STRESS RECOGNITION IN EVERYDAY LIFE

by

Deniz Ekiz

B.S., Computer Engineering, Istanbul Bilgi University, 2015

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

Graduate Program in Computer Engineering

Boğaziçi University

2019

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisor Prof. Cem Ersoy for the continuous support of my M.S. study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I want to thank my previous advisor Prof. Bert Arnrich for helping me to start my study.

I have spent amazing years in my graduate study. I would like to thank my teammates: Yekta Said Can and Niaz Chalabianloo. I would like to thank every member of NETLAB for their ultimate support.

I would like to thank my family: my wife Gözde Ekiz, my mother Ayşin Ekiz, my father Hüseyin Ekiz and my sister Aslı Moral, my brother in law Resat Moral and my grandparents Nurcan and Kemal Özaş for supporting me throughout writing this thesis and my life in general.

This work is supported by the Turkish Ministry of Development under the TAM Project number DPT2007K120610. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 722022.

ABSTRACT

STRESS RECOGNITION IN EVERYDAY LIFE

In the last decades, most of the diseases in modern society are caused by stress. This is the reason researchers want to detect and alleviate stress in daily life as early as possible. With the advance of technology, smartphones, smartbands, watches have become integral items of our daily lives. The research question that whether detecting stress with these widely used wearable devices is possible has arisen. The research has started in laboratory environments and recently a number of works have taken a step outside the laboratory to real life. In this thesis, we employed two different case studies. First, we collected 339 hours of physiological data and 7119 workload survey questions from 17 participants in their real-life environments with Samsung Gear S2 smartwatch. The duration of this experiment for the participants was a month on average. Heart rate variability and accelerometer features are used to evaluate the level of stress. Second, we conducted a context-driven stress measurement experiment, we collected 672 hours (in 9 days) of physiological data from 21 participants of an algorithmic competition event. This event has free, lecture and contest sessions. By using heart rate, skin conductance, and accelerometer signals, we achieved approximately 98% accuracy of discriminating contest stress, the cognitive load (lecture) and relaxed activities by using machine learning methods.

ÖZET

GÜNLÜK HAYATTA STRES TANIMA

Son on yılda, modern toplumdaki hastalıkların çoğu stresden kaynaklanmaktadır. Bu yüzden araştırmacılar günlük yaşamdaki stresi mümkün olduğunca erken tespit etmek ve azaltmak istemektedir. Teknolojinin gelişmesiyle birlikte akıllı telefonlar, akıllı bileklikler, akıllı saatler günlük hayatımızın ayrılmaz bir parçası haline geldi. Bu yaygın olarak kullanılan giyilebilir cihazlar ile stresin tespit edilip edilemeyeceğinin araştırma konusu olmuştur. Araştırmalar, laboratuvar ortamında başlamış ve son zamanlarda laboratuvar dışında gerçek hayat uygulamaları olarak devam etmektedir.. Bu tez çalışmasında iki farklı durum çalışması yaptık. İlk önce, Samsung Gear S2 akıllı saati ile gerçek yaşam ortamlarında 17 katılımcıdan 339 saat fizyolojik veri ve 7119 iş yükü anketi sorusu topladık. Katılımcılar genellikle bu deneyi bir ay içerisinde tamamladılar. Kalp atış hızı değişkenliği ve ivmeölçer özellikleri stres seviyelerini değerlendirmek için kullanıldı. İkinci olarak, içerik odaklı stress ölçümü deneyi yaptık. Algoritma kampında 21 katılımcıdan 672 saat (9 gün içinde) fizyolojik veri topladık. Bu etkinlikte serbest, ders ve yarışma oturumları vardır. Kalp atış hızı, cilt iletkenliği ve ivmeölçer sinyalleri ile makine öğrenmesi yöntemlerini kullanarak yarışma, ders ve serbest zaman aktivitelerinden yaklaşık %98 başarım elde ettik. İkinci olarak,

TABLE OF CONTENTS

ACKNOWLEDGEMENTS . . . iii
ABSTRACT . . . iv
ÖZET . . . v
LIST OF FIGURES . . . viii
LIST OF TABLES . . . x
LIST OF ACRONYMS/ABBREVIATIONS . . . xiii
1. INTRODUCTION . . . 1
2. BACKGROUND . . . 4
2.1. What is Stress? . . . 4
2.2. Automatic Measurement of Stress . . . 6
2.2.1. Subjective Assessment . . . 6
2.2.2. Laboratory Procedure . . . 7
2.3. Physiological Signals . . . 9
2.3.1. Acceleration and Movement . . . 9
2.3.2. Skin Temperature . . . 9
2.3.3. Electrodermal Activity (EDA) . . . 9
2.3.4. Heart Rate Activity . . . 10
2.4. Unobtrusive Physiological Measurement . . . 11
3. RELATED WORKS . . . 13
4. STRESS LEVEL DETECTION SYSTEM . . . 18
4.1. Samsung Gear Series . . . 21
4.2. Empatica E4 Smartband . . . 23
4.3. Ethical issues & Informed consent forms . . . 24
4.4. Problems Related to the Movement and Improper Placement of Devices 24
4.5. Electrodermal Activity Signal Preprocessing and Feature Extraction Tools 25
4.6. Heart Activity Signal Preprocessing and Feature Extraction Tools . . . 26
4.7. Accelerometer Processing and Feature Extraction . . . 28
4.7.1. Machine Learning Tools . . . 28
5. CASE STUDY: WORKLOAD AND STRESS DETECTION IN THE WILD 30

5.1. Description of the Case Study 30
5.2. Methodology 31
5.2.1. Features of Collected Data of Samsung Gear S2 31
5.2.2. Self Report Questionnaire as Ground Truth 32
5.2.3. Generation of Class Labels 32
5.2.4. The Quality of Heart Rate Data 33
5.3. Results 34
5.3.1. Classification Results 34
5.3.2. Regression Results 37
5.4. Discussion 40
6. CONTINUOUS STRESS DETECTION USING WEARABLE SENSORS IN REAL LIFE SCENARIOS: ALGORITHMIC SUMMER CAMP CASE STUDY 41
6.1. Real-life group experiment setup 41
6.2. Data Collection 43
6.3. Preprocessing 43
6.4. Machine Learning Classifiers 44
6.5. Results 45
6.6. Device Type Comparison 45
6.7. Effect of Artifact Detection Percentage Threshold, Interpolation and Accumulation Window on Accuracy 48
6.8. Person independent and dependent models 51
6.9. Effect of Different Physiological Modalities 52
6.10. TLX clustering vs. Known Context, Effect of Ground Truth on Accuracy 53
6.11. Discussion 55
7. CONCLUSION 56
REFERENCES 58

LIST OF FIGURES

Figure 2.1. The heuristic model diagram of the stress mechanism [1]. The dashed lines demonstrates the feedback mechanism. 5

Figure 2.2. The 6 questions of paper based Nasa Task Load Index (TLX) [2]. 7

Figure 2.3. Screenshots of the official NASA-TLX questionnaire Iphone application interfaces. 8

Figure 2.4. An example of ECG signal. Original image from [3]. 10

Figure 2.5. The positions of the wrist-worn devices are shown. These devices are worn on left hand of the participant. The Empatica E4, the Samsung Gear S, the Samsung Gear S2 Classic and the Samsung Gear S3 Classic are presented with the physiological signals [4]. . 11

Figure 3.1. Biopac-Mp36 physiological signal recording device [5]. 14

Figure 4.1. The system design figure of stress level detection system 19

Figure 4.2. The system design figure of stress level detection system with Samsung Gear S and S2. 20

Figure 4.3. The system design figure of stress level detection system with Empatica E4 wristband. 22

Figure 4.4. The Empatica E4 smartband. Original figure from [6] 24

Figure 4.5. Decomposed EDA Signal from Empatica E4 wristband by applying cvxEDA tool. 26

Figure 5.1. An experimental procedure for daily workload sessions. 31

Figure 5.2. The histogram of RR intervals quality in terms of the ratio of non-interpolated RR intervals in percentage (%). Each quality measure is calculated for each session. 34

Figure 5.3. The boxplot of RR intervals quality in terms of the ratio of non-interpolated RR intervals in percentage (%). Each quality measure is calculated for each session. 35

Figure 6.1. Daily training classes before the contest. 42

Figure 6.2. Inzva Final contests prize ceremony. 43

Figure 6.3. A view of smart watches and wristbands before data collection, charged and ready to use. 44

Figure 6.4. Percentage of the remaining data (for both device types) after the artifacts are removed versus the different percentage thresholds of artifact detection. 48

LIST OF TABLES

Table 4.1. Heart rate variability features and their definitions. 27

Table 5.1. The distribution of class labels for each label generation technique 33

Table 5.2. The three class classification results of labels generated from weighted score with LDA, SVM and MLP classifiers 35

Table 5.3. The three class classification results of labels generated from frustration score with LDA, SVM and MLP classifiers 36

Table 5.4. The three class classification results of labels generated from mental demand score with LDA, SVM and MLP classifiers 36

Table 5.5. The two class classification results of labels generated from weighted score with LDA, SVM and MLP classifiers 36

Table 5.6. The two class classification results of labels generated from frustration score with LDA, SVM and MLP classifiers 36

Table 5.7. The two class classification results of mental demand with LDA, SVM and MLP classifiers 37

Table 5.8. Regression results on balanced dataset with heart rate variability and accelerometer features. 38

Table 5.9. Regression results on balanced dataset with heart rate variability and accelerometer features, feature selection is applied. 39

Table 5.10. Regression results on balanced dataset with heart rate variability features 39

Table 5.11. Regression results on balanced dataset with heart rate variability features, feature selection is applied 39

Table 6.1. Stress Level Classification Accuracy, f-Measure, precision and recall values with Different ML Algorithms - 3 class - HR + EDA +ACC - Empatica E4. 46

Table 6.2. Stress Level Classification Accuracy, f-Measure, precision and recall values with Different ML Algorithms - 3 class - HR +ACC - Empatica E4. 46

Table 6.3. Stress Level Classification Accuracy, f-Measure, precision and recall values with Different ML Algorithms - 3 class - HR +ACC - All. . 47

Table 6.4. Effect of the used device to 3-class stress level classification accuracy when heart activity and accelerometer data are used together (with context HR + ACC). 47

Table 6.5. Effect of the used device to 3-class stress level classification accuracy when only heart activity signal is used (without context only HR data). 47

Table 6.6. Effect of the Length of the Aggregation Window on Classification Accuracies. 49

Table 6.7. Classification accuracies vs. changing percentage based artifact detection and filtering rules. 50

Table 6.8. Classification accuracies when removed artifacts are replaced with interpolation vs. when they are removed. 51

Table 6.9. Classification accuracies of General and Person Specific Models. . . 52

Table 6.10. Stress Detection Accuracies with Different ML Algorithms - 3 class classification. On the left side, stress recognition results which are only using HR and EDA signals are presented. On the right side, context information with accelerometer data is also added. 53

Table 6.11. Classification accuracies when different ground truths are used. On the left, known context information (Free:1, Lecture:2, Contest:3) is used as class labels. On the right, subjective ground truths are used as class labels. 54

LIST OF ACRONYMS/ABBREVIATIONS

ACC	Accelerometer
ANS	Autonomic Nervous System
BVP	Blood Volume Pressure
ECG	Electrocardiogram
EDA	Electrodermal Activity
EEG	Electroencephalogram
FFT	Fast Fourier Transform
HF	High Frequency
HR	Heart Rate
HRM	Heart Rate Monitor
HRV	Heart Rate Variability
GPS	Global Positioning System
GSR	Galvanic Skin Response
IBI	Interbeat Interval
LF	Low Frequency
PCA	Principal Component Analysis
pHF	Prevalent High Frequency
pLF	Prevalent Low Frequency
PPG	Photoplethysmogram
PNS	Parasympathetic Nervous System
RF	Random Forest
RMSS	Root Mean Square of Successive Difference of the RR intervals
TSST	Trier Social Stress Test
TINN	Triangular Interpolation of RR interval histogram
IMU	Inertial Measurement Unit
kNN	K-nearest Neighbour
ML	Machine Learning
MLP	Multilayer Perceptron

NN	Neural Network
SVM	Support Vector Machine
VHF	Very High Frequency
VLF	Very Low Frequency

1. INTRODUCTION

Smart-sensing, pervasive and ubiquitous technologies become more accessible to the population during the last decade. There are a lot of sensing and smart devices emerging in the market. Most of them offer new opportunities to gather physiological data of the users. New sensing technologies offer more personal health monitoring options. Smartphones, smartwatches and smart bands are equipped with many sensors. These devices are available in the market with different price levels. Some examples to these devices can be Samsung Gear S, S2 and S3 smartwatches [7] and Empatica E4 smartband [6]. Three-dimensional acceleration unit, pedometer, heart rate measuring unit, barometer, the global positioning system can be found in most the modern smartphones and smartwatches. Thanks to these sensing units, deductions can be made about about an individual's daily routines and health condition.

Work-related stress causes physical and mental health problems. Heavy workload, long working hours, poor support at work can be some of the many causes of work-related stress. There is a need for robust stress policy in workplaces. The chronic stress should be eliminated, otherwise, it can lead to diseases [8]. The World Health Organization concluded that psychological stress is one of the most significant health problems in the 21st-century [9].

Tracking the psychological stress levels of individuals is not straight forward for several reasons. First, psychological stress or mental stress is highly subjective. The response of an individual to a stressor event may alter in another. Second, in real life settings, it is challenging to find the ground truth for assessing the stress level. Due to the tremendous subjectivity and the perpetual intrinsic of the stress mechanism, it is hard to determine the beginning, the span, and severity of a stress incident. Last, in order to measure the stress multimodal sensing is required. Because, there are three components of stress which are affective, behavioral and biological responses [1].

