D. Gokcay, "Role of pupil dilation and facial temperature features in stress detection", *2014 22nd Signal Processing and Communications Applications Conference (SIU)*, pp. 1259–1262, April 2014.

87. Mozos, O. M., V. Sandulescu, S. Andrews, D. Ellis, N. Bellotto, R. Dobrescu and J. M. Ferrandez, "Stress Detection Using Wearable Physiological and Sociometric Sensors", *International Journal of Neural Systems*, Vol. 27, No. 02, p. 1650041, 2017.

88. Zubair, M., C. Yoon, H. Kim, J. Kim and J. Kim, "Smart Wearable Band for Stress Detection", *2015 5th International Conference on IT Convergence and Security (ICITCS)*, pp. 1–4, Aug 2015.

89. Martinez, R., E. Irigoyen, A. Arruti, J. Martin and J. Muguerza, "A real-time stress classification system based on arousal analysis of the nervous system by an F-state machine", *Computer methods and programs in biomedicine*, Vol. 148, pp. 81–90, 2017.

90. Hong, J.-H., J. Ramos and A. K. Dey, "Understanding Physiological Responses to Stressors During Physical Activity", *Proceedings of the 2012 ACM Conference on*

Ubiquitous Computing, UbiComp '12, pp. 270–279, ACM, New York, NY, USA, 2012.

91. Ramos, J., J.-H. Hong and A. K. Dey, "Stress Recognition-A Step Outside the Lab.", *PhyCS*, pp. 107–118, 2014.

92. Cho, Y., N. Bianchi-Berthouze, S. J. Julier and N. Marquardt, "ThermSense: Smartphone-based Breathing Sensing Platform using Noncontact Low-Cost Thermal Camera", *arXiv preprint arXiv:1710.05044*, 2017.

93. Cho, Y., N. Bianchi-Berthouze and S. J. Julier, "DeepBreath: Deep Learning of Breathing Patterns for Automatic Stress Recognition using Low-Cost Thermal Imaging in Unconstrained Settings", *arXiv preprint arXiv:1708.06026*, 2017.

94. Akhonda, M. A. B. S., S. M. F. Islam, A. S. Khan, F. Ahmed and M. M. Rahman, "Stress detection of computer user in office like working environment using neural network", *Computer and Information Technology (ICCIT), 2014 17th International Conference on*, pp. 174–179, IEEE, 2014.

95. Mohd, M. H., M. Kashima, K. Sato and M. Watanabe, "Mental stress recognition based on non-invasive and non-contact measurement from stereo thermal and visible sensors", *International Journal of Affective Engineering*, Vol. 14, No. 1, pp. 9–17, 2015.

96. Chen, T., P. Yuen, M. Richardson, G. Liu and Z. She, "Detection of Psychological Stress Using a Hyperspectral Imaging Technique", *IEEE Transactions on Affective Computing*, Vol. 5, No. 4, pp. 391–405, Oct 2014.

97. Huang, M. X., J. Li, G. Ngai and H. V. Leong, "StressClick: Sensing Stress from Gaze-Click Patterns", *Proceedings of the 2016 ACM on Multimedia Conference*, MM '16, pp. 1395–1404, ACM, New York, NY, USA, 2016.

98. Egilmez, B., E. Poyraz, W. Zhou, G. Memik, P. Dinda and N. Alshurafa, "US-

tress: Understanding college student subjective stress using wrist-based passive sensing", *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pp. 673–678, March 2017.

99. Hernandez, J., R. R. Morris and R. W. Picard, "Call Center Stress Recognition with Person-specific Models", *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction - Volume Part I*, ACII'11, pp. 125–134, Springer-Verlag, 2011.

100. Garcia-Ceja, E., V. Osmani and O. Mayora, "Automatic Stress Detection in Working Environments From Smartphones x2019; Accelerometer Data: A First Step", *IEEE Journal of Biomedical and Health Informatics*, Vol. 20, No. 4, pp. 1053–1060, July 2016.