This study aims to automatically measure the stress level of an individual. Emotions and stress can induce bodily reactions which can be measured with sensory units and converted to physiological signals. Analysis of these signals can be used to measure the stress and negative emotions. The proposed system has some contributions to the related works in the literature which are as follows.

- *Unobtrusive and comfortable design* is an important building block of this system. Solutions which are not comfortable makes the adoption more difficult. Utilizing uncomfortable and bulky devices cannot be applied to most of the population. Our methodology is also robust in real-life environments such as the office, public transportation, the school, and the public places.
- *Low-cost solution* is important for the population. A solution should not be very expensive in order to be applicable to the mass population. Smartwatches are the most popular type of wearable computers [10]. Our solution can be applied to the commercially available Samsung Gear Series. Comparing to the high-end sensor systems, Empatica E4 smartband is also a relatively low-cost solution. Currently, Empatica E4 aims the research market, in the future a commercial device with such properties would be cheaper.
- *Platform independent method* is aimed in the proposed system. This system can be applicable to the devices which have the same sensory units. Thanks to this property, this system is easily portable to other similar platforms.
- *Performances of different types of devices* such as Empatica E4 smartband, Samsung Gear S and S2 smartwatches in terms of the stress level measurement are compared in this study.
- *A solution for data loss* is proposed in this work. In real-life environments, sensory units can make faulty measurements called artifacts. Since modern smartwatches are not as precise as the high-end sensor systems, a lot of artifacts can occur. Therefore, loss of data can occur if we delete the recordings that contain artifacts. In this thesis, we propose a methodology to handle the data with artifacts.

- *Efficient way for subjective assessment of workload* is proposed in this work. Different modalities of a subjective assessment questionnaire are discussed throughout the thesis.
- *Simultaneous data collection* is proposed in this work. This is the first study in automatic stress detection which provides a methodology to conduct context-driven stress measurement simultaneously. Conducting the experiment separately for each subject is expensive. Therefore, we propose a low-cost solution in this field.

The rest of the thesis is organized as follows. In Chapter 2, the basic information about the stress, objective and subjective measurement of stress, controlled stress inducing techniques and physiological signals that are used in this study are presented. In Chapter 3, the related works in the literature on the emotion and stress measurement are presented and discussed. In Chapter 4, the proposed system for measuring the stress level is explained with its limitations and considerations. In Chapter 5, we used the system proposed in Chapter 4, the training of the system is only dependent on the self-reported questionnaires, in other words, no objective context information is given. We validated our system on 339 hours of physiological data and 7119 workload survey questions from 17 subjects in their real-life environments with Samsung Gear S2 smartwatches. For each device type, data collection procedure and implementation are described. In Chapter 6, we present an implementation of the system proposed in Chapter 4. In order to validate the system, 672 hours of physiological data and 7119 workload questions in 9 days is collected from 21 subjects using Samsung Gear S and S2 smartwatches, and the Empatica E4 smartband. The case study is made in an algorithmic summer camp for university and high school students. The performance of the system is analyzed and minor modifications of the system are made for each device type. Finally, in Chapter 7, conclusions and future works are presented.

2. BACKGROUND

In this chapter, we present the required background information for automatic stress detection which consists of several building blocks. First, we need to define the psychological stress, when the mechanism starts and how the recovery is done. We also present the difference between benign stress and chronic stress which can lead to diseases and poor health practices. Second, we present assessment and controlled alleviation techniques of stress. We will propose an automatic stress measurement system in Chapter 4, therefore, we should have a validation procedure to measure the performance of the system. Third, we provide the information on the physiological signals that the devices used in this study can collect. These are accelerometer, skin temperature, electrodermal activity (EDA), heart rate activity. Finally, since, one of the important components of the system proposed in this study is the unobtrusive design, unobtrusive sensing tools are presented. Emerging technologies create an opportunity to sense physiological responses with highly compact devices instead of uncomfortable sensor systems.

2.1. What is Stress?

Stress is a broadly employed subject however it is not agreeable for experts because it is biased and difficult to determine [11]. Still, if specialists cannot describe stress, quantifying it is not possible. The definition of stress is "a physical, chemical, or emotional factor that causes bodily or mental tension and may be a factor in disease causation". We can informally describe the stress as the organism's behaviour of acting to a difficult or uncertain event. [12].

Cerebellum starts the stress response to observations of receptive glands, for example the ear, eye and nose. While the organism detects a menace which might be existing or made-up, protective operation of the organism starts a fast, automated process termed as the stress, quantifying it is not possible. The definition of stress is "fight-or-flight" reflex or the stress response to defend itself. Cerebellum instantly

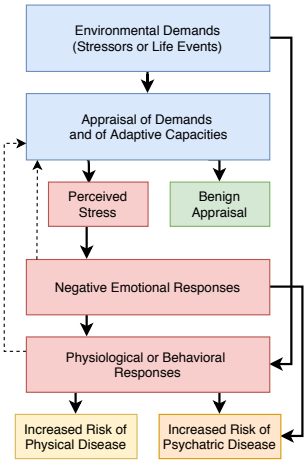

Figure 2.1: The heuristic model diagram of the stress mechanism [1]. The dashed lines demonstrates the feedback mechanism.

senses a distress signal to the hypothalamus. Control centre of the cerebellum is known as Hypothalamus which controls uncontrolled organism performs through the autonomic nervous system (ANS) which forms from couple parts which are the sympathetic nervous system (SNS) and the parasympathetic nervous system (PNS). SNS can be regarded as to an accelerator in an automobile. Still, if specialists cannot describe stress, quantifying it is not possible. The definition of stress is "fight-or-flight" circumstances, after receiving the distress signal, hypothalamus activates SNS. SNS delivers stress hormones such as epinephrine and cortisol, which stimulate the body to function in pressure circumstances. The pulse rises, muscles strain, blood pressure increases, breathing rate rises. Oxygen in the cerebellum rises and this causes functions to become sharper. Increased level of glucose in blood causes a person to be more active. Abovementioned turns enhance strength and endurance, reduces the time of the response, and increases the concentration. If the stress response works as desired, it helps an individual to keep awake and concentrated. During the threat disappears, PNS reduces the stress response. Stress assists the person to endure in serious situations. Stress might be necessary during challenging circumstances some of which are a presentation at the workplace, a test in the school. In important circumstances, stress can assist and aid us. Though, following a certain level, stress is no longer useful, contrarily, stress begins ruins the well-being, mood, fertility and life quality of a person. The mechanism of the stress is shown in Figure 2.1.

The reason for the distress is that the human nervous system cannot discriminate within emotional and physical threats [13]. In emotional circumstances including, but not limited to an argument with a colleague or a deadline for a project, the nervous system acts like the organism is in fatal circumstances. In case that, this situation shifts chronic and an individual stresses out more, organism would be stressed frequently, and this situation may become into severe health complications.

2.2. Automatic Measurement of Stress

In order to validate automatic stress measurement system, researchers need a definition of the ground-truth. However, such ground-truth is still an open question. In this thesis, some case studies including self-reported questionnaires, implementation of laboratory procedures and in the wild measurements were applied.

2.2.1. Subjective Assessment

NASA Task Load Index (TLX) [14] is used to measure the perceived workload of individuals. First, the subject has to rate each workload phase with 6 items on a scale from 1 to 20 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, the subject is asked to indicate which of the items represents the most important contributor to the workload. Based on these ratings, the total workload was computed as a weighted average. Nasa Task Load Index (TLX) can be implemented on mobile phone [15] , paper [14] or computer [16] [17]. The six questions of the paper-based Nasa-TLX is shown in Figure 2.2.

In this thesis, we used the official application of the NASA-TLX [15] for Iphone . The most important six screens of the application are presented in Figure 2.3. The first screen is the start screen of the application where user can start the NASA-TLX questionnaire by taping *Start* button. The application brings the study participant to the instruction page. [15]. The participant advances to the next view by taping the *Next* button and can go back to the previous screen by tapping the *Back* button

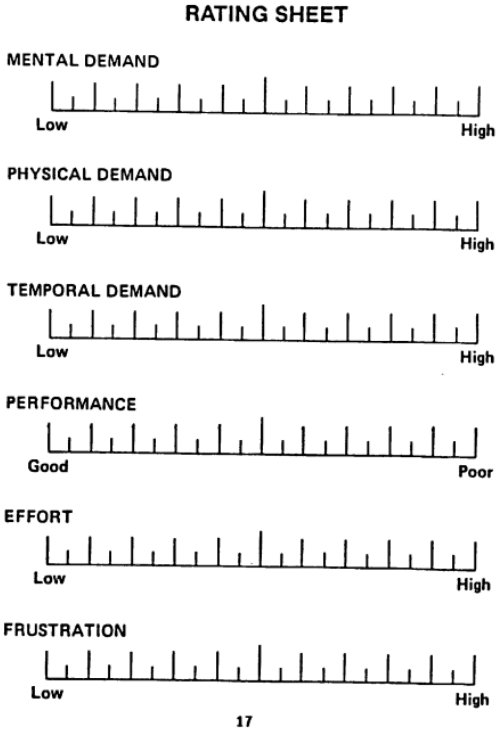

Figure 2.2: The 6 questions of paper based Nasa Task Load Index (TLX) [2].

on the upper left-hand corner of the view. The participant may quit the NASA-TLX questionnaire at any time by tapping *Quit* on the upper right-hand corner. After the instructions screen, if Pairwise Comparisons are enabled in the settings, the application will show two workload factors in a single page, the user can look at the definition of each workload factor by clicking the *Info* button at the upper right-hand of the screen. After the user taps the selection of 15 Pairwise Comparisons, the application shows the second instructions screen about the following 6 Rating scale questions. The participant advances to the next view by taping the *Next* button. For each six rating factors, the application shows a rating scale in a single page to the user who can move the red marker by taping to the rating scale of each workload factor.

2.2.2. Laboratory Procedure

Trier social stress test is a standardized laboratory protocol used to reliably induce stress in human research participants [18]. It is a combination of procedures that were previously known to induce stress, but previous procedures did not do so reliably. There are some versions of the Trier Social Stress Test (TSST) [19]. Most recently

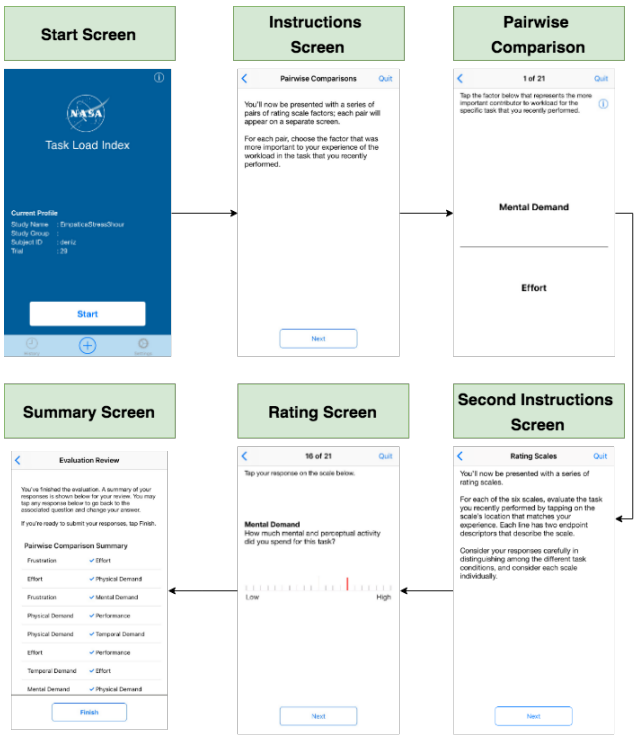

Figure 2.3: Screenshots of the official NASA-TLX questionnaire Iphone application interfaces.

applied TSST takes about 15 minutes which includes three 5 minutes long parts. The participant is equipped with a heart rate monitor (HRM). The blood sample of the participant is collected. The experiment starts when the participant enters the interview room where there is a panel of three jury members are sitting, along with a video camera and audio recorder. In the first five minutes, the participant is directed to make a presentation in front of the jury. Judges are trained to maintain neutral expressions during the test. If the participant finishes the presentation before the five minute ends, the jury asks him or her to continue until the first five minute ends. In the second five minute, the participant is asked to make an arithmetic task. Generally, the jury asks the participant to count backwards from 1022 in steps of 13. If a mistake is made, he/she must start again from the beginning, until the five-minute interval ends. In the last five minute, the jury debriefs the participant that it was a test on purpose for measuring the stress level, it was the reason that they tried to create stress and no way a reflection on his or her personal abilities. This period is called recovery. After the five minutes, saliva and blood samples are collected. Studies including different implementations also exist [20].

2.3. Physiological Signals

Physiological signals can be used to automatically detect the emotions. In this section, we present the physiological signals that we used to automatically detect the stress level of individuals. Thanks to emerging sensor technology all of these signals can be embedded in an unobtrusive device.

2.3.1. Acceleration and Movement

Body and head movements can be used to detect the emotions and arousal level [21]. Montepare *et al.* [22] demonstrated video recordings of actors expressing emotions with body movements to the 82 younger and older adults. The face of actors are blurred, thus understanding the facial expression from the video recording is not possible. All videos were silent. Participants correctly identified the emotions expressed by actors in video recordings.

Smartwatches and smartbands are equipped with inertial measurement units which can record the 3D acceleration and gyroscope signal. Thanks to these signals, body movements can be automatically recognised.

2.3.2. Skin Temperature

Skin temperature can be measured with a temperature sensor attached to the wrist. Empatica E4 is equipped with a temperature sensor, therefore an additional temperature sensor is not needed. Anger can increase the skin temperature and fear can decrease it [23].

2.3.3. Electrodermal Activity (EDA)

The body raises sweating gland activity when psychological arousal happens. Electrodermal activity (EDA) analysis is a process of estimating the electrical conductance of the epithelium which changes with its condensation amount [24]. This

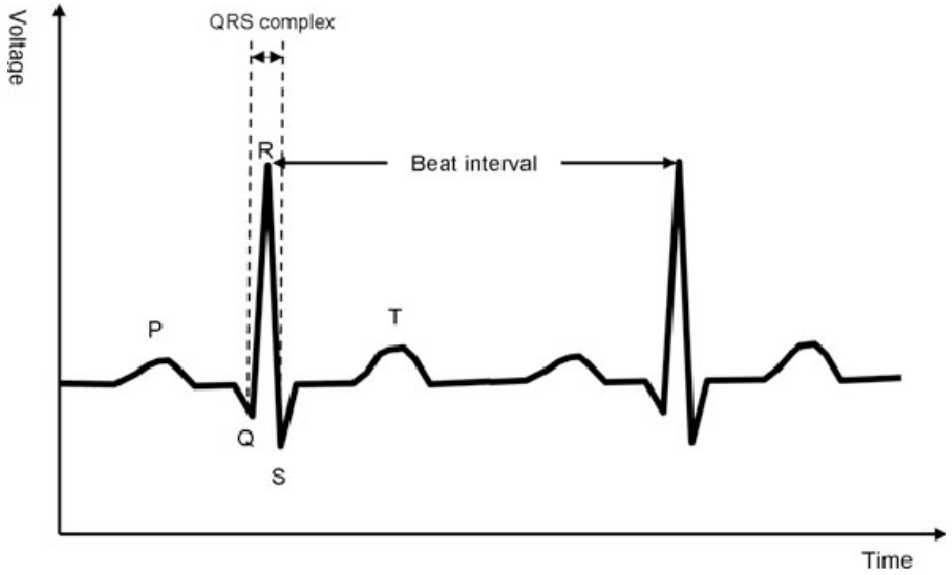

Figure 2.4: An example of ECG signal. Original image from [3].

measurement methodology is also known as the galvanic skin response (GSR), the skin conductance, the electrodermal response, the psychogalvanic reflex in the literature. There is a latency between the stimulus and galvanic skin response, in which the amount varies with respect to the type of the stimulus [25].

2.3.4. Heart Rate Activity

Traditionally heart rate activity can be measured with the electrocardiogram (ECG). An example of the ECG signal is presented in Figure 2.4. ECG provides a sinusoidal signal with recurrent peaks. The heart beats are measured by calculating the difference of R peaks, this is named as RR intervals, beat interval or inter-beat intervals (IBI). The current sensing technology of smartwatches and smartbands measures the heart rate activity with Photoplethysmography (PPG) sensor which uses a light-based technology. In commercially available devices, this technology is called the heart rate monitoring unit (HRM) that are placed in the bottom of the device it touches the wrist.

2.4. Unobtrusive Physiological Measurement

High-end sensor systems provide accurate measurement of physiological signals. These systems are still important health-care systems in hospital or diagnosis settings [26]. In real life applications for majority of the population, high end sensor systems are not suitable due to comfort and cost.

In order to collect data and deploy data collection systems in daily life of individuals, stress measurement devices should be unobtrusive [27]. People should wear these devices without being uncomfortable in their daily routines, during sleeping, meeting and everyday activities. If subjects become uncomfortable while wearing these devices, these systems will fail in the daily life deployment. Obtrusiveness can add extra stress on participants. The ideal system should collect massive amounts of data without the user even being aware of it [28] .

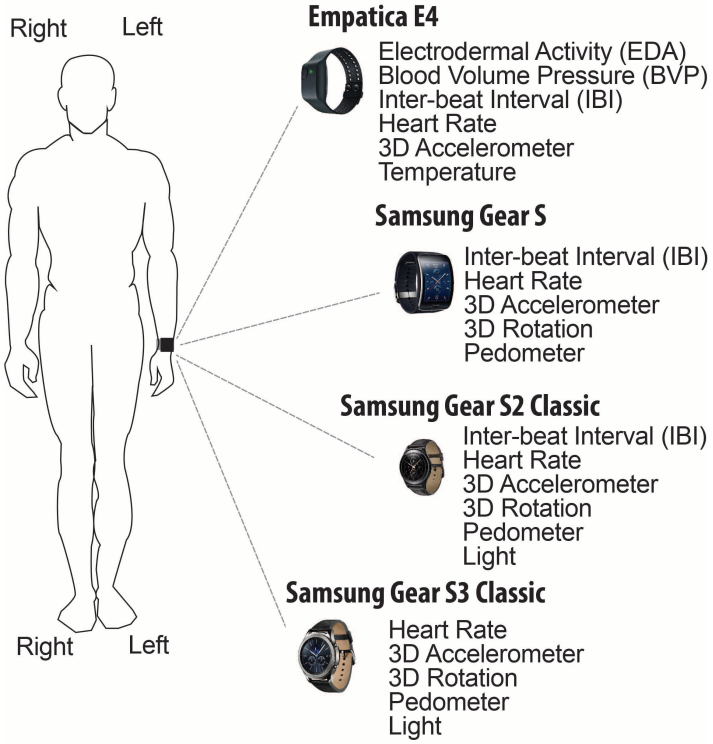

Figure 2.5: The positions of the wrist-worn devices are shown. These devices are worn on left hand of the participant. The Empatica E4, the Samsung Gear S, the Samsung Gear S2 Classic and the Samsung Gear S3 Classic are presented with the physiological signals [4].

Heart activity and skin conductance sensors provide important features for detecting stress. These physiological changes can be measured with ECG and EDA. This is known as the physiological measure of mental stress [29]. Thanks to the emerging technology, smartbands and smartwatches on the market are accessible with affordable prices. These devices can record heart activity (HR), daily activity such as the number of steps, instantaneous acceleration, galvanic skin response (GSR), ambient light. The battery life of such devices in consumer electronics is not suitable for all day recordings. For example, Samsung Gear S2 battery drains out in 2 hours when all of the sensors continuously record data. Therefore, researchers have to charge the device several times in a day. During charging, sensory parts of the device are blocked and it is not possible to record data. Samsung Gear S3 runs more than 4 hours however, due to energy consumption of Tizen operating system it does not provide inter-beat data which is important for heart rate variability (HRV). Empatica E4 can run more than 48 hours, however the price is high for the current market [6]. In this thesis, case studies including Empatica E4 and Samsung Gear S2 are given. The devices are placed on the wrist. They look like an ordinary watch. An example of placement of sensors can be seen in Figure 2.5.

3. RELATED WORKS

Ubiquitous sensing tools can be used to monitor an individual's mental and physical stress levels. It can support preventing mental disorders and mental health. Most of the existing studies about stress conditions are made in laboratory conditions [30]. Henelius *et al.* [31] showed that short-term (140 seconds) heart rate variability (HRV) metrics can be used to discriminate between a high and low level of mental workload. Hjortskov *et al.* [32] studied the heart rate variability and mental load of twelve female participants while working with computers. They reported an increase in the ratio low frequency to high frequency ratio (LF/HF) and reduction in high frequency (HF) component during mental load session compared to the relaxed condition. However, many features can be derived from the current sensing technology and examining each of them with traditional statistical methods is complex and is not suitable for automated pervasive health systems. In order to find the optimal set of features, approaches such as machine learning algorithms are required.

Continuous monitoring of work-related stress or mental workload is still in the exploratory stage. Morris *et al.* [33] proposed that each subject's baseline and stress threshold should be established in a laboratory setting using a protocol to alternatively evoke stress responses that can be used to discriminate between stress and non-stress in everyday life. Cinaz *et al.* [34] proposed a system by using a chest belt in order to get HRV parameters with the help of individual's calibration results. They presented how mental workload levels in everyday life scenarios can be discriminated by incorporating individual's calibration results. In order to build an everyday life application, minimum sensor setup and calibration are required for comfort reasons. However, their settings did not provide a working model with high classification accuracy in real office work settings. They selected participants only from the male gender.

Martinez textitet al. [35] conducted their research in a laboratory on 166 university students. They used Biopac-Mp36 which is a laboratory level physiological signal recording device as shown in Figure 3.1. They conducted one of the biggest study in

terms of number of participant. However, the device used in this study is not suitable for real-life environments.

Figure 3.1: Biopac-Mp36 physiological signal recording device [5].

Smartwatch and smartband is a more comfortable and acceptable technology than chest belts for longer periods of monitoring. In previous studies [36] [37], it is shown that Polar V800 with Polar H7 offers an alternative to the ECG for obtaining inter-beat interval time series data, and the derived HRV features reflect the HRV metrics of ECG in supine and standing positions. The modern smartphones and smart devices make HRV metrics more accessible to the population. There is still very little information about the HRV metrics of individuals in everyday-life scenarios. Ollander *et al.* [38] points that the Empatica E4 wristband had a significant loss in terms of detecting inter-beat intervals, but those time-domain features such as the mean heart rate and standard deviation of the heart rate were still well estimated, with good stress discrimination power under the Trier Social Stress Test protocol in classical laboratory settings. The loss generally occurs when the subject is performing a task, this situation affects the frequential HRV features most. The results come from 12 subjects. They did not apply an inter-beat correction algorithm, in other words, they analyzed the data directly as it was given.

The literature on real-life mood and stress detection with wrist-worn band or watches is still limited. Jaques *et al.* [39] tried to model 68 students well-being and recorded physiological signals, location, smartphone logs, and survey responses to behavioural questions over a month. They removed medium scores by excluding 40% of scores and achieved 70% classification accuracy of self-reported happiness. Sano *et*

al. [40] collected five days of data for 18 participants. The data includes a wrist sensor (accelerometer and skin conductance), mobile phone usage (call, short message service, location and screen on/off) and surveys (stress, mood, sleep, tiredness, general health, alcohol or caffeinated beverage intake and electronics usage). They achieved 75% accuracy by using mobile phone features. However, due to privacy concerns of individuals mobile phone usage data is not well suited for daily life applications. In [41], they applied a laboratory experiment on 10 participants for 30 minutes with unobtrusive physiological sensors and achieved 75% classification accuracy between low and high stress.

Ciman *et al.* [42] proposed a stress detection system by using statistics derived from smart mobile phone usage of the participants. The trial is partitioned into two sections which are conducted in the controlled laboratory and real-life environments. An Android application with search and typing missions is deployed. Participant's typing, swiping, scrolling and tapping gestures are recorded. The stress stimuli are giving as a task. They applied the Experience Sampling Method using the Likert scale [43] with 5-points to 13 subjects. They applied different classifiers which are support vector machine, decision tree, neural network and k-nearest neighbour. The maximum accuracy is reported as 80%. In real life, the dataset formed with physical activity, real-time screen brightness and screen interaction occurrences. A smart mobile phone is used as a part of subjects daily activities. The maximum accuracy is reported as 70%. In this study, the relation between stress and application type could not be identified. They plan to repeat the same trial with different wearable devices and a remote assessment system.

Gjoreski *et al.* [44] proposed a method for continuous detection of stressful events using data provided from a commercial wrist device in both laboratory and real-life. They used the Empatica [6] wrist device. In the laboratory setting, a mental arithmetic task is given to the participant in order to induce stress. 63 feature from blood volume pressure (BVP), heart rate (HR), skin temperature (ST), galvanic skin response (GSR) and inter-beat (RR) intervals signals, were computed. They achieved 83% classification accuracy on two classes problem. Moreover, Gjoreski *et al.* proposed a model

with three stress levels which are zero, low and high. They achieved 72% classification accuracy with that model. An activity recognition model which discriminates sitting, walking, running and cycling, was used for the everyday life settings. The intention following the use of activity recognition model is to distinguish an intense movement from a stressful case by giving the context information to the everyday life stress detectors as new information. The activity recognition model is used to differentiate high physical activity from a mental stress elevated case. The day is divided into segments which the duration is an hour. Participants can record mental stress elevated cases by pushing a button. They achieved 92% classification accuracy with the model that includes the activity recognition. From 5 participants, daily life data collected for 11 days. As a future work, they plan to discriminate also the stress level. They planned a person tailored model depending on age, gender, fitness condition as a future work. They made a bigger case study in [45] where Gjoreski *et al.* created a stress recognition system for the laboratory environment with an activity recognition model. The physiological signals were recorded during the experiment. First, they asked subjects to stay relax. Then they gave a mental arithmetic task to the subjects. The calibration of the subjective assessment is made with the scores gathered from the mental arithmetic task phase. They trained their model with these data. The model make a decision every 20 minutes. Electrodermal activity (EDA), blood volume pressure (BVP), heart rate (HR), temperature, and inter-beat interval (IBI) data were obtained. The best performing model achieved a 70% recall and 95% precision. Gimpel *et al.* [46] proposed a stress recognition method by collecting smartphone data. The researchers declared that the most significant distinction from the previous studies is that their method is not based on the user declaration or supplementary wearables. Gimpel *et al.* developed an Android application and investigated the smartphone dataset. 36 features derived from the collected data. The results are not published. Though, they observed that too much smartphone usage, average battery temperature, the maximum amount of running software and rate of turning the display on are related to high stress. Additionally, authors showed that the relationship between stress and smartphone usage is not observed. Moreover, perceived stress may not reflect real stress.

Syosev *et al.* presented a study using smartphones [47]. They used audio, gyroscope, accelerometer, ambient light sensor data, screen mode changing frequency, self-assessment and activity type. The raw version of the NASA-TLX was used as a reference. In the raw version, only the 6 rating scale questions were asked. They combined an activity recognition system with a stress detection system to increase the performance of stress detection scheme. The best performing model accomplished approximately 77% classification accuracy. The addition of activity recognition model to the stress detection system contributed 3.8%.

Maier *et al.* [48] presented their work in progress. They used motion and context information data to improve the classification performance of the stress recognition scheme which uses HRV derived from ECG. When the system detects a high-stress level of participants, they can get out from the stress-inducing environment or some relaxation methods are provided to them. Their modal adapts to the user's behavior and the system performance increases in time. The application gets data from the participant about their sentiment and physical well-being. Bodily motion from the accelerometer, the position from GPS, variety of position and "time of the day" data was added to HRV in order to improve the performance of the proposed system. They applied a special kind of neural network classifier (BINN) proposed by [49]. They did not provide the performance results. As a future work, they intend to conduct another study on mental patients by adding a Perceived Stress Questionnaire (PSS).