101. Cinaz, B., B. Arnrich, R. Marca and G. Tröster, "Monitoring of Mental Workload Levels During an Everyday Life Office-work Scenario", *Personal Ubiquitous Comput.*, Vol. 17, No. 2, pp. 229–239, Feb. 2013.

102. Kocielnik, R., N. Sidorova, F. M. Maggi, M. Ouwerkerk and J. H. D. M. Westerink, "Smart technologies for long-term stress monitoring at work", *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, pp. 53–58, June 2013.

103. Lee, D. S., T. W. Chong and B. G. Lee, "Stress Events Detection of Driver by Wearable Glove System", *IEEE Sensors Journal*, Vol. 17, No. 1, pp. 194–204, Jan 2017.

104. Lee, B. G. and W. Y. Chung, "Wearable Glove-Type Driver Stress Detection Using a Motion Sensor", *IEEE Transactions on Intelligent Transportation Systems*, Vol. PP, No. 99, pp. 1–10, 2016.

105. Chen, L.-l., Y. Zhao, P.-f. Ye, J. Zhang and J.-z. Zou, "Detecting driving stress in physiological signals based on multimodal feature analysis and kernel classifiers",

Expert Systems with Applications, Vol. 85, pp. 279–291, 2017.

106. Munla, N., M. Khalil, A. Shahin and A. Mourad, "Driver stress level detection using HRV analysis", *2015 International Conference on Advances in Biomedical Engineering (ICABME)*, pp. 61–64, Sept 2015.

107. Ghaderi, A., J. Frounchi and A. Farnam, "Machine learning-based signal processing using physiological signals for stress detection", *2015 22nd Iranian Conference on Biomedical Engineering (ICBME)*, pp. 93–98, Nov 2015.

108. Keshan, N., P. V. Parimi and I. Bichindaritz, "Machine learning for stress detection from ECG signals in automobile drivers", *2015 IEEE International Conference on Big Data (Big Data)*, pp. 2661–2669, Oct 2015.

109. Wang, J.-S., C.-W. Lin and Y.-T. C. Yang, "A k-nearest-neighbor classifier with heart rate variability feature-based transformation algorithm for driving stress recognition", *Neurocomputing*, Vol. 116, pp. 136 – 143, 2013, advanced Theory and Methodology in Intelligent ComputingSelected Papers from the Seventh International Conference on Intelligent Computing (ICIC 2011).

110. Goldberger, A. L., L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet", *Circulation*, Vol. 101, No. 23, pp. e215–e220, 2000.

111. Healey, J. A. and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors", *IEEE Transactions on Intelligent Transportation Systems*, Vol. 6, No. 2, pp. 156–166, June 2005.

112. Castaldo, R., W. Xu, P. Melillo, L. Pecchia, L. Santamaria and C. James, "Detection of mental stress due to oral academic examination via ultra-short-term HRV analysis", *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 3805–3808, Aug 2016.

113. Gjoreski, M., H. Gjoreski, M. Lutrek and M. Gams, "Automatic Detection of Perceived Stress in Campus Students Using Smartphones", *2015 International Conference on Intelligent Environments*, pp. 132–135, July 2015.

114. Bogomolov, A., B. Lepri, M. Ferron, F. Pianesi and A. S. Pentland, "Pervasive stress recognition for sustainable living", *2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS)*, pp. 345–350, March 2014.

115. Bauer, G. and P. Lukowicz, "Can smartphones detect stress-related changes in the behaviour of individuals?", *2012 IEEE International Conference on Pervasive Computing and Communications Workshops*, pp. 423–426, March 2012.

116. Wang, R., F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev and A. T. Campbell, "StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students Using Smartphones", *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '14, pp. 3–14, ACM, New York, NY, USA, 2014.

117. Bogomolov, A., B. Lepri, M. Ferron, F. Pianesi and A. S. Pentland, "Daily Stress Recognition from Mobile Phone Data, Weather Conditions and Individual Traits", *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, pp. 477–486, ACM, New York, NY, USA, 2014.