Kostopoulos *et al.* [50], used a data collection application for a smartphone to assess stress levels. Thanks to their application, they collected context information and surveys from the smartphone. Castaldo *et al.* tried to find a threshold to cut down the duration of HRV recording by analyzing different length of inter-beat intervals data for detection of mental stress [51]. They carried out three experiments in real life and in a laboratory, in which the Stroop Color Word tests was used. Sierra *et al.* [52] applied Fuzzy logic in order to discriminate low and high stress on HRV and EDA data. They introduces their application as real-time stress assessment tool. The experiments made until this work, were focused on offline stresss asessment.

4. STRESS LEVEL DETECTION SYSTEM

In this thesis, we propose a stress level detection system. This system allows the user to learn their stress levels in their daily activities without creating any interruption or restriction. The only requirement to use this system is wearing a wrist-worn watch-like device. These devices are known as smartwatches or smartbands. The overall diagram of the proposed system is presented in Figure 4.1. The system starts with the informed consent form provided to the user. Once the user accepts the terms in the informed consent form, he/she can use the application. After this step, the data collection is started from a variety of sensors and stored in the memory of the device. Then, the artifacts of physiological signals are detected and interpolated. Next, the features are extracted from the sensory signals and fed to the machine learning algorithm for prediction. In order to use this system, pre-trained machine learning models are required. For the training of models, machine learning algorithms ran on the feature vectors with generated class labels.

We present two variants of this system in this thesis because we have different types of physiological sensing devices, these are Empatica E4, Samsung Gear S3, Samsung Gear S2 and Samsung Gear S. In this thesis, we do not propose a system for Samsung Gear S3 which does not provide RR intervals.

The diagram of the first variant of the proposed system is described in Figure 4.2. This system proposes a stress detection solution for Samsung Gear S2 and Samsung Gear S. This system can also be applied to any device which provides IBI and the acceleration signal. The first part of this system is data collection where IBI and acceleration data is gathered from sensors and recorded. IBI signals are filtered for artifacts by calculating the difference of successive RR intervals, those are more than 20% are considered as artifacts. Artifacts are interpolated with cubic spline interpolation. Acceleration and IBI signals are aligned and divided into 2 minutes long sliding windows with 50% overlap. For each window, 22 features are calculated. Respectively, the feature extraction of IBI and acceleration signals are described in Subsection 4.6

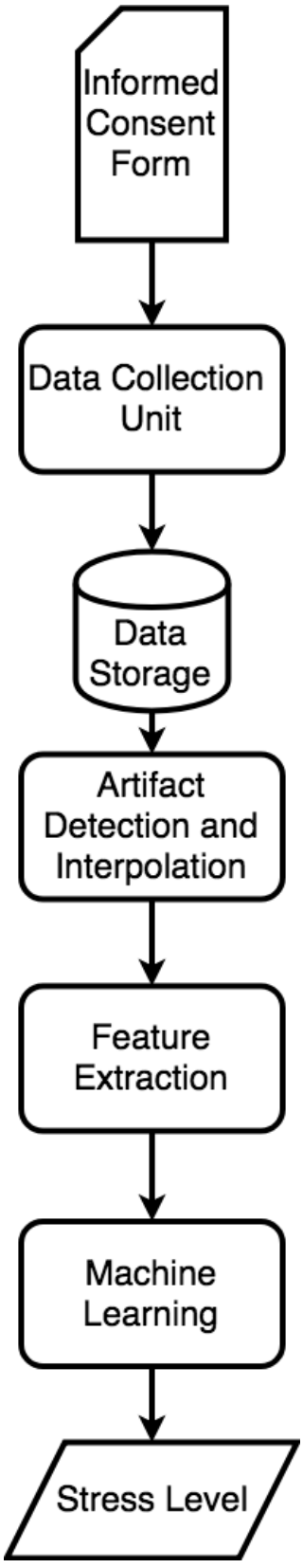

Figure 4.1: The system design figure of stress level detection system

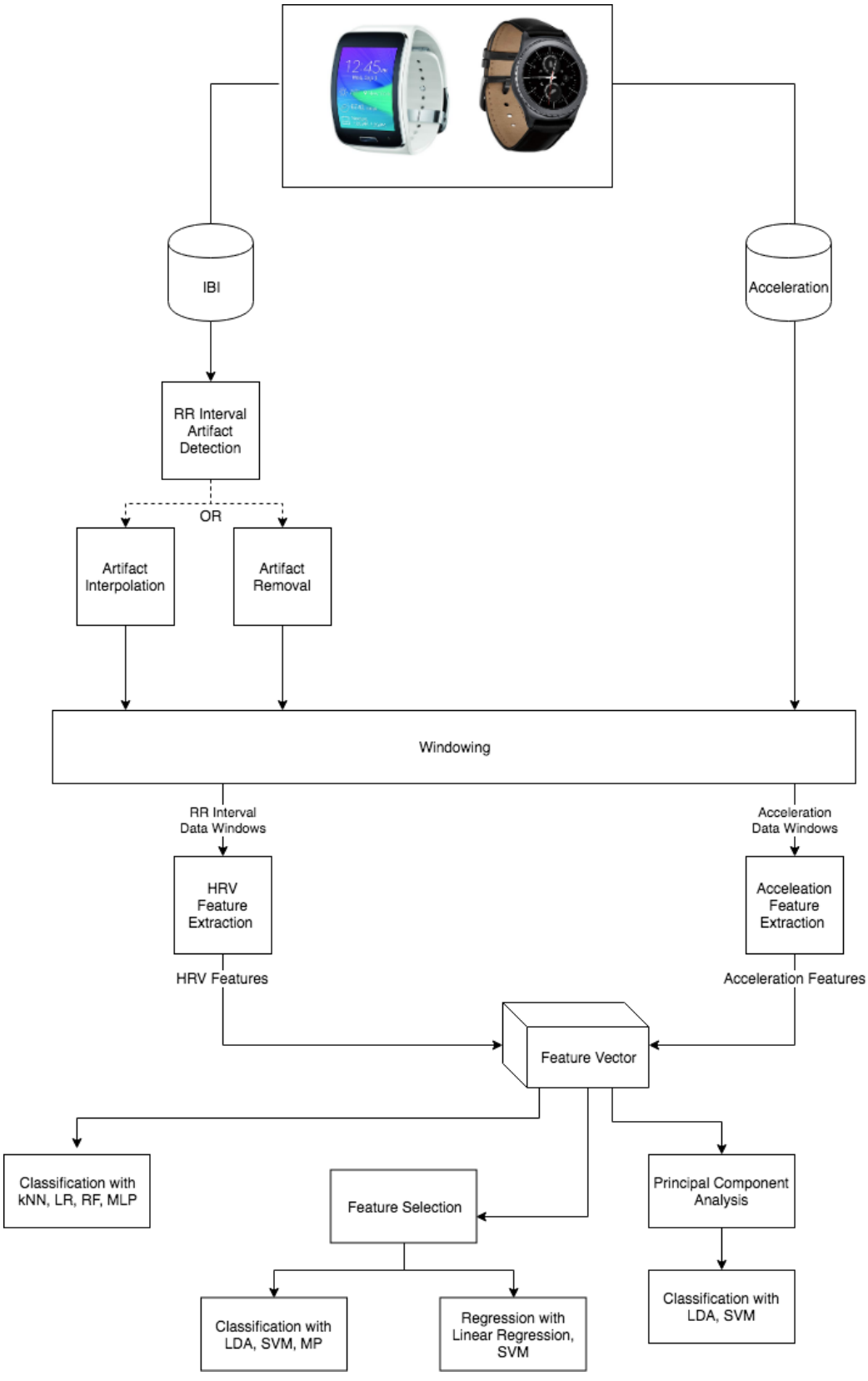

Figure 4.2: The system design figure of stress level detection system with Samsung Gear S and S2.

and 4.7. The most important features derived from the IBI signal are the HRV. The class labels are generated by using the NASA-TLX or the context information. Then, there are several possibilities, feature selection based on classifier subset evaluation or principal component analysis can be made. The system can also pass the feature vector without applying the feature selection to the next step. Finally, a machine learning algorithm is employed.

The diagram of the second variant of the proposed system is described in Figure 4.3. This system proposes a stress detection solution for the Empatica E4 smartband. This system requires IBI, acceleration, EDA and Skin Temperature. The system detects and interpolated the artifacts in the IBI. EDA artifacts are cleaned using a artifact detection model which requires the acceleration and the skin temperature signal in addition to the EDA signal. EDA, acceleration and IBI signals are aligned and divided into 2 minutes long sliding windows with 50% overlap. For each window, 29 features are calculated. Respectively, the feature extraction of EDA, IBI and acceleration signals is described in Subsection 4.5,4.6 and 4.7. The class labels are generated by using the NASA-TLX or the context information. Finally, a machine learning algorithm is employed with feature selection methods.

4.1. Samsung Gear Series

Samsung is one of the biggest commercial smartwatch producer in the market [7]. They have manufactured Samsung Gear S, Samsung Gear S2 and Samsung Gear S3 smartwatches in 2014, 2015 and 2016. These devices are equipped with the ambient light sensor, PPG sensor for heart rate monitoring, 3D inertial measurement unit with 3D accelerometer and gyroscope and pedometer. Their prices are more affordable then research smart bands. Samsung Gear Series run on the Tizen platform [53].

A data collection application for the Tizen Framework is developed by our research group [54] [55]. This application can run on Samsung Gear S, Samsung Gear S2 and Samsung Gear S3. Tizen platform comes with a Tizen IDE for development. It has two types of applications which are native and web applications. In the native

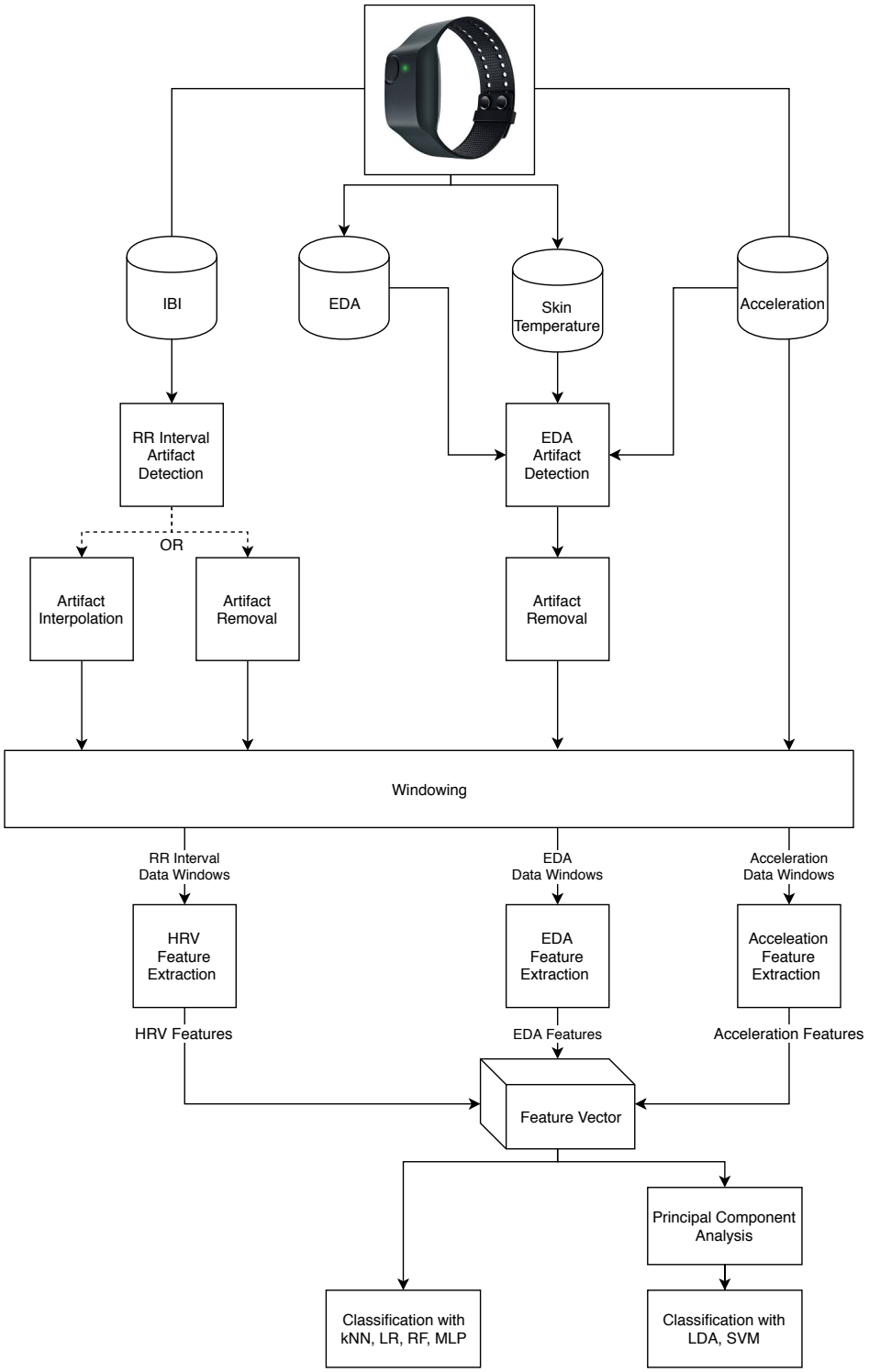

Figure 4.3: The system design figure of stress level detection system with Empatica E4 wristband.

application c++ is used to develop an application. In the web application, javascript is used to develop an application. Two types of applications are developed in order to compare the performance. In this study, a web application in the Tizen platform is used for the data collection due to its efficiency in adjusting the sampling rates of sensors and easy handling of sensors. Three-dimensional acceleration (20Hz), inter-beat information from heart rate monitoring unit, pedometer measurements is obtained thanks to our application.

The battery life of the Samsung Gear Series is not suitable for 24 hours of continuous recording. The application developed by our research group can run up to 2 hours on Samsung Gear S, 3 hours on Samsung Gear S2 and 4 hours on Samsung Gear S3. In order to record 24 hours of data, duty cycling can be applied.

4.2. Empatica E4 Smartband

Empatica E4 is a wristband manufactured by Empatica Inc. It is an unobtrusive and comfortable device. Empatica E4 is equipped with a PPG sensor for the heart rate measurement, a 3D accelerometer sensor, an EDA sensor for measurement of the response of the skin and temperature sensor for the skin. This device can record 48 hours in the offline mode with a fully charged battery. The flash memory for the offline mode is up to 60 hours of recording. It is equipped with a Bluetooth Low Energy Smart unit. The wristband can be paired with IOS or Android operating systems. There are available official applications to record the real-time data with the smartphone. Empatica also provides an API to develop custom applications on smartphones using Android or IOS. Once the device is paired with the smartphone, it enters the streaming mode. In the streaming mode, the data is recorded in real-time and can be sent to the Empatica cloud. The wristband can run up to 24 hours in the recording mode. Empatica E4 can be charged by connecting to any USB port. It takes about 45 minutes to be fully charged.

Figure 4.4: The Empatica E4 smartband. Original figure from [6]

4.3. Ethical issues & Informed consent forms

The procedure of the experiments used in this thesis is approved by the Ethical Committee of Boğaziçi University (INAREK). Prior to the data acquisition, each participant received a consent form which explains the experiment procedure and its benefits to both the society and the subject. The data collection procedure and all of the interventions in this research fully meet the 1964 Declaration of Helsinki [56]. Data is stored anonymously.

4.4. Problems Related to the Movement and Improper Placement of Devices

Nowadays, off-the-shelf wearable devices provide us with high-quality data standards according to the definition of high-quality by [57]. However, certain conditions must be satisfied for high-quality data acquisition. Electrodes should be properly placed obeying the instructions of the device, devices must be tightly worn, body movements should be limited. Otherwise, signals are contaminated by noise, loosely worn devices, and body movements [58]. These are the problems researchers will have to face when they take a step outside the lab. To remove this noise from the signal,

some signal processing techniques must be applied. Every problem creates multiple options for researchers. To give an illustration, if a subject wears the device loosely, and for some period the data could not be acquired, the researcher may opt to ignore this time period or interpolate the data with some compatible function compatible with the data. Another example would be the choice of handling data artifacts due to unconstrained movement of a subject in daily life data recording. Most of the state-of-the art tools for feature extraction of physiological signals come with artifact handling methods.

4.5. Electrodermal Activity Signal Preprocessing and Feature Extraction Tools

Electrodermal Activity signal is affected from increased physical activity and temperature changes. In these situations, obtained signal is contaminated and should be filtered. To this end, we employed the EDA Explorer tool from Taylor *et al.* [59]. This tool has a built-in classifer model to detect artifacts. In order to build it, the EDA signals are manually labeled from the experts. By applying the SVM (Support Vector Machine) classifier with the accelerometer and temperature data, this tool achieves 95% accuracy on detecting artifacts in the EDA signals. After cleaning the artifacts from the signals, features were extracted. The EDA signal has two components phasic and tonic namely. Skin Conductance Level (Tonic) component includes more long term slow changes whereas phasic components include faster (event-related) changes in the signal. When evaluating the mean, standard deviation and percentile features, researchers use tonic component because they do not want to overestimate these long term changes with event-related fast changes. The phasic part is subtracted and features are calculated. On the other hand, some peak related features like peak per 100 seconds, peak amplitude, strong peak(more than) per 100 are calculated from phasic element.

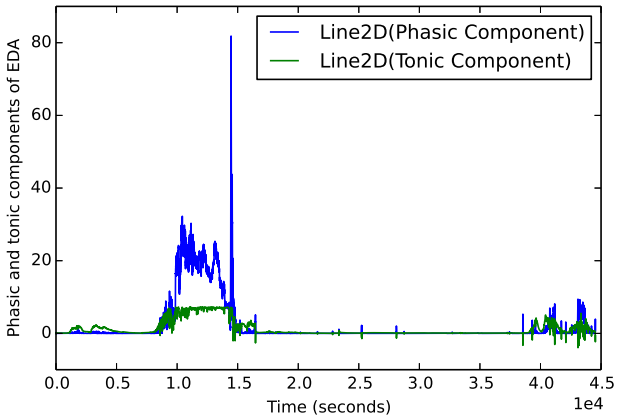

Figure 4.5: Decomposed EDA Signal from Empatica E4 wristband by applying cvxEDA tool.

We decomposed the EDA signal by applying the *cvxEDA* tool [60] on the EDA signal which makes use of a convex optimization approach to decompose EDA signal. After that, we extracted seven features from the EDA signal: mean, standard deviation, peak, strong peak, 20th percentile, 80th percentile and quartile deviation (75th percentile – 25 percentile).

4.6. Heart Activity Signal Preprocessing and Feature Extraction Tools

The heart rate activity signal is also sensitive to the movement of the subjects and loosely worn wrist devices. In order to cope with these problems and clean the artifacts from the signal, our research group developed a preprocessing tool in MATLAB. With this tool, we employed an artifact detection percentage threshold between the data and local average and delete the data points that are not satisfying this rule. In the literature, this threshold is generally set as 20% [34]. The user can choose to remove these data points or replace them with shape preserving cubic spline interpolation by changing parameters. If the users choose to remove these data points, they can set new rules on the remaining data that they need the amount of consecutive data points (non-interrupted with deleted artifacts) to evaluate the segment meaningful. They can set some minimum consecutive time rules similarly for these parameters.

Table 4.1: Heart rate variability features and their definitions.

Feature	Description
Mean RR	Mean value of the inter-beat (RR) intervals
STD RR	standard deviation of the inter-beat interval
RMSSD	Root mean square of successive difference of the RR intervals
pNN50	Percentage of the number of successive RR intervals varying more than 50ms from the previous interval
HRV triangular index	Total number of RR intervals divided by the height of the histogram of all RR intervals measured on a scale with bins of 1/128 s
TINN	Triangular interpolation of RR interval histogram
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

The tool has also a batch processing feature. Length of local mean, percentage of artifact detection threshold, minimum consecutive time and data sample constraints can be obtained from parameters. For feature extraction, we used MATLAB built-in tools along with Marcus Vollmer HRV toolbox [61] along with our preprocessing tool.

The employed time domain features are the mean value of heart rate (Mean HR), the standard deviation of 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 50ms 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).

We also applied Fast Fourier Transform (FFT) and Lomb-Scargle periodogram and the following frequency domain features are calculated: 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), (From Lomb-Scargle) LF, HF, LF/HF. There are many heart rate variability toolboxes available, one of them is Kubios HRV [62].

4.7. Accelerometer Processing and Feature Extraction

The Accelerometer sensor in Samsung Gear models and Empatica E4 records 3-axis acceleration with gravity. 4th axis called acceleration magnitude is calculated by taking the square root of squares of 3-axis. We employed the accelerometer modality in two ways. Firstly, to detect artifacts in the EDA data, the accelerometer data is used along with the skin temperature data. Secondly, we extracted statistical features from this sensor such as the mean value of each axis. Lastly, FFT is applied to the acceleration magnitude and energy is calculated.

4.7.1. Machine Learning Tools

For the classification and regression of the data, we employed the Weka toolkit [63]. The last column of the feature vector is defined as categorical class for classification and numeric class for regression. Since, our data set is unbalanced in terms of membership of class instances, we applied different techniques in order to overcome class imbalance problem. Therefore, we prevented classifiers from biasing towards classes with more instances.

The system uses several classification and regression algorithms. The classification algorithms are presented below.

- Linear Discriminant Analysis (LDA)
- Support Vector Machine with radial based function kernel (SVM)
- K-Nearest Neighbours (n=1) (kNN)
- Logistic Regression (LR)
- Random Forest with 100 tress (RF)
- Multilayer Perceptron (MLP)

The regression algorithms are presented below.

- Linear Regression
- Support Vector Machine (SVM) for regression

Feature selection techniques are also available in Weka toolkit. We used principal component analysis (PCA) and classifier based feature selection [64].

5. CASE STUDY: WORKLOAD AND STRESS DETECTION IN THE WILD

In the literature, researchers trained the model in the laboratory and later applied to the real-life settings. In this study, we train the model in real life settings with physiological data and the stress levels generated from questionnaires and predict the level of stress in another real life session. We collected the data by not providing an interrupt to their daily life and not changing the environment of the individuals, in order to discover the performance of the proposed system without any restrictions. We did not train the model in a so-called calibration session, but the model is trained with an hour session data from another day of an individual. Because we want our system to be easily trainable by mass populations. Gathering the real context information with a lot of questionnaires can create an interrupt or with video recordings and labeling creates privacy concerns. All the context information that we have for validation is dependant on 21 questions of NASA-TLX. This is the one of the first studies, where the data quality of a commercially available smartwatch is discovered in real-life settings.

5.1. Description of the Case Study

17 participants joined this study. Answers of 7119 questions from NASA-TLX questionnaire and 339 hours of physiological signals is collected from 17 participants. However, five of them dropped the study due to misuse of the application. The data is collected over a month for each participant. Participants were asked to wear a Samsung Gear S2 smartwatch during their everyday life. We asked to open the application that we developed whenever they want in a day. Since the battery life of the smartwatch is limited, we wanted them to start our application once a day for an hour. 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 subjective score of mental workload is determined with the NASA-TLX questionnaire. The experimental procedure is described in Figure 5.1. Participants are allowed to

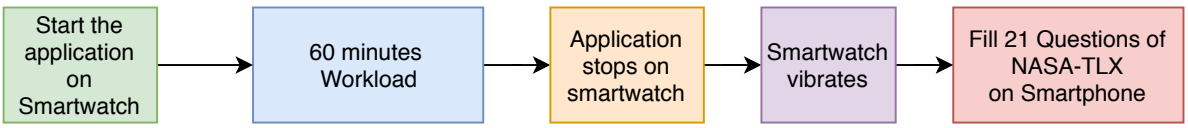

Figure 5.1: An experimental procedure for daily workload sessions.

start the application more than one in a day. We wanted them to complete at least 20 sessions. Since our study is totally voluntary based and participants did not receive any form of payment, mostly they could finish 20 sessions in a month. We discover the possibility to automatically detect stress by training the system with subjectively reported questionnaires. Thanks to the nature of this system, it can be scalable easily because the user only needs to use the application to discover his/her stress level and if they want to they can also donate their data with their NASA-TLX form. After dropping the misused participants, the answers of 5544 questions from NASA-TLX questionnaire and 264 hours of physiological data from 12 participants is left.

5.2. Methodology

5.2.1. Features of Collected Data of Samsung Gear S2

The features of Samsung Gear S2 are presented in Section 2.4. We used the accelerometer and heart rate variability features. In addition to these features, we presented a new feature called the ratio of the interpolated RR intervals. The window size is selected as 2 minutes and overlap is defined as 50%. The mean value of each feature in resulting segments is calculated and stored as session. For example, from 60 minutes of data, we extract 59 windows. For each window, 22 physiological features are computed and added as a row in a feature vector. Then, the mean value of each column is calculated and class label attached as the 23rd column.

5.2.2. Self Report Questionnaire as Ground Truth

NASA-TLX questionnaire consists of 21 questions, we used the full version of the NASA-TLX. First 6 questions ask the level of mental demand, physical demand, effort, performance and frustration in a scale of 0-100, following 15 questions compare every item with each other. The scores are always between 0 and 100. A score can be computed from each question and their combinations. In this study, we measured the performance of the three scores. Because in such an environment gathering the ground truth is still a challenge.

- Weighted score computed from 21 questions.
- Frustration score coming from one question.
- Mental demand score coming from one question.

Weighted scores are calculated as follows: Weights coming from the last 15 questions are multiplied with the 6 scores coming from the first 6 questions. Frustration scores are calculated by using the rating coming from the question about the frustration level. Mental demand scores are calculated by using the rating coming from the question about mental demand level.

5.2.3. Generation of Class Labels

The class labels are created using the NASA-TLX scores. The following rule is applied for the class labels.

- NASA-TLX scores ≤ 33 as *low*
- $66 \geq$ NASA-TLX scores > 33 as *medium*
- $100 \geq$ NASA-TLX scores > 66 as *high*

Table 5.1: The distribution of class labels for each label generation technique

Label Generation Technique	Low	Medium	High
Weighted TLX	67	146	51
Frustration	113	76	75
Mental Demand	82	104	78

After removal of sessions where participants forgot to fill the questionnaire or misused the application, 264 sessions of data with 21 features are left. Since we gathered data in real life settings, our dataset introduces a class imbalance. This problem is handled differently in classification and regression. The distribution of the class labels for each label generation technique is presented in Table 5.1.

5.2.4. The Quality of Heart Rate Data

Samsung Gear S2 is equipped with a PPG sensor. The data is gathered with Tizen Web application developed by our research group. The physiological data contains artifacts this might be due to the power consumption policy of the Tizen platform, the interrupt of the sensor or the contact problem with the skin. Therefore, these artifacts should be handled. We removed and interpolated every RR interval which exceeds more than 20% of its predecessor with a cubic spline interpolation. The amount of interpolation is an important aspect of our analysis because these removals and interpolations affect the heart rate variability measures. We created another feature using the ratio of non-interpolated RR intervals in percentage for each session of recordings. We name this feature as the quality of RR intervals. In Figures 5.2 and 5.3, we plotted the histogram and boxplot of the data quality. We do not see any sessions where the data quality is less than 30 percent. However, these results show that collecting very high-quality interbeat-interval data with Samsung Gear S2 is not possible.