118. Ciman, M. and K. Wac, "Individuals' stress assessment using human-smartphone interaction analysis", *IEEE Transactions on Affective Computing*, Vol. PP, No. 99, pp. 1–1, 2016.

119. Gjoreski, M., H. Gjoreski, M. Luštrek and M. Gams, "Continuous Stress Detection Using a Wrist Device: In Laboratory and Real Life", *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, UbiComp '16, pp. 1185–1193, ACM, New York, NY, USA, 2016.

120. Gjoreski, M., M. Luštrek, M. Gams and H. Gjoreski, "Monitoring stress with a wrist device using context", *Journal of biomedical informatics*, Vol. 73, pp. 159–170, 2017.

121. Sysoev, M., A. Kos and M. PogaăźNik, "Noninvasive Stress Recognition Considering the Current Activity", *Personal Ubiquitous Comput.*, Vol. 19, No. 7, pp. 1045–1052, Oct. 2015.

122. Muaremi, A., A. Bexheti, F. Gravenhorst, B. Arnrich and G. Troster, "Monitoring the impact of stress on the sleep patterns of pilgrims using wearable sensors", *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pp. 185–188, June 2014.

123. Sano, A. and R. W. Picard, "Stress Recognition Using Wearable Sensors and Mobile Phones", *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 671–676, Sept 2013.

124. Muaremi, A., B. Arnrich and G. Tröster, "Towards measuring stress with smartphones and wearable devices during workday and sleep", *BioNanoScience*, Vol. 3, No. 2, pp. 172–183, 2013.

125. Hovsepian, K., M. al'Absi, E. Ertin, T. Kamarck, M. Nakajima and S. Kumar, "cStress: Towards a Gold Standard for Continuous Stress Assessment in the Mobile Environment", *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pp. 493–504, ACM, New York, NY, USA, 2015.

126. Lu, H., D. Frauendorfer, M. Rabbi, M. S. Mast, G. T. Chittaranjan, A. T. Campbell, D. Gatica-Perez and T. Choudhury, "StressSense: Detecting Stress in Unconstrained Acoustic Environments Using Smartphones", *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pp. 351–360, ACM, New York, NY, USA, 2012.

127. Vildjiounaite, E., J. Kallio, V. Kyllönen, M. Nieminen, J. Mäntyjärvi and G. Gimel'farb, "Unobtrusive stress detection on the basis of smartphone usage data", *Personal and Ubiquitous Computing*, Jan 2018.

128. Gimpel, H., C. Regal and M. Schmidt, "myStress: Unobtrusive Smartphone-Based Stress Detection.", *European Conference on Information Systems*, 2015.

129. Maier, E., U. Reimer, E. Laurenzi, M. Ridinger and T. Ulmer, "A Mobile Solution for Stress Recognition and Prevention", *Proc. Int'l Conf. Health Informatics (HealthInf)*, pp. 428–433, 2014.

130. Kostopoulos, P., A. I. Kyritsis, M. Deriaz and D. Konstantas, "Stress Detection Using Smart Phone Data", *eHealth 360°*, pp. 340–351, Springer, 2017.

131. Posner, J., J. A. Russell and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology", *Development and Psychopathology*, Vol. 17, No. 3, p. 715–734, 2005.

132. Reimer, U., E. Laurenzi, E. Maier and T. Ulmer, "Mobile Stress Recognition and Relaxation Support with SmartCoping: User-Adaptive Interpretation of Physiological Stress Parameters", *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2017.

133. Taylor, S., N. Jaques, W. Chen, S. Fedor, A. Sano and R. Picard, "Automatic identification of artifacts in electrodermal activity data", *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1934–1937, IEEE, 2015.

134. Greco, A., G. Valenza, A. Lanata, E. P. Scilingo and L. Citi, "cvxEDA: A Convex Optimization Approach to Electrodermal Activity Processing", *IEEE Transactions on Biomedical Engineering*, Vol. 63, No. 4, pp. 797–804, April 2016.