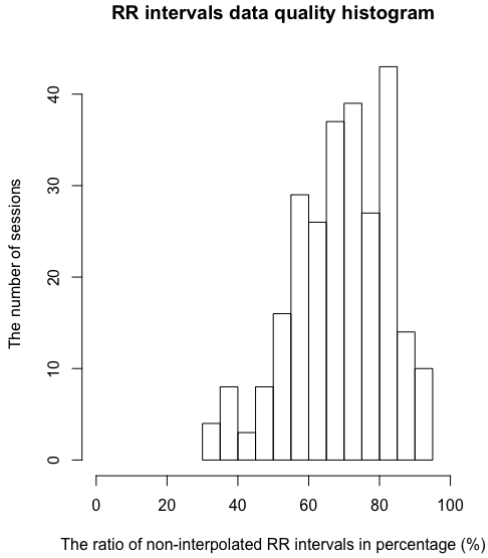

Figure 5.2: The histogram of RR intervals quality in terms of the ratio of non-interpolated RR intervals in percentage (%). Each quality measure is calculated for each session.

Our goal is to discriminate the different levels of stress, therefore we report how well and in which precision can we achieve this goal using a commercially available unobtrusive device for end-users without a need to restrict their daily activities. A concrete example of these restrictions can be only using the application when their hands do not move or do not let them walk. This restriction will limit the usage of the application. Every results that we present in this thesis covers the usage of Samsung Gear S2 without any restrictions.

5.3. Results

5.3.1. Classification Results

The classification accuracy is measured by using the generated class labels coming from the NASA-TLX. In order to overcome the class-imbalance problem, we resampled the minority class instances of each training set in every iteration 10 fold cross validation, therefore they became uniform. We did not apply resampling to the test set, so our dataset is not biased. The feature selection is applied based on the subset of a

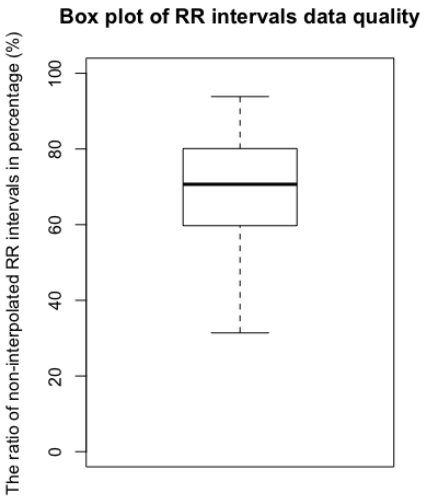

Figure 5.3: The boxplot of RR intervals quality in terms of the ratio of non-interpolated RR intervals in percentage (%). Each quality measure is calculated for each session.

Table 5.2: The three class classification results of labels generated from weighted score with LDA, SVM and MLP classifiers

Algorithm	**Accuracy (%)**	**Precision**	**Recall**	**F-Measure**
LDA	45.83%	0.521	0.458	0.464
SVM	42.04%	0.510	0.420	0.429
MLP	39.77%	0.436	0.398	0.407

classifier proposed and implemented in Weka by [65]. The classifier is selected as LDA and the measure to optimize is selected as the classification accuracy. We iterated over every classifiers in feature selection and we find the best classifier for feature selection as LDA.

The performance of the following three classifiers are evaluated.

- Linear Discriminant Analysis (LDA)
- Support Vector Machine (SVM) with linear kernel
- Multilayer Perceptron (MLP)

Table 5.3: The three class classification results of labels generated from frustration score with LDA, SVM and MLP classifiers

Algorithm	**Accuracy (%)**	**Precision**	**Recall**	**F-Measure**
LDA	42.8%	0.451	0.428	0.432
SVM	42.04%	0.395	0.420	0.392
MLP	42.8%	0.453	0.428	0.435

Table 5.4: The three class classification results of labels generated from mental demand score with LDA, SVM and MLP classifiers

Algorithm	**Accuracy (%)**	**Precision**	**Recall**	**F-Measure**
LDA	40.53%	0.410	0.405	0.407
SVM	45.45%	0.455	0.455	0.453
MLP	43.56%	0.439	0.436	0.437

Table 5.5: The two class classification results of labels generated from weighted score with LDA, SVM and MLP classifiers

Algorithm	**Accuracy (%)**	**Precision**	**Recall**	**F-Measure**
LDA	68.64	0.689	0.686	0.687
SVM	70.33	0.702	0.703	0.702
MLP	63.55	0.638	0.636	0.637

Table 5.6: The two class classification results of labels generated from frustration score with LDA, SVM and MLP classifiers

Algorithm	**Accuracy (%)**	**Precision**	**Recall**	**F-Measure**
LDA	63.82%	0.650	0.638	0.642
SVM	68.61%	0.685	0.686	0.686
MLP	64.89%	0.652	0.649	0.650

Table 5.7: The two class classification results of mental demand with LDA, SVM and MLP classifiers

Algorithm	**Accuracy (%)**	**Precision**	**Recall**	**F-Measure**
LDA	59.37%	0.594	0.594	0.594
SVM	52.51%	0.524	0.525	0.520
MLP	63.75%	0.637	0.638	0.637

Three class results are presented in Table 5.2, Table 5.3 and Table 5.4. For three classes the best result that we achieve is 45.83% weighted score with the LDA classifier. This accuracy is not feasible for real life applications.

Two class results are presented in Table 5.5, Table 5.6 and Table 5.7. For two class the best result that we achieve is 70.33% weighted score with the SVM classifier. This is an acceptable accuracy because, in [45] they achieved 70% accuracy with Empatica E4 which is far better device than Samsung Gear S2 in terms of battery life, the signal quality and variety. Even though, we received many complaints about the questions of the NASA-TLX, weighted scores coming from 21 questions are important for classification.

5.3.2. Regression Results

In Section 5.3.1, we measured the performance of classification algorithms. However, in order to classify, we created class labels with thresholds. The selection of thresholds can change the accuracy of the measurement. Appropriate thresholds are still unknown for this context. Therefore, we apply numeric regression in order to overcome this a-priori selection.

The best scores for classification came from the weighted NASA-TLX in Section 5.3.1. The regression performance is measured by using the weighted scores come from 21 questions of the NASA-TLX for each instance. We tried to predict different ratings coming from the weighted NASA-TLX with regression.

Table 5.8: Regression results on balanced dataset with heart rate variability and accelerometer features.

Algorithm	**MAE**	**RMSE**	**Corr. Coeff.**
Linear Regression	19.88	23.95	0.230
SVM for regression	19.43	23.75	0.280

In order to overcome the imbalanced dataset problem, the NASA-TLX scores are discretized using equal-width discretization to establish pseudo classes for weighting. Then, we reweighted the instances in the data so that each class has the same total weight. We selected width as 10. Leave-one session out cross validation is applied. The performance of the following regression models are evaluated.

- Linear Regression
- Support Vector Machine (SVM) for regression

In our system, SVM for regression uses a normalized polynomial kernel. Because linear and radial basis kernels are not implemented in Weka for SVM for regression The dataset is normalized before the application of SVM.

Correlation coefficient (Corr. Coeff.), mean absolute error (MAE) and root mean squared error (RMSE) are reported in Table 5.8, Table 5.9, Table 5.10 and Table 5.11. After getting the first results, we applied feature selection by evaluating the subsets of features in terms of MAE and the feature set is decreased to eight features. The results are shown in Table 5.9. Attribute selection outperforms the results coming from without attribute selection. The SVM for regression algorithm gives better results than the linear regression. The best result we achieve is RMSE of 23.13 and MAE of 18.80 with SVM for regression with feature selection presented in Table 5.9.

We also discarded accelerometer features in order to see the effect on accuracy. These results are presented in Table 5.10 and Table 5.11. Feature selection increases the performance of the regression algorithm.

Table 5.9: Regression results on balanced dataset with heart rate variability and accelerometer features, feature selection is applied.

Algorithm	**MAE**	**RMSE**	**Corr. Coeff.**
Linear Regression	19.75	23.72	0.245
SVM for regression	18.80	23.13	0.280

Table 5.10: Regression results on balanced dataset with heart rate variability features

Algorithm	**MAE**	**RMSE**	**Corr. Coeff.**
Linear Regression	20.43	24.38	0.1755
SVM for regression	20.16	24.55	0.1645

Table 5.11: Regression results on balanced dataset with heart rate variability features, feature selection is applied

Algorithm	**MAE**	**RMSE**	**Corr. Coeff.**
Linear Regression	20.02	23.96	0.2058
SVM for regression	20.1903	24.53	0.1743

5.4. Discussion

For two classes classifier, the best result that we achieve is 70.33% weighted score coming from 21 questions of NASA-TLX questionnaire with SVM classifier with a linear kernel. For three classes classifier, we could not achieve more than 45% accuracy. The performance of the classifier is not acceptable for real-life applications. This might be due to subjective labels and low quality of data of Samsung Gear S2.

Regression results provide an interpolation of the data because introducing the class labels relying on thresholds may be biased and better performance might be achieved. The best result we achieve is RMSE of 23.13 and MAE 18.80 of with SVM for regression.

All the results are coming from the person independent model. Since we do not have enough data for each person to create such a personalized model in this study.

In the Chapter 6, we reapplied the same procedure with Empatica E4 smartbands in order to analyze if better data quality improves the performance of the proposed system described in Chapter 4.

We also received many complaints about the 15 pairwise comparison questions of the full version of the NASA-TLX questionnaire. The participants found them repetitious and hard to answer. However, the best results we achieved was with the full version of the NASA-TLX questionnaire. If one cannot recruit participants due to this problem, he/she can only ask the question on frustration rating since it is scale that we the best results excluding the weighted scale.

6. CONTINUOUS STRESS DETECTION USING WEARABLE SENSORS IN REAL LIFE SCENARIOS: ALGORITHMIC SUMMER CAMP CASE STUDY

In this chapter, we conducted a context-driven stress analysis in AN algorithmic competition summer camp which is organized each year in Istanbul, Turkey [66]. Since, in Chapter 5 we could not achieve meaningful results with the information only depending on 21 questions of NASA-TLX questionnaire and the only device we used was Samsung Gear S2, we bought four Empatica E4 smartwatches in addition to Samsung Gear smartwatches and we provided the context information by monitoring each individual during the recordings. In the literature, experiments are conducted separetly for each individual which is highly costly comparing to our methodology. In this work, we present a procedure to conduct analysis on simultaneously. The data is gathered from 21 participants simultaneously in 10 days.

6.1. Real-life group experiment setup

The algorithmic contest is conducted in three levels, expert, advanced and foundation. 84 students with different levels of expertise gathered to participate in this algorithmic contest. Almost all of the attendees come from high-ranking Turkish universities. These students enter the contest after passing an entrance exam in which their programming and coding levels are assessed. They would be permitted to participate in the event only if they could get a passing grade from the entrance exam. Algorithmic contest camp was held for 10 days. The data was collected from the 21 participants in the Foundation level.

Training is the major part of the event which has not been neglected during the competition so that each day, there were training classes before the contests. The program for the whole 9 days was scheduled to be held from 10:00 to 17:00. First, attendees had training classes with professors from the field of Computer Science from

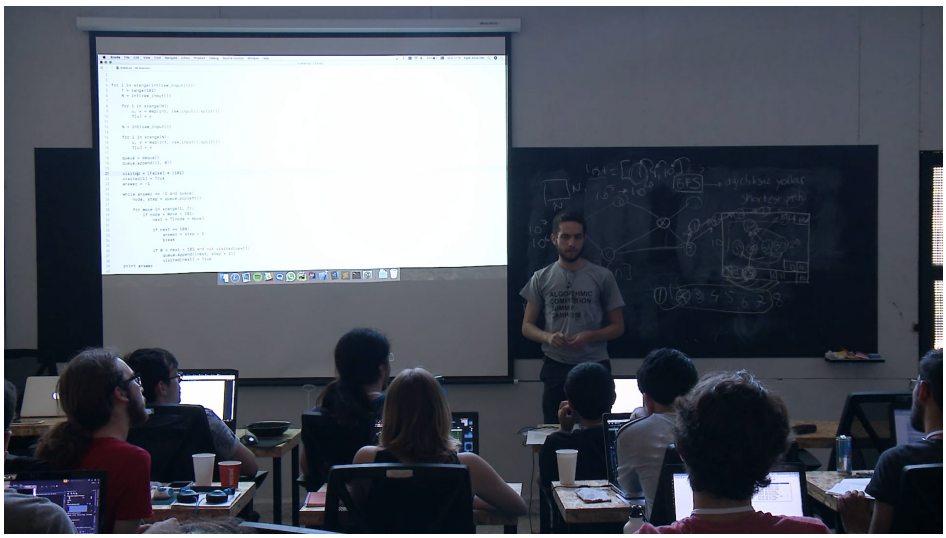

Figure 6.1: Daily training classes before the contest.

high-ranking universities in Turkey for two hours. The participants entered daily problem-solving contests which were in regards to the training lectures afterwards. An example of the lecture is shown in Figure 6.1. The same schedule was repeated for the whole 8 days. However, on the last day of the contest, the program was entirely different. On the closing day, all of the participants entered the final contest and they should solve challenging questions which were asked from all of the topics covered in the total 8 days of training. In order to collect more points, participants should solve more questions in the shorter time period than their opponents. As a result, mental demand increased as well as the temporal demand which encouraged the participants to gain more points in a shorter time and to achieve a higher position in the final ranking. We demonstrate the prize ceremony of the event in Figure 6.2, in order to describe the event.

In the field, our team consists of three testers. One of them provided the questionnaires after each session. Since the break between sessions is only 10 minutes. Providing the questionnaires and gathering the answers from 21 participants were challenging. The other scientist was responsible for the devices and he was making sure that the devices' batteries are full and everybody wears the device correctly since improper usage and wearing results a decrease in data quality. The organization provided us a desk to put our devices as shown in Figure 6.3.

Figure 6.2: Inzva Final contests prize ceremony.

6.2. Data Collection

Study subjects were selected from a similar expertise level considering they would be attending the entire week with approximately the same conditions. 21 volunteers from the foundation level have been selected to take part in our study. Tutorials and further guidelines were presented to all of them concerning how to use the devices and how to fill in the questionnaires. For data extraction, collecting the forms and battery recharge procedures which were being administered by our team, a schedule was set up and participants followed this schedule regularly.

A unique number has been assigned to each participant and to each watch, for the whole week participants used the watch which has the same number as their own participant number however for privacy reasons we did not collect the participant names and all of the information is kept anonymous.

6.3. Preprocessing

In this study, we used the EDA, heart rate and accelerometer toolboxes that are presented in Section 4.5, Section 4.6 and Section 4.7. In this work, different window sizes and artifact correction thresholds are selected for evaluation. These evaluations are described in the next subsections of this work.

Figure 6.3: A view of smart watches and wristbands before data collection, charged and ready to use.

6.4. Machine Learning Classifiers

In this study, we evaluated the performance of the six well-known classifiers listed below.

- Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA)
- PCA and Support Vector Machine with radial basis function kernel (SVM)
- K-Nearest Neighbours (n=1) (kNN)
- Logistic Regression
- Random Forest with 100 tress
- Multilayer Perceptron

We applied 10-fold cross validation. We report the results of the three class problem where the classes are free day, lecture and contest, therefore we expect to detect low, medium and high stress levels.

6.5. Results

We developed a 3-class stress detection system. The system can differentiate the stress level of free day, lecture and contest sessions. It can further differentiate three levels of perceived stress (See Section 5.6). The detailed accuracy results, f-measure, precision and recall values are presented in Tables 6.1, 6.2 and 6.3. Table 6.1 presents the classification accuracy results obtained from HR, EDA and ACC signals. The apparatus was the Empatica E4 device in this table. 32 hours of data was collected from four participants. The Multilayer Perceptron algorithm achieved the best classification accuracy which is 92.15%.

In Table 6.2, we leave the EDA signal collected from Empatica E4 out. Since Samsung Galaxy Gear devices do not have EDA sensors. In Table 6.3, we demonstrated the results from the data collected from 18 participants for 32 hours (9 days) by using both 4 Empatica E4 and 14 Samsung Gear devices (All Devices). Note that we also collected data with Samsung Gear S3 classic smartwatches from three participants. However, we do not use this data since they do not provide the raw RR interval data. The Multilayer Perceptron algorithm achieved the best result (92.19%) from HR and ACC signals collected using Empatica E4 whereas the Random Forest algorithm gives the best classification accuracy (88.26%) with the HR and ACC data collected from all devices.

6.6. Device Type Comparison

We compared Empatica E4 and Samsung Gear S2 devices. Samsung Gear S2 devices are commercial type, relatively cheaper smart watches. On the other hand, Empatica E4 is a more precise, relatively more expensive research device. We compared the classification accuracies and data quality on both of these devices. In the literature, RR intervals which differ more than 20% from the local average are removed [67]. This is called the artifact detection with percentage threshold. We change the value of this threshold from 10% to 25% and observe the remaining clean data in terms of percentage.

Table 6.1: Stress Level Classification Accuracy, f-Measure, precision and recall values with Different ML Algorithms - 3 class - HR + EDA +ACC - Empatica E4.

Algorithm	**HR + EDA + ACC (Empatica E4)**			
	Accuracy	**f-Measure**	**Precision**	**Recall**
PCA + LDA	82.35	82.2	82.6	82.4
PCA + SVM (radial)	82.35	82.5	83.3	82.4
kNN	80.39	80.4	80.8	80.4
Logistic Regression	90.19	90.1	90.2	90.2
Random Forest	86.27	86.2	86.2	86.3
Multilayer Perceptron	92.15*	92.20*	92.30*	92.2*

Table 6.2: Stress Level Classification Accuracy, f-Measure, precision and recall values with Different ML Algorithms - 3 class - HR +ACC - Empatica E4.

Algorithm	**HR + ACC (Empatica E4)**			
	Accuracy	**f-Measure**	**Precision**	**Recall**
PCA + LDA	72.54	71.6	71.8	72.5
PCA + SVM (radial)	86.27	86.2	86.9	86.3
kNN	84.31	84.1	84.6	84.3
Logistic Regression	86.27	86.2	86.9	86.3
Random Forest	88.25	88.0	88.1	88.2
Multilayer Perceptron	92.19*	90.3*	91.4*	90.2*

As seen in Figure 6, Empatica E4 devices have approximately 25% more remaining data for all of the different artifact detection percentage thresholds. We deduce that the quality of RR intervals of Empatica E4 devices is higher than Samsung Gear S2 devices. We further investigate the effect of data quality on stress level classification accuracies. We demonstrated that classification accuracies obtained from the data collected with Empatica E4 is higher than Samsung devices with all classifiers in Table 6.4 and Table 6.5. From these results, we can observe that the data quality has an effect on the stress level classification accuracies.

Table 6.3: Stress Level Classification Accuracy, f-Measure, precision and recall values with Different ML Algorithms - 3 class - HR +ACC - All.

Algorithm	**HR + ACC (All Devices)**			
	Accuracy	**f-Measure**	**Precision**	**Recall**
PCA + LDA	59.12	59.8	60.1	59.6
PCA + SVM (radial)	76.99	77.1	77.3	77.0
kNN	87.32	87.2	87.3	87.3
Logistic Regression	65.25	65.0	65.0	65.3
Random Forest	88.26*	88.20*	88.2*	88.30*
Multilayer Perceptron	83.09	83.0	83.2	83.1

Table 6.4: Effect of the used device to 3-class stress level classification accuracy when heart activity and accelerometer data are used together (with context HR + ACC).

Algorithm	**Empatica E4**	**Samsung Gear S2**	**All Devices**
PCA + LDA	88.88	72.60	59.12
PCA + SVM (rad)	92.06	78.60	76.91
kNN	87.30	85.30	87.3
Logistic Regression	90.47	83.30	65.25
Random Forest	90.40	88.60*	88.30*
Multilayer Perception	95.23*	87.30	83.10

Table 6.5: Effect of the used device to 3-class stress level classification accuracy when only heart activity signal is used (without context only HR data).

Algorithm	**Empatica E4**	**Samsung Gear S2**	**All Devices**
PCA + LDA	65.07	55.33	52.58
PCA + SVM (rad)	90.40*	73.33	62.60
kNN	88.88	82.00	82.15
Logistic Regression	84.90	66.66	66.66
Random Forest	87.30	84.67*	82.62*
Multilayer Perception	88.88	78.00	71.36

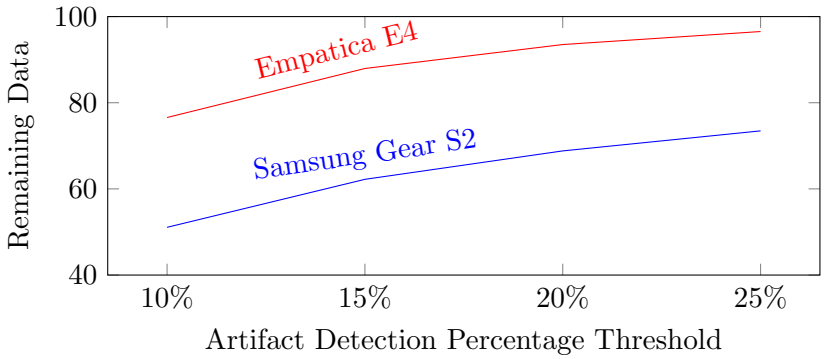

Figure 6.4: Percentage of the remaining data (for both device types) after the artifacts are removed versus the different percentage thresholds of artifact detection.

6.7. Effect of Artifact Detection Percentage Threshold, Interpolation and Accumulation Window on Accuracy

Physiological signals are sensitive to the movements of the subjects. Especially the quality of the heart rate data declines very drastically in the case of intense physical activities. We applied a few preprocessing techniques and filters to remove the contamination of the heart rate data. In this subsection, we investigated the effects of artifact detection percentage threshold, interpolation and the aggregation window length. Artifact detection percentage threshold is the minimum percentage difference between a data point and the local average in order to evaluate the data point as an artifact. If the value of the artifact correction percentage threshold increases, the filter loosen i.e. the number of detected artifacts will be decreased. Furthermore, aggregation window is the data segment in which features are extracted and averaged for the whole session to get the features of the session.

We applied artifact correction percentage thresholds from 10% to 25% and investigate the stress level classification accuracies in Table 6.7. We were unable to observe a pattern when we applied different classifiers and changed the artifact detection percentage thresholds. We can infer that changing this threshold between 10% and 25% does not have a clear effect on the classification accuracy. Because even the 25% rule can detect most of the artifacts.

Table 6.6: Effect of the Length of the Aggregation Window on Classification Accuracies.

Algorithm	**Aggregation Window Size (Seconds)**			
	120	**300**	**600**	**1200**
PCA + LDA	59.62	62.24	54.14	63.02*
PCA + SVM (radial)	76.32	77.94	77.27	88.33*
kNN	87.32	88.30	88.38*	85.41
Logistic Regression	65.25	69.90	72.22	76.16*
Random Forest	88.26*	86.76	87.87	84.14
Multilayer Perceptron	83.09	86.76	81.81	88.54*

We further examined the effect of the aggregation window on the stress level classification accuracy. We change the length of the aggregation window from 2 minutes to 20 minutes. We observed that the behavior changes between different ML algorithms. Researchers should take the ML algorithm and its performance of different aggregation window sizes when deciding the optimum window length. Gjoreski *et al.* found that the aggregation window lengths between 10 minutes and 17.5 minutes have better accuracy performance in general [45] which is similar to our results.

As we mentioned previously, we provide a parameter in our heart rate preprocessing tool that decides whether to interpolate the removed artifacts or remove and apply some minimum consequtive rules. The minimum consequtive rules could be either the minimum required number of samples or the time interval for a segment to extract features. We further investigate the effect of removal and interpolation to the classification accuracies. In Table 6.8, we can see that applying interpolation achieves higher results than filtering for some machine learning methods (removal and minimum consequtive filter) and lower results for other algorithms. This decision depends on the applied ML method.

Table 6.7: Classification accuracies vs. changing percentage based artifact detection and filtering rules.

Algorithm	**10 percent**	**15 percent**	**20 percent**	**25 percent**
PCA + LDA	64.28*	62.38	59.62	63.80
PCA + SVM(rad)	80.95*	78.57	77.0	79.52
kNN	87.61*	86.66	87.32	85.2
Logistic Regression	73.80*	61.90	66.25	66.19
Random Forest	89.0	88.09	89.26*	82.6
Multilayer Perception	80.0	78.57	83.09*	80.95

The user is provided an option to remove the artifacts or replace detected artifacts with cubic spline interpolation. The effect of this decision is examined in this subsection. The effect of aggregation window on the classification accuracies is also investigated.

Table 6.8: Classification accuracies when removed artifacts are replaced with interpolation vs. when they are removed.

Algorithm	**Filtering**	**Interpolation**
PCA + LDA	72.72*	50.75
PCA + SVM	89.39*	89.39*
kNN	95.45	97.72*
Logistic Regression	83.33	89.39*
Random Forest	95.45*	93.93
Multilayer Perception	89.39	95.45*

6.8. Person independent and dependent models

We developed two different stress detection systems. The first one is the general (person independent) model. In this model, all the data is divided into training and test segments. For both system we employed 10-fold cross-validation, the accuracy of the system is determined independently from any individual's data. The second model is the person dependent model. In this model, the classifier is trained and validated with the data coming from the same participant and different sessions.

We presented the accuracy results in Table 6.9. We can observe that person specific stress detection models have higher classification accuracies than general models as expected. Furthermore, we achieve the highest classification accuracy on person specific models with Empatica E4 devices when the Random Forest algorithm is applied (97.92%) to features from all signals. With all algorithms, HR, EDA and ACC signal combination with Empatica E4 devices have higher accuracies than with all devices in person specific models. These results demonstrate that stress level detection schemes

should give more weight to the individuals data than data from other people while building models.

Table 6.9: Classification accuracies of General and Person Specific Models.

Algorithm	General		Person Specific	
	HR+EDA+ACC-E4	**HR+ACC-All**	**HR+EDA+ACC-E4**	**HR+ACC-All**
PCA + LDA	82.35	59.12	95.83	87.6
PCA + SVM (radial)	82.35	76.99	93.75	85.98
kNN	80.39	87.3	95.83	89.91
Logistic Regression	90.19	65.25	95.83	90.17
Random Forest	86.27	88.2*	97.92*	90.17
MLP	92.15*	83.2	95.83	91.54*

6.9. Effect of Different Physiological Modalities

Multi-modality of any stress detection scheme is proven to improve the accuracy and the performance of the systems. Having said that, the effect of each modality and the combinations of them are different when the performance is taken into consideration. We examined the effect of each modality. Heart Activity alone, heart activity - accelerometer combination, heart activity - electrodermal activity combination and all of the modalities together were investigated and presented in Table 6.10. We achieve the highest stress level accuracies when MLP is applied on the features from all the modalities.

We further investigated the stress level detection schemes using different modalities into two groups: stress only and stress activity. With the addition of the data from accelerometer sensor, information about the activity and context of individuals are also evaluated. This data increases the accuracies of other physiological signals when combined with them in all cases (See Table 6.10). Furthermore, when we combined HR and EDA signals, the accuracy is higher than both signals alone in almost all cases (in RF it is equal to HR) We can infer from that using multiple modalities increases the performance of the stress level detection schemes.