135. Bornoiu, I. V. and O. Grigore, "2013 8th International Symposium on Advanced Topics in Electrical Engineering (ATEE)", pp. 1–4, May 2013.

136. Tarvainen, M. P., J.-P. Niskanen, J. A. Lipponen, P. O. Ranta-Aho and P. A. Karjalainen, "Kubios HRV–heart rate variability analysis software", *Computer methods and programs in biomedicine*, Vol. 113, No. 1, pp. 210–220, 2014.

137. Vollmer, M., *MarcusVollmer/HRV Toolbox*, 2018, `https://www.github.com/MarcusVollmer`, accessed at December 2018.

138. Lomb, N. R., "Least-squares frequency analysis of unequally spaced data", *Astrophysics and space science*, Vol. 39, No. 2, pp. 447–462, 1976.

139. Hall, M. A. and G. Holmes, "Benchmarking attribute selection techniques for discrete class data mining", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 15, No. 6, pp. 1437–1447, Nov 2003.

140. Holmes, G., A. Donkin and I. H. Witten, "WEKA: a machine learning workbench", *Proceedings of ANZIIS '94 - Australian New Zealnd Intelligent Information Systems Conference*, pp. 357–361, Nov 1994.

141. Kotsiantis, S., D. Kanellopoulos, P. Pintelas *et al.*, "Handling imbalanced datasets: A review", *GESTS International Transactions on Computer Science and Engineering*, Vol. 30, No. 1, pp. 25–36, 2006.

142. Eibe, F., M. Hall and I. Witten, "The WEKA Workbench. Online Appendix for" Data Mining: Practical Machine Learning Tools and Techniques", *Morgan Kaufmann*, 2016.

143. *Python Keras. [Online]*, 2019, `https://keras.io/installation`, accessed at February 2020.

144. Muaremi, A., A. Bexheti, F. Gravenhorst, B. Arnrich and G. Tröster, "Moni-

toring the impact of stress on the sleep patterns of pilgrims using wearable sensors", *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pp. 185–188, June 2014.

145. Can, Y. S., D. Gokay, D. R. Kılıç, D. Ekiz, N. Chalabianloo and C. Ersoy, "How Laboratory Experiments Can Be Exploited for Monitoring Stress in the Wild: A Bridge Between Laboratory and Daily Life", *Sensors*, Vol. 20, No. 3, 2020.

146. Collins, J., H. Regenbrecht, T. Langlotz, Y. Said Can, C. Ersoy and R. Butson, "Measuring Cognitive Load and Insight: A Methodology Exemplified in a Virtual Reality Learning Context", *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 351–362, Oct 2019.

147. Zangróniz, R., A. Martínez-Rodrigo, J. Pastor, M. López and A. Fernández-Caballero, "Electrodermal activity sensor for classification of calm/distress condition", *Sensors*, Vol. 17, No. 10, p. 2324, 2017.

148. Cohen, S., T. Kamarck and R. Mermelstein, "A global measure of perceived stress", *Journal of health and social behavior*, pp. 385–396, 1983.

149. Can, Y. S., N. Chalabianloo, D. Ekiz, J. Fernandez-Alvarez, G. Riva and C. Ersoy, "Personal stress-level clustering and decision-level smoothing to enhance the performance of ambulatory stress detection with smartwatches", *IEEE Access*, pp. 1–1, 2020.

150. Can, Y. S., B. Arnrich and C. Ersoy, "Stress detection in daily life scenarios using smart phones and wearable sensors: A survey", *Journal of Biomedical Informatics*, Vol. 92, p. 103139, 2019.

151. Brantley, P. J. and G. N. Jones, *Daily stress inventory: Professional manual*, Psychological Assessment Resources Odessa, FL, 1989.