Algorithm	Stress Only			Stress and Activity		
	HR	**EDA**	**HR + EDA**	**HR + EDA + ACC**	**HR + ACC**	**EDA + ACC**
PCA + LDA	49.01	52.94	62.70	82.35	72.5	80.39
PCA + SVM (radial)	80.39	62.74	84.31	82.35	86.27	80.39
kNN	82.35	84.31*	86.27	80.39	84.31	80.39
Logistic Regression	84.21	60.78	92.15*	90.19	86.27	78.43
Random Forest	86.27*	80.39	86.27	86.27	90.19*	84.31*
Multilayer Perception	86.27*	68.62	90.19	92.15*	90.19*	82.35

Table 6.10: Stress Detection Accuracies with Different ML Algorithms - 3 class classification. On the left side, stress recognition results which are only using HR and EDA signals are presented. On the right side, context information with accelerometer data is also added.

6.10. TLX clustering vs. Known Context, Effect of Ground Truth on Accuracy

As seen in the literature, in the same context, the perceived stress and physiological stress of individuals can be different. We investigated two different ground truths in this subsection. The first one is the known context as the ground truth. In our case, the contest context is assumed to induce stress, lecture context gives some cognitive load and lower amount of stress and free times are assumed to be relaxed sessions. In this calculation, the ground truth is enumerated from the known context as free day:1, lecture:2 and contest:3. When we examine the physiological signals we can differentiate these three levels with high classification accuracies (see Table 6.11). Perceived stress of individuals was also measured. We asked participants the following question to learn their stress level:

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

The answers increase by 5 and changes from 0 to 100. We determined the stress level as 1 if the answer is between 0 and 30. Stress level is assigned to 2 if the answer is between 40-70. The highest stress level is assigned is the answer is equal or above 80 which is 3.

When we look at the classification accuracies of the perceived stress level, for all machine learning algorithms and signal combinations, they are lower than those correspond to th physiological stress. This is because perceived stress is subjective, depends on individuals. The survey answers are also prone to error and this might be another reason for the decrease in the stress level detection accuracies. The correlation between the known context and perceived stress labels is computed to be 0.356 which is a moderate relation. The relation between the perceived stress and physiological stress is not investigated thoroughly in the literature. Development of personal perceived stress level models and filtering out the outlier survey answers might increase the performance of the classifiers.

Table 6.11: Classification accuracies when different ground truths are used. On the left, known context information (Free:1, Lecture:2, Contest:3) is used as class labels. On the right, subjective ground truths are used as class labels.

Algorithm	**Accuracy wrt. known Context**		**Accuracy wrt. Subjective Ground Truth**	
	HR+ACC-All	**HR+EDA+ACC-E4**	**HR+ACC-All**	**HR+EDA+ACC-E4**
PCA + LDA	59.12	82.35	54.46	50.98
PCA + SVM (radial)	76.99	82.35	69.01	72.55
kNN	87.3	80.39	85.44	78.44*
Logistic Regression	65.25	90.19	57.27	78.43
Random Forest	88.2*	86.27	86.38*	76.47
MLP	83.2	92.15*	80.28	68.62

6.11. Discussion

In this research, we developed a stress detection scheme for real-life data. We applied and measured the performance of context-driven stress recognition in an algorithmic summer camp. From there, we obtained labelled sessions for 21 subjects. We first described the event. We mentioned the difficulties, which do not occur in laboratory environments, of data collection in real life. After describing our algorithm, we presented the results. From the discussions in Section 5, we can deduce that the data quality of the devices increases the performance of the stress level classifier. Another significant finding is that the combination of modalities increases the performance of our system. Finally, we observed that the subjectively reported stress level classification has lower classification accuracy than the corresponding accuracies for the physiological stress level. Personalized perceived stress level models from the ground truth surveys and outlier answer removals could increase the performance. However, creating personalized models requires a lot of training data since multimodality is very important for such complex phenomena.

7. CONCLUSION

In this thesis, we proposed an unobtrusive design that recognizes the stress in real life with physiological signals. With the help of this information, the system analyzes the stress level of the user and warns the user to consult a professional in order to control the stress level. Several advantages and disadvantages of smart wearable devices are discussed throughout this thesis. The most unobtrusive system is formed. The patients can make their selection with respect to their budget.

In this thesis, we are able to discriminate different stress levels with a smartwatch application. This thesis has some contributions to the literature by proposing first stress level recognition application for a commercial smartwatch (Samsung Gear Series) and on a high-end wristband (Empatica E4). We completed one of the biggest simultaneous case study of this domain in both controlled and uncontrolled real-life environment with a smartwatch and a smartband. The possible problems of such a system are discussed throughout the thesis. For example, misusage of application results as loss of data, improper attachment of the smartwatch cause low-quality data. Therefore, some features may not come to exist. These problems decrease the performance of the machine learning classifiers.

The results show that an improvement in terms of sensor quality by extending the duration of the battery is required for smartwatches in the commercial consumer electronics market. We believe that in the future the sensor quality of such unobtrusive devices will be enhanced and the cost of these high-end devices will decrease. In this thesis, we explain the procedure to develop a stress level detection system for everyday life. We present the current performance of state-of-the-art data reprocessing techniques, feature extraction tools and machine learning algorithms in this application area.

Getting rid of these problems may be possible with visual or audio feedback. Thanks to the data collected with this thesis, our stress detection system can be used to motivate the participants to see their stress levels and they can provide more physiological data to train our system. During this thesis, we did not provide any feedback to the participants about their stress levels and other information about their mental health. Providing these pieces of information will help the system to collect a mass amount of data from all smartwatch users. This system can be implemented on other commercially available systems and embedded in one platform. This platform may play a big role in the mental health care system. Stress reduction is not discussed through this thesis and we only focused on stress detection. In addition to this system, stress reduction support system can be added as an extension. As a future work, our research group can create such a platform and mass populations can benefit from.

REFERENCES

1. Cohen, S., R. C. Kessler and L. U. Gordon, *Measuring stress: A guide for health and social scientists*, Oxford University Press on Demand, 1997.

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

3. Walford, N., J. Phillips, A. Hockey and S. Pratt, "Assessing the needs of older people in urban settings: integration of emotive, physiological and built environment data", *Geo: Geography and Environment*, Vol. 4, p. e00037, 01 2017.

4. Chalabianloo, N., D. Ekiz, Y. S. Can and C. Ersoy, "Smart watch based stress detection in real life", *11th International Symposium on Health Informatics and Bioinformatics*, p. 39, Antalya.

5. *BIOPAC*, 2018, `https://www.biopac.com/product/upgrade-to-mp36-system/`, accessed at December 2018.

6. *Empatica*, 2018, `https://www.empatica.com/`, accessed at December 2018.

7. *Samsung*, 2018, `https://www.samsung.com/`, accesed at December 2018.

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

9. Fink, G., *Stress: Concepts, Definition and History*, 01 2017.

10. Chuah, S. H.-W., P. A. Rauschnabel, N. Krey, B. Nguyen, T. Ramayah and S. Lade, "Wearable technologies: The role of usefulness and visibility in smartwatch adoption", *Computers in Human Behavior*, Vol. 65, pp. 276 – 284, 2016.

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

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

13. *Understanding the stress response*, 2018, http://www.health.harvard.edu/, accessed at December 2018.

14. 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.

15. 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.

16. 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.

17. 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.

18. Scholz, U., R. L. Marca, U. M. Nater, U. E. I. Aberle, R. Hornung, M. Martin and M. Kliegel, "Go no-go performance under psychosocial stress: Beneficial effects of implementation intentions.", *Neurobiology of Learning and Memory*, 2009.

19. Miller, R. and C. Kirschbaum, *Trier Social Stress Test*, pp. 2005–2008, Springer New York, New York, NY, 2013.

20. Ollander, S. and C. Godin, ""A comparison of wearable and stationary sensors for

stress detection", *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2015.

21. Ekman, P., "Differential communication of affect by head and body cues.", *Journal of Personality and Social Psychology*, Vol. 2, No. 5, pp. 726–735, 1965.

22. Montepare, J., E. Koff, D. Zaitchik and M. Albert, *Journal of Nonverbal Behavior*, Vol. 23, No. 2, pp. 133–152, 1999.

23. Ekman, P., R. W. Levenson and W. V. Friesen, "Autonomic nervous system activity distinguishes among emotions.", *Science*, Vol. 221 4616, pp. 1208–10, 1983.

24. Kappeler-Setz, C., *Multimodal emotion and stress recognition*, ETH Zurich, 2012.

25. Grings, W. W. and A. M. Schell, "Magnitude of electrodermal response to a standard stimulus as a function of intensity and proximity of a prior stimulus.", *Journal of Comparative and Physiological Psychology*, Vol. 67, No. 1, pp. 77–82, 1969.

26. Lamichhane, B., U. Großekathöfer, G. Schiavone and P. Casale, "Towards Stress Detection in Real-Life Scenarios Using Wearable Sensors: Normalization Factor to Reduce Variability in Stress Physiology", K. Giokas, L. Bokor and F. Hopfgartner (Editors), *eHealth 360°*, pp. 259–270, Springer International Publishing, Cham, 2017.

27. Postolache, O., P. S. Girão, E. Pinheiro and G. Postolache, *Unobtrusive and Non-invasive Sensing Solutions for On-Line Physiological Parameters Monitoring*, pp. 277–314, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

28. Postolache, O., P. S. Girão and G. Postolache, *Pervasive Sensing and M-Health: Vital Signs and Daily Activity Monitoring*, pp. 1–49, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

29. Cegarra, J. and A. Chevalier, "Theoretical and Methodological Considerations in

the Comparison of Performance and Physiological Measures of Mental Workload", D. Harris (Editor), *Engineering Psychology and Cognitive Ergonomics*, pp. 264–268, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

30. Arnrich, B., C. Setz, R. L. Marca, G. Tröster and U. Ehlert, "What Does Your Chair Know About Your Stress Level?", *IEEE Transactions on Information Technology in Biomedicine*, Vol. 14, 2010.

31. Henelius, A., K. Hirvonen, A. Holm, J. Korpela and K. Muller, "Mental workload classification using heart rate metrics", *Conf Proc IEEE Eng Med Biol Soc*, 2009.

32. Hjortskov, N., D. Rissén, A. K. Blangsted, N. Fallentin, U. Lundberg and K. Sogaard, "The effect of mental stress on heart rate variability and blood pressure during computer work.", *European Journal of Applied Physiology*, 2004.

33. Morris, M. and F. Guilak, "Mobile heart health: project highlight", *IEEE Perv Comput*, 2009.

34. Cinaz, B., B. Arnrich, R. Marca and G.Tröster, "Monitoring of mental workload levels during an everyday life office-work scenario", *Personal Ubiquitous Comput.*, 2013.

35. 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.

36. Giles, D., N. Draper and W. Neil, "Validity of the Polar V800 heart rate monitor to measure RR intervals at rest", *European Journal of Applied Physiology,*, 2015.

37. Board, E. M., T. Ispoglou and L. Ingle, "Validity of Telemetric-Derived Measures of Heart Rate Variability: A Systematic Review", *Journal of Exercise Physiology*, 2016S. Ollander, C. Godin, A. Campagne, S. Charbonnier, "A comparison

of wearable and stationary sensors for stress detection", 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Budapest, 2016, pp. 004362-004366.

38. Ollander, S., C. Godin, A. Campagne and S. Charbonnier, "A comparison of wearable and stationary sensors for stress detection", *IEEE International Conference on Systems*, 2016.

39. Jaques, N., S. Taylor, A. Azaria, A. Ghandeharioun, A. Sano and R. Picard, "Predicting students' happiness from physiology, phone, mobility, and behavioral data", pp. 222–228, Sept 2015.

40. 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.

41. Ciabattoni, L., F. Ferracuti, S. Longhi, L. Pepa, L. Romeo and F. Verdini, "Real-time mental stress detection based on smartwatch", *2017 IEEE International Conference on Consumer Electronics (ICCE)*, pp. 110–111, Jan 2017.

42. 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.

43. Carifio, J. and R. J. Perla, "Ten Common Misunderstandings, Misconceptions, Persistent Myths and Urban Legends about Likert Scales and Likert Response Formats and their Antidotes", *Journal of Social Sciences*, Vol. 3, No. 3, pp. 106–116, mar 2007.

44. 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.

45. 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.

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

47. 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.

48. 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.

49. Reimer, U., E. Maier, S. Streit, T. Diggelmann and M. Hoffleisch, "Learning a Lightweight Ontology for Semantic Retrieval in Patient-Centered Information Systems", *Int. J. Knowl. Manag.*, Vol. 7, No. 3, pp. 11–26, Jul. 2011.

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

51. 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.

52. 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.

53. *Tizen*, 2018, `https://www.tizen.org/`, accessed at December 2018.

54. Ekiz, D., G. E. Kaya, S. Buğur, S. Güler, B. Buz, B. Kosucu and B. Arnrich, "Sign sentence recognition with smart watches", *2017 25th Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, May 2017.

55. Akyazi, O., S. Batmaz, B. Kosucu and B. Arnrich, "SmokeWatch: A smartwatch smoking cessation assistant", *2017 25th Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, May 2017.

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

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. 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.

59. Taylor, J. N. C. W. F. S. S. A., S. and R. Picard, "Automatic Identification of Artifacts in Electrodermal Activity Data", *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Vol. 7, p. 1934, 2015.

60. 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.

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

62. Tarvainen, M. P., J. P. Niskanen, J. A. Lipponen, P. O. Ranta-aho and P. A. Karjalainen, "Kubios HRV — A Software for Advanced Heart Rate Variability Analysis", J. Vander Sloten, P. Verdonck, M. Nyssen and J. Haueisen (Editors),

4th European Conference of the International Federation for Medical and Biological Engineering, pp. 1022–1025, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

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

64. Peng, H., F. Long and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy", *IEEE Transactions on pattern analysis and machine intelligence*, Vol. 27, No. 8, pp. 1226–1238, 2005.

65. 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.

66. *INZVA algorithmic summer camp*, 2018, `https://inzva.com`, accessed at December 2018.

67. 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.