152. Hokanson, J., S. Stader, H. Flynn and R. Tate, "The Daily Experiences Sur-

vey: An instrument for daily recordings of multiple variables associated with psychopathology", *Unpublished manuscript, Florida State University*, 1992.

153. Hard, S. and Stavenland, "Development of NASA-TLX (task load index): results of empirical and theoretical research.", *Advances in Psychology*, Vol. 52, pp. 139 – 183, 1988.

154. Meyer, T. J., M. L. Miller, R. L. Metzger and T. D. Borkovec, "Development and validation of the penn state worry questionnaire", *Behaviour research and therapy*, Vol. 28, No. 6, pp. 487–495, 1990.

155. Ganor, T., N. Mor and J. D. Huppert, "Development and validation of a State-Reappraisal Inventory (SRI).", *Psychological assessment*, Vol. 30, No. 12, p. 1663, 2018.

156. Marchetti, I., N. Mor, C. Chiorri and E. H. Koster, "The brief state rumination inventory (BSRI): Validation and psychometric evaluation", *Cognitive Therapy and Research*, Vol. 42, No. 4, pp. 447–460, 2018.

157. Hart, S. G., "NASA Task load Index (TLX). Volume 1.0; Paper and pencil package", *NASA Ames Research Center*, 1986.

158. Lewis, E. J., K. L. Yoon and J. Joormann, "Emotion regulation and biological stress responding: associations with worry, rumination, and reappraisal", *Cognition and Emotion*, Vol. 32, No. 7, pp. 1487–1498, 2018.

159. World Medical Association, "World medical association declaration of Helsinki: Ethical principles for medical research involving human subjects", *Journal of the American Medical Association*, Vol. 310, No. 20, pp. 2191–2194, 2013.

160. Howarth, E. and M. S. Hoffman, "A multidimensional approach to the relationship between mood and weather", *British Journal of Psychology*, Vol. 75, No. 1, pp. 15–23, Feb 1984.

161. Sanders, J. L. and M. S. Brizzolara, "Relationships between Weather and Mood", *The Journal of General Psychology*, Vol. 107, No. 1, pp. 155–156, jul 2010.

162. *Windguru. [Online]*, 2019, `https://www.windguru.cz/`, accessed at February 2020.

163. *Physical Activity Reduces Stress*, 2018, `https://adaa.org/understanding-anxiety/related-illnesses/other-related-conditions/stress/physical-activity-reduces-st`, accessed at December 2018.

164. Smith, A. P., E. J. K. Wadsworth, C. Shaw, S. Stansfeld, K. Bhui and K. Dhillon, "Ethnicity, work characteristics, stress and health", , 2005, `https://www.hse.gov.uk/research/rrpdf/rr308.pdf`, accessed at: February 2020.

165. Gore, B. F. and R. H. Kim, *NASA TLX for iOS User Guide v1.0*, 2018, `https://humansystems.arc.nasa.gov/groups/TLX/downloads/NASA-TLX-for-iOS-User-Guide-Final.pdf`, accessed at December 2018.

166. Cao, A., K. K. Chintamani, A. K. Pandya and R. D. Ellis, "NASA TLX: Software for assessing subjective mental workload", *Behavior Research Methods*, Vol. 41, No. 1, pp. 113–117, Feb 2009.

167. Sharek, D., "A Useable, Online NASA-TLX Tool", *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 55, No. 1, pp. 1375–1379, 2011.

168. Cinaz, B., B. Arnrich, R. Marca and G.Tröster, "Monitoring of mental workload levels during an everyday life office-work scenario", *Personal Ubiquitous Comput.*, Vol. 17, p. 229.

169. Sano, A., S. Taylor, A. W. McHill, A. J. Phillips, L. K. Barger, E. Klerman and R. Picard, "Identifying Objective Physiological Markers and Modifiable Behaviors

for Self-Reported Stress and Mental Health Status Using Wearable Sensors and Mobile Phones: Observational Study", *J Med Internet Res*, Vol. 20, No. 6, p. e210, Jun 2018